We thank the reviewer for the positive and constructive comments. In the following, please find our point-by-point responses.

Question1: Did the authors only used those grid cells with ozone observations for training, and how many of those grid cells are there?

Response: Yes, only grid cells surface with ozone observations were used for training. Since there is mismatch between the site and grid scales, observations of all sites within a grid cell were averaged to be consistent with grid-level predictor variables. Finally, we obtained a total of 643 grid cells with surface ozone observations across China, the number of which was 857 less than the number of sites (1500). We are sorry for not clearly stating this pre-processing step in the manuscript and will clarify this point in the revision of the manuscript.

Question2: Why did the authors chose a time window of 24 hours for training, is other time window possible?

Response: To determine the optimal time window for training, we conducted several experiments with eight different lookback windows, including 1, 2, 6, 12, 24, 36, 48 and 72 hours. The model was found to perform best at the 24-hour time window, with the highest coefficient of determination (R^2 = 0.75) value and the lowest root mean square error (RMSE = 12.84 ppb; mean = 32.58 ppb). The detailed experiment results using different lookback windows are shown in the table below. The above statements will be included in the revised version.

Table R1. Detailed results of model tests at eight lookback windows.

Lookback windows (hour)	R^2	RMSE (ppb)
1	0.69	14.49
2	0.68	14.74
6	0.73	13.56
12	0.74	13.22
24	**0.75**	**12.84**
36	0.71	13.98
48	0.74	13.09
72	0.74	13.13

Question3: The authors carried out a 10-fold cross validation for hyperparameter optimization, which should be followed by a model performance evaluation with the testing data. I might have missed the information on model testing, but how does the model perform with the testing data?

Response: The reviewer seems to have misunderstood our parameter tuning process. To determine the optimal hyperparameters (including epoch, batch size, number of neurons and optimizer), we first conducted a sensitivity analysis to identify the

importance of each hyperparameter. Specifically, each hyperparameter was assigned a prior range and the whole dataset was partitioned into the training data and the validation data using a ratio of 9:1. We adopted a one-at-a-time (OAT) strategy, i.e., changing one parameter at a fixed interval while keeping others unchanged, to avoid too much consumption of the high-performance computer resources. The results showed that changes in hyperparameter had minor effects on model performance (the R^2 and RMSE values were nearly stable being around 0.7 and 10 ppb, respectively). Thus, the mean value of the specific range for each of the hyperparameters was used. Figure R1 shows the convergence of the loss function using the final hyperparameters. The detailed hyperparameter settings can refer to Section 2.2 of the manuscript (Lines 224–237).

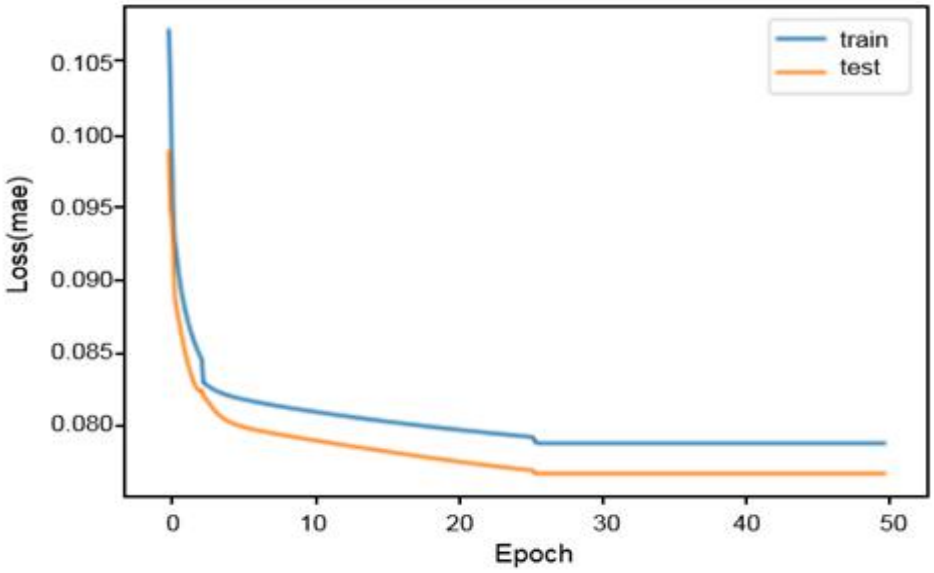

Figure R1. Loss functions of the LSTM model for training data and testing data, respectively, at each epoch.

The 10-fold cross-validation was used to evaluate the predictability capability and robustness of the LSTM algorithm and the determined hyperparameters in varied conditions of data availability. The results (Figure R2) were satisfied and showed no degradation of the model performance. But as the reviewer pointed out, an independent testing data is essentially needed to ensure the generic function of the final trained LSTM model. Here we collected extra data at over 900 monitoring sites across China in 2014 and used them as test data. The model performance is shown in Figure R3 (R^2 = 0.62, RMSE = 18.75 ppb; it is noted that the available data is much less than those in the years 2015-2020), comparable to the final LSTM model trained using data during 2015-2020 and the 10-fold cross-validation. In the revision version, we will also try to train the LSTM model using data between 2015-2019 and use the data in 2014 and 2020 as an independent data.

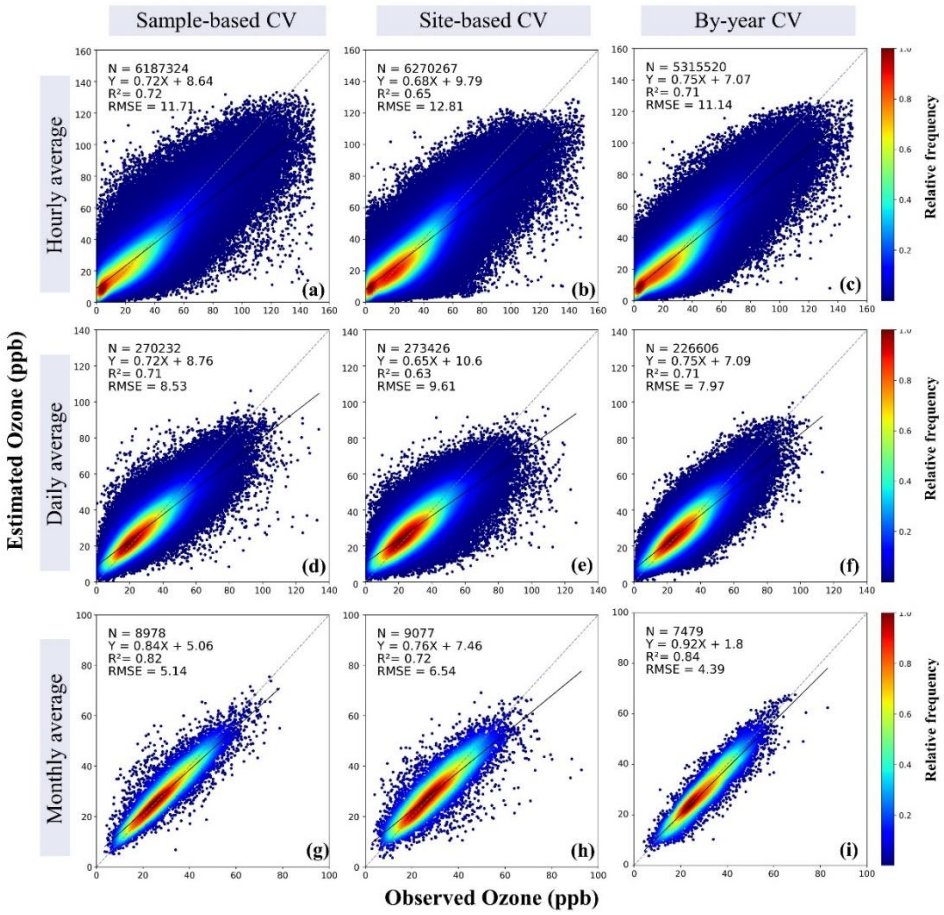

Figure R2. Comparisons between model estimated surface ozone concentrations and observations across China. The panels are sample-based cross validations at hourly, daily and monthly time-steps (a, d, g), site-based cross validations at hourly, daily and monthly time-steps (b, e, h), and by-year cross validations at hourly, daily and monthly time-steps (c, f, i). The dashed and black lines represent the 1:1 lines and the linear regression lines, respectively.

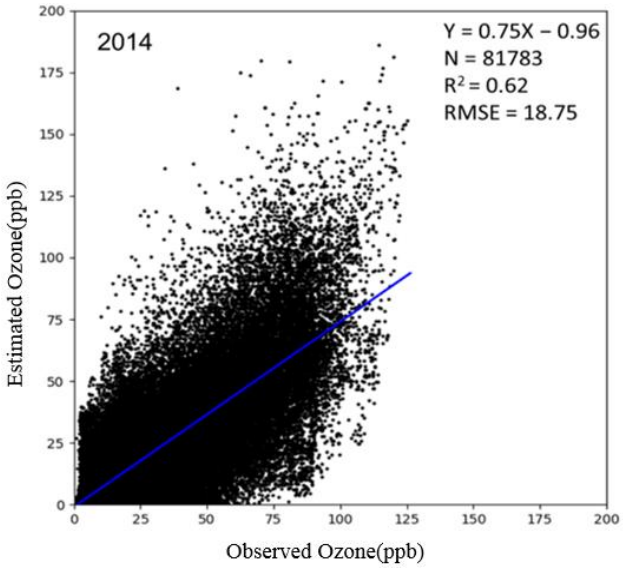

Figure R3. Comparisons between model estimated surface ozone concentrations and observations across China in 2014.