In this paper, the authors introduce an air pollutant (ozone and PM2.5) forecasting model system which based on the deep-learning method. With the implementation of ground observations and the outputs of 3D chemical transport model, this model can forecast more accurate concentrations of ozone and PM2.5. Further, this model system can extend the prediction of air pollutants from individual station to a regional forecasting by considering the temporal characteristics of the time series and spatial relationships among different stations. By comparing the result with the observations, the results of this model show a more accurate and reasonable distributions of ozone and PM2.5, which indicates that this model system can work as a feasible and efficient option to improve current forecast performance. I think the authors did an interesting work. And this study is within the scope of GMD journal. Some problems need to be solved before it can be published.

[Response]: We want to express our sincere thanks to Anonymous Referee #1 for acknowledging the significance of our study. Moreover, the valuable comments from the referee have also offered us great help in improving the quality of the manuscript. Please refer to the following point-to-point response to the comments. The corresponding changes have been reflected in the revised version of the paper.

Comments:

"The ground monitoring stations with at least 90% valid records" (Line 92~93) and "the ground monitoring stations with at least 95% valid records ……were selected as the source stations" (Line 95~96). How are these threshold values determined? Please describe detailed information about these threshold values determination methods.

[Reponse]: Thank you for your question. Generally speaking, the threshold values have been determined adaptively from the nature of the dataset, based on our experience from previous studies of similar nature (Lu et al., 2021; Sun et al., 2021).

The completeness of data is critical to the quality of our study. For many stations, long periods of continuous invalid records exist, which indicates that the measurement stations have undergone a systematic failure or closure. These periods may last for months or even years, so applying interpolation methods to fill these gaps is improper. Otherwise, the interpolated data may harm the training process, especially as our model uses LSTM to process the time-series data. Moreover, large fault rates indicate that the ground monitoring stations may be unreliable, in the sense that even those data that are not invalid are still untrustworthy, which may harm the model's performance profoundly. Therefore, we set the threshold values relatively high, 90% for the target stations, and even higher (95%) for the source stations, to filter out the stations that may harm the training.

Moreover, the threshold for the source stations is higher than that for target stations because the source stations are relatively more important than the target stations: the source stations provide data as part of the input to the Broadcasting model and are a part of the model. In contrast, the target stations only provide the ground truth values as a reference during the training and testing procedures (please see our response to question 2 for details). Theoretically, the trained model could be more reliable if all the source stations could reach a 100% valid rate. However, in practice, the 100% valid rate is unrealistic, and we set the value to 95% to keep an adequate number of stations. In the ideal case, if all source stations have 100% valid records, the model's performance could be further improved.

Despite the integrity of the data, we also have to consider the consequence of filtering out too many stations. Figures 1 and 2 show the threshold values for each source station and training target station to be filtered out. If the threshold for source stations (95%) is set higher (e.g., 97%), many source stations will be filtered out, and the prediction in a vast portion of the target region would be rendered questionable. Similarly, if the threshold value for target stations (90%) is set higher (e.g., 95%), the easternmost cluster of stations (around 23.5°N, 111.2°E) would be filtered out, which will cause the model to be not well-trained for the surrounding areas.

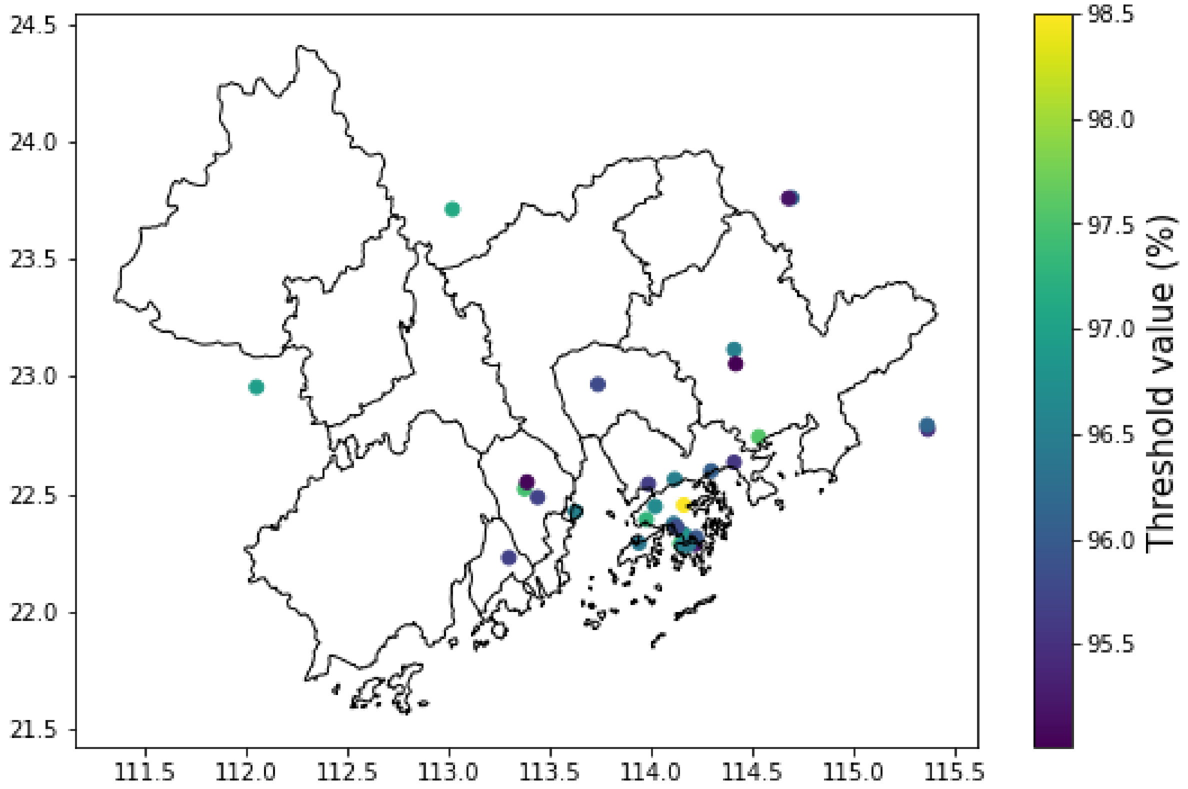

Figure 1: The threshold values for each source station to be filtered out.

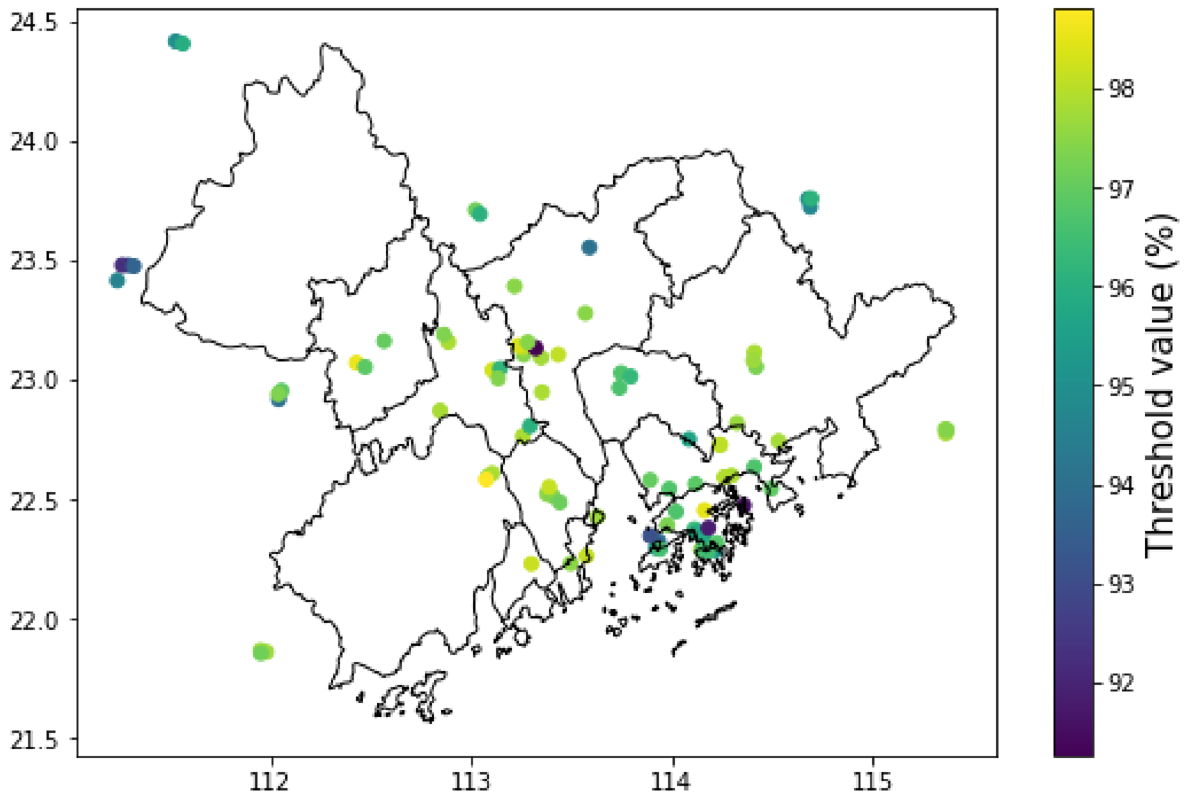

Figure 2: The threshold values for each training target station to be filtered out.

To further clarify the selection process of the source and target stations, we have added the following sentences in lines 107-109:

"Note that the threshold values of the selection criterion are determined adaptively from the nature of the dataset. The values were set relatively high such that the quality of the data could be guaranteed. However, to ensure that an adequate amount of stations are selected to represent different areas of the target region, the threshold values could not be set too close to 100%."

2. The data of source stations is also used for model training. What is the difference between the data of source stations and which of training stations?

[Response]: Thank you for the question. There is a primary difference between the source stations and the training target stations (referred to as *training stations* by the reviewer). The source stations serve as part of the **input** to the model **throughout the whole pipeline**, while the training target stations only serve as the **output during the training phase**.

As shown in Figure 3 of the preprint, each source station is associated with an LSTM encoder-decoder. Therefore, when making a forecast (for the target region), the model expects the input from all the source stations. That is to say, the set of source stations is **fixed** and **part of the Broadcasting model**.

On the other hand, the training target stations only play a role in the training phase. The ground observation values of the two target pollutants in the training target stations ($PM_{2.5}$ and O_3) for the future two days are used as reference in training. After the training, the training target stations are no longer relevant, and the model can forecast for **any place in the target region**, rather than limited to the training target stations or testing target stations.

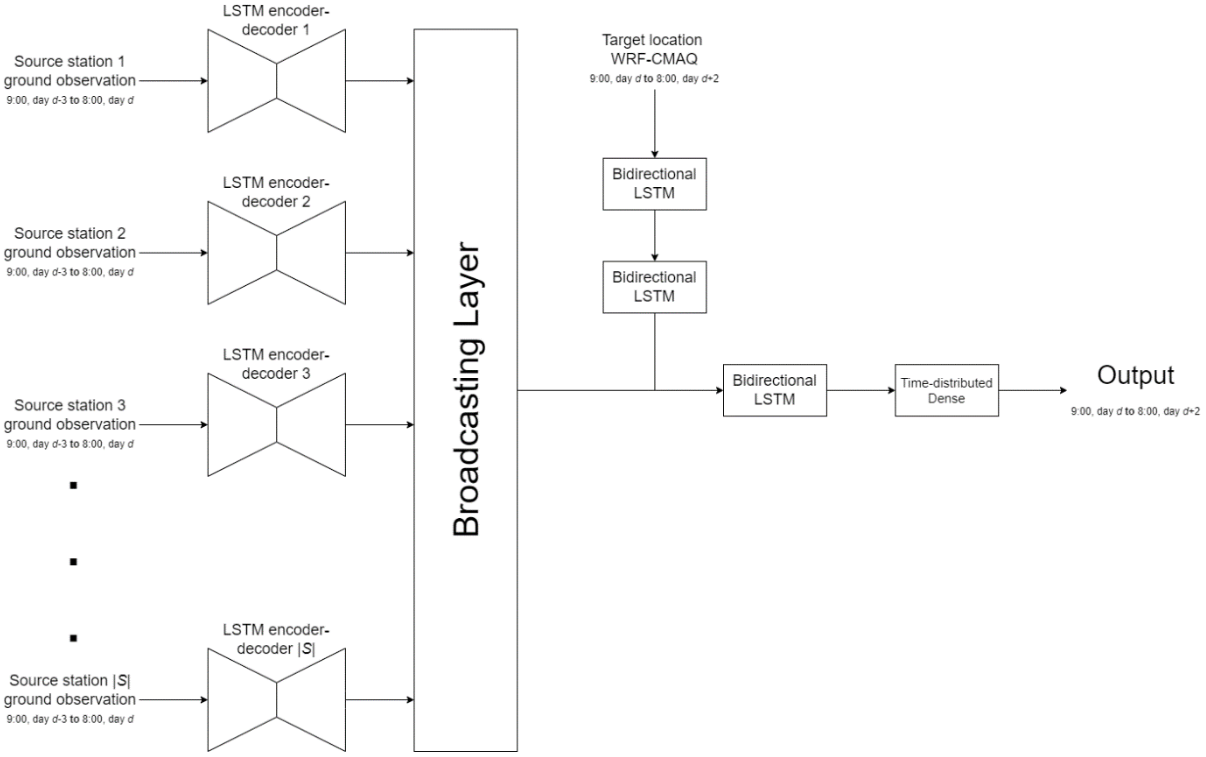
LSTM encoder-decoder 1
Source station 1 ground observation
9:00, day *d*-3 to 8:00, day *d*
LSTM encoder-decoder 2
Source station 2 ground observation
9:00, day *d*-3 to 8:00, day *d*
LSTM encoder-decoder 3
Source station 3 ground observation
9:00, day *d*-3 to 8:00, day *d*
LSTM encoder-decoder |*S*|
Source station |*S*| ground observation
9:00, day *d*-3 to 8:00, day *d*
Broadcasting Layer
Target location WRF-CMAQ
9:00, day *d* to 8:00, day *d*+2
Bidirectional LSTM
Bidirectional LSTM
Bidirectional LSTM
Time-distributed Dense
Output
9:00, day *d* to 8:00, day *d*+2

Figure 3: LSTM-broadcasting model structure (also Figure 3 in the manuscript)

3. Line 114. I am confused that why the authors use the WRF-CMAQ data with the next two days to training the model?

[Response]: Thank you for the question. Recall that the WRF-CMAQ model can also used to do the forecast. Therefore, its result is also available for the future two days. Concretely, following the setting in lines 119-126, when making the forecast for 9 am, day d to 8 am, day $d + 2$, the WRF-CMAQ results for these 48 hours will be available.

As mentioned in lines 110-114, the WRF-CMAQ results for the coming 48 h carry valuable information about this period's weather and air pollution conditions. Therefore, including it in the input can significantly improve forecasting accuracy, as verified by previous studies (Sun et al., 2021; Lu et al., 2020).

We have added the following sentence in lines 122-124 to enhance the clarity:

"Note that the WRF-CMAQ model can also work as a forecasting model, and therefore these data are available and can be used before the beginning of the forecast."

4. In this model system, large quantities of observations are used for training model. Please discuss the impact of the number of training stations and source stations on the model performance?

[Response]: Thank you for the question. To answer this question, we will first focus on the Broadcasting layer, the main deep-learning structure connecting the source stations and the training target stations.

The development Broadcasting layer is based on Tobler's first law of geography, which states that *"everything is related to everything else, but near things are more related than distant things."* The design of the Broadcasting layer, as described in Section 2.4, has the property that the impact of a source station on a target location decreases as its distance increases, which follows the law. The Broadcasting layer is a deep-learning alternative to the traditional spatial interpolation methods (e.g., inverse distance weighted, kriging), which also follow the law. Therefore, the Broadcasting layer can model the spatial distribution of air pollution caused by the relative locations and becomes the first deep learning structure to model this law to the best of our knowledge.

However, the spatial distribution of $PM_{2.5}$ and O_3 may be complex and depends on multiple factors (e.g., the meteorological field, terrain, etc.). Therefore, a substantial density of source stations and training target stations is necessary for modelling the air pollution in different areas of the target region:

If an area in the target region has few to no source stations, then the model would not gain sufficient information from the past meteorological conditions and air pollution in this area. Therefore, the forecasting accuracy may be impacted in areas with lower source station density. On the other hand, if the training target stations in an area are low in density, the model will not be trained to perform well in that area. This is because the training mechanism will put a larger weight on the areas with denser distributions of training target stations, and the other regions may be "ignored" by the training algorithm. The impact of the number of source stations and training target stations has been shown in Figure 8 and discussed in Lines 293-301.

Therefore, a relatively uniform distribution source and training target stations may also improve the model's performance. As a result, other selection strategies may be developed considering this factor in future works to improve the regional forecasting accuracy further.

To make our discussion more comprehensive, we have added the following content in lines 348-352:

"In our setting, the source and training target stations play an essential role in the model's accuracy. The forecast quality generally increases as the number of source and training target stations increases. Therefore, the model's performance has been uneven across different areas of the target region: for example, as shown in 3.3, the performance in Hong Kong is generally better than that in other regions. In future works, other selection criteria of source stations and training target stations, in place of those introduced in Section 2.1, may be developed to resolve this issue."

Reference:

Lu, X., Sha, Y. H., Li, Z., Huang, Y., Chen, W., Chen, D., ... & Fung, J. C. (2021). Development and application of a hybrid long-short term memory–three dimensional variational technique for the improvement of PM2.5 forecasting. *Science of The Total Environment*, *770*, 144221.

Sun, H., Fung, J. C., Chen, Y., Chen, W., Li, Z., Huang, Y., ... & Lu, X. (2021). Improvement of PM2.5 and O3 forecasting by integration of 3D numerical simulation with deep learning techniques. *Sustainable Cities and Society*, 103372.