Multi-stage Ensemble-learning-based Model Fusion for Surface Ozone Simulations: A Focus on CMIP6 Models

Zhe Sun[1] and Alexander Archibald[2]

[1]Centre for Atmospheric Science
[2]Centre for Atmospheric Sciences, Yusuf Hamied Department of Chemistry, University of Cambridge

November 26, 2022

Abstract

Accurately simulating global surface ozone has long been one of the principal components of chemistry-climate modelling, but divergences in simulation outcomes have been reported as a result of the mechanistic complexity of tropospheric ozone budget. Settling the cross-model discrepancies to achieve higher accuracy thus is a task of priority. Building on the Coupled Model Intercomparison Project Phase 6 (CMIP6), we have transplanted a conventional ensemble learning approach, and also constructed an innovative 2-stage enhanced space-time Bayesian neural network to fuse an ensemble of 57 simulations together with a prescribed ozone dataset, both of which have realised outstanding performances (R-square > 0.95, RMSE < 2.12 ppbV). The conventional ensemble learning approach is computationally cheaper and results in higher overall performance, but at the expense of oceanic ozone being overestimated and the learning process being uninterpretable. The Bayesian approach performs better in spatial generalisation and enables perceivable interpretability, but requires heavier computational burdens. Both of these multi-stage learning-based approaches provide frameworks for improving the fidelity of composition-climate model outputs for use in future impact studies.

Title

Multi-stage Ensemble-learning-based Model Fusion for Surface Ozone Simulations: A Focus on CMIP6 Models

Authors

Zhe Sun[1,2*], Alexander T. Archibald[1,3*]

Affiliations

[1] Centre for Atmospheric Science, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK

[2] Department of Earth Sciences, University of Cambridge, Cambridge CB2 3EQ, UK

[3] National Centre for Atmospheric Science, Cambridge CB2 1EW, UK

[*] Corresponding authors:

Zhe Sun (zs347@cam.ac.uk) and Alexander T. Archibald (ata27@cam.ac.uk)

ABSTRACT

Accurately simulating global surface ozone has long been one of the principal components of chemistry-climate modelling, but divergences in simulation outcomes have been reported as a result of the mechanistic complexity of tropospheric ozone budget. Settling the cross-model discrepancies to achieve higher accuracy thus is a task of priority. Building on the Coupled Model Intercomparison Project Phase 6 (CMIP6), we have transplanted a conventional ensemble learning approach, and also constructed an innovative 2-stage enhanced space-time Bayesian neural network to fuse an ensemble of 57 simulations together with a prescribed ozone dataset, both of which have realised outstanding performances ($R^2 > 0.95$, RMSE < 2.12 ppbv). The conventional ensemble learning approach is computationally cheaper and results in higher overall performance, but at the expense of oceanic ozone being overestimated and the learning process being uninterpretable. The Bayesian approach performs better in spatial generalisation and enables perceivable interpretability, but requires heavier computational burdens. Both of these multi-stage learning-based approaches provide frameworks for improving the fidelity of composition-climate model outputs for use in future impact studies.

Keywords

CMIP6; CCM; surface ozone; model ensemble; space-time Bayesian neural network; data fusion

1 INTRODUCTION

Tropospheric ozone (O_3) is a trace-gas, near-term climate forcer with global mean lifetime ~23 days, and also a major air pollutant being of detrimental defects on human and ecosystem health.[1-3] Besides warming the atmosphere as a greenhouse gas, ground-level O_3 also reduces crop yields.[4-6] Laboratory experiments have confirmed O_3 exposure to cause oxidative stress, inflammatory responses and immunologic diseases.[7] Epidemiological studies report that short-term exposures to high-level ozone are significantly associated with the exacerbation of asthma[8] and have increased hospitalisations among children,[9] while long-term ozone exposure is linked to respiratory diseases like chronic obstructive pulmonary disease, cardiovascular diseases, and even premature deaths.[10-14] Global Burden of Disease (GBD) reported over 0.36 million premature deaths globally in 2019 from exposure to ambient O_3;[15] and high O_3 exposure could exacerbate the $PM_{2.5}$-mortality risk associations.[16] These results underscore the pressing need for research linking population exposure assessment to surface O_3 and its impacts on human health.

Satellite-based observations cannot provide accurate measurements for O_3 at the surface since surface O_3 will be obscured by the climbing O_3 abundance in high-layer atmosphere thus cannot be measured directly from remote-sensing; while the ground-level station-based observation sites are still rather limited in spatial coverage.[17, 18] The demands for full-coverage surface O_3 concentrations have promoted the application of model simulations, which have been being improved as our understanding of the mechanisms behind tropospheric O_3 has improved.[19-21] But model simulations are not perfect, due to imperfections of O_3 chemistry mechanisms built in the models, biases and errors in the underlying emissions, and uncertainties caused by the discretisation and numerical treatment of a non-linear complex system. Archibald et al. have shown that for future evolution projections of the tropospheric column O_3, model differences are a leading order term of uncertainty over decadal scales.[21] There are various types of models used to simulate surface O_3. Chemical transport models (CTM) perform satisfactorily especially in regional-level simulations;[22-25] and are considered to be free of biases in meteorology due to the use of prescribed meteorology. But these models lack important feedbacks from atmospheric composition on to the model meteorology and climate, hence atmospheric composition-climate models (CCM) have been developed; and when coupled with land, sea, and sea-ice modules into earth system models (ESM), it is feasible to simulate multi-decadal or even centennial scale changes in atmosphere.[26-29]

To evaluate and compare the coupled models, a number of research institutes have contributed to the Coupled Model Inter-comparison Project Phase 6 (CMIP6) with a range of experiments conducted by a series of state-of-the-art coupled CCMs and ESMs. The same inputs are used, including emission inventories and land properties.[30-33] CMIP6 has endorsed a total of 23 MIPs to answer a wide range of scientific questions in atmospheric chemistry and climate, among which the Aerosols and Chemistry Model Intercomparison Project (AerChemMIP) involves a collection of simulations targeted at reactive gases and aerosols including tropospheric O_3.[34] Large discrepancies have been detected across models; beyond figuring out the mechanistic causes for these differences,[31, 35] an urgent challenge is how to calibrate and make the maximum use of the simulation ensemble.

Applying frontier machine learning algorithms to assimilate the outputs from multi-source modelling activities like MIPs and observation databases, known as data “assimilation” or data “fusion”, is an important part of environmental research in the big data era. Studies which enhance the prediction accuracy of ambient air pollution concentrations by ensemble learning have emerged in recent years.[36-39] However, these studies only used no more than one model simulation integrated with predictor variables contributing to the budget of O_3, without involving fusing multiple simulation ensembles like CMIP6. In addition, the conventional machine- or deep-learning approaches aim purely at brute-force fitting into high accuracy while sacrificing the interpretability of the training processes, so have long been criticised as “black-box” and contradict the nature of mechanism-driven sciences like atmospheric modelling.[40-42] Under these circumstances, reaching a performance-interpretability balance for multi-source data fusion following credible observations will be of high value in atmospheric research.

Our current study is an innovative exploration on this issue, emphasising on developing innovative ensemble-learning frameworks to assimilate the multiple CMIP6 model simulation ensembles and TOAR observations to obtain one single surface O_3 dataset capturing the spatiotemporal variabilities as accurate as possible. Fusing a collection of simulation ensembles rather than just using the output from one simulation can give more prominence to the mechanism-driven models so as to avoid brute-force overfitting resulting from external predictor variables, especially when any given model simulation could be largely biased. The primary innovation of this study is in transplanting the conventional ensemble-learning data-assimilation methodology onto multi-source data fusion, and optimising an enhanced 2-stage space-time Bayesian neural network to assimilate the CMIP6 simulation ensemble. The advantages of the conventional approach include a much lower computation burden and higher accuracy in observation-covered regions, while the merits of the innovative Bayesian approach lie in its better spatial generalisability and intuitive perception of spatiotemporal model weighting. In either case, the multi-model fused surface O_3 concentration can fill in the observational gaps and enable further relevant researches in the long term. As an example we show here using Fourier-series function to fit the temporal surface O_3 variability provides a feasible way to effectively summarise periodic air pollutant concentrations. Detailed evaluations and comparisons on CMIP6 model ensemble, and deeper discussions on model revision insights from deep learning-based calibration processes are beyond the scope of this study.

2 METHODOLOGY AND DATA SOURCES

2.1 CMIP6 simulation ensemble

We collect 14 coupled earth system models having finished the "historical" simulations (1850-2014) of tropospheric O_3 as listed in Table S1, of which 8 models use interactive chemistry schemes. A prescribed O_3 concentration dataset is used for all 4 non-interactive chemistry models (AWI-ESM,[43] BCC-CSM2,[44-46] IPSL-CM6A,[47, 48] and MPI-M-ESM1.2[49-52]) and 2 CNRM models are not considered for fusion due to the simplified treatment of O_3 chemistry.[53-57] A total of 8 models, including BCC-ESM1,[58, 59] MPI-ESM1.2-HAM,[60] MRI-ESM2.0,[61-63] NASA-GISS-E2.1,[64-66] NCAR-CESM2-WACCM6,[67, 68] NCC-NorESM,[69] NOAA-GFDL-ESM4,[70, 71] and UKESM1-0-LL, [19, 28, 72-75] consisting of 57 individual simulation experiments (i.e. realisations in terms of CCM simulation labelled as $r_n i_n p_n f_n$) and 1 prescribed input dataset (from Inputs4MIPs)[76] are recruited for data fusion. The multiple ensemble members under one model allow for capturing the uncertainties in the chaotic coupled chemistry-climate system; and because of the free-running nature of the simulations, each of the 57 individual simulations is treated separately with no cross-ensemble averaging clustering into each model involved. All simulation outputs are averaged to monthly time frequency for assimilation with observations. Detailed information of the participant research institutes, design of atmosphere module settings, and experiment labelling rules are illustrated in the Supporting Information.

2.2 Observations

The tropospheric ozone assessment report (TOAR) programme has archived high-quality ground-level O_3 measurements over the period 1990-2014,[17] which are used as “standard” for physical and statistical model evaluation; our study period is thus selected as 1990-2014. To support analyses at the planar spatial resolution of the CCMs involved in this study, TOAR sites are aggregated into 2°×2° latitude-longitude grid as plotted in Figure S1, including 585 spatial grids with a total of 5,322 different observational sites; and averaged to monthly temporal interval for the robustness of model-observation evaluation. Such spatiotemporal aggregations can also strengthen the stability of grid-level observation-simulation evaluation, and to some extent abate the statistical compromises by excluding the observation missing records for some certain sites in the early years of the dataset (ca. 1990s). Only spatial grids in which there is at least one observation site are used. Throughout the study, the gridded TOAR observations are used as supervised learning labels.

2.3 Additional auxiliary predictors

Higher prediction accuracy can be achieved when integrating additional features into statistical models.[36-38] Comprehensively considering the O_3 budget mechanisms, experiences from previous relevant studies, and statistical correlations with surface O_3, we screen out 13 variables as assistant predictors as: CMIP6 simulated concentrations of surface $PM_{2.5}$, NO_2, higher layers of O_3 (vertical O_3 column), and ambient air temperature obtained from the World Climate Research

Programme (WCRP) Earth System Grid Federation (ESGF) CMIP6 database (https://esgf-node.llnl.gov/search/cmip6); emissions of biogenic VOCs, NO*x*, CO, black carbon (BC) and organic carbon (OC) together with urbanised land proportions, collected from input datasets for Model Intercomparison Projects (https://esgf-node.llnl.gov/search/input4mips); surface elevation downloaded from the Global Multi-resolution Terrain Elevation Data (GMTED);[77] and gridded urban and rural populations linearly interpolated with corrections towards the actual annual world total populations into year-precision from United Nation's World Population Prospects (UN WPP) Adjusted Population Density and Gridded Population of the World (GPW) operated by NASA Socioeconomic Data and Applications Centre (SEDAC).[78]

2.4 Multi-model Fusion Frameworks

We use "physical model" to refer to the CMIP6 mechanism-driven atmospheric models, and "statistical model" for the data-oriented machine- or deep-learning frameworks to avoid confusion in terminology. No transformations are made for either the observations or model simulations as they follow the Gaussian distribution well with slight temporal imbalance. Following literatures,[36-38] an adjusted ensemble learning-based multi-model fusion framework is constructed as presented in the upper panel of Figure 1. In this approach, raw simulations (i.e. 57 CMIP6 historical simulations and 1 prescribed O_3 dataset, noted as "57+1 ensemble" hereafter) together with the normalised additional predictor variables are first re-gridded onto the 2°×2° TOAR observation grids, following procedures graphically presented in Figure S2. Then, all the model simulation ensembles, external predictors, and 6 space-time indices (i.e. 3 Euclidean spherical coordinates in analytic geometry, and 3 helix-shape trigonometricised month sequence t as [$\cos(2\pi t T^{-1})$, $\sin(2\pi t T^{-1})$, t] where T is prescribed as 1 year)[79] are mixed together as inputs for random forest, gradient boost decision tree, and convolutional neural network regression models separately; and outputs from the 3 algorithms are finally blended by L2-regularisation-based weighting (ridge regression). This approach is entitled as "aggressive" approach because this methodology respects the observations (i.e. labels for supervision) more than the physical models, hence during the process of training, the concentrations in each grid are treated individually so as to compromise the spatiotemporal continuous structure of the original physical model simulations, leading to inexplicability. The aggressive approach involves at least two stages of ensemble: the first CMIP6 multi-model ensemble and second multi-algorithm ensemble, where the random forest regressor essentially is another layer of ensemble learning. The random forest regressor is a large collection of separate decision trees with individual of which generating a single prediction and the final prediction given by averaging all trees, thus the random forest is perceived as an ensemble learning method.[80]

Contrarily, in order to maintain the interpretability of the deep learning processes, we also adopt an enhanced 2-stage space-time Bayesian neural network (BNN) framework as illustrated in the lower panel of Figure 1. Space-time indices and additional predictors are put into a 10-layer 1024-node BNN to generate spatiotemporal variant re-scaling factors (k), bias correctors (b) and the randomised noises (σ), under the supervision of TOAR observations to pre-calibrate the raw re-gridded CMIP6 simulations. Then, spatiotemporal variant model weights (α) are estimated by 5-layer 256-node BNN merely from the 6 space-time indices, to finally reach the weighted average ensemble surface O_3 concentration predictions. This approach is named as the "conservative" approach as throughout the process of prediction enhancement, all parameters are clamped by space-time indices with presumed distributions, thus this framework respects the raw simulations more and might be highly biased on extreme observations. All involved parameters can be thoroughly separated from the framework and presented intuitively by mapping, so that the whole process of assimilation is traceable and interpretable. We construct the two-stage BNN instead of single-stage because the divergences still exist among the calibrated CMIP6 models in the first-stage and hence further mixing is required. Directly using the second-stage BNN will lose the chance to observe the calibration features for individual physical models; and different degrees of initial biases will cast higher weights onto the smaller biased models, possibly leading to undesirable feature monopolisation.

Statistical principles of naïve space-time BNN (i.e. single-stage space-time BNN) are illustrated in details by a recent report.[79] Mathematically speaking, solutions of the spatiotemporal parameters (i.e. k, b, and α) are not unique, but it is reasonable to assume the observation covered and uncovered regions are of homogeneity in distribution of these parameters, which requires a Bayesian method to replace the single value of parameters with a distribution. The 6 space-time indices can assist in capturing the spatiotemporal autocorrelation of the surface O_3. 10,000 times of Monte Carlo simulation ensembles are applied to approximate the distribution, so as to guarantee the robustness of BNN estimation, thence the conservative approach involves 3-stage ensemble: first in multi-model ensemble and the latter two in the 2-stage Bayesian parameter generation. For the final predictions based on the optimised distribution parameters trained through the BNN, 69.2% fall into 1 standard deviation (σ) range, 96.2% into 2σ and 99.9% into 3σ, conforming to the regularity of Gaussian distribution and thus justifying our Bayesian model presumption.

To evaluate the performance of 2 approaches, 10-fold cross-validation (CV) assessment is applied, and 7:3 training-test split is used through the full dataset during 1990-2014. An additional temporal extrapolation test is conducted by manually setting the 1990-2009 TOAR observations with grid-corresponding physical model simulations as training set and 2010-2014 as test set. Three manual cross-validation tests are conducted by splitting the whole dataset into training-testing sets with regional integrity as i) Europe-training for North-America-testing; ii) North-America-training for Europe-testing; and iii) Europe-North-America-training for East-Asia-testing, so as to evaluate the spatial extrapolation capability of the 2 statistical models. Decomposition of model-observation errors follow a previous research.[81] The neural network trainings are accomplished by Adam stochastic optimisation algorithm, setting the initial anchor values from observations and the learning rate as 10^{-4} after centric normalisation.

The complex machine learning frameworks are constructed instead of using simple statistical models owing to their limitations in handling the i) similarities across multiple physical models (i.e. collinearity in statistical term); ii) interaction effects between the input variables; iii) spatiotemporal auto-correlations and discrepancies in calibration parameters; and iv) propensity of overfitting when introducing high-order polynomial terms. Additionally, this cross-disciplinary study closely follows the trends of applying the cutting-edge data sciences onto environmental studies, hence only machine- and deep-learning approaches are transplanted, enhanced and discussed here.

2.5 Other relevant statistics

Fourier-series sinusoid functions theoretically can fit any periodical variables,[82] so are used to capture the location-specific seasonal periodic variations of surface O_3 in this study to parametrically interpret the final assimilated surface O_3 concentrations by revealing the intra- and inter-year variability quantitatively with perceivable mapping. Akaike Information Criteria (AIC) is used for statistical model selection, taking the realistic explicability altogether into consideration as listed in Table S2. Given TOAR observations and model outputs are monthly averaged, the final Fourier function is chosen as

$$f(t) = a_0 e^{a_1 t} + (b_0 + b_1 t)\, sin\left(\frac{\pi}{6} t + \varphi_1\right) + c_0\, sin\left(\frac{2\pi}{6} t + \varphi_2\right),$$

where t represents the month-sequence; a_0 as starting-point surface O_3 concentration (January 1990); $12a_1$ as annual average change rates; $2b_0$ as the baseline and $24b_1$ as annual change of seasonal variation amplitude (i.e. peak-valley difference); and c_0 as the fine-tuning parameter which can modify the sinusoidal shape, but usually the absolute values are rather small, thus not considered for interpretation. An exponential term for the annual average surface O_3 is applied instead of linear term as the long-term simulations have reported exponential increasing trend of the tropospheric O_3 over centennial scales,[31] regardless of the fact that the AIC values vote for the linear model.

3 RESULTS

3.1 Raw simulation evaluations

Raw CMIP6 surface O_3 simulations generally perform fairly well across all TOAR covered areas in terms of synchronicity (Figure 2), as the correlations between observations and the 57+1 ensemble averages are 0.74 ± 0.18 (inter-quartile range, IQR: [0.67, 0.87], Range: [-0.58, 0.96]). Overestimations are observed at 4.1 ± 2.0 (IQR: [5.1, 13.1], Range: [-22.2, 31.1]) ppbv across all TOAR covered spatial grids, hence the normalised mean biases (NMB) are high at 9.7 ± 6.3 (IQR: [4.2, 13.5], Range: [-28.1, 48.9]) %. Some regions like west Australia coastline even report negative correlations (Pearson's ρ = -0.58).

The synchronicity and bias for realisation-ensembled model outputs are also evaluated in Figure S3 and Figure S4. NASA-GISS-E2.1 reports negative synchronicity in the USA-Canada border, while NCC-NorESM fails to reproduce the temporal variabilities in most of the studied sites. UKESM1-0-LL predicts closely to the measurements, but underestimates the surface O_3 around the USA-Canada border; while all the rest models present overestimations. Divergences are found between the individual models (Figure S5), and the high simulation discrepancies are mainly aggregated in the intertropical convergence zone (ITCZ) and eastern China, where the standard deviations exceeded 20% of the ensemble means. The barely satisfactory synchronicities and high overestimation biases indicate that the raw surface O_3 simulation might not be suitable for direct application in health impact studies, verifying the necessity of calibrations, at least statistically.

3.2 Performance of multi-model ensemble fusion

Both aggressive and conservative multi-model fusion perform well in prediction enhancement (Figure 2). The model-observation correlations are high at 0.98 ± 0.01 (IQR: [0.97, 0.99]) and 0.95 ± 0.08 (IQR: [0.95, 0.98]) for the aggressive and conservative approach, respectively; and NMBs of the aggressive model are 0.29 ± 3.06 (IQR: [-1.22, 1.54]) %, marginally smaller than the conservative model at 0.40 ± 3.57 (IQR: [-1.72, 1.93]) %. The general overestimation issues of the raw CMIP6 simulations have been handled well, but there are still some sporadic high NMBs detected in Asia, Africa, and South America, where the ground-based monitoring sites are rare and spatially scarce.

The full-range fitting R^2 (Table 1) of the aggressive and conservative approaches are 0.96 and

0.95, respectively, both indicating plausibility of the multi-model fusion with calibration; while the conservative predictions follow more loosely to the observations, especially in the low-concentration ranges (Figure S6), resulting in relatively higher root mean squared error (RMSE) at 2.12 ppbv compared with 1.81 ppbv for the aggressive approach. However, the conservative approach performs better in 1:1 model-observation calibration criteria according to the closer-to-one slope factor ($k_c^{-1} < k_a$, $0.97^{-1} < 1.05$) and closer-to-zero systematic bias ($|b_c| < |b_a|$, $|0.71| < |-1.35|$). This is because directly involving additional features (i.e. the aggressive approach) can possibly introduce noise into the calibration, as their association with surface O_3 are not simply linear, especially in higher concentration ranges, so that the 1:1 model-calibration line is deviated.

Both approaches calibrate the physical models effectively, with the conventional aggressive approach performing slightly better than the innovatively established conservative model, which however, is already good. The spatiotemporal stability of the two approaches are also assessed in Table 1, concluding that the aggressive approach performs better in the later years of the dataset, while the conservative approach performs consistently well across the 25-year period. This is because the aggressive approach depends so largely on the observations that defects of observation coverage in early years will compromise the learning effects. However, the aggressive approach performs well across different continents ($R^2 > 0.90$), but the conservative approach performs slightly worse in the southern hemisphere ($R^2 > 0.83$), as a result of insufficient observations. This data sparsity results in the inter model-spread in the raw simulations being, to some extent, retained, as this could not be addressed by the BNN-based weighted linear combination; instead, additional features in the prediction-oriented aggressive approach brute-forcedly correct the large observation-simulation gaps. Both approaches perform well across seasons.

3.3 Extrapolation generalisability

Due to the limitations of lacking systematic observations in China, India, Africa and oceanic regions during 1990-2014, there are no means to verify the simulations in these areas directly; but this problem can be explored indirectly by checking the extrapolation potential on the observation-uncovered locations. Three regional cross-validation tests are graphically summarised in Figure S7, all of which reveal better generalisation capability of the conservative approach than aggressive. Neither underfitting nor overfitting issues are detected on the conservative approaches (i.e. CV and test scores are quite close); while underfitting is apparent for the aggressive approach in these regions, mainly reflected by failures in capturing extreme O_3 concentrations. The temporal extrapolation tests of two statistical models reveal high generalisability on the most recent 5-year test sets during 2010-2014 as $R^2 = 0.91$ (CV-$R^2 = 0.88$, test-$R^2 = 0.82$) for the aggressive approach and $R^2 = 0.92$ (CV-$R^2 = 0.89$, test-$R^2 = 0.85$) for the conservative approach. The temporal extrapolation performances are better than spatial generalisation, because the temporal periodic variations of surface O_3 are of a more stable pattern than regional divergences. In a nutshell, the conservative BNN approach wins over towards spatial and temporal generalisability, and we thus regard the conservative BNN results as "standard" for further interpretation.

3.4 Differences between ensemble approaches

Comparisons between the "standard" and aggressive approach outcomes are graphically summarised in Figure S8, revealing most of the global regions are of high congruity ($\rho = 0.85 \pm 0.17$, IQR: [0.81, 0.96]), while the divergences mostly occur on the ITCZ and Arabian-African areas ($\rho < 0.02$). Small relative biases have also justified the similarity between the aggressive and conservative approaches, as the NMBs (defined as aggressive – conservative) are 1.38 ± 4.61

(IQR: [-1.59, 3.77]) %. The positive differences mainly aggregate in Africa, Antarctica, Oceania and most of the oceanic basins, while the negative differences cluster in Asia, Europe and America.

The simplest fusion, the arithmetic average, of CMIP6 simulation ensemble would be used as a compromise were there no ground-based observations as used by precedent studies,[31] which factually could lead to high biases if the real surface O_3 exposure assessment is the main research interest. This study aims to develop innovative approaches to fuse both model simulations and observations, and by comparing with the simplest fusion, advantages of new methods can be highlighted. The conservatively ensembled surface O_3 concentrations are of higher synchronicity (ρ = 0.97 ± 0.06, IQR: [0.97, 0.99]) with the simple ensemble average than the aggressive approach (ρ = 0.87 ± 0.14, IQR: [0.83, 0.96]), as the BNN is essentially an enhanced linear combination of multiple model simulations without substantial changes to the spatiotemporal auto-correlation. The ensemble average exceeds the aggressive fusion by 5.9 ± 9.7 (IQR: [-7.9, 14.3]) %, and the overestimations cluster regularly on land surface, especially the high-population-density regions; but surpass the conservative fusion by 9.6 ± 10.5 (IQR: [0.81, 20.2]) %, with the overestimations mainly detected in the wide-coverage northern-hemisphere without apparent land-ocean distinguishment. In conclusion, the simple ensemble average can lead to overestimations, especially in the northern hemispheric land surface; and the differences also reveal that the aggressive fusion model has modified the spatial auto-correlation of the raw CMIP6 simulation to a larger extent than the conservative approach.

3.5 Bayesian spatiotemporal weights

The differences between the two approaches can also be partially attributed to the different weighting schemes of the raw individual simulations. The 57+1 ensembles occupy 93.9% weights in the aggressive approach while the additional assistant variables only contribute 6.1%. Generally, for the aggressive approach, 4 among the 58 simulations contribute dominantly by over 10%, as UKESM1-0-LL-$r_3i_1p_1f_2$ (18.6%), the prescribed O_3 (17.4%), NASA-GISS-E2.1-G-$r_1i_1p_3f_1$ (14.7%) and NCAR-CESM2-WACCM6- $r_1i_1p_1f_1$ (14.1%), while 36 ensemble members contribute less than 0.1%, as graphical presented in Figure S9. On the contrary, the conservative approach results in relatively more even weights, where the prescribed O_3 (2.1%), UKESM1-0-LL (1.9%) and NASA-GISS-E2.1 (1.8%).

Besides the physical model weights, the space-time BNN also generates spatiotemporal variant weights, which can reflect the regions of skill for each individual physical model as presented in Figure S10: UKESM1-0-LL and NCAR-CESM2-WACCM6 are weighted higher in northern hemisphere over land, while the prescribed O_3 dataset, NASA-GISS-E2.1, and NOAA-GFDL-ESM4 contribute more in southern hemisphere over land. The temporal variations of the spatial weights are generally small and of regular regional clustering trends, indicating that the physical models have captured the seasonal variability well.

BNN-based multi-model fusion treats the assistant variables independently with the CMIP6 model simulations, so that the weights of these additional features are not at the same level as the physical models like in the aggressive approach. Direct comparisons of the weights of the assistant variables between the two approaches reveal quite similar patterns of using these additional features for model calibration as shown in Figure S11 which indicates that urban-rural populations, ambient air temperature and elevation are important factors. We suggest further work pay more attention to the role of model surface temperature, which is not fixed in these free-running simulations. High contributions of the space-time indices also indicate that more additional features need to be included for further consideration.

3.6 Long-term surface ozone variations

Spatiotemporal variabilities of the BNN-fused surface O_3 are summarised parametrically using Fourier-series functions (Figure 3). The fitting quality R^2 has reached 0.81 ± 0.12 (IQR: [0.77, 0.87]), where the poor performances ($R^2 < 0.50$) concentrate in ITCZ and the coastlines. The global annual average increasing rate of the surface O_3 is estimated to be 0.23 (95% CI: [0.21, 0.25]) % yr^{-1}, and the highest increasing rates are detected in south Asia, South America, and continental Europe. Decreasing trends are also discovered in eastern China and eastern US. The average intra-year seasonal variation is 13.9 (IQR: [2.1, 49.5]) ppbv, and the highest amplitude differences cluster in eastern US, Africa, Europe, and eastern China. The annual changes of seasonal variations also demonstrate regional variabilities: widening in eastern China by maximum as 1.8 ppbv per year while narrowing in western countries by extreme to -0.8 ppbv per year. The intra-year peak and valley concentrations are generally ascending, as the peaks increase by 8.8 ± 1.1 (IQR: [-6.8, 16.1]) ppbv per year, and the valleys ascend by 0.6 ± 0.8 (IQR: [-7.0, 8.3]) ppbv per year.

4. DISCUSSION

4.1 Multi-model fusion improvement potentials

Decomposition of model-observation errors (Figure S12) can assist in evaluating the optimisation potentials for both the physical and statistical models.[83] The overall RMSE for the aggressive approach is 1.81 ppbv, among which the irreducible root-noise is 1.42 ± 0.47 ppbv, occupying 66.1 ± 16.7 % of the total errors; while the averaged error of the conservative approach is 2.58 ppbv, where the root-noise is 1.87 ± 0.70 ppbv, accounting for 62.2 ± 25.4 %. The noises together with the biases by conservative approach are generally higher than the aggressive approach, while their proportions are close except for the African regions, as listed in Table S3. Most of the unsolvable noises take over more than half of the errors, indicating that both fusion approaches have well approached the realistic observations.

The variances, also known as cross-model divergences, are comparable or even greater than biases for the aggressive approach, while for conservative approach the variances are several folds lower than biases, accounting for less than 10% except for South America (17%). This indicates the conservative fusion model is more robust. The model variances can be statistically perceived as discrepancies of model construction by random draws of the training subset, so that higher model variances represent severe dependences on training inputs, revealing higher sensitivity and lower generalisability.

The current crux of the conservative fusion model falls on the biases, suggesting higher optimisation potentials than the aggressive approach. The biases originate from the inherent systematic biases in physical models, and also the insufficient inclusion of assistant features to enhance the prediction statistically. Comparatively, due to the relatively higher statistical model variances, the aggressive approach shall no longer be the prevalent stream for multi-model fusion, as changes in observation coverage (i.e. labels for supervision in machine learning) will affect the stability of the statistical model substantially.

4.2 Differences in spatial extrapolation

The better spatial generalisation ability of the conservative space-time BNN multi-model fusion is an advantage over the aggressive approach. Paradoxically, the aggressive approach actually performs well on capturing the extreme values. This shall be attributed to overfitting on the assistant features added directly into the fusion processes, so that the predictions are excessively reliant on these external variables. However, due to the complexity of the mechanisms controlling O_3, the statistical associations between physical models, auxiliary predictors, and the realistic concentrations recognised by the aggressive approach will be superfluous and of localised boundedness so that might be drastically different across regions. Excluding these features from aggressive multi-model fusion alleviates the poor performance in spatial extrapolation, as for each regional cross-validation test, R^2 rise to 0.81, 0.83, 0.74, and RMSE decline to 3.64, 3.97, 5.95 ppbv for North America, Europe and East Asia, respectively. To put it briefly, the external assistant features can increase the fitting quality in statistical training, but also serve as the limiting factors for model generalisation. This presents an issue towards understanding the processes of aggressive multi-model fusion, as conservative predictions manifested as underfitting by aggressive approach should be ascribed to the overfitting in the additional feature-assisted aggressive pathway. It suggests that conventional ensemble deep-learning approaches respecting the observations as supervision and linking the input variables only statistically rather than respecting the physical and chemistry mechanisms are of rather limited use, hence it is the second reason that the novel conservative multi-model fusion approach by space-time BNN is preferred.

4.3 Cross-approach divergences

Most discrepancies between the two fusion approaches and the simple ensemble average are located in tropics (Figure S8), which is primarily attributable to the lack of observations as training data, and the variations in raw simulations (Figure S5) resulting from the difficulty in capturing O_3 in this region as a result of complexity in the precursor emissions like biogenic VOCs, soil NO_x, lightning NO_x, etc.[31] We highlight in particular the need for long-term continuous ground-based measurements of O_3 in the tropics as a research priority.

The differences between the simple ensemble average and the aggressive fusion approach (Figure S8) indicate that the aggressive approach only addressed the systematic overestimations on the land surface; the additional variables lead to a land-ocean contrast (e.g. the population, ambient air temperature, O_3 precursor emissions), which are used as key nodes in the tree-structure regressions, so that the calibrations are only effective over the land rather than the whole global surface. The conservative approach respects the raw simulations more by calibrating uniformly for both lands and oceans, so that the average-conservative differences are more spatially uniform (Figure S8).

4.4 Systematic overestimation

Direct use of the raw CMIP6 surface O_3 simulation ensemble mean, as commonly used in the literature[2, 31, 35], causes positive biases around 5-10%, equal to 3.6 ± 4.4 ppbv, with some regions like India high-biased by +40% (+22.7 ppbv), consistent with recent multi-model ensemble studies in this region.[84] Such large biases have important implications on the use of raw ensemble mean data for work related to public health and pollution control policy studies in these regions, reiterating the necessity of observation-supervised calibration. The systematic overestimations across CMIP6 simulations speculate the major cause as the inadequate vertical stratification in atmospheric module. Essentially speaking, the lowest layers of CMIP6 model simulations are used to approximate the surface O_3, but the layer actually refers to a vertical average. Tropospheric O_3 concentration rises with the altitude,[31] thus resulting in overestimation. UKESM1-0-LL stratifies

85 vertical layers,[19] which is the most among 8 interactive chemistry CMIP6 models (Table S1), and lowest overestimations are found, with even underestimations observed in quite a few regions (Figure S4). Further experiments by adjusting the vertical stratifications to observe the changes in surface O_3 simulation performances are suggested to rigorously check this speculation.

4.5 Rationality of enhanced space-time BNN

Our enhanced space-time BNN is optimised from the traditional naïve space-time BNN, without additional feature involvement.[79] The enhancement in part comes from overcoming the inconsistence between the overall and location-specific observation-simulation linear relationships: each simulation cell at different time requires a unique set of *k*-*b* parameters for calibration as $y_{l,t}^{obs} = k_{l,t} \cdot y_{l,t}^{mod} + b_{l,t} + \varepsilon_{l,t}$, where the subscripts *l* and *t* represent location and time indices, so that using a fixed slope *k* and intercept *b* to calibrate all simulation cells is of limited use. However, the calculated sets of parameters are spatially limited to the observations, thus a naïve space-time BNN framework is required for spatial extrapolation onto the full global space.

The BNN generates the space-time variant calibration slopes and intercepts for each CMIP6 model in the pilot attempts, with which the assistant features are significantly correlated Figure S13, indicating these additional factors can contribute to the calibration parameters. For the purpose of increasing the prediction accuracy, the enhanced 2-stage Bayesian neural network regression-based multi-model fusion framework is constructed by firstly incorporating the assistant features into the multi-layer perception structure to generate the calibrated individual simulations, and secondly fusing them up by another naïve space-time BNN without involving any external features.

4.6 Sensitivity Analysis

Considering the cross-realisation variations (0.5 ± 0.1 ppbv) are much lower than the cross-model deviations (4.6 ± 1.7 ppbv, Figure S5), we conduct an additional sensitivity analysis by firstly averaging the multi-realisations within each model and then putting the 8 realisation-averaged model simulations together with the prescribed O_3 (hereafter noted as 8+1 models) into the aggressive and conservative model as input layer. The results of these new fused data are very similar to the previous calculations, with $R^2 = 0.94$, RMSE = 2.24 ppbv for the aggressive approach, and $R^2 = 0.93$, RMSE = 2.67 ppbv for the conservative approach. It shows that different numbers of realisations for each model will not significantly affect the fusion performance, indicating that the disparity in the number of realisations for a given model (e.g. 21 realisations for NASA-GISS-E2.1 while only a single realisation for NOAA-GFDL-ESM4) is not a significant issue when it comes to model data fusion. It also suggests averaging the multi-realisation ensemble before multi-model fusion takes place will still result in accurate results. This is particularly important if the model-data fusion approach is computationally expensive, as is the case for the conservative approach we have used.

One-dropout sensitivity analysis shows removing one model (with all its realisations) can achieve accuracy R^2 as 0.91 – 0.93 with RMSE 2.49 – 2.82 ppbv with the aggressive approach, and R^2 ranging 0.89 – 0.93 with RMSE 2.97 – 3.46 ppbv by conservative approach; results which are insignificantly lower than using all 8+1 CMIP6 models. However, the multi-model fusion performances are substantially reduced when only 2 models are kept (keeping only one single model will be inappropriate for the basic idea of *multi-model fusion*), as $R^2 = 0.83 – 0.87$, RMSE = 3.68 – 5.14 ppbv with the aggressive approach, and $R^2 = 0.71 – 0.78$, RMSE = 4.79 – 8.02 ppbv with the conservative approach. The aggressive-conservative performance gap converges when fusing >9 realisations, or >4 realisation-averaged models. It exposes the critical limitation of the conservative approach and that the innovative enhanced space-time BNN will not perform satisfactorily when only a few models are used for fusion, because different models have used different chemistry mechanisms, or simplifications, or have other physical differences,[85] so that limited numbers of models cannot capture the full variations of the realistic surface O_3 by BNN-based linear-combination. It also further justifies the necessity of the CMIP6 multi-model study from the perspective of raising the signal-noise ratio and enabling more credible surface O_3 datasets (the more models used in the fusion process the better the performance). We keep the aggressively and conservatively-fused outcomes separately as 2 ultimate achievements of this study, instead of mixing them up into a single dataset, because of our aim of maintaining the interpretability of the BNN-fusion processes instead of purely focusing on brute-force fitting.

4.7 Merits and Limitations

Five major merits of our study are highlighted. First, we establish an enhanced 2-stage space-time Bayesian neural network regression-based deep-learning framework to fuse multi-ensemble surface O_3 simulation, which is verified to be of high accuracy and accessible interpretability in spatiotemporal weighting. Second, we verify the better spatial extrapolation generalisability of our newly developed approach than the conventional method; and owing to the commendable spatial and temporal extrapolation potentials, our ensemble learning frameworks can be applied to a wide temporal range of surface O_3 studies. Third, as far as we are aware, our study is the first study to fuse CMIP6 model simulations for surface O_3 over the 25 historical year period of 1990-2014 by machine learning techniques, and such long-term global studies are still rather rare. Fourth, the fused and calibrated surface O_3 concentration dataset can be used by further researchers for further cross-disciplinary studies. Last but not the least, we innovatively apply Fourier-series functions for the purpose of parametrising and visualising the complex temporal periodical variations of surface O_3. However, our studies are still of several limitations. First, the model evaluation-calibration resolution is coarse as 2°×2°, and some heavily polluted regions like China, India and Africa are still lacking of observations. Second, the additional assistant features to enhance the statistical model prediction are still limited, and more variables shall be considered in further studies. Third, more detailed and deeper discussions concerning the parametric model calibration by 2-stage space-time BNN regression could have been replenished and excavated, but not included in this current paper as it is beyond the scope of this study. We aim to address some of these issues in our further research.

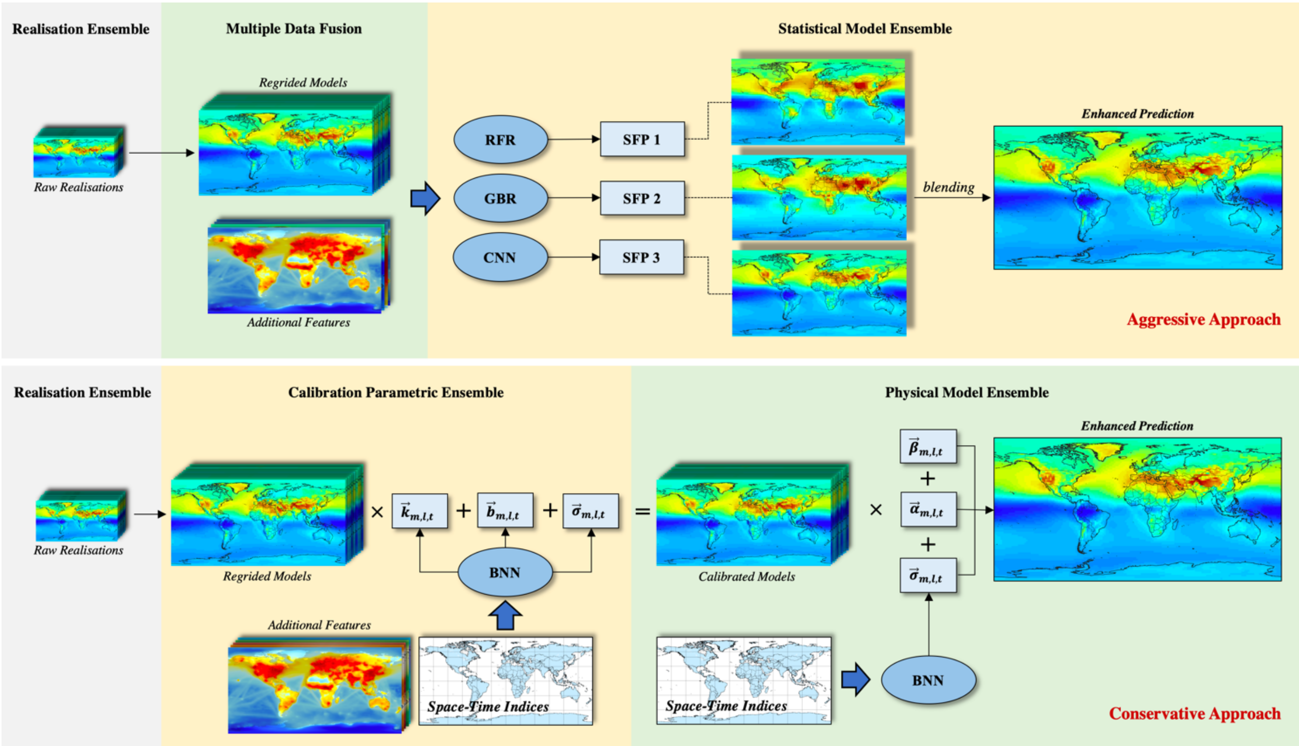

Figure 1 Schematic diagram of machine learning-based multi-model fusion by aggressive and conservative approaches. The stacking of source data layers refers to the collections of datasets with the same level in training models; the ellipses indicate elemental machine learning methodologies; and the rectangles represent the raw outputs from machine learning treatments. A total of 57 physical model simulations and 1 prescribed O_3 concentration dataset (Inputs4MIPs) are considered.

Abbreviations and denotations: RFR, *random forest regression*; GBR, *gradient boosting decision tree regression*; CNN, *convolutional neural network regression*; SFP, *semi-final product*; BNN, *Bayesian neural network regression*; $\boldsymbol{k}$, *re-scaling factor*; $\boldsymbol{b}$, *systematic bias corrector*; $\boldsymbol{\alpha}$, *individual model weight*; $\boldsymbol{\beta}$, *bias corrector*; $\boldsymbol{m}$, *physical model identifier*; $\boldsymbol{l}$, *location index*; $\boldsymbol{t}$, *temporal index*; $\boldsymbol{\sigma}$, *random noise*.

Table 1 Evaluation summary of aggressive and conservative multi-model fusion for surface ozone. The model evaluation metrics include the cross-validation (CV), test and full dataset overall coefficient of determination (R^2), the root mean squared error (RMSE), the normalised mean bias (NMB), and the linear regression slope (k) and intercept (b). Both two statistical models are evaluated separately for each 5-year period, season and continent to assess the spatiotemporal performances.

	Aggressive Approach							**Conservative Approach**						
	CV-R^2	*test-R^2*	*full-R^2*	RMSE	NMB	*k*	*b*	*CV-R^2*	*test-R^2*	*full-R^2*	RMSE	NMB	*k*	*b*
Period														
1990-1994	0.91	0.90	0.94	2.00	3.41	1.11	-1.62	0.92	0.91	0.93	2.00	0.02	0.98	0.59
1995-1999	0.90	0.90	0.94	1.74	1.71	1.09	-1.26	0.92	0.91	0.92	2.10	0.84	0.97	0.66
2000-2004	0.91	0.91	0.95	1.71	0.88	1.09	-1.16	0.91	0.91	0.93	2.28	0.71	0.97	0.95
2005-2009	0.91	0.91	0.96	1.68	1.11	1.09	-1.17	0.91	0.91	0.91	2.22	0.83	0.97	0.82
2010-2014	0.94	0.93	0.96	1.71	0.88	1.09	-1.16	0.92	0.91	0.94	2.28	0.71	0.97	0.95
Region														
Europe	0.91	0.91	0.94	1.94	2.40	1.12	-1.61	0.92	0.91	0.92	2.02	1.27	0.98	0.37
North America	0.93	0.93	0.96	1.61	1.27	1.08	-1.19	0.91	0.91	0.93	1.96	-0.04	0.97	0.94
Latin America and the Caribbean	0.90	0.87	0.95	1.22	3.12	1.10	-0.89	0.83	0.81	0.83	2.55	3.06	0.92	1.51
Asia	0.92	0.92	0.95	2.14	4.03	1.12	-1.65	0.90	0.90	0.92	2.96	1.85	0.96	0.90
Africa	0.90	0.86	0.90	2.13	2.82	1.19	-2.33	0.82	0.80	0.84	3.69	-3.81	0.93	2.88
Oceania	0.94	0.91	0.96	0.91	0.68	1.08	-0.78	0.83	0.81	0.84	2.13	-1.05	0.88	2.65
Season														
March-May	0.93	0.90	0.97	1.91	0.84	1.13	-0.65	0.94	0.91	0.96	2.06	0.89	0.99	0.97
June-August	0.94	0.92	0.98	1.78	1.12	1.09	-0.86	0.94	0.92	0.95	2.14	0.74	0.97	0.75
September-November	0.93	0.89	0.98	1.75	3.09	1.12	-0.57	0.93	0.90	0.95	2.07	0.10	0.98	0.69
December-February	0.93	0.90	0.98	1.80	3.05	1.14	-0.60	0.93	0.90	0.95	2.19	0.54	0.98	0.51
TOAR	0.94	0.89	0.96	1.81	2.01	1.05	-1.35	0.90	0.88	0.95	2.12	0.57	0.97	0.71

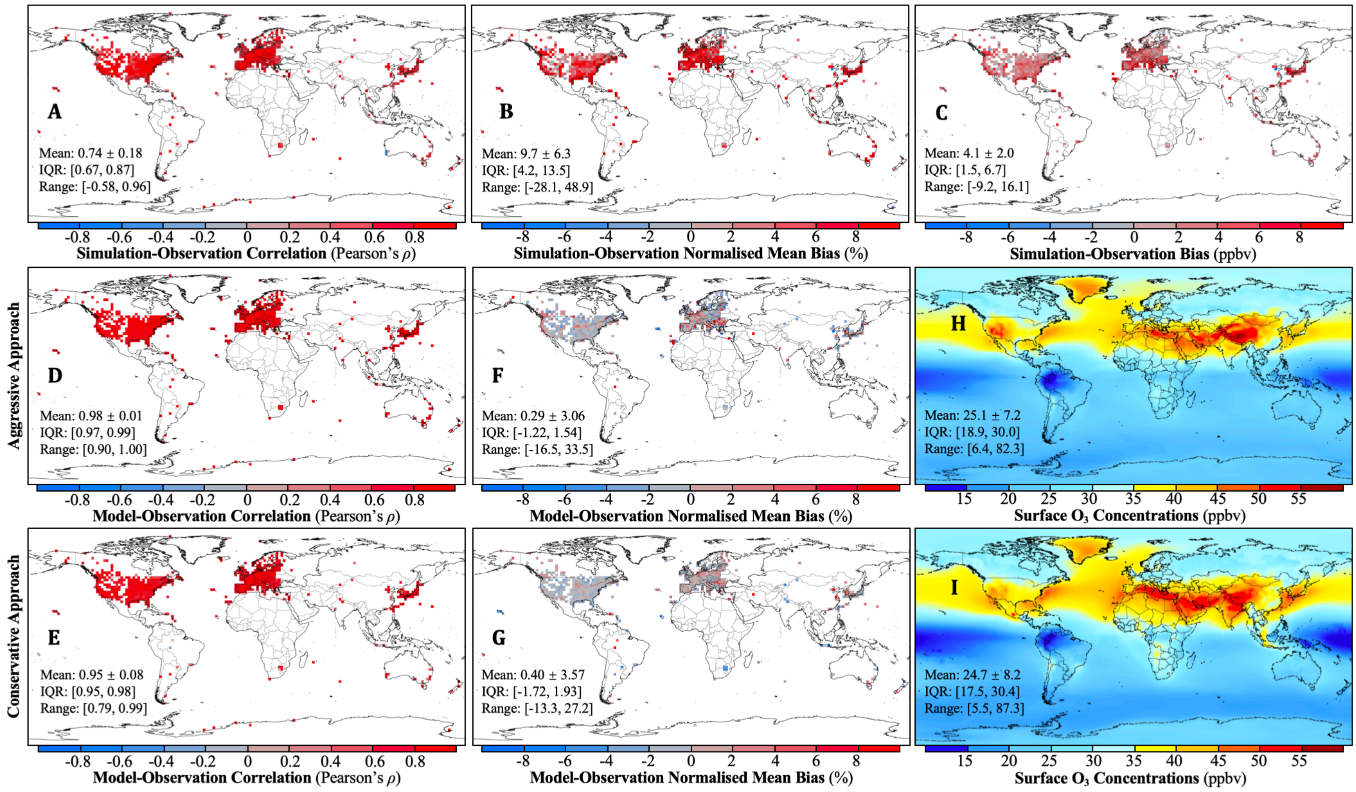

Figure 2 Model-observation evaluation for the raw CMIP6 surface ozone simulation-ensemble and multi-model fusion by both aggressive and conservative approaches. A-C: Simulation-observation synchronicity, absolute and relative biases for 57+1 CMIP6 simulation ensemble. Model evaluations are conducted on TOAR observation covered sites across 1990-2014. **D-G:** Evaluations of aggressively and conservatively integrated surface ozone concentrations in terms of the overall model-observation synchronicity and bias. **H-I:** Multi-model and TOAR-observation assimilated historical global surface ozone concentrations by aggressive and conservative approaches. The 25-year average surface ozone concentrations during 1990-2014 are mapped as summary. All spatial resolutions are set as 2°×2°, and the temporal interval is set to month.

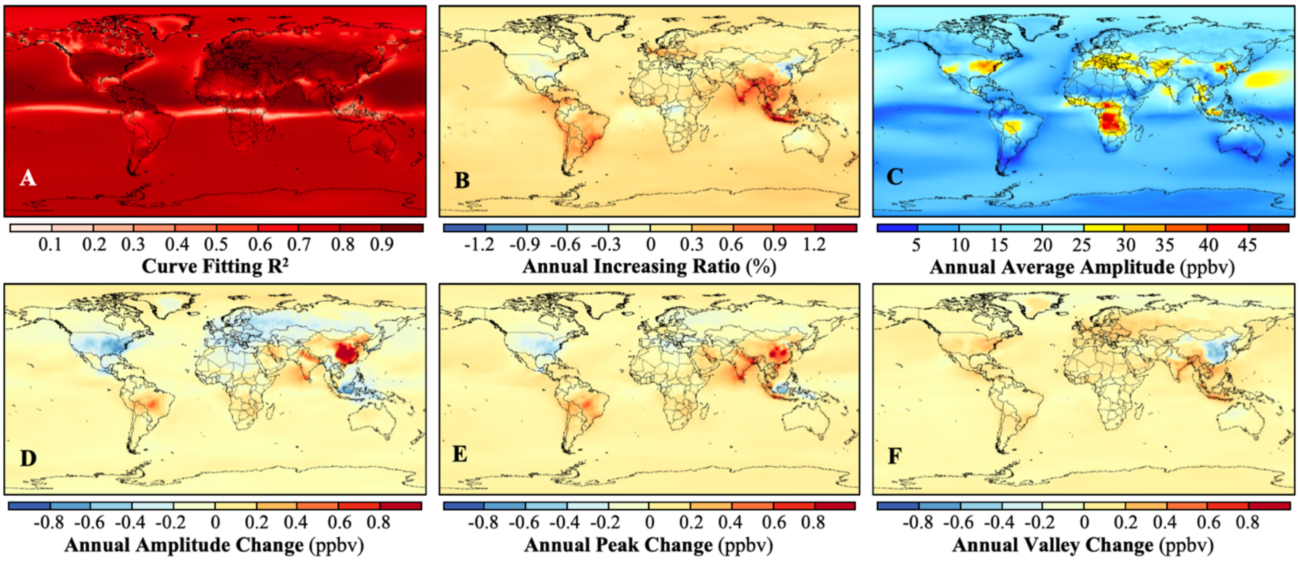

Figure 3 Spatiotemporal variability parametrisation for CMIP6 multi-model ensemble assimilated surface ozone concentrations during 1990-2014 by the conservative approach. The ensemble-learning predicted concentrations are clustered by month. **A:** Fourier-series function-based curve-fitting quality for grid-specific surface ozone variabilities against temporal sequence, quantified by R^2. **B:** Annual increasing ratio for yearly average surface ozone concentrations, estimated by $12a_1$. **C:** Annual average intra-year amplitude as the peak-valley differences, estimated by $2b_0$. **D:** Annual average linear change rates of the intra-year amplitudes, estimated by $24b_1$. **E-F:** Averaged annual change rates of peak and valley concentrations, deduced from the fitted second-order Fourier-series function.

Supporting Information. Further detailed information on CMIP6 AerChemMIP Surface O_3 historical simulation participant research institutes, and annotations on atmospheric module settings. A total of 13 supplementary figures and 3 tables.

Acknowledgments

The authors are funded by Natural Environment Research Council (NERC) and National Centre for Atmospheric Science (NCAS). We thank Youngsub Matthew Shin (University of Cambridge) for advising the data collection and pre-processing, Ushnish Sengupta (University of Cambridge) and Matt Amos (Lancaster University) for sharing their Python space-time Bayesian neural network core, and Mingtao Xia (University of California, Los Angeles) for scrutinising the code optimisation. We are also grateful to the editor and 3 anonymous reviewers for their insightful revision comments to substantially improve the manuscript.

Data and code availability

Core Python codes to construct the first-stage calibration-oriented and second-stage assimilation-targeted Bayesian neural network regressions are available at: https://github.com/csuen27/BayesNN, scheduled with regular upgrades every half-year to fit into the latest deep learning frameworks. The CMIP6 simulations with associated metadata can be accessed at: https://esgf-node.llnl.gov/search/cmip6. CMIP6 collaborators keep updating the simulation repository, whether adding new ensemble experiments or retracting ones when constructive improvements are to be made, and correspondingly data fusion works will be updated. The up-to-date assimilated surface O_3 concentrations can be shared by the authors for academic use upon request.

REFERENCES

1. Stocker, T., *Climate change 2013: the physical science basis: Working Group I contribution to the Fifth assessment report of the Intergovernmental Panel on Climate Change*. Cambridge University Press: 2014.
2. Young, P. J.; Archibald, A. T.; Bowman, K. W.; Lamarque, J. F.; Naik, V.; Stevenson, D. S.; Tilmes, S.; Voulgarakis, A.; Wild, O.; Bergmann, D.; Cameron-Smith, P.; Cionni, I.; Collins, W. J.; Dalsøren, S. B.; Doherty, R. M.; Eyring, V.; Faluvegi, G.; Horowitz, L. W.; Josse, B.; Lee, Y. H.; MacKenzie, I. A.; Nagashima, T.; Plummer, D. A.; Righi, M.; Rumbold, S. T.; Skeie, R. B.; Shindell, D. T.; Strode, S. A.; Sudo, K.; Szopa, S.; Zeng, G., Pre-industrial to end 21st century projections of tropospheric ozone from the Atmospheric Chemistry and Climate Model Intercomparison Project (ACCMIP). *Atmos Chem Phys* **2013,** *13*, (4), 2063-2090.
3. Wilson, S. R.; Madronich, S.; Longstreth, J. D.; Solomon, K. R., Interactive effects of changing stratospheric ozone and climate on tropospheric composition and air quality, and the consequences for human and ecosystem health. *Photochem Photobiol Sci* **2019,** *18*, (3), 775-803.
4. Monks, P. S.; Archibald, A. T.; Colette, A.; Cooper, O.; Coyle, M.; Derwent, R.; Fowler, D.; Granier, C.; Law, K. S.; Mills, G. E.; Stevenson, D. S.; Tarasova, O.; Thouret, V.; von Schneidemesser, E.; Sommariva, R.; Wild, O.; Williams, M. L., Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer. *Atmos Chem Phys* **2015,** *15*, (15), 8889-8973.
5. Avnery, S.; Mauzerall, D. L.; Liu, J.; Horowitz, L. W., Global crop yield reductions due to surface ozone exposure: 1. Year 2000 crop production losses and economic damage. *Atmos Environ* **2011,** *45*, (13), 2284-2296.
6. Ghude, S. D.; Jena, C.; Chate, D. M.; Beig, G.; Pfister, G. G.; Kumar, R.; Ramanathan, V., Reductions in India's crop yield due to ozone. *Geophys Res Lett* **2014,** *41*, (15), 5685-5691.
7. Wiegman, C. H.; Li, F.; Clarke, C. J.; Jazrawi, E.; Kirkham, P.; Barnes, P. J.; Adcock, I. M.; Chung, K. F., A comprehensive analysis of oxidative stress in the ozone-induced lung inflammation mouse model. *Clin Sci (Lond)* **2014,** *126*, (6), 425-440.
8. McConnell, R.; Berhane, K.; Gilliland, F.; London, S. J.; Islam, T.; Gauderman, W. J.; Avol, E.; Margolis, H. G.; Peters, J. M., Asthma in exercising children exposed to ozone: a cohort study. *Lancet* **2002,** *359*, (9304), 386-391.
9. Sheffield, P. E.; Zhou, J.; Shmool, J. L.; Clougherty, J. E., Ambient ozone exposure and children's acute asthma in New York City: a case-crossover analysis. *Environ Health* **2015,** *14*, (1), 25.
10. Ghude, S. D.; Chate, D. M.; Jena, C.; Beig, G.; Kumar, R.; Barth, M. C.; Pfister, G. G.; Fadnavis, S.; Pithani, P., Premature mortality in India due to $PM_{2.5}$ and ozone exposure. *Geophys Res Lett* **2016,** *43*, (9), 4650-4658.
11. Qiu, X.; Wei, Y.; Wang, Y.; Di, Q.; Sofer, T.; Awad, Y. A.; Schwartz, J., Inverse probability weighted distributed lag effects of short-term exposure to $PM_{2.5}$ and ozone on CVD hospitalisations in New England Medicare participants - Exploring the causal effects. *Environ Res* **2020,** *182*, 109095.
12. Turner, M. C.; Jerrett, M.; Pope, C. A., 3rd; Krewski, D.; Gapstur, S. M.; Diver, W. R.; Beckerman, B. S.; Marshall, J. D.; Su, J.; Crouse, D. L.; Burnett, R. T., Long-Term Ozone Exposure and Mortality in a Large Prospective Study. *Am J Respir Crit Care Med* **2016,** *193*, (10), 1134-42.
13. Di, Q.; Wang, Y.; Zanobetti, A.; Wang, Y.; Koutrakis, P.; Choirat, C.; Dominici, F.; Schwartz, J. D., Air Pollution and Mortality in the Medicare Population. *N Engl J Med* **2017,** *376*, (26), 2513-2522.
14. Bell, M. L.; McDermott, A.; Zeger, S. L.; Samet, J. M.; Dominici, F., Ozone and short-term mortality in 95 US urban communities, 1987-2000. *JAMA* **2004,** *292*, (19), 2372-8.
15. Institute for Health Metrics and Evaluation., GBD Compare Data Visualization. http://vizhub.healthdata.org/gbd-compare (*accessed on December 18, 2020*).
16. Weichenthal, S.; Pinault, L. L.; Burnett, R. T., Impact of oxidant gases on the relationship between outdoor fine particulate air pollution and nonaccidental, cardiovascular, and respiratory mortality. *Scientific Reports* **2017,** *7*, (1), 1-10.
17. Schultz, M. G.; Schröder, S.; Lyapina, O.; Cooper, O.; Galbally, I.; Petropavlovskikh, I.; Von Schneidemesser, E.; Tanimoto, H.; Elshorbany, Y.; Naja, M.; Seguel, R.; Dauert, U.; Eckhardt, P.; Feigenspahn, S.; Fiebig, M.; Hjellbrekke, A.-G.; Hong, Y.-D.; Christian Kjeld, P.; Koide, H.; Lear, G.; Tarasick, D.; Ueno, M.; Wallasch, M.; Baumgardner, D.; Chuang, M.-T.; Gillett, R.; Lee, M.; Molloy, S.; Moolla, R.; Wang, T.; Sharps, K.; Adame, J. A.; Ancellet, G.; Apadula, F.; Artaxo, P.; Barlasina, M.; Bogucka, M.; Bonasoni, P.; Chang, L.; Colomb, A.; Cuevas, E.; Cupeiro, M.; Degorska, A.; Ding, A.; Fröhlich, M.; Frolova, M.; Gadhavi, H.; Gheusi, F.; Gilge, S.; Gonzalez, M. Y.; Gros, V.; Hamad, S. H.; Helmig, D.; Henriques, D.; Hermansen, O.; Holla, R.; Huber, J.; Im, U.; Jaffe, D. A.; Komala, N.; Kubistin, D.; Lam, K.-S.; Laurila, T.; Lee, H.; Levy, I.; Mazzoleni, C.; Mazzoleni, L.; McClure-Begley, A.; Mohamad, M.; Murovic, M.; Navarro-Comas, M.; Nicodim, F.; Parrish, D.; Read, K. A.; Reid, N.; Ries, L.; Saxena, P.; Schwab, J. J.; Scorgie, Y.; Senik, I.; Simmonds, P.; Sinha, V.; Skorokhod, A.; Spain, G.; Spangl, W.; Spoor, R.; Springston, S. R.; Steer, K.; Steinbacher, M.; Suharguniyawan, E.; Torre, P.; Trickl, T.; Weili, L.; Weller, R.; Xu, X.; Xue, L.; Zhiqiang, M., Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations. *Elementa-Sci Anthrop* **2017,** *5*, 58-83.

18. Zoogman, P.; Jacob, D. J.; Chance, K.; Worden, H. M.; Edwards, D. P.; Zhang, L., Improved monitoring of surface ozone by joint assimilation of geostationary satellite observations of ozone and CO. *Atmos Environ* **2014,** *84*, 254-261.

19. Archibald, A. T.; O'Connor, F. M.; Abraham, N. L.; Archer-Nicholls, S.; Chipperfield, M. P.; Dalvi, M.; Folberth, G. A.; Dennison, F.; Dhomse, S. S.; Griffiths, P. T.; Hardacre, C.; Hewitt, A. J.; Hill, R.; Johnson, C. E.; Keeble, J.; Köhler, M. O.; Morgenstern, O.; Mulchay, J. P.; Ordóñez, C.; Pope, R. J.; Rumbold, S.; Russo, M. R.; Savage, N.; Sellar, A.; Stringer, M.; Turnock, S.; Wild, O.; Zeng, G., Description and evaluation of the UKCA stratosphere-troposphere chemistry scheme (StratTrop vn 1.0) implemented in UKESM1. *Geosci Model Dev* **2019**, 1223–1266.

20. Wang, T.; Xue, L.; Brimblecombe, P.; Lam, Y. F.; Li, L.; Zhang, L., Ozone pollution in China: A review of concentrations, meteorological influences, chemical precursors, and effects. *Sci Total Environ* **2017,** *575*, 1582-1596.

21. Archibald, A.; Neu, J.; Elshorbany, Y.; Cooper, O.; Young, P.; Akiyoshi, H.; Cox, R.; Coyle, M.; Derwent, R.; Deushi, M., Tropospheric Ozone Assessment ReportA critical review of changes in the tropospheric ozone burden and budget from 1850 to 2100. *Elementa-Sci Anthrop* **2020,** *8*, (1), 34-86.

22. Byun, D.; Schere, K. L., Review of the Governing Equations, Computational Algorithms, and Other Components of the Models-3 Community Multiscale Air Quality (CMAQ) Modeling System. *Appl Mech Rev* **2006,** *59*, (2), 51-77.

23. Appel, K. W.; Napelenok, S. L.; Foley, K. M.; Pye, H. O. T.; Hogrefe, C.; Luecken, D. J.; Bash, J. O.; Roselle, S. J.; Pleim, J. E.; Foroutan, H.; Hutzell, W. T.; Pouliot, G. A.; Sarwar, G.; Fahey, K. M.; Gantt, B.; Gilliam, R. C.; Heath, N. K.; Kang, D.; Mathur, R.; Schwede, D. B.; Spero, T. L.; Wong, D. C.; Young, J. O., Description and evaluation of the Community Multiscale Air Quality (CMAQ) modeling system version 5.1. *Geosci Model Dev* **2017,** *10*, (4), 1703-1732.

24. Tesche, T. W.; Morris, R.; Tonnesen, G.; McNally, D.; Boylan, J.; Brewer, P., CMAQ/CAMx annual 2002 performance evaluation over the eastern US. *Atmos Environ* **2006,** *40*, (26), 4906-4919.

25. Mar, K. A.; Ojha, N.; Pozzer, A.; Butler, T. M., Ozone air quality simulations with WRF-Chem (v3.5.1) over Europe: model evaluation and chemical mechanism comparison. *Geosci Model Dev* **2016,** *9*, (10), 3699-3728.

26. Mann, G. W.; Carslaw, K. S.; Spracklen, D. V.; Ridley, D. A.; Manktelow, P. T.; Chipperfield, M. P.; Pickering, S. J.; Johnson, C. E., Description and evaluation of GLOMAP-mode: a modal global aerosol microphysics model for the UKCA composition-climate model. *Geosci Model Dev* **2010,** *3*, (2), 519-551.

27. McLaren, A.; Banks, H.; Durman, C.; Gregory, J.; Johns, T.; Keen, A.; Ridley, J.; Roberts, M.; Lipscomb, W.; Connolley, W., Evaluation of the sea ice simulation in a new coupled atmosphere-ocean climate model (HadGEM1). *Journal of Geophysical Research: Oceans* **2006,** *111*, (C12), 14-30.

28. Sellar, A. A.; Walton, J.; Jones, C. G.; Wood, R.; Abraham, N. L.; Andrejczuk, M.; Andrews, M. B.; Andrews, T.; Archibald, A. T.; Mora, L.; Dyson, H.; Elkington, M.; Ellis, R.; Florek, P.; Good, P.; Gohar, L.; Haddad, S.; Hardiman, S. C.; Hogan, E.; Iwi, A.; Jones, C. D.; Johnson, B.; Kelley, D. I.; Kettleborough, J.; Knight, J. R.; Köhler, M. O.; Kuhlbrodt, T.; Liddicoat, S.; Linova-Pavlova, I.; Mizielinski, M. S.; Morgenstern, O.; Mulcahy, J.; Neininger, E.; O'Connor, F. M.; Petrie, R.; Ridley, J.; Rioual, J. C.; Roberts, M.; Robertson, E.; Rumbold, S.; Seddon, J.; Shepherd, H.; Shim, S.; Stephens, A.; Teixiera, J. C.; Tang, Y.; Williams, J.; Wiltshire, A.; Griffiths, P. T., Implementation of U.K. Earth System Models for CMIP6. *J Adv Model Earth Syst* **2020,** *12*, (4), e2019MS001946.

29. Young, P. J.; Naik, V.; Fiore, A. M.; Gaudel, A.; Guo, J.; Lin, M. Y.; Neu, J. L.; Parrish, D. D.; Rieder, H. E.; Schnell, J. L.; Tilmes, S.; Wild, O.; Zhang, L.; Ziemke, J.; Brandt, J.; Delcloo, A.; Doherty, R. M.; Geels, C.; Hegglin, M. I.; Hu, L.; Im, U.; Kumar, R.; Luhar, A.; Murray, L.; Plummer, D.; Rodriguez, J.; Saiz-Lopez, A.; Schultz, M. G.; Woodhouse, M. T.; Zeng, G.; Helmig, D.; Lewis, A., Tropospheric Ozone Assessment Report: Assessment of global-scale model performance for global and regional ozone distributions, variability, and trends. *Elementa-Sci Anthrop* **2018,** *6*, (1), 10-50.

30. Eyring, V.; Bony, S.; Meehl, G. A.; Senior, C. A.; Stevens, B.; Stouffer, R. J.; Taylor, K. E., Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organisation. *Geosci Model Dev* **2016,** *9*, (5), 1937-1958.

31. Griffiths, P. T.; Murray, L. T.; Zeng, G.; Archibald, A. T.; Emmons, L. K.; Galbally, I.; Hassler, B.; Horowitz, L. W.; Keeble, J.; Liu, J.; Moeini, O.; Naik, V.; O'Connor, F. M.; Shin, Y. M.; Tarasick, D.; Tilmes, S.; Turnock, S. T.; Wild, O.; Young, P. J.; Zanis, P., Tropospheric ozone in CMIP6 Simulations. *Atmos Chem Phys* **2021,** *21*, (5), 4187–4218.

32. Stouffer, R. J.; Eyring, V.; Meehl, G. A.; Bony, S.; Senior, C.; Stevens, B.; Taylor, K. E., CMIP5 Scientific Gaps and Recommendations for CMIP6. *B Am Meteorol Soc* **2017,** *98*, (1), 95-105.

33. Feng, L.; Smith, S. J.; Braun, C.; Crippa, M.; Gidden, M. J.; Hoesly, R.; Klimont, Z.; van Marle, M.; van den Berg, M.; van der Werf, G. R., The generation of gridded emissions data for CMIP6. *Geosci Model Dev* **2020,** *13*, (2), 461-482.

34. Collins, W. J.; Lamarque, J.-F.; Schulz, M.; Boucher, O.; Eyring, V.; Hegglin, M. I.; Maycock, A.; Myhre, G.; Prather, M.; Shindell, D.; Smith, S. J., AerChemMIP: quantifying the effects of chemistry and aerosols in CMIP6. *Geosci Model Dev* **2017,** *10*, (2), 585-607.

35. Turnock, S. T.; Allen, R. J.; Andrews, M.; Bauer, S. E.; Deushi, M.; Emmons, L.; Good, P.; Horowitz, L.; John, J. G.; Michou, M.; Nabat, P.; Naik, V.; Neubauer, D.; O'Connor, F. M.; Olivié, D.; Oshima, N.; Schulz, M.; Sellar, A.; Shim, S.; Takemura, T.; Tilmes, S.; Tsigaridis, K.; Wu, T.; Zhang, J., Historical and future changes in air pollutants from CMIP6 models. *Atmos Chem Phys* **2020,** *20*, (23), 14547-14579.

36. Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M. B.; Choirat, C.; Koutrakis, P.; Lyapustin, A., Assessing

NO2 Concentration and Model Uncertainty with High Spatiotemporal Resolution across the Contiguous United States Using Ensemble Model Averaging. *Environ Sci Technol* **2019,** *54*, (3), 1372-1384.

37. Di, Q.; Amini, H.; Shi, L.; Kloog, I.; Silvern, R.; Kelly, J.; Sabath, M. B.; Choirat, C.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Mickley, L. J.; Schwartz, J., An ensemble-based model of $PM_{2.5}$ concentration across the contiguous United States with high spatiotemporal resolution. *Environ Int* **2019,** *130*, 104909.

38. Di, Q.; Kloog, I.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Schwartz, J., Assessing $PM_{2.5}$ Exposures with High Spatiotemporal Resolution across the Continental United States. *Environ Sci Technol* **2016,** *50*, (9), 4712-21.

39. Lyu, B.; Hu, Y.; Zhang, W.; Du, Y.; Luo, B.; Sun, X.; Sun, Z.; Deng, Z.; Wang, X.; Liu, J.; Wang, X.; Russell, A. G., Fusion Method Combining Ground-Level Observations with Chemical Transport Model Predictions Using an Ensemble Deep Learning Framework: Application in China to Estimate Spatiotemporally-Resolved $PM_{2.5}$ Exposure Fields in 2014-2017. *Environ Sci Technol* **2019,** *53*, (13), 7306-7315.

40. Ren, X.; Mi, Z.; Georgopoulos, P. G., Comparison of Machine Learning and Land Use Regression for fine scale spatiotemporal estimation of ambient air pollution: Modeling ozone concentrations across the contiguous United States. *Environ Int* **2020,** *142*, 105827.

41. Rudin, C., Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* **2019,** *1*, (5), 206-215.

42. Wang, A.; Xu, J.; Tu, R.; Saleh, M.; Hatzopoulou, M., Potential of machine learning for prediction of traffic related air pollution. *Transport Res D-Tr E* **2020,** *88*, 102599.

43. Danek, C.; Shi, X.; Stepanek, C.; Yang, H.; Barbi, D.; Hegewald, J.; Lohmann, G., AWI AWI-ESM1.1-LR model output prepared for CMIP6 CMIP historical. In Earth System Grid Federation: 2020.

44. Wu, T.; Lu, Y.; Fang, Y.; Xin, X.; Li, L.; Li, W.; Jie, W.; Zhang, J.; Liu, Y.; Zhang, L.; Zhang, F.; Zhang, Y.; Wu, F.; Li, J.; Chu, M.; Wang, Z.; Shi, X.; Liu, X.; Wei, M.; Huang, A.; Zhang, Y.; Liu, X., The Beijing Climate Center Climate System Model (BCC-CSM): the main progress from CMIP5 to CMIP6. *Geosci Model Dev* **2019,** *12*, (4), 1573-1600.

45. Wu, T.; Yu, R.; Lu, Y.; Jie, W.; Fang, Y.; Zhang, J.; Zhang, L.; Xin, X.; Li, L.; Wang, Z., BCC-CSM2-HR: A High-Resolution Version of the Beijing Climate Center Climate System Model. *Geosci Model Dev* **2020,** *14*, (5), 2977-3006.

46. Wu, T.; Chu, M.; Dong, M.; Fang, Y.; Jie, W.; Li, J.; Li, W.; Liu, Q.; Shi, X.; Xin, X.; Yan, J.; Zhang, F.; Zhang, J.; Zhang, L.; Zhang, Y., BCC BCC-CSM2-MR model output prepared for CMIP6 CMIP piControl. In Earth System Grid Federation: 2018.

47. Boucher, O.; Servonnat, J.; Albright, A. L.; Aumont, O.; Balkanski, Y.; Bastrikov, V.; Bekki, S.; Bonnet, R.; Bony, S.; Bopp, L., Presentation and evaluation of the IPSL-CM6A-LR climate model. *J Adv Model Earth Syst* **2020,** *12*, (7), e2019MS002010.

48. Boucher, O.; Denvil, S.; Levavasseur, G.; Cozic, A.; Caubel, A.; Foujols, M.-A.; Meurdesoif, Y.; Cadule, P.; Devilliers, M.; Ghattas, J.; Lebas, N.; Lurton, T.; Mellul, L.; Musat, I.; Mignot, J.; Cheruy, F., IPSL IPSL-CM6A-LR model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2018.

49. Mauritsen, T.; Bader, J.; Becker, T.; Behrens, J.; Bittner, M.; Brokopf, R.; Brovkin, V.; Claussen, M.; Crueger, T.; Esch, M.; Fast, I.; Fiedler, S.; Flaschner, D.; Gayler, V.; Giorgetta, M.; Goll, D. S.; Haak, H.; Hagemann, S.; Hedemann, C.; Hohenegger, C.; Ilyina, T.; Jahns, T.; Jimenez-de-la-Cuesta, D.; Jungclaus, J.; Kleinen, T.; Kloster, S.; Kracher, D.; Kinne, S.; Kleberg, D.; Lasslop, G.; Kornblueh, L.; Marotzke, J.; Matei, D.; Meraner, K.; Mikolajewicz, U.; Modali, K.; Mobis, B.; Muller, W. A.; Nabel, J.; Nam, C. C. W.; Notz, D.; Nyawira, S. S.; Paulsen, H.; Peters, K.; Pincus, R.; Pohlmann, H.; Pongratz, J.; Popp, M.; Raddatz, T. J.; Rast, S.; Redler, R.; Reick, C. H.; Rohrschneider, T.; Schemann, V.; Schmidt, H.; Schnur, R.; Schulzweida, U.; Six, K. D.; Stein, L.; Stemmler, I.; Stevens, B.; von Storch, J. S.; Tian, F.; Voigt, A.; Vrese, P.; Wieners, K. H.; Wilkenskjeld, S.; Winkler, A.; Roeckner, E., Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO_2. *J Adv Model Earth Syst* **2019,** *11*, (4), 998-1038.

50. Müller, W. A.; Jungclaus, J. H.; Mauritsen, T.; Baehr, J.; Bittner, M.; Budich, R.; Bunzel, F.; Esch, M.; Ghosh, R.; Haak, H., A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). *J Adv Model Earth Syst* **2018,** *10*, (7), 1383-1413.

51. Gutjahr, O.; Putrasahan, D.; Lohmann, K.; Jungclaus, J. H.; von Storch, J.-S.; Brüggemann, N.; Haak, H.; Stössel, A., Max Planck Institute Earth System Model (MPI-ESM1.2) for the High-Resolution Model Intercomparison Project (HighResMIP). *Geosci Model Dev* **2019,** *12*, (7), 3241-3281.

52. von Storch, J.-S.; Putrasahan, D.; Lohmann, K.; Gutjahr, O.; Jungclaus, J.; Bittner, M.; Haak, H.; Wieners, K.-H.; Giorgetta, M.; Reick, C.; Esch, M.; Gayler, V.; de Vrese, P.; Raddatz, T.; Mauritsen, T.; Behrens, J.; Brovkin, V.; Claussen, M.; Crueger, T.; Fast, I.; Fiedler, S.; Hagemann, S.; Hohenegger, C.; Jahns, T.; Kloster, S.; Kinne, S.; Lasslop, G.; Kornblueh, L.; Marotzke, J.; Matei, D.; Meraner, K.; Mikolajewicz, U.; Modali, K.; Müller, W.; Nabel, J.; Notz, D.; Peters, K.; Pincus, R.; Pohlmann, H.; Pongratz, J.; Rast, S.; Schmidt, H.; Schnur, R.; Schulzweida, U.; Six, K.; Stevens, B.; Voigt, A.; Roeckner, E., MPI-M MPI-ESM1.2-XR model output prepared for CMIP6 HighResMIP. In Earth System Grid Federation: 2017.

53. Voldoire, A.; Saint-Martin, D.; Sénési, S.; Decharme, B.; Alias, A.; Chevallier, M.; Colin, J.; Guérémy, J. F.; Michou, M.; Moine, M. P., Evaluation of CMIP6 deck experiments with CNRM-CM6-1. *J Adv Model Earth Syst* **2019,** *11*, (7), 2177-2213.

54. Michou, M.; Nabat, P.; Saint-Martin, D.; Bock, J.; Decharme, B.; Mallet, M.; Roehrig, R.; Séférian, R.; Sénési, S.; Voldoire, A., Present-day and historical aerosol and ozone characteristics in CNRM CMIP6 simulations. *J Adv Model Earth Syst* **2020,** *12*,

(1), e2019MS001816.

55. Voldoire, A., CNRM-CERFACS CNRM-CM6-1 model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2018.
56. Séférian, R.; Nabat, P.; Michou, M.; Saint-Martin, D.; Voldoire, A.; Colin, J.; Decharme, B.; Delire, C.; Berthet, S.; Chevallier, M., Evaluation of CNRM Earth System Model, CNRM-ESM2-1: Role of Earth System Processes in Present-Day and Future Climate. *J Adv Model Earth Syst* **2019,** *11*, (12), 4182-4227.
57. Seferian, R., CNRM-CERFACS CNRM-ESM2-1 model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2018.
58. Wu, T.; Zhang, F.; Zhang, J.; Jie, W.; Zhang, Y.; Wu, F.; Li, L.; Yan, J.; Liu, X.; Lu, X.; Tan, H.; Zhang, L.; Wang, J.; Hu, A., Beijing Climate Center Earth System Model version 1 (BCC-ESM1): model description and evaluation of aerosol simulations. *Geosci Model Dev* **2020,** *13*, (3), 977-1005.
59. Zhang, J.; Wu, T.; Shi, X.; Zhang, F.; Li, J.; Chu, M.; Liu, Q.; Yan, J.; Ma, Q.; Wei, M., BCC BCC-ESM1 model output prepared for CMIP6 CMIP piControl. In Earth System Grid Federation: 2018.
60. Neubauer, D.; Ferrachat, S.; Siegenthaler-Le Drian, C.; Stoll, J.; Folini, D. S.; Tegen, I.; Wieners, K.-H.; Mauritsen, T.; Stemmler, I.; Barthel, S.; Bey, I.; Daskalakis, N.; Heinold, B.; Kokkola, H.; Partridge, D.; Rast, S.; Schmidt, H.; Schutgens, N.; Stanelle, T.; Stier, P.; Watson-Parris, D.; Lohmann, U., HAMMOZ-Consortium MPI-ESM1.2-HAM model output prepared for CMIP6 AerChemMIP. In Earth System Grid Federation: 2019.
61. Yukimoto, S.; Kawai, H.; Koshiro, T.; Oshima, N.; Yoshida, K.; Urakawa, S.; Tsujino, H.; Deushi, M.; Tanaka, T.; Hosaka, M.; Yabu, S.; Yoshimura, H.; Shindo, E.; Mizuta, R.; Obata, A.; Adachi, Y.; Ishii, M., The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and Basic Evaluation of the Physical Component. *J Meteorol Soc Jpn* **2019,** *97*, (5), 931-965.
62. Yukimoto, S.; Adachi, Y.; Hosaka, M.; Sakami, T.; Yoshimura, H.; Hirabara, M.; Tanaka, T. Y.; Shindo, E.; Tsujino, H.; Deushi, M.; Mizuta, R.; Yabu, S.; Obata, A.; Nakano, H.; Koshiro, T.; Ose, T.; Kitoh, A., A New Global Climate Model of the Meteorological Research Institute: MRI-CGCM3 Model Description and Basic Performance. *J Meteorol Soc Jpn* **2012,** *90A*, (2), 23-64.
63. Yukimoto, S.; Koshiro, T.; Kawai, H.; Oshima, N.; Yoshida, K.; Urakawa, S.; Tsujino, H.; Deushi, M.; Tanaka, T.; Hosaka, M.; Yoshimura, H.; Shindo, E.; Mizuta, R.; Ishii, M.; Obata, A.; Adachi, Y., MRI MRI-ESM2.0 model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2019.
64. Shindell, D. T.; Pechony, O.; Voulgarakis, A.; Faluvegi, G.; Nazarenko, L.; Lamarque, J. F.; Bowman, K.; Milly, G.; Kovari, B.; Ruedy, R.; Schmidt, G. A., Interactive ozone and methane chemistry in GISS-E2 historical and future climate simulations. *Atmos Chem Phys* **2013,** *13*, (5), 2653-2689.
65. NASA-GISS, NASA-GISS GISS-E2.1G model output prepared for CMIP6 ISMIP6. In Earth System Grid Federation: 2018.
66. NASA-GISS, NASA-GISS GISS-E2.1H model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2018.
67. Gettelman, A.; Mills, M. J.; Kinnison, D. E.; Garcia, R. R.; Smith, A. K.; Marsh, D. R.; Tilmes, S.; Vitt, F.; Bardeen, C. G.; McInerny, J.; Liu, H. L.; Solomon, S. C.; Polvani, L. M.; Emmons, L. K.; Lamarque, J. F.; Richter, J. H.; Glanville, A. S.; Bacmeister, J. T.; Phillips, A. S.; Neale, R. B.; Simpson, I. R.; DuVivier, A. K.; Hodzic, A.; Randel, W. J., The Whole Atmosphere Community Climate Model Version 6 (WACCM6). *J Geophys Res-Atmos* **2019,** *124*, (23), 12380-12403.
68. Danabasoglu, G., NCAR CESM2-WACCM model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2019.
69. Seland, Ø.; Bentsen, M.; Olivié, D. J. L.; Toniazzo, T.; Gjermundsen, A.; Graff, L. S.; Debernard, J. B.; Gupta, A. K.; He, Y.; Kirkevåg, A.; Schwinger, J.; Tjiputra, J.; Aas, K. S.; Bethke, I.; Fan, Y.; Griesfeller, J.; Grini, A.; Guo, C.; Ilicak, M.; Karset, I. H. H.; Landgren, O. A.; Liakka, J.; Moseid, K. O.; Nummelin, A.; Spensberger, C.; Tang, H.; Zhang, Z.; Heinze, C.; Iversen, T.; Schulz, M., NCC NorESM2-LM model output prepared for CMIP6 CMIP historical. In Earth System Grid Federation: 2019.
70. Krasting, J.; John, J.; Blanton, C.; McHugh, C.; Nikonov, S.; Radhakrishnan, A.; Rand, K.; Zadeh, N.; Balaji, V.; Durachta, J., NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 CMIP. In Earth System Grid Federation, 2018.
71. Horowitz, L. W.; Naik, V.; Sentman, L.; Paulot, F.; Blanton, C.; McHugh, C.; Radhakrishnan, A.; Rand, K.; Vahlenkamp, H.; Zadeh, N. T.; Wilson, C.; Ginoux, P.; He, J.; John, J. G.; Lin, M.; Paynter, D. J.; Ploshay, J.; Zhang, A.; Zeng, Y., NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 AerChemMIP. In Earth System Grid Federation: 2018.
72. Mulcahy, J. P.; Jones, C.; Sellar, A.; Johnson, B.; Boutle, I. A.; Jones, A.; Andrews, T.; Rumbold, S. T.; Mollard, J.; Bellouin, N.; Johnson, C. E.; Williams, K. D.; Grosvenor, D. P.; McCoy, D. T., Improved Aerosol Processes and Effective Radiative Forcing in HadGEM3 and UKESM1. *J Adv Model Earth Syst* **2018,** *10*, (11), 2786-2805.
73. Sellar, A. A.; Jones, C. G.; Mulcahy, J. P.; Tang, Y.; Yool, A.; Wiltshire, A.; O'connor, F. M.; Stringer, M.; Hill, R.; Palmieri, J., UKESM1: Description and evaluation of the UK Earth System Model. *J Adv Model Earth Syst* **2019,** *11*, (12), 4513-4558.
74. Tang, Y.; Rumbold, S.; Ellis, R.; Kelley, D.; Mulcahy, J.; Sellar, A.; Walton, J.; Jones, C., MOHC UKESM1.0-LL model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2019.
75. Yool, A.; Palmiéri, J.; Jones, C.; Sellar, A.; de Mora, L.; Kuhlbrodt, T.; Popova, E.; Mulcahy, J.; Wiltshire, A.; Rumbold, S. T., Spin-up of UK Earth System Model 1 (UKESM1) for CMIP6. *J Adv Model Earth Syst* **2020**, e2019MS001933.

76. Hegglin, M.; Kinnison, D.; Lamarque, J.-F.; Plummer, D., CCMI ozone in support of CMIP6 - version 1.0. In Earth System Grid Federation: 2016.

77. Danielson, J. J.; Gesch, D. B., Global multi-resolution terrain elevation data 2010 (GMTED2010). US Department of the Interior, US Geological Survey: 2011.

78. Lloyd, C. T.; Sorichetta, A.; Tatem, A. J., High resolution global gridded data for use in population studies. *Scientific Data* **2017,** *4*, (1), 1-17.

79. Sengupta, U.; Amos, M.; Hosking, S.; Rasmussen, C. E.; Juniper, M.; Young, P., Ensembling geophysical models with Bayesian Neural Networks. *Advances in Neural Information Processing Systems* **2020,** *33*.

80. Sexton, J.; Laake, P., Standard errors for bagged and random forest estimators. *Computational Statistics & Data Analysis* **2009,** *53*, (3), 801-811.

81. Solazzo, E.; Galmarini, S., Error apportionment for atmospheric chemistry-transport models – a new approach to model evaluation. *Atmos Chem Phys* **2016,** *16*, (10), 6263-6283.

82. Kovač-Andrić, E.; Brana, J.; Gvozdić, V., Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods. *Ecological Informatics* **2009,** *4*, (2), 117-122.

83. Solazzo, E.; Galmarini, S., Error apportionment for atmospheric chemistry-transport models: a new approach to model evaluation. *Atmos Chem Phys* **2016,** *16*, (10), 6263-6283.

84. Hakim, Z. Q.; Archer-Nicholls, S.; Beig, G.; Folberth, G. A.; Sudo, K.; Abraham, N. L.; Ghude, S.; Henze, D. K.; Archibald, A. T., Evaluation of tropospheric ozone and ozone precursors in simulations from the HTAPII and CCMI model intercomparisons– a focus on the Indian subcontinent. *Atmos Chem Phys* **2019,** *19*, (9), 6437-6458.

85. Derwent, R. G.; Parrish, D. D.; Archibald, A. T.; Deushi, M.; Bauer, S. E.; Tsigaridis, K.; Shindell, D.; Horowitz, L. W.; Khan, M. A. H.; Shallcross, D. E., Intercomparison of the representations of the atmospheric chemistry of pre-industrial methane and ozone in earth system and other global chemistry-transport models. *Atmos Environ* **2021**, 118248.

SUPPORTING INFORMATION

Multi-stage Ensemble-learning-based Model Fusion for Surface Ozone Simulations: A Focus on CMIP6 Models

Authors

Zhe Sun[1,2*], Alexander T. Archibald[1,3*]

Affiliations

[1] Centre for Atmospheric Science, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK

[2] Department of Earth Sciences, University of Cambridge, Cambridge CB2 3EQ, UK

[3] National Centre for Atmospheric Science, Cambridge CB2 1EW, UK

[*] Corresponding authors:
Zhe Sun (zs347@cam.ac.uk) and Alexander T. Archibald (ata27@cam.ac.uk)

21 pages, 13 figures, and 3 tables in total.

CMIP6 Surface O_3 historical simulation participant research institutes

The 13 research institutes are Alfred Wegener Institute (AWI), Beijing Climate Centre (BCC), the HAMMOZ-Consortium (consisting of Swiss Federal Institute of Technology Zurich, Max Planck Institute for Meteorology, Forschungszentrum Jülich, University of Oxford, Finnish Meteorological Institute, Leibniz Institute for Tropospheric Research and Centre for Climate Systems Modelling at ETH Zurich), Institute Pierre-Simon Laplace (IPSL), Met Office Hadley Centre (MOHC), Met Office Natural Environment Research Council (MO-NERC), Max Planck Institute for Meteorology (MPI-M), Japan Meteorological Research Institute (MRI), NASA Goddard Institute for Space Studies (NASA-GISS), National Centre for Atmospheric Research (NCAR), Norwegian Climate Centre (NCC), National Institute of Meteorological Sciences – Korea Meteorological Administration (NIMS-KMA), and the Geophysical Fluid Dynamics Laboratory of the National Oceanic and Atmospheric Administration (NOAA-GFDL).

Detailed annotations on atmospheric module settings

Multiple sub-experiments are noted with different chaotic climate realisations (r), initialisations (i), aerosol physics (p) and forcing (f). Reproducing the multicentury global-scale weather and climate conditions shall be crucial to the CCMs which have amalgamated the atmospheric component-climate interaction effects. However, the Earth climate is so complex and chaotic that it is rather difficult to simulate the decadal realistic climate system, past, present and future. In this sense, CMIP6 completed a series of experiments to realise the near-equilibrium pre-industrial state of the climate system as control (denoted as *piControl* in DECK),[1] starting from initial climate spin-ups centuries ago and gradually reaching balance with forcing by the selected starting year for the modelling of atmospheric components, as 1850 in this current study.[2]

Considering the butterfly effect in the climate system that small disturbances in temperature, wind, humidity and other weather features even in a small place will possibly lead to different future evolution paths of the whole climate system,[3] unique initial climate conditions are realised from different simulation starting dates with the same coupled model and same atmospheric physics. All these realisations are kept, as the simulation ensemble can be able to reproduce the weather events with similar frequency as observed records, and the initial climate conditions should also have fully represented the Interdecadal Pacific Oscillation[4-8] and the Atlantic Multidecadal Oscillation,[9-12] manifested by basin-scale variability in sea surface temperature (SST).[13, 14] In addition, these multiple unique but reasonable realisations can verify how the small disturbance in SST will affect the atmospheric chemistry processes. Though the individual ensemble members of initial climate simulations differed from each other, they are still of high similarity in a relatively longer time-scale, so that are all used for CMIP6 atmospheric historical simulations.[2, 15] The initialisation time point is consistently set as 1850 within the scope of this research, so that all the experiments are marked as i_1.

CMIP6 historical simulations consider 3 aerosol microphysics as the calculation of atmospheric compositions can be realised by non-interactive (NINT) read-in of pre-computed transient aerosol and ozone fields (p_1),[16] One-Moment Aerosol scheme (OMA, p_3) with the "Tracers, Chemistry, Aerosols Direct and Indirect Effect (TCADI)" configuration,[16, 17] and more complex Multi-configuration Aerosol TRacker of mIXing state (MATRIX, p_5),[16, 18] among which p_3 and p_5 considered cloud impacts. The major difference between the OMA and TOMAS is the calculation algorithm of the particle size distribution (PSD), in which the one-moment scheme will only involve the mass concentration while the two-moment scheme can consider both the mass and the number concentrations.[18-20] As p_5 can be regarded as an extended setting of p_4 without much more computation burdens, the CMIP6 historical and future scenario simulation experiments did not include p_4, so that comparisons will only be made between p_1, p_3, and p_5 within the scope of this current study, and till present (2021), different aerosol micro-physics configurations are only considered by NASA-GISS-E2.1 model series.

Three radiative forcing prescriptions are used and noted as f_1, f_2, and f_3. Pattern f_1 represents the compositional forcing derived from OMA simulations, based on which the ozone-related heterogeneous reactions are corrected into pattern f_2; f_3 is the same as f_2, except the stratospheric surface area density of aerosols (SAD) is fixed at 1850 values.

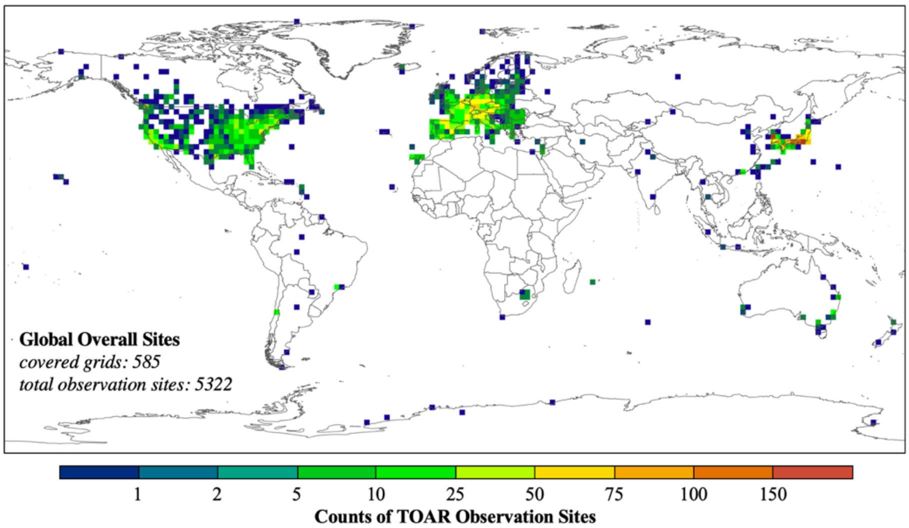

Figure S1 Global distribution of TOAR sites during 1990-2014. The TOAR ground monitoring sites are clustered into coordinational grids with 2°×2° spatial resolution, and the counts of observational sites in each grid are marked in different colours.

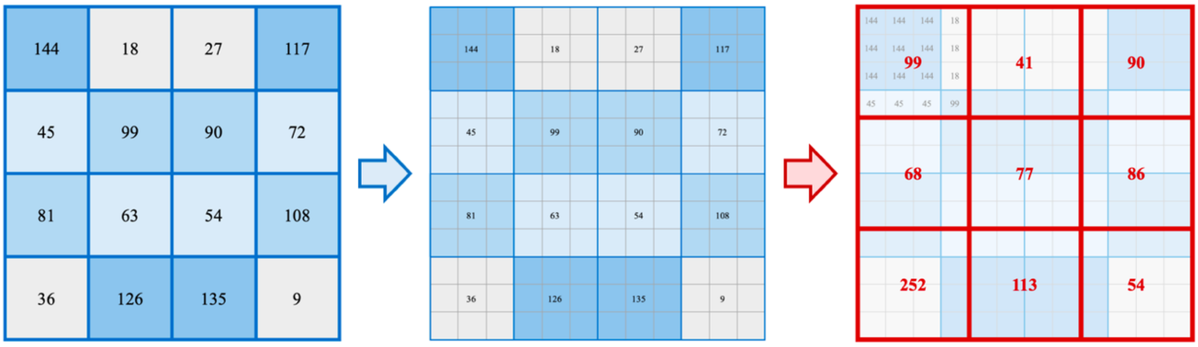

Figure S2 **Schematic graph for re-gridding process.** A faked dataset is used to illustrate the segmentation and re-aggregation process. During segmentation, interpolation is performed by deploying the same values of the larger grid for the spatial continuous variables (i.e. pollution concentrations), and using the equally distributed values for the discrete variables (i.e. population), so that during re-aggregation, values for the grids are calculated as average for continuous variables and sum for discrete variables correspondingly.

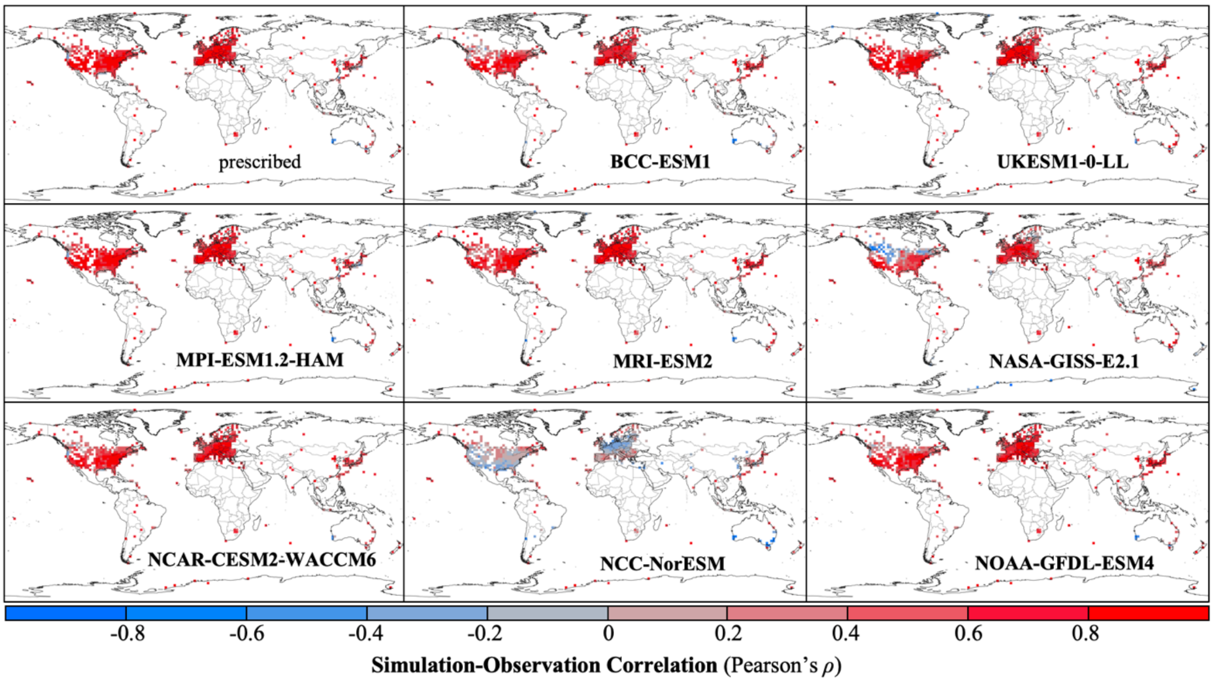

Figure S3 Correlation distributions between TOAR observations and CMIP6 multi-realisation-ensembled individual model simulations of surface ozone concentrations during 1990-2014. A total of 8 individual models together with the prescribed ozone are considered, and the grid-specific correlations are mapped onto all covered TOAR monitoring sites and scaled in blue-red colour bar where red spatial cells indicate higher positive correlations, the grey-coloured cells indicate failure in capturing the temporal variations of surface ozone concentrations, and blue-coloured cells refer to the reversed (negative) correlation between observations and simulations.

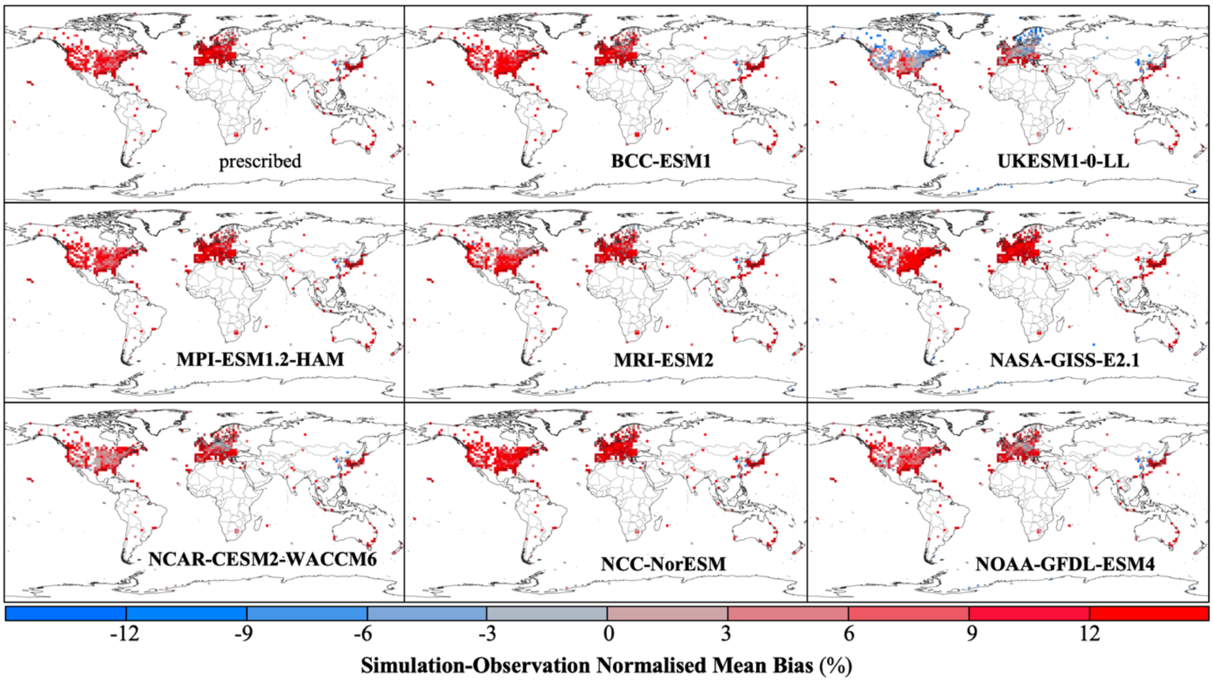

Figure S4 Global distributions of normalised mean biases between individual CMIP6 simulations and TOAR observations of surface ozone through 1990-2014. The normalised mean biases (NMB) are defined on the cell-specific 25-year full-duration (1990-2014) overall average concentrations for each individual model, mapped onto all covered TOAR monitoring sites with red-coloured cells indicating over-estimations and grey cells reflecting underestimations.

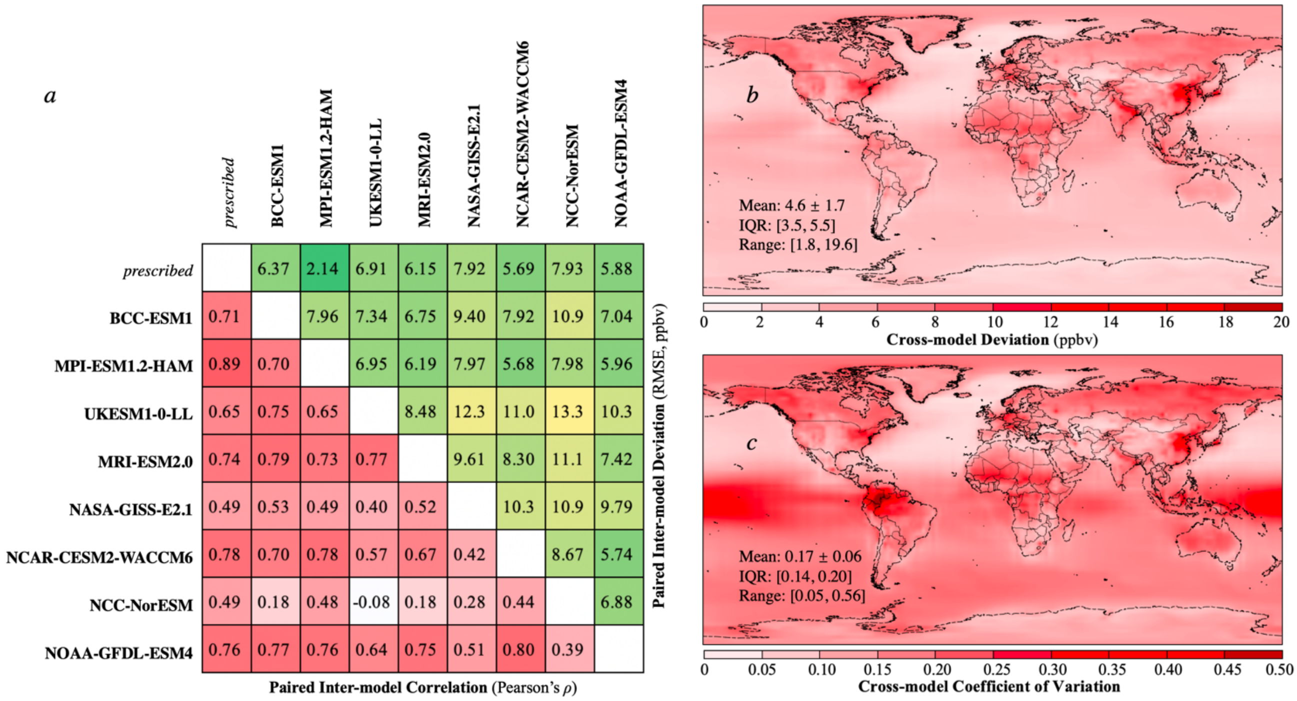

Figure S5 Inter-model similarity and divergence of CMIP6 simulation ensemble. The paired linear correlations defined as Pearson's ρ and biases quantified by root mean squared error (RMSE) were presented in lattice (panel *a*), where red correlations and green deviations representing higher similarity (i.e. higher Pearson's ρ and lower RMSE). The model-paired correlations and deviations are calculated upon the 9 multi-realisation ensemble average for the concision of summarisation, also owing to the high similarities across the realisations. The cross-model standard deviations through 1990-2014 are calculated from 58 individual simulations and mapped in panel *b*, with the relative deviations quantified in coefficient of variation (standard deviations divided by arithmetic mean) mapped in panel *c*.

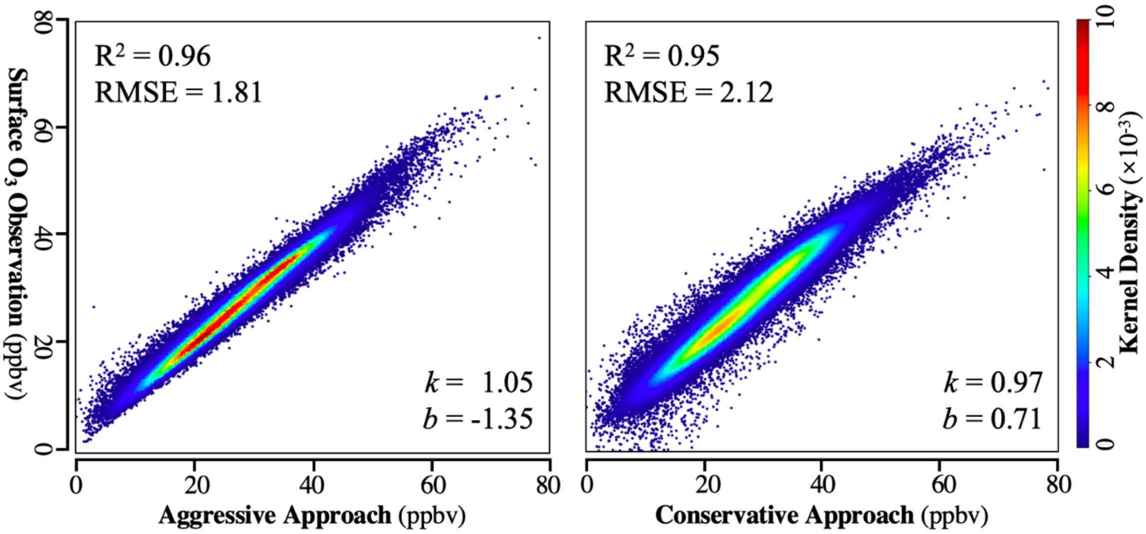

Figure S6 Model-observation evaluation for multi-model fusion by aggressive and conservative approach. Gaussian kernel density estimation is applied to construct the scatter plot. Evaluations are conducted on all observation covered regions and periods during 1990-2014. The overall R^2, RMSE, relative slope k and bias b are given for either multi-model fusion approach.

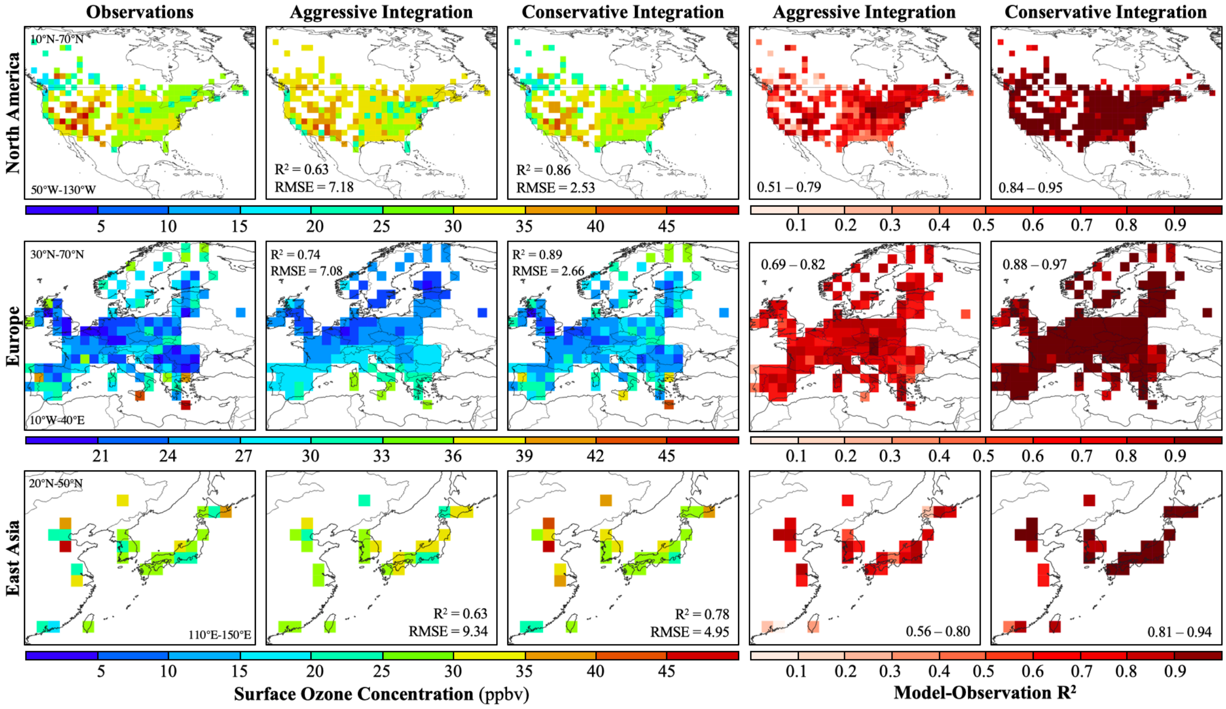

Figure S7 Regional extrapolation evaluation of aggressive and conservative multi-model ensemble integration. The spatial extrapolation experiments are conducted by 3 different training-*test* combinations, as Europe-*North America*, North America-*Europe*, and Western-*East Asia*. The first column presents the total average observed surface ozone concentrations measured by TOAR through 1990-2014. The second and third columns map the fused multi-model simulations by aggressive and conservative approaches, respectively, with the overall coefficient of determination (R^2) and root mean squared error (RMSE). The fourth and fifth columns quantify the spatial grid-specific model-observation fitting R^2 of aggressive and conservative approaches, respectively, indicated with the inter-quartile range (IQR). Overall, the conservative approach integrated performances for North America ($R^2 = 0.86$, RMSE = 2.53), Europe ($R^2 = 0.89$, RMSE = 2.66), and Asia ($R^2 = 0.78$, RMSE = 4.95) are better than aggressive approach integrated performances for North America ($R^2 = 0.63$, RMSE = 7.18), Europe ($R^2 = 0.74$, RMSE = 7.08), and Asia ($R^2 = 0.63$, RMSE = 9.34).

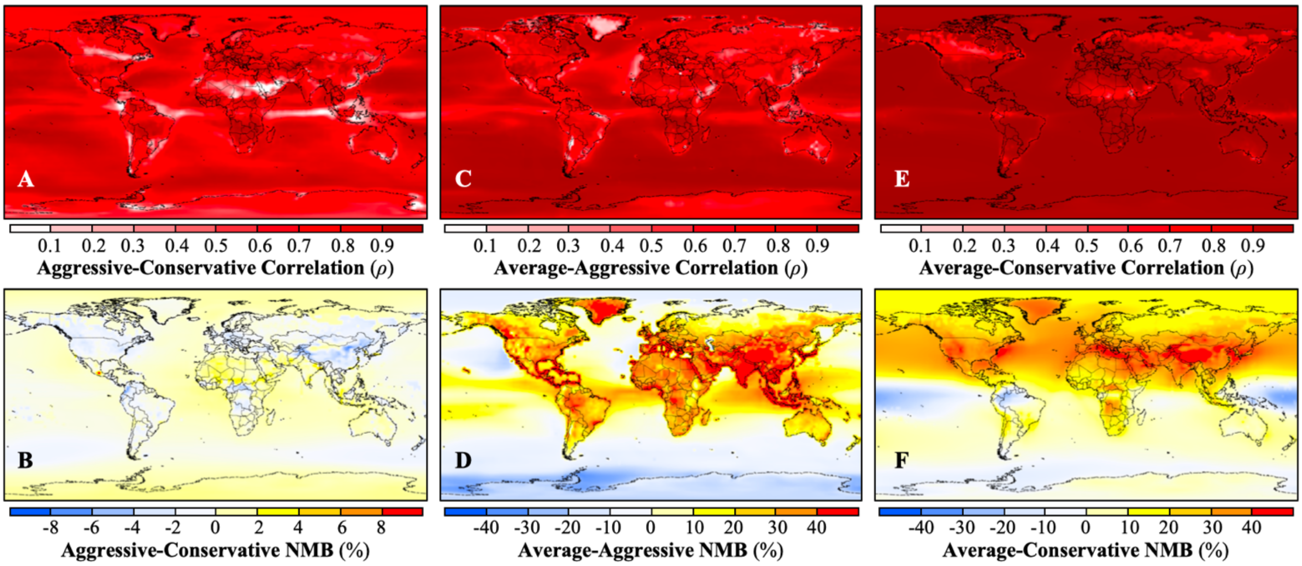

Figure S8 Similarity and discrepancy between 3 different multi-model ensemble approaches: arithmetic average, aggressive and conservative integration. The similarities are quantified as synchronicity by Pearson's correlation coefficient ρ; while the discrepancies are scaled by normalised mean bias (NMB, %). The statistics are drawn by inter-comparisons on 25-year ensembled surface ozone during 1990-2014, and summarised in average for mapping. The NMBs defined with $A - B$ are calculated as $\sum (A_i - B_i) / \sum B_i$, setting B as the reference.

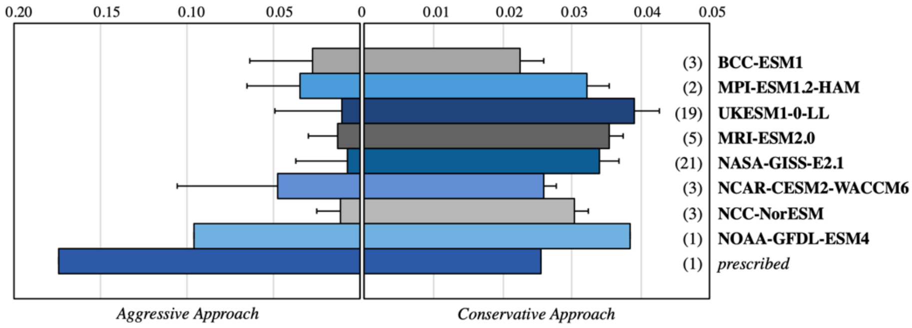

Figure S9 Individual model ensemble weights by aggressive and conservative integration approaches for 25-year simulation during 1990-2014. The 18 individual models were grouped into 4 major categories as non-interactive chemistry (NINT), UKESM1 chemistry schemes, NASA-GISS series, and other interactive chemistry driven models (INT-Chem). The weight uncertainties characterised by error bars of aggressive approach were estimated from different random seeds for ensemble learning model construction, while of conservative approach were given as the spatiotemporal variations instead, as the variability from Bayesian MCMC simulations were rather small. The weights presented in histogram referred to the model weights on TOAR covered locations, and the global weights given by Bayesian neural network regressions were also listed in the inserted columns.

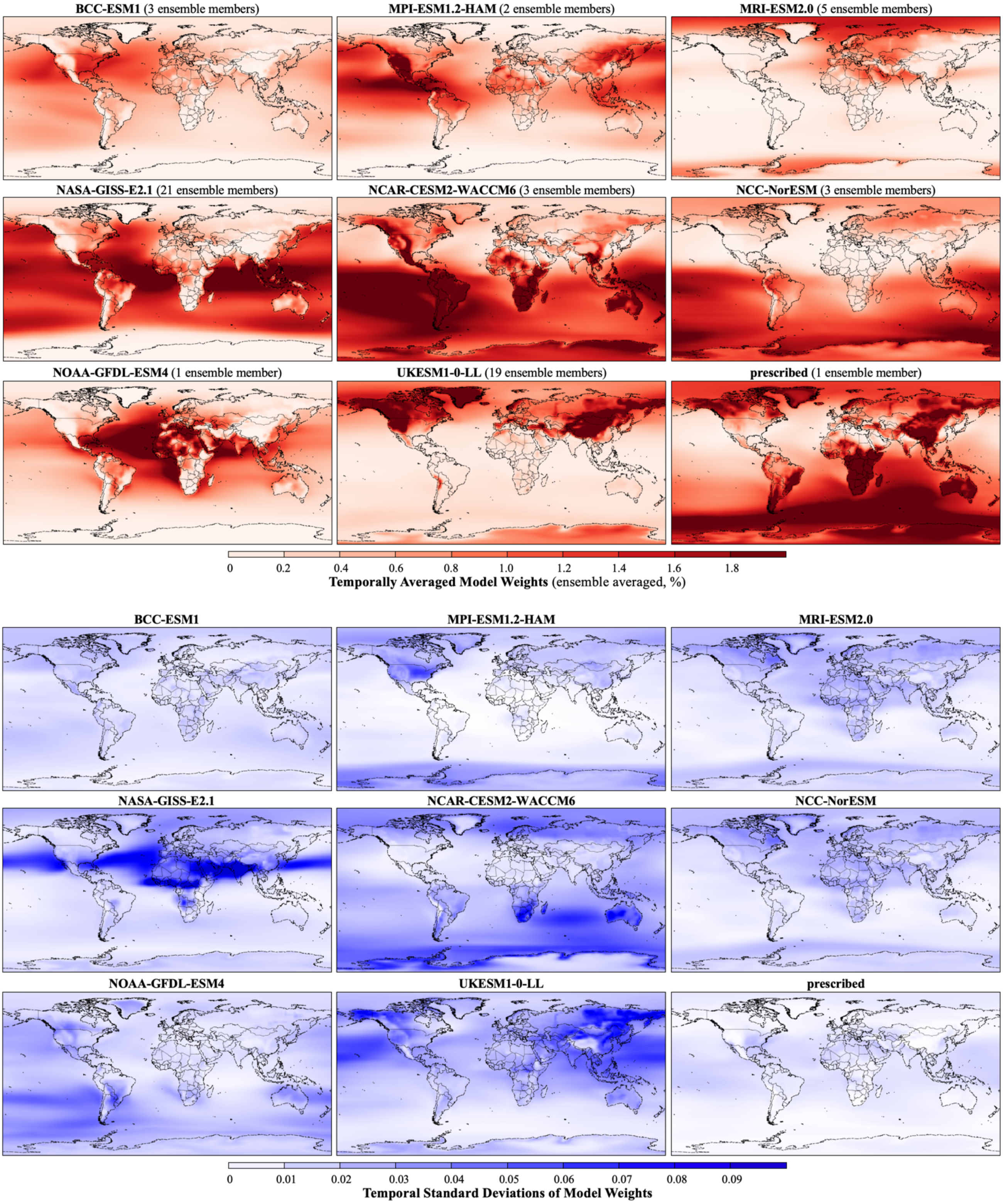

Figure S10 Temporal averaged global distribution of individual model weights by conservative space-time Bayesian neural network regression-based integration. The arithmetically averaged weights and standard deviations for linear combination of the 58 involved CMIP6 model simulation ensembles are calculated across 1990-2014, with darker reds indicating higher contribution weights and darker blue representing higher temporal variability. The spatial weights are also summarised as multi-realisation ensemble average for the concision of summarisation.

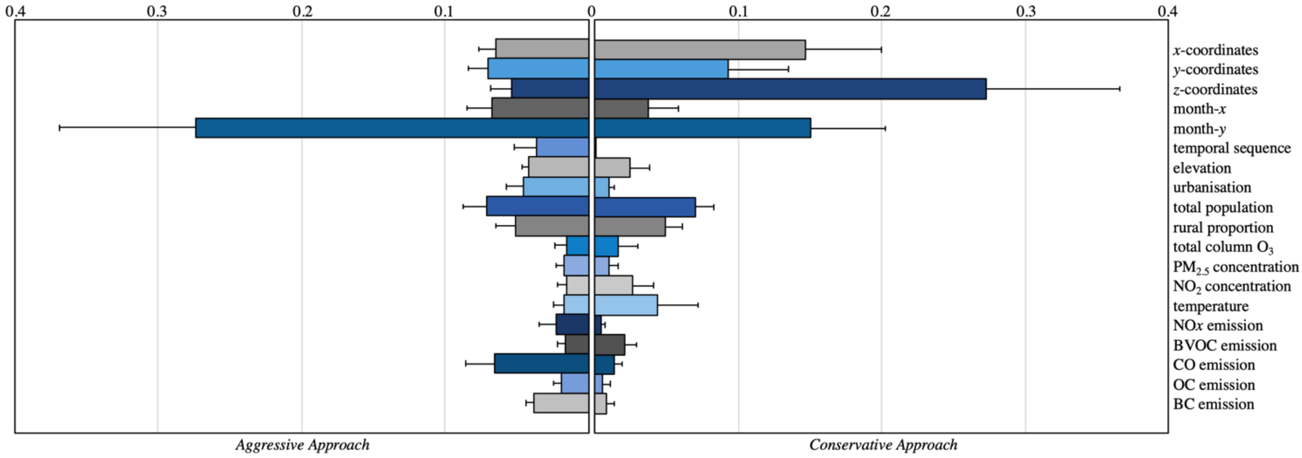

Figure S11 Contribution weights of additional assistant features for ensemble deep learning-based CMIP6 model prediction enhancement by aggressive and conservative multi-model integration approach. The weights are normalised by excluding the CMIP6 individual model contributions and shown in histograms with error bars representing the variations across different algorithms for aggressive approach and the spatiotemporal variability for conservative approach.

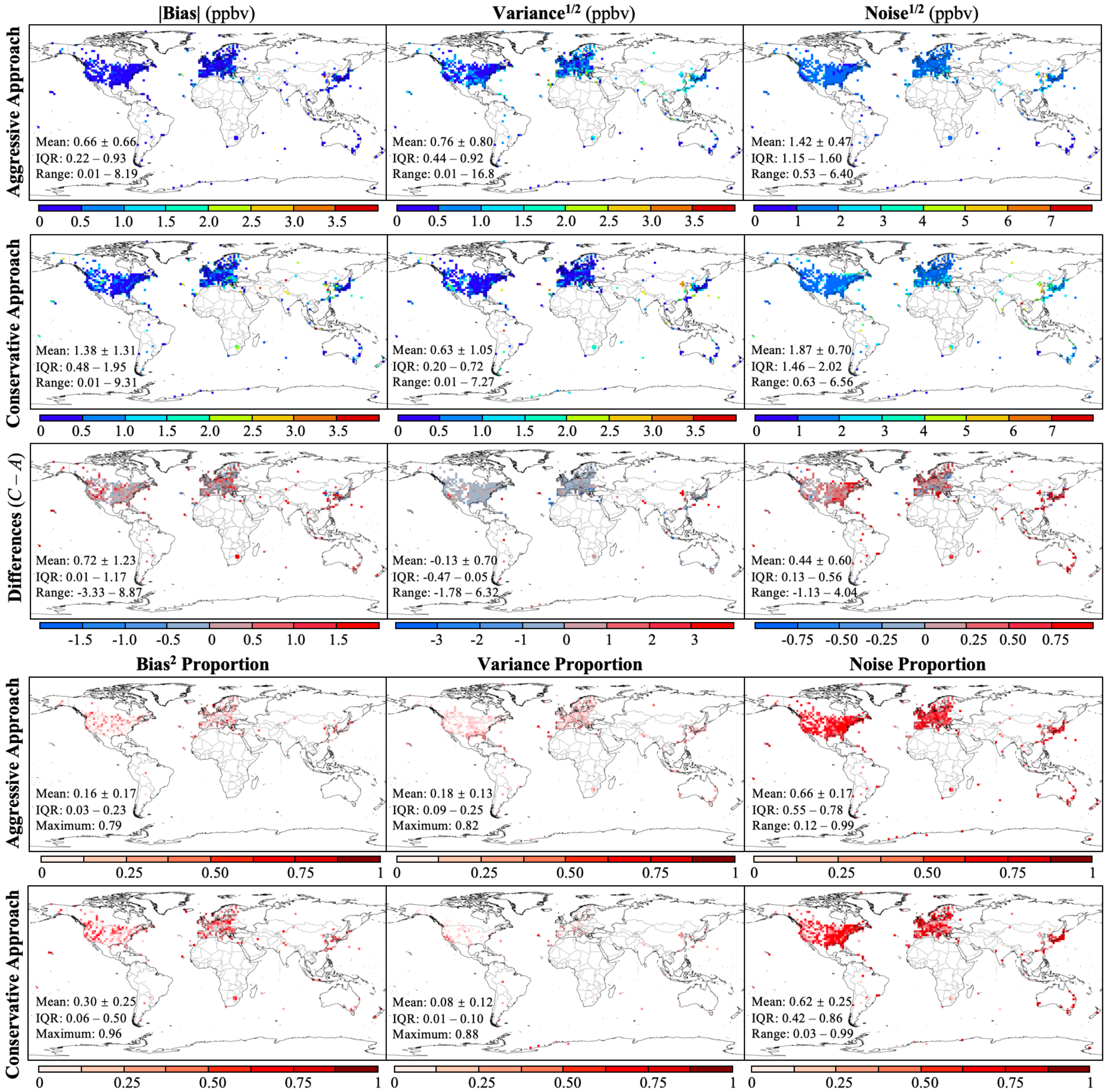

Figure S12 Decomposition of model-observation errors into bias, variance and noise in aggressive and conservative multi-model ensemble for surface ozone simulation. By definition, MSE = bias^2 + model variance + noise. Absolute values of bias, root variance and root noise for both aggressive and conservative approaches together with the differences, and the proportions of the squared bias, variance and noise are mapped individually. Major statistics include arithmetic mean, standard deviation, full range and inter-quartile range (IQR) are summarised for each metric. For the proportions of bias^2 and variance, the ranges are replaced by maximums, as the minimums are extremely small.

	prescribed	BCC-ESM1	UKESM1-0-LL	MPI-ESM1.2-HAM	MRI-ESM2.0	NASA-GISS-E2.1	NCAR-CESM2-WACCM6	NCC-NorESM	NOAA-GFDL-ESM4
x-coordinates	6.14	7.47	2.20	8.74	7.39	1.95	9.10	7.20	6.14
y-coordinates	-5.02	-2.30	-5.62	-4.72	-5.06	-5.39	-0.27	-6.16	-1.30
z-coordinates	-12.2	-7.96	7.27	-8.07	-6.58	-10.3	-6.31	-19.0	-7.41
month-x	-1.62	-0.19	-2.36	-1.67	-0.85	-2.48	-1.84	-0.36	-1.81
month-y	-7.06	-5.70	-8.60	-8.62	-7.86	-1.85	-7.24	-10.3	-7.78
temporal sequence	<0.01	<0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.02
elevation	0.54	0.54	2.17	1.43	1.20	0.03	1.21	1.23	0.22
urbanisation	0.81	0.51	0.31	0.83	-0.06	0.15	0.57	1.23	0.35
total population	1.16	0.70	1.33	1.02	0.95	-2.08	1.50	-0.17	1.15
rural proportion	1.77	0.61	1.73	1.76	0.12	3.75	2.24	2.57	2.55
total column O_3	-1.25	-0.47	-1.85	-1.70	-1.66	0.48	-0.50	-0.83	-0.25
$PM_{2.5}$ concentration	-0.50	-0.26	-0.33	-0.25	-0.75	-0.43	-0.35	-0.23	-0.52
NO_2 concentration	-2.70	-2.42	-2.13	-2.28	-3.50	-2.93	-2.73	-4.18	-3.25
temperature	-1.80	-1.71	-1.39	-1.48	-2.29	-2.15	-2.17	-2.96	-2.22
NOx emission	0.12	0.20	0.04	-0.05	0.14	0.17	-0.01	0.26	0.10
BVOC emission	1.15	1.01	0.99	1.29	0.24	-0.26	0.83	1.33	0.97
CO emission	0.64	0.29	0.47	0.58	0.81	0.80	0.33	0.45	0.76
OC emission	0.07	-0.29	0.54	-0.02	0.51	0.73	-0.23	-0.31	0.34
BC emission	-0.32	0.04	0.48	-0.27	0.15	0.57	-0.26	0.06	-0.32

k

	prescribed	BCC-ESM1	UKESM1-0-LL	MPI-ESM1.2-HAM	MRI-ESM2.0	NASA-GISS-E2.1	NCAR-CESM2-WACCM6	NCC-NorESM	NOAA-GFDL-ESM4
x-coordinates	-3.22	-3.87	-2.19	-5.15	-4.03	-1.59	-4.06	-3.78	-2.64
y-coordinates	2.93	2.55	2.11	3.10	2.54	2.18	1.13	4.76	1.47
z-coordinates	3.41	2.76	0.09	3.25	1.99	2.15	1.11	3.76	1.19
month-x	-1.47	-2.46	-2.28	-1.07	-1.12	-1.92	-1.09	-0.90	-1.31
month-y	6.44	6.32	5.39	7.05	6.54	6.09	6.60	8.09	6.20
temporal sequence	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
elevation	-1.52	-1.46	-1.83	-1.59	-1.62	-1.35	-1.52	-1.60	-1.55
urbanisation	0.15	0.10	<0.01	0.21	0.32	0.26	0.33	0.33	0.17
total population	-0.15	-0.13	-0.26	-0.44	-0.35	-0.06	-0.36	-0.20	-0.19
rural proportion	0.07	0.08	-0.09	0.06	0.06	0.09	0.17	<0.01	0.10
total column O_3	-0.04	-0.09	-0.16	-0.02	-0.07	0.03	-0.02	-0.19	-0.02
$PM_{2.5}$ concentration	0.77	0.84	0.88	0.79	0.82	0.80	0.71	0.89	0.76
NO_2 concentration	0.32	0.76	0.64	0.38	0.55	0.69	0.55	0.34	0.50
temperature	-0.71	-0.90	-1.32	-0.77	-0.86	-0.90	-0.47	-0.81	-0.63
NOx emission	-0.13	-0.20	-0.21	-0.14	-0.18	-0.32	-0.15	-0.16	-0.18
BVOC emission	-0.33	-0.52	-0.67	-0.48	-0.27	-0.47	-0.42	-0.37	-0.41
CO emission	-0.56	-0.71	-0.51	-0.53	-0.70	-0.81	-0.55	-0.49	-0.62
OC emission	-0.69	-0.76	-0.67	-0.73	-0.75	-0.74	-0.69	-0.88	-0.58
BC emission	-0.17	-0.21	-0.47	-0.21	-0.30	-0.16	-0.14	-0.12	-0.24

b

Figure S13 Correlation strength between the assistant features and single model calibration parameters by Bayesian neural network regression. The regressions are conducted on the standardised variables, with the left panel referring to the re-scaling parameter (slope k) and the right referring to the bias-correction parameter (intercept b). The values represent the regression coefficients, among which 10^{-2} is scaled for k. The positive correlations are shown in red while the negative correlations are in blue, with darker colours representing stronger associations. Grey cells indicate insignificant correlations.

Table S1 Summarisation of CMIP6 historical project participant institutes and models with chemistry schemes, spatial gridding, and experiment realisation, physics, and forcing scenarios. The names of institutes and coupled earth system models are listed in abbreviation. The three-dimensional spatial resolutions are represented in longitudinal-latitudinal-vertical grids. The tropospheric and stratospheric chemistry schemes are denoted as interactive (I), prescribed (P) and none (N) in "Trop" and "Strat" columns. The realisation, physics and forcing indices identify ensemble experiment members. The "Fusion" column indicates whether the simulation experiments are included into multi-model fusion. Full names of the CMIP6 participant research institutes and detailed information of coupling components of earth system models are listed in Supplementary Information.

Institute	Model	Trop	Strat	Grids	Realisations	Physics	Forcing	Fusion	Refs
AWI	**ESM**‖	P#	P	192×96×47	$r_1^{\perp}$	p_1	f_1		21
BCC	**ESM1**	I	P	128×64×26	r_1, r_2, r_3	p_1	f_1	√	22, 23
	CSM2	P	P	320×160×19	r_1	p_1	f_1		24-26
CNRM*	**CM6.1**	N	I	256×128×91	r_{1-5}	p_1	f_2		27-29
	ESM2.1	N	I	256×128×91	r_1, r_2, r_3	p_1	f_2		28, 30, 31
HAMMOZ§	**MPI-ESM1.2-HAM**	I	P	192×96×47	r_1, r_2	p_1	f_1	√	32
IPSL	**CM6A**	P	P	144×143×79	r_{1-10}	p_1	f_1		33, 34
MOHC	**UKESM1-0-LL†**	I	I	192×144×85	r_{10-12}, r_{14-19}	p_1	f_2	√	
	UKESM1-0-LL	I	I	192×144×85	r_{5-7}	p_1	f_3	√	15, 35-39
MO-NERC	**UKESM1-0-LL**	I	I	192×144×85	r_{1-4}, r_{8-9}	p_1	f_2	√	
MPI-M	**ESM1.2-HR**	P	P	384×192×95	r_{1-10}	p_1	f_1		40-43
MRI	**ESM2.0**	I	I	128×64×80	r_{1-5}	p_1	f_1	√	44-46
NASA-GISS	**E2.1-G**	I	I	144×90×40	r_{1-10}	p_3	f_1	√	
	E2.1-G	I	I	144×90×40	r_1, r_2, r_3	p_5	f_1	√	47-49
	E2.1-H	I	I	144×90×40	r_{1-5}	p_3	f_1	√	
	E2.1-H	I	I	144×90×40	r_1, r_2, r_3	p_5	f_1	√	
NCAR	**CESM2-WACCM6**	I	I	288×192×70	r_1, r_2, r_3	p_1	f_1	√	50, 51
NCC	**NorESM-MM‡**	I	P	288×192×32	r_1, r_2, r_3	p_1	f_1	√	52
NIMS-KMA	**UKESM1-0-LL**	I	I	192×144×85	r_{13}	p_1	f_2	√	53
NOAA-GFDL	**ESM4**	I	I	288×180×49	r_1	p_1	f_1	√	54, 55

‖ The earth system models are unique for each institute, but coincidently are named the same as ESM with version numbers, thus are named by institute + model name hereafter in this paper for distinguishment (i.e. CNRM-ESM2.1 is not an updated version of BCC-ESM1, but a new version of CNRM-ESM1)[56].

AWI-ESM, BCC-CSM2, IPSL-CM6A, and MPI-M-ESM1.2-HR use the same prescribed ozone for the whole earth system modelling instead of simulating the ozone, so that the surface ozone concentrations reported by these 4 models are essentially the same. In this sense, the single prescribed ozone (input4MIPs)[57] is used in place of the 4 models to avoid duplication.

⊥ All the realisations of the climate equilibrium started since 1850, so that are marked with the same initialisation index, i_1. The ensemble experiment variant serial numbers are defined by a combination of realisation, initialisation, physics, and forcing, e.g. $r_1 i_1 p_1 f_1$.

* The 2 CNRM models are not considered for surface ozone multi-model fusion as they do not include tropospheric ozone module.

§ Full name as HAMMOZ-Consortium, marked as HAM in model name.

† MOHC, MO-NERC and NIMS-KMA ran the same UKESM1 model with same configuration, but contributed different ensemble experiments, so that are referred collectively as UKESM1-0-LL hereafter in this paper.

‡ NCC ran the NorESM in two different coupling resolutions, as low atmospheric-medium ocean resolution (LM) and median atmospheric-medium ocean resolution (MM). In order to achieve higher performance in multi-model fusion, only the higher spatial-resolution simulation, MM, is considered so as to avoid duplication.

Table S2 Diagnostic features of model fitting regression for the seasonal oscillations of TOAR monthly average surface ozone concentrations. A total of 25-year 300-month observations were included in the TOAR project. The mathematical formulae were given together with abbreviated models names as index. The coefficient of determinacy (R^2), rooted mean squared error (RMSE), and Akaike Information Criteria (AIC) were calculated as model selection criteria. The temporal term t was set as the serial month since January 1990.

Model	Equation	R^2	RMSE	k†	AIC
F1-L-L‡	$\hat{y}(t) = (a_0 + a_1 t) + (b_0 + b_1 t) \times sin(\frac{\pi}{6}t + \varphi_1)$	0.9103	1.55	5	266.79
F1-L-E	$\hat{y}(t) = (a_0 + a_1 t) + (b_0 e^{b_1 t}) \times sin(\frac{\pi}{6}t + \varphi_1)$	0.9101	1.55	5	267.56
F1-E-L	$\hat{y}(t) = (a_0 e^{a_1 t}) + (b_0 + b_1 t) \times sin(\frac{\pi}{6}t + \varphi_1)$	0.9101	1.55	5	267.39
F1-E-E	$\hat{y}(t) = (a_0 e^{a_1 t}) + (b_0 e^{b_1 t}) \times sin(\frac{\pi}{6}t + \varphi_1)$	0.9099	1.55	5	268.11
F2-L-L-C	$\hat{y}(t) = (a_0 + a_1 t) + (b_0 + b_1 t) \times sin(\frac{\pi}{6}t + \varphi_1) + c_0 \times sin(\frac{2\pi}{6}t + \varphi_2)$	0.9563	1.08	7	55.41
F2-L-L-L	$\hat{y}(t) = (a_0 + a_1 t) + (b_0 + b_1 t) \times sin(\frac{\pi}{6}t + \varphi_1) + (c_0 + c_1 t) \times sin(\frac{2\pi}{6}t + \varphi_2)$	0.9565	1.08	8	55.75
F2-L-L-E	$\hat{y}(t) = (a_0 + a_1 t) + (b_0 + b_1 t) \times sin(\frac{\pi}{6}t + \varphi_1) + (c_0 e^{c_1 t}) \times sin(\frac{2\pi}{6}t + \varphi_2)$	0.9565	1.08	8	55.66
F2-L-E-C	$\hat{y}(t) = (a_0 + a_1 t) + (b_0 e^{b_1 t}) \times sin(\frac{\pi}{6}t + \varphi_1) + c_0 \times sin(\frac{2\pi}{6}t + \varphi_2)$	0.9546	1.11	7	66.77
F2-L-E-L	$\hat{y}(t) = (a_0 + a_1 t) + (b_0 e^{b_1 t}) \times sin(\frac{\pi}{6}t + \varphi_1) + (c_0 + c_1 t) \times sin(\frac{2\pi}{6}t + \varphi_2)$	0.9547	1.11	8	68.19
F2-L-E-E	$\hat{y}(t) = (a_0 + a_1 t) + (b_0 e^{b_1 t}) \times sin(\frac{\pi}{6}t + \varphi_1) + (c_0 e^{c_1 t}) \times sin(\frac{2\pi}{6}t + \varphi_2)$	0.9549	1.10	8	66.75
F2-E-L-C	$\hat{y}(t) = (a_0 e^{a_1 t}) + (b_0 + b_1 t) \times sin(\frac{\pi}{6}t + \varphi_1) + c_0 \times sin(\frac{2\pi}{6}t + \varphi_2)$	0.9564	1.08	7	54.18
F2-E-L-L	$\hat{y}(t) = (a_0 e^{a_1 t}) + (b_0 + b_1 t) \times sin(\frac{\pi}{6}t + \varphi_1) + (c_0 + c_1 t) \times sin(\frac{2\pi}{6}t + \varphi_2)$	0.9562	1.09	8	57.84
F2-E-L-E	$\hat{y}(t) = (a_0 e^{a_1 t}) + (b_0 + b_1 t) \times sin(\frac{\pi}{6}t + \varphi_1) + (c_0 e^{c_1 t}) \times sin(\frac{2\pi}{6}t + \varphi_2)$	0.9562	1.09	8	57.49
F2-E-E-C	$\hat{y}(t) = (a_0 e^{a_1 t}) + (b_0 e^{b_1 t}) \times sin(\frac{\pi}{6}t + \varphi_1) + c_0 \times sin(\frac{2\pi}{6}t + \varphi_2)$	0.9559	1.09	7	57.58
F2-E-E-L	$\hat{y}(t) = (a_0 e^{a_1 t}) + (b_0 e^{b_1 t}) \times sin(\frac{\pi}{6}t + \varphi_1) + (c_0 + c_1 t) \times sin(\frac{2\pi}{6}t + \varphi_2)$	0.9561	1.09	8	58.36
F2-E-E-E	$\hat{y}(t) = (a_0 e^{a_1 t}) + (b_0 e^{b_1 t}) \times sin(\frac{\pi}{6}t + \varphi_1) + (c_0 e^{c_1 t}) \times sin(\frac{2\pi}{6}t + \varphi_2)$	0.9561	1.09	8	58.19

† k represents the degree of freedom, which is equal to the number of parameters.

‡ The model abbreviations were defined by the order(s) of Fourier Series, and the sub-types of each term, sequenced as the intercept, 1st- and 2nd- order series. F1, 1st-order Fourier Series models. F2, 2nd-order Fourier Series models. L, linear coefficient. E, exponential coefficient. C, constant coefficient.

Table S3 Model-observation error decomposition of aggressive and conservative multi-model ensemble integration approaches for each continent. The metrics include the values and ratios of bias, root variance and root noise, summarised with arithmetic mean, median and inter-quartile range (IQR, 25-75%ile). The continental statistics are summarised based on the TOAR realistic measurement-covered sites.

			Europe		North America		South America		Asia		Africa		Oceania	
			mean (median)	*IQR*	*mean (median)*	*IQR*	*mean (median)*	*IQR*	*mean (median)*	*IQR*	*mean (median)*	*IQR*	*mean (median)*	*IQR*
Aggressive	**Value**	*Bias*	0.78 (0.63)	0.36 – 1.05	0.58 (0.49)	0.23 – 0.79	0.38 (0.17)	0.13 – 0.23	0.93 (0.47)	0.31 – 1.24	0.83 (0.63)	0.22 – 0.77	0.17 (0.17)	0.06 – 0.22
		Variance$^{1/2}$	0.80 (0.75)	0.59 – 0.92	0.57 (0.53)	0.35 – 0.74	0.73 (0.91)	0.42 – 0.95	1.02 (0.93)	0.71 – 1.27	1.27 (1.14)	1.07 – 1.43	0.46 (0.41)	0.33 – 0.56
		Noise$^{1/2}$	1.50 (1.37)	1.22 – 1.66	1.33 (1.29)	1.14 – 1.50	1.02 (1.00)	0.80 – 1.12	1.69 (1.62)	1.19 – 1.97	1.55 (1.56)	1.48 – 1.58	0.77 (0.77)	0.68 – 0.93
	Ratio	*Bias*	0.18 (0.14)	0.04 – 0.27	0.16 (0.09)	0.03 – 0.24	0.09 (0.02)	0.01 – 0.07	0.18 (0.11)	0.03 – 0.25	0.14 (0.08)	0.02 – 0.11	0.05 (0.04)	0.01 – 0.09
		Variance$^{1/2}$	0.19 (0.17)	0.12 – 0.25	0.14 (0.12)	0.06 – 0.20	0.30 (0.33)	0.17 – 0.39	0.24 (0.24)	0.09 – 0.31	0.31 (0.35)	0.22 – 0.37	0.24 (0.23)	0.15 – 0.32
		Noise$^{1/2}$	0.63 (0.66)	0.52 – 0.74	0.70 (0.71)	0.61 – 0.84	0.61 (0.59)	0.50 – 0.77	0.58 (0.58)	0.49 – 0.69	0.56 (0.63)	0.49 – 0.70	0.71 (0.75)	0.59 – 0.81
Conservative	**Value**	*Bias*	1.31 (1.22)	0.54 – 1.87	1.05 (0.88)	0.33 – 1.48	1.15 (0.60)	0.42 – 2.08	2.83 (1.96)	0.81 – 4.30	4.56 (4.56)	4.28 – 4.70	1.30 (0.76)	0.32 – 2.24
		Variance$^{1/2}$	0.51 (0.43)	0.25 – 0.66	0.43 (0.29)	0.15 – 0.64	1.39 (0.56)	0.25 – 1.55	1.30 (0.72)	0.31 – 1.61	1.27 (1.48)	1.28 – 1.57	0.24 (0.14)	0.02 – 0.53
		Noise$^{1/2}$	1.74 (1.58)	1.43 – 1.95	1.67 (1.61)	1.41 – 1.84	2.56 (2.62)	1.47 – 3.01	2.96 (2.41)	2.05 – 3.84	3.02 (2.43)	2.35 – 3.71	1.47 (1.48)	1.07 – 1.79
	Ratio	*Bias*	0.33 (0.29)	0.10 – 0.52	0.27 (0.20)	0.04 – 0.47	0.17 (0.10)	0.03 – 0.33	0.34 (0.24)	0.09 – 0.55	0.65 (0.68)	0.64 – 0.72	0.35 (0.26)	0.03 – 0.59
		Variance$^{1/2}$	0.07 (0.04)	0.01 – 0.08	0.06 (0.02)	0.01 – 0.08	0.17 (0.03)	0.02 – 0.31	0.10 (0.04)	0.01 – 0.10	0.06 (0.09)	0.04 – 0.09	0.03 (0.01)	0.01 – 0.04
		Noise$^{1/2}$	0.61 (0.59)	0.42 – 0.80	0.67 (0.71)	0.48 – 0.89	0.67 (0.84)	0.23 – 0.95	0.56 (0.56)	0.33 – 0.87	0.28 (0.26)	0.20 – 0.32	0.62 (0.70)	0.41 – 0.89

REFERENCES

1. Stouffer, R. J.; Weaver, A. J.; Eby, M., A method for obtaining pre-twentieth century initial conditions for use in climate change studies. *Climate Dynamics* **2004,** *23*, (3-4), 327-339.
2. Eyring, V.; Bony, S.; Meehl, G. A.; Senior, C. A.; Stevens, B.; Stouffer, R. J.; Taylor, K. E., Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geosci Model Dev* **2016,** *9*, (5), 1937-1958.
3. Dantas-Torres, F., Climate change, biodiversity, ticks and tick-borne diseases: The butterfly effect. *Int J Parasitol Parasites Wildl* **2015,** *4*, (3), 452-61.
4. Power, S.; Casey, T.; Folland, C.; Colman, A.; Mehta, V., Inter-decadal modulation of the impact of ENSO on Australia. *Climate Dynamics* **1999,** *15*, (5), 319-324.
5. Zhang, Y.; Wallace, J. M.; Battisti, D. S., ENSO-like interdecadal variability: 1900-1993. *Journal of Climate* **1997,** *10*, (5), 1004-1020.
6. Folland, C. K.; Renwick, J. A.; Salinger, M. J.; Mullan, A. B., Relative influences of the Interdecadal Pacific Oscillation and ENSO on the South Pacific Convergence Zone. *Geophysical Research Letters* **2002,** *29*, (13), 21-1-21-4.
7. Meehl, G. A.; Hu, A. X.; Arblaster, J. M.; Fasullo, J.; Trenberth, K. E., Externally Forced and Internally Generated Decadal Climate Variability Associated with the Interdecadal Pacific Oscillation. *Journal of Climate* **2013,** *26*, (18), 7298-7310.
8. Salinger, M. J.; Renwick, J. A.; Mullan, A. B., Interdecadal Pacific Oscillation and South Pacific climate. *International Journal of Climatology* **2001,** *21*, (14), 1705-1721.
9. Kerr, R. A., A north atlantic climate pacemaker for the centuries. *Science* **2000,** *288*, (5473), 1984-5.
10. Knight, J. R.; Folland, C. K.; Scaife, A. A., Climate impacts of the Atlantic Multidecadal Oscillation. *Geophysical Research Letters* **2006,** *33*, (17).
11. Knudsen, M. F.; Seidenkrantz, M. S.; Jacobsen, B. H.; Kuijpers, A., Tracking the Atlantic Multidecadal Oscillation through the last 8,000 years. *Nature Communications* **2011,** *2*, (1), 1-8.
12. Zhang, R.; Delworth, T. L., Impact of the Atlantic Multidecadal Oscillation on North Pacific climate variability. *Geophysical Research Letters* **2007,** *34*, (23).
13. Goes, J. I.; Saino, T.; Oaku, H.; Ishizaka, J.; Wong, C. S.; Nojiri, Y., Basin scale estimates of Sea Surface Nitrate and New Production from remotely sensed Sea Surface Temperature and Chlorophyll. *Geophysical Research Letters* **2000,** *27*, (9), 1263-1266.
14. Kushnir, Y., Interdecadal Variations in North-Atlantic Sea-Surface Temperature and Associated Atmospheric Conditions. *Journal of Climate* **1994,** *7*, (1), 141-157.
15. Sellar, A. A.; Walton, J.; Jones, C. G.; Wood, R.; Abraham, N. L.; Andrejczuk, M.; Andrews, M. B.; Andrews, T.; Archibald, A. T.; Mora, L.; Dyson, H.; Elkington, M.; Ellis, R.; Florek, P.; Good, P.; Gohar, L.; Haddad, S.; Hardiman, S. C.; Hogan, E.; Iwi, A.; Jones, C. D.; Johnson, B.; Kelley, D. I.; Kettleborough, J.; Knight, J. R.; Köhler, M. O.; Kuhlbrodt, T.; Liddicoat, S.; Linova-Pavlova, I.; Mizielinski, M. S.; Morgenstern, O.; Mulcahy, J.; Neininger, E.; O'Connor, F. M.; Petrie, R.; Ridley, J.; Rioual, J. C.; Roberts, M.; Robertson, E.; Rumbold, S.; Seddon, J.; Shepherd, H.; Shim, S.; Stephens, A.; Teixiera, J. C.; Tang, Y.; Williams, J.; Wiltshire, A.; Griffiths, P. T., Implementation of U.K. Earth System Models for CMIP6. *J Adv Model Earth Syst* **2020,** *12*, (4), e2019MS001946.
16. Bauer, S. E.; Tsigaridis, K.; Faluvegi, G.; Kelley, M.; Lo, K. K.; Miller, R. L.; Nazarenko, L.; Schmidt, G. A.; Wu, J., Historical (1850-2014) aerosol evolution and role on climate forcing using the GISS ModelE2.1 contribution to CMIP6. *J Adv Model Earth Syst* **2020,** *12*, (8), e2019MS001978.
17. Koch, D.; Jacob, D.; Tegen, I.; Rind, D.; Chin, M., Tropospheric sulfur simulation and sulfate direct radiative forcing in the Goddard Institute for Space Studies general circulation model. *J Geophys Res-Atmos* **1999,** *104*, (D19), 23799-23822.
18. Bauer, S. E.; Wright, D. L.; Koch, D.; Lewis, E. R.; McGraw, R.; Chang, L. S.; Schwartz, S. E.; Ruedy, R., MATRIX (Multiconfiguration Aerosol TRacker of mIXing state): an aerosol microphysical module for global atmospheric models. *Atmos Chem Phys* **2008,** *8*, (20), 6003-6035.
19. Lee, Y. H.; Adams, P. J., A Fast and Efficient Version of the TwO-Moment Aerosol Sectional (TOMAS) Global Aerosol Microphysics Model. *Aerosol Science and Technology* **2012,** *46*, (6), 678-689.
20. Liu, X. H.; Penner, J. E.; Herzog, M., Global modeling of aerosol dynamics: Model description, evaluation, and interactions between sulfate and nonsulfate aerosols. *J Geophys Res-Atmos* **2005,** *110*, (D18), 206-242.
21. Danek, C.; Shi, X.; Stepanek, C.; Yang, H.; Barbi, D.; Hegewald, J.; Lohmann, G., AWI AWI-ESM1.1LR model output prepared for CMIP6 CMIP historical. In Earth System Grid Federation: 2020.
22. Wu, T. W.; Zhang, F.; Zhang, J.; Jie, W. H.; Zhang, Y. W.; Wu, F. H.; Li, L.; Yan, J. H.; Liu, X. H.; Lu, X.; Tan, H. Y.; Zhang, L.; Wang, J.; Hu, A. X., Beijing Climate Center Earth System Model version 1 (BCC-ESM1): model description and evaluation of aerosol simulations. *Geosci Model Dev* **2020,** *13*, (3), 977-1005.
23. Zhang, J.; Wu, T.; Shi, X.; Zhang, F.; Li, J.; Chu, M.; Liu, Q.; Yan, J.; Ma, Q.; Wei, M., BCC BCC-ESM1 model output prepared for CMIP6 CMIP *piControl*. In Earth System Grid Federation: 2018.

24. Wu, T. W.; Lu, Y. X.; Fang, Y. J.; Xin, X. G.; Li, L.; Li, W. P.; Jie, W. H.; Zhang, J.; Liu, Y. M.; Zhang, L.; Zhang, F.; Zhang, Y. W.; Wu, F. H.; Li, J. L.; Chu, M.; Wang, Z. Z.; Shi, X. L.; Liu, X. W.; Wei, M.; Huang, A. N.; Zhang, Y. C.; Liu, X. H., The Beijing Climate Center Climate System Model (BCC-CSM): the main progress from CMIP5 to CMIP6. *Geosci Model Dev* **2019,** *12*, (4), 1573-1600.
25. Wu, T.; Yu, R.; Lu, Y.; Jie, W.; Fang, Y.; Zhang, J.; Zhang, L.; Xin, X.; Li, L.; Wang, Z., BCC-CSM2-HR: A High-Resolution Version of the Beijing Climate Center Climate System Model. *Geosci Model Dev* **2021**, *14*, (5), 2977-3006.
26. Wu, T.; Chu, M.; Dong, M.; Fang, Y.; Jie, W.; Li, J.; Li, W.; Liu, Q.; Shi, X.; Xin, X.; Yan, J.; Zhang, F.; Zhang, J.; Zhang, L.; Zhang, Y., BCC BCC-CSM2MR model output prepared for CMIP6 CMIP *piControl*. In Earth System Grid Federation: 2018.
27. Voldoire, A.; Saint-Martin, D.; Sénési, S.; Decharme, B.; Alias, A.; Chevallier, M.; Colin, J.; Guérémy, J. F.; Michou, M.; Moine, M. P., Evaluation of CMIP6 deck experiments with CNRM-CM6-1. *J Adv Model Earth Syst* **2019,** *11*, (7), 2177-2213.
28. Michou, M.; Nabat, P.; Saint-Martin, D.; Bock, J.; Decharme, B.; Mallet, M.; Roehrig, R.; Séférian, R.; Sénési, S.; Voldoire, A., Present-day and historical aerosol and ozone characteristics in CNRM CMIP6 simulations. *J Adv Model Earth Syst* **2020,** *12*, (1), e2019MS001816.
29. Voldoire, A., CNRM-CERFACS CNRM-CM6-1 model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2018.
30. Séférian, R.; Nabat, P.; Michou, M.; Saint-Martin, D.; Voldoire, A.; Colin, J.; Decharme, B.; Delire, C.; Berthet, S.; Chevallier, M., Evaluation of CNRM Earth System Model, CNRM-ESM2-1: Role of Earth System Processes in Present-Day and Future Climate. *J Adv Model Earth Syst* **2019,** *11*, (12), 4182-4227.
31. Seferian, R., CNRM-CERFACS CNRM-ESM2-1 model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2018.
32. Neubauer, D.; Ferrachat, S.; Siegenthaler-Le Drian, C.; Stoll, J.; Folini, D. S.; Tegen, I.; Wieners, K.-H.; Mauritsen, T.; Stemmler, I.; Barthel, S.; Bey, I.; Daskalakis, N.; Heinold, B.; Kokkola, H.; Partridge, D.; Rast, S.; Schmidt, H.; Schutgens, N.; Stanelle, T.; Stier, P.; Watson-Parris, D.; Lohmann, U., HAMMOZ-Consortium MPI-ESM1.2-HAM model output prepared for CMIP6 AerChemMIP. In Earth System Grid Federation: 2019.
33. Boucher, O.; Servonnat, J.; Albright, A. L.; Aumont, O.; Balkanski, Y.; Bastrikov, V.; Bekki, S.; Bonnet, R.; Bony, S.; Bopp, L., Presentation and evaluation of the IPSL-CM6A-LR climate model. *J Adv Model Earth Syst* **2020,** *12*, (7), e2019MS002010.
34. Boucher, O.; Denvil, S.; Levavasseur, G.; Cozic, A.; Caubel, A.; Foujols, M.-A.; Meurdesoif, Y.; Cadule, P.; Devilliers, M.; Ghattas, J.; Lebas, N.; Lurton, T.; Mellul, L.; Musat, I.; Mignot, J.; Cheruy, F., IPSL IPSL-CM6A-LR model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2018.
35. Archibald, A. T.; O'Connor, F. M.; Abraham, N. L.; Archer-Nicholls, S.; Chipperfield, M. P.; Dalvi, M.; Folberth, G. A.; Dennison, F.; Dhomse, S. S.; Griffiths, P. T.; Hardacre, C.; Hewitt, A. J.; Hill, R.; Johnson, C. E.; Keeble, J.; Köhler, M. O.; Morgenstern, O.; Mulchay, J. P.; Ordóñez, C.; Pope, R. J.; Rumbold, S.; Russo, M. R.; Savage, N.; Sellar, A.; Stringer, M.; Turnock, S.; Wild, O.; Zeng, G., Description and evaluation of the UKCA stratosphere-troposphere chemistry scheme (StratTrop vn 1.0) implemented in UKESM1. *Geosci Model Dev* **2020,** *13*, (3), 1223-1266.
36. Mulcahy, J. P.; Jones, C.; Sellar, A.; Johnson, B.; Boutle, I. A.; Jones, A.; Andrews, T.; Rumbold, S. T.; Mollard, J.; Bellouin, N.; Johnson, C. E.; Williams, K. D.; Grosvenor, D. P.; McCoy, D. T., Improved Aerosol Processes and Effective Radiative Forcing in HadGEM3 and UKESM1. *J Adv Model Earth Syst* **2018,** *10*, (11), 2786-2805.
37. Sellar, A. A.; Jones, C. G.; Mulcahy, J. P.; Tang, Y.; Yool, A.; Wiltshire, A.; O'connor, F. M.; Stringer, M.; Hill, R.; Palmieri, J., UKESM1: Description and evaluation of the UK Earth System Model. *J Adv Model Earth Syst* **2019,** *11*, (12), 4513-4558.
38. Tang, Y.; Rumbold, S.; Ellis, R.; Kelley, D.; Mulcahy, J.; Sellar, A.; Walton, J.; Jones, C., MOHC UKESM1.0-LL model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2019.
39. Yool, A.; Palmiéri, J.; Jones, C.; Sellar, A.; de Mora, L.; Kuhlbrodt, T.; Popova, E.; Mulcahy, J.; Wiltshire, A.; Rumbold, S. T., Spin-up of UK Earth System Model 1 (UKESM1) for CMIP6. *J Adv Model Earth Syst* **2020**, e2019MS001933.
40. Mauritsen, T.; Bader, J.; Becker, T.; Behrens, J.; Bittner, M.; Brokopf, R.; Brovkin, V.; Claussen, M.; Crueger, T.; Esch, M.; Fast, I.; Fiedler, S.; Flaschner, D.; Gayler, V.; Giorgetta, M.; Goll, D. S.; Haak, H.; Hagemann, S.; Hedemann, C.; Hohenegger, C.; Ilyina, T.; Jahns, T.; Jimenez-de-la-Cuesta, D.; Jungclaus, J.; Kleinen, T.; Kloster, S.; Kracher, D.; Kinne, S.; Kleberg, D.; Lasslop, G.; Kornblueh, L.; Marotzke, J.; Matei, D.; Meraner, K.; Mikolajewicz, U.; Modali, K.; Mobis, B.; Muller, W. A.; Nabel, J.; Nam, C. C. W.; Notz, D.; Nyawira, S. S.; Paulsen, H.; Peters, K.; Pincus, R.; Pohlmann, H.; Pongratz, J.; Popp, M.; Raddatz, T. J.; Rast, S.; Redler, R.; Reick, C. H.; Rohrschneider, T.; Schemann, V.; Schmidt, H.; Schnur, R.; Schulzweida, U.; Six, K. D.; Stein, L.; Stemmler, I.; Stevens, B.; von Storch, J. S.; Tian, F.; Voigt, A.; Vrese, P.; Wieners, K. H.; Wilkenskjeld, S.; Winkler, A.; Roeckner, E., Developments in the MPI-M Earth System Model version 1.2 (MPI-ESM1.2) and Its Response to Increasing CO_2. *J Adv Model Earth Syst* **2019,** *11*, (4), 998-1038.
41. Müller, W. A.; Jungclaus, J. H.; Mauritsen, T.; Baehr, J.; Bittner, M.; Budich, R.; Bunzel, F.; Esch, M.; Ghosh, R.; Haak, H., A Higher-resolution Version of the Max Planck Institute Earth System Model (MPI-ESM1.2-HR). *J Adv Model Earth Syst* **2018,** *10*, (7), 1383-1413.

42. Gutjahr, O.; Putrasahan, D.; Lohmann, K.; Jungclaus, J. H.; von Storch, J. S.; Bruggemann, N.; Haak, H.; Stossel, A., Max Planck Institute Earth System Model (MPI-ESM1.2) for the High-Resolution Model Intercomparison Project (HighResMIP). *Geosci Model Dev* **2019,** *12*, (7), 3241-3281.
43. von Storch, J.-S.; Putrasahan, D.; Lohmann, K.; Gutjahr, O.; Jungclaus, J.; Bittner, M.; Haak, H.; Wieners, K.-H.; Giorgetta, M.; Reick, C.; Esch, M.; Gayler, V.; de Vrese, P.; Raddatz, T.; Mauritsen, T.; Behrens, J.; Brovkin, V.; Claussen, M.; Crueger, T.; Fast, I.; Fiedler, S.; Hagemann, S.; Hohenegger, C.; Jahns, T.; Kloster, S.; Kinne, S.; Lasslop, G.; Kornblueh, L.; Marotzke, J.; Matei, D.; Meraner, K.; Mikolajewicz, U.; Modali, K.; Müller, W.; Nabel, J.; Notz, D.; Peters, K.; Pincus, R.; Pohlmann, H.; Pongratz, J.; Rast, S.; Schmidt, H.; Schnur, R.; Schulzweida, U.; Six, K.; Stevens, B.; Voigt, A.; Roeckner, E., MPI-M MPI-ESM1.2-XR model output prepared for CMIP6 HighResMIP. In Earth System Grid Federation: 2017.
44. Yukimoto, S.; Kawai, H.; Koshiro, T.; Oshima, N.; Yoshida, K.; Urakawa, S.; Tsujino, H.; Deushi, M.; Tanaka, T.; Hosaka, M.; Yabu, S.; Yoshimura, H.; Shindo, E.; Mizuta, R.; Obata, A.; Adachi, Y.; Ishii, M., The Meteorological Research Institute Earth System Model Version 2.0, MRI-ESM2.0: Description and Basic Evaluation of the Physical Component. *Journal of the Meteorological Society of Japan* **2019,** *97*, (5), 931-965.
45. Yukimoto, S.; Adachi, Y.; Hosaka, M.; Sakami, T.; Yoshimura, H.; Hirabara, M.; Tanaka, T. Y.; Shindo, E.; Tsujino, H.; Deushi, M.; Mizuta, R.; Yabu, S.; Obata, A.; Nakano, H.; Koshiro, T.; Ose, T.; Kitoh, A., A New Global Climate Model of the Meteorological Research Institute: MRI-CGCM3-Model Description and Basic Performance. *Journal of the Meteorological Society of Japan* **2012,** *90A*, 23-64.
46. Yukimoto, S.; Koshiro, T.; Kawai, H.; Oshima, N.; Yoshida, K.; Urakawa, S.; Tsujino, H.; Deushi, M.; Tanaka, T.; Hosaka, M.; Yoshimura, H.; Shindo, E.; Mizuta, R.; Ishii, M.; Obata, A.; Adachi, Y., MRI MRI-ESM2.0 model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2019.
47. Shindell, D. T.; Pechony, O.; Voulgarakis, A.; Faluvegi, G.; Nazarenko, L.; Lamarque, J. F.; Bowman, K.; Milly, G.; Kovari, B.; Ruedy, R.; Schmidt, G. A., Interactive ozone and methane chemistry in GISS-E2 historical and future climate simulations. *Atmos Chem Phys* **2013,** *13*, (5), 2653-2689.
48. NASA-GISS, NASA-GISS GISS-E2.1G model output prepared for CMIP6 ISMIP6. In Earth System Grid Federation: 2018.
49. NASA-GISS, NASA-GISS GISS-E2.1H model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2018.
50. Gettelman, A.; Mills, M. J.; Kinnison, D. E.; Garcia, R. R.; Smith, A. K.; Marsh, D. R.; Tilmes, S.; Vitt, F.; Bardeen, C. G.; McInerny, J.; Liu, H. L.; Solomon, S. C.; Polvani, L. M.; Emmons, L. K.; Lamarque, J. F.; Richter, J. H.; Glanville, A. S.; Bacmeister, J. T.; Phillips, A. S.; Neale, R. B.; Simpson, I. R.; DuVivier, A. K.; Hodzic, A.; Randel, W. J., The Whole Atmosphere Community Climate Model Version 6 (WACCM6). *J Geophys Res-Atmos* **2019,** *124*, (23), 12380-12403.
51. Danabasoglu, G., NCAR CESM2-WACCM model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2019.
52. Seland, Ø.; Bentsen, M.; Olivié, D. J. L.; Toniazzo, T.; Gjermundsen, A.; Graff, L. S.; Debernard, J. B.; Gupta, A. K.; He, Y.; Kirkevåg, A.; Schwinger, J.; Tjiputra, J.; Aas, K. S.; Bethke, I.; Fan, Y.; Griesfeller, J.; Grini, A.; Guo, C.; Ilicak, M.; Karset, I. H. H.; Landgren, O. A.; Liakka, J.; Moseid, K. O.; Nummelin, A.; Spensberger, C.; Tang, H.; Zhang, Z.; Heinze, C.; Iversen, T.; Schulz, M., NCC NorESM2-LM model output prepared for CMIP6 CMIP historical. In Earth System Grid Federation: 2019.
53. Byun, Y.-H., NIMS-KMA UKESM1.0-LL model output prepared for CMIP6 AerChemMIP hist-piNTCF. In Earth System Grid Federation: 2020.
54. Krasting, J.; John, J.; Blanton, C.; McHugh, C.; Nikonov, S.; Radhakrishnan, A.; Rand, K.; Zadeh, N.; Balaji, V.; Durachta, J., NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 CMIP. In Earth System Grid Federation: 2018.
55. Horowitz, L. W.; Naik, V.; Sentman, L.; Paulot, F.; Blanton, C.; McHugh, C.; Radhakrishnan, A.; Rand, K.; Vahlenkamp, H.; Zadeh, N. T.; Wilson, C.; Ginoux, P.; He, J.; John, J. G.; Lin, M.; Paynter, D. J.; Ploshay, J.; Zhang, A.; Zeng, Y., NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 AerChemMIP. In Earth System Grid Federation: 2018.
56. Séférian, R.; Delire, C.; Decharme, B.; Voldoire, A.; Salas y Melia, D.; Chevallier, M.; Saint-Martin, D.; Aumont, O.; Calvet, J.-C.; Carrer, D., Development and evaluation of CNRM Earth system model–CNRM-ESM1. *Geosci Model Dev* **2016,** *9*, (4), 1423-1453.
57. Hegglin, M.; Kinnison, D.; Lamarque, J.-F.; Plummer, D., CCMI ozone in support of CMIP6 - version 1.0. In Earth System Grid Federation: 2016.