We thank Anthony Fishwick for the constructive comments, which have helped us greatly improve the quality of our manuscript. Please see the point-to-point answer listed below; the corresponding changes have been reflected in our manuscript.

Q1. The format.

1) Figure 1 no north arrow and scale bar.

Thank you for the suggestion. We have added the north arrow and the scale bar in Figure 1.

2) Tables 1&2 For the concentration values, please round to one or two decimal numbers because the models cannot predict so accurate till four decimal numbers. SC (NN) or SC(nearest)

Thank you for pointing out the issues with Tables 1 and 2. We have corrected the typo and switched to 2 significant digits.

3) Figures 2&3 The outline line was covered. Please replot it. Also, descriptions should be added to the figure caption because a figure is independent from the text. Also for the Tables.

Thanks for the suggestion. We have followed the journal's formatting guidelines when making the tables, the plots, and other parts of the composition, making them easy to understand. The numbers of tables and figures have also been indicated in the text when they are referred to. We have added the description of Figure 2, please see Line 133-144. The description of Figure 3 can be found in Lines 178-186.

4) Figure 5 The fonts are too small then it is hard to read. Please replot the figure.

Thank you for pointing out the issue. We have replotted Figure 5 with a larger font size.

5) Table 3 Here just three quartiles not four

Thank you for pointing out the issue. However, Table 3 is meant to show the values of the 25%, 50% and 75% quantile, dividing 0% to 100% evenly into four quartiles. For example, in the first row, the 25%, 50%, and 75% quantiles are 9.833, 15.68, and 24.10, respectively, which means that the four quantiles are $[0, 9.833), [9.833, 15.68), [15.68, 24.10), [24.10, +\infty)$. We have changed Table 3 to the interval format to avoid confusion.

Q2. There is totally no information on the computation cost. For the broadcasting model in this work, what is the computation burden for different cases? Please add one section on this, a trade-off between accuracy and computation cost should be considered rather than just the accuracy. This is important for the real application of air quality forecast (Zhang 2012, Atmospheric Environment; Lee 2020 Geosci. Model Dev. 1055-1073).

Thank you for pointing out the issue. With GPU support, the Broadcasting model takes a few seconds to complete the regional forecast for each day (e.g., Section 3.3 and Figure 8). This constitutes the ignorable overhead compared with the time needed to obtain the WRF-CMAQ simulation results. For the SC models, depending on the nature of the interpolation methods, the regional forecast may take seconds if the interpolation can be parallelized (e.g., NN, IDW) or minutes otherwise (e.g., kriging). Therefore, the broadcasting model does not cause a significant computation burden, which satisfies the requirements for real applications (Zhang et al., 2012; Lee et al., 2020).

There is no significant trade-off relationship between the accuracy and the computational cost. We have added the following sentences about the efficiency of the Broadcasting model in lines 304-308:

"Moreover, the running time of the Broadcasting model is also reasonable. With the GPU (K80 in the Google Colab environment) support, it only takes several seconds to finish the computation for the regional forecasting of one day after the ground observation results and WRF-CMAQ data are available. Therefore, the Broadcasting model satisfies the efficiency requirements of real applications (Lee et al., 2020; Zhang et al., 2012). On the other hand, SC may take several seconds (NN and IDW) to about 3~5 minutes (Kriging), depending on whether interpolation methods can be fully parallelized."

Reference:

Lee, K., Yu, J., Lee, S., Park, M., Hong, H., Park, S. Y., ... & Song, C. H. (2020). Development of Korean Air Quality Prediction System version 1 (KAQPS v1) with focuses on practical issues. *Geoscientific Model Development*, *13*(3), 1055-1073.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., & Baklanov, A. (2012). Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. *Atmospheric environment*, *60*, 656-676.

Q3. For the established model framework, why did the authors choose LSTM? The authors should mention the reasons/advantages.

Thanks for the question. LSTM is one of the most successful and widely-used deep-learning structures for processing time series (Greff et al., 2016; Hochreiter & Schmidhuber, 1997; Karim et al., 2017; Siami-Namini et al., 2018). As a variant of RNN, it resolves the inherent problems of exploding and vanishing gradients, extending the effective forecasting horizon. Different variations of LSTM have been adopted in various scenarios of time series forecasting, depending on the diverse natures of the problems. We have chosen to use LSTM because our task heavily involves processing time series. As mentioned in lines 40 – 53, LSTM has been widely adopted in other literature on deep-learning-based air pollution prediction. In particular, as mentioned in Sections 2.2 and 2.3, LSTM encoder-decoders are suitable for the case when the output succeeds the input in temporal order, which applies to the ground observation input and the forecast; bidirectional LSTM, on the other hand, is suitable for processing the WRF-CMAQ input, which is in the same temporal space as the output. Therefore, we have combined these two variations of LSTM to construct our model.

We have added the following paragraph in Lines 323-329 to address our usage of LSTM:

"Also, our study has extensively exploited the power of LSTM in time-series-related deep learning tasks. LSTM is one of the most powerful deep-learning tools for time-series forecasting (Greff et al., 2016; Karim et al., 2017; Siami-Namini et al., 2018). As a variation of RNN, it resolves the inherent gradient explosion and vanishing problem, significantly extending the forecast horizon. By carefully examining the nature of different input and output components, we proposed combining two variations of LSTM – LSTM encoder-decoders and bi-directional LSTMs to construct the model, and achieved relatively good results. From this, we find that in complex time-series-related deep-learning tasks, careful and ad hoc analysis of the nature of the different input and output time series is needed to construct the most effective model and achieve higher accuracy."

Reference:

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, *28*(10), 2222-2232.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Karim, F., Majumdar, S., Darabi, H., & Chen, S. (2017). LSTM fully convolutional networks for time series classification. *IEEE access*, *6*, 1662-1669.

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018, December). A comparison of ARIMA and LSTM in forecasting time series. In *2018 17th IEEE international conference on machine learning and applications (ICMLA)* (pp. 1394-1401). IEEE.

Q4. I know that there are some similar previous studies (e.g., Sayeed 2021 Scientific Reports; Sayeed 2021 Atmospheric Environment; Lu 2021 Atmos. Pollut. Res. 101066; Lyu 2019 Environmental Science and Technology etc.). Therefore, the authors should clearly identify the scholarship of this work compared with previous studies and publications within your own research team.

Thank you for raising the issue. As we have mentioned in our literature review of previous works on air pollution prediction using machine learning (lines 40 - 80) and the introduction of our contributions (lines 81 - 88), our study was the first to propose an **end-to-end deep learning model** for air pollution forecasting. In particular, the proposed broadcasting layer is a novel deep learning structure that can supersede the traditional spatial interpolation methods and achieve better forecasting accuracy in regional air pollution forecasting tasks.

In particular, our study also has significant novelty compared to the listed studies, which the reviewer suggested may be similar to ours. Sayeed et al. (2021a, 2021b) and Lu et al. (2021a) studies, although combining the ground observation data and CMAQ model with the deep learning techniques, focus on improving the forecasts **at the ground monitor stations**, which is different from our work which focuses on high-quality **regional forecasts**. On the other hand, although Lyu et al. (2019) have developed a deep-learning model for regional forecasts, the generalization from the scattered ground monitor stations to the whole region still depends on kriging interpolation, which is meant to be superseded by our study.

Similarly, as we have mentioned in the manuscript, this study also improves the regional air pollution forecasting of the LSTM-3D-VAR-CAMx (Lu et al., 2021b) and the LSTM-WRF-CMAQ (Sun et al., 2021) models, which were developed by our team previously. Experiments (Section 3) have shown that this study has significant advantages over our previous studies in both accuracy and efficiency.

To further clarify the novelty of this study, we have added the following sentences in lines 63 – 67:

"Sayeed et al. (2021a, 2021b) and Lu et al. (2021a) improved the accuracy and horizon CMAQ forecast by ground observations using deep learning techniques, but the improvements were still limited to the ground monitor stations rather than the whole region. On the other hand, Lyu et al. (2019) developed an ensemble model that combines the chemical transport models and the ground observations, but the regional forecast still depends on the traditional kriging method."

Reference:

Lu, H., Xie, M., Liu, X., Liu, B., Jiang, M., Gao, Y., & Zhao, X. (2021a). Adjusting prediction of ozone concentration based on CMAQ model and machine learning methods in Sichuan-Chongqing region, China. *Atmospheric Pollution Research*, *12*(6), 101066.

Lu, X., Sha, Y. H., Li, Z., Huang, Y., Chen, W., Chen, D., ... & Fung, J. C. (2021b). 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.

Lyu, B., Hu, Y., Zhang, W., Du, Y., Luo, B., Sun, X., ... & Russell, A. G. (2019). Fusion method combining ground-level observations with chemical transport model

predictions using an ensemble deep learning framework: application in China to estimate spatiotemporally-resolved PM2.5 exposure fields in 2014–2017. Environmental science & technology, 53(13), 7306-7315.

Sayeed, A., Choi, Y., Eslami, E., Jung, J., Lops, Y., Salman, A. K., ... & Choi, M. H. (2021a). A novel CMAQ-CNN hybrid model to forecast hourly surface-ozone concentrations 14 days in advance. *Scientific reports*, *11*(1), 1-8.

Sayeed, A., Lops, Y., Choi, Y., Jung, J., & Salman, A. K. (2021b). Bias correcting and extending the PM forecast by CMAQ up to 7 days using deep convolutional neural networks. *Atmospheric Environment*, *253*, 118376.

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.