Environmental Science and Ecotechnology 8 (2021) 100124

Contents lists available at ScienceDirect

Environmental Science and Ecotechnology

journal homepage: www.journals.elsevier.com/environmental-science-and-ecotechnology/

Multi-stage ensemble-learning-based model fusion for surface ozone simulations: A focus on CMIP6 models

Zhe Sun [a, b, *], Alexander T. Archibald [a, c, *]

[a] Centre for Atmospheric Science, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW, UK
[b] Department of Earth Sciences, University of Cambridge, Cambridge, CB2 3EQ, UK
[c] National Centre for Atmospheric Science, Cambridge, CB2 1EW, UK

ARTICLE INFO

Article history:
Received 23 July 2021
Received in revised form
1 September 2021
Accepted 9 September 2021

Keywords:
CMIP6
CCM
Surface ozone
Model ensemble
Space-time Bayesian neural network
Data fusion

ABSTRACT

Accurately simulating the geographical distribution and temporal variability of global surface ozone has long been one of the principal components of chemistry-climate modelling. However, the simulation outcomes have been reported to vary significantly as a result of the complex mixture of uncertain factors that control the tropospheric ozone budget. Settling the cross-model discrepancies to achieve higher accuracy predictions of surface ozone is thus a task of priority, and methods that overcome structural biases in models going beyond naïve averaging of model simulations are urgently required. 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 induces heavier computational burdens. Both of these multi-stage machine learning-based approaches provide frameworks for improving the fidelity of composition-climate model outputs for uses in future impact studies.

© 2021 The Author(s). Published by Elsevier B.V. on behalf of Chinese Society for Environmental Sciences, Harbin Institute of Technology, Chinese Research Academy of Environmental Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

Tropospheric ozone (O_3) is a trace-gas, near-term climate forcer with global mean lifetime ~23 days, and a major air pollutant with detrimental effects 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 O_3 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 including chronic obstructive pulmonary disease, cardiovascular diseases, preterm delivery and even premature deaths [10–15]. Global Burden of Disease (GBD) reported over 0.36 million premature deaths globally in 2019 from exposure to ambient O_3 [16]; high O_3 exposure may also exacerbate $PM_{2.5}$-mortality risk associations [17]. These results underscore the pressing need for research which links population exposure assessment to surface O_3 and its impacts on human health.

Satellite-based observations are limited in their ability to provide accurate measurements for O_3 at the surface, since ambient air O_3 is obscured by increased O_3 abundance in the upper atmosphere which prevents direct measurement by remote-sensing. Ground-level station-based observation sites are excellent tools for recording and monitoring surface O_3 but still suffered from limited spatial coverage [18,19]. The demand for full-coverage surface O_3 concentrations have promoted the application of model simulations, which have improved alongside our mechanistic understanding of tropospheric O_3 [20–22]. However, model simulations are also limited, due to imperfect O_3 chemistry mechanisms built

* Corresponding authors. Centre for Atmospheric Science, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge, CB2 1EW, UK.
E-mail addresses: zs347@cam.ac.uk (Z. Sun), ata27@cam.ac.uk (A.T. Archibald).

https://doi.org/10.1016/j.ese.2021.100124
2666-4984/© 2021 The Author(s). Published by Elsevier B.V. on behalf of Chinese Society for Environmental Sciences, Harbin Institute of Technology, Chinese Research Academy of Environmental Sciences. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

into 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 tropospheric column O_3, model differences are a leading order term of uncertainty over decadal scales [22]. There are various types of models used to simulate surface O_3. Chemical transport models (CTM) perform satisfactorily, especially in regional-level simulations [23–26], and are considered to be free of biases in meteorological dynamics due to the use of prescribed weather features. However, these models lack important atmospheric composition feedbacks to the modelled meteorology and climate, hence development of atmospheric composition-climate models (CCM); 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 of the atmosphere [27–30].

To evaluate and compare the coupled models, a number of research institutes have contributed to the Coupled Model Intercomparison 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 [31–34]. 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. [35] Large discrepancies have been detected across models; in addition to identifying the mechanistic causes for these differences [32,36], an urgent challenge is the calibration and utilisation of the simulation ensemble. Another prominent strength of free-running CCMs is their capacity for decadal simulations without nudging by observations, so the harmonisation of cross-model disagreements, in future years beyond observations, will benefit further long-term scenario projection studies.

Applying frontier machine learning algorithms to “assimilate” the outputs from multi-source modelling activities like MIPs and observation databases is an important part of environmental research in the big data era. However, unlike retrospective analysis (abbreviated as re-analysis) products like Modern-Era Retrospective analysis for Research and Applications (MERRA), which involves observational nudging in the simulation processes as a representative type of data assimilation [37], there are opportunities for post-simulation data mining by integrating various relevant databases, which is customarily named “data fusion” [38–40]. Data fusion can be thought of an aspect of bias correction, but we treat it here as a specific machine learning task. Data fusion studies, which use ensemble learning to enhance the prediction accuracy of ambient air pollution concentrations beyond observation sites, have been emerging in recent years [41–44]. However, these studies have used no more than one model simulation, integrated with predictor variables contributing to the budget of O_3, without fusing multiple simulation ensembles such as CMIP6. In addition, the conventional machine- or deep-learning approaches aim purely at brute-force fitting to high accuracy while sacrificing the interpretability of training processes, so have long been criticised as “black-box” and contradict the nature of mechanism-driven sciences like atmospheric modelling [45–47]. Under these circumstances, a performance-interpretability balance for multi-source data fusion following credible observations is of high value in atmospheric research.

Our current study is an innovative exploration on this issue, emphasising the development of innovative ensemble-learning frameworks to assimilate the multiple CMIP6 model simulation ensembles and TOAR observations to obtain a single surface O_3 dataset which captures spatiotemporal variabilities as accurately as possible. Our ultimate goal is to establish data assimilation approaches which can be projected onto future scenario forecasts, hence historical observational products only serve as labels for supervised training rather than inputs, since such inputs for future years are counterfactual. Fusing a collection of simulation ensembles, rather than using outputs from one simulation, gives more prominence to the mechanism-driven models in order to avoid brute-force overfitting resulting from external predictor variables, especially when any given model simulation may be largely biased. The primary innovation of this study lies in transplanting the conventional ensemble-learning methodology onto multi-source data fusion, and optimising an enhanced 2-stage space-time Bayesian neural network to assimilate the CMIP6 simulation ensemble. Advantages of the conventional approach include a much lower computation burden and higher accuracy in observation-covered regions, while the innovative Bayesian approach offers better spatial generalisability and intuitive perception of spatiotemporal model weighting. In either case, the multi-model fused surface O_3 concentration can fill observational gaps and enable further relevant research. As an example, we show that using a Fourier-series function to fit temporal surface O_3 variability provides a feasible way to effectively summarise periodical changes in air pollutant concentrations. Detailed evaluations and comparisons on the 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 1, 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 [48], BCC-CSM2 [49–51], IPSL-CM6A [52,53], and MPI-M-ESM1.2 [54–57]) and 2 CNRM models are not considered for fusion due to the simplified treatment of O_3 chemistry [58–62]. The prescribed O_3 is an average of 2 earlier generation atmospheric models, CESM1-WACCM and Canadian Middle Atmosphere Model (CMAM), under the auspices of the IGAC/SPARC Chemistry-Climate Model Initiative (CCMI) [63]. A total of 8 models, including BCC-ESM1 [64,65], MPI-ESM1.2-HAM [66], MRI-ESM2.0 [67–69], NASA-GISS-E2.1 [70–72], NCAR-CESM2-WACCM6 [73,74], NCC-NorESM [75], NOAA-GFDL-ESM4 [76,77], and UKESM1-0-LL [20,78–82], 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) [63] 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 Appendix.

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

Table 1
Summarisation of CMIP6 historical project participant institutes and models with chemistry schemes, spatial gridding, and experiment realisation, physics, and forcing settings. 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 are listed in the Appendix.

Institute	Model	Trop	Strat	Grids	Realisations	Physics	Forcing	Fusion	Refs
AWI	**ESM**[‖]	P[#]	P	192 × 96 × 47	$r_1^{\perp}$	p_1	f_1		[48]
BCC	**ESM1**	I	P	128 × 64 × 26	r_1, r_2, r_3	p_1	f_1	✓	[65,95]
	CSM2	P	P	320 × 160 × 19	r_1	p_1	f_1		[96–98]
CNRM[*]	**CM6.1**	N	I	256 × 128 × 91	$r_{1\text{-}5}$	p_1	f_2		[58–60]
	ESM2.1	N	I	256 × 128 × 91	r_1, r_2, r_3	p_1	f_2		[59,61,62]
HAMMOZ[§]	**MPI-ESM1.2-HAM**	I	P	192 × 96 × 47	r_1, r_2	p_1	f_1	✓	[66]
IPSL	**CM6A**	P	P	144 × 143 × 79	$r_{1\text{-}10}$	p_1	f_1		[52,53]
MOHC	**UKESM1-0-LL**[†]	I	I	192 × 144 × 85	$r_{10\text{-}12}, r_{14\text{-}19}$	p_1	f_2	✓	[80,81,99–102]
	UKESM1-0-LL	I	I	192 × 144 × 85	$r_{5\text{-}7}$	p_1	f_3	✓	
MO-NERC	**UKESM1-0-LL**	I	I	192 × 144 × 85	$r_{1\text{-}4}, r_{8\text{-}9}$	p_1	f_2	✓	
MPI-M	**ESM1.2-HR**	P	P	384 × 192 × 95	$r_{1\text{-}10}$	p_1	f_1		[54–57,103]
MRI	**ESM2.0**	I	I	128 × 64 × 80	$r_{1\text{-}5}$	p_1	f_1	✓	[69,104,105]
NASA-GISS	**E2.1-G**	I	I	144 × 90 × 40	$r_{1\text{-}10}$	p_3	f_1	✓	[70,106,107]
	E2.1-G	I	I	144 × 90 × 40	r_1, r_2, r_3	p_5	f_1	✓	
	E2.1-H	I	I	144 × 90 × 40	$r_{1\text{-}5}$	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	✓	[74,108]
NCC	**NorESM-MM**[‡]	I	P	288 × 192 × 32	r_1, r_2, r_3	p_1	f_1	✓	[75]
NIMS-KMA	**UKESM1-0-LL**	I	I	192 × 144 × 85	r_{13}	p_1	f_2	✓	[109]
NOAA-GFDL	**ESM4**	I	I	288 × 180 × 49	r_1	p_1	f_1	✓	[76,77]

[‖] 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) [110].
[#] 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) [63] 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. $r_1i_1p_1f_1$.
[*] 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.

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 Fig. 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 solely as supervised learning labels to train models that can be applied onto wider temporal-range CMIP6 CCMs (e.g. 2015–2100), rather than as inputs.

2.3. Additional auxiliary predictors

Higher prediction accuracy can be achieved when integrating additional features into statistical models, which are nominated as *auxiliary variables*, *covariates*, *predictor variables*, or *assistant features* exchangeably in terminology reported by literature [42,43,83]. Comprehensively considering the O_3 budget mechanisms, experiences from previous relevant studies, and statistical correlations with surface O_3 using generalised linear model (GLM) stepwise backward selection, 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) [84]; 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) [85]. The excluded auxiliary variables might be of mechanistic relationships with O_3 budget (e.g. concentrations of VOCs, humidity), but post-simulation data fusion studies respect the GLM stepwise screening results more, in line with the parsimony principle when enhancing prediction accuracy.

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 [42,43,83], an adjusted ensemble learning-based multi-model fusion framework is constructed as presented in the upper panel of Fig. 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 auxiliary predictor variables are first re-gridded onto the $2^\circ \times 2^\circ$ TOAR observation grids as the first stage, following procedures graphically presented in Fig. S2. In the second stage, 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 $[\cos(2\pi tT^{-1}), \sin(2\pi tT^{-1}), t]$ where T is prescribed as 1 year) [40] are mixed together as inputs for random forest, gradient boosting decision tree, and convolutional neural network regression models separately; and in the third stage, outputs from the 3 algorithms are finally blended by L_2-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 accompanied by compromising 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 [86]. Combination of random forest, gradient boosting decision tree, and convolutional neural network follows the design of previous studies [43,44].

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 Fig. 1. Space-time indices and additional predictors are put into a 10-layer 1024-node at maximum 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 at maximum BNN merely from the 6 space-time indices, to finally reach the weighted average ensemble surface O_3 concentration predictions. The numbers of hidden layers and nodes are determined by pilot tuning experiments. 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

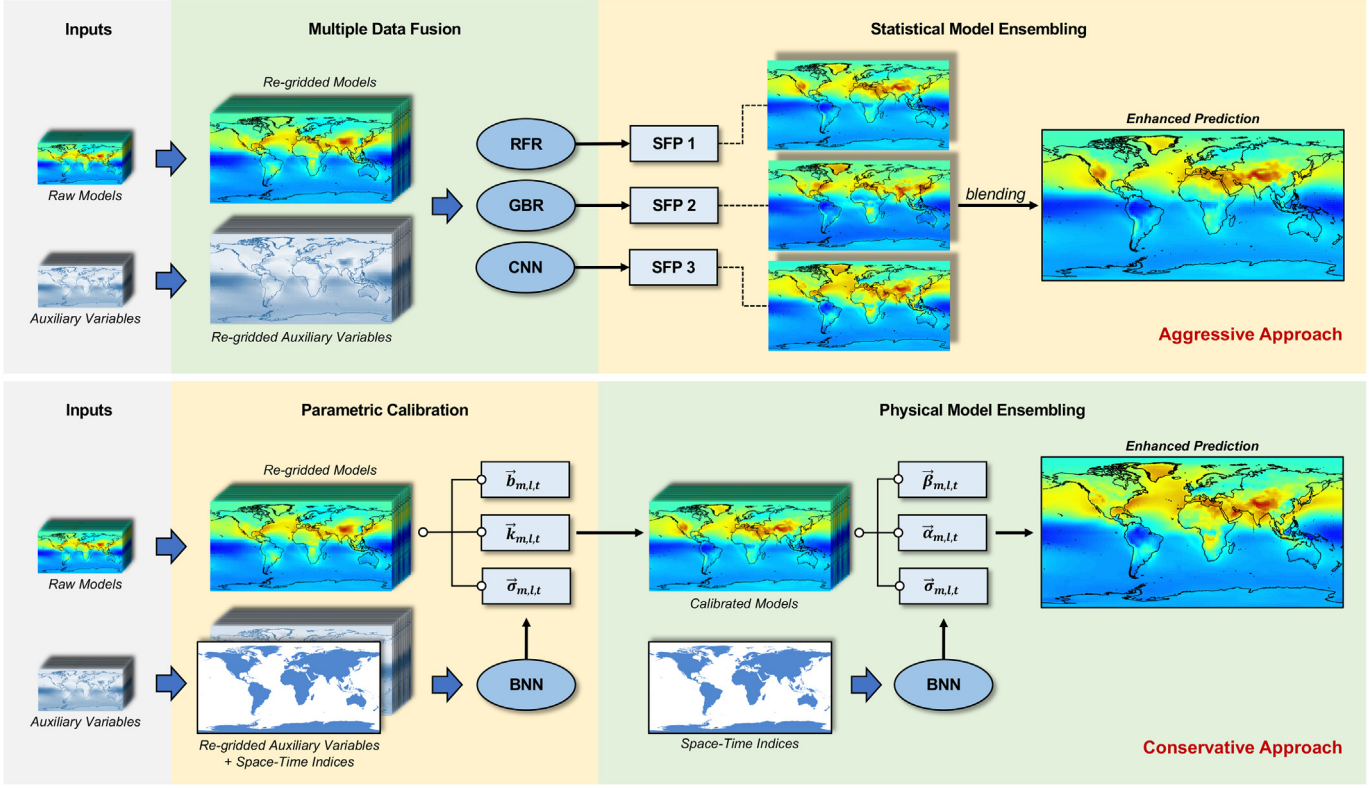

Fig. 1. Schematic diagram of machine learning-based multi-model fusion by aggressive and conservative approaches. The stacking of source data layers refers to the collections of datasets with the same level in training models; the ellipses indicate elemental machine learning methodologies; and the rectangles 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; β, bias corrector; m, physical model identifier; l, location index; t, temporal index; $\boldsymbol{\sigma}$, random noise.

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 [40]. 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 Markov-Chain Monte-Carlo (MCMC) simulation ensembles are applied to approximate the distribution, so as to guarantee the robustness of BNN estimation, thence the conservative approach involves a total of 3-stage ensembles: 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 [87]. 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.

Including space-time indices as inputs enables the two frameworks to achieve space- and time-specific simulation calibrations, rather than a simple homogeneous correction. 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 [88], so are used to capture the location-specific long-term 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. The aggressive and conservative approaches are post-simulation data fusion frameworks without any influences onto physical modelling mechanisms, and Fourier fittings are post-fusion descriptive statistics not independent from CCM simulations and fusion processes. Akaike Information Criteria (AIC) is used for statistical model selection, taking the realistic explicability altogether into consideration as listed in Table S1. Given TOAR observations and model outputs are monthly averaged, the final Fourier function is chosen as

$$f(t) = a_0 e^{a_1 t} + (b_0 + b_1 t)\sin\left(\frac{\pi}{6}t + \phi_1\right) + c_0 \sin\left(\frac{2\pi}{6}t + \phi_2\right),$$

where t represents the month-sequence; a_0 as starting-point surface O_3 concentration (January 1990); $12a_1$ as log-transformed annual average change rates (given the temporal interval is defined as month, 12 should be multiplied for converting into yearly metric); $2b_0$ as the baseline and $24b_1$ as annual change of seasonal variation amplitude (i.e. peak-valley difference; since b_0 refers to the peak-centricity or centricity-valley gaps, the peak-valley disparities need to be doubled); 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 [32], 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 (Fig. 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 $\rho = -0.58$).

The synchronicity and bias for realisation-ensembled model outputs are also evaluated in Fig. S3 and Fig. 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 (Fig. S5), and the high simulation discrepancies are mainly aggregated in the intertropical convergence zone (ITCZ) and eastern China, where the standard deviations exceed 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. In addition, since the model simulation losses are of spatiotemporal variant patterns, simple calibration approaches like subtracting a constant overestimation bias from multi-model average might be of limited use. On this condition, more complicated post-simulation statistical models are required, using a series of full spatiotemporal coverage variables to capture the patterns from observation-covered sites and project the patterns onto regions beyond observation.

3.2. Performance of multi-model ensemble fusion

Both aggressive and conservative multi-model fusion perform well in prediction enhancement (Fig. 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

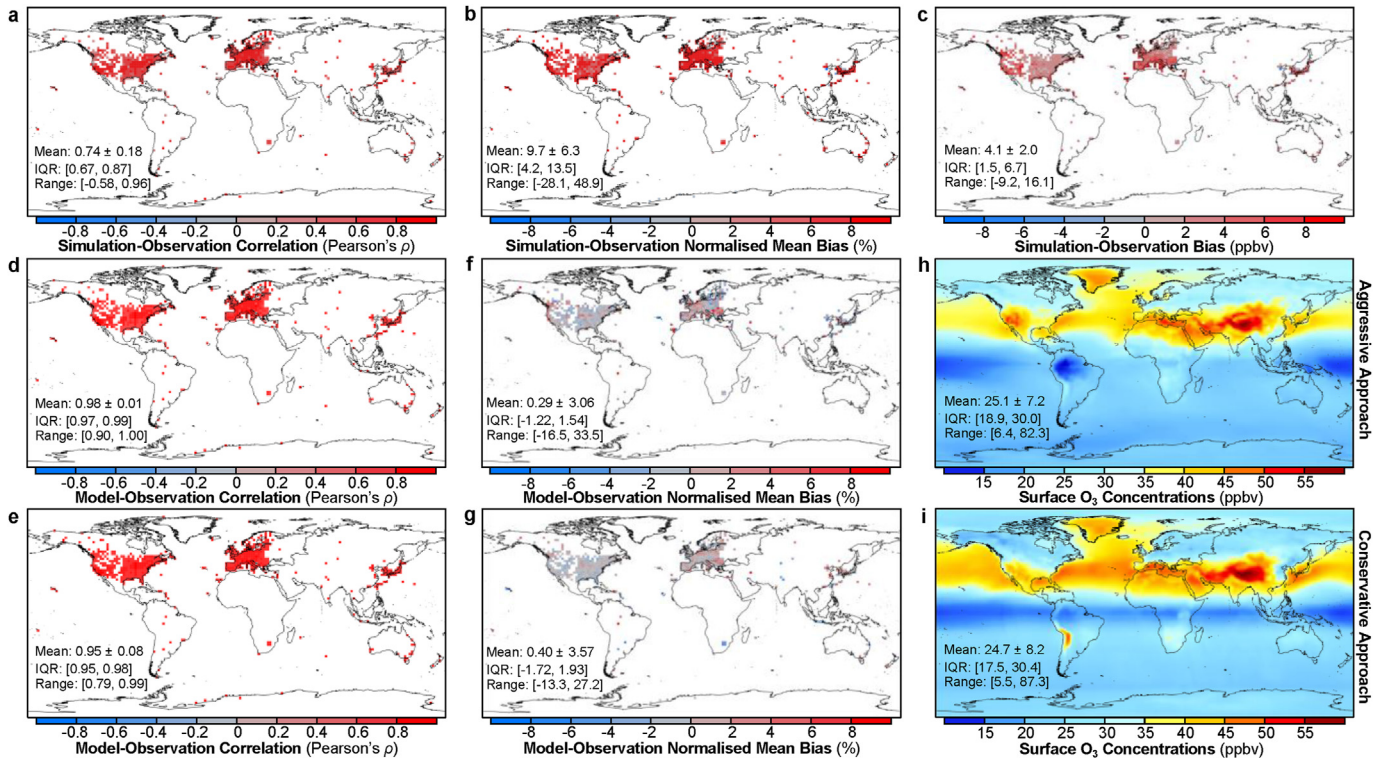

Fig. 2. Model-observation evaluation for the raw CMIP6 surface ozone simulation-ensemble and multi-model fusion by both aggressive and conservative approaches. a-c: Simulation-observation synchronicity, absolute and relative biases for 57 + 1 CMIP6 simulation ensemble. Model evaluations are conducted 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.

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 2) 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 (Fig. S6), resulting in relatively higher root mean squared error (RMSE) as 2.12 ppbv compared with 1.81 ppbv for the aggressive approach. However,

Table 2
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
South America	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

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 auxiliary 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. The crude planar and longitudinal resolutions can smooth the observational noises and extreme cases, resulting in higher prediction accuracies than other similar data fusion studies with finer spatiotemporal precisions [44,89,90].

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 2, 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.

The extreme cases of observed surface O_3 are defined as outliers exceeding 1^{st}-99^{th} percentiles, equally 8.3–50.6 ppbv. Both 2 approaches perform closely well on the low-concentration extreme O_3 as RMSE <1.92 ppbv, but the conservative approach fails in the high-concentrations owing to substantial low biases (RMSE = 6.16 ppbv, NMB = −7.67%), inferior to the aggressive approach (RMSE = 4.64 ppbv, NMB = −5.08%). This is because the Bayesian approach will "restrict" predictions into the prior probabilistic distribution, hence labelled as "*conservative approach*".

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 Fig. 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 Fig. 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 minus 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 [32], 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 ($\rho = 0.97 \pm 0.06$, IQR: [0.97, 0.99]) with the simple ensemble average than the aggressive approach ($\rho = 0.87 \pm 0.14$, 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 Fig. 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 Fig. 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 Fig. 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 (Fig. 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 potential

Decomposition of model-observation errors (Fig. S12) can assist in evaluating the potential optimisations of statistical models, as it is worthwhile to analyse the sources of prediction losses and how these may be theoretically reduced [91]. The overall RMSE for the aggressive approach is 1.81 ppbv, among which the irreducible square 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 square 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 S2. 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 the biases for the aggressive approach, while conservative approach 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 from the training subset, so that higher model variances represent severe dependence 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 the insufficient inclusion of assistant features to enhance the prediction statistically. Comparatively, due to the relatively higher statistical model variances, the aggressive approach should not be the prevalent stream for multi-model fusion, as changes in observation coverage (i.e. labels for supervision in machine learning) will substantially affect the stability of the statistical model.

4.2. Differences in spatial extrapolation

Divergences exist in the multi-model fused and calibrated surface O_3 by two approaches, especially in observation-uncovered areas. The better spatial generalisation capacity of the conservative space-time BNN multi-model fusion is an advantage over the aggressive approach. Paradoxically, the aggressive approach actually performs well in capturing extreme values. This is attributed to overfitting on

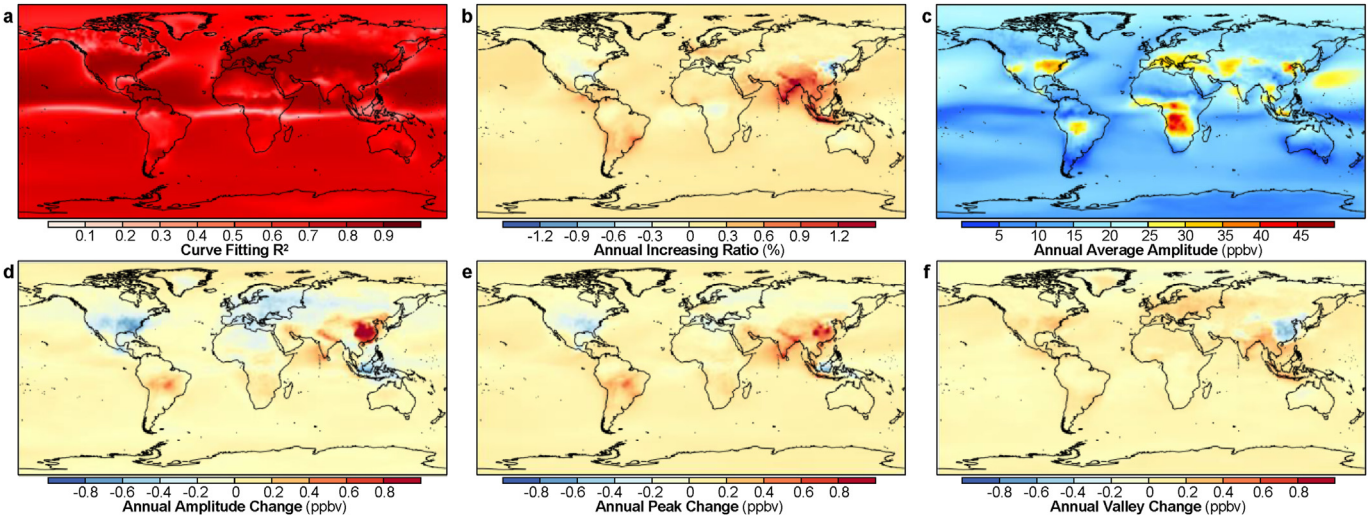

Fig. 3. Spatiotemporal variability parametrisation for CMIP6 multi-model ensemble assimilated surface ozone concentrations during 1990–2014 by the conservative approach. The ensemble-learning predicted concentrations are clustered by month. **a:** Fourier-series function-based curve-fitting quality for grid-specific surface ozone variabilities against temporal sequence, quantified by R^2. **b:** Annual increasing ratio for yearly average surface ozone concentrations, estimated by $\exp(12a_1)$-1. **c:** Annual average intra-year variation amplitude as the peak-valley gaps, estimated by $2b_0$. **d:** Annual average linear change rates of the intra-year variation 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.

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 realistic concentrations recognised by the aggressive approach are superfluous and of localised boundedness so that they may drastically differ 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 rises to 0.81, 0.83, 0.74, and RMSE declines 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, since conservative predictions manifested as underfitting by the aggressive approach should be ascribed to overfitting in the additional feature-assisted aggressive pathway. It suggests that conventional ensemble deep-learning approaches respect the observations as supervision and link the input variables only statistically, rather than respecting that the physical and chemical mechanisms are of limited use. Hence, this 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 the tropics (Fig. S8). This is primarily attributable to the lack of observations as training data, and the variations in raw simulations (Fig. S5) which result from the difficulty in capturing O_3 in this region due to complexity in the precursor emissions, including biogenic VOCs, soil NO*x*, lightning NO*x*, etc. [32] 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 (Fig. 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 land rather than the whole global surface. The conservative approach respects the raw simulations more by calibrating uniformly for both land and oceans, so that the average-conservative differences are more spatially uniform (Fig. S8).

4.4. Systematic overestimation

Direct averaging the raw CMIP6 surface O_3 simulation ensembles, as commonly used in the literature [2,32,36], 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 [92]. Subtracting a constant systematic bias-offset still cannot handle the regional variant biases. Such large biases have detrimental influences 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 to be the inadequate vertical stratification in atmospheric modules. Essentially speaking, the lowest layers of CMIP6 model simulations are used to approximate surface O_3, but the layer actually refers to a vertical average. Tropospheric O_3 concentration rises with altitude [32], thus resulting in overestimation. UKESM1-0-LL stratifies 85 vertical layers [20], which is the most among 8 interactive chemistry CMIP6 models (Table 1), and lowest overestimations are found, with underestimations observed in a few regions (Fig. S4). Further experiments to adjust vertical stratifications and observe the changes in surface O_3 simulation performance are suggested to rigorously check this speculation.

4.5. Rationality of enhanced space-time BNN

We design our enhanced 2-stage space-time BNN optimised from a classical naïve space-time BNN which does not consider additional feature involvement [40]. The enhancement in part stems from overcoming the inconsistence between the overall and location-specific observation-simulation linear relationships: each simulation cell at different time points 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 observations, thus a naïve space-time BNN framework is required for spatial extrapolation onto the full global space.

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

4.6. Comparisons with relevant studies

There are 2 other recent studies exploring possible means to fuse multiple CMIP6 simulations for surface O_3 [38,39]. Chang et al. (2019) developed an M^3Fusion method which combined a convolutional neural network (CNN) to capture the spatial auto-correlations and a recurrent neural network (RNN) to recognise the temporal dependences, so that both spatiotemporal variabilities can be reproduced in multi-scale, multi-temporal and multi-modality weighting-based data fusion with bias correction, which is a praiseworthy leap in data-driven environmental studies [39]. The main weaknesses are its opaque model training processes and lack of direct evaluation of spatiotemporal auto-correlated residuals. DeLang et al. (2021) made an improvement by adhering Bayesian Maximum Entropy (BME), a pure posterior statistical algorithm without involving machine learning, as the second-stage operation to calibrate the spatiotemporal auto-correlated residuals after M^3Fusion [38]. BME is of high statistical interpretability and efficient computation, but the prediction accuracies are not comparable with machine learning models because heavy reliance on the *priori* of spatiotemporal autocorrelation patterns may over-smooth the final predictions [93].

Our 2-stage space-time BNN in the conservative approach leverages the high prediction capacities of deep learning frameworks and core principles of maximum entropy information theory to achieve comprehensive interpretability, with the cost of extremely high computational burdens as the BNN uses probabilistic distribution estimation to replace deterministic calculations. Additionally, the space-time BNN does not perform well in extreme cases like the aggressive approach, which computes much more quickly at the expense of ignoring residual reproduction and interpretability. Comprehensively considering the pros and cons, the

aggressive approach is preferable for regional data fusion with sufficient observation coverage, while the conservative approach is of higher value for longer-term larger-scale studies which require more extrapolations, if computation expense permits.

4.7. Sensitivity analysis

Considering that cross-realisation variations (0.5 ± 0.1 ppbv) are much lower than cross-model deviations (4.6 ± 1.7 ppbv, Fig. S5), we conduct an additional sensitivity analysis by firstly averaging the multi-realisations within each model, 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 the input stacking layer, so that potential influences from imbalances in realisation, physics and forcing numbers can be thoroughly eliminated. 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. This sensitivity analysis experiment shows that unequal numbers of intra-model realisations do not significantly affect fusion performance, indicating that 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 the research target is post-simulation data fusion. It also suggests that averaging the multi-realisation ensemble before the 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 that removing one model (with all its realisations) can achieve impressive accuracy, with $R^2 = 0.91–0.93$, RMSE = 2.49–2.82 ppbv using the aggressive approach, and $R^2 = 0.89–0.93$, RMSE = 2.97–3.46 ppbv by the conservative approach, both insignificantly lower than using all 8 + 1 CMIP6 models. This evidence also corroborates the limited interference from inequality of realisation numbers towards data fusion. However, the multi-model fusion performances are substantially reduced when only 2 models are kept (keeping a single model would be inappropriate for the basic idea of *multi-model fusion*), with $R^2 = 0.83–0.87$, RMSE = 3.68–5.14 ppbv using 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 [94], 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, rather than combine into a single dataset, because of our aim to maintain the interpretability of the BNN-fusion processes instead of purely focusing on brute-force fitting.

4.8. 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 improved spatial extrapolation generalisability of our newly developed approach compared to 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, to the best of our knowledge, this shall be the first study to fuse CMIP6 model simulations for surface O_3 over the 25-year historical period of 1990–2014 using machine learning techniques, and such long-term global studies continue to be rare. Fourth, the fused and calibrated surface O_3 concentration dataset can be used further for cross-disciplinary studies. Finally, 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 have several limitations. First, the model evaluation-calibration resolution is coarse at $2^\circ \times 2^\circ$, and some heavily polluted regions including China, India and Africa still lack observations. Second, the additional assistant features used 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, and mechanistic influences from different physics and forcing settings, could have been replenished and excavated, but this is beyond the scope of the current study. We aim to address some of these issues in future research.

5. Conclusion

To explore the possibility of harmonising the cross-model simulation discrepancies and more accurately predicting the surface O_3 concentrations in decadal scales, two parallel multi-stage ensemble-learning frameworks have been developed: i) an aggressive approach using ensemble learning of random forest, gradient boosting decision tree, and convolutional neural network, and ii) a conservative approach constructing 2-stage space-time Bayesian neural networks. Both the aggressive and conservative approaches perform satisfactorily in fusing multiple CMIP6 free-running CCM surface O_3 simulations under supervision of observations, assisted with auxiliary datasets. The innovative Bayesian neural network framework is of better interpretability and higher spatiotemporal extrapolation capacity than the conventional ensemble learning model at the expense of high computation burdens. The Bayesian method is also able to present the parametric calibration and weighting layers intuitively, which can inspire further mechanistic model revisions and help improve surface O_3 modelling with CCMs in the future. Besides the development of the two machine learning frameworks as methodological frameworks for post-simulation data assimilation research, the multi-model fused surface O_3 concentrations with bias calibration also contribute to the literature for further impact and policy studies.

Declaration of competing interest

We have no conflicts of interest to disclose.

Acknowledgments

The authors are funded by Natural Environment Research Council (NERC), National Centre for Atmospheric Science (NCAS), and Fulbright Scholarship. 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, Mingtao Xia (University of California, Los Angeles) for scrutinising the code optimisation, and Michelle Wan (University of Cambridge) for polishing the language. We are also grateful to the editors and 7 anonymous reviewers for their insightful revision comments to substantially improve the manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ese.2021.100124.

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] T. Stocker, 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] P.J. Young, A.T. Archibald, K.W. Bowman, J.F. Lamarque, V. Naik, D.S. Stevenson, S. Tilmes, A. Voulgarakis, O. Wild, D. Bergmann, P. Cameron-Smith, I. Cionni, W.J. Collins, S.B. Dalsøren, R.M. Doherty, V. Eyring, G. Faluvegi, L.W. Horowitz, B. Josse, Y.H. Lee, I.A. MacKenzie, T. Nagashima, D.A. Plummer, M. Righi, S.T. Rumbold, R.B. Skeie, D.T. Shindell, S.A. Strode, K. Sudo, S. Szopa, G. Zeng, Pre-industrial to end 21st century projections of tropospheric ozone from the atmospheric chemistry and climate model Intercomparison project (ACCMIP), Atmos. Chem. Phys. 13 (4) (2013) 2063–2090.
[3] S.R. Wilson, S. Madronich, J.D. Longstreth, K.R. Solomon, 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. 18 (3) (2019) 775–803.
[4] P.S. Monks, A.T. Archibald, A. Colette, O. Cooper, M. Coyle, R. Derwent, D. Fowler, C. Granier, K.S. Law, G.E. Mills, D.S. Stevenson, O. Tarasova, V. Thouret, E. von Schneidemesser, R. Sommariva, O. Wild, M.L. Williams, Tropospheric ozone and its precursors from the urban to the global scale from air quality to short-lived climate forcer, Atmos. Chem. Phys. 15 (15) (2015) 8889–8973.
[5] S. Avnery, D.L. Mauzerall, J. Liu, L.W. Horowitz, Global crop yield reductions due to surface ozone exposure: 1. Year 2000 crop production losses and economic damage, Atmos. Environ. 45 (13) (2011) 2284–2296.
[6] S.D. Ghude, C. Jena, D.M. Chate, G. Beig, G.G. Pfister, R. Kumar, V. Ramanathan, Reductions in India's crop yield due to ozone, Geophys. Res. Lett. 41 (15) (2014) 5685–5691.
[7] C.H. Wiegman, F. Li, C.J. Clarke, E. Jazrawi, P. Kirkham, P.J. Barnes, I.M. Adcock, K.F. Chung, A comprehensive analysis of oxidative stress in the ozone-induced lung inflammation mouse model, Clin. Sci. (Lond.) 126 (6) (2014) 425–440.
[8] R. McConnell, K. Berhane, F. Gilliland, S.J. London, T. Islam, W.J. Gauderman, E. Avol, H.G. Margolis, J.M. Peters, Asthma in exercising children exposed to ozone: a cohort study, Lancet 359 (9304) (2002) 386–391.
[9] P.E. Sheffield, J. Zhou, J.L. Shmool, J.E. Clougherty, Ambient ozone exposure and children's acute asthma in New York City: a case-crossover analysis, Environ. Health 14 (1) (2015) 25.
[10] S.D. Ghude, D.M. Chate, C. Jena, G. Beig, R. Kumar, M.C. Barth, G.G. Pfister, S. Fadnavis, P. Pithani, Premature mortality in India due to $PM_{2.5}$ and ozone exposure, Geophys. Res. Lett. 43 (9) (2016) 4650–4658.
[11] X. Qiu, Y. Wei, Y. Wang, Q. Di, T. Sofer, Y.A. Awad, J. Schwartz, Inverse probability weighted distributed lag effects of short-term exposure to $PM_{2.5}$ and ozone on CVD hospitalizations in New England Medicare participants - exploring the causal effects, Environ. Res. 182 (2020) 109095.
[12] M.C. Turner, M. Jerrett, C.A. Pope 3rd, D. Krewski, S.M. Gapstur, W.R. Diver, B.S. Beckerman, J.D. Marshall, J. Su, D.L. Crouse, R.T. Burnett, Long-term ozone exposure and mortality in a large prospective study, Am. J. Respir. Crit. Care Med. 193 (10) (2016) 1134–1142.
[13] Q. Di, Y. Wang, A. Zanobetti, Y. Wang, P. Koutrakis, C. Choirat, F. Dominici, J.D. Schwartz, Air pollution and mortality in the medicare population, N. Engl. J. Med. 376 (26) (2017) 2513–2522.
[14] M.L. Bell, A. McDermott, S.L. Zeger, J.M. Samet, F. Dominici, Ozone and short-term mortality in 95 US urban communities, 1987-2000, J. Am. Med. Assoc. 292 (19) (2004) 2372–2378.
[15] Z. Sun, L. Yang, X. Bai, W. Du, G. Shen, J. Fei, Y. Wang, A. Chen, Y. Chen, M. Zhao, Maternal ambient air pollution exposure with spatial-temporal variations and preterm birth risk assessment during 2013-2017 in Zhejiang Province, China, Environ. Int. 133 (Pt B) (2019) 105242.
[16] Institute for health metrics and evaluation GBD compare data visualization. http://vizhub.healthdata.org/gbd-compare.
[17] S. Weichenthal, L.L. Pinault, R.T. Burnett, Impact of oxidant gases on the relationship between outdoor fine particulate air pollution and non-accidental, cardiovascular, and respiratory mortality, Sci. Rep. 7 (1) (2017) 1–10.
[18] M.G. Schultz, S. Schröder, O. Lyapina, O. Cooper, I. Galbally, I. Petropavlovskikh, E. Von Schneidemesser, H. Tanimoto, Y. Elshorbany, M. Naja, R. Seguel, U. Dauert, P. Eckhardt, S. Feigenspahn, M. Fiebig, A.-G. Hjellbrekke, Y.-D. Hong, P. Christian Kjeld, H. Koide, G. Lear, D. Tarasick, M. Ueno, M. Wallasch, D. Baumgardner, M.-T. Chuang, R. Gillett, M. Lee, S. Molloy, R. Moolla, T. Wang, K. Sharps, J.A. Adame, G. Ancellet, F. Apadula, P. Artaxo, M. Barlasina, M. Bogucka, P. Bonasoni, L. Chang, A. Colomb, E. Cuevas, M. Cupeiro, A. Degorska, A. Ding, M. Fröhlich, M. Frolova, H. Gadhavi, F. Gheusi, S. Gilge, M.Y. Gonzalez, V. Gros, S.H. Hamad, Helmig, D. A1, D. Henriques, O. Hermansen, R. Holla, J. Huber, U. Im, D.A. Jaffe, N. Komala, D. Kubistin, K.-S. Lam, T. Laurila, H. Lee, I. Levy, C. Mazzoleni, L. Mazzoleni, A. McClure-Begley, M. Mohamad, M. Murovic, M. Navarro-Comas, F. Nicodim, D. Parrish, K.A. Read, N. Reid, L. Ries, P. Saxena, J.J. Schwab, Y. Scorgie, I. Senik, P. Simmonds, V. Sinha, A. Skorokhod, G. Spain, W. Spangl, R. Spoor, S.R. Springston, K. Steer, M. Steinbacher, E. Suharguniyawan, P. Torre, T. Trickl, L. Weili, R. Weller, X. Xu, L. Xue, M. Zhiqiang, Tropospheric ozone assessment report: database and metrics data of global surface ozone observations, Elementa-Sci Anthrop 5 (2017) 58–83.
[19] P. Zoogman, D.J. Jacob, K. Chance, H.M. Worden, D.P. Edwards, L. Zhang, Improved monitoring of surface ozone by joint assimilation of geostationary satellite observations of ozone and CO, Atmos. Environ. 84 (2014) 254–261.
[20] A.T. Archibald, F.M. O'Connor, N.L. Abraham, S. Archer-Nicholls, M.P. Chipperfield, M. Dalvi, G.A. Folberth, F. Dennison, S.S. Dhomse, P.T. Griffiths, C. Hardacre, A.J. Hewitt, R. Hill, C.E. Johnson, J. Keeble, M.O. Köhler, O. Morgenstern, J.P. Mulchay, C. Ordóñez, R.J. Pope, S. Rumbold, M.R. Russo, N. Savage, A. Sellar, M. Stringer, S. Turnock, O. Wild, G. Zeng, Description and evaluation of the UKCA stratosphere-troposphere chemistry scheme (StratTrop vn 1.0) implemented in UKESM1, Geosci. Model Dev. (GMD) (2019) 1223–1266.
[21] T. Wang, L. Xue, P. Brimblecombe, Y.F. Lam, L. Li, L. Zhang, Ozone pollution in China: a review of concentrations, meteorological influences, chemical precursors, and effects, Sci. Total Environ. 575 (2017) 1582–1596.
[22] A. Archibald, J. Neu, Y. Elshorbany, O. Cooper, P. Young, H. Akiyoshi, R. Cox, M. Coyle, R. Derwent, M. Deushi, Tropospheric Ozone Assessment ReportA critical review of changes in the tropospheric ozone burden and budget from 1850 to 2100, Elementa: Science of the Anthropocene 8 (1) (2020) 34–86.
[23] D. Byun, K.L. Schere, Review of the governing equations, computational algorithms, and other components of the models-3 community multiscale Air quality (CMAQ) modeling system, Appl. Mech. Rev. 59 (2) (2006) 51–77.
[24] K.W. Appel, S.L. Napelenok, K.M. Foley, H.O.T. Pye, C. Hogrefe, D.J. Luecken, J.O. Bash, S.J. Roselle, J.E. Pleim, H. Foroutan, W.T. Hutzell, G.A. Pouliot, G. Sarwar, K.M. Fahey, B. Gantt, R.C. Gilliam, N.K. Heath, D. Kang, R. Mathur, D.B. Schwede, T.L. Spero, D.C. Wong, J.O. Young, Description and evaluation of the Community Multiscale Air Quality (CMAQ) modeling system version 5.1, Geosci. Model Dev. (GMD) 10 (4) (2017) 1703–1732.
[25] T.W. Tesche, R. Morris, G. Tonnesen, D. McNally, J. Boylan, P. Brewer, CMAQ/CAMx annual 2002 performance evaluation over the eastern US, Atmos. Environ. 40 (26) (2006) 4906–4919.
[26] K.A. Mar, N. Ojha, A. Pozzer, T.M. Butler, Ozone air quality simulations with WRF-Chem (v3.5.1) over Europe: model evaluation and chemical mechanism comparison, Geosci. Model Dev. (GMD) 9 (10) (2016) 3699–3728.
[27] G.W. Mann, K.S. Carslaw, D.V. Spracklen, D.A. Ridley, P.T. Manktelow, M.P. Chipperfield, S.J. Pickering, C.E. Johnson, Description and evaluation of GLOMAP-mode: a modal global aerosol microphysics model for the UKCA composition-climate model, Geosci. Model Dev. (GMD) 3 (2) (2010) 519–551.
[28] A. McLaren, H. Banks, C. Durman, J. Gregory, T. Johns, A. Keen, J. Ridley, M. Roberts, W. Lipscomb, W. Connolley, Evaluation of the sea ice simulation in a new coupled atmosphere-ocean climate model (HadGEM1), J. Geophys. Res.: Oceans 111 (C12) (2006) 14–30.
[29] A.A. Sellar, J. Walton, C.G. Jones, R. Wood, N.L. Abraham, M. Andrejczuk, M.B. Andrews, T. Andrews, A.T. Archibald, L. Mora, H. Dyson, M. Elkington, R. Ellis, P. Florek, P. Good, L. Gohar, S. Haddad, S.C. Hardiman, E. Hogan, A. Iwi, C.D. Jones, B. Johnson, D.I. Kelley, J. Kettleborough, J.R. Knight,

M.O. Köhler, T. Kuhlbrodt, S. Liddicoat, I. Linova-Pavlova, M.S. Mizielinski, O. Morgenstern, J. Mulcahy, E. Neininger, F.M. O'Connor, R. Petrie, J. Ridley, J.C. Rioual, M. Roberts, E. Robertson, S. Rumbold, J. Seddon, H. Shepherd, S. Shim, A. Stephens, J.C. Teixiera, Y. Tang, J. Williams, A. Wiltshire, P.T. Griffiths, Implementation of U.K. Earth system models for CMIP6, J. Adv. Model. Earth Syst. 12 (4) (2020), e2019MS001946.

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

[31] V. Eyring, S. Bony, G.A. Meehl, C.A. Senior, B. Stevens, R.J. Stouffer, K.E. Taylor, Overview of the coupled model Intercomparison project Phase 6 (CMIP6) experimental design and organization, Geosci. Model Dev. (GMD) 9 (5) (2016) 1937–1958.

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

[33] R.J. Stouffer, V. Eyring, G.A. Meehl, S. Bony, C. Senior, B. Stevens, K.E. Taylor, CMIP5 scientific gaps and recommendations for CMIP6, Bull. Am. Meteorol. Soc. 98 (1) (2017) 95–105.

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

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

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

[37] R. Gelaro, W. McCarty, M.J. Suarez, R. Todling, A. Molod, L. Takacs, C. Randles, A. Darmenov, M.G. Bosilovich, R. Reichle, K. Wargan, L. Coy, R. Cullather, C. Draper, S. Akella, V. Buchard, A. Conaty, A. da Silva, W. Gu, G.K. Kim, R. Koster, R. Lucchesi, D. Merkova, J.E. Nielsen, G. Partyka, S. Pawson, W. Putman, M. Rienecker, S.D. Schubert, M. Sienkiewicz, B. Zhao, The Modern-Era retrospective analysis for research and applications, version 2 (MERRA-2), J. Clim. 30 (13) (2017) 5419–5454.

[38] M.N. DeLang, J.S. Becker, K.-L. Chang, M.L. Serre, O.R. Cooper, M.G. Schultz, S. Schröder, X. Lu, L. Zhang, M. Deushi, Mapping yearly fine resolution global surface ozone through the bayesian maximum entropy data fusion of observations and model output for 1990-2017, Environ. Sci. Technol. 55 (8) (2021) 4389–4398.

[39] K.-L. Chang, O.R. Cooper, J.J. West, M.L. Serre, M.G. Schultz, M. Lin, V. Marécal, B. Josse, M. Deushi, K. Sudo, J. Liu, C.A. Keller, A new method (M^3Fusion v1) for combining observations and multiple model output for an improved estimate of the global surface ozone distribution, Geosci. Model Dev. (GMD) 12 (3) (2019) 955–978.

[40] U. Sengupta, M. Amos, S. Hosking, C.E. Rasmussen, M. Juniper, P. Young, Ensembling geophysical models with bayesian neural networks, Adv. Neural Inf. Process. Syst. 33 (2020).

[41] Q. Di, H. Amini, L. Shi, I. Kloog, R. Silvern, J. Kelly, M.B. Sabath, C. Choirat, P. Koutrakis, A. Lyapustin, Assessing NO_2 concentration and model uncertainty with high spatiotemporal resolution across the contiguous United States using ensemble model averaging, Environ. Sci. Technol. 54 (3) (2019) 1372–1384.

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

[43] Q. Di, I. Kloog, P. Koutrakis, A. Lyapustin, Y. Wang, J. Schwartz, Assessing $PM_{2.5}$ exposures with high spatiotemporal resolution across the continental United States, Environ. Sci. Technol. 50 (9) (2016) 4712–4721.

[44] B. Lyu, Y. Hu, W. Zhang, Y. Du, B. Luo, X. Sun, Z. Sun, Z. Deng, X. Wang, J. Liu, X. Wang, A.G. Russell, 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. 53 (13) (2019) 7306–7315.

[45] X. Ren, Z. Mi, P.G. Georgopoulos, 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. 142 (2020) 105827.

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

[47] A. Wang, J. Xu, R. Tu, M. Saleh, M. Hatzopoulou, Potential of machine learning for prediction of traffic related air pollution, Transport Res D-Tr E 88 (2020) 102599.

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

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

[50] T. Wu, R. Yu, Y. Lu, W. Jie, Y. Fang, J. Zhang, L. Zhang, X. Xin, L. Li, Z. Wang, BCC-CSM2-HR: a high-resolution version of the Beijing Climate Center climate system model, Geosci. Model Dev. (GMD) 14 (5) (2020) 2977–3006.

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

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

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

[54] T. Mauritsen, J. Bader, T. Becker, J. Behrens, M. Bittner, R. Brokopf, V. Brovkin, M. Claussen, T. Crueger, M. Esch, I. Fast, S. Fiedler, D. Flaschner, V. Gayler, M. Giorgetta, D.S. Goll, H. Haak, S. Hagemann, C. Hedemann, C. Hohenegger, T. Ilyina, T. Jahns, D. Jimenez-de-la-Cuesta, J. Jungclaus, T. Kleinen, S. Kloster, D. Kracher, S. Kinne, D. Kleberg, G. Lasslop, L. Kornblueh, J. Marotzke, D. Matei, K. Meraner, U. Mikolajewicz, K. Modali, B. Mobis, W.A. Muller, J. Nabel, C.C.W. Nam, D. Notz, S.S. Nyawira, H. Paulsen, K. Peters, R. Pincus, H. Pohlmann, J. Pongratz, M. Popp, T.J. Raddatz, S. Rast, R. Redler, C.H. Reick, T. Rohrschneider, V. Schemann, H. Schmidt, R. Schnur, U. Schulzweida, K.D. Six, L. Stein, I. Stemmler, B. Stevens, J.S. von Storch, F. Tian, A. Voigt, P. Vrese, K.H. Wieners, S. Wilkenskjeld, A. Winkler, E. Roeckner, 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. 11 (4) (2019) 998–1038.

[55] W.A. Müller, J.H. Jungclaus, T. Mauritsen, J. Baehr, M. Bittner, R. Budich, F. Bunzel, M. Esch, R. Ghosh, H. Haak, A higher-resolution version of the max planck institute earth system model (MPI-ESM1.2-HR), J. Adv. Model. Earth Syst. 10 (7) (2018) 1383–1413.

[56] O. Gutjahr, D. Putrasahan, K. Lohmann, J.H. Jungclaus, J.-S. von Storch, N. Brüggemann, H. Haak, A. Stössel, Max planck institute earth system model (MPI-ESM1.2) for the high-resolution model Intercomparison project (HighResMIP), Geosci. Model Dev. (GMD) 12 (7) (2019) 3241–3281.

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

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

[59] M. Michou, P. Nabat, D. Saint-Martin, J. Bock, B. Decharme, M. Mallet, R. Roehrig, R. Séférian, S. Sénési, A. Voldoire, Present-day and historical aerosol and ozone characteristics in CNRM CMIP6 simulations, J. Adv. Model. Earth Syst. 12 (1) (2020), e2019MS001816.

[60] A. Voldoire, CNRM-CERFACS CNRM-CM6-1 model output prepared for CMIP6 CMIP, in: Earth System Grid Federation, 2018.

[61] R. Séférian, P. Nabat, M. Michou, D. Saint-Martin, A. Voldoire, J. Colin, B. Decharme, C. Delire, S. Berthet, M. Chevallier, 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. 11 (12) (2019) 4182–4227.

[62] R. Seferian, CNRM-CERFACS CNRM-ESM2-1 model output prepared for CMIP6 CMIP, in: Earth System Grid Federation, 2018.

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

[64] T. Wu, F. Zhang, J. Zhang, W. Jie, Y. Zhang, F. Wu, L. Li, J. Yan, X. Liu, X. Lu, H. Tan, L. Zhang, J. Wang, A. Hu, Beijing Climate Center Earth System Model version 1 (BCC-ESM1): model description and evaluation of aerosol simulations, Geosci. Model Dev. (GMD) 13 (3) (2020) 977–1005.

[65] J. Zhang, T. Wu, X. Shi, F. Zhang, J. Li, M. Chu, Q. Liu, J. Yan, Q. Ma, M. Wei, BCC BCC-ESM1 model output prepared for CMIP6 CMIP piControl, in: Earth System Grid Federation, 2018.

[66] D. Neubauer, S. Ferrachat, C. Siegenthaler-Le Drian, J. Stoll, D.S. Folini, I. Tegen, K.-H. Wieners, T. Mauritsen, I. Stemmler, S. Barthel, I. Bey, N. Daskalakis, B. Heinold, H. Kokkola, D. Partridge, S. Rast, H. Schmidt, N. Schutgens, T. Stanelle, P. Stier, D. Watson-Parris, U. Lohmann, HAMMOZ-Consortium MPI-ESM1.2-HAM model output prepared for CMIP6

AerChemMIP, in: Earth System Grid Federation, 2019.
[67] S. Yukimoto, H. Kawai, T. Koshiro, N. Oshima, K. Yoshida, S. Urakawa, H. Tsujino, M. Deushi, T. Tanaka, M. Hosaka, S. Yabu, H. Yoshimura, E. Shindo, R. Mizuta, A. Obata, Y. Adachi, M. Ishii, 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. 97 (5) (2019) 931–965.
[68] S. Yukimoto, Y. Adachi, M. Hosaka, T. Sakami, H. Yoshimura, M. Hirabara, T.Y. Tanaka, E. Shindo, H. Tsujino, M. Deushi, R. Mizuta, S. Yabu, A. Obata, H. Nakano, T. Koshiro, T. Ose, A. Kitoh, A new global climate model of the meteorological research institute: MRI-CGCM3 model description and basic performance, J. Meteorol. Soc. Jpn. 90A (2) (2012) 23–64.
[69] S. Yukimoto, T. Koshiro, H. Kawai, N. Oshima, K. Yoshida, S. Urakawa, H. Tsujino, M. Deushi, T. Tanaka, M. Hosaka, H. Yoshimura, E. Shindo, R. Mizuta, M. Ishii, A. Obata, Y. Adachi, MRI MRI-ESM2.0 model output prepared for CMIP6 CMIP, in: Earth System Grid Federation, 2019.
[70] D.T. Shindell, O. Pechony, A. Voulgarakis, G. Faluvegi, L. Nazarenko, J.F. Lamarque, K. Bowman, G. Milly, B. Kovari, R. Ruedy, G.A. Schmidt, Interactive ozone and methane chemistry in GISS-E2 historical and future climate simulations, Atmos. Chem. Phys. 13 (5) (2013) 2653–2689.
[71] NASA-GISS, NASA-GISS GISS-E2.1G model output prepared for CMIP6 ISMIP6, in: Earth System Grid Federation, 2018.
[72] NASA-GISS, NASA-GISS GISS-E2.1H model output prepared for CMIP6 CMIP, in: Earth System Grid Federation, 2018.
[73] A. Gettelman, M.J. Mills, D.E. Kinnison, R.R. Garcia, A.K. Smith, D.R. Marsh, S. Tilmes, F. Vitt, C.G. Bardeen, J. McInerny, H.L. Liu, S.C. Solomon, L.M. Polvani, L.K. Emmons, J.F. Lamarque, J.H. Richter, A.S. Glanville, J.T. Bacmeister, A.S. Phillips, R.B. Neale, I.R. Simpson, A.K. DuVivier, A. Hodzic, W.J. Randel, The whole atmosphere community climate model version 6 (WACCM6), J. Geophys. Res. Atmos. 124 (23) (2019) 12380–12403.
[74] G. Danabasoglu, NCAR CESM2-WACCM model output prepared for CMIP6 CMIP, in: Earth System Grid Federation, 2019.
[75] Ø. Seland, M. Bentsen, D.J.L. Oliviè, T. Toniazzo, A. Gjermundsen, L.S. Graff, J.B. Debernard, A.K. Gupta, Y. He, A. Kirkevåg, J. Schwinger, J. Tjiputra, K.S. Aas, I. Bethke, Y. Fan, J. Griesfeller, A. Grini, C. Guo, M. Ilicak, I.H.H. Karset, O.A. Landgren, J. Liakka, K.O. Moseid, A. Nummelin, C. Spensberger, H. Tang, Z. Zhang, C. Heinze, T. Iversen, M. Schulz, NCC NorESM2-LM model output prepared for CMIP6 CMIP historical, in: Earth System Grid Federation, 2019.
[76] J. Krasting, J. John, C. Blanton, C. McHugh, S. Nikonov, A. Radhakrishnan, K. Rand, N. Zadeh, V. Balaji, J. Durachta, NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 CMIP, in: Earth System Grid Federation, 2018.
[77] L.W. Horowitz, V. Naik, L. Sentman, F. Paulot, C. Blanton, C. McHugh, A. Radhakrishnan, K. Rand, H. Vahlenkamp, N.T. Zadeh, C. Wilson, P. Ginoux, J. He, J.G. John, M. Lin, D.J. Paynter, J. Ploshay, A. Zhang, Y. Zeng, NOAA-GFDL GFDL-ESM4 model output prepared for CMIP6 AerChemMIP, in: Earth System Grid Federation, 2018.
[78] J.P. Mulcahy, C. Jones, A. Sellar, B. Johnson, I.A. Boutle, A. Jones, T. Andrews, S.T. Rumbold, J. Mollard, N. Bellouin, C.E. Johnson, K.D. Williams, D.P. Grosvenor, D.T. McCoy, Improved aerosol processes and effective radiative forcing in HadGEM3 and UKESM1, J. Adv. Model. Earth Syst. 10 (11) (2018) 2786–2805.
[79] A.A. Sellar, C.G. Jones, J.P. Mulcahy, Y. Tang, A. Yool, A. Wiltshire, F.M. O'connor, M. Stringer, R. Hill, J. Palmieri, UKESM1: description and evaluation of the UK earth system model, J. Adv. Model. Earth Syst. 11 (12) (2019) 4513–4558.
[80] A.A. Sellar, J. Walton, C.G. Jones, R. Wood, N.L. Abraham, M. Andrejczuk, M.B. Andrews, T. Andrews, A.T. Archibald, L. Mora, H. Dyson, M. Elkington, R. Ellis, P. Florek, P. Good, L. Gohar, S. Haddad, S.C. Hardiman, E. Hogan, A. Iwi, C.D. Jones, B. Johnson, D.I. Kelley, J. Kettleborough, J.R. Knight, M.O. Köhler, T. Kuhlbrodt, S. Liddicoat, I. Linova-Pavlova, M.S. Mizielinski, O. Morgenstern, J. Mulcahy, E. Neininger, F.M. O'Connor, R. Petrie, J. Ridley, J.C. Rioual, M. Roberts, E. Robertson, S. Rumbold, J. Seddon, H. Shepherd, S. Shim, A. Stephens, J.C. Teixiera, Y. Tang, J. Williams, A. Wiltshire, P.T. Griffiths, Implementation of U.K. Earth system models for CMIP6, J. Adv. Model. Earth Syst. 12 (4) (2020), e2019MS001946.
[81] Y. Tang, S. Rumbold, R. Ellis, D. Kelley, J. Mulcahy, A. Sellar, J. Walton, C. Jones, MOHC UKESM1.0-LL model output prepared for CMIP6 CMIP, in: Earth System Grid Federation, 2019.
[82] A. Yool, J. Palmiéri, C. Jones, A. Sellar, L. de Mora, T. Kuhlbrodt, E. Popova, J. Mulcahy, A. Wiltshire, S.T. Rumbold, Spin-up of UK earth system model 1 (UKESM1) for CMIP6, J. Adv. Model. Earth Syst. (2020), e2019MS001933.
[83] Q. Di, H. Amini, L. Shi, I. Kloog, R. Silvern, J. Kelly, M.B. Sabath, C. Choirat, P. Koutrakis, A. Lyapustin, Assessing NO_2 concentration and model uncertainty with high spatiotemporal resolution across the contiguous United States using ensemble model averaging, Environ. Sci. Technol. 54 (3) (2019) 1372–1384.
[84] J.J. Danielson, D.B. Gesch, Global Multi-Resolution Terrain Elevation Data 2010 (GMTED2010), US Department of the Interior, US Geological Survey, 2011.
[85] C.T. Lloyd, A. Sorichetta, A.J. Tatem, High resolution global gridded data for use in population studies, Scientific Data 4 (1) (2017) 1–17.
[86] J. Sexton, P. Laake, Standard errors for bagged and random forest estimators, Comput. Stat. Data Anal. 53 (3) (2009) 801–811.
[87] E. Solazzo, S. Galmarini, Error apportionment for atmospheric chemistry-transport models – a new approach to model evaluation, Atmos. Chem. Phys. 16 (10) (2016) 6263–6283.
[88] E. Kovač-Andrić, J. Brana, V. Gvozdić, Impact of meteorological factors on ozone concentrations modelled by time series analysis and multivariate statistical methods, Ecol. Inf. 4 (2) (2009) 117–122.
[89] T. Xue, Y. Zheng, G. Geng, Q. Xiao, X. Meng, M. Wang, X. Li, N. Wu, Q. Zhang, T. Zhu, Estimating spatiotemporal variation in ambient ozone exposure during 2013-2017 using a data-fusion model, Environ. Sci. Technol. 54 (23) (2020) 14877–14888.
[90] T. Xue, Y. Zheng, D. Tong, B. Zheng, X. Li, T. Zhu, Q. Zhang, Spatiotemporal continuous estimates of $PM_{2.5}$ concentrations in China, 2000-2016: a machine learning method with inputs from satellites, chemical transport model, and ground observations, Environ. Int. 123 (2019) 345–357.
[91] E. Solazzo, S. Galmarini, Error apportionment for atmospheric chemistry-transport models: a new approach to model evaluation, Atmos. Chem. Phys. 16 (10) (2016) 6263–6283.
[92] Z.Q. Hakim, S. Archer-Nicholls, G. Beig, G.A. Folberth, K. Sudo, N.L. Abraham, S. Ghude, D.K. Henze, A.T. Archibald, 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. 19 (9) (2019) 6437–6458.
[93] Y. Xu, M.L. Serre, J. Reyes, W. Vizuete, Bayesian maximum entropy integration of ozone observations and model predictions: a national application, Environ. Sci. Technol. 50 (8) (2016) 4393–4400.
[94] R.G. Derwent, D.D. Parrish, A.T. Archibald, M. Deushi, S.E. Bauer, K. Tsigaridis, D. Shindell, L.W. Horowitz, M.A.H. Khan, D.E. Shallcross, 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.
[95] T.W. Wu, F. Zhang, J. Zhang, W.H. Jie, Y.W. Zhang, F.H. Wu, L. Li, J.H. Yan, X.H. Liu, X. Lu, H.Y. Tan, L. Zhang, J. Wang, A.X. Hu, Beijing Climate Center Earth System Model version 1 (BCC-ESM1): model description and evaluation of aerosol simulations, Geosci. Model Dev. (GMD) 13 (3) (2020) 977–1005.
[96] T.W. Wu, Y.X. Lu, Y.J. Fang, X.G. Xin, L. Li, W.P. Li, W.H. Jie, J. Zhang, Y.M. Liu, L. Zhang, F. Zhang, Y.W. Zhang, F.H. Wu, J.L. Li, M. Chu, Z.Z. Wang, X.L. Shi, X.W. Liu, M. Wei, A.N. Huang, Y.C. Zhang, X.H. Liu, The Beijing Climate Center climate system model (BCC-CSM): the main progress from CMIP5 to CMIP6, Geosci. Model Dev. (GMD) 12 (4) (2019) 1573–1600.
[97] T. Wu, R. Yu, Y. Lu, W. Jie, Y. Fang, J. Zhang, L. Zhang, X. Xin, L. Li, Z. Wang, BCC-CSM2-HR: a high-resolution version of the Beijing Climate Center climate system model, Geosci. Model Dev. (GMD) 14 (5) (2020) 2977–3006.
[98] T. Wu, M. Chu, M. Dong, Y. Fang, W. Jie, J. Li, W. Li, Q. Liu, X. Shi, X. Xin, J. Yan, F. Zhang, J. Zhang, L. Zhang, Y. Zhang, BCC BCC-CSM2-MR model output prepared for CMIP6 CMIP piControl, in: Earth System Grid Federation, 2018.
[99] A.T. Archibald, F.M. O'Connor, N.L. Abraham, S. Archer-Nicholls, M.P. Chipperfield, M. Dalvi, G.A. Folberth, F. Dennison, S.S. Dhomse, P.T. Griffiths, C. Hardacre, A.J. Hewitt, R. Hill, C.E. Johnson, J. Keeble, M.O. Köhler, O. Morgenstern, J.P. Mulchay, C. Ordóñez, R.J. Pope, S. Rumbold, M.R. Russo, N. Savage, A. Sellar, M. Stringer, S. Turnock, O. Wild, G. Zeng, Description and evaluation of the UKCA stratosphere-troposphere chemistry scheme (StratTrop vn 1.0) implemented in UKESM1, Geosci. Model Dev. (GMD) 13 (3) (2019) 1223–1266.
[100] J.P. Mulcahy, C. Jones, A. Sellar, B. Johnson, I.A. Boutle, A. Jones, T. Andrews, S.T. Rumbold, J. Mollard, N. Bellouin, C.E. Johnson, K.D. Williams, D.P. Grosvenor, D.T. McCoy, Improved aerosol processes and effective radiative forcing in HadGEM3 and UKESM1, J. Adv. Model. Earth Syst. 10 (11) (2018) 2786–2805.
[101] A.A. Sellar, C.G. Jones, J.P. Mulcahy, Y. Tang, A. Yool, A. Wiltshire, F.M. O'connor, M. Stringer, R. Hill, J. Palmieri, UKESM1: description and evaluation of the UK earth system model, J. Adv. Model. Earth Syst. 11 (12) (2019) 4513–4558.
[102] A. Yool, J. Palmiéri, C. Jones, A. Sellar, L. de Mora, T. Kuhlbrodt, E. Popova, J. Mulcahy, A. Wiltshire, S.T. Rumbold, Spin-up of UK earth system model 1 (UKESM1) for CMIP6, J. Adv. Model. Earth Syst. (2020), e2019MS001933.
[103] O. Gutjahr, D. Putrasahan, K. Lohmann, J.H. Jungclaus, J.S. von Storch, N. Bruggemann, H. Haak, A. Stossel, Max planck institute earth system model (MPI-ESM1.2) for the high-resolution model Intercomparison project (HighResMIP), Geosci. Model Dev. (GMD) 12 (7) (2019) 3241–3281.
[104] S. Yukimoto, H. Kawai, T. Koshiro, N. Oshima, K. Yoshida, S. Urakawa, H. Tsujino, M. Deushi, T. Tanaka, M. Hosaka, S. Yabu, H. Yoshimura, E. Shindo, R. Mizuta, A. Obata, Y. Adachi, M. Ishii, 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. 97 (5) (2019) 931–965.
[105] S. Yukimoto, Y. Adachi, M. Hosaka, T. Sakami, H. Yoshimura, M. Hirabara, T.Y. Tanaka, E. Shindo, H. Tsujino, M. Deushi, R. Mizuta, S. Yabu, A. Obata, H. Nakano, T. Koshiro, T. Ose, A. Kitoh, A new global climate model of the meteorological research institute: MRI-CGCM3-Model description and basic performance, J. Meteorol. Soc. Jpn. 90A (2012) 23–64.
[106] NASA Goddard Institute for Space Studies, NASA-GISS GISS-E2.1G model output prepared for CMIP6 ISMIP6, in: Earth System Grid Federation, 2018.
[107] NASA Goddard Institute for Space Studies, NASA-GISS GISS-E2.1H model output prepared for CMIP6 CMIP, in: Earth System Grid Federation, 2018.
[108] A. Gettelman, M.J. Mills, D.E. Kinnison, R.R. Garcia, A.K. Smith, D.R. Marsh, S. Tilmes, F. Vitt, C.G. Bardeen, J. McInerny, H.L. Liu, S.C. Solomon,

L.M. Polvani, L.K. Emmons, J.F. Lamarque, J.H. Richter, A.S. Glanville, J.T. Bacmeister, A.S. Phillips, R.B. Neale, I.R. Simpson, A.K. DuVivier, A. Hodzic, W.J. Randel, The whole atmosphere community climate model version 6 (WACCM6), J. Geophys. Res. Atmos. 124 (23) (2019) 12380–12403.

[109] Y.-H. Byun, NIMS-KMA UKESM1-0-LL model output prepared for CMIP6 AerChemMIP hist-piNTCF, in: Earth System Grid Federation, 2020.

[110] R. Séférian, C. Delire, B. Decharme, A. Voldoire, D. Salas y Melia, M. Chevallier, D. Saint-Martin, O. Aumont, J.-C. Calvet, D. Carrer, Development and evaluation of CNRM Earth system model—CNRM-ESM1, Geosci. Model Dev. (GMD) 9 (4) (2016) 1423–1453.