Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2017-1189-AC1, 2018

Atmospheric Chemistry and Physics
Discussions

Interactive comment on "Identification of new particle formation events with deep learning" *by* Jorma Joutsensaari et al.

Jorma Joutsensaari et al.

jorma.joutsensaari@uef.fi

Received and published: 18 June 2018

We thank the referee for valuable comments and suggestions to improve the manuscript (MS). We have considered the comments and will modify MS accordingly. Our detailed responses to the referee's comments are below.

Referee's comment 1:

First, the article claimed that in the abstract and throughout the paper (section 1: lines 29-30) that this is the first time that their method classified successfully NPF events. Of course this statement is not true. In fact, Junninen et al. (2007), Kulmala et al. (2012) and later Zaidan et al. (2017) also succeed classifying this automatically. Although they were using different methods and aiming slightly for different NPF classes, but the

primary aim is the same: automatic NPF event classification. Zaidan et al. (2017) also obtained the accuracy about 84% in their recent study, which they may considered also a success. Therefore, the word "first time" here undermines the previous contributions. Instead, you should mention how different this method and/or how this method is better compared to the previous methods.

Authors' response:

We thank the referee to point out the valuable work of other researcher and the conference paper by Junninen et al., 2007, which we have not noticed before. In the revised version of MS, we will point out that this is the first time when a deep learning method, i.e. transfer learning of a deep neural network, has been successfully applied in NPF identification using the unprocessed data. In addition, we will compare in more detail our method with previous studies (Junninen et al., 2007; Kulmala et al., 2012; Zaidan et al., 2017). We would also note that the "protocol" for event analysis described by Kulmala et al. (2012) does not include the classification method by Junninen et al. (2007), even if Junninen is a co-author in the protocol paper.

Referee's comment 2:

In the lines 21-22 (section 1), the article said that "both of these studies were able to construct models to predict the probability of NPF occurrence with reasonable accuracy". This statement is not quite true. In fact, the paper by Hyvonen et al (2005) used data mining, such as ML classifier, to find relevant atmospheric variables to NPF. It is not to construct models to predict the probability of NPF. This statement should be revised.

Authors' response:

Hyvönen et al. states already in their Abstract "Using these two variables it was possible to derive a nucleation probability function." Therefore, even though the main purpose of the paper was to find variables related to NPF they still reported the probability function and thus the sentence in referred lines is correct. However, we will rephrase the sentence in the revised MS as follows: "Both of these studies were able to find the characteristic conditions for NPF event days in each site and it was seen that the conditions differ significantly. They were also able to construct models to predict the probability of NPF occurrence with reasonable accuracy, and this approach has also been used in the day-to-day planning of a complex airborne measurement campaign (Nieminen et al., 2015)"

Referee's comment 3:

The introduction section seems redundant and too long. There is a mixed reviews between the important of aerosol study, the use of data driven method and the algorithm review that is used in atmospheric study. This section should be narrowed down, by focusing only motivation of aerosol study and the use of data mining algorithms, such as CNN, in the field of atmospheric sciences and related discipline. General algorithm review can be pointed out to a specific reference sources.

Authors' response:

We agree that the introduction section is quite long. However, be believe that a general review of algorithms suitable for dataset classification would be interesting for audience of the journal. We will make the introduction section more compact and move details of CNN to an appendix in the revised version of MS.

Referee's comment 4:

In section 2 , lines 7-9, is it possible to get the latest statistics? The figure is 11 years old, this may be interesting to present the latest one, because the analyzed dataset was started from March 2017 (section 2: line 23).

Authors' response:

We have revised the statistic by the latest results by Nieminen et al. (2018).

Referee's comment 5:

Also, I may misunderstood your statement in section 2, lines: 23-25. Could you please clarify how the total days was 5534 from 24 March 2012 until 16 May 2017? (since in a year, there is only 365/366 days).

Authors' response:

The correct starting time is 24 March 2002, the error will be corrected in the revised version of MS.

Referee's comment 6:

For section 2.3, the first paragraph seems containing a lot of new details about the properties of CNN. This needs more clarification, for example, what are kernels, RELU and sofmax functions. Please also include other relevant mathematical details in the section 2.3? For example, in addition to Figure 2, it is good to include the mathematical representation of CNN.

Authors' response:

We will move detailed descriptions about CNN and training processes to the appendix in the revised MS including their short definitions. However, the detailed mathematical representation of a CNN will be quite long and include many variables and thus, we think, will most likely confuse the readers. The readers can read detailed description from textbooks of the subject, referred in the new version of the MS (e.g. Buduma and Locascio, 2017; Duda et al., 2012, Ch. 6.6.).

Referee's comment 7:

Any justification why you used standard procedure and options? (line 17, section 2.3)

Authors' response:

Our focus was to study suitability of a CNN based method for the event identification and we did not like to fine tune learning parameters. Therefore, we used procedures and options introduced in deep learning examples by Matlab because they worked well. In general, an optimization of the training parameters would be a very time consuming study. However, we will study an effect of training parameters in detail in the future.

Referee's comment 8:

Section 2.4, line 13, what happened to the first reference?

Authors' response:

We do not understand this comment. Both references (Venables and Ripley, 2002; Huber, 1981) are listed in the references. ANOVA is analysis of variance.

Referee's comment 9:

Another part of your data pre-processing is to uniform the pixel size of the images, could you explain how this has been done? Is it an automatic process?

Authors' response:

Resizing images to 227 x 227 pixels was done automatically by a standard imresize -function using default bicubic interpolation in the Matlab code (see https://se.mathworks.com/help/images/ref/imresize.html). This will be mentioned in the MS

Referee's comment 10:

What is the value to estimate class 0? because human might put some ambiguous days into this class., which they are not sure if the days are event and non-event. This class is very subjective, the CNN learning in this subjective class may confuse the model and bias the results.

Authors' response:

This is a very good question. In this study, we used the same classes than in the

manual classification in order to compare human-made and CNN based results. In fact, we made some tests by ignoring Class BD but not Class 0. Ignoring Class 0 in the CNN analysis and classifying days that are not clearly event or non-event days (or bad data) into Class 0, it would be a very good idea to improve the model. We will test this in the future studies.

Referee's comment 11:

What do you think if you filter out the bad data before you feed this into CNN learning and analysis? In this case, the CNN learning can be simplified by reducing the number of classes.

Authors' response:

We made some tests by ignoring Class BD class but this did not significantly change the results (accuracy was practically the same). We feel that Class BD should be included to the CNN analysis because it is one way to recognize when the instrument was not working properly and also the accuracy to find those days is high (ca. 85 %). In fact, BD class helps us to ignore automatically days with a system malfunction during data analysis. In general, reducing the number of classes is an interesting way of simplifying the learning process, but it means that we have less data in the training set, making the learning stage more prone to overfitting.

Referee's comment 12:

There is no method that is perfect. Describe the weakness of your method in the conclusion part?

Authors' response:

Of course, there are still some weaknesses in the developed CNN-based NPF identification method, e.g., quality and quantity of data is crucial in the training process, supervised learning is needed, the method needs some computing power (GPU), the identification is not perfect, particle formation and growth rates cannot be determined, etc. We will mention those weaknesses of the method in more detail in the revised MS.

References:

Buduma, N., and Locascio, N.: Fundamentals of Deep Learning: Designing Next-Generation Machine Intelligence Algorithms, O'Reilly Media, Sebastopol, CA, USA, 298 pp., 2017.

Duda, R., Hart, P., and Stork, D.: Pattern classification, 2nd edition, John Wiley & Sons, Inc., 2012.

Hyvönen, S., Junninen, H., Laakso, L., Dal Maso, M., Grönholm, T., Bonn, B., Keronen, P., Aalto, P., Hiltunen, V., Pohja, T., Launiainen, S., Hari, P., Mannila, H., and Kulmala, M.: A look at aerosol formation using data mining techniques, Atmos. Chem. Phys., 5, 3345-3356, 10.5194/acp-5-3345-2005, 2005.

Junninen, Heikki, et al. "An Algorithm for Automatic Classification of Two dimensional Aerosol Data." Nucleation and Atmospheric Aerosols. Springer, Dordrecht, 2007. 957-961.

Kulmala, Markku, Tuukka Petäjä, Tuomo Nieminen, Mikko Sipilä, Hanna E Manninen, Katrianne Lehtipalo, Miikka Dal Maso, et al. 2012. "Measurement of the nucleation of atmospheric aerosol particles." Nature protocols 7 (9): 1651–1667.

Nieminen, T., Kerminen, V.-M., Petäjä, T., Aalto, P. P., Arshinov, M., Asmi, E., Baltensperger, U., Beddows, D. C. S., Beukes, J. P., Collins, D., Ding, A., Harrison, R. M., Henzing, B., Hooda, R., Hu, M., Hõrrak, U., Kivekäs, N., Komsaare, K., Krejci, R., Kristensson, A., Laakso, L., Laaksonen, A., Leaitch, W. R., Lihavainen, H., Mihalopoulos, N., Németh, Z., Nie, W., O'Dowd, C., Salma, I., Sellegri, K., Svenningsson, B., Swietlicki, E., Tunved, P., Ulevicius, V., Vakkari, V., Vana, M., Wiedensohler, A., Wu, Z., Virtanen, A., and Kulmala, M.: Global analysis of continental boundary layer new particle formation based on long-term measurements, Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-304, in review, 2018.

Zaidan, M.A, V. Haapasilta, R. Relan, H. Junninen, P.P. Aalto, F.F. Canova, L. Laurson, and A.S. Foster. Neural network classifier on time series features for predicting atmospheric particle formation days. In The 20th International Conference on Nucleation and Atmospheric Aerosols, 2017

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2017-1189, 2018.