Learning models to predict climate-relevant properties of atmospheric aerosols

Jeff Curtis, Nicole Riemer, Matthew West

1 Introduction

The Intergovernmental Panel on Climate Change stated in the most recent assessment (`https://www.ipcc.ch/report/ar6/wg2/`) that the uncertainties in estimating the anthropogenic climate impact are largely due to atmospheric aerosol particles. These particles are tiny, ranging from nanometers to micrometers in diameter. A major source of difficulty in simulating their climate impact is due to scale interactions: Aerosol processes occur at the microscale, but have far reaching impacts on the macroscale radiative properties of the atmosphere.

Atmospheric aerosols are evolving mixtures of many different chemical species. In global climate models (GCMs), this "aerosol mixing state" is represented in a highly simplified manner. This can introduce large errors in the estimates of climate-relevant aerosol properties, such as the concentration of cloud condensation nuclei or the aerosol optical properties. To overcome the current limitations in aerosol modeling, our group has developed the particle-resolved aerosol model WRF-PartMC, a new model capability that uses a 3D Eulerian grid for the atmospheric fluid flow, while explicitly resolving the evolution of individual aerosol particles per grid cell. This method is highly accurate, but it is so expensive that even on large supercomputers, we are only able to sample a few regions on the globe at a few points in time.

For further background information regarding the subject, see Zheng et al. (2021).

Challenge. The objective of this project is to create a machine learning model trained on accurate WRF-PartMC data that predicts climate-relevant aerosol properties from only the features that current GCMs can output.

Training Data Provided. We provide model output from a WRF-PartMC simulation as training data. This simulation covers the area of Northern California, as shown in Figure 1. The dataset consists of 4-dimensional (x, y, z position and time t) arrays of different trace gases and aerosol-related quantities (e.g., total aerosol number concentration), temperature, relative humidity, and altitude above sea level. Examples of a few horizontal distributions are shown in Figure 1.

What an AI model should do. Predict the quantities listed in Table 1 based on quantities listed in Table 2.

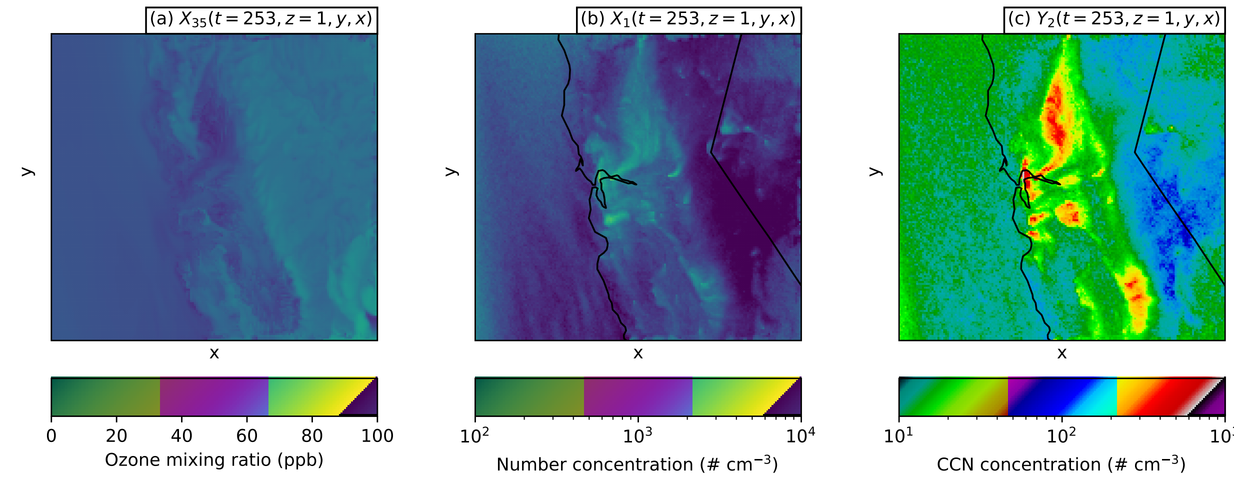

Figure 1: 2D slices of input and output data. These show x-y spatial distribution at $t = 253$ and $z = 1$ of (a) gaseous ozone (X_{35}), (b) total aerosol number (X_1) concentration and (c) CCN concentration at critical supersaturation 0.3% (Y_1) after 42 hours of simulation time.

Table 1: Quantities that we want to predict.

	Variable name	**Description**	**Importance to the project**
Y_1	ccn_001	cloud condensation nuclei (CCN) concentration at 0.1% supersaturation	high
Y_2	ccn_003	CCN concentration at 0.3% supersaturation	high
Y_3	ccn_006	CCN concentration at 0.6% supersaturation	high
Y_4	CHI	Mixing state index based on CCN properties	high
Y_5	CHI_CCN	Mixing state index based on chemical composition	high
Y_6	D_ALPHA	α-diversity	medium
Y_7	D_GAMMA	γ-diversity	medium
Y_8	D_ALPHA_CCN	α-diversity based on CCN properties	medium
Y_9	D_GAMMA_CCN	γ-diversity based on CCN properties	medium
Y_{10}	PM25	Aerosol mass concentration of particles smaller than 2.5 μm	low

Table 2: Quantities that can be used as features.

	Variable name	**Description**
X_1	TOT_NUM_CONC	Total aerosol number concentration
X_2	TOT_MASS_CONC	Total aerosol mass concentration
X_3	pmc_SO4	SO4 mass concentration
X_4	pmc_NO3	NO3 mass concentration
X_5	pmc_Cl	Cl mass concentration
X_6	pmc_NH4	NH4 mass concentration
X_7	pmc_ARO1	ARO1 mass concentration
X_8	pmc_ARO2	ARO2 mass concentration
X_9	pmc_ALK1	ALK1 mass concentration
X_{10}	pmc_OLE1	OLE1 mass concentration
X_{11}	pmc_API1	API1 mass concentration
X_{12}	pmc_API2	API2 mass concentration
X_{13}	pmc_LIM1	LIM1 mass concentration
X_{14}	pmc_LIM2	LIM2 mass concentration
X_{15}	pmc_OC	OC mass concentration
X_{16}	pmc_BC	BC mass concentration
X_{17}	pmc_H2O	H2O mass concentration
X_{18}	P+PB	Total pressure
X_{19}	TEMPERATURE	Temperature
X_{20}	REL_HUMID	Relative humidity
X_{21}	ALT	Inverse density of air
X_{22}	Z	Height
X_{23}	XLAT	Latitude
X_{24}	XLON	Longitude
X_{25}	h2so4	h2so4 mixing ratio
X_{26}	hno3	hno3 mixing ratio
X_{27}	hcl	hcl mixing ratio
X_{28}	nh3	nh3 mixing ratio
X_{29}	no	no mixing ratio
X_{30}	no2	no2 mixing ratio
X_{31}	no3	no3 mixing ratio
X_{32}	n2o5	n2o5 mixing ratio
X_{33}	hono	hono mixing ratio
X_{34}	hno4	hno4 mixing ratio
X_{35}	o3	o3 mixing ratio
X_{36}	o1d	o1d mixing ratio
X_{37}	O3P	O3P mixing ratio
X_{38}	oh	oh mixing ratio
X_{39}	ho2	ho2 mixing ratio
X_{40}	h2o2	h2o2 mixing ratio
X_{41}	co	co mixing ratio
X_{42}	so2	so2 mixing ratio
X_{43}	ch4	ch4 mixing ratio
X_{44}	c2h6	c2h6 mixing ratio
X_{45}	ch3o2	ch3o2 mixing ratio

X_{46}	ethp	ethp mixing ratio
X_{47}	hcho	hcho mixing ratio
X_{48}	ch3oh	ch3oh mixing ratio
X_{49}	ANOL	ANOL mixing ratio
X_{50}	ch3ooh	ch3ooh mixing ratio
X_{51}	ETHOOH	ETHOOH mixing ratio
X_{52}	ald2	ald2 mixing ratio
X_{53}	hcooh	hcooh mixing ratio
X_{54}	RCOOH	RCOOH mixing ratio
X_{55}	c2o3	c2o3 mixing ratio
X_{56}	pan	pan mixing ratio
X_{57}	aro1	aro1 mixing ratio
X_{58}	aro2	aro2 mixing ratio
X_{59}	alk1	alk1 mixing ratio
X_{60}	ole1	ole1 mixing ratio
X_{61}	api1	api1 mixing ratio
X_{62}	api2	api2 mixing ratio
X_{63}	lim1	lim1 mixing ratio
X_{64}	lim2	lim2 mixing ratio
X_{65}	par	par mixing ratio
X_{66}	AONE	AONE mixing ratio
X_{67}	mgly	mgly mixing ratio
X_{68}	eth	eth mixing ratio
X_{69}	OLET	OLET mixing ratio
X_{70}	OLEI	OLEI mixing ratio
X_{71}	tol	tol mixing ratio
X_{72}	xyl	xyl mixing ratio
X_{73}	cres	cres mixing ratio
X_{74}	to2	to2 mixing ratio
X_{75}	cro	cro mixing ratio
X_{76}	open	open mixing ratio
X_{77}	onit	onit mixing ratio
X_{78}	rooh	rooh mixing ratio
X_{79}	ro2	ro2 mixing ratio
X_{80}	ano2	ano2 mixing ratio
X_{81}	nap	nap mixing ratio
X_{82}	xo2	xo2 mixing ratio
X_{83}	xpar	xpar mixing ratio
X_{84}	isop	isop mixing ratio
X_{85}	isoprd	isoprd mixing ratio
X_{86}	isopp	isopp mixing ratio
X_{87}	isopn	isopn mixing ratio
X_{88}	isopo2	isopo2 mixing ratio
X_{89}	api	api mixing ratio
X_{90}	lim	lim mixing ratio
X_{91}	dms	dms mixing ratio
X_{92}	msa	msa mixing ratio
X_{93}	dmso	dmso mixing ratio

X_{94}	dmso2	dmso2 mixing ratio
X_{95}	ch3so2h	ch3so2h mixing ratio
X_{96}	ch3sch2oo	ch3sch2oo mixing ratio
X_{97}	ch3so2	ch3so2 mixing ratio
X_{98}	ch3so3	ch3so3 mixing ratio
X_{99}	ch3so2oo	ch3so2oo mixing ratio
X_{100}	ch3so2ch2oo	ch3so2ch2oo mixing ratio
X_{101}	SULFHOX	SULFHOX mixing ratio

1.1 A winning solution

We want you to come up with creative solutions for this challenge. This section can serve you as a measure of your success.

- Presentation quality. Results should be summarized in a slide deck (.pptx or .pdf), and with code (Jupyter notebooks) documented in a GitHub repository.
- Prediction accuracy. Both the root mean square error (L_2) and the mean average error (L_1) should be reported.
- Computational efficiency. Training and inference efficiency should be reported.
- Sparsity of inputs. While we have a long list of variables that could be used as features, it is desired to develop a ML model that uses only a small number of features.

2 Technical Description

The data is stored in a netCDF4 file. NetCDF (`https://www.unidata.ucar.edu/software/netcdf/`) is a set of software libraries and machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data. For this project, the netCDF file may be imported using various python packages such as netCDF4 (`https://unidata.github.io/netcdf4-python/`), xarray (`https://docs.xarray.dev/en/stable/`), or a package specific for WRF data (`https://wrf-python.readthedocs.io/en/latest/`).

The majority of data variables are stored as 4D arrays that are size (n_t, n_z, n_y, n_x) where n_t is the number of output time steps (145), n_z is the number of vertical levels (39), n_y is the number of grid cells in the south-north direction (159) and n_x is the number of grid cells in the west-east direction (169). Variables XLAT and XLON represent the latitude and longitude respectively of the grid cells and are stored as 3D arrays (n_z, n_y, n_x) but are constant in time in for this particular dataset. The 4D variable Z contains the altitude of each grid cell and these values will change over the course of the simulation due to the evolving pressure (the underlying grid uses pressure as a vertical coordinate rather than altitude). We will provide a simple example Jupyter notebook that shows how to read in the data and make some basic plots.

Important note: Before you start training the model, drop the data from grid cells at the boundary of the model domain in south-north ($x = 1$, $x = n_x$) and west-east direction ($y = 1$, $y = n_y$).

References

Zheng, Z., J. H. Curtis, Y. Yao, J. T. Gasparik, V. G. Anantharaj, L. Zhao, M. West, and N. Riemer. Estimating submicron aerosol mixing state at the global scale with machine learning and earth system modeling. *Earth and Space Science*, 8(2):e2020EA001500, 2021. doi:https://doi.org/10.1029/2020EA001500.