**Sensitivity analysis**

Long-term ambient O3 tracking covers earlier years beyond the satellite-based remote sensing measurements or chemical reanalysis (i.e. 1990–2002), indicating predictions would merely relied on the CMIP6 numerical simulations for this period. We therefore extended a sensitivity analysis for the first-stage space-time Bayesian neural network-based data assimilation during 2003–2019 under two scenarios, as fusing 8 CMIP6 models with (ScA) and without (ScB) a machine-learning-calibrated remote-sensing measurements and chemical reanalysis outputs,1 assisted with over 40 auxiliary features.2

We then evaluated the accuracies of 10-fold cross-validation tests by random split (70% dataset matched with observations), external validation tests (the rest 30%), and overall fitting, as summarised in **Supplementary Fig. 1**. We compared the developed ambient O3 datasets under the two scenarios by coefficient of variation (CoV): standard deviation divided by the arithmetic mean. We concluded that the deep-learning-based prediction accuracies by solely using CMIP6 simulations were as competitive as fusing additional measurements, and no substantial discrepancies were observed between ScA and ScB (CoV = 1.0%, spatiotemporal 5–95th%ile: 0.1–2.8%).

We furtherly split the full dataset manually for cross-validation tests under ScB, maintaining the temporal coherence: i) 2003–2012 for training and 2013–2019 for testing; ii) 2003–2007 and 2015–2019 for training and 2008–2014 for testing; and iii) 2010–2019 for training and 2003–2009 for testing. All three temporally staged cross-validation tests had revealed good performances (R2 = 0.90, 0.92, 0.92; RMSE =2.86, 2.71, 2.70 ppb, respectively for the three tests)**.** The constrained cross-scenario divergences and stable temporal generalisability verified the credibility of model-based ambient O3 tracking in the earlier years.

**Supplementary Table 1 Evaluations of accuracies of deep-learning-based data assimilation with (ScA) and without (ScB) satellite-based remote-sensing measurements and chemical reanalysis outputs.**

Accuracy evaluations include coefficient of determination (R2) and root-mean-square error (RMSE, ppb) for 10-fold cross-validation tests using 70% observation-matched dataset by random split, external validation tests using 30% dataset, and overall model fitting for the two scenarios respectively. ~~Panel (b) summarises the accruacies for staged temporal extrapolation tests: i) 2003–2012 for training and 2013–2019 for testing (ScB-I); ii) 2003–2007 and 2015–2019 for training and 2008–2014 for testing (ScB-II); and iii) 2010–2019 for training and 2003–2009 for testing (ScB-III). Panel (c) and (d) show the overall deep-learning-based fitting quality under the two scenarios.~~ Given systematic *in situ* observations were unavailable in earlier years of China, and CNEMC sites were allocated in urban and rural environments disproportionally, model fitting and performance evaluations are conducted on global scale.

|  |  |  |
| --- | --- | --- |
| **Evaluation Metrics** | **ScA** | **ScB** |
| Cross-validation R2 | 0.883 | 0.882 |
| Cross-validation RMSE | 3.887 | 3.876 |
| External validation R2 | 0.885 | 0.883 |
| External validation RMSE | 3.879 | 3.868 |
| Overall fitting R2 | 0.969 | 0.968 |
| Overall fitting RMSE | 2.550 | 2.542 |

REFERENCES

1. Liu X, Zhu YJ, Xue L, Desai AR, Wang HK. Cluster-Enhanced Ensemble Learning for Mapping Global Monthly Surface Ozone From 2003 to 2019. *Geophys Res Lett* 2022; **49**(11): e2022GL097947.

2. Sun H, Shin YM, Xia M, et al. Spatial Resolved Surface Ozone with Urban and Rural Differentiation during 1990-2019: A Space-Time Bayesian Neural Network Downscaler. *Environ Sci Technol* 2022; **56**(11): 7337-49.