This manuscript (gmd-2022-164) proposed a new approach to refine air pollutant forecasting using deep-learning techniques. Based on the LSTM technique, a novel broadcasting layer was introduced to provide results with spatial coverage. This study not only extends the LSTM-3D-VAR method (Lu et al., 2021), but also demonstrated an alternative to achieve spatial coverage. The ability to extend single site results to multiple locations is useful for real-world applications and of great interest to the atmospheric community. The performance of the new integrated modal was carefully evaluated, and the results are promising. The reviewer thinks the manuscript can be published on GMD after addressing the following issues.

[Response]: We want to express our sincere thanks to Anonymous Referee #2 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.

1. Line 144. Please provide the full name of "The SC method".

[Response]: We thank the referee for pointing out this issue. The full name of the SC method, **spatial correction**, has been added in the corresponding location as specified by the reviewer. Please see Line 156.

2. The authors mentioned that the LSTM-3D-VAR model (Lu et al., 2021) required substantial computation power. How about the computational efficiency of LSTM-Broadcasting compared with LSTM-3D-VAR? Does LSTM-Broadcasting consume less computation power than the LSTM-3D-VAR model? If possible, please include the direct comparison of the two models developed by the authors.

[Response]: Thank you for the question. LSTM-3D-VAR-CAMx will cost about 90 minutes when the numerical forecast was performed in a cluster machine with 40 cores and 128GB memory, after the ground observation and the numerical models (WRF-CAMx) simulation results are available. For this LSTM-Broadcasting deep-learning framework, with the GPU (Google Colab K80) support, it only took several seconds to finish the computation when making the forecast for each day after receiving the ground observation data and WRF-CMAQ results. Therefore, the LSTM-Broadcasting model does not constitute any significant computational overhead and is much more efficient when compared to the LSTM-3D-VAR-

CAMx scheme developed by the authors before. We have added below sentences to Lines 304-311 in the manuscript:

"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 forecast 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. By contrast, the LSTM-3D-VAR-CAMx will cost about 90 minutes (tested on a cluster machine with 40 cores and 128GB of memory) given the ground observation and WRF-CAMx results as input, which may render the approach infeasible when instant forecasts are needed."

3. Line 281. "As in Section 3.3, with GPU acceleration…". Please check, section 3.3 cannot be found in the manuscript.

[Response]: We thank the referee for pointing out this issue. We have corrected the formatting error, and please see line 282 in the manuscript.