This paper addresses a relevant and important topic related to improved air quality forecasting. It uses deep neural networks for that and applies to data from a single air quality station in Seoul. Three different experimental configurations are included, namely forcing with observed air quality measurements only, forecasting with observed air quality and forecasted weather data, and forecasting with observed air quality, weather, and predicted air quality from a physics model.

There are a couple of issues with the paper that hinders understanding:

1) It's not clear what the contributions of the paper are versus what already exists. Did the authors run the WRF and CMAQ models to generate the training data or were these data obtained from some other source?
2) Similarly there is no justification for the choice of model. I would like to see the approach benchmarked against simpler models (ARIMA, Random Forest, … basically anything from the statistical or machine learning family to compare against the deep learning approach). The volume of data that the model is trained on are not huge so it is not apparent that a DNN is the best choice of algorithm
3) Many details on the DNN model setup are presented with no real justification for their choice e.g. using the membership function for temporal features, the choice of DNN architecture such as number of layers is not explained. On line 50 – 55 the authors 1) note the advantages of RNN for time series forecasting and 2) that Kim et al. (2019) developed an RNN model to predict PM2.5 concentrations at two locations in Seoul. Why was RNN not considered for this study rather than DNN and how does the performance of this model compare to that reported by Kim et. al. Similarly it is not clear if the autoregressive features of the data were expressed in any form? Of course RNN expresses these implicitly but in other models it can be advantageous to feature engineer the autoregressive dependencies. Were any feature combinations other than those reported explored in the paper. The authors need to justify the choice of algorithm and how the DNN was designed detailing such information as feature selection, number of layers/nodes.
4) The manuscript could be improved to enhance readability and replicability of the study. I appreciate the authors making code and data available on Zenodo. I would however encourage them to create a GitHub repository with some documentation to allow people easily replicate the results. As mentioned in 1) authors could be more descriptive when detailing data sources. Some parts of the paper could be explained better, e.g. line 161 "average weather and air quality prediction data". What is meant by average here? Spatial or temporal. Are WRF and CMAQ data extracted from the entire Seoul area domain or subset corresponding to location of the observation point. Line 166 – 167 "ensure that the training data were not biased" – feature scaling does not ensure unbiased datasets, it simply helps the model learn better. The data could still be biased. Line 173: "undergoes feature scaling through the backpropagation algorithm" – not clear what is meant by feature scaling in this context. Line 128 "16 meteorological forecast variables were created by the WRF model" – I believe what is meant here is that 16 variables were extracted as features but many more variables were generated by the WRF model.
5) I really don't see the relevance of section 4.2. The models have already been compared and evaluated in terms of predictive skill in regression. Then you take the same models and evaluate in terms of a classification model but only whether they

predicted within those bounds (i.e. the model and results are the same the only thing that changes are the interpretation)

Other more minor comments:

1) What is the membership function defined in line 144? Is this the generation of temporal features described in subsequent lines? Why was time data encoded in this manner? It seems more standard to represent as integer values or to convert those integer values to cyclic features (i.e. so that month 12 and month 1 are close to each other rather than far away). I haven't seen this approach used previously and would like to understand the motivation and/or justification.
2) The test period is quite short – 3 months out of 51 months. Was there a reason for this?
3) Line 159 – 163: This is a quite confusing way to present forecast horizons. I'd suggest to just use hours and present forecast horizons as T06, T12, T18, T24, ... Mixing days + hours and having different chunks within each day is confusing to the reader.
4) Line 167 – 168: I don't quite understand why data was both standardised and normalised? Did this improve performance versus just using normalisation (if you wished to have bounded between 0 and 1) or indeed versus the unscaled data? Generally people chose either standardisation or normalisation so I'm curious why you did both
5) Figure 4: What does Epoch_n = Epoch_n-1 + 1 mean? What does Epoch_n-1 of validation cost > Epoch_n of validation cost mean? Should it be Validation cost of Epoch_n?
6) Equation 11 and 12, I’m not sure the use of both MSE and RMSE is necessary and could probably drop one.
7) Line 225 – 230: In classification problems, accuracy, precision and recall are the standard metrics presented. I would suggest in this paper you also include (you already do accuracy so I suggest adding precision and recall)
8) Line 280 – 281: This is a difficult narrative to support. You are using the CMAQ model output as the training data and then saying the usage of that training data allows the model to better represent the long-term-transport-induced phenomenon.
9) What do the dashed lines in the residual figures represent in Figure 7.