Review of ‘Development of daily PM10 and PM2.5 prediction system using a deep long short-term memory neural network model by Kim et al.

Kim et al. describe the development of a long short-term memory (LSTM) artificial recurrent neural network model trained to predict concentrations of PM10 and PM2.5 at seven Korean sites. They find that their LSTM model predicts PM concentrations with an accuracy better than (or at least comparable to) a chemical transport model.
The use of machine learning for air quality applications is a hot topic and the use of LSTM for the local prediction of PM2.5/PM10 is an interesting idea. As such, the manuscript offers a meaningful addition to the discussion on machine learning in atmospheric chemistry. The paper is well written and generally easy to follow. However, in its current form the manuscript lacks context and the quality of the LSTM is not entirely convincing. The following points need to be addressed in more detail before the paper can be recommended for publication:

1. The authors should better clarify what the intended use-case is for the machine learning model. In particular, there is some ambiguity in the word ‘prediction’ as it means different things in machine learning (where it means the ‘guess’ of the model) and atmospheric chemistry (where it refers to concentration estimates in the future): is the goal of the LSTM model to make a prediction of PM2.5/PM10 concentrations 24 hours from now - based on current conditions? If so, I assume the inputs/outputs have been prepared in such a way that they incorporate this 24-hour time lag? Or is the LSTM designed to make an optimal prediction of the concentration at a given time based on current conditions? In this case, it is not fully clear what the use-case for such a model would be.
In general, the mapping between input features and the predictor variable needs more explanation. For example, based on Figures 9 and 10 one would conclude that the variables needed to make a model prediction are the current meteorological conditions as well as the previous day pollutant concentrations? If this is the case, was there a rationale for this choice? Also, PM2.5 concentration is not used as an input for the PM10 prediction model (Figure 9), but seems to be used for the prediction of PM2.5 (Figure 10)?

2. The motivation to choose LSTM over another architecture should be discussed in more detail. Was LSTM selected because urban PM concentrations are expected to be dominated by local processes (e.g., emissions) and thus have a local, time-persistent signal? This would be a reasonable argument, but possibly also limits the usefulness of this approach to (urban) areas where PM concentrations are primarily determined by local processes? Based on the current version of the manuscript, it is not obvious why a simpler architecture (e.g., XGBoost) wouldn’t yield a comparable (or even better) result.

3. The authors should clarify whether they trained just one LSTM model (for all 7 locations combined) or an individual LSTM for each station. If the former, can the LSTM model then also be used for PM predictions for a different city? This would be a powerful argument for this methodology and worthwhile testing.

4. Did the authors consider to use the logarithmic of NO_2 and SO_2 before normalizing the inputs? These species are often log-normally distributed and applying the regular normalization function to them (Eq. 1) might not be optimal. The generated missing values for both NO_2 and SO_2 are much worse than the predictions for the other four species (Figure S2), which might be further indication that these two species are not treated optimally. At the very least, a justification for using non-logarithmic concentration values for NO_2 and SO_2 should be provided.

5. The paragraph on model overfitting is confusing (page 6, line 14ff.): an overfitted model will produce better skill scores against the training data vs. the validation data since it has learned to fit well to the training data, but the model doesn't generalize. The results shown in Table 1 are not particularly encouraging in that regard and need more explanation.
It would also be helpful to provide more information on the network architecture, in particular the number of hidden nodes. Given that the number of input features is relatively small (11 variables per station per hour) and the training period only covers 2.3 years, it seems plausible that a complex model with too many modes will (a) overfit or (b) not converge to a (local) minimum in time because the training sample is too small. With regards to the latter, it would be instructive to show the MSE as a function of training cycles.

6. Another issue that should be addressed in the context of overfitting is the correlation of input variables: I assume some of the input features are highly correlated (e.g. NO_2 and SO_2, PM2.5 and PM10, temperature and O_3, etc.). While this is not a problem for the LSTM, per se, it lowers the amount of (independent) information contained in the training data and will likely slow convergence of the LSTM model as the model 'wastes time' learning these correlations first.
In that regard it is surprising to see that, for a number of stations, the PM2.5 prediction strongly depends on the previous day PM2.5 concentration but shows little dependency (or even a negative dependency) on PM10 concentration (Figure 10). Is this an expected result?

7. The CTM used in this study was run at 15x15 km^2 horizontal resolution, which can make it challenging to compare its output against ground-based observations due to representation error. This is particularly true for urban sites that might be heavily influenced by local, small-scale emission sources that are difficult to capture at this model resolution. As such, the comparison between CTM vs. LSTM predictions is somewhat unfair as it seems likely that a CTM with a local bias correction applied to it would perform significantly better. While this might be difficult to quantify, it should at least be addressed in the revised version of the manuscript.

Minor comments:

- Page 4, line 12: it would be helpful to provide the number of missing values (in %) for the pollutant concentrations.

- Page 6, line 17: I assume the authors mean 'overtuned', not 'overturned'…

- Page 21/22: the authors should explain why the LSTM predictions are missing for Daejeon from approximately 5/27 to 6/7.

- Appendix, equation S4: Isn't the sigmoid function defined as: $\sigma(x) = 1 / (1 + e^{-x})$?