

Review of paper

Capturing synoptic-scale variations in surface aerosol pollution using deep learning with meteorological data

by E. Feng et al.

submitted to *Atmos. Chem. Phys.*

Meteorology dominates day-to-day variations in aerosol concentrations, and air quality forecasts require the estimation of aerosol concentrations and their variations from meteorological data. The authors introduce a DL model to predict synoptic-scale variations in daily PM_{2.5} concentrations in China, based on a “deep” Weather Index for Aerosols (deepWIA). The authors claim that their DL approach to predict PM_{2.5} performs better than WRF-Chem simulations for eight aerosol-polluted cities in China, as well as reported semi-empirical meteorological indices and ML-based PM_{2.5} concentration retrievals based on observations. As far as I understand the approaches, the method introduced here seems to have great potential for air quality forecasting. However, the paper is too brief when describing the different approaches that are compared, and I often have the impression that the comparisons are not totally fair because apples are compared to pears. The authors should clarify this (see specific points below) and also provide more information about other important aspects that are currently treated too superficially. I therefore recommend reconsidering the paper after major revisions.

Major comment

- A) Section 2.1 introduces the input variables of the DL model. However, this appears very much ad hoc. How did you choose these variables? Why is this the best possible choice? Why are they called “feature variables”? The ASI variables are particularly unclear to me. I understand that more information is given in the Supplement, but the reader needs still more explanation about the physical concept of these variables in the paper. Why do you need “day-of-year”? I understand that seasonality matters, but this is already captured by including daily meteorological variables.
- B) In Section 2 it is very unclear to me whether the approach works on the daily timescale or shorter. Are ERA5 values used as daily means or hourly fields? Is PM_{2.5} predicted every hour, or rather daily means? This should be clarified early in the paper.
- C) Section 2.3 is not understandable to me, and I fear it is similar for many other atmospheric science colleagues. I invite the authors to explain the basic ideas of their approach with less DL jargon.

- D) L151: how do you get 1.6 million samples? Is this the number of stations times the number of days? 1.6 million seems a lot to me ...
- E) L166: I don't understand how you can compare your values with those from other studies that used WRF-Chem. I assume that they used different stations and time periods (and therefore meteorological conditions)? If two methods are not applied to exactly the same dataset, then it is very risky to compare their performance.
- F) L210: I am not sure that this comparison is fair. If I understand correctly, then WRF-Chem is run in forecast mode, whereas deepWIA is based on ERA5, i.e., on analyzed meteorological fields that are based on observations. It seems clear to me that WRF-Chem performs worse as its meteorological input suffers from forecast errors. If deepWIA was applied with forecasted meteorological fields, its performance would also be lower. I also don't fully understand how deepWIA can be called a prediction. It is based on ERA5 and therefore can only be run weeks after the actual situation.
- G) Section 4.2: Again, I am not sure if this comparison is fair, it seems to me that different methods are/were applied to different time periods and stations. This would not make for a fair comparison.
- H) Section 5: I don't understand the purpose of this section. What does the reader learn from this section and how does it relate to the rest of the paper?
- I) Overall, what I am missing, is a discussion about the parameters that are relevant for the accuracy of the DL model. As noted in my comment A), it appears to me the model uses many input parameters (maybe selected with the strategy "the more the better"), but it would be very interesting to know which input parameters really provide predictive information. Can the authors say more about this?
- J) L324: This is an interesting outlook, but it is very brief. Can the authors explain a bit more how they could turn their approach in a real forecasting system?

Minor comments

L44: "apply" should read "be applied"

L52: I think "singly" should read "individually"

L54: what is meant by "surging"?

L59: “more spatiotemporal meteorological features and a more powerful nonlinear capability ...”: this is not clear to me, “more” compared to what?

L62: I am missing a clearer description of the purpose of the model: what are input and what are output variables? What exactly is predicted on what temporal and spatial scale?

L77: “ERA-5” should read “ERA5”

L78: four categories are announced, but then only 3 are mentioned.

L94: what is meant by “and four variables of daily maxima and minima”? of what?

L98: I doubt that ERA5 has good information about vegetation height. How can this quantity be estimated on a scale of about 25 km?

L180: what is meant by “with the deepWIA model being strongly robust with the test dataset”?