We appreciate the reviewer's valuable comments and constructive suggestions which help us improve the quality of the manuscript. We have carefully revised the manuscript according to these comments. Point-to-point responses are provided in the attachment. The reviewers' comments are in black, our responses are in blue, and the quotes from our manuscript are in italics.

Reviewer #1

[Comment]: How does the authors ensure the robustness of the model?

[Response]: We thank the reviewer for the valuable comment. We ensure the robustness of the model from three aspects:

a) Model structure. Inspired by computer vision tasks, we adopt the batch-normalization (Ioffe and Szegedy, 2015), dropout (Srivastava et al., 2014), L2 regularization (Zhang et al., 2016) to improve the generalization and robustness.

b) Early stop. When we train the NN-CTM, we split the data into train dataset and validation dataset. As introduced in Sec. 3.1, we trained NN-CTM on the data of the first 22 days in January, April, July, and October 2015 and tested it on the remaining successive 8 days of each month. We stop the model training when the evaluation in validation dataset does not improve within 1000 iterations.

c) Data augmentation. During training, we employ the noise injection, random rescaling, random rotation method to avoid the overfitting in training dataset.

We have clarified the model robustness in the revised manuscript, as follows:

(Section 2.2, Paragraph 5) *"Model robustness. We ensure the robustness of the model from three aspects: 1) Model structure. Inspired by computer vision tasks, we adopt the batch-normalization (Ioffe and Szegedy, 2015), dropout (Srivastava et al., 2014), L2 regularization (Zhang et al., 2016) to improve the generalization and robustness. 2) Early stop. When we train the NN-CTM, we split the data into train dataset and validation dataset, and we stop the model training when the evaluation in validation dataset does not improve within 1000 iterations. 3) Data augmentation. During training, we employ the noise injection, random rescaling, random rotation method to avoid the overfitting in training dataset."*

Reference:

Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by

Reducing Internal Covariate Shift. JMLR.org 2015.
Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 2014; 15: 1929-1958.
Zhang C, Be Ngio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization, 2016.

[Comment]: The authors use the observation data to update the emissions, however they do not mention what happens in case more than one observation station is in the grid. 27 km × 27 km is a large grid size and hence would include many observation stations in one grid. The averaged observed concentration of all stations if used won't serve the purpose to accurately update the emissions at a station.

[Response]: We thank the reviewer for the valuable comment. As mentioned in Section 2.3, we use the average value in case of multiple observation stations in a 27 km × 27 km grid. We use the same processing method for observations when calculating MAE. We focus on the emission estimation in one grid, which will be limited by the grid size. If we want to get the higher resolution emission inventory estimation result (such as focus on one typical region instead of whole China domain), we should use a finer-grained emission inventory as the input. What's more, the lack of observation data in some regions limits our updating, so we are more concerned about making good use of existing observation data.

[Comment]: The entire premise of the model depends on availability of observation data, what happens if data is very sparsely available e.g. say out of 4 neighboring grids only one has observation data how are the emissions in other 3 grids updated?

[Response]: We thank the reviewer for the valuable comment. When we train the NN-CTM, the long short term memory (LSTM) block is employed to capture the temporal information, and the convolution (U-Net) is employed to capture the spatial information (e.g. the emission inventory, meteorological information, and geographic information of its neighbor grid). That is to say, in NN-CTM, the convolution neural network will capture the surrounding grids' information within the receptive field, and we make a detailed introduction about the receptive field in the answer of next comment which represents the transmission between different grids. Therefore, as shown in Fig. 1, if only the red gird has observation data, the surrounding blue grids' emission inventory within the receptive field will also be updated. At the same time, the grids with a longer distance will have a lower update weight. In extreme circumstances, if we have no observation data, our method will not work as we have no more information to adjust the emission inventory. If the observation data is denser, the emission inventory estimation is more accurate as it can consider more observation data.

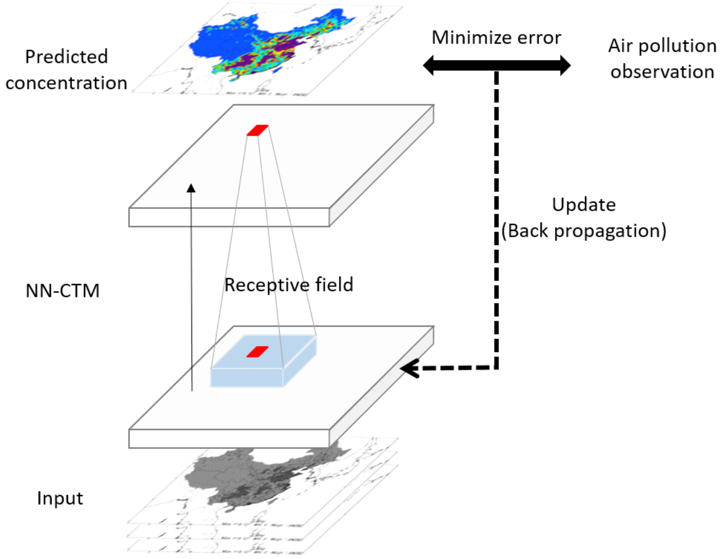

Figure 1: The visualization of neighbor emission update.

We have clarified the relation between observation data and emission inventory in the revised manuscript, as follows:

(Section 2.3, Paragraph 1) *"The observation data will help update the surrounding grids' emission inventory within the receptive field. However, in extreme circumstances, if we have no observation data, our method will not work as we have no more information to adjust the emission inventory. If the observation data is denser, the emission inventory estimation is more accurate as it can consider more observation data."*

[Comment]: Does the deep learning process consider the impact of transmission between different grids? The authors are suggested to explain this point in detail.

[Response]: We thank the reviewer for the valuable suggestion to improve the quality of the paper. The deep learning process has considered the impact of transmission between different grids. The NN-CTM, which refers to U-Net branch in particular, employs the convolution neural network to utilize neighbor information effectively. We visualize a demo case of 3×3 convolution and 5×5 convolution in Fig. 2. In U-Net, the stacked of convolution can get the neighbor information with a bigger receptive field (e.g. stacking 5×5 convolution and 5×5 convolution can get a 9×9 convolution), the non-linear function (P-RELU) is employed to improve model fitting with nearly zero extra computational cost and little overfitting risk, and the batch normalization and dropout are employed to enhance the robustness of the model. We calculate that the receptive field of our model is 38×38 grid. In other words, the predicted pollutant concentration is related to its surrounding 38×38 grid's information, which represents the transmission between different grids. Meanwhile, the closer the distance, the greater the contribution.

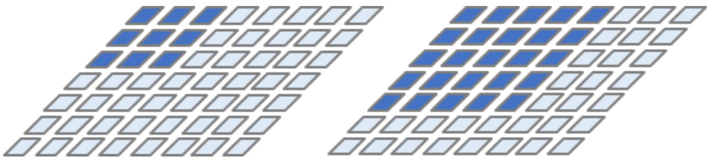

Figure 2: The visualization of convolution neural network (left: 3*3 kernel size, right: 5*5 kernel size).

We have clarified the impact of transmission between different grids in the revised manuscript, as follows:

(Section 2.2, Paragraph 3) *"In U-Net, the stacked of convolution can get the neighbor information with a bigger receptive field (e.g. stacking 5×5 convolution and 5×5 convolution can get a 9×9 convolution), the non-linear function (P-RELU) is employed to improve model fitting with nearly zero extra computational cost and little overfitting risk, and the batch normalization and dropout are employed to enhance the robustness of the model. We calculate that the receptive field of our model is 38×38 grid. In other words, the predicted pollutant concentration is related to its surrounding 38×38 grid's information, which represents the transmission between different grids. Meanwhile, the closer the distance, the greater the contribution."*

[Comment]: Lines 24-26, Abstract. Please be specific on the simulation year and the emission inventory you applied.

[Response]: We apologize for missing this information, and we have added year 2015 in Abstract.

[Comment]: Line 310, Page 15. I suggest the authors add more description for Figure 8, such as explaining why the performance of using the new emission inventory worsened at some sites.

[Response]: We appreciate the reviewer for the valuable suggestion. We have added more explanations accordingly as follows:

(Section 3.4, Paragraph 2) *"The model performance of most stations has been improved, and a small number of stations with worsen performance show the link between compound pollutants. For example, stations with larger deviations between $PM_{2.5}$ simulation results and observations tend to have greatly improved O_3 performance, and vice versa."*

[Comment]: The language of the manuscript needs to be further polished.

[Response]: We thank the reviewer for the comment, and we have further polished the manuscript and checked grammar carefully. All modifications will be marked in the revised manuscript.