atmosphere

MDPI

Supplementary Materials

A simple and effective Random Forest refit to map the spatial distribution of NO_2 concentrations

Yufeng Chi[1,*] and Yu Zhan[2]

To supplement the content of the main article, the Supplementary Section includes Method S1, Figures S2–S9 and Supplement discussion.

1. Method S1 (Iterate TWS)

Recently, we developed a moving small window two−step (TWS) model for recovering missing data from multiple remote sensing products. This model reports presents an excellent performance for AOD data recovery. At the same time, the model can be applied to the recovery of space−time gaps of most remote sensing products (Chi et al. 2020). Studies have shown that a single operation of TWS can reduce the AOD missing rate from 88% to 10%, which has been cross−validated with the ground AERONET network, with R=0.87 and RMSE=0.23. There is no Aerosol Robotic Network (AERONET) site in SWFJ. Therefore, data gaps are randomly established, and CV is used to verify the recovery results and the original data. The first step of the TWS model uses the LightGBM machine learning method, and the second step uses a multimode moving window spatiotemporal interpolation method (STW). TWS can be used in two steps in combination or independently. Among them, MAIAC AOD uses the first step and iterative second step of TWS, and the OMI NO_2-column is recovered through the iterative second step of TWS. The technical route is shown in Fig.S1. The TWS details are as follows:

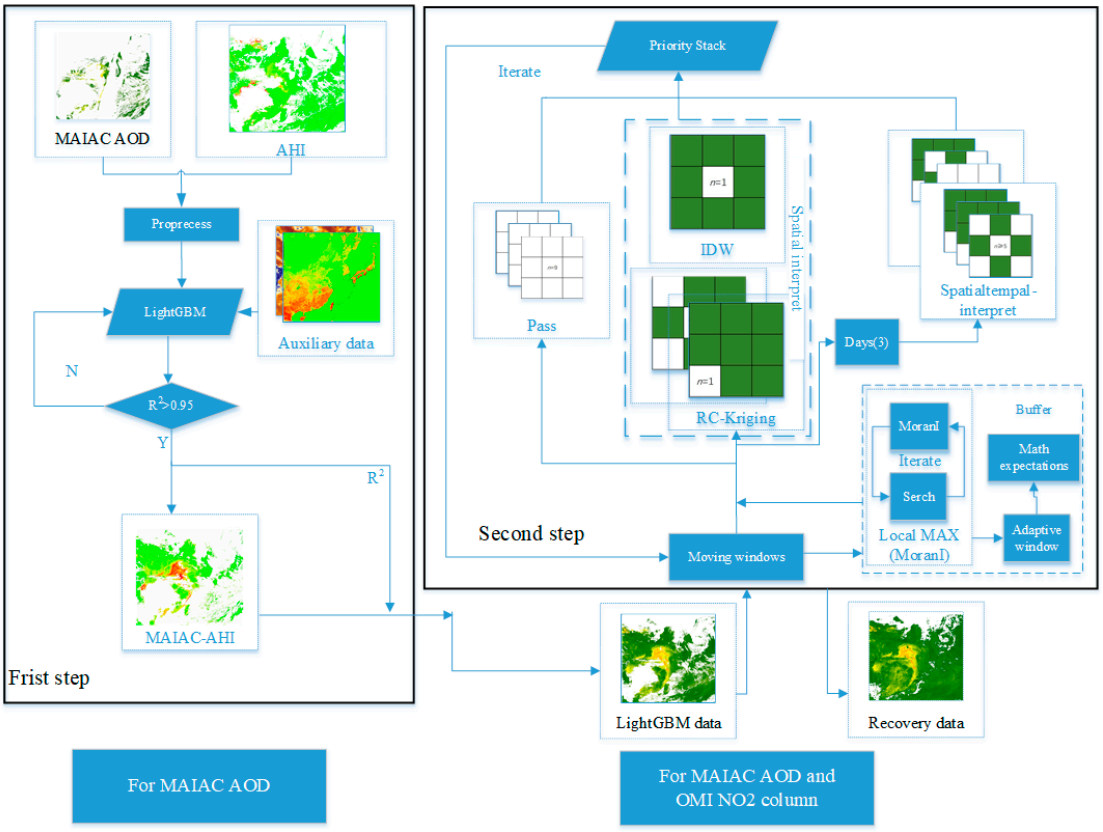

Figure S1. Iterate TWS technology roadmap.

1.1. First Step of TWS

The formula to restore MAIAC AOD using LightGBM is as follows:

$$AOD_{maiacpre} = Lg\ (AHI_{470nm}, RID, MET, ELE, SL, POP, NDVI, RL, LU, DOY) \quad \text{(S1)}$$

where, Lg represents the LightGBM; AHI_{470nm}, RID, MET, ELE, SL, POP, $NDVI$, RL, LU, and DOY represents the 470 nm AHI AOD, random ID, meteorological parameters (temperature, air pressure, wind speed, humidity), altitude, slope, NDVI, road length, land use, and day of year, respectively.

1.2. Second Step of TWS

1.2.1. Design of Moving Window Size and Selection of Interpolation Mode

The size of the mobile window is 3*3. Set four scenarios for TWS:

(1) Use Inverse Distance Weight interpolation (IDW) interpolation when center pixel is missing in moving window.

(2) The RC−Kriging method is used when five or fewer pixels are missing from the moving window.

(3) We used spatiotemporal weight interpolation when the number of missing cells of Day 2 was greater than or equal to 5 and the number of valid pixels of Day 1 or Day 3 was greater than or equal to 5.

(4) When there were too few pixels in the moving window for three consecutive days (Day 2 had no valid pixels and the number of valid pixels for Days 1 and 3 were fewer than 5 pixels), we ignored this part of the calculation.

1.2.2. Buffer Factor

The mathematical expectation of the moving window pixels is used as a buffer factor to correct the bias. The formula is as follows:

$$\text{MoranI} = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} G_{ij}(p_i - \bar{p})\left(p_j - \bar{p}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{n} G_{ij}\sum_{i=1}^{n}(p_i - \bar{p})^2} \quad \text{(S2)}$$

$$G_{ij} = 1/\sqrt{(i_x - j_x)^2 + \left(i_y - j_y\right)^2}$$

$$w \leftarrow Scope\ Window \leftrightarrow\ Max(\mathrm{MoranI}_{w-1}, \mathrm{MoranI}_{w}, \mathrm{MoranI}_{w+1})$$

$$E_w = (\sum_{i=1}^{w*w} S_i)/w^2$$

$$P_{(S_{tk},E_{t2})} = \frac{\sum_{j=1}^{n}(S_{tk} - E_{t2w})\,(S_{tk} - \overline{\tau_{t2}})}{\sqrt{\sum_{j=1}^{n}(S_{tk} - \overline{\tau_{tk}})^2 \sum_{j=1}^{n}(S_{tk} - \overline{\tau_{t2}})^2}} \quad k \in (1,3)$$

where MoranI represents the Global Moran's I. Here, n represents the number of valid pixels; p_i and p_j represent the values of the two pixels, i and j ; $\bar{x}$ represents the average value of the pixels; $dis(i,j)$represents the spatial distance between the two pixels, i and j; $G_{i,j}$ represents the inverse distance weight; $Scope\ Window$ represents the window that corresponds to the maximum local MoranI, $Scope\ Window$ is a square; w represents the number of pixels on one side of the square a $Scope\ Window$; $\leftrightarrow$ represents iterative search for the $Scope\ Window$; $\leftarrow$ represents obtaining w; $S_i$represents the value in the $Scope\ Window$; S_{tk} represents the value in the $Scope\ Window$ on day tk; E_w represents the mathematical expectation in the $Scope\ Window$ (buffer factor); and $P_{(S_{tk},E_{t2})}$ represents the Spearman correlation coefficient between day tk and day $t2$.

1.2.3. Spatial Interpolation Method (IDW and RC Kriging)

The formulas of IDW and RC Kriging are as follows:

$$\begin{cases} Z_1 = \left[\sum_{i=1}^{N}\sum_{j=1}^{N} G_{i,j}(S_{i,j} - E_w)\right] + E_w \\ \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda 2_{i,j} \times Cov(s_{i,j}) - \mu = Cov(s_{j,i}) \\ \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda 2_{i,j} = 1 \\ Z_2 = \left[\sum_{i=1}^{N}\sum_{j=1}^{N} \lambda 2_{i,j}(S_{i,j} - E_w)\right] + E_w \end{cases} \quad (S3)$$

Where Z_1 and Z_2 represent the estimates produced by IDW and RC Kriging interpolation, $G_{i,j}$ represents the inverse distance weight, $s_{i,j}$ represents the value at points i and j, μ presents the Lagrange multiplier, $\lambda 2_{i,j}$ represents the weight, , $Cov(s_{i,j})$ and $Cov(s_{j,i})$ represent the covariance of $s_{i,j}$ and $s_{j,i}$, and E_w represents the mathematical expectation in the $Scope\ Window$(buffer factor).

1.2.4. Spatiotemporal Weight Interpolation (STW)

The formulas of STW are as follows:

$$Z_{ST_0} = \sum_{tc=1}^{3}\left(\sum_{j=1}^{N_t}\left(\left[\sum_{i=1}^{N_t}\left(\lambda_{tk_{i,j}}\left(S_{t_{i,j}} - E_{tc}\right)\right)\right] + E_{tc}\right)\right)$$

$$\lambda_{tn} = \sum_{j=1}^{N}\sqrt{\left(1 - \left[{}^{(P_{(S_{t1},E_{tn})} + P_{(S_{tn},E_{t3})})}/_{2}\right]\right)\left(\frac{{}^{1}/_{distance(tn_i,tn_j)}}{\sum_{i=1}^{N}\left({}^{1}/_{distance(tn_i,tn_j)}\right)}\right)} \quad \mathrm{n} = 2 \quad (S4)$$

$$\lambda_{tn} = \lambda_{(tn,t2)} = \sum_{j=1}^{N} \sqrt{\left(\frac{P_{(S_{tn},E_{t2})}}{2}\right)^2 + \left(\frac{1/distance(tn_i,tn_j)}{\sum_{i=1}^{N}\left(1/distance(tn_i,tn_j)\right)}\right)^2} \quad n \in (1,3)$$

$$distance(i,j) = \sqrt{(i_x - j_x)^2 + (i_y - j_y)^2}$$

where Z_{ST_0} represents the 470 nm MAIAC AOD and NO_2-column estimated by STW, T represents time, $t1$ is the day before the estimated AOD result, $t2$ is the date when the AOD/NO_2-column is estimated, and $t3$ is the second day when the AOD/NOcolumn is estimated, S_t represents the value of effective AOD, E is the mathematical expectation, $E_{tc}$is the global mathematical expectation of Day T, and $(P_{(S_{t1},E_{tn})})$ represents the R between the $t1$ and tn estimated AOD/NOcolumn, λ_{tn} represents the time weight of the nth day (n ∈ (1,2,3)), N is the number of pixels in the moving window (the size of the moving window is 7 pixels), $distance(tn_i,tn_j)$ represents the spatial distance between tn_i and tn_j.

1.2.5. Priority Setting of Overlapping Pixels

Set the priority to IDW > RC Kriging > STW. When pixel restoration causes overlap, fill in missing values according to their priority. Furthermore, if the restoration results are pixel−overlapped in the same way, the average of the restoration result overlaps should be determined as the final result.

Supplement figures

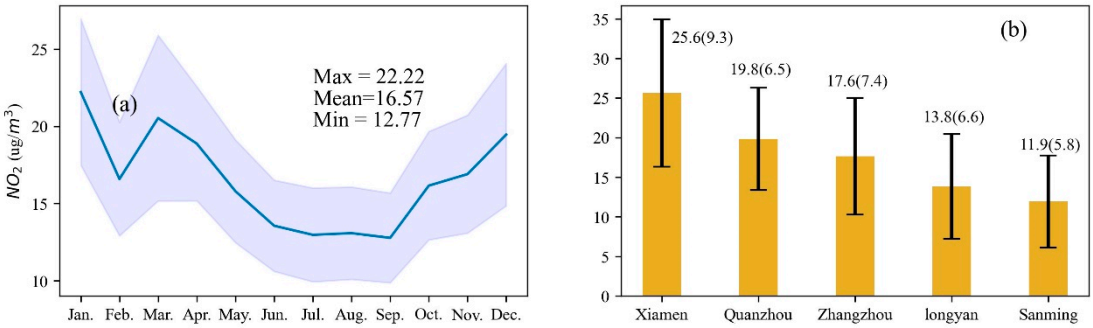

Figure S2. Monthly average of NO2 monitoring data in SWFJ in 2018.

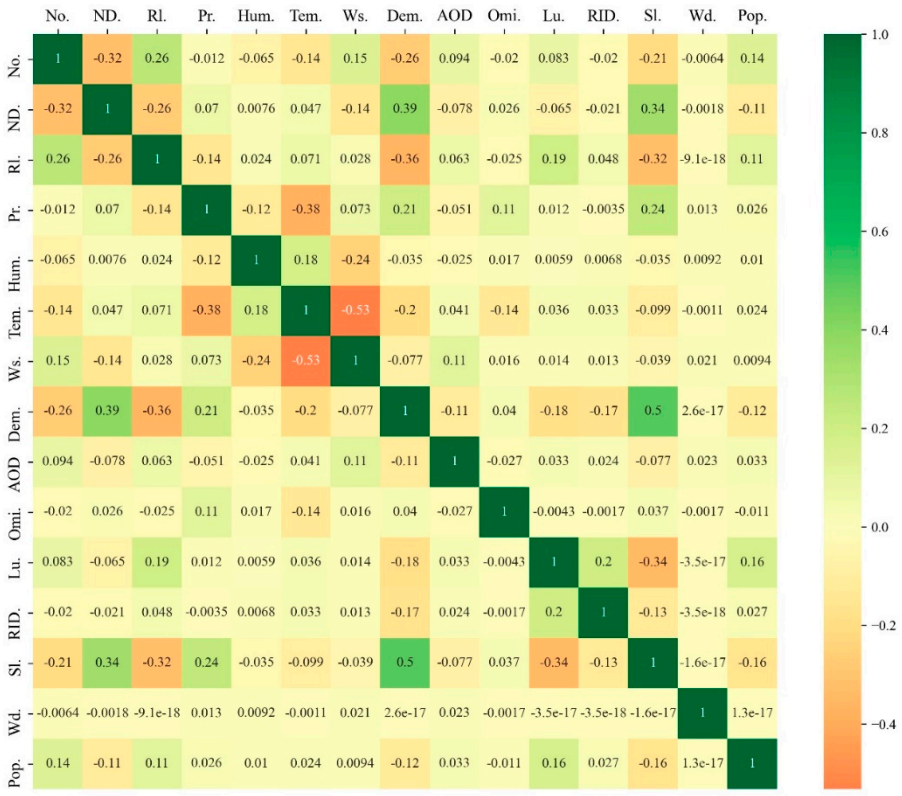

Figure S3. Correlation coefficient between NO2 monitoring data and remote sensing products in SWFJ in 2018.

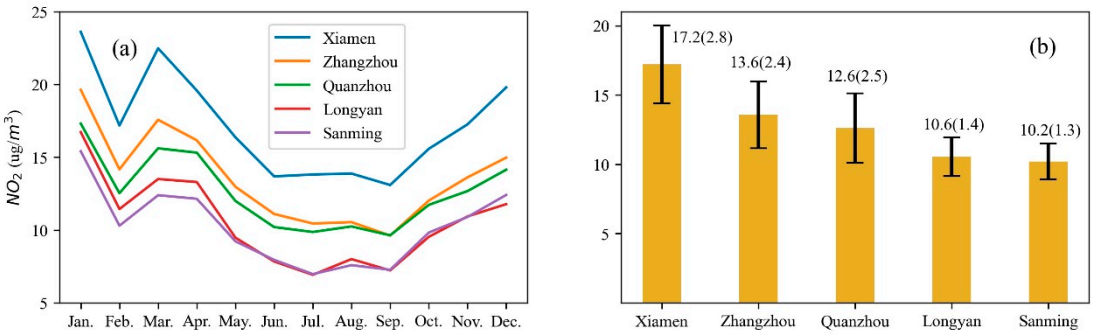

Figure S4. Monthly and annual averages of NO2 in different cities in SWFJ. (A) Monthly average trend chart of NO2 in different cities; (b) annual average and 1/2 standard deviation of different cities. The value and the value in parentheses represent the annual average and 0.5 times the standard deviation, respectively.

The cross−validation (CV) method of interpolation includes two types. The first is to divide the observed data into training and validation data proportionally. This method is not prone to overfitting. However, since the interpolation algorithm is usually linear fitting, the empirical data will directly affect the interpolation effect. This verification method needs to lose part of the data, which affects the fitting of the data. The second case is the error between the cross−validation interpolation results and the observed results. Although this method is prone to local overfitting, more training data can improve the overall interpolation effect [1,2]. Therefore, we choose the second interpolation verification method. The result is in Fig. S5.

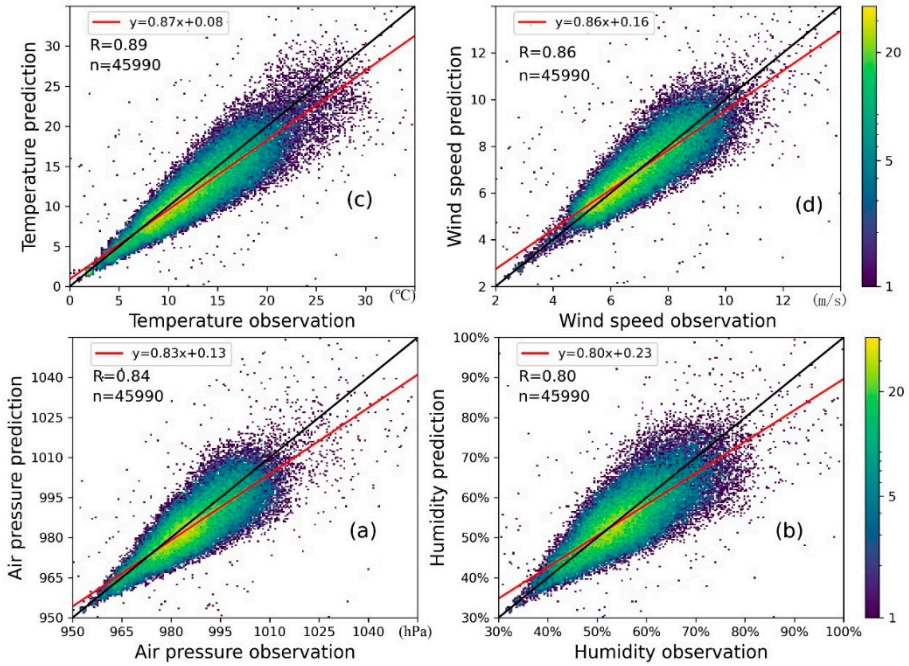

Figure S5. (a), (b), (c), and (d) represent the cross−validation of the spatial interpolation of air pressure, humidity, air temperature, and wind speed, respectively. The horizontal axis represents the observation results, and the vertical axis represents the interpolation results. R represents the correlation coefficient, and n represents the number of samples. The black line represents the 1:1 ratio line, the solid red line represents the first−order linear fitting function curve, and the color bar represents the point density.

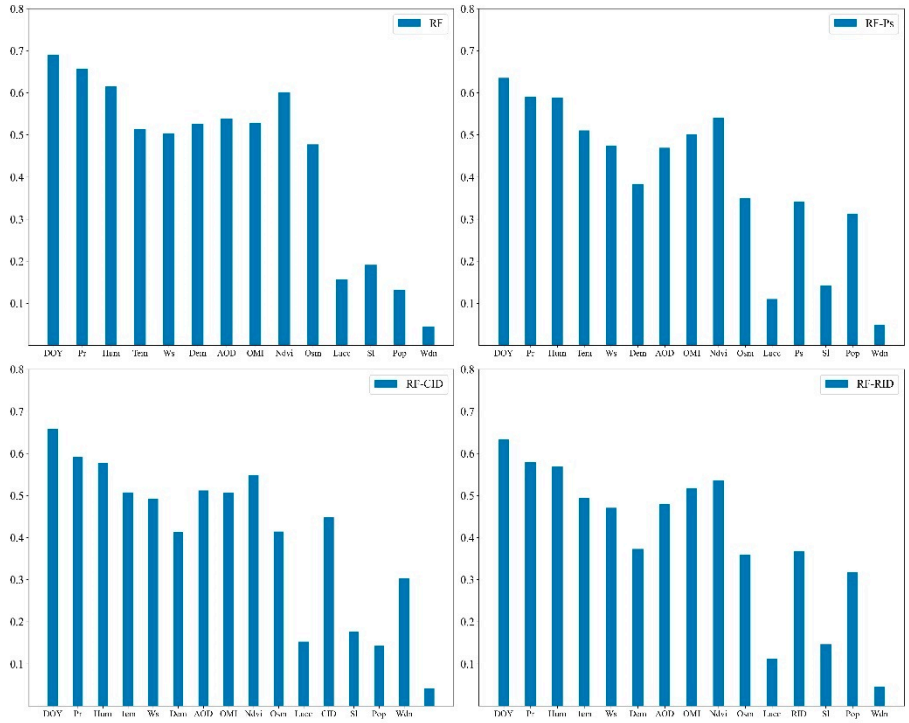

Figure S6. RF, RF−Ps, RF−CID and RF−RID feature importance. The x−axis represents the factors used to build the different models. The y−axis represents importance values.

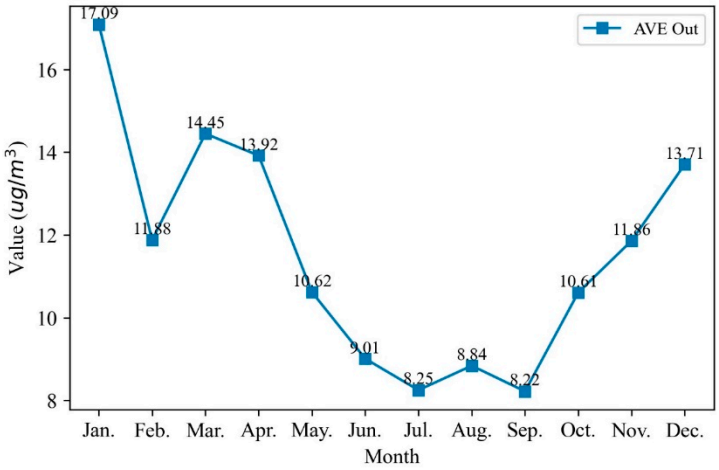

Figure S7. 2018 RF−RID monthly Average. The x−axis represents the different months of 2018. The y−axis represents the mean NO2 concentration.

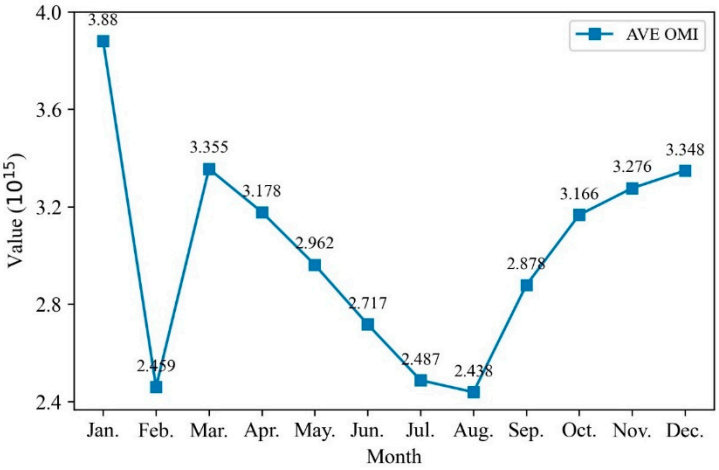

Figure S8. 2018 OMI monthly Average. The x−axis represents the different months of 2018. The y−axis represents the mean OMI value.

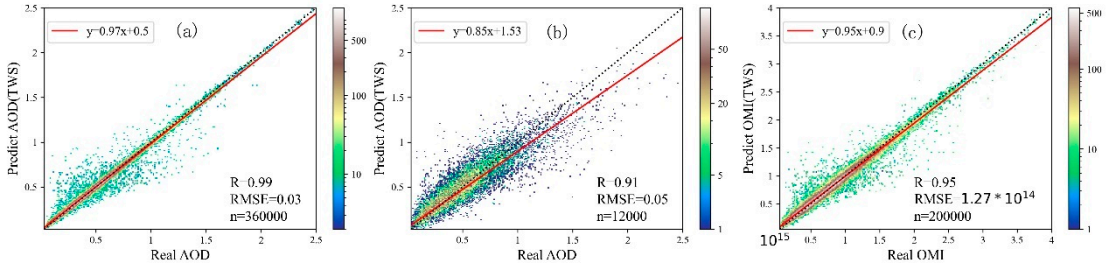

Figure S9. The CV scatterplot of TWS recovery OMI and AOD. (a) represents the cross−validation of the first step of TWS to restore AOD. (b) represents the cross−validation of the AOD recovery in the second step of TWS. (c) represents the cross−validation of TWS second−step recovery OMI.

Supplement discussion

Compared with the other three models, the RF−RID model achieved better CV in the 7−day and weekly forecasts. However, the accuracy of continuous prediction results is lower than that of random prediction. The main reason is that in the process of continuous prediction, the continuous absence of some independent variables has a more significant impact on the machine learning model [3]. Taking DOY as an example, the feature importance of the four models is ranked, and DOY occupies the highest position. However,

in the continuous prediction, 7 (weekly) to 31 (monthly) DOY values will be extracted for model training, and these extracted data cannot participate in the model training process. Some DOY values are missing from the continuous predictions, making the model unable to learn enough features in the missing parts. Therefore, the CV value is reduced when making predictions. In addition, the way we ended up simulating the spatial distribution of NO_2 is closer to random validation, so the reduced accuracy of continuous predictions has less of an impact on the simulation process. Under circumstances RF−RID obtains better CV in random sampling or continuous sample tests, indicating that RF−RID is more robust than other models.

Reference

1. Chen, G.; Guo, J. Spatial interpolation techniques: their applications in regionalizing climate−change series and associated accuracy evaluation in Northeast China. *Geomatics, Natural Hazards and Risk.* **2017**, *8*, 689−705, https://doi.org/10.1080/19475705.2016.1255669.
2. Tustison, B.; Harris, D.; Foufoula−Georgiou, E. Scale issues in verification of precipitation forecasts. *Journal of Geophysical Research: Atmospheres.* **2001**, *106*, 11775−11784, https://doi.org/10.1029/2001JD900066.
3. Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing.* **2018**, *300*, 70−79, https://doi.org/10.1016/j.neucom.2017.11.077.