Dr. Omar Torres
Associated Editor
Atmospheric Measurement Techniques

Re: "Machine learning as an inversion algorithm for aerosol profile and property retrieval from multi-axis differential absorption spectroscopy measurements: a feasibility study."
Amt-2019-368

Dear Dr. Torres,

Please see attached a revised version of Amt-2019-368, with the changes tracked by the Microsoft Word. We have responded to the comments from the two reviewers. Reviewers' comments are in bold italics; our responses are in regular Times font.

Thank you for your consideration.

Best Regards,
Yun Dong, Elena Spinei and Anuj Karpatne

Responses to the comments by Reviewer 1.

Changes recommended:

(I) The weight of the manuscript needs to be on the ML approach, this is currently not the case.

We have removed some details about the aerosols and have added introduction to machine learning to section 1. (lines 37 to 137 in the new version)

(II) Normally, for ML, the data is split into three sets: (1) a training dataset (2) an evaluation dataset used during training to identify when the training results in overfitting (3) a completely new set of data for testing. ... The authors seem to have only used a validation data set (25% of the total data set) for the testing but no proper testing with parameters outside the training range (so not only "not this specific combination") was performed.

We thank the reviewer for bringing this point on the correctness of our evaluation setup and we would like to clarify that no part of the test data was used in any way during training, thus ensuring the validity of our test results in representing the performance of our ML model on samples outside the training set. Note that in any supervised ML experiment, it is very important to ensure that there is no overlap between the training and test sets, so that the performance on the test set is a true indicator of generalization performance, i.e., the performance on "unseen" instances never seen before during training (also known as out-of-sample instances). This is generally done by holding off a fraction of the overall data during training, thus partitioning the overall data into two sets: a "training set" used only for model building, and a "test set" used only for model evaluation (for further details on evaluating supervised ML models, see Tan et al., 2018 and Fiedman et al., 2001). A common approach for partitioning the overall data into training and test sets is to consider random sampling, also known as the random holdout method [ak1]. In our experiments, we randomly partitioned the overall data into a training set (comprising of 75% instances) and a test set (comprising of 25% instances). Further, an optional procedure that is sometimes followed is to hold a certain fraction of the training set as the "validation set" and monitor the performance on the validation set during training to either avoid overfitting or to tune the hyper-parameters of the ML model. Since the validation set is used during model building (although indirectly) it is no longer considered as a representation of "unseen" instances, and hence, the validation performance is not a true indicator of generalization performance. Note that in our work, we did not make use of any validation set, as the values of all hyper-parameters in our ML model were kept constant across all experiments. Instead, we only report our results on the test set that was not used during training, either directly or indirectly.

We have added the following text to the revised paper to address this comment: *Section 4 (lines 315 to 318 in the new version):* "ML algorithm was trained on 75% randomly selected measurement simulations (1094400 samples) and model performance was tested on the remaining 25%. Note, that no validation data was held off from the 75% training set for tuning hyper-parameters of our ML model, as all ML hyper-parameters were kept constant across all experimental settings in this paper."

Section 5 (lines 378 to 385 in the new version): "We trained the model on 75% of the dataset for 124 epochs with a batch size of 640. The following choice of hyperparameters was used: choice of optimizer=RMSprop, lr=0.001, rho=0.9, epsilon=None, and decay=0.0. We did not perform any hyper-parameter tuning on a separately held validation set inside the training set, and the values of all hyper-parameters in our ML model were kept constant throughout all experiments in the paper on the test set. In order to ensure that there was no overlap between the training and testing steps, we did not make use of the test data either directly or indirectly during the training phase, either for learning parameter weights or selecting hyper-parameters."

General comments:

(1) There is a lengthy (and maybe not super accurate, see below) description on the aerosol phase function and the asymmetry parameter, both in the introduction and in Sect.4. However, there is no information on ML in the intro. This does not at all reflect the title. After all, this manuscript claims to be about the ML as inversion algorithm. Suggestion: Bundle the aerosol information from here and from Sect. 4 in the section about training/validation/test data creation (which is currently Sect. 4) and include some paragraph or two on ML use in inverse modelling and general ML.

We've shorten the aerosol part and added general description of ML and detailed description of the ML model. See reply to general comments (1)

(2) I suggest a different ordering: (1) general introduction including advances in ML, aerosol importance in general, current retrieval techniques and why ML should be applied to aerosol retrieval (2) MAXODAS method description (3) Aerosol properties and modelling and forward modelling with VLIDORT (4) Overview of the methodology of the 3 necessary steps (instead of selling it as two steps as done in Sect. 3, where the

first of the two has itself two steps) and a detailed description of the specific ML setup and choice of hyperparameters. (5...) as before

See reply to general comments (1)

(3) While what is written about OEM and parametrized methods is true, most of it is true for ML as well (i.e regarding e.g. the T/P profile). This section paints an overly dark image of OEM and parametrized codes. I think that the main problem with "traditional" methods is indeed the time they take, and this should be clearly (even more clearly) stated, since this is the one huge advantage of using ML. Also, especially around line 136, it gives the impression of full profile retrieval of asy and ssa, while in fact, it is "only" the aod profile and single scalar values asy and ssa valid for all layers.

We have reworded this part (*lines 203-206*):

"Aerosol extinction coefficient profiles are inverted while aerosol single scattering albedo and asymmetry factor are typically assumed based on the co-located AERONET measurements. They also require external information about the atmosphere (e.g. temperature and pressure profiles) that might not be readily available at the measurement time scales, and a priori information that does not typically exist."

(4)

(a) Which backend was used? Tensorflow, Theano? Some other? Why the mentioning of the yupiler notebook? Why was it used at all? Certainly no web-based interactivity is needed? Why wasn't it simply put in a plain python script?

TensorFlow backend was used. We mention Jupyter Notebook just because the code is implemented in Jupyter Notebook. Yes, web-based interactivity is not needed and of course we can use plain python script. We used Jupyter notebooks just for easy sharing of code, analysis of results, and reproducibility of experiments.

(b) CNN is normally used in ML for image recognition, why is it used here? Why is LSTM used? Maybe some intro on recurrent neural networks in general is needed. This seems to indicate … that scans are not considered separately, but as a function of time.... (so a scan from now and then from 10 minutes, not one from here and now and the next one from tomorrow and somewhere else). However, this seems not to fit your introduction and abstract where you very specifically write about a single scan. This is very confusing and needs explanation. Also maybe, you can start with explaining what a SimpleRNN layer is and why this was not chosen?

Different from image recognition in which 2D CNN is usually used, what we use in our model is 1D CNN which is good for capturing features from 1D-sequences. We've also added general introduction of LSTM. As mentioned above, we consider the profile as a sequence, that's the reason we use the LSTM. We do not use the LSTM in a typical way where the input or output sequence is a time series. In our case, it is nothing related to time but a series of partial AOD values at sequential heights. Simple RNN is inadequate to capture long-term memory effects where the inputs-outputs at a given element of the sequence can affect the outputs at another element of the sequence separated by a long interval. Actually we've tried simple RNN and it does not work as well as LSTMs.

We have added the following text to the paper in Section 5 (lines 330 to 340) to address this comment:

"Note, that in our supervised ML formulation, there are sequences in both the input signals and output signals, namely $\Delta AMF^{aerosol}$ sequence and partial AOD sequence, respectively. Further note that the input and output signals used in our problem setting are of very different types and thus have different dimensionalities (e.g., $\Delta AMF^{aerosol}$ takes 16 values at varying VZAs while partial AOD takes 23 values at varying atmospheric layer depths). We thus first apply a 1-dimensional CNN to extract features from the sequence part of the input signals. Note that our input signals are not image-based, which is one of the common types of input data for which CNNs are used. Instead, our input data is structured as a 1D sequence, and the convolution operations of CNN help in extracting sequence-based features from the input signals that are then fed into subsequent ANN components. We also use an LSTM to model the sequence part of the output signals. Note that our data contains no time dimension as we are only working with single scan data. However, it is the sequence-based nature of the output signals that motivated us to use LSTM models for sequence-based output prediction. Furthermore, the dataset we use for training is produced by a physical model (VLIDORT), where the relationship between the inputs and outputs are known."

(b) Why was it decided to split for profile and ssa/asy retrieval?

We split profile and SSA/ASY retrieval because we consider the profile as a sequence (the partial AODs at adjacent layers are related) that needs to be modeled using an LSTM, but the SSA/ASY are scalars that can be modeled using Dense layers. We've tried a lot of architectures and find that combining profile and SSA/ASY as a single output sequence results in inferior performance.

We have added the following text in the paper in Section 5 (lines 346 to 357) to address this comment:

"To extract sequence-based features from MAX-DOAS inputs, a 1-dimensional Convolutional Neural Network (CNN, Fukushima, 1980; LeCun et al., 1999) is first applied on the sequence of inputs (we concatenate $\Delta AMF^{aerosol}$ sequence with SZA and RAA to obtain an 18-length input sequence), which results in a sequence of preliminary hidden features. These preliminary hidden features are then sent to two different branches of 1D-CNN layers that perform further compositions of convolution operators to produce non-linear hidden features for predicting two different types of outputs: (a) scalar outputs: SSA and ASY, and (b) sequence-based outputs: aerosol extinction profile. For the branch corresponding to scalar outputs, the features extracted from 1D-CNN layers are simply passed on to a fully-connected dense layer to produce a two-dimensional output of SSA and ASY. For the branch corresponding to sequence-based outputs, the features extracted from 1D-CNN layers are fed to a Long Short-Term Memory network (LSTM, Hochreiter and Schmidhuber, 1997) to produce a sequence of partial AOD values at varying atmospheric layers."

(c) What were the choices of the hyperparameters? Which batch size was used? Which lr was used for the RMSprop? Where there any drop out layers? Which activation function was used? There is no information on any of the parameters. How many nodes do the layers have?

We’ve added a plot of the detailed architecture of the ML model in the supplement with all the information.
We have also added the following text in the paper in Section 5 (lines 375 to 380) to provide more details about the hyper-parameters of our model:
“RMSprop was chosen as the optimizer and the mean squared error was used as the loss function (Hinton, 2012). We trained the model on 75% of the dataset for 124 epochs with a batch size of 640. The following choice of hyperparameters was used: choice of optimizer=RMSprop, lr=0.001, rho=0.9, epsilon=None, and decay=0.0.”

(5) what happens if the network gets data that is by no means covered by the training data (i.e. completely outside the range in one or more parameters?) What is the effect of measurement noise (also including "noise" from situations that are not 1 dimensional)?

Though the outputs of the test set are not outside the range of the training data, however, the mappings contained in the test set are different. And different combination of SSA/ASY/profile produces different values of radiance. The model hasn’t seen these input values and output combinations of SSA/ASY/profile before. As for the point you mentioned here, there are next steps of our work. ML itself is a technique learning from the statistics of the data. If applying on the dataset which is too different from the training set, of course with high probability it cannot provide reliable predictions. The more the ML model ‘see’, the better it works. Thus, we need to include more realistic aerosol inputs and radiative transfer simulations as mentioned in the ‘Conclusion and Future Work’ section. We will also consider noise in future work. For this work, our key point is the ‘feasibility’, which aiming at demonstrating that it is feasible to use ML technique into MAX-DOAS aerosol retrieval.

Specific comments:

(1) page 1, line 23 "... and have relative short lifetimes..." –> relative to what? Also, few minutes to few weeks spans about 5 order of magnitude in time, while one end of this span can be considered as short, the other cannot really. Please specify "relative".

- we replaced “relatively short” with “variable”.

(2) page 1, line 26: apart from all the properties already listed, what else do you mean with "physical properties" as opposed to optical? This is very unclear.

- we removed this sentence since it did not add any new information: “The aerosol classification depends on the aerosol source, composition, size and number distribution, aging processes, and optical and physical properties.”

(3) page 1, line 28 "The spatial and temporal distribution of aerosols ... is greatly affected by ... the type of aerosols". I think this is incorrect, the correct verb here is "depends on".

- we replaced “is greatly affected by” for “greatly depends on”

(4) page 2, line 39–40: If you put this statement, then you need to explain more. I also cannot see any connection of this statement to the rest of the paper. The minimum that should be added is how it depends on the surface albedo.

- to shorten the aerosol discussion in introduction we removed line 37 – 63 on p.2.

(5) page 2, line 41–42: "escpecially of anthropogenic origin" "of"? or "for"? This sentence does not make too much sense like it is, reformulation needed.

- to shorten the aerosol discussion in introduction we removed line 37 – 63 on p.2.

(6) page 2, eq2 and eq3: I would think that the range of the asymmetry parameter as such depends on the normalization of the phase function, so you need to have integral(phase function) over 3D angle = 1. If so, then the first moment <cos theta> is the asymmetry btw. forward and backward scattering. So with this, would you not have a factor of 1/4pi missing in the HG phase function? Maybe you could check the normalization factors for consistence btw. g and P.

- to shorten the aerosol discussion in introduction we removed line 37 – 63 on p.2.

(7) page 2, eq. 4: You seem to use tensor notation to make a difference btw. covariant and contravariant tensors and apply Einstein summation convention. However, you still put the summation sign, but without indicating what you are actually summing over.

- to shorten the aerosol discussion in introduction we removed line 37 – 63 on p.2.

(8) page 4, line 101: "approximately known"? Please clarify.

- we added (e.g., temperature and pressure profiles from atmospheric sounding or models)

(9) page 5, around line 136: Since it was highlighted before that

- not sure how to interpret this comment

(10) page 5, line 153..154: both input and output states run to N, one of them should have a different limit, maybe... M? Otherwise it is confusing, especially because it is written that x has 67 layers, but y has "only" 16 angles.

- we replaced y number of elements with M

(11) page 7, line 196: Although VLIDORT has as direct input the viewing zenith angle, most people in the MAXDOAS community are more familiar with the elevation angle. Maybe it is an idea to change this to make it easier to connect to.

- we agree that "elevation angle" is a more familiar term but the MAX-DOAS community is well aware of the zenith angle definition.

(12) page 8, line 199, 201, and other listings in the text of parameters that are summarized in the Table 1: I do not think that they need to be repeated, I think it is enough if they are in the table.

- we replaced the exact listing with the following: "… and nineteen viewing zenith angles between 0 and 89^{o} (see Table 1). To ensure that the training dataset contains all observation geometries feasible for MAX-DOAS sky scans we have included: nineteen relative azimuth angles (0 to 180^{o}, 10^{o} step), and twelve solar zenith angles (0 to 85^{o}, 89^{o} see Table 1)."

(13) page 8, table1: Can you comment on how the direct sun cases for raa=0, sza=vza are handled?

- it is not handled in any special way. We do recognize that no meaningful profile information is available from such geometries and the forward scattering has large uncertainties.

(14) page 9, line 223: why do you need ozone absorption?

- strictly speaking we do not need ozone absorption, but since there is no harm in its presence we left it in.

(15) page 9, line 230: maybe a small sketch to explain the aerosol profile parameters (with the two components of the profile) would be helpful

- we added: "Figure 11 demonstrates the aerosol profile samples, where the near surface aerosol partial optical depth profiles are described by the exponential function and the layers aloft are described by the Gaussian function with various widths and heights added to the exponential function profile."

(16) page 9, line 237: The 25% were fixed between the 20 realizations, or not? It would be really good to see some plots here of the evaluation loss as a function of epoch. Also, please comment on how over-fitting was mitigated.

- we did not perform the hyper-parameter optimization in a formal way, so no cross evaluation was done. However, we did monitor training loss and it converged. To eliminate the confusion, we have replaced "evaluated" with "performance tested".

(17) page 9, line 236: this height is the middle height or the height of the upper boundary? This is not clear.

- we replaced layer "heights" with "depths"

(18) page 9, line 247ff: I would certainly describe the architecture of the network here, not only the Fig. 3. Also, dense layers are not explained. Also, how many nodes in the layers? Do you use maxpooling layers btw. your conv1d layers? What is the size of your convolution window? And again, how was the architecure chosen? Why does it make sense to separate the SSA and the ASY the way you do? Do you extract the SZA and RAA as well? They should certainly be == the input? Is there a test on this?

- We've added a plot of model architecture in the supplement.

(19) page 9, line 259: While you do use 25% for test (or do you actually use this for evaluation? Not really, because you use it to test the network. What was used for the evaluation then?)

- Hyper-parameter selection was done offline. We have not seen a significant difference between the hyper-parameter choices for the selected architecture and did not include the cross-evaluation at all during the training so 25% of the data that the model NEVER saw are used for testing the performance of the model.

Because you use the same type of parametrization, this is not a good test. A different, unseen set of data should be used.

- The model never saw the test data so the test is valid

How do features that were not included in the training dataset at all (by all means outside the parameter range) affect the result?

- Ideally, the final MAX-DOAS inversion algorithm based on ML would "see" most of the possible ranges, for example, profiles from the LIVAS database, but optical properties varied across all aerosol types for the same profile so the solution is more reliable.

What about thin cloud layers above 4 km, do they affect the result?

- Friess et al 2018 used NASA real-time aerosol retrieval algorithm that is the basis for the dAMF (-dAMF= AMF – $AMF_{Rayleigh}$) analysis in this study. It was shown that the method is not sensitive to the aerosol/cloud layers above 4 km. We assume the same applies to the ML-based algorithm.

The tests included here are not very useful.

- We disagree with this statement. Most studies evaluating the performance of MAX-DOAS algorithms (e.g. Fiess et al, 2018) have significantly simpler and smaller data sets from both profile variability, observation geometry and optical properties that were tested in this study.

(20) page 10, line 275 & 285: given the range of parameters, using eq. 270, the maximum error is about 20%, not 100%. This puts these low numbers in context.

- This is assuming that the ML-based algorithm always retrieves the ranges of the training data set. The fact that the ranges were within the realistic roam of aerosol profiles and properties is not a weakness, but a goal for a robust inversion ML algorithm.

(21) page 11, Fig. 5: When you wrote earlier that the mean error is "-0.14", you really just took the mean over all angles? What is the significance of this? If it were in have the parameter space +50% and in the other -50%, its mean is still 0... and the model really not good, so what is the significance of the mean error here?

- We agree, that as a stand-alone mean error over all observation geometries and all aerosol scenarios is somewhat meaningless unless the goal is to detect any systematic biases. That is why we also show 2 standard deviation results and dependency on observation geometry.

(22) page 12, Fig. 7: please explain the box whisker plot. Is the line mean or median? The box is how much percentage? The whisker? There are different conventions...

- we added the following text to the Figure 5 caption: "The central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol."

(23) page16, line351: This paper does not present the ML-based algorithm. It presents some of the results and that's it. There is not enough information on the ML model. This sentence is not summarizing the paper.

- we have made modifications to expend the discussion of the ML-based algorithm

(24) page 16, line 363: maybe this is because of the choice of training parameters as a linear distributed AOD?

- while the AOD itself is linearly distributed the dAMF used in training is not. The more realistic reason for the lower accuracy for low AOD is probably a smaller signal in dAMF.

Technical corrections and suggestions:

1. ***Many times, there are definite articles missing (e.g. page 3 line 65 "The MAXDOAS..", page 3 line 84: "The DOAS technique", page 4 line 91 "The offset term...")***

- thank you for pointing this out

(2) Eq. 5 on page 4 is not referred to in the text.

- we added a reference to (Eq. 5) on line 90.

(3) page7, line204: I highly doubt that Clemer et al 2010 is the only code here. I would add a "e.g.".

- added

(4) page 9, line 229: I think you miss the AOD=0 case in this list

- we assumed that the algorithm will retrieve the properties perfectly in the absence of noise and AOD = 0 (dAMF = 0) and did not want to skew the data. However, we will include very small AOD in the next version of the model with the real data.

(5) page 10ff: I suggest to use an equi-distant grid for the raa-sza plots, as they are now, it gives a biased impression to the eye.

- we replaced Fig 4.

(6) page 12, line 311 f: I cannot quite understand the sentence "This error contribution..." maybe you can reformulate

- the phrase was replaced with "Layer partial AOD retrieval error relative to the total AOD"

References

P. Tan, M. Steinbach, A. Karpatne, and V. Kumar "Introduction to Data Mining (2^{nd} Ed.)," Pearson Addison–Wesley, ISBN-13: 978-0133128901, 2018.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1, no. 10. New York: Springer series in statistics, 2001.

Responses to the comments by Reviewer 2.

1. Not enough information about the machine learning (ML) algorithm itself are given. The introduction focuses on aerosols and MAX-DOAS only without introducing ML. In section 5, there is no explanation why the individual ML steps were chosen as they are.

- We have removed some details about the aerosols and have added the following section to the introduction: See response to comment 1 by Reviewer 1.

2. The validation section appears to be insufficient to assess the performance of the algorithm:

(a) Why not changing the testing dataset to realistic profiles which are not included in the training data? How can you be sure that you do not over-fit your results?

-The mappings contained in the test set are different from those in the training set. And different combination of SSA/ASY/profile produces different values of radiance. The model hasn't seen these input values and output combinations of SSA/ASY/profile before. We split the data into two sets, the training set and the test set, no automatic model selection process using validation set. The training loss converges. We use 75% of the entire dataset for training and then directly apply the model on the remaining 25% for testing. We use ReLU unit (through sparsity) and Max Pooling layer (through reducing parameters) to control overfitting.

(b) Why not using larger aerosol loads?

- Lower to medium AOD loadings are more common. We will use LIVAS data base in future studies with more realistic profiles and global AOD loadings

(c) Why did you use 16 different elevation angles for the testing dataset even though this number is much too high for most measurement locations? What happens if you just use 8 or 10 elevation angles? Does the algorithm still perform well?

- The main goal of the current work is to evaluate feasibility of ML as a retrieval method. We have included more angles than typical for MAX-DOAS to explore the maximum information content of the measurements and the inversion method. We have also tried a scan with10 angles and the ML methods performed well.

(d) The training dataset was created by using an US standard atmosphere. This is mostly a poor representation of the true atmosphere. What happens if the conditions change?

We have not evaluated the effect of temperature and pressure profiles on the retrieval at this stage. This will be done in the future studies

(e) Why not testing the algorithm on real data?

We do not have real data at 360 nm with accurate partial AOD profiles, ASY and SSA retrieved in the same direction as MAX-DOAS pointing.

3. Are there plans to extend the training dataset by more wavelengths, SSA/asymmetry factors, albedo, profiles, trace gases (as suggested in Sec. 7)?

- Yes

4. Note that many articles are missing in the manuscript.

- we have gone through the references

Specific comments

P2, L36-42: There is no need to show the equation of SSA and its detailed description. I suggest to change those 6 lines into one sentence only.

- we have removed details about SSA from the manuscript

P2, L43-62: This part about the aerosol phase functions is much too long, especially since there is no further discussion about this topic in your paper. The lines about the Legendre expansion could be completely removed without loosing important information for the understanding of the manuscript.

- we have removed details about ASY from the manuscript

P2, L58: When kept, please change the index L of PL(cosθ) P3, L65: "The" MAX-DOAS

- we have removed details about ASY from the manuscript

Figure 1: A single sky scan...

- corrected

P3, L84: "The" DOAS technique...

- corrected

P4, L91: "The" offset term...

- corrected

P4, L101: "Forward model parameters that are considered approximately". I guess you are referring, among other parameters, to a priori knowledge when using "approximately" here? Please change the wording or reformulate this sentence.

- this is reference to such parameters as temperature and pressure profiles. Clarified in the text now: (e.g., temperature and pressure profiles from atmospheric soundings or models).

P4, L114-115: "A priori information about...". I find this sentence to be rather confusing. What do you want to say here?

- we added two commas to improve readability: "…distribution, before the measurements are made…"

P5, L123-125: "None of the algorithms perform perfectly". That depends on what you understand as "perfect". I don't think that it is possible at all to retrieve the true atmosphere in a extremely high vertical resolution. However, as far as I know, the second part of this sentence is correct. I would suggest to reformulate the first part.

- We removed this sentence.

Note that you also used external information about the atmosphere for creating your training dataset. You directly applied a priori knowledge by using exponentially decreasing profiles including Gaussian's for your dataset. And I would not say that a priori knowledge does not exist. You look out of the window and know that it is a hazy day so you adapt the a priori. Sometimes you know about local sources or have ancillary measurements available. I would strongly suggest to change this paragraph.

- a priori information, as applied in MAP, is typically a climatological distribution of the parameters of interest so adjusting the a priori according to the observation at the moment is not an appropriate use of the technique. The only technique that is available to measure aerosol extinction coefficient vertical profiles at 355 nm is LIDAR but even in this case the aerosol property information is very limited, so we do not agree that the true a priori for AOD, ASY and SSA at 360 nm exists for many locations. However, since there might be some locations where some of this information is available we have replaced the sentence with: "They also require external information about the atmosphere (e.g. temperature and pressure profiles) that might not be readily available at the measurement time scales, and a priori information might not exist'

P7, L176: AMF represents → An AMF represents

- changed

P7, L178: observations → observation

- changed

P7, L181: Where "the" vertical column density

- changed

P7, L183-191: This part could be shortened as you have already introduced aerosols before.

- we removed some of the introductory aerosol information.

P7, L205: "The" VLIDORT code

- changed

Table 1: Why do your Gaussian profiles don't have center heights higher than 2km? Since the vertical sensitivity for higher altitudes is an issue for common algorithms, I am wondering if your algorithm performs better here?

- in our opinion the limiting factor for the retrieval of elevated layers is the MAX-DOAS approach not the specific algorithm itself. By subtracting the zenith AMF we are removing some information. This information is also reduced by typically considering only a "single species" (O_4) at a "single wavelength". Because of this it made sense to limit the tests to 2 km. Further studies will include actual measurements compiled in LIVAS database.

What is the scaling height of your exponential functions?

- they recalculated depending on the total loading and partitioning between the Gaussian and exponential AOD.

P9, L235: What was the reason for changing the grid step width to a coarser resolution for higher altitudes? I have the feeling that your choice of Gaussian profile center heights and retrieval grid steps might deteriorate retrieval results for higher altitudes (as indicated in Fig. 9 and 11).

- Yes, this is correct. Since this is a feasibility study we first wanted to demonstrate that the method works before performing much more elaborate RT and retrieval modeling. We plan to expand the study including higher vertical resolution.

Section 5: I think it would be nice to add more information to this section to explain also the in-between steps and parameters of your CNN and LSTM.

- we have added more details about the ML in the text and supplement

P9, L251: "RMSprop was chosen...". Please explain.

- It is a consensus in the CS community that RMSprop works well on recurrent networks such as LSTM, but RMSprop is an unpublished optimization algorithm.

Figure 4: It would be interesting for the reader to see a similar plot describing the profile shape distribution. You could show 3 more plots for different partitioning, showing Gaussian center heights and width on x and y axis, respectively. Furthermore, the number of profiles with a certain total AOD would also be interesting (especially when looking at Figure 7). I fear that the reader might loose the connection to the actual profile shape due to the rather statistical analysis in the following paragraphs.

- while this is an interesting information we feel it does not add any additional insight into the results;

Figure 5, 6 and 8: It is interesting to see that mean error and standard deviation show areas with high or low values at certain geometries. I was wondering if this is a matter of the scattering angle (angle between incident and outgoing photon assuming single scattering)? Could you please create a plot showing the scattering angle versus the respective
error/standard deviation? Since e.g. RAA = 30◦ and a low SZA is equivalent to a large scattering angle (e.g. Fig 5c) it might show issues for certain scattering geometries. In addition, you could check if there are certain profile shapes or aerosol parameters more frequent for areas with a large standard deviation or high mean errors compared to other geometries. I was also wondering about the outliers in all three histograms. Any reason for that?

- Thank you for the recommendation! Since Henyey-Greenstein approximation has a poor representation of the forward and backward scattering we will apply the suggested analysis to the future more realistic aerosol modeling.

P11, L292: I agree that OEM methods also struggle with data inversion measured at small RAA but I was wondering why your synthetic analysis fails?

- We believe this is due to the RT at small RAA, where the photon paths are very “direct” and MAX-DOAS is not “benefitting” from the low elevation angles as much.

P12, L295: "The total AOD retrieval..." or "The retrieval of total..."

- changed

P12, L297: In general, "the" ML algorithm

- changed

P12, L299: What is the reason for the second peak in the histogram?

- *we do not know*

Figure 7: Please explain all depicted quantities (mean, median, percentiles...) in the caption of this figure.

- added: “The central mark indicates the median, the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol.”

Here, it would also be interesting to see if the largest underestimations correspond to certain profiles or parameters.

- figure 11 shows some of the worst cases that correspond to low AOD, RAA <= 10^o and large SZA >= 80^o.

Figure 9: Why is the error larger for 1.5 km than for 2 km? Since you also included Gaussian's with Peak heights around 2km, I would expected the largest error at higher altitudes. Especially when considering the higher sensitivity of MAX-DOAS measurements for aerosol loads closer to the surface (which can be seen for altitudes lower than 1.5km).

- This is potentially an artifact of the layer depths changes at 1 and 3 km

Figure 10: It appears that there is also an underestimation of the predicted AOD for all sub-figures with true partial AOD's larger than 0.2. For example in the upper left subfigure, but also in the second row (first figure), the third row (2nd and 3rd fig). Do these underestimations correspond to problematic scenarios/geometries/parameters?

- We have not explored the details.

P16, L353: Training and evaluation of "the" ML

- changed

P16-17, L365-372: Points 1 and 2 are valid but only a demand for near real-time applications. I doubt that there is a need for science to have profiles immediately after the measurement. For point 3, I was wondering if this is a major advantage. The dependence of profiling results on SSA is rather small and the Henyey-Greenstein approximation is in most cases a poor representation of the scattering distribution of aerosols. So why should the reader decide for your algorithm when an AERONET station nearby measures "real" phase functions and SSA?

- this is a feasibility study and by no means suggests that the presented algorithm should be used as is. However, what this study does suggest is that more elaborate RT modeling with more realistic RT settings and realistic profiles can open the possibilities to fast and potentially more accurate algorithms using multiple wavelengths and multiple species.

In point 4 you even diminish the potential of your approach by saying that it might be used as an initial guess for other algorithms. To me, this does not sound as if the authors are convinced of the capability of ML algorithms.

- The goal of this research is not to convert the entire community to use ML-based approach but rather to explore its feasibility and possibilities

If this is true, why not? If not, why are there no strong arguments in favor of your approach?

- We personally believe that physics based ML methods can be very effective in accurate inversions, especially of MAX-DOAS data. However, the quality of training data is very important. Ideally, 3D models should be used with complete physics and exhaustive atmospheric conditions. Availability of such dataset is the part that we are mostly skeptical about. Another issue is the validation of the actual retrievals. There are no other profilers that "sample" in a similar way.

track changes version:

A feasibility study to use machine learning as an inversion algorithm for aerosol profile and property retrieval from multi-axis differential absorption spectroscopy measurements.

Yun Dong[1], Elena Spinei[1], Anuj Karpatne[2]

[1]Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg, VA 24060, USA
[2]Department of Computer Science, Virginia Tech, Blacksburg, VA 24060, USA

Correspondence to: Elena Spinei (eslind@vt.edu)

Abstract. In this study, we explore a new approach based on machine learning (ML) for deriving aerosol extinction coefficient profiles, single scattering albedo and asymmetry parameter at 360 nm from a single MAX-DOAS sky scan. Our method relies on a multi-output sequence-to-sequence model combining Convolutional Neural Networks (CNN) for feature extraction and Long Short-Term Memory networks (LSTM) for profile prediction. The model was trained and evaluated using data simulated by VLIDORT v2.7, which contains 1459200 unique mappings. 75% randomly selected simulations were used for training and the remaining 25% for validation. The overall error of estimated aerosol properties for (1) total AOD is -1.4 ± 10.1 %, (2) for single scattering albedo is 0.1 ± 3.6 %; and (3) asymmetry factor is -0.1 ± 2.1 %. The resulting model is capable of retrieving aerosol extinction coefficient profiles with degrading accuracy as a function of height. The uncertainty due to the randomness in ML training is also discussed.

1. Introduction

Aerosols play an important role in the Earth-atmosphere system by modifying the global energy balance, participating in cloud formation and atmospheric chemistry, and fertilizing land and ocean. Aerosols are widely spread in the troposphere and are emitted by anthropogenic and natural processes (primary aerosols), and are formed by gas-to-particle conversion mechanisms (secondary aerosols). Aerosols are removed from the atmosphere by dry (gravitational settling and turbulent) deposition and wet deposition, and have variable lifetimes ranging from a few minutes to a few weeks (Haywood and Boucher, 2000).

The spatial and temporal distribution of aerosols in the lower troposphere is highly variable and greatly depends on the proximity to the sources, type of aerosols, meteorological conditions, and photochemical processes. Horizontal and vertical heterogeneity of the aerosol distribution, their properties and processes pose a serious challenge for modeling aerosol induced radiative forcing and is an important source of uncertainties in the climate modeling results (Intergovernmental Panel on Climate Change, 2014).

Deleted: relatively short

Deleted: The aerosol classification depends on the aerosol source, composition, size and number distribution, aging processes, and optical and physical properties.

Deleted: is

Deleted: affected by

Macroscopic aerosol optical properties required for modeling aerosol radiative forcing include single scattering albedo, scattering phase function, and aerosol optical thickness (AOD), (Dubovik et al., 2002).

This paper investigates the potential of using advances in machine learning to invert aerosol properties (aerosol extinction coefficient profiles, single scattering albedo and scattering phase function) from a hyperspectral remote sensing technique called multi-axis differential optical absorption.

Machine learning (ML) is a branch of artificial intelligence that derives its roots from pattern recognition and statistics. The goal of ML is to build statistical (or mathematical) models of a real-world phenomenon by relying on training examples. For instance, in supervised ML, a model is first presented with a set of paired examples (termed as the training set), where every training example contains a pair of input variables and output variables, and the goal of ML algorithms is to find the statistical structure of mapping from the input variables to the output variables that match with the training examples and can be generalized to unseen examples (termed as test set). The learned mapping (or the model) can be applied to the inputs of test examples to make predictions on their outputs. There are several advantages of using ML. Firstly, it can sift through vast amounts of training data and discover patterns that are not apparent to humans. Secondly, ML algorithms can have continuous improvement in accuracy and efficiency with increasing amount of training data. Thirdly, ML algorithms are usually very fast to apply on test examples since the time-consuming training process of ML models is offline and one-time. With these advantages as well as the availability of faster hardware, ML has soon become the most popular data analytic technique since the 1990s. In recent years, it has also been applied to the field of remote sensing (Efremenko et al., 2017; Hedelt et al., 2019).

Artificial neural networks (ANN) are methods studied in the ML field, successfully applied to a number of commercial problems such as image detection, text translation, and speech recognition. It is inspired by the biological neural networks constituting animal brains. As an analogy to a biological brain, an ANN is based on artificial neurons. An artificial neuron is a mathematical function receiving and processing input signals and producing outputs signals or activations. Each neuron comprises of weighted inputs, an activation function, and an output. Weights of the neuron are parameters to be adjusted, while the activation function defines the relationship from the input signals to the output signals. When multiple neurons are composed together in a layered manner (where the output signals of neurons in a given layer are used as inputs for the neurons in the next layer), we call it an artificial neural network (ANN). A common algorithm for training ANNs is the backpropagation algorithm, that passes the gradients of errors on the training set from the output layer to inner layers to refine the weights at all layers in an incremental way. The backpropagation algorithm converges when there is no change in ANN weights across all layers beyond a certain threshold. There are several optimization methods that are used for performing backpropagation and are behind standalone ANN packages commonly used by the ML community. ANNs

Deleted: These parameters depend on aerosol chemical composition, aerosol mixing, particle shape and size distribution, and particle orientation.

Deleted: Single scattering albedo, ω(λ), is defined as the ratio of scattering optical depth ($\tau_{scattering}$) to the total optical depth ($\tau_{scattering}$+ $\tau_{absorption}$) at wavelength λ (Eq. (1)): ... [1]

Deleted: (4)

have many different types depending on the specifics of the neuron arrangement or architecture. A simple type of ANN is a multilayer perceptron (MLP), where all neurons at a given layer are fully connected with all neurons of the next layer, also termed as dense layers. Other complex types of ANN include convolutional neural network (CNN) and recurrent neural network (RNN). Two important types of artificial neural networks used in this study are the convolutional neural networks (CNN) (Fukushima, 1980; LeCun et al., 1999) and the Long short-term memory (LSTM) neural networks (Hochreiter and Schmidhuber, 1997), which are variants of recurrent neural networks.

Convolutional neural network (CNN) is a class of deep neural networks that uses the convolution operation to define the type of connections from one layer to another. While they have shown impressive results in extracting complex features from images in computer vision applications (Krizhevsky et al., 2012; Simonyan and Zisserman, 2015), they are relevant in many other applications involving structured input data, e.g., 1D-sequences. A CNN is composed of an input layer, multiple hidden layers and an output layer. The hidden layers usually consist of several convolutional layers, followed by pooling layers, fully connected layers (dense layers) and normalization layers. Figure 1 shows a simple example of CNN. The input vector (or sequence) is first passed through a convolutional layer where it is convolved with 3 filters (convolution kernels) of size 3 using the same padding to produce three 6x1 feature maps. Since the ReLU function $(f(\boldsymbol{x}) = max\,(0, \boldsymbol{x}))$ is commonly chosen as the activation function in CNNs, the feature maps only contain positive values. Then the max pooling layer picks the maximum value every 3 elements for each feature map, generating three 2 x 1 vectors. After passing through a flatten layer, the max pooling output is reshaped into a 6 x 1 vector, which is followed by a dense (fully connected) layer with 2 nodes. The dense layer multiplies its input by a weight matrix and add a bias vector for generating the output of the model. The computer adjusts the model's convolutional kernel values or weights through a training process called backpropagation, a class of algorithms utilizing the gradient of loss function to update weights. For the case in Figure 1, there are 26 tunable parameters. ($(3+1)\times 3 = 12$ from convolution kernels and $(6+1)\times 2 = 14$ from the dense layer.)

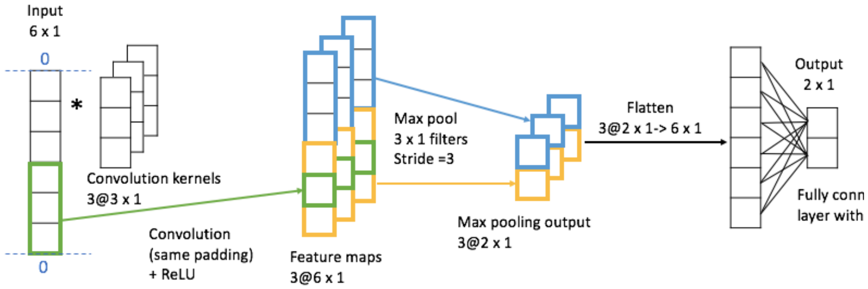


Figure 1. Schematics of a simple CNN

Formatted: Font:Bold

Long short-term memory (LSTM) neural networks have many applications such as speech recognition (Li and Wu, 2015) and handwriting recognition (Graves et al., 2008; Graves and Schmidhuber, 2009). They are a special kind of ANNs termed as recurrent neural networks (RNNs). RNNs are designed for modeling sequence dependent behavior (e.g., in time). They are called "recurrent" because they perform the same operation for every element of a sequence, with the output at a given element dependent on previous computations at earlier elements (Britz, 2015). This is different from traditional neural networks wherein all the input-output examples are assumed to be independent of each other.

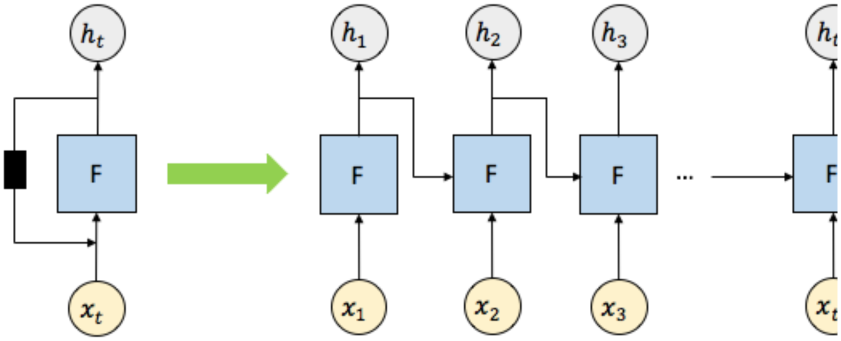


Figure 2. Unrolled recurrent neural network.

Formatted: Font:Bold

Figure 2 shows a diagram of an unrolled RNN with t input nodes, where "unrolled" means showing the network for the full sequence of inputs and outputs. The RNNs work as follows. At the first element of the sequence, the set of input signals x_1 (which can be multi-dimensional) is fed into the neural network F to produce an output h_1. At the next element of the sequence, the same neural network F takes both the next input x_2 and previous output h_1, generating the next output h_2. This recurrent computation continues for t times to produce the output at the tth element of the sequence, h_t. While RNNs are powerful architectures for modeling sequence behavior, classical RNNs are inadequate to capture long-term memory effects where the inputs-outputs at a given element of the sequence can affect the outputs at another element of the sequence separated by a long interval. Long-short-term memory (LSTM) models are variants of RNNs that are able to overcome this challenge and are efficient at capturing long-term dependencies as well as short-term dependencies. It does so by introducing an internal memory state that is operated by neural network layers termed as gates, such as the "input gate," that adds new information from the input signals to the memory state, the "forgot gate," that erases content from the memory state depending on the input signals, and the "output gate," that transforms information contained in the input signals and the memory state to produce output signals.

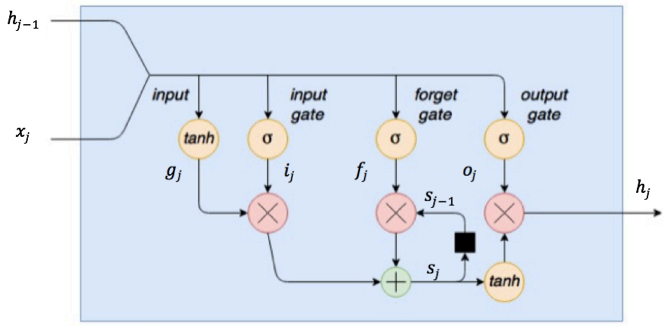


Figure 3. LSTM cell diagram (modified from Thomas, 2018).

An example of an LSTM cell is illustrated in Figure 3, of which the update rules are:

$$g_j = \tanh\,(b^g + x_jU^g + h_{j-1}V^g)$$

$$i_j = \sigma(b^i + x_jU^i + h_{j-1}V^i)$$

$$f_j = \sigma(b^f + x_jU^f + h_{j-1}V^f)$$

$$s_j = s_{j-1} \circ f_j + g_j \circ i_j$$

$$o_j = \sigma(b^o + x_jU^o + h_{j-1}V^o)$$

$$h_j = \tanh\,(s_j) \circ o_j$$

where j is the element index, σ(x) represents the sigmoid function, and tanh (x) represents the hyperbolic tangent function. $x \circ y$ denotes the element-wise product of x and y. U^g, U^i, U^f, U^o are the weights for the input x_j, while V^g, V^i, V^f, V^o are the weights for the other input h_{j-1}, and b^g, b^i, b^f, b^o are the scalar terms (termed as bias). The term g_j is the input modulation gate, which modulates the input $b^g + x_jU^g + h_{j-1}V^g$ by a hyperbolic tangent function, squashing the input between -1 to 1. The term i_j is the input gate, which applies a sigmoid function to its input, limiting the output values between 0 and 1. The input gate i_j determines which inputs are switched on or off when multiplying the modulated inputs ($g_j \circ i_j$). The term s_j is the internal cell state that provides an internal recurrence loop to learn the sequence dependence. The terms f_j and o_j are the forgot gate and output gate, respectively. They have similar function to the input gate i_j, regulating the information into and out of the LSTM cell. The term h_j is the output at step j.

Formatted: Font:(Default) Times New Roman, 10 pt

Deleted:

2. Multi-Axis Differential Optical Absorption (MAX-DOAS) technique

Deleted: .

The MAX-DOAS technique has been widely used to derive vertical aerosol extinction coefficient profiles in the lower troposphere. This is typically done from ground-based measurements of oxygen collision complex (O_2O_2) absorption (for a detailed list of references see Table 1 in Wagner et al., (2018)). Since the oxygen volume mixing ratio ($\chi_{O2} = 0.209$) is considered constant, the O_2O_2 abundance depends only on the total number of air molecules (pressure, temperature and to a small degree humidity) and can be easily calculated. More than 93% of O_2O_2 is located below 10 km (scale height ~ 4 km). Any deviation in measured O_2O_2 absorption from this molecular (Rayleigh) scattering case is only due to the change in the photon path through the O_2O_2 layer. Aerosols and clouds are the main causes of such photon path modification for ground-based measurements. O_2O_2 has several absorption bands in the ultraviolet (UV) and visible (VIS) parts of the electromagnetic spectrum (band peaks at 343, 360, 380, 477, 577, 630 nm (Thalman and Volkamer, 2013).

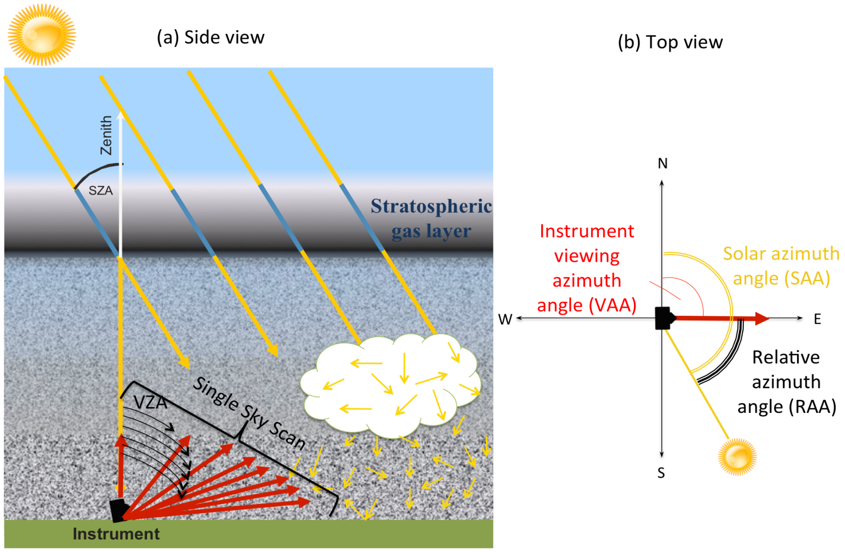

(a) Side view
(b) Top view
Zenith
SZA
Stratospheric gas layer
VZA
Single Sky Scan
Instrument
N
S
W
E
Instrument viewing azimuth angle (VAA)
Solar azimuth angle (SAA)
Relative azimuth angle (RAA)

Figure 4. Demonstration of the MAX-DOAS principle: (a) side view and (b) top view. Simplified photon paths through the atmosphere are shown in yellow. A single sky scan sequence for profile retrieval consists of multiple viewing zenith angles (VZA) in a specific direction (viewing azimuth angle, VAA) at a specific solar zenith angle (SZA) and is shown in red.

Deleted: 1

Deleted: s

The MAX-DOAS technique consists of measuring sky-scattered UV-VIS solar photons at multiple, primarily, low elevation angles (Fig. 4). MAX-DOAS shows a large sensitivity to the tropospheric gases due to increased photon path length through the lower troposphere (Platt and Stutz, 2008). To eliminate the contribution from the upper atmosphere solar spectra measured at low elevation angles are divided by the reference spectrum collected from the zenith direction. The DOAS technique has the advantage of not needing an absolute radiometric calibration.

Deleted: spectra

Deleted: 1

The first step of the DOAS retrieval is a spectral evaluation to calculate the differential slant column density

($\Delta SCD_{measured}$ = SCD - $SCD_{reference}$) of O_2O_2. This step is accomplished through the simultaneous non-linear least-squares fitting of the absorption by species *i*, low-order polynomial function (P_{LO}) and offset to the difference between the logarithms of the attenuated (*I*) and reference ($I_{reference}$) spectra (Eq. 5). P_{LO} estimates combined attenuation due to molecular scattering and aerosol total extinction (scattering and absorption). The offset term approximates instrumental stray light and residual dark current.

Deleted: Offset

$$ln\left(I_{reference}(\lambda)\right) - ln\left(I(\lambda) - offset(\lambda)\right) = \left(\sum_s \sigma_i(\lambda) \cdot \Delta SCD_i\right) + P_{LO}, \qquad (1)$$

Deleted: 5

The second step of the MAX-DOAS analysis is the conversion of a single sky scan (multiple viewing angles) $\Delta SCD(O_2O_2)$ into a vertical aerosol extinction coefficient profile. The physical relationship between the measured ΔSCD and the desired aerosol extinction coefficient profile and aerosol properties is complex, and, in general, can be expressed mathematically by Eq. (6) (Rodgers, 2004):

Deleted: equation

$$\boldsymbol{y} = f(\boldsymbol{x}, \boldsymbol{b}) + \boldsymbol{\varepsilon}, \qquad (2)$$

Deleted: 6

Where, the measured quantities (measurement vector $\boldsymbol{y}$) are described by a forward model $f(\boldsymbol{x}, \boldsymbol{b})$ and the measurement error vector ($\boldsymbol{\varepsilon}$). The forward model, $f(\boldsymbol{x}, \boldsymbol{b})$, is a model that estimates physical processes that relate the measured parameter ($\boldsymbol{y}$), the unknown quantity to be retrieved (state vector ($\boldsymbol{x}$)), and forward model parameters ($\boldsymbol{b}$) that are considered approximately known (e.g., temperature and pressure profiles from atmospheric soundings or models). Under most conditions, there are more unknowns than measurements, and as a result equation (6) does not have a unique solution.

The inversion of Eq. (6) is often done in the framework of Bayes' theorem, which allows for the assignment of probability density functions to all possible states given measurements and prior knowledge of the state. However, in reality, we are not interested in all possible solutions, but rather a single, the most "probable" solution with its error estimation. Equation (7) shows a Transfer Function that defines an estimated solution ($\hat{\boldsymbol{x}}$) as a function of the measurement system and retrieval method (Rodgers, 2004):

Deleted: equation

$$\hat{\boldsymbol{x}} = R\left(f(\boldsymbol{x}, \boldsymbol{b}) + \boldsymbol{\varepsilon}, \hat{\boldsymbol{b}}, \boldsymbol{x_a}, \boldsymbol{c}\right), \qquad (3)$$

Deleted: 7

where *R* is a retrieval method, $f(\boldsymbol{x}, \boldsymbol{b})$ is a forward function with the true state ($\boldsymbol{x}$) and true parameters ($\boldsymbol{b}$), $\hat{\boldsymbol{b}}$ is the estimated forward model parameter vector, $\boldsymbol{x_a}$ is the a priori estimate of state vector ($\boldsymbol{x}$), and $\boldsymbol{c}$ is a retrieval method parameter vector (e.g. convergence criteria). For nonlinear problems the solution to equation (7) cannot be found explicitly, and iterative numerical methods are required. A maximum a posteriori (MAP) approach has been widely applied to moderately nonlinear problems with Gaussian distribution of both measurement errors and a priori state errors. A priori information about the state vector distribution before the measurements are made is used to constrain the solution of the ill-posed problems (Rodgers, 2004). It is essential to use the best estimate of the state available since in the MAP approach the retrieved state is proportional to the weighted mean of the actual state and the a priori state. In addition, an appropriate covariance matrix for the a priori state vector has to be constructed. This a priori information for aerosol vertical extinction coefficient profiles, however, is rarely available.

In addition to the optimal estimation method (OEM), briefly described above, parameterized (Beirle et al., 2019; Vlemmix et al., 2015) and analytic (Spinei et al 2019, in preparation) inversion algorithms were developed. Frieß et al., (2019) provide a detailed intercomparison of currently available state-of-the-art inversion algorithms for the MAX-DOAS measurements. Most of the current algorithms take between 3 to 216 seconds to process a single MAX-DOAS sky scan (Frieß et al., 2019) mainly due to the iterative inversion step. Aerosol extinction coefficient profiles are inverted while aerosol single scattering albedo and asymmetry factor are typically assumed based on the co-located AERONET measurements. They also require external information about the atmosphere (e.g. temperature and pressure profiles) that might not be readily available at the measurement time scales, and a priori information that does not typically exist. With an increasing number of MAX-DOAS 2-D instruments worldwide capable of sunrise to sunset measurements (e.g. Pandonia Global Network) fast methods are needed that can harvest full information from the MAX-DOAS hyperspectral measurements.

Deleted: None of the algorithms perform perfectly and none of them estimate asymmetry factor or single scattering albedo in addition to aerosol extinction coefficient profiles.

Deleted: ,

Deleted: aerosol single scattering albedo and asymmetry factor

Deleted:)

This study describes and evaluates a fast novel machine learning (ML) approach for retrieving aerosol extinction coefficient profiles, asymmetry factor and single scattering albedo at 360 nm from ΔSCD(O_2O_2) observations within a single MAX-DOAS sky scan. The basic idea of our approach is: (1) develop an "inverse model" by one-time offline training of a supervised ML algorithm on simulated MAX-DOAS data and corresponding atmospheric aerosol conditions, and (2) use the relationships derived in the first step to estimate the aerosol extinction profile, asymmetry factor, and single scattering albedo from the MAX-DOAS ΔSCD(O_2O_2) measurements. We specifically leverage recent advances in ML, e.g., deep learning methods, to automatically extract the inverse mapping from the observations ($\boldsymbol{y}$) to the state vectors ($\boldsymbol{x}$), using a collection of ($\boldsymbol{x}, \boldsymbol{y}$) pairs available for training. Different machine learning algorithms were successfully used in remote sensing applications (Schulz et al., 2018, Schilling et al., 2018, Efremenko et al., 2017; Hedelt et al., 2019).

The rest of the paper is organized in the following sections. Section 3 provides an overview of the new retrieval algorithm. Section 4 focuses on training data generation using the radiative transfer model (VLIDORT). Section 5 details ML implementation. Section 6 provides an extensive comparison of ML predicted versus "true" macroscopic aerosol properties outside the training dataset. Section 7 summarizes the findings.

3. Overview of the Methodology

Our approach consists of three stages: (1) training set generation; (2) a one-time training that results in an inverse ML model $R(\widehat{\Theta})$ with appropriate architecture and parameters $\widehat{\Theta}$; and (3) an inversion stage, where the trained ML model $R(\widehat{\Theta}\,)$ is applied to MAX-DOAS measurements to retrieve aerosol properties. Figure 5 provides a schematic overview of the three stages.

First, a training set containing simulated measurements $\{\boldsymbol{y}_i | i = 1,2, \dots, M\}$ is generated by a forward model (VLIDORTv2.7) given atmospheric states $\{\boldsymbol{x}_i | i = 1,2, \dots, N\}$. The model describes atmospheric radiative

Deleted: two

Deleted: stage

Deleted: 2

Deleted: 2

Deleted: both

Deleted: The offline training stage comprises of two key steps.

Deleted: N

transfer processes connecting the atmospheric states and the measurements. Second, both the atmospheric states and the simulated measurements are fed into the ML model for learning the inverse mapping from the measurement space to the state space. This is based on solving an optimization problem that minimizes the mean squared error (MSE) between the retrieved values ($\{\hat{\boldsymbol{x}}_i | i = 1,2,\dots,N\}$) and the true values ($\{\boldsymbol{x}_i | i = 1,2,\dots,N\}$). We specifically chose artificial neural network (ANN) models to learn the inverse mapping from $\boldsymbol{y}$ to $\boldsymbol{x}$. By iteratively adjusting the parameters of the ANN model using gradient descent (backpropagation) algorithms (Johansson et al., 1991), we are able to arrive at ANN model parameters $\widehat{\Theta}$ that provide a local optimum performance in terms of MSE on the training data. The result of the training stage is an inverse model $R(\widehat{\Theta})$ whose architecture and parameters are saved in an HDF5 file (1.3 MB). The trained model $R(\widehat{\Theta})$ is an inversion operator that transforms measurements vector $\boldsymbol{y}$ into the state vector $\hat{\boldsymbol{x}}$ through a set of simple linear and nonlinear operations. The inverse model provides a convenient and fast way for retrieval of aerosol properties from ΔSCD(O_2O_2) measurements during the inversion stage. It takes ~0.15 ms for the retrieval of the studied aerosol properties from a single MAX-DOAS sky scan ΔSCD(O_2O_2) on a single CPU core.

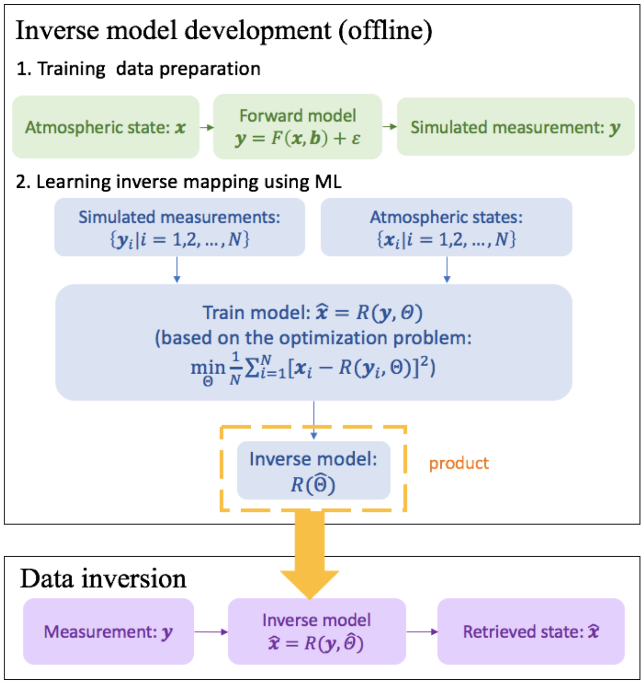

Inverse model development (offline)
1. Training data preparation
Atmospheric state: $\boldsymbol{x}$
Forward model
$\boldsymbol{y} = F(\boldsymbol{x}, \boldsymbol{b}) + \varepsilon$
Simulated measurement: $\boldsymbol{y}$
2. Learning inverse mapping using ML
Simulated measurements:
$\{\boldsymbol{y}_i | i = 1,2,\dots,N\}$
Atmospheric states:
$\{\boldsymbol{x}_i | i = 1,2,\dots,N\}$
Train model: $\hat{\boldsymbol{x}} = R(\boldsymbol{y}, \theta)$
(based on the optimization problem:
$\min_{\Theta} \frac{1}{N} \sum_{i=1}^{N} [\boldsymbol{x}_i - R(\boldsymbol{y}_i, \Theta)]^2$)
Inverse model:
$R(\widehat{\Theta})$
product
Data inversion
Measurement: $\boldsymbol{y}$
Inverse model
$\hat{\boldsymbol{x}} = R(\boldsymbol{y}, \hat{\theta})$
Retrieved state: $\hat{\boldsymbol{x}}$

Figure 5. Schematics of the machine learning inversion algorithm.

Deleted: 2

4. Training data preparation

The success of any ML model depends on the quality of the training data. Since there is no reliable dataset that combines simultaneous MAX-DOAS measurements and observations of aerosol macrophysical properties and vertical extinction coefficient profiles at 360 nm we use a radiative transfer model to simulate MAX-DOAS measurements. In this study, we train our ML model on air mass factors (AMF) calculated from the simulated solar radiances at the bottom of the atmosphere.

AMF represents a ratio between the true average path that photons take through a gas layer before detection by a MAX-DOAS instrument and the vertical path. Since O_2O_2 absorption in the reference (zenith scattered) spectrum is not precisely known, a differential AMF at a specific wavelength λ and observations geometry μ (relative azimuth angle, solar zenith angle, and viewing zenith angle), is determined as:

$$\Delta AMF(O_2O_2,\lambda,\mu) = \frac{\Delta SCD_{measured}(O_2O_2,\lambda,\mu)}{VCD(O_2O_2)_{calculated}} = \frac{\ln\left(I_{reference}(\lambda,\mu_o)\right) - \ln\left(I(\lambda,\mu)\right)}{VCD(O_2O_2)_{calculated} \cdot \sigma(O_2O_2,\lambda)}, \qquad (4)$$

Deleted: 8

Where vertical column density of O_2O_2 (VCD) is estimated as the squared oxygen number density integrated from the surface to the top of the atmosphere; and σ(λ) is the molecular absorption cross-section of O_2O_2.

In the absence of aerosols and clouds only air molecules (mainly oxygen and nitrogen) scatter solar photons in the Earth's atmosphere. This molecular only (Rayleigh) scattering process is considered to be well understood (Bodhaine et al., 1999) and $\Delta AMF^{Rayleigh}$ can be calculated from the simulated intensities. In the presence of aerosols, dust and clouds not only air molecules but also particles and cloud droplets scatter solar photons. This type of scattering can be generally described by the T-matrix theory. In this study we consider only spherical aerosols (Lorenz-Mie theory), whose scattering phase function is approximated according to the Henyey-Greenstein approach using the asymmetry factor *g*. $\Delta AMF^{aerosol+Rayleigh}$ are determined from simulated downwelling radiances for atmosphere with different aerosol types and their extinction coefficient profiles. The change in AMF due to aerosol presence can be described by $\Delta AMF^{aerosol}$:

$$\Delta AMF^{aerosol} = \Delta AMF^{Rayleigh} - \Delta AMF^{aerosol+Rayleigh}, \qquad (5)$$

Deleted: 9

$\Delta AMF^{aerosol}$ for O_2O_2 at 360 nm for different observation geometries and scattering conditions is used for ML training in this feasibility study. A single MAX-DOAS measurement considered here is $\Delta AMF^{aerosol}$ set from the full sky scan at a single solar zenith angle, single relative azimuth angles, and *nineteen viewing zenith angles between 0° and 89° (see Table 1). To ensure that the training dataset contains all observation geometries feasible for MAX-DOAS sky scans we have included: nineteen relative azimuth angles (0° to 180°, 10° step), and twelve solar zenith angles (0° to 85°, see Table 1).* Solar radiances at the bottom of the atmosphere were simulated using VLIDORT v.2.7 (Spurr, 2008). VLIDORT is a discrete-ordinate radiative transfer model that has been successfully applied to simulate radiances and weighting functions for forward models in optimal estimation inversion (e.g., Clémer et al., 2010) and machine learning algorithms (Efremenko et al., 2017, Hedelt et al., 2019). VLIDORT code applies pseudo-spherical approximation to direct solar beam attenuation in a curved atmosphere. All scattering processes are estimated using the plane-parallel approximation in a stratified atmosphere. Precise single scattering computation is performed using Nakajima/Tanaka ansatz and delta-M scaling. VLIDORT v.2.7 calculates analytically derived Jacobians

Deleted: the following viewing zenith angles: 0, 40, 50, 60, 65, 70, 75, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89°. To ensure that the training dataset contains all observation geometries feasible for MAX-DOAS sky scans we have included the following: ... [2]

(radiance weighting functions) with respect to any profile/column/surface variables. VLIDORT computes elastic scattering by molecules to all orders (Spurr, 2008).

Table 1. Radiative transfer model settings

General Model Settings	Physical and Observation Geometry Inputs

NO Refraction correction;
Scalar calculations;
Only elastic scattering;
Aerosol scattering phase function estimation using Henyey-Greenstein approximation from the asymmetry factor (g).
	Observation Geometry:
	Viewing zenith angle scan: 0, 40, 50, 60, 65, 70, 75, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89^o;
	Relative azimuth angles: 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180^o
	Solar Zenith angles: 0, 10, 20, 30, 40, 50, 60, 65, 70, 75, 80, 85, 86, 87, 88, 89^o
	Wavelength: 360 nm;
	Vertical grid (67 layers):
	100 m up to 4 km, 500 m from 4 to 8 km, 1 km from 8 to 12km, 2 km from 12 to 30km, 5 km from 30 to 60 km
	Atmospheric air density:
	Pressure [hPa]: US1976 standard atmosphere
	Temperature [K]: US1976 standard atmosphere
	Gas volume mixing ratio profiles:
	O_3 profile: climatology over Cabauw in September
	O_3 molecular absorption cross-section: Daumont
	O_2O_2 profile: from temperature and pressure
	O_2O_2 molecular absorption cross-section: Thalman and Volkamer (2011)
	Aerosol properties:
	Single scattering albedo: 0.775, 0.825, 0.875, 0.925, 0.975
	Henyey-Greenstein asymmetry factor: 0.675, 0.725, 0.775, 0.825
	Aerosol extinction coefficient profiles [1/km] as a function of altitude;
	Exponential function at the surface combined with “sliding” Gaussian function above;
	Total AOD: 0, 0.15, 0.3, 0.45, 0.6, 0.75;
	Gaussian profile center height: 0.5, 1, 1.5, 2 km;
	Gaussian width: 0.1, 0.2, 0.3, 0.5 km;
	Partitioning between exponential and Gaussian attributed AOD: 0.3, 0.6, 0.9
	Surface reflectivity:
	Lambertian albedo at 0.04

VLIDORT models radiative transfer processes at a specific wavelength in a stratified atmosphere. It requires geometrical and “optical” information about the atmospheric layers and the underlying ground surface. These include layer heights, pressure and temperature at layer boundaries for refractive geometry calculations, solar zenith, viewing zenith direction and relative azimuth angles between the viewing direction and solar position. Each atmospheric layer is described by total optical thickness, total single scatter albedo, and the set of Greek matrices specifying the total scattering law.

VLIDORT simulations were performed for the US 1976 standard atmosphere divided into 67 layers (same as in Frieß et al., 2019) with 0.1 km layers from the surface to 4 km; 0.5 km layers from 4 to 8 km and varying width up to 60 km. Since surface reflectivity has a small effect on ground-based MAX-DOAS measurements we performed simulations only for a single Lambertian albedo of 0.04. Absorption only by two gases was considered in this study: ozone and O_2O_2. Light polarization, direct beam refraction, and inelastic scattering were not included in this study. Table 1 summarizes VLIDORT inputs and general settings.

Aerosol types in this study are described by a single scattering albedo and asymmetry factor combination with total 20 “types”: (1) Single scattering albedo: 0.775, 0.825, 0.875, 0.925, 0.975; (2) Henyey-Greenstein asymmetry factor: 0.675, 0.725, 0.775, 0.825. Aerosol extinction coefficient profiles were generated by combining an exponential function at the surface with a “sliding” Gaussian function above. The aerosol total optical depth was partitioned between the exponential and Gaussian functions. Total AOD cases included 0.15, 0.3, 0.45, 0.6, and 0.75 with exponential to Gaussian partitioning fractions of 0.3, 0.6 and 0.9. The Gaussian function peak center height was varied from 0.5 km to 2 km in steps of 0.5 km. The Gaussian function peak width was varied too: 0.1, 0.2, 0.3, and 0.5 km. This results in 4800 aerosol cases and a total of 1459200 measurement simulations (sky scan). Figure 14 demonstrates the aerosol profile samples, where the near surface aerosol partial optical depth profiles are described by the exponential function and the layers aloft are described by the Gaussian function with various widths and heights added to the exponential function profile. While VLIDORT simulations were performed for an atmosphere divided into 67 layers, ML training was done by resampling onto 23 layers only. The new layer depths are: 100 m from the surface to 1km, 200 m from 1 km to 3 km, 500 m from 3 km to 4 km, and the last layer is 56 km high. The new layer partial AODs were generated by adding the neighboring layer partial aerosol optical depths. ML algorithm was trained on 75% randomly selected measurement simulations (1094400 samples) and model performance was tested on the remaining 25%. Note, that no validation data was held off from the 75% training set for tuning hyper-parameters of our ML model, as all ML hyper-parameters were kept constant across all experimental settings in this paper.

Deleted: heights

Deleted: evaluated

5. Learning inverse mapping using ML

We employ a supervised ML formulation for our problem of aerosol profile retrieval, where the goal is to learn the mapping from input variables to output variables given a training set of paired data instances. In our formulation, every data instance corresponds to a single MAX-DOAS sky scan at a fixed Relative Azimuth Angle (RAA) and Solar Zenith Angle (SZA), where the inputs of the data instance comprise of: (a) RAA scalar value, (b) SZA scalar value, and (c) a sequence of $\Delta AMF^{aerosol}$ values at 16 VZAs. The output variables at a data instance correspond to the aerosol properties we are interested in predicting given the inputs, which are: (a) Single Scattering Albedo (SSA) scalar value, (b) Asymmetry factor (ASY) scalar value, and (c) a sequence of partial Aerosol Optical Depth (AOD) values at 23 vertical layers of the atmosphere, termed as the aerosol extinction profile.

Deleted: al

Deleted: depths

Deleted:

Note, that in our supervised ML formulation, there are sequences in both the input signals and output signals, namely $\Delta AMF^{aerosol}$ sequence and partial AOD sequence, respectively. Further note that the input and output signals used in our problem setting are of very different types and thus have different dimensionalities (e.g.,

$\Delta AMF^{aerosol}$ takes 16 values at varying VZAs while partial AOD takes 23 values at varying atmospheric layers). We thus first apply a 1-dimensional CNN to extract features from the sequence part of the input signals. Note that our input signals are not image-based, which is one of the common types of input data for which CNNs are used. Instead, our input data is structured as a 1D sequence, and the convolution operations of CNN help in extracting sequence-based features from the input signals that are then fed into subsequent ANN components. We also use an LSTM to model the sequence part of the output signals. Note, that our data contains no time dimension as we are only working with single scan data, assuming the atmosphere does not change during the scan time. However, it is the sequence-based nature of the output signals that motivated us to use LSTM models for sequence-based output prediction. Furthermore, the dataset we use for training is produced by a physical model (VLIDORT), where the relationship between the inputs and outputs are known.

Figure 6 illustrates the novel multi-output sequence-to-sequence model for learning the inverse mapping from MAX-DOAS measurements to aerosol optical properties. To extract sequence-based features from MAX-DOAS inputs, a 1-dimensional Convolutional Neural Network (CNN, Fukushima, 1980; LeCun et al., 1999) is first applied on the sequence of inputs (we concatenate $\Delta AMF^{aerosol}$ sequence with SZA and RAA to obtain an 18-length input sequence), which results in a sequence of preliminary hidden features. These preliminary hidden features are then sent to two different branches of 1D-CNN layers that perform further compositions of convolution operators to produce non-linear hidden features for predicting two different types of outputs: (a) scalar outputs: SSA and ASY, and (b) sequence-based outputs: aerosol extinction profile. For the branch corresponding to scalar outputs, the features extracted from 1D-CNN layers are simply passed on to a fully-connected dense layer to produce a two-dimensional output of SSA and ASY. For the branch corresponding to sequence-based outputs, the features extracted from 1D-CNN layers are fed to a Long Short-Term Memory network (LSTM, Hochreiter and Schmidhuber, 1997) to produce a sequence of partial AOD values at varying atmospheric layers.

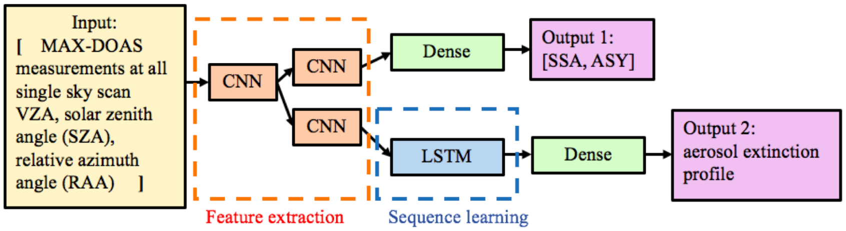

Input:
[MAX-DOAS measurements at all single sky scan VZA, solar zenith angle (SZA), relative azimuth angle (RAA)]
CNN
CNN
CNN
Dense
Output 1:
[SSA, ASY]
LSTM
Dense
Output 2:
aerosol extinction profile
Feature extraction
Sequence learning

Figure 6. Schematics of the multi-output sequence-to-sequence model for deriving aerosol optical properties from MAX-DOAS measurements.

Figure S1 shows the detailed architecture of the multi-output sequence-to-sequence model. The CNNs consist of eight 1D convolutional layers (c_1 to c_8) and four max-pooling layers (p_1 to p_4). For convolutional layers c_1 to c_6, the activation function is the Rectified Linear Unit (ReLU) function. For layers c_7 and c_8, it

Deleted: depth

Deleted:

Deleted: 3

Deleted: A

Deleted: to learn

Deleted: is illustrated in Figure 3

Deleted: The inputs to the ML model comprise of a sequence of $\Delta AMF^{aerosol}$ at 16 VZAs, as well as two scalar inputs: the solar zenith angle (SZA) and the relative azimuth angle (RAA) of the single sky scan.

Deleted: data

Deleted: applies a learnable convolution operator

Deleted: MAX-DOAS values

Deleted:

Deleted: t

Deleted: produce a sequence of hidden features. These hidden features are then fed into the subsequent neural network layers to predict a sequence of partial AOD values, along with two scalars: single scattering albedo and asymmetry factor. Given the sequential nature of the partial AOD outputs, we employ a

Deleted: that is able to capture varying scales of memory in the

Deleted: levels

Deleted:

Deleted: depths

Deleted: The complete ML model is implemented in the Jupyter Notebook by using the Keras library. RMSprop was chosen as the optimizer and the mean squared error as the loss function (Hinton, 2012).

Deleted: 3

Deleted: An example of an

is a hyperbolic tangent function (tanh). We set the kernel size of the convolution operation to be the typical value of 5 and use the same padding for all $c_k, \forall k \in \{1, 2, \ldots, 8\}$. ReLU and Max pooling layers help to reduce overfitting through model sparsity and parameter reduction. The convolution kernel weights are initialized using a "Glorot uniform" method (Glorot and Bengio, 2010).

Extracted feature vector from the p_1 layer is sent into two different branches. In the branch for profile prediction, we take a one-to-many LSTM (Fig. 3) with 23 layer steps and a hidden size of 128 to capture the correlation between the partial AODs at different layers. We simply duplicate the feature vector learned from CNNs for 23 times to generate the inputs for the LSTM model. The sequential output $\{y_1, y_2, \ldots, y_{23}\}$ of the LSTM (after passing through a flatten layer and an ReLU layer) is interpreted as the 23-layer aerosol extinction profile. For the SSA/ASY branch, 1D convolutional layers and dense layers are combined for the prediction. The reason for taking a two-output architecture is that SSA and ASY are independent scalar outputs that cannot be treated as a sequence, in contrast to the aerosol extinction profile.

We implemented our ML model in the Jupyter Notebook using the Keras library, which is a commonly used deep learning library for Python. RMSprop was chosen as the optimizer and the mean squared error was used as the loss function (Hinton, 2012). We trained the model on 75% of the dataset for 124 epochs with a batch size of 640. The following choice of hyperparameters was used: choice of optimizer = RMSprop, lr = 0.001, rho = 0.9, epsilon = None, and decay = 0.0. We did not perform any hyper-parameter tuning on a separately held validation set inside the training set, and the values of all hyper-parameters in our ML model were kept constant throughout all experiments in the paper on the test set. In order to ensure that there was no overlap between the training and testing steps, we did not make use of the test data either directly or indirectly during the training phase, either for learning parameter weights or selecting hyper-parameters.

6. Results

Evaluation of the accuracy of ML mapping rules derived during the training stage for MAX-DOAS data inversion was done by comparing the "true" atmospheric aerosol properties to the ML inverted properties. The evaluation data set consists of 364800 MAX-DOAS simulated sky scans that are outside of the training set. The number of simulations in the evaluation data set as a function of solar zenith angle (SZA) and relative azimuth angle (RAA) are shown in Figure 7. Between 1100 and 1300 aerosol scenarios are present in each SZA-RAA bin.

Deleted: aerosol profile,

Deleted: variables

Deleted: The model is implemented using Keras, which is a commonly used deep learning library for Python.

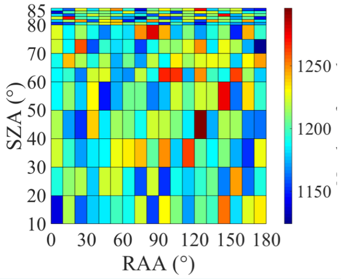
Deleted: selection

Deleted: default values of the

Deleted: was used without looking at the remainder 25% of the data that was used for testing.

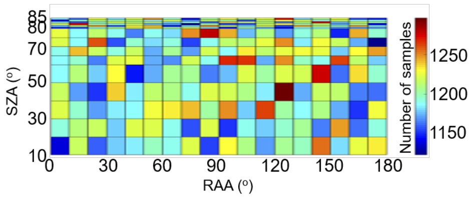
Deleted:

Deleted: 4



Deleted:

Formatted: Font:(Default) Times New Roman, 10 pt



Formatted: Font:(Default) Times New Roman, 10 pt

Figure 7. Number of simulations in the evaluation data set as a function of solar zenith (SZA) angle and relative azimuth angle (RAA).

Deleted: 4

The following ML predicted aerosol properties were evaluated: (1) asymmetry factor, (2) single scattering albedo, (3) total aerosol optical thickness, and (4) partial aerosol optical thickness for each layer from 0 to 4 km. A relative error ϵ of the retrieved by ML parameter $\hat{x}$ relative to the "true" value x is calculated according to Eq. (10):

$$\epsilon \equiv \frac{\hat{x}-x}{x}\cdot 100\% \,, \qquad (6)$$

Deleted: 10

The relative error evaluation presented in the subsequent sections was performed on the retrievals from a single ML training. Since ML itself introduces randomness during the training stage, we retrained the model 20 times with the same hyperparameters for evaluating the uncertainty of the ML training.

6.1. Asymmetry factor at 360 nm

The ML-based approach shows an ability to invert aerosol asymmetry factor with a mean error of -0.14% and two standard deviations of 2.04% and nearly normal error distribution (Fig. 8(a)). To evaluate if any dependence of the asymmetry factor retrieval exists on SZA and RAA the mean error and the two standard deviations are shown in Fig. 8(b, c). These distributions suggest that dependence of the asymmetry factor retrieval on SZAs and RAAs is relatively small. However, systematically higher relative errors are observed around SZA of 65° and RAA of 30-40°. The cause of these elevated errors is not clear at this point.

Deleted: 5

Deleted: 5

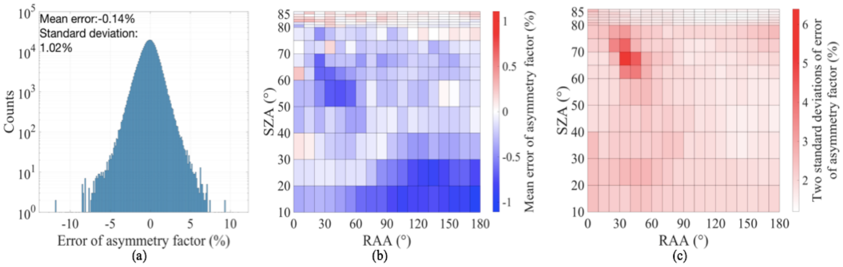


Figure 8. Asymmetry factor retrieval errors: (a) error histogram; (b) mean error as a function of SZA and RAA; (c) two standard deviations as a function of SZA and RAA.

Deleted: 5

6.2. Single scattering albedo at 360 nm

Similar high accuracy is achieved for ML retrieval of the single scattering albedo with a mean error of 0.19% and two standard deviations of 3.46% and nearly normal error distribution, somewhat positively skewed (Fig. 9). Slightly higher errors are observed at RAA smaller than 60° and most SZA.

Deleted: 6

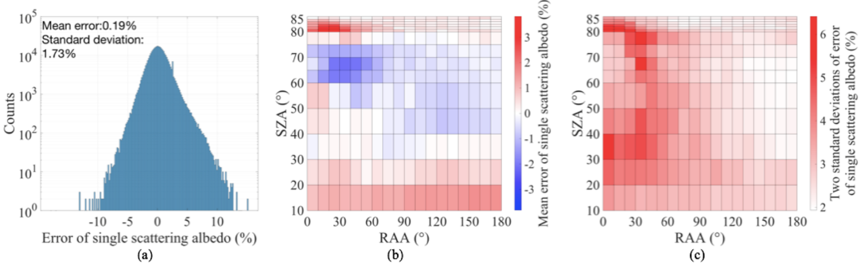


Figure 9. Single scattering albedo retrieval errors: (a) error histogram (b) mean error as a function of SZA and RAA (c) two standard deviations as a function of SZA and RAA.

Deleted: 6

Mean errors are also larger at small RAA and SZA > 85°. Traditional optimal estimation techniques also struggle with the MAX-DOAS data inversion at small RAA due to uncertainty in aerosol forward and backward scattering.

6.3. Total aerosol optical depth at 360 nm

Total AOD retrieval is more challenging for the ML model than the single scattering albedo or asymmetry factor, especially at lower total AOD levels. Box plots of the total AOD error for different "true" total AOD values are given in Fig. 10. In general, ML algorithm tends to underestimate total AOD from the mean error ± 2 standard deviations of -8.39 ± 8.81% (total AOD 0.15) to -1.52 ± 3.10% (total AOD of 0.75). Total AOD retrieval error distribution over all cases is close to Gaussian distribution, but with two peaks (Fig. 11). The mean error (± two standard deviations) is -3.58% ± 7.68%. The bias of the model does not have much dependence on SZAs and RAAs (Fig. 11(b)). Still, lager errors and uncertainties can be observed at higher SZAs and lower RAAs (Fig. 11(c)).

Deleted: 7

Deleted: 8

Deleted: 8

Deleted: 8

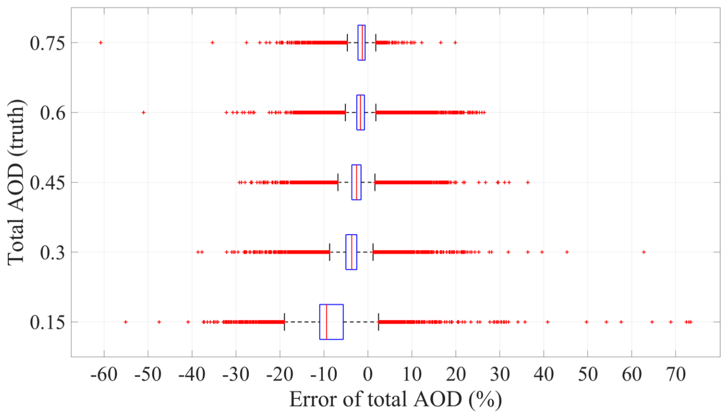


Figure 10. Box plots of total AOD prediction errors for each "true" total AOD value. The box central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol.

Deleted: 7

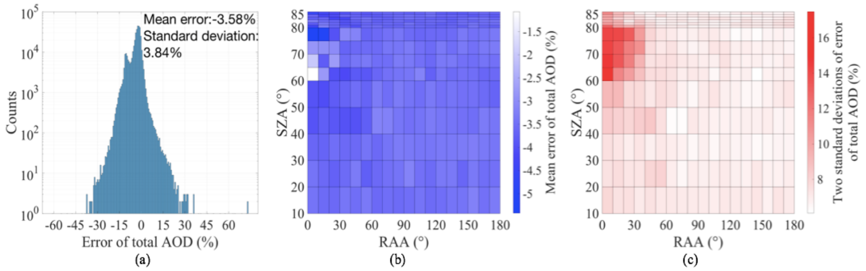


Figure 11. Total AOD retrieval errors: (a) error histogram (b) mean error as a function of SZA and RAA (c) two standard deviation as a function of SZA and RAA.

Deleted: 8

6.4. Partial aerosol optical depth profile from 0 to 4 km

The contribution of partial AOD retrieval error at each atmospheric layer from 0 to 4 km to the total AOD is shown in Fig. 12. Layer partial AOD retrieval error relative to the total AOD depends on the absolute amount of aerosols and its altitude and on average is less than 1% per layer. Just like OEM methods, the ML method has lower accuracy of retrieving elevated aerosol layers especially corresponding to smaller total AOD. The larger distribution of relative errors in partial AOD at 1.5 km and 2 km is mainly due to the presence of elevated layers in the training data that peaked at those heights. If the aerosol were also present in meaningful amounts above those altitudes the error distribution would have been larger above 2 km.

Deleted: 9

Deleted: This error contribution to the total

Deleted: AOD error

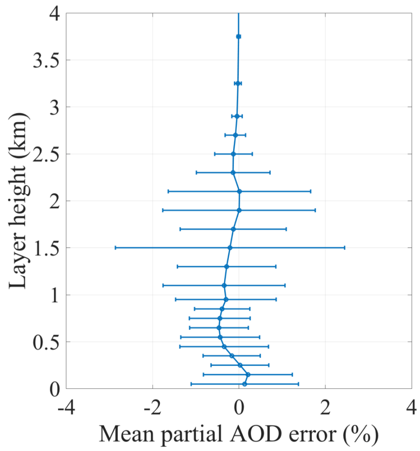


Figure 12. Mean partial layer AOD error ± one standard deviation.

Deleted: 9

A linear regression analysis of the “true” versus the retrieved partial AOD was performed using the least-squares fitting for each layer from 0 to 2.2 km (Fig. 13). Intercepts of linear regression analysis for all layers were zero with RMS ≤ 0.01. High R^2 values (0.93 – 0.99) and slopes (m) close to one suggest that the ML method relatively accurately estimates partial AOD at the layers between 0 and 2.2 km. As was noted earlier lower retrieval accuracy is observed at the higher altitudes.

Deleted: 10

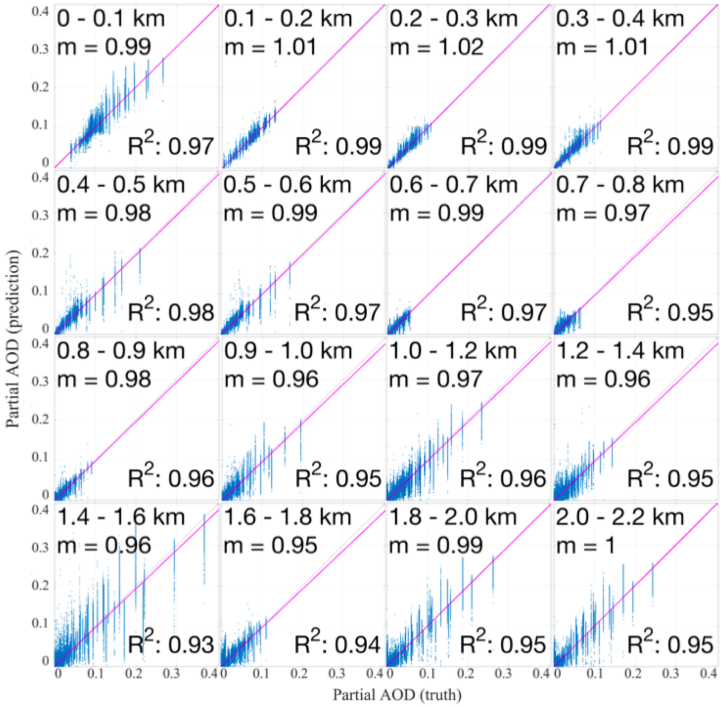


Figure 13. Correlation between the retrieved partial AOD and the “true” partial AOD for each layer from 0-2.2 km ($retrieved\ partial\ AOD = m \cdot "true"\ partial\ AOD + intercept$). The intercept of all linear regression analyses is 0 with RMS < 0.01.

Deleted: 10

Figure 14 shows some examples of the partial AOD profiles retrieved by the ML inversion model. Panels (a)-(h) in Fig. 14 contain randomly selected profiles out of the tested pool. While panels (i)-(l) contain some of the worst predictions. These examples show that the ML model is able to predict the elevated aerosol layers and even in those cases having large discrepancies, the model is still capturing the correct shape.

Deleted: 11

Deleted: 11

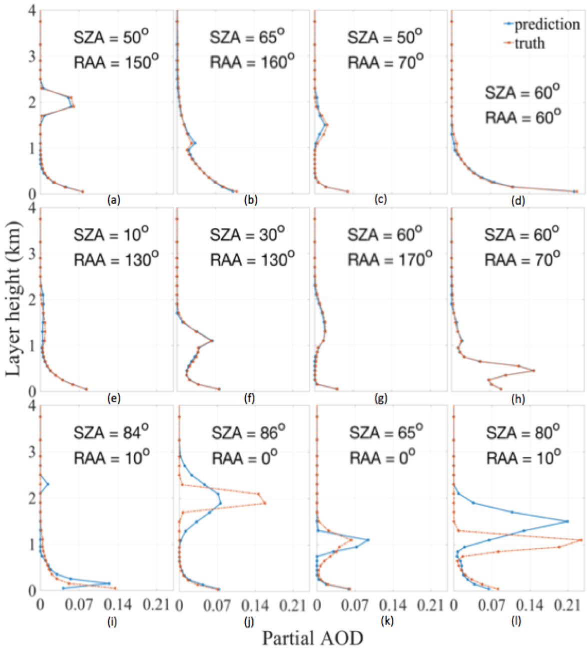


Figure 14. Examples of predicted partial layer AOD profiles: (a)-(h) randomly selected examples and (i)-(l) bad predictions

Deleted: 11

6.5. Effect of random noise in ML training on the retrievals

To estimate retrieval uncertainties due to random noise in ML training on the aerosol properties we reran the ML training stage 20 times. Mean errors and standard deviations for total AOD, single scattering albedo and asymmetry factor for each trained model are shown in Fig. 15.

Deleted: 12

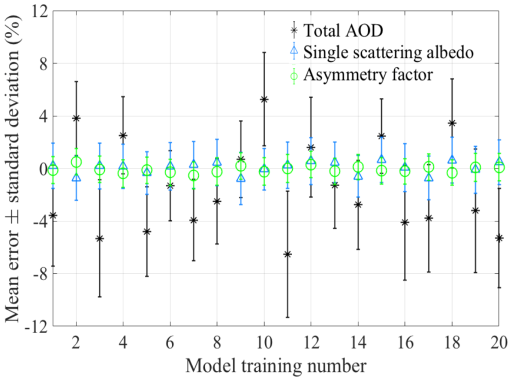


Figure 15. Effect of random noise in model training on the retrieved aerosol properties.

Deleted: 12

Table 2 summarizes the effect of random model training noise on the retrieved properties. In general, most ML models result in a normal distribution of errors with an additional bias in the mean. Since the individual model training has a very small effect on error distribution (small changes in standard deviation between the different training runs) we add the variation in bias with standard deviation in quadrature to estimate the total error of the ML model including the random error of the training as:

(1) Total AOD error ± 2 standard deviations = -1.4 ± 10.1 %;
(2) Single scattering albedo error ± 2 standard deviations = 0.1 ± 3.6 %;
(3) Asymmetry factor error ± 2 standard deviations = -0.1 ± 2.1 %.

Table 2. Statistics of aerosol property error analysis from 20 ML models (20 different training runs)

Optical property	**bias ± std, %**	**Standard deviation ± std, %**
Total AOD error	-1.43 ± 3.54	3.56 ± 0.64
Single scattering albedo error	0.06 ± 0.47	1.72 ± 0.10
Asymmetry factor error	-0.08 ± 0.25	1.01 ± 0.03

7. Conclusions and future work

This paper presents a fast ML-based algorithm for the inversion of $\Delta SCD(O_2O_2)$ from a single MAX-DOAS sky scan into aerosol partial optical depth profile, single scattering albedo and asymmetry factor at 360 nm. Training and evaluation of ML algorithm are performed using VLIDORT simulations of $\Delta AMF(O_2O_2)$ for about 1.45 million scenarios with 75% randomly selected cases for training and 25% (~ 365 thousand cases) for evaluation.

Evaluation of four retrieved aerosol properties (asymmetry factor, single scattering albedo, total AOD and partial AOD for each layer from 0 to 4 km) shows good performance of the ML algorithm with small biases and normal distribution of the errors. 95.4% of the retrieved optical properties have errors within the following ranges: (-1.4 ± 10.1) % for total AOD, (0.1 ± 3.6) % for single scattering albedo, and (-0.1 ± 2.1) % for asymmetry factor. Linear regression analysis using the least-squares fitting method between the "true" and retrieved layer partial AODs resulted in high correlation coefficients ($R^2 = 0.93 – 0.99$), slopes near unity (0.95 – 1.02) and zero intercepts with RMS ≤ 0.01 for each layer from 0 to 2.2 km. The ML algorithm, in general, has less accuracy retrieving low total AOD scenarios and their corresponding profiles. Even in those scenarios with less accuracy, the ML model is still capable of capturing the correct profile shape.

Application of ML-based algorithm to real data inversion has the following advantages:

(1) Fast real-time data inversion of the aerosol optical properties;
(2) Simple implementation by using an HDF file with the model coefficients in open source codes such as Python;
(3) Ability to retrieve single scattering albedo and asymmetry factor;

(4) Use of the ML algorithm retrieved aerosol extinction coefficient profiles; single scattering albedo and asymmetry factor as initial guess inputs in more formal inversion algorithms (with radiative transfer simulations).

To verify that the ML retrievals are representative of the physical processes we suggest simulating ΔSCD(O_2O_2) using a radiative transfer model (e.g. VLIDORT) with the ML retrieved properties as inputs (aerosol extinction coefficient profile, single scattering albedo, and asymmetry). Deviations from the measured and simulated ΔSCD(O_2O_2) should be included in error analysis.

To make the ML model more robust the training data should include more realistic aerosol inputs and radiative transfer simulations including 1) Rotational Raman scattering simulations to add Ring measurements from MAX-DOAS; 2) different surface albedos; 3) more realistic aerosol profiles (e.g. from a 3-D multi-wavelength aerosol/cloud database based on CALIPSO and EARLINET aerosol profiles, LIVAS (Amiridis et al., 2015)); 4) multiple wavelengths.

Code/Data availability

All data used in this study (radiative transfer simulations and ML model from a single training) are available from (Dong et al., 2019).

Author contribution

Elena Spinei conceived the original idea of the algorithm and performed radiative transfer simulations to generate training and test data sets. Yun Dong developed the machine learning (ML) algorithm, conducted training and data inversion, performed error analysis and visualization. Anuj Karpatne guided the design of the ML model architecture. Elena Spinei supervised the project. All authors discussed the results and contributed to the final manuscript.

Competing interests

The authors declare that they have no conflict of interest.

References

Amiridis, V., Marinou, E., Tsekeri, A., Wandinger, U., Schwarz, A., Giannakaki, E., Mamouri, R., Kokkalis, P., Binietoglou, I., Solomos, S., Herekakis, T., Kazadzis, S., Gerasopoulos, E., Proestakis, E., Kottas, M., Balis, D., Papayannis, A., Kontoes, C., Kourtidis, K., Papagiannopoulos, N., Mona, L., Pappalardo, G., Le Rille, O. and Ansmann, A.: LIVAS: a 3-D multi-wavelength aerosol/cloud database based on CALIPSO and EARLINET, Atmospheric Chemistry and Physics, 15(13), 7127–7153, doi:10.5194/acp-15-7127-2015, 2015.

Beirle, S., Dörner, S., Donner, S., Remmers, J., Wang, Y. and Wagner, T.: The Mainz profile algorithm (MAPA), Atmospheric Measurement Techniques, 12(3), 1785–1806, doi:https://doi.org/10.5194/amt-12-1785-2019, 2019.

Bodhaine, B. A., Wood, N. B., Dutton, E. G. and Slusser, J. R.: On Rayleigh Optical Depth Calculations, Journal of Atmospheric and Oceanic Technology, 16(11), 1854–1861, doi:10.1175/1520-0426(1999)016<1854:ORODC>2.0.CO;2, 1999.

Britz, D.: Recurrent Neural Networks Tutorial, Part 1 – Introduction to RNNs, WildML [online] Available from: http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-rnns/ (Accessed 15 January 2020), 2015.

Clémer, K., Van Roozendael, M., Fayt, C., Hendrick, F., Hermans, C., Pinardi, G., Spurr, R., Wang, P. and De Mazière, M.: Multiple wavelength retrieval of tropospheric aerosol optical properties from MAXDOAS measurements in Beijing, Atmospheric Measurement Techniques, 3(4), 863–878, doi:10.5194/amt-3-863-2010, 2010.

Dong, Y., Spinei, E. and Karpatne, A.: amt-2019-368, University Libraries, Virginia Tech [online] Available from https://doi.org/10.7294/6A3T-ZV25, 2019.

Dubovik, O., Holben, B., Eck, T. F., Smirnov, A., Kaufman, Y. J., King, M. D., Tanré, D. and Slutsker, I.: Variability of Absorption and Optical Properties of Key Aerosol Types Observed in Worldwide Locations, Journal of the Atmospheric Sciences, 59(3), 590–608, doi:10.1175/1520-0469(2002)059<0590:VOAAOP>2.0.CO;2, 2002.

Efremenko, D. S., Loyola R., D. G., Hedelt, P. and Spurr, R. J. D.: Volcanic SO2 plume height retrieval from UV sensors using a full-physics inverse learning machine algorithm, International Journal of Remote Sensing, 38(sup1), 1–27, doi:10.1080/01431161.2017.1348644, 2017.

Frieß, U., Beirle, S., Alvarado Bonilla, L., Bösch, T., Friedrich, M. M., Hendrick, F., Piters, A., Richter, A., van Roozendael, M., Rozanov, V. V., Spinei, E., Tirpitz, J.-L., Vlemmix, T., Wagner, T. and Wang, Y.: Intercomparison of MAX-DOAS vertical profile retrieval algorithms: studies using synthetic data, Atmospheric Measurement Techniques, 12(4), 2155–2181, doi:10.5194/amt-12-2155-2019, 2019.

Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybernetics, 36(4), 193–202, doi:10.1007/BF00344251, 1980.

Glorot, X. and Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks, 8, 2010.

Graves, A. and Schmidhuber, J.: Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks, in Advances in Neural Information Processing Systems 21, edited by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, pp. 545–552, Curran Associates, Inc. [online] Available from: http://papers.nips.cc/paper/3449-offline-handwriting-recognition-with-multidimensional-recurrent-neural-networks.pdf (Accessed 4 January 2020), 2009.

Graves, A., Liwicki, M., Bunke, H., Schmidhuber, J. and Fernández, S.: Unconstrained On-line Handwriting Recognition with Recurrent Neural Networks, in Advances in Neural Information Processing Systems 20, edited by J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, pp. 577–584, Curran Associates, Inc. [online] Available from: http://papers.nips.cc/paper/3213-unconstrained-on-line-handwriting-recognition-with-recurrent-neural-networks.pdf (Accessed 4 January 2020), 2008.

Haywood, J. and Boucher, O.: Estimates of the direct and indirect radiative forcing due to tropospheric aerosols: A review, Reviews of Geophysics, 38(4), 513–543, doi:10.1029/1999RG000078, 2000.

Hedelt, P., Efremenko, D. S., Loyola, D. G., Spurr, R. and Clarisse, L.: SO2 Layer Height retrieval from Sentinel-5 Precursor/TROPOMI using FP_ILM, Atmospheric Measurement Techniques Discussions, 1–23, doi:10.5194/amt-2019-13, 2019.

Hinton, G.: Neural Networks for Machine Learning Lecture 6a, [online] Available from: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf (Accessed 16 March 2019), 2012.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9(8), 1735–1780, 1997.

Intergovernmental Panel on Climate Change, Ed.: Evaluation of Climate Models, in Climate Change 2013 - The Physical Science Basis, pp. 741–866, Cambridge University Press, Cambridge., 2014.

Johansson, E. m., Dowla, F. u. and Goodman, D. m.: Backpropagation learning for multilayer feed-forward neural networks using the conjugate gradient method, Int. J. Neur. Syst., 02(04), 291–301, doi:10.1142/S0129065791000261, 1991.

Krizhevsky, A., Sutskever, I. and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, in Advances in Neural Information Processing Systems 25, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, pp. 1097–1105, Curran Associates, Inc. [online] Available from: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf (Accessed 4 January 2020), 2012

LeCun, Y., Haffner, P., Bottou, L. and Bengio, Y.: Object Recognition with Gradient-Based Learning, in Shape, Contour and Grouping in Computer Vision, edited by D. A. Forsyth, J. L. Mundy, V. di Gesú, and R. Cipolla, pp. 319–345, Springer Berlin Heidelberg, Berlin, Heidelberg., 1999.

Li, X. and Wu, X.: Constructing Long Short-Term Memory based Deep Recurrent Neural Networks for Large Vocabulary Speech Recognition, arXiv:1410.4281 [cs] [online] Available from: http://arxiv.org/abs/1410.4281 (Accessed 16 January 2020), 2015.

Platt, U. and Stutz, J.: Differential optical absorption spectroscopy: principles and applications, Springer, Berlin., 2008.

Rodgers, C. D.: Inverse methods for atmospheric sounding: theory and practice, Reprinted., World Scientific, Singapore., 2004.

Schilling, H., Bulatov, D., Niessner, R., Middelmann, W. and Soergel, U.: Detection of Vehicles in Multisensor Data via Multibranch Convolutional Neural Networks, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11, 4299–4316, doi:10.1109/JSTARS.2018.2825099, 2018.

Schulz, K., Hänsch, R. and Sörgel, U.: Machine learning methods for remote sensing applications: an overview, in Earth Resources and Environmental Remote Sensing/GIS Applications IX, vol. 10790, p. 1079002, International Society for Optics and Photonics., 2018.

Simonyan, K. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv:1409.1556 [cs] [online] Available from: http://arxiv.org/abs/1409.1556 (Accessed 4 January 2020), 2015.

Spurr, R.: LIDORT and VLIDORT: Linearized pseudo-spherical scalar and vector discrete ordinate radiative transfer models for use in remote sensing retrieval problems, in Light Scattering Reviews 3, edited by A. A. Kokhanovsky, pp. 229–275, Springer Berlin Heidelberg, Berlin, Heidelberg., 2008.

Thalman, R. and Volkamer, R.: Temperature dependent absorption cross-sections of O2–O2 collision pairs between 340 and 630 nm and at atmospherically relevant pressure, Physical Chemistry Chemical Physics, 15(37), 15371, doi:10.1039/c3cp50968k, 2013.

Thomas, A.: Keras LSTM tutorial - How to easily build a powerful deep learning language model, Adventures in Machine Learning [online] Available from: https://adventuresinmachinelearning.com/keras-lstm-tutorial/ (Accessed 16 March 2019), 2018.

Vlemmix, T., Eskes, H. J., Piters, A. J. M., Schaap, M., Sauter, F. J., Kelder, H. and Levelt, P. F.: MAX-DOAS tropospheric nitrogen dioxide column measurements compared with the Lotos-Euros air quality model, Atmospheric Chemistry and Physics, 15(3), 1313–1330, doi:https://doi.org/10.5194/acp-15-1313-2015, 2015.

Wagner, T., Beirle, S., Benavent, N., Bösch, T., Chan, K. L., Donner, S., Dörner, S., Fayt, C., Frieß, U., García-Nieto, D., Gielen, C., González-Bartolome, D., Gomez, L., Hendrick, F., Henzing, B., Jin, J. L., Lampel, J., Ma, J., Mies, K., Navarro, M., Peters, E., Pinardi, G., Puentedura, O., Puķīte, J., Remmers, J., Richter, A., Saiz-Lopez, A., Shaiganfar, R., Sihler, H., Van Roozendael, M., Wang, Y. and Yela, M.: Is a scaling factor required to obtain closure between measured and modelled atmospheric O4 absorptions? A case study for two days during the MADCAT campaign, Atmospheric Measurement Techniques Discussions, 1–85, doi:10.5194/amt-2018-238, 2018.

Deleted: Amiridis, V., Marinou, E., Tsekeri, A., Wandinger, U., Schwarz, A., Giannakaki, E., Mamouri, R., Kokkalis, P., Binietoglou, I., Solomos, S., Herekakis, T., Kazadzis, S., Gerasopoulos, E., Proestakis, E., Kottas, M., Balis, D., Papayannis, A., Kontoes, C., Kourtidis, K., Papagiannopoulos, N., Mona, L., Pappalardo, G., Le Rille, O. and Ansmann, A.: LIVAS: a 3-D multi-wavelength aerosol/cloud database based on CALIPSO and EARLINET, Atmospheric Chemistry and Physics, 15(13), 7127–7153, doi:10.5194/acp-15-7127-2015, 2015. ... [3]

Page 16: [1] Deleted lind 3/17/20 1:30:00 AM

Single scattering albedo, $\omega(\lambda)$, is defined as the ratio of scattering optical depth ($\tau_{scattering}$) to the total optical depth ($\tau_{scattering}+\tau_{absorption}$) at wavelength λ (Eq. (1)):

$$\omega(\lambda) = \frac{\tau_{scattering}}{\tau_{scattering}+\tau_{absorption}} \ , \qquad (1)$$

The magnitude of $\omega(\lambda)$ determines whether the aerosols have a cooling or warming effect depending on the underlying surface albedo. Since $\omega(\lambda)$ mainly depends on the aerosol composition (complex part of the refractive index) and size, it is difficult to characterize for aerosol mixtures, especially of the anthropogenic origin.

Scattering phase function describes the angular intensity distribution of electromagnetic radiation scattered by the aerosol. It depends on the aerosol size compared to the incident electromagnetic radiation wavelength (λ), aerosol particle shape, and composition (relative refractive index m at λ). In the Lorenz-Mie formalism, applied in this study, wavelength-aerosol size dependence is expressed by the size parameter (α) as the ratio of the spherical particle circumference to the wavelength (Seinfeld and Pandis, 2016).

The scattering phase function, $P(\theta,\alpha,m)$, at a scattering angle θ for spheres is calculated by normalizing the scattered intensity into angle θ by the intensity integrated over all scattering directions. The dominating scattering direction is described by the asymmetry factor (g), which is defined as the phase function weighted cosine of the scattering angles integrated from 0° (forward direction) to 180° (backward direction):

$$g(\alpha,m) = \frac{1}{2}\int_0^{\pi} cos\,(\theta)\cdot P(\theta,\alpha,m)\cdot sin\,(\theta)d\theta \ , \qquad (2)$$

The asymmetry factor ranges from -1 (backscattering) to +1 (forward scattering). Henyey and Greenstein (1941) proposed a simplified "fitting" technique to calculate $P(\theta)$ using solely the asymmetry factor:

$$\mathrm{P_{HG}}(\cos\theta) = \frac{1-g^2}{(1+g^2-2\cdot g\cdot\cos(\theta))^{\frac{3}{2}}} \ , \qquad (3)$$

Several methods used to solve the radiative transfer equation in the atmosphere (e.g. δ-M, discrete ordinate, and Monte Carlo) require scattering phase function expansion into a finite series of Legendre polynomials ($PL(\cos\theta)$) to account for the dependence of the radiation field on azimuth (Spurr, 2008). Lorenz-Mie type codes output the Legendre expansion coefficients. The expansion of the Henyey-Greenstein phase function into Legendre polynomials (P_L) is given by a simple relationship shown in Eq. (4), where $(2L+1)g^L$ is its Legendre moments (expansion coefficients).

$$P_{HG}(\cos\theta) = \sum(2L+1)\cdot g^L\cdot P_L(cos\theta),$$

Page 24: [2] Deleted lind 3/17/20 3:06:00 PM

the following viewing zenith angles: 0, 40, 50, 60, 65, 70, 75, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89°. To ensure that the training dataset contains all observation geometries feasible for MAX-DOAS sky scans we have included the following:

Relative azimuth angles: 0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150, 160, 170, 180°, and

Solar zenith angles: 0, 10, 20, 30, 40, 50, 60, 65, 70, 75, 80, 85°.

Page 39: [3] Deleted Microsoft Office User 6/16/20 3:02:00 AM

Amiridis, V., Marinou, E., Tsekeri, A., Wandinger, U., Schwarz, A., Giannakaki, E., Mamouri, R., Kokkalis, P., Binietoglou, I., Solomos, S., Herekakis, T., Kazadzis, S., Gerasopoulos, E., Proestakis, E., Kottas, M., Balis, D., Papayannis, A., Kontoes, C., Kourtidis, K., Papagiannopoulos, N., Mona, L., Pappalardo, G., Le Rille, O. and Ansmann, A.: LIVAS: a 3-D multi-wavelength aerosol/cloud database based on CALIPSO and EARLINET, Atmospheric Chemistry and Physics, 15(13), 7127–7153, doi:10.5194/acp-15-7127-2015, 2015.

Beirle, S., Dörner, S., Donner, S., Remmers, J., Wang, Y. and Wagner, T.: The Mainz profile algorithm (MAPA), Atmospheric Measurement Techniques, 12(3), 1785–1806, doi:https://doi.org/10.5194/amt-12-1785-2019, 2019.

Bodhaine, B. A., Wood, N. B., Dutton, E. G. and Slusser, J. R.: On Rayleigh Optical Depth Calculations, Journal of Atmospheric and Oceanic Technology, 16(11), 1854–1861, doi:10.1175/1520-0426(1999)016<1854:ORODC>2.0.CO;2, 1999.

Clémer, K., Van Roozendael, M., Fayt, C., Hendrick, F., Hermans, C., Pinardi, G., Spurr, R., Wang, P. and De Mazière, M.: Multiple wavelength retrieval of tropospheric aerosol optical properties from MAXDOAS measurements in Beijing, Atmospheric Measurement Techniques, 3(4), 863–878, doi:10.5194/amt-3-863-2010, 2010.

Dong, Y., Spinei, E. and Karpatne, A.: amt-2019-368, University Libraries, Virginia Tech [online] Available from https://doi.org/10.7294/6A3T-ZV25, 2019.

Dubovik, O., Holben, B., Eck, T. F., Smirnov, A., Kaufman, Y. J., King, M. D., Tanré, D. and Slutsker, I.: Variability of Absorption and Optical Properties of Key Aerosol Types Observed in Worldwide Locations, Journal of the Atmospheric Sciences, 59(3), 590–608, doi:10.1175/1520-0469(2002)059<0590:VOAAOP>2.0.CO;2, 2002.

Efremenko, D. S., Loyola R., D. G., Hedelt, P. and Spurr, R. J. D.: Volcanic SO2 plume height retrieval from UV sensors using a full-physics inverse learning machine algorithm, International Journal of Remote Sensing, 38(sup1), 1–27, doi:10.1080/01431161.2017.1348644, 2017.

Frieß, U., Beirle, S., Alvarado Bonilla, L., Bösch, T., Friedrich, M. M., Hendrick, F., Piters, A., Richter, A., van Roozendael, M., Rozanov, V. V., Spinei, E., Tirpitz, J.-L., Vlemmix, T., Wagner, T. and Wang, Y.: Intercomparison of MAX-DOAS vertical profile retrieval algorithms: studies using synthetic data, Atmospheric Measurement Techniques, 12(4), 2155–2181, doi:10.5194/amt-12-2155-2019, 2019.

Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biol. Cybernetics, 36(4), 193–202, doi:10.1007/BF00344251, 1980.

Haywood, J. and Boucher, O.: Estimates of the direct and indirect radiative forcing due to tropospheric aerosols: A review, Reviews of Geophysics, 38(4), 513–543, doi:10.1029/1999RG000078, 2000.

Hedelt, P., Efremenko, D. S., Loyola, D. G., Spurr, R. and Clarisse, L.: SO2 Layer Height retrieval from Sentinel-5 Precursor/TROPOMI using FP_ILM, Atmospheric Measurement Techniques Discussions, 1–23, doi:10.5194/amt-2019-13, 2019.

Henyey, L. C. and Greenstein, J. L.: Diffuse radiation in the Galaxy, The Astrophysical Journal, 93, 70, doi:10.1086/144246, 1941.

Hinton, G.: Neural Networks for Machine Learning Lecture 6a, [online] Available from: https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf (Accessed 16 March 2019), 2012.

Hochreiter, S. and Schmidhuber, J.: Long Short-Term Memory, Neural Computation, 9(8), 1735–1780, 1997.

Intergovernmental Panel on Climate Change, Ed.: Evaluation of Climate Models, in Climate Change 2013 - The Physical Science Basis, pp. 741–866, Cambridge University Press, Cambridge., 2014.

Johansson, E. m., Dowla, F. u. and Goodman, D. m.: Backpropagation learning for multilayer feed-forward neural networks using the conjugate gradient method, Int. J. Neur. Syst., 02(04), 291–301, doi:10.1142/S0129065791000261, 1991.

LeCun, Y., Haffner, P., Bottou, L. and Bengio, Y.: Object Recognition with Gradient-Based Learning, in Shape, Contour and Grouping in Computer Vision, edited by D. A. Forsyth, J. L. Mundy, V. di Gesú, and R. Cipolla, pp. 319–345, Springer Berlin Heidelberg, Berlin, Heidelberg., 1999.

Platt, U. and Stutz, J.: Differential optical absorption spectroscopy: principles and applications, Springer, Berlin., 2008.

Rodgers, C. D.: Inverse methods for atmospheric sounding: theory and practice, Reprinted., World Scientific, Singapore., 2004.

Schilling, H., Bulatov, D., Niessner, R., Middelmann, W. and Soergel, U.: Detection of Vehicles in Multisensor Data via Multibranch Convolutional Neural Networks, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11, 4299–4316, doi:10.1109/JSTARS.2018.2825099, 2018.

Schulz, K., Hänsch, R. and Sörgel, U.: Machine learning methods for remote sensing applications: an overview, in Earth Resources and Environmental Remote Sensing/GIS Applications IX, vol. 10790, p. 1079002, International Society for Optics and Photonics., 2018.

Seinfeld, J. H. and Pandis, S. N.: Atmospheric chemistry and physics: from air pollution to climate change, Third edition., John Wiley & Sons, Hoboken, New Jersey., 2016.

Spurr, R.: LIDORT and VLIDORT: Linearized pseudo-spherical scalar and vector discrete ordinate radiative transfer models for use in remote sensing retrieval problems, in Light Scattering Reviews 3, edited by A. A. Kokhanovsky, pp. 229–275, Springer Berlin Heidelberg, Berlin, Heidelberg., 2008.

Thalman, R. and Volkamer, R.: Temperature dependent absorption cross-sections of O2–O2 collision pairs between 340 and 630 nm and at atmospherically relevant pressure, Physical Chemistry Chemical Physics, 15(37), 15371, doi:10.1039/c3cp50968k, 2013.

Vlemmix, T., Eskes, H. J., Piters, A. J. M., Schaap, M., Sauter, F. J., Kelder, H. and Levelt, P. F.: MAX-DOAS tropospheric nitrogen dioxide column measurements compared with the Lotos-Euros air quality model, Atmospheric Chemistry and Physics, 15(3), 1313–1330, doi:https://doi.org/10.5194/acp-15-1313-2015, 2015.

Wagner, T., Beirle, S., Benavent, N., Bösch, T., Chan, K. L., Donner, S., Dörner, S., Fayt, C., Frieß, U., García-Nieto, D., Gielen, C., González-Bartolome, D., Gomez, L., Hendrick, F., Henzing, B., Jin, J. L., Lampel, J., Ma, J., Mies, K., Navarro, M., Peters, E., Pinardi, G., Puentedura, O., Puķīte, J., Remmers, J., Richter, A., Saiz-Lopez, A., Shaiganfar, R., Sihler, H., Van Roozendael, M., Wang, Y. and Yela, M.: Is a scaling factor required to obtain closure between measured and modelled atmospheric O4 absorptions? A case study for two days during the MADCAT campaign, Atmospheric Measurement Techniques Discussions, 1–85, doi:10.5194/amt-2018-238, 2018.