Supplement Information for

Unleashing the Potential of Geostationary Satellite Observations in Air Quality Forecasting Through Artificial Intelligence Techniques

Chengxin Zhang[1], Xinhan Niu[1], Hongyu Wu[2], Zhipeng Ding[2], Ka Lok Chan[3], Jhoon Kim[4], Thomas Wagner[5], Cheng Liu[1,6,7*]

[1]Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, 230026, China
[2]School of Environmental Science and Optoelectronic Technology, University of Science and Technology of China, Hefei, 230026, China
[3]Rutherford Appleton Laboratory Space, Harwell Oxford, United Kingdom
[4]Department of Atmospheric Sciences, Yonsei University, Seoul, Republic of Korea
[5]Satellite Remote Sensing Group, Max Planck Institute for Chemistry, Mainz, Germany
[6]Key Laboratory of Environmental Optics and Technology, Anhui Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Hefei, 230031, China
[7]Key Laboratory of Precision Scientific Instrumentation of Anhui Higher Education Institutes, University of Science and Technology of China, Hefei, 230026, China

[*]Correspondence: Cheng Liu (chliu81@ustc.edu.cn)

This file contains supplementary text S1-S2 and figures S1-S12.

Supplementary Text

S1. Data pre-processing

Outlier Handling

We conducted outlier handling for each GeoNet input datasets using z-scores, wherein data normalization was performed based on the mean and standard deviation. Data points exceeding a certain threshold of z-scores were discarded. The calculation formula is as follows:

$$z(x) = \frac{x - \mu_x}{\sigma_x}$$

Where x is the data value, μ_x and σ_x are the mean average and standard deviation.

Missing Value Handling

Due to meteorological factors, the GEMS dataset used in this study contains many missing values. Fig. S1 presents the overall missing ratio of GEMS satellite NO_2 retrieval for each ground pixel in 2021.

To enhance data availability, the GEMS dataset underwent imputation procedures. Various data imputation methods were employed to assess their impact on the dataset, including zero imputation, WRF data imputation, and CAMS data imputation. Specifically, missing data points were replaced with either zero or corresponding data from the WRF and CAMS datasets at the respective spatiotemporal positions. For other datasets, missing values were addressed through spatiotemporal interpolation using multidimensional linear interpolation.

Resampling

Due to variations in spatiotemporal resolutions among different datasets, it was necessary to ensure data consistency and facilitate model computation by resampling all datasets in both time and space domains. Resampling operations involved both upsampling and downsampling. Upsampling was achieved through interpolation, while downsampling was performed using local mean aggregation. Following resampling, the temporal resolution of all datasets was standardized to 1 hour, and the spatial resolution to 0.1 degrees.

Normalization

The normalization process applied here is beneficial for overcoming overfitting issues during model training and dealing with heterogeneous data of different scales, thereby potentially accelerating training speed. This process is essential for bringing each variable to a comparable scale, ensuring that each feature carries similar importance. In this study, min-max normalization was applied to all datasets. In this method, the maximum value of the data is transformed to 1, the minimum value to 0, and other values are scaled to decimals between 0 and 1. The calculation method is as follows:

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}}$$

Where $x,\ x_{max}, x_{min}$ is the data value, maximum, and minimum, respectively.

S2. The configuration and optimization of GeoNet

For the GeoNet model, the model configurations and hyperparameters such as the optimizer, loss function, L1 or L2 regularization, dropout, training steps, and epochs can make a difference to the model performance including the prediction accuracy and generalizability. Thus, several scenarios of model hyper-parameters have been tested during the model training phase. The model accuracy on validation datasets and the learning rate curve were used to diagnose the model hyperparameters. We used the following metrics of model performance in this study:

The coefficient of determination (R^2):

$$R^2 = \frac{\sum_{i=1}^{m}(f(x_i) - \bar{y})^2}{\sum_{i=1}^{m}(y_i - \bar{y})^2}$$

The root mean square error (RMSE):

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2}$$

The mean absolute error (MAE):

$$\frac{1}{n}\sum_{i=1}^{n}|\hat{y_i} - y_i|$$

The mean absolute percentage error (MAPE):

$$\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y_i} - y_i}{y_i}\right|$$

The model parameters mainly include the number of layers and the dimensions of the hidden layers, both control the model's capacity. If the model capacity is relatively small, underfitting may occur; overfitting may exist if it is too large. Therefore, selecting an appropriate model capacity is crucial for improving model performance. During the pre-training process, the model is trained by combining different numbers of layers and dimensions of the hidden layers. The Mean Squared Error (MSE) Loss is recorded for each training iteration, and a heatmap is generated as shown in Fig. S2. From the heatmap, it can be observed that when the number of layers is 2 and the dimension of the hidden layer is 256, the model achieves the minimum MSE Loss. Fig. S3 shows the sensitivity test results of model loss varying with different batch size settings, indicating that a batch size of 64 is optimal. Based on the model's MSE loss under different hyperparameter configurations, the best fitting model can be selected.

The Adam optimization algorithm controls the learning rate, which can design independent adaptive learning rates for different parameters. The three initialization parameters ϵ, $\rho 1$, and $\rho 2$ of the Adam algorithm are set to be 0.0001, 0.9, and 0.99, respectively. For the epoch, its size is controlled by the early stop method. The early stop method monitors the change of the model's loss function on the validation set during the training process and stops the model training immediately when the validation loss of the model starts to become larger. Due to the fluctuation of the loss function, a threshold p is set for the early stopping method in practice, and when the validation loss of the model becomes large for p consecutive epochs, the model is rolled back to the lowest validation loss and the training is stopped, and the threshold p is set to 10 in this paper. Fig. S4 shows a typical learning curve of the MSE loss in training and validation data sets for different learning steps in training an optimal model. Such diagnostics can be used to avoid the model overfitting.

Supplementary Figures

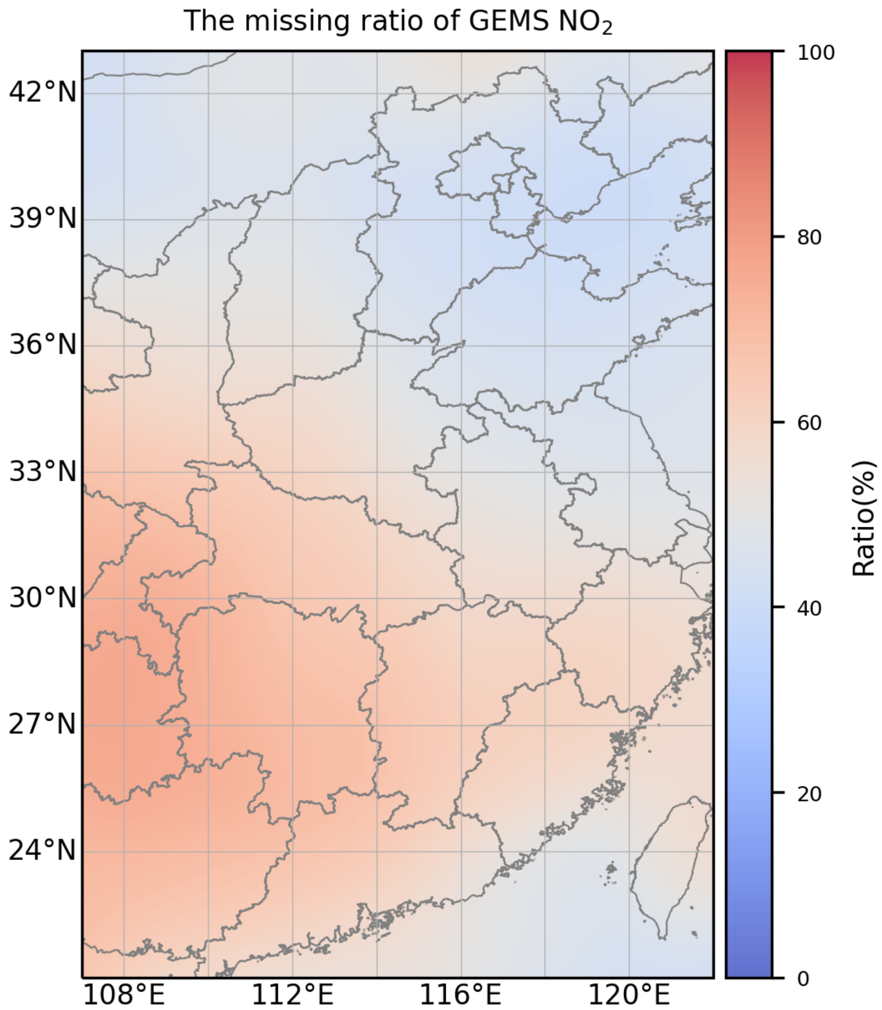

Figure S1. The ratio of missing data for hourly GEMS NO_2 retrievals over East China in 2021.

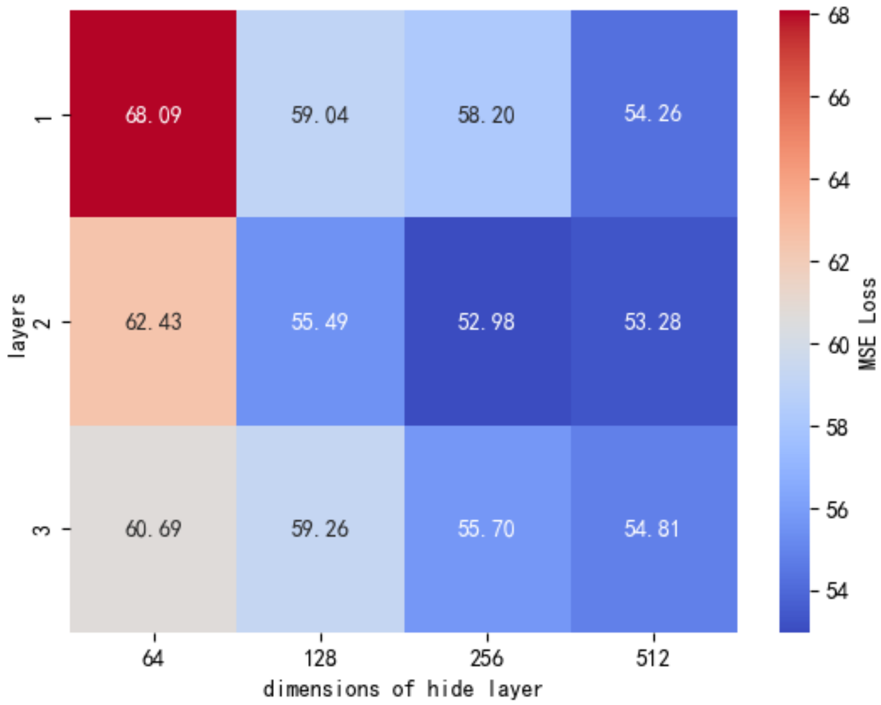

Figure S2. The influence of model hyperparameters including both ConvLSTM layers and dimensions of hide layer on the MSE loss of GeoNet prediction.

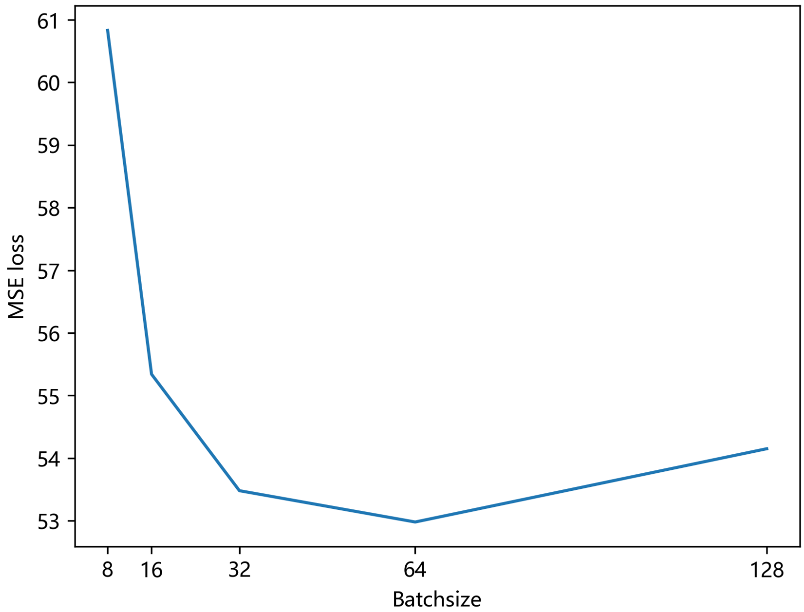

Figure S3. The impact of batch size on the MSE loss of GeoNet prediction.

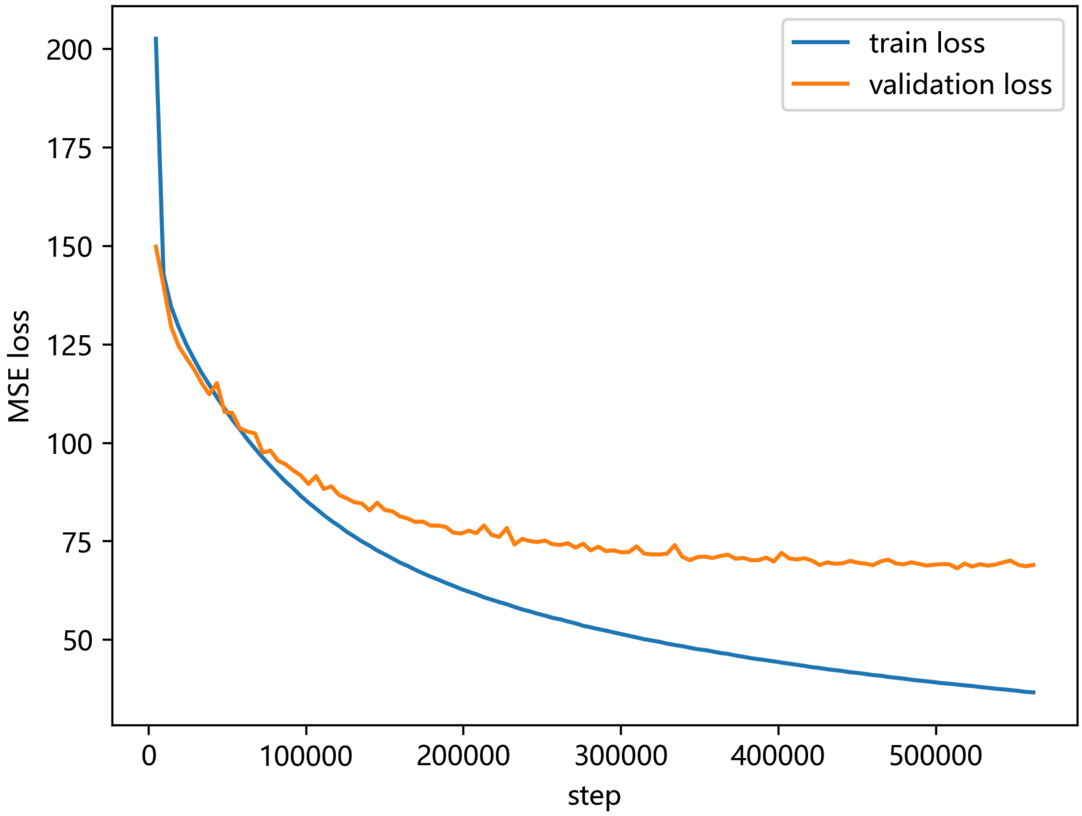

Figure S4. The learning curve of model loss in validation and training datasets for different steps.

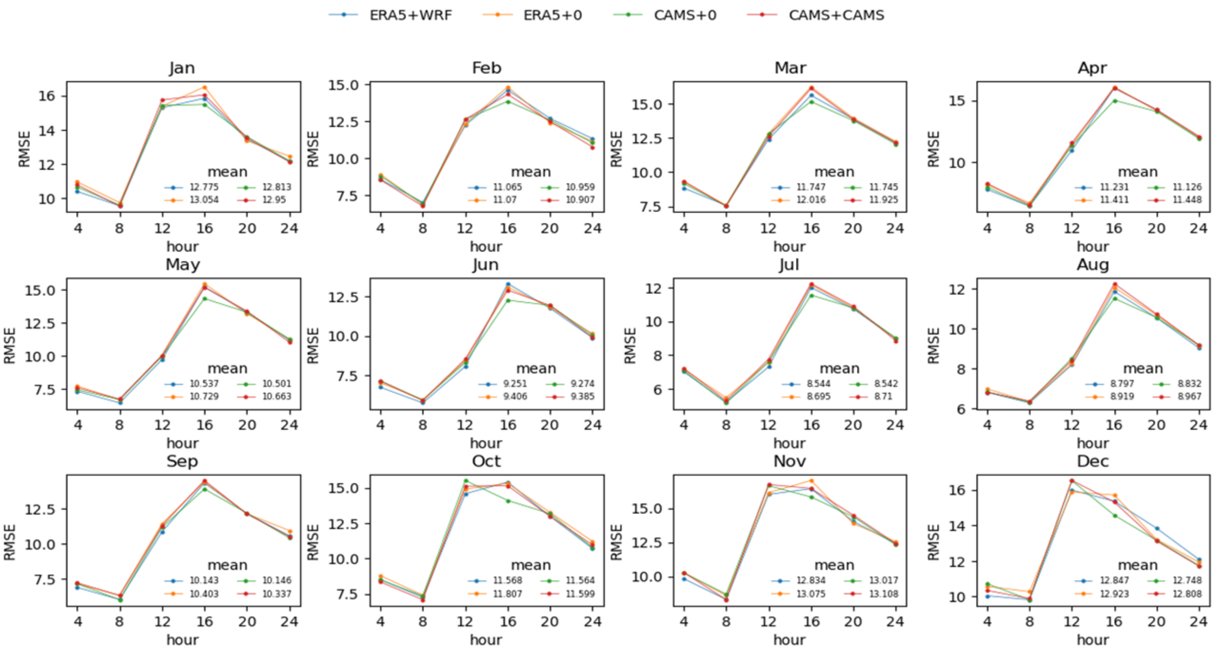

Figure S5. The RMSE of GeoNet predicted-NO_2 varys with different prediction step from t+4h to t+24h, for different months.

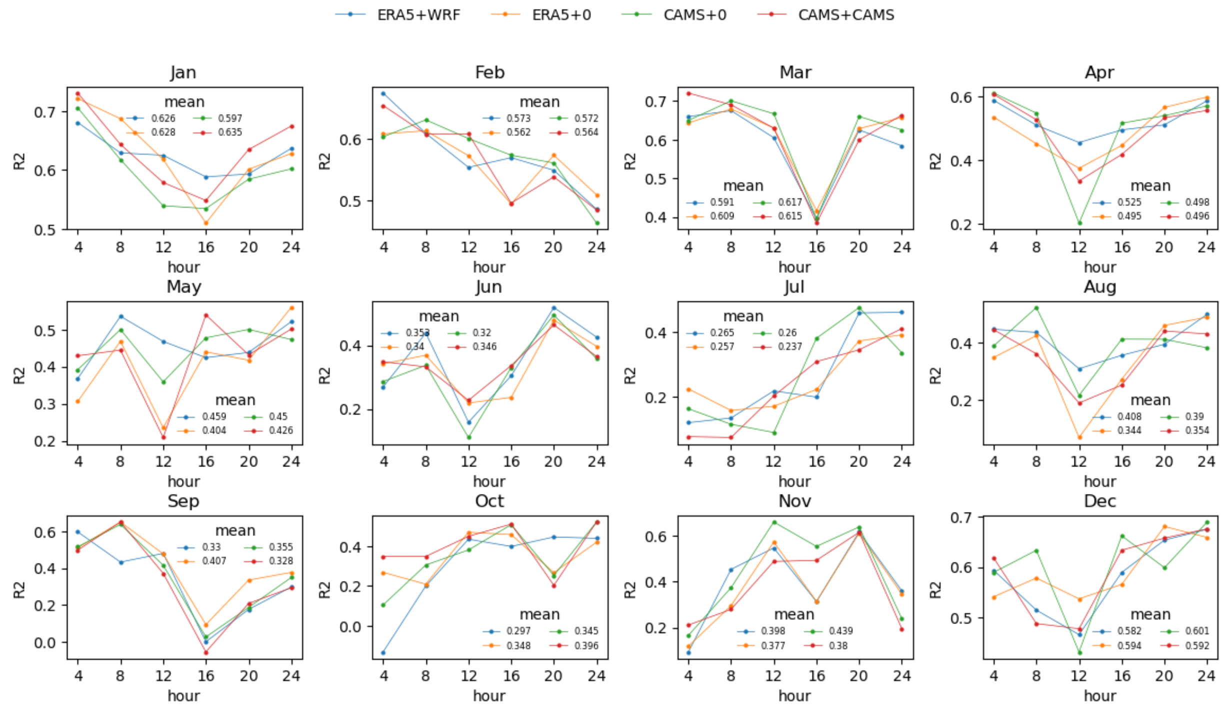

Figure S6. Similar to Fig. S5, but for R^2.

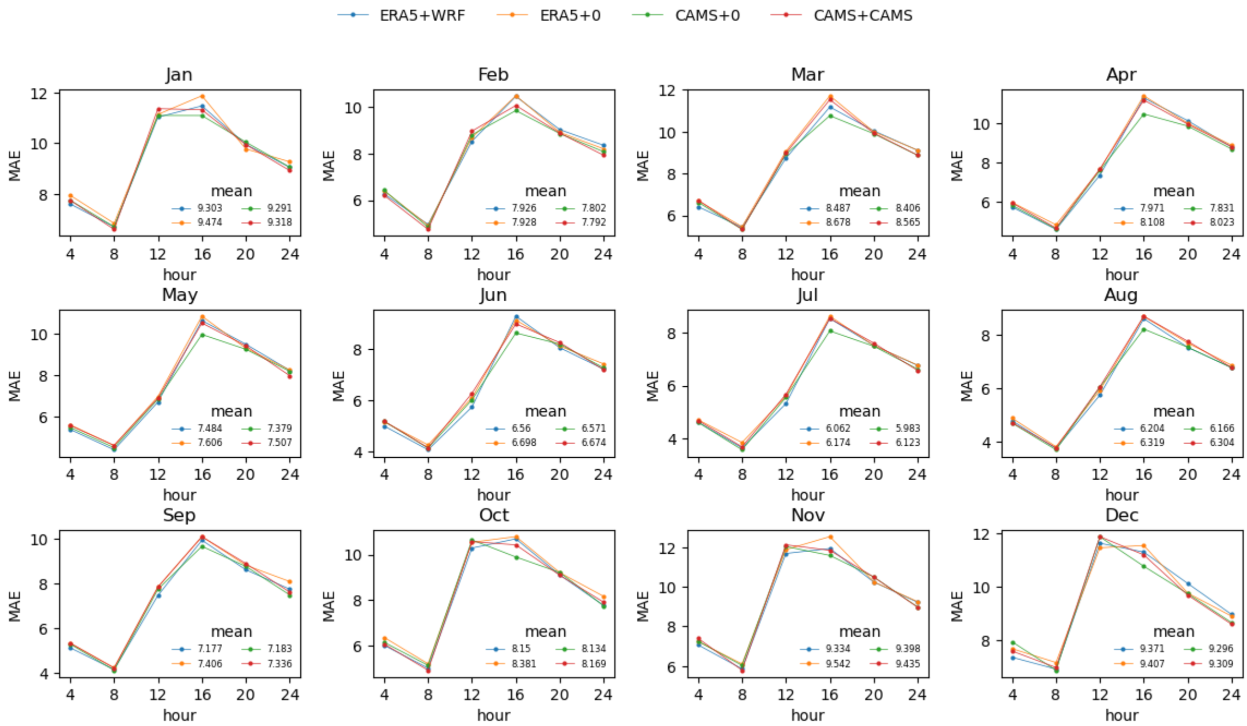

Figure S7. Similar to Fig. S5, but for MAE.

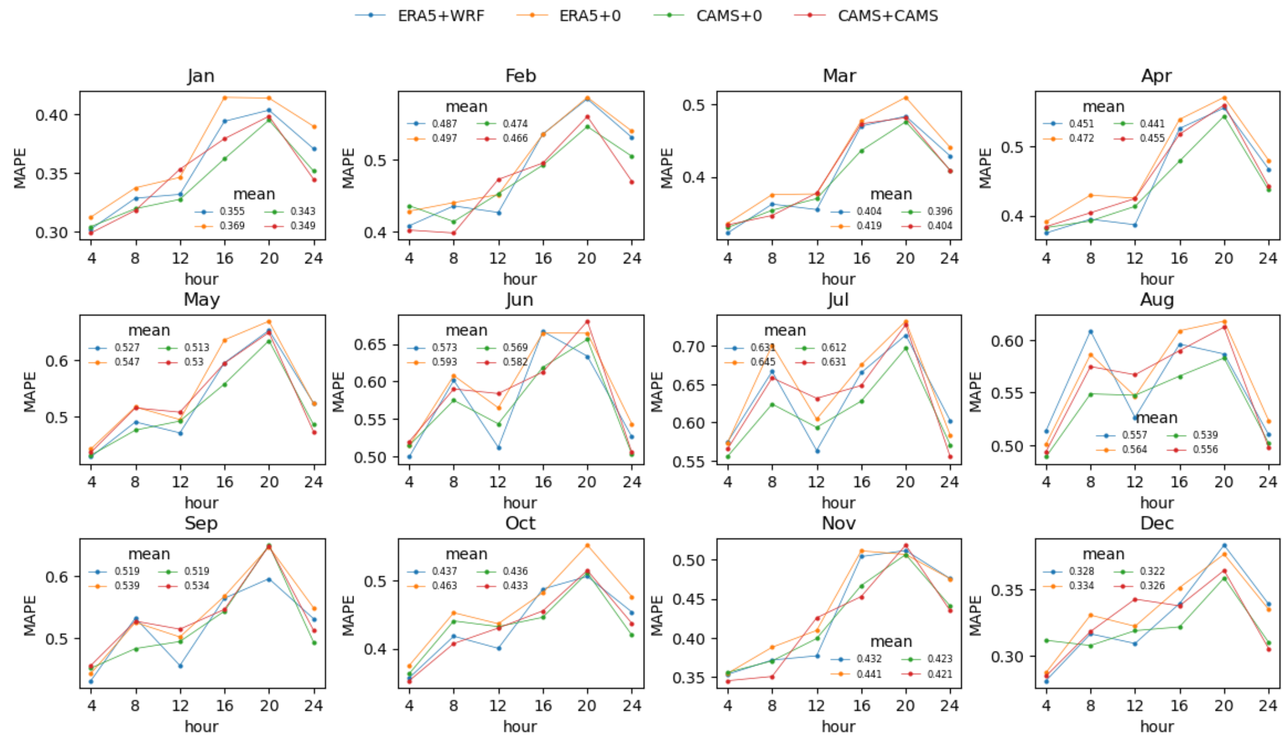

Figure S8. Similar to Fig. S5, but for MAPE.

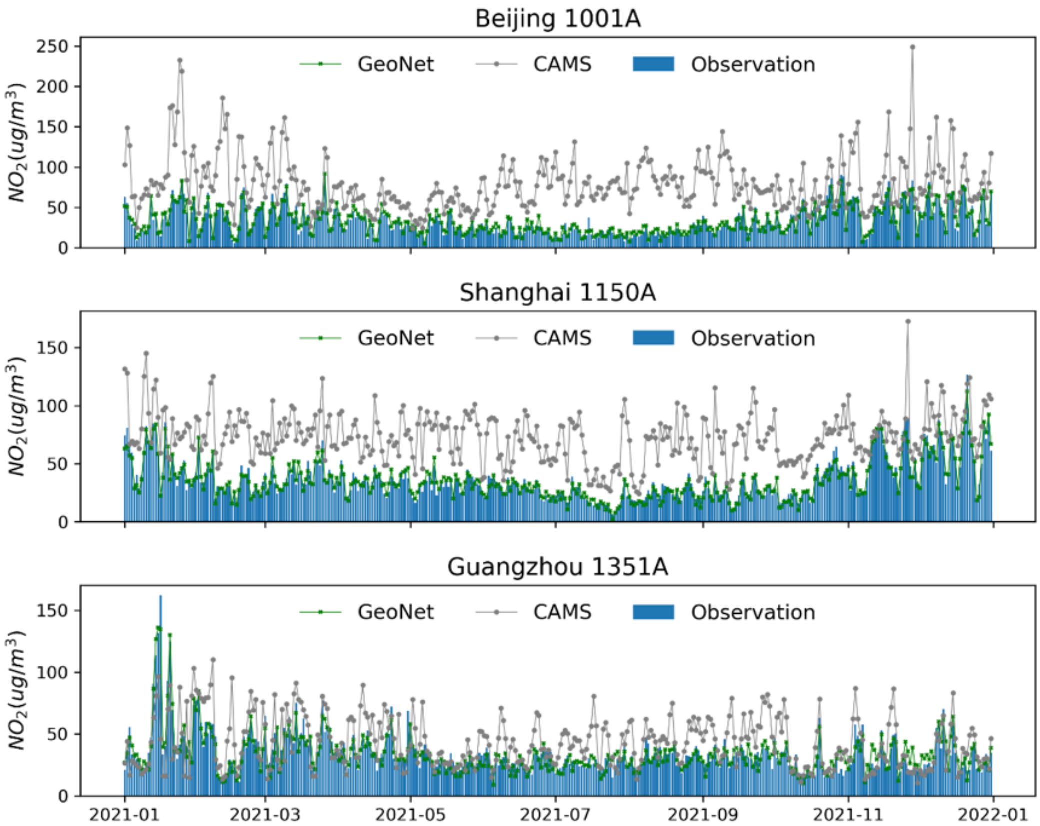

Figure S9. Time series comparison of daily t+4h prediction of surface NO_2 concentration among GeoNet and CAMS prediction, as well as the CNEMC measurements. These results are shown for one typical site in (**a**) Beijing, (**b**) Shanghai, and (**c**) Guangzhou, respectively.

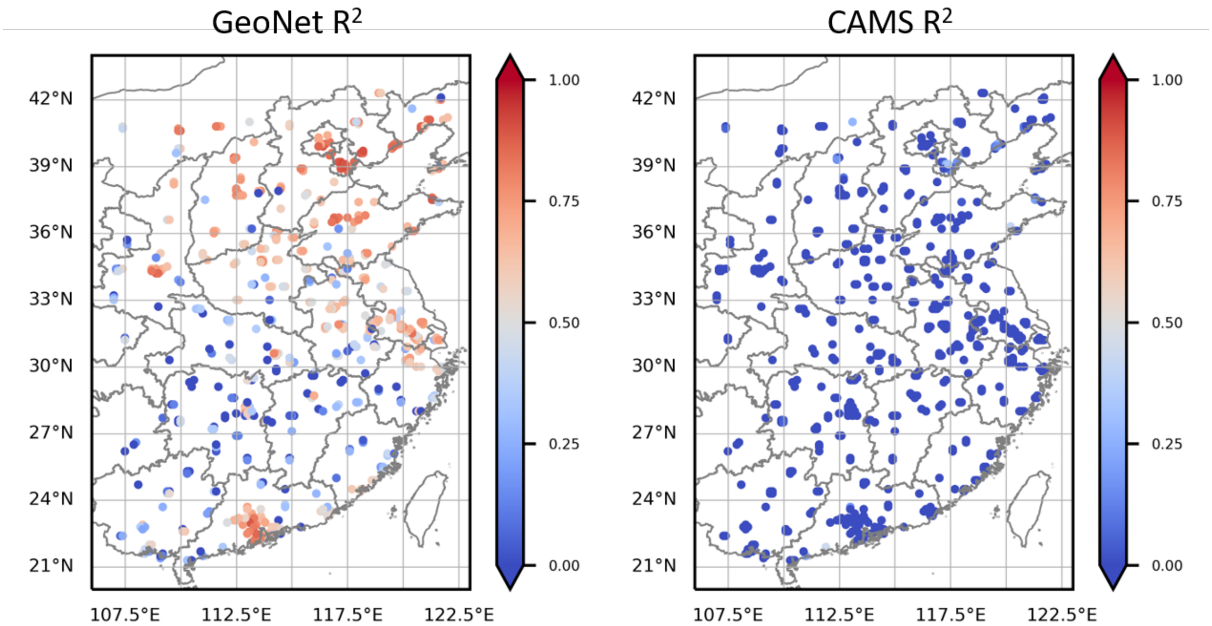

Figure S10. The site-specific Pearson's R^2 between the CNEMC measurements and NO_2 prediction by (a) GeoNet, and (b) CAMS over East China.

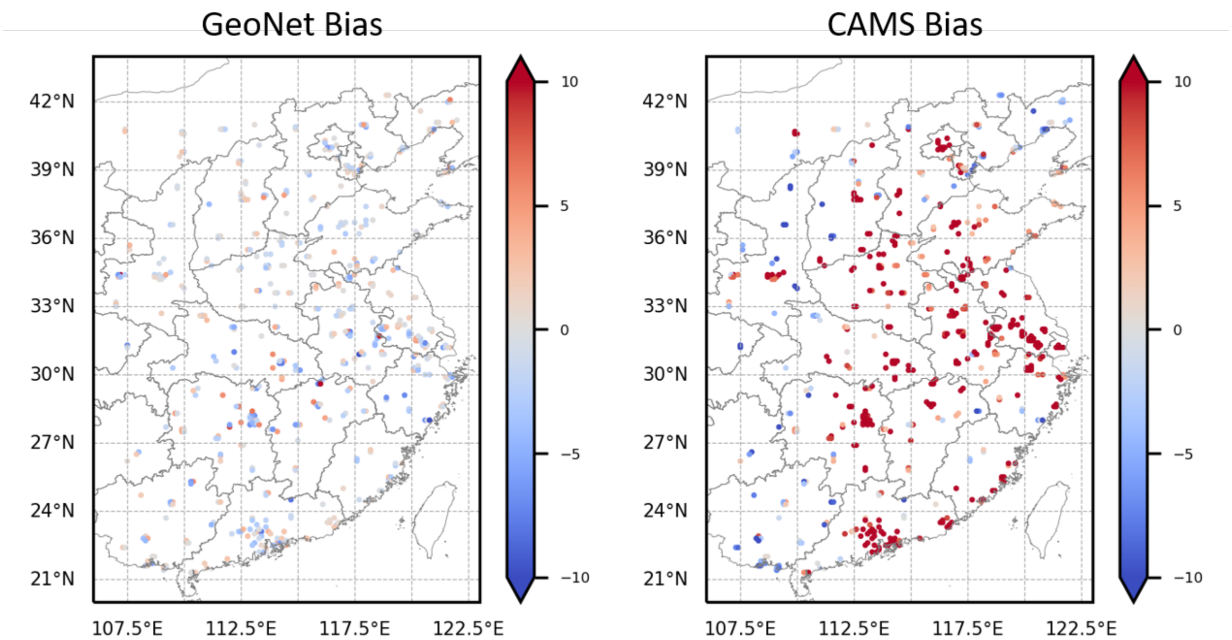

Figure S11. Similar to Fig. S10, but for RMSE.

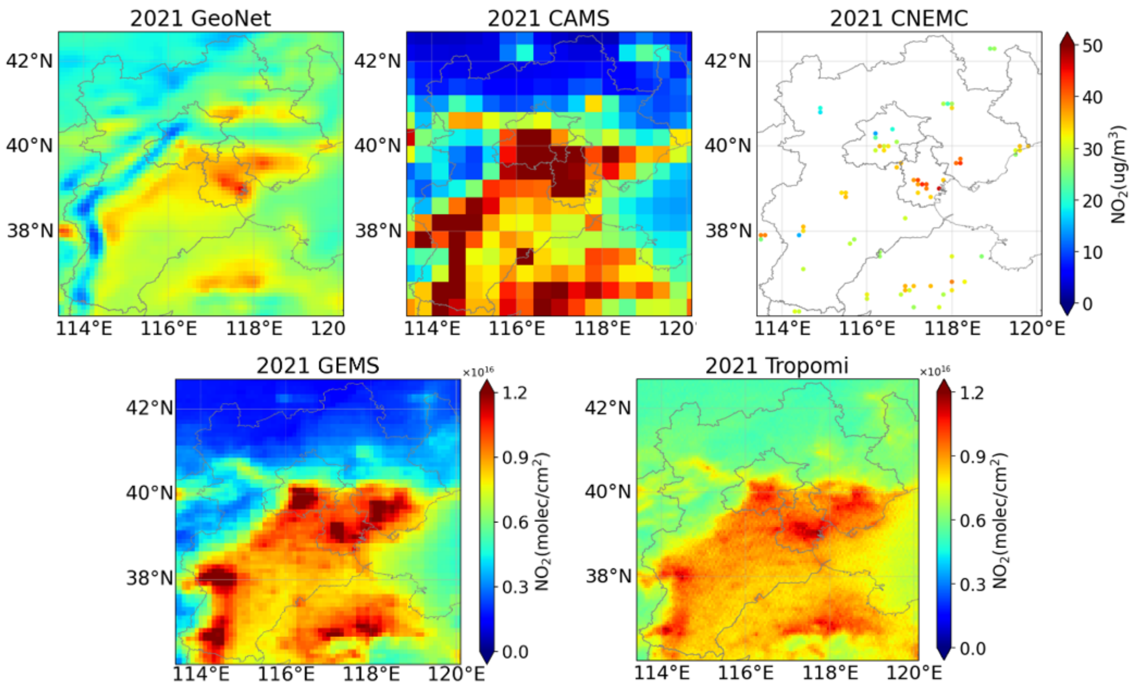

Figure S12. The comparisons of annual NO_2 distribution among GeoNet, CAMS, and CNEMC (top panel), as well as the tropospheric NO_2 column from GEMS and TROPOMI over East China in 2021 (bottom panel).