**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. 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 (ppb) | 3.887 | 3.876 |
| External validation R2 | 0.885 | 0.883 |
| External validation RMSE (ppb) | 3.879 | 3.868 |
| Overall fitting R2 | 0.969 | 0.968 |
| Overall fitting RMSE (ppb) | 2.550 | 2.542 |

**Supplementary Table 2 | Evaluation of spatial and temporal extrapolation accuracy by space-time Bayesian neural network downscaler with urban-rural differentiation.**

Different from classical cross-validation tests by randomly splitting the dataset, spatiotemporal generalisability validation tests manually divide the initial dataset by location or time period. Region-clustered spatial generalisability tests use observations in aggregated regions for algorithm training, and assign observations in other aggregated regions for testing, including four sub-experiments (cross-validation for spatial generalisability, cvs1: training on North America, testing on Europe; cvs2: training on Europe, testing on North America; cvs3: training on North America and Europe, testing on Asia; and cvs4: training on locations outside China, testing on China). Period-staged temporal generalisability tests treat six consecutive years as testing subset based on trainings from the rest 24-year global-scale dataset, including five sub-experiments (cross-validation for temporal generalisability, cvt1: training on 1990–2013, testing on 2014–2019; cvt2: training on 1990–2007 and 2014–2019, testing on 2008–2013; cvt3: training on 1990–2001 and 2008–2019, testing on 2002–2007; cvt4: training on 1990–1995 and 2002–2019, testing on 1996–2001; cvt5: training on 1996–2019, testing on 1990–1995). Evaluation statistics include crude R2 and RMSE (in ppb) without 1:1 linear regression calibration, linear regression slope (*k*) and intercept (*b*).

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Urban | | | |  | Rural | | | |
| Spatial extrapolation | ***R*2** | **RMSE (ppb)** | ***k*** | ***b*** |  | ***R*2** | **RMSE (ppb)** | ***k*** | ***b*** |
| cvs1 | 0.86 | 6.3 | 0.89 | 4.14 |  | 0.82 | 6.7 | 0.93 | 4.43 |
| cvs2 | 0.86 | 6.0 | 0.92 | 4.28 |  | 0.81 | 7.3 | 0.88 | 3.66 |
| cvs3 | 0.84 | 5.1 | 0.85 | 7.15 |  | 0.80 | 7.9 | 0.82 | 5.01 |
| cvs4 | 0.88 | 4.9 | 0.80 | 9.65 |  | 0.81 | 6.6 | 0.87 | 2.84 |
| Temporal extrapolation |  |  |  |  |  |  |  |  |  |
| cvt1 | 0.90 | 5.7 | 0.92 | 1.65 |  | 0.88 | 4.7 | 1.07 | –0.51 |
| cvt2 | 0.88 | 5.0 | 0.93 | 1.89 |  | 0.82 | 5.3 | 1.05 | –0.52 |
| cvt3 | 0.91 | 4.9 | 0.92 | 1.44 |  | 0.80 | 4.6 | 1.02 | –0.53 |
| cvt4 | 0.87 | 5.1 | 0.91 | 1.67 |  | 0.87 | 4.4 | 1.02 | –0.56 |
| cvt5 | 0.85 | 4.7 | 0.91 | 1.38 |  | 0.83 | 4.8 | 1.01 | –0.29 |

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.