Atmospheric Measurement Techniques Discussions

Atmos. Meas. Tech. Discuss., referee comment RC1
https://doi.org/10.5194/amt-2021-37-RC1, 2021

Comment on amt-2021-37

Anonymous Referee #2

Referee comment on "Data imputation in in situ-measured particle size distributions by means of neural networks" by Pak Lun Fung et al., Atmos. Meas. Tech. Discuss., https://doi.org/10.5194/amt-2021-37-RC1, 2021

To my understanding, this manuscript proposes a neural network method to fill for missing/invalid values in particle size distributions (PSDs) measured in situ, based on correct values measured at other size bins and on knowledge of meteorological parameters. They train a number of neural networks in which one out of 23 bins in the considered size distributions is considered missing, and is predicted based on meteorological parameters and/or on the other 22 size bins.
The study seems to show that, if PSD values are known at 22 size bins, meteorological data do not add much information for the prediction of the missing value. This finding does not sound particularly surprising to me, as I would expect PSDs to be relatively smooth, and knowing their values at 22 size bins should indeed provide a good deal of information about its value at some neighbouring bins.

MAIN COMMENTS

• While the technical work of setting up and training the neural networks looks correct as far as I can see, the main problems I see with this paper are the following:

1) The first problem is that I had to carefully peruse the manuscript several times before understanding what the main goal exactly was. This means that the Authors should be more upfront in describing the goals of the study and its setup in the introduction. In particular, looking at the Introduction from Line 146 onwards, I had the impression that the scope was much broader than just interpolating a PSD at some missing values. Maybe also the title can be made more informative: instead of reading “Neural network modelling … on other particle sections, etc.” it may be something like “Replacement of missing values in in situ measured particle size distributions by means of neural networks”. This would give the reader a more immediate idea of what they are going to find in this paper.

2) The second problem is that I am not entirely sure what the scientific significance of this study is. Training a neural network to fill in a single value in a PSD based on knowledge of other 22 values looks like a rather standard technical exercise to me. I am not sure if it makes sense to write a scientific paper about this, but since I am not part of the in situ measurement community I may not be the most appropriate person to comment on this. In general, when you use a NN to fill for missing values in the PSD you are basically developing a functional relationship between the PSD value at the missing bin and its values at other bins, and it is quite obvious that this relationship will work well or less well depending on how the real PSD actually behaves. I am not sure what general conclusions can be drawn from the analysis presented in section 4, except for the one – widely expected in my opinion – that knowing already the PSD at almost all size bins helps a lot in predicting the missing value, compared to only knowing some meteorological information.

• To have a clearer idea of the added value of using a NN to replace a missing size bin in a PSD, it would be interesting to see how the NN method performs compared to a simpler approaches such as linear interpolation.

• While it is fair to say that a NN predicting PSD solely based on meteorological inputs constitutes indeed a predictive model (you can try to predict the PSD anytime and anywhere you have the needed meteorological input), I am not sure the same holds for the NN that also uses measurements of the PSD itself. I mean, if you are not measuring the PSD I guess you cannot operate such NN. Therefore, I would say that the NN using PSD values as input is not a "model to estimate the PSD" but just a "NN to interpolate missing values in the PSD". Can you please elaborate on the impact that using PSD values as input has on the operational applicability of your NN?

OTHER COMMENTS

• P1, L18, "which are able to deposit…". What is the subject of this sentence?

• P1, L21, "invert" -> "solve"

• P2, L47. I guess Dp is the particle diameter. You should state this explicitly in the text.

• P3, L109 and L118. What are these "kernel functions", and what does it mean that they are "not optimally configured"?

• P3, L112, and P5, L184. It looks to me like the acronym CPC has not been defined anywhere.

• P4, L146, “objectives” -> “objective”

• P5, L182 and L185. I think the acronym DMA is also not defined anywhere in the paper.

• P6, L211. I guess you mean “linearly interpolated in time” here

• Section 2.3. Please summarize the inputs you used in your NNs (which and how many), preferably in a table. This information is hard to find in the paper as it is now, and should be readily available to the reader.

• P6, L226. What do you mean by "generate a signal or remain silent"? Are you using a binary threshold activation function? Later on you say that you are using a sigmoid.

• P10, several instances. What do you mean by “mutual size distribution”? “mutual” between what?

• P11, L384, I guess “which leads to a lower predictably” contains some typo. What do you actually mean?

• P12, L455, by “trivial” do you mean “negligible”?

• P13, L466, “ill-configured” -> “misconfiguration” ?