ARTICLE OPEN

Check for updates

Physics informed deep neural network embedded in a chemical transport model for the Amazon rainforest

Himanshu Sharma[1], Manish Shrivastava[1]✉ and Balwinder Singh[1]

Secondary organic aerosols (SOA) are fine particles in the atmosphere, which interact with clouds, radiation and affect the Earth's energy budget. SOA formation involves chemistry in gas phase, aqueous aerosols, and clouds. Simulating these chemical processes involve solving a stiff set of differential equations, which are computationally expensive steps for three-dimensional chemical transport models. Deep neural networks (DNNs) are universal function approximators that could be used to represent the complex nonlinear changes in aerosol physical and chemical processes; however, key challenges such as generalizability to extended time periods, preservation of mass balance, simulating sparse model outputs, and maintaining physical constraints have limited their use in atmospheric chemistry. Here, we develop an approach of using a physics-informed DNN that overcomes previous such challenges and demonstrates its applicability for the chemical formation processes of isoprene epoxydiol SOA (IEPOX-SOA) over the Amazon rainforest. The DNN is trained with data generated by simulating IEPOX-SOA over the entire atmospheric column, using the Weather Research and Forecasting Model coupled with Chemistry (WRF-Chem). The trained DNN is then embedded within WRF-Chem to replace the computationally expensive default solver of IEPOX-SOA formation. The trained DNN predictions generalizes well with the default model simulation of the IEPOX-SOA mass concentrations and its size distribution (20 size bins) over several days of simulations in both dry and wet seasons. The embedded DNN reduces the computational expense of WRF-Chem by a factor of 2. Our approach shows promise in terms of application to other computationally expensive chemistry solvers in climate models.

npj Climate and Atmospheric Science (2023)6:28 ; https://doi.org/10.1038/s41612-023-00353-y

INTRODUCTION

Understanding impacts of secondary organic aerosol (SOA) on the Earth's energy budget[1,2], human health[3,4] and air quality[5] has been an active area of research for the last few decades. Isoprene oxidation products contribute significantly to SOA formation mainly through the reactive chemical uptake of isoprene epoxydiols (IEPOX) in acidic aqueous aerosols[4]. Isoprene is the most abundant non-methane hydrocarbon emitted by vegetation with a global emissions rate of 500 Tg/y[6]. IEPOX-SOA formation involves the processes of gas-phase diffusion, particle-phase diffusion that is kinetically limited by the viscosity of organic aerosol coatings, and chemical reactions of IEPOX with an aqueous inorganic aerosol core. Particle–phase diffusion within the SOA shell varies in the atmosphere with temperature, relative humidity and composition of the SOA coatings, and chemical reactions within the inorganic core depend on its composition involving sulfate, acidity, and aerosol water content[7–10]. The complex reacto-diffusive processes of IEPOX-SOA formation could be simulated by a coupled set of differential equations[11,12]. However, these processes are computationally expensive, especially when applied to predict IEPOX-SOA formation and the resulting changes in the particle-size distribution over several size sections. In our recent work, the WRF-Chem model was used to simulate particles and IEPOX-SOA formation in 20 size sections ranging from 1 nm to 10 μm[11]. The explicit IEPOX-SOA solver needs to take small sub-timesteps (1–5 s) to accurately simulate IEPOX-SOA. Here, we develop and test an emulator of the complex recto-diffusive processes causing IEPOX-SOA formation using machine learning.

Recently, atmospheric chemistry has seen increasing interest in using machine-learning approaches to develop surrogates for atmospheric phenomena[13,14]. The work of ref. [14] trained a neural network model to emulate the Carbon Bond Mechanism (CBM-Z) gas-phase chemical mechanism. The neural network aimed to predict change in concentrations of reacting gas-phase chemical species an hour in future. Although the neural network was an order of magnitude faster than the reference model in ref. [15], its errors accumulated in the longer term, leading to predictions of nonphysical concentrations. Keller et al.[13] developed a random forest-based emulator for an air quality model (AQM). The model was tested for a short simulation horizon on pre-trained conditions only. However, the simulations of the ML embedded model were slower than those of the reference model, when used under long range and different conditions.

More recently, Kelp et al.[14] used the recurrent encoder-decoder architecture to train a model to represent chemical reactions on multiple timescales. The work used a CBM-Z gas-phase chemistry mechanism combined with the Model for Simulating Aerosol Interactions and Chemistry (MOSAIC) model[15,16] as a ground-truth reference model that did not include emission, deposition, advection, or any atmospheric processes other than chemistry and microphysics. Training data was generated by running the ground-truth model to generate the required input and target features. The study aimed to address the challenges of developing a machine-learning model that is *stable* and *general* so that the emulator can be used for atmospheric modeling. The recurrent encoder-decoder architecture showed a high speedup in comparison to the traditional reference solver without a significant decrease in the prediction accuracy, but was tested for a short simulation window. Further, the model developed in the study was not coupled within a three-dimensional regional chemical transport models, therefore, it is not clear if the machine-learning

[1]Pacific Northwest National Laboratory, Richland, WA 99354, USA. ✉email: manishkumar.shrivastava@pnnl.gov

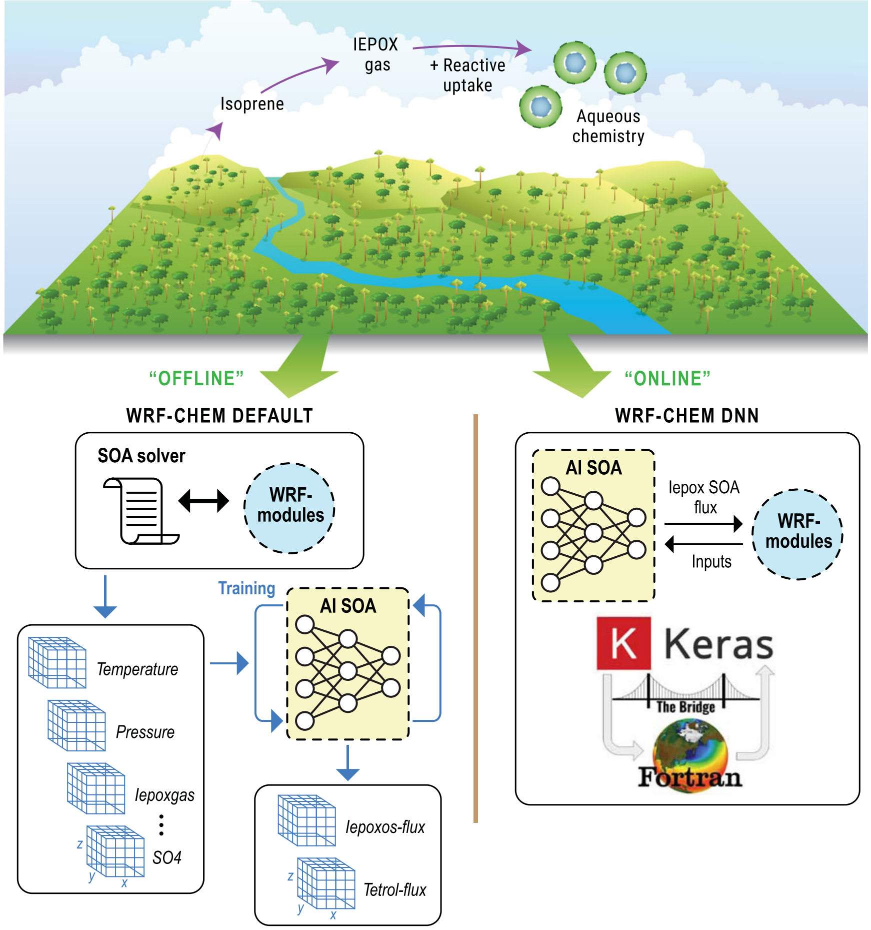

Fig. 1 Schematic illustrating the simulation of IEPOX-SOA using the WRF-Chem Default model and the WRF-Chem DNN model. The default WRF-Chem model is used to generate training data over a 7 h timescale, which is used to train the DNN to output IEPOX-SOA fluxes over each of the WRF-Chem chemistry timesteps (5 min). The trained DNN is then embedded in WRF-Chem to replace the default IEPOX-SOA solver and used to simulate IEPOX-SOA over 6 days.

model will provide similar performance and accuracy as observed in uncoupled mode.

Here, we develop a physics-informed Deep Neural Network (DNN) emulator that (a) has been embedded and coupled within the three-dimensional regional model WRF-Chem, (b) can make accurate predictions over longer time periods in a coupled mode once trained with a relatively short high-fidelity simulation data, and (c) does not experience exponential error propagation while running for much longer timescales beginning from different initial conditions.

Our aim with this work is to demonstrate the applicability of a trained DNN (AI-SOA) to represent the WRF-Chem simulations of diffuso-reactive processes of IEPOX-SOA formation. Training features input to the DNN are informed by the known physics and chemistry of IEPOX-SOA formation and their size distribution. The training data are generated by running three-dimensional simulations using the WRF-Chem model. The trained emulator is then used to replace the WRF-Chem computationally expensive IEPOX-SOA solver, and is run in-line within WRF-Chem. Since we developed and trained the emulator using a reference simulation from WRF-Chem, we call this step the *offline* step. Once the emulator is trained, we embed the DNN emulator in WRF-Chem to interact with the WRF modules at every simulation timestep. Figure 1 schematically illustrates the aforementioned steps.

Due to three-dimensional variations in concentrations of gas-phase IEPOX and particle-phase sulfate, acidity, and aerosol water content that are key variables affecting multiphase chemistry of IEPOX-SOA, the IEPOX-SOA concentrations in the computational domain are localized in specific regions, resulting in a highly skewed data distribution, which is challenging for training a DNN model. In this work, we address this challenge by a transformation of the training data to ensure that the developed emulator is robust for predicting the skewed target distributions accurately. An additional challenge with AI emulator predictions is the inability to obey physical conservation laws, leading to research in the area of "*physics-informed*" machine learning[17]. For our current work, we ensure that the predictions of the DNN model are bounded by using the flux data-based training approach presented in ref. [18]. Our results show that the embedded WRF-DNN model accurately simulates IEPOX-SOA components (2-methyltetrols and organosulfates) within each of the 20 size bins, and maintains mass balance between gas and particle phases.

Using the flux-based information additionally helps in generalizing the embedded model. In our experiments, we run 6-day simulations, although the model was only trained on 7 h WRF-Chem data. Our results demonstrate that the DNN emulators embedded in WRF-chem solutions agree with the WRF-Chem default solver with marginal errors. We show that the DNN emulator model embedded in WRF-Chem is generalizable to longer timescale simulation periods, since it predictions agree with the reference model predictions unseen by the DNN model during training. Furthermore, we found that the embedded DNN models result in a ~2X speedup in the computational time.

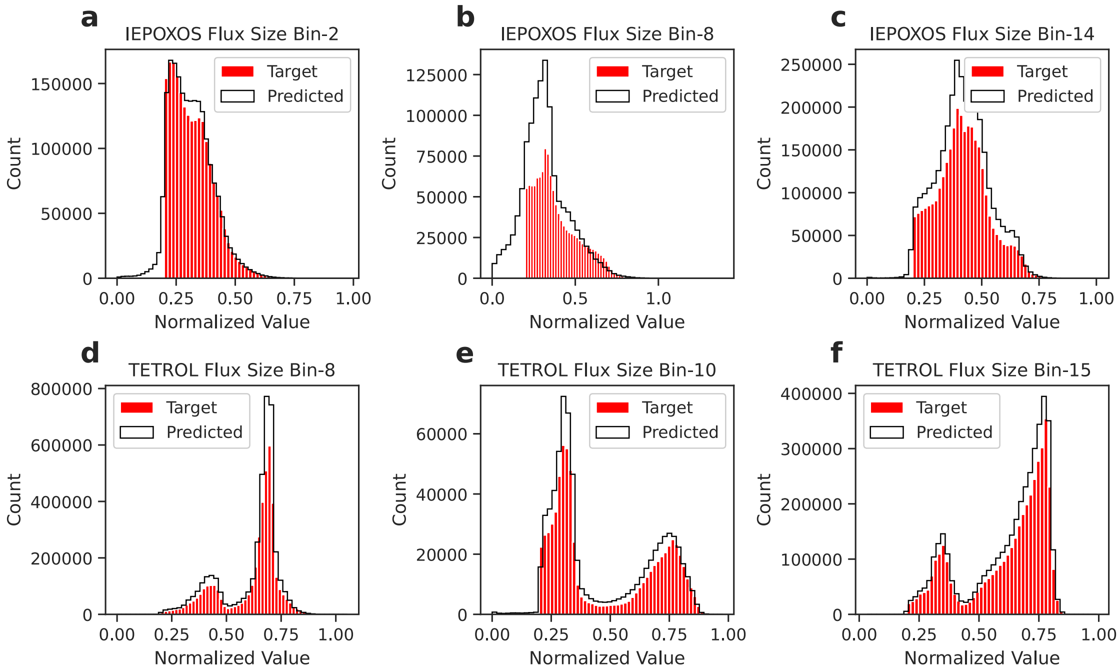

Fig. 2 Rate of change (flux) of IEPOX-SOA components due to aqueous chemistry from the test target data and those predicted by the uncoupled DNN model. (a-c) IEPOXOS particle fluxes in size bins 2, 8, 14 and (**d**–**f**) Tetrol particle fluxes in size bins 8, 10, 15.

RESULTS

DNN model performance

We evaluate our model before using it as an emulator in the WRF-Chem. The trained DNN model is evaluated to predict the unseen 10% of test data kept aside during training. Note, that in the DNN architecture (Supplementary Fig. 3) each size-bin model takes size-bin-specific inputs to predict the corresponding size-bin-specific flux targets i.e., fluxes for the 2 IEPOX-SOA components: organosulfates (iepoxos) and 2-methyltetrols (tetrols).

In Fig. 2, we plot the test target data distribution overlayed on the predictions of the DNN emulator model for a few selected size bins. The target fluxes representing the rate of change of IEPOX-SOA particulate components (IEPOXOS and Tetrol) due to reactive uptake of IEPOX are shown in Fig. 2 since the flux distribution is highly skewed with many values close to zero, we only plot values with magnitudes greater than 0.2 for improving the clarity of the figure. But our spatial contour plots in (Figs. 3 and 4) show that the model performs the challenging task of predicting zeroes as well as significant flux magnitudes in the right locations, which provides additional confidence in our results. All size-bin plots are shown in Supplementary Figs. 4 and 5. It can be seen that the model prediction accurately captures the target data distribution. This shows the qualitative performance of the standalone model prediction on the test data. To quantitatively evaluate and compare target and DNN model predictions we calculate the coefficient of determination (R^2) which are tabulated in Supplementary Table 1 for all the size-bin models. Our trained model showed high R^2 scores close to 0.9, indicating a successful parameterization and training of the models. The a priori analysis of the model provides additional confidence in its optimal performance when embedded in the WRF-Chem simulation.

DNN-embedded WRF-simulation experiment

To test our DNN model emulator in an embedded setting, we simulated a six-day duration from 2014-09-23 00:00 UTC to 2014-09-29 00:00 UTC. Typically, WRF-Chem is mostly run for 1–2 weeks due to its computational expense[19] when it has many details of SOA processes included like in this work. But similar SOA formation processes are expected to apply over the entire season. Our objective is to show that compared to a short training timescale of 7 h of data that we used for training the DNN, the model test performance was good for a much longer testing time of 6 days. Thus, the ratio of testing to training timescale was 20:1, providing a strong evidence of the generalizability of the model to long timescales. This long simulation timescale is chosen so that the WRF-Chem DNN performance can be validated with respect to the following criteria:a) the WRF-Chem DNN output should generalize well when it is applied to predict unseen input data, i.e., input data not used for training; (b) for long-timescale simulations, the error between WRF-Chem DNN and WRF-Chem default models should not accumulate and become very large at the end. (c) The WRF-Chem DNN model should provide reasonable accuracies regardless of any arbitrary non-zero initial conditions of inputs and outputs.

Figure 3 compares the spatial distribution of the total IEPOX-SOA concentrations simulated by the WRF-Chem Default and WRF-Chem-DNN models summed across the 20 size bins as a 6-day average at two different altitudes (Fig. 3a, b 0.7 km, and Fig. 3c, d 15 km altitudes). During the 6 days, the WRF-Chem DNN model predicted IEPOX-SOA within each of the 20 size sections for a range of weather conditions over the Amazon for each of the 24 h period i.e., both daytime and nighttime and all the way from the surface to the upper troposphere i.e., upto 14-km altitudes. Note that meteorological conditions (temperature, RH, pressure, winds) experienced by the WRF-Chem DNN-embedded model during the 6-day testing period include distributions unseen by the model during training and cover a wide range e.g., RH ranging 0–100%, temperatures ranging 200–320 K, winds ranging 0–20 m/s and pressure levels spanning the altitudes from surface to 14 km as shown in Supplementary Fig. 6. Similarly, the target chemical fluxes of IEPOX-SOA components tetrols and iepoxos span a wide range including zeroes and finite values that vary across the 20 size sections. It can be seen that the WRF-Chem-DNN model shows excellent agreement with the default WRF-Chem model and predicts spatial variations accurately. We calculated the root mean square error (RMSE) of IEPOX-SOA between the WRF-Chem Default and WRF-Chem DNN. At altitudes of 0.7 km and 15.0 km,

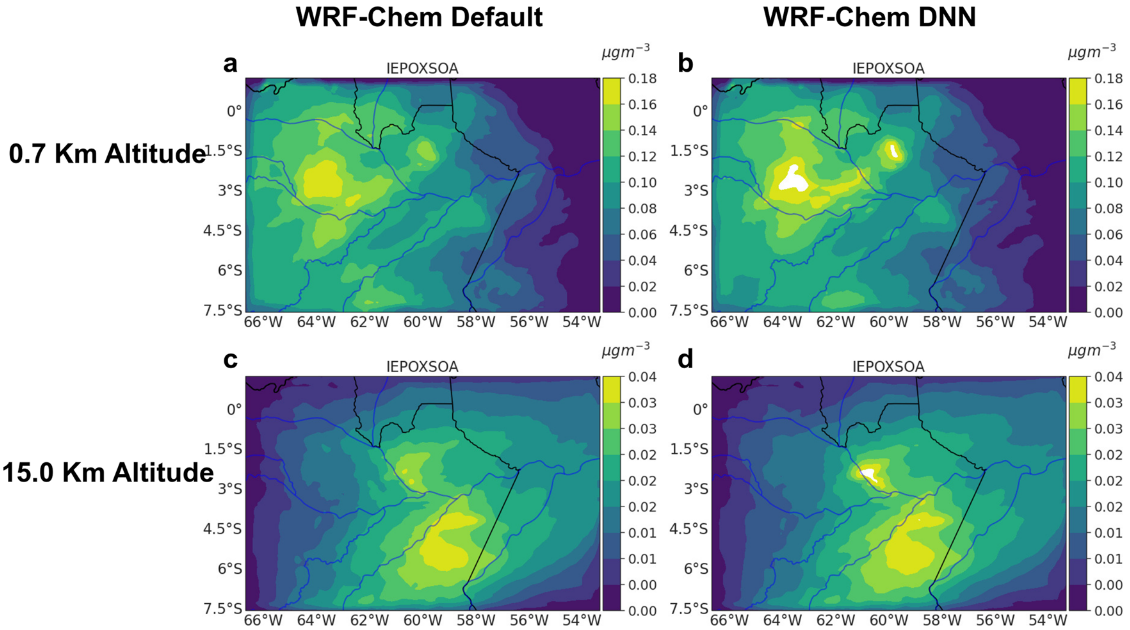

Fig. 3 Comparison of WRF-Chem default and WRF-Chem DNN model simulations at two different altitudes. Simulated time-averaged IEPOX-SOA concentrations over horizontal cross sections at (**a**, **b**) 0.7 km altitude. (**c**, **d**) 15.0 km altitude.

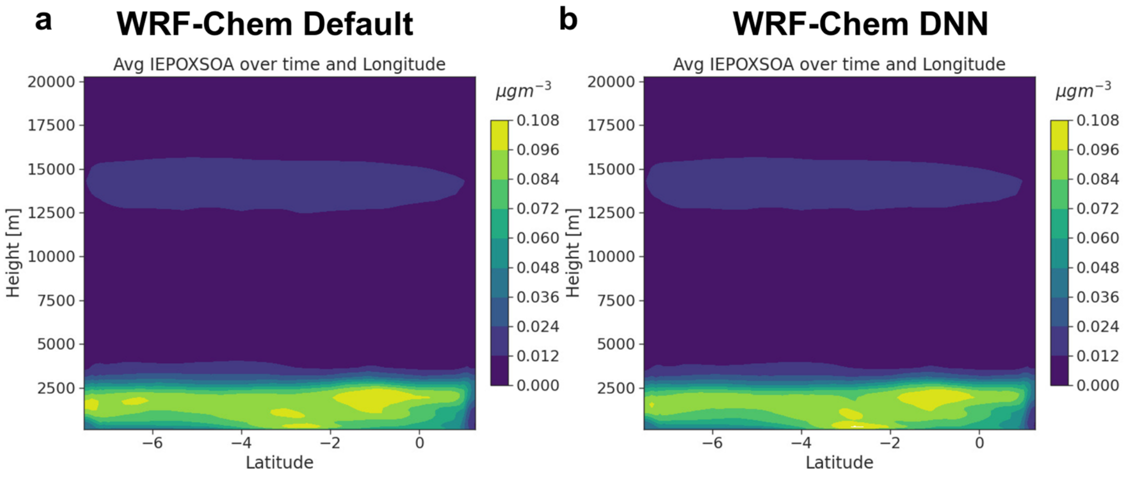

Fig. 4 WRF-Chem default and WRF-Chem DNN predictions of zonal and time-averaged IEPOX-SOA concentrations. Total IEPOX-SOA (summed across all size bins) and averaged across six days, including Sep 23–28 2014 over the Amazon (**a**) WRF-Chem Default and (**b**) WRF-Chem DNN. WRF-Chem DNN shows excellent agreement with WRF-Chem Default near surface and in the upper troposphere where IEPOX-SOA concentrations are significant.

the RMSE was low, i.e., 0.00388 μgm^{-3} and 0.000549 μgm^{-3}, respectively. Low RMSE quantitatively validates that the trained WRF-Chem DNN model captured the IEPOX-SOA simulated by WRF-Chem Default over much longer timescales (days) compared to the training data (hours) near the surface and upper troposphere. As discussed in our previous study[11], IEPOX-SOA formation is greatest near the surface due to the emissions of SOA precursors by the forest, while deep convection transports IEPOX-SOA to the upper troposphere (>10 km), resulting in almost negligible SOA in the middle troposphere. The mean absolute percent error (MAPE) in IEOPOX-SOA computation between two simulations for 0.7 km and 15 km altitude is 5.07% and 4.07%, respectively. The low MAPE value establishes that model performance did not decline even when the emulator was used on simulation days that were not part of the training set. In addition, the error accumulation is low throughout the simulation period, demonstrating the solver convergence. We further assess the generality of our DNN model, by performing new simulations with WRF-Chem DNN during another season, the wet rainy season, i.e., during March 2014. The wet season is characterized by more frequent rains, cloudy conditions and higher RH distributions compared to the dry season. For the wet season, we performed simulations for March 9–16, 2014, and compared the IEPOX-SOA predictions from WRF-Chem default and WRF-Chem DNN-embedded simulation. Note that we did not retrain the DNN model with wet season data. Thus, we challenged the DNN model to predict spatial distributions of IEPOX-SOA in different meteorological conditions, characterized by higher RH and more frequent rains, compared to the dry season where it was trained. Supplementary Fig. 7 compares the spatial distribution of the total IEPOX-SOA simulated by the WRF-Chem Default and WRF-Chem-DNN models summed across the 20 size bins as a 5-day

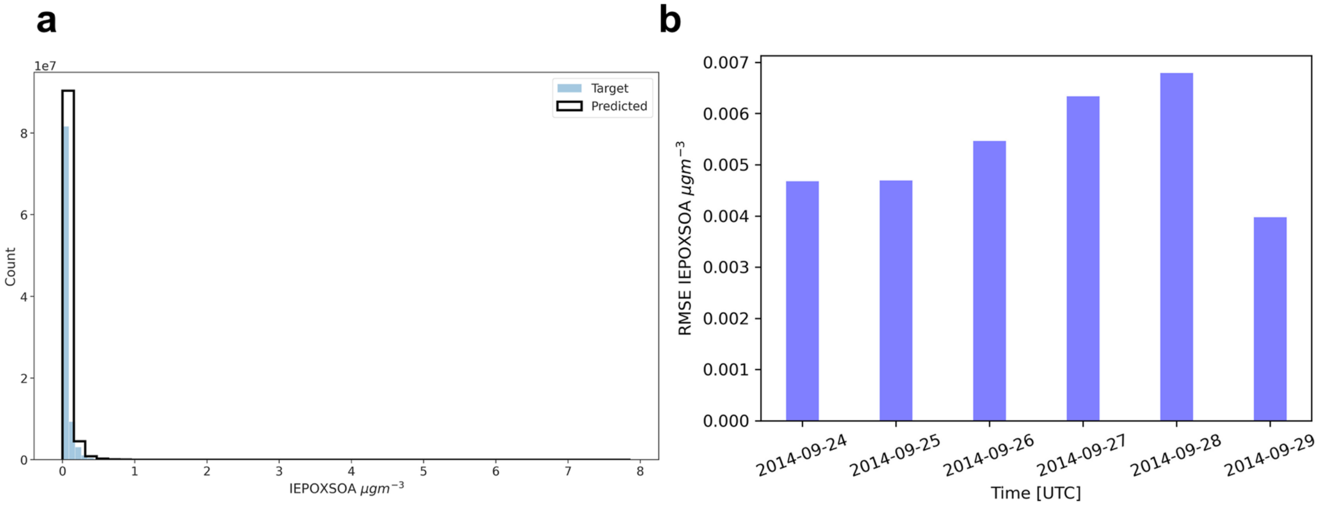

Fig. 5 Target and predicted IEPOX-SOA distributions and RMSE over time. a IEPOX-SOA distribution in the computational domain for a 6-day simulation for WRF-Chem Default (target) and the WRF-Chem DNN (predicted) run. **b** The RMSE between WRF-Chem Default and WRF-Chem DNN computation of IEPOX-SOA at 00:00 UTC of every day for the 6-day simulation.

average at two different altitudes (Supplementary Fig. 7a, b) 0.7 km, and Supplementary Fig. 7c, d 15-km altitudes). Although at both lower and higher altitude, the DNN model moderately overpredicts the IEPOX-SOA concentrations (within a factor of 2) compared to the WRF-Chem default, the DNN model captures the spatial distribution of IEPOX-SOA throughout the domain. These results are very encouraging since they indicate the DNN model trained with just 7 h of data during a completely different time and season, predicts the spatial distributions of IEPOX-SOA in different meteorological conditions i.e., during the wet season.

Most of the IEPOX-SOA formation due to multiphase chemistry occurs near the surface where concentrations of IEPOX gases, aerosols, sulfate, and their water content are greater compared to higher altitudes. As a result of warm temperatures and high RH near the surface, organic aerosols that coat the aqueous inorganic core are liquid-like; hence, diffusion limitations in the particle phase are small and do not affect IEPOX-SOA formation significantly. However, at high altitudes (10–15 km altitude) where temperatures are −50 C and RH is low, aerosols lack liquid water and organic coatings are solid imposing strong diffusion limitations for the formation of IEPOX-SOA[11]. While we included the emissions of biogenic volatile organic compounds (VOCs including isoprene, terpenes), and anthropogenic and biomass-burning emissions of trace gases and particles in the current study as documented in ref. [11], we did not include our newly discovered process of direct emissions of gas-phase 2-methyltetrols that were shown to explain IEPOX-SOA at high altitudes in our previous study[11]. Our focus in this study was to simulate the aqueous chemistry of IEPOX-SOA using the WRF-Chem-DNN model without any direct emissions of IEPOX-SOA components. But this direct emissions source is outside the aqueous chemistry module in WRF-Chem, therefore, it could easily be incorporated in future studies. WRF-Chem simulated IEPOX-SOA at high altitudes is mostly transported due to deep convection from the surface. IEPOX-SOA concentrations are mostly negligible in the middle troposphere (2–12 km altitude) since the WRF-Chem convective parameterization predicts a mixture of cloud tops at low levels (1.5–2.0 km) and deep convection that extends to the upper troposphere (greater than 12 km) during the period of interest, with lesser amounts of clouds at intervening levels. Figure 4 compares the zonal average distributions of the total IEPOX-SOA (summed over all sizes) as a 6-day average between WRF-Chem Default and WRF-Chem DNN models. WRF-Chem DNN simulations perfectly capture the variations in IEPOX-SOA with altitude similar to the WRF-Chem Default model. The bimodal peaks in the simulated IEPOX-SOA (which occur mainly below 2.5 km altitude and between 12 and 15 km) with concentrations approaching zero in the middle troposhere are well captured by the WRF-Chem DNN model. The MAPE between the two simulations for the zonal average plots (Fig. 6) is 3.88%. The small error establishes that the altitude variations of IEPOX-SOA is also well captured by the DNN model emulator embedded in WRF-Chem. Simulating both zeros and significantly higher concentrations at the two altitude regions is difficult for DNN models due to the large sparsity in training data. However, our approaches of normalizing the training data and predicting molar fluxes instead of concentrations as outputs are able to overcome these challenges and show great promise for future applications of DNN models to other similarly skewed training datasets.

In Fig. 5a, we compare the prediction accuracy of WRF-Chem and the WRF-Chem DNN model in the 3D computational domain. We compare the distributions of all the IEPOX-SOA simulated data in 3D for all six days for both the simulations WRF-Chem Default (blue) and WRF-Chem DNN (black). Figure 5a shows that the WRF-Chem DNN model successfully captures the highly skewed IEPOX-SOA distribution with even small steps in the tail of the distribution, accurately. For a duration of 6 days in the 3D domain, the IEPOX-SOA RMSE is 0.021 μgm^{-3}. Figure 5b shows the calculated RMSE in IEPOX-SOA for each day of the 6-day simulation period. The accumulation of RMSE between the WRF-Chem default and the WRF-Chem DNN remains small and does not increase significantly over time, indicating that the embedded WRF-Chem DNN emulator is generalizable over longer-time simulations with only small accumulation of errors.

Although we have demonstrated the success of the WRF-Chem DNN model in emulating the 3D distributions of total IEPOX- SOA predicted by the WRF-Chem Default model over long timescales, it is also important to simulate the size distribution of IEPOX-SOA accurately, since particle size governs its interactions with clouds and radiation. In Figure 6, we compare the predictions from WRF-Chem Default and WRF-Chem DNN models for a few representative individual size bins (size bins 6, 8, 12, 16, 18, 20) of particulate Tetrols, which constitute more than 90% of total simulated IEPOX-SOA concentrations. We compare the distributions for the full 3D WRF-Chem data over the six-day period of simulations. Simulated tetrol sizes have a large number of zeroes in the spatial 3D

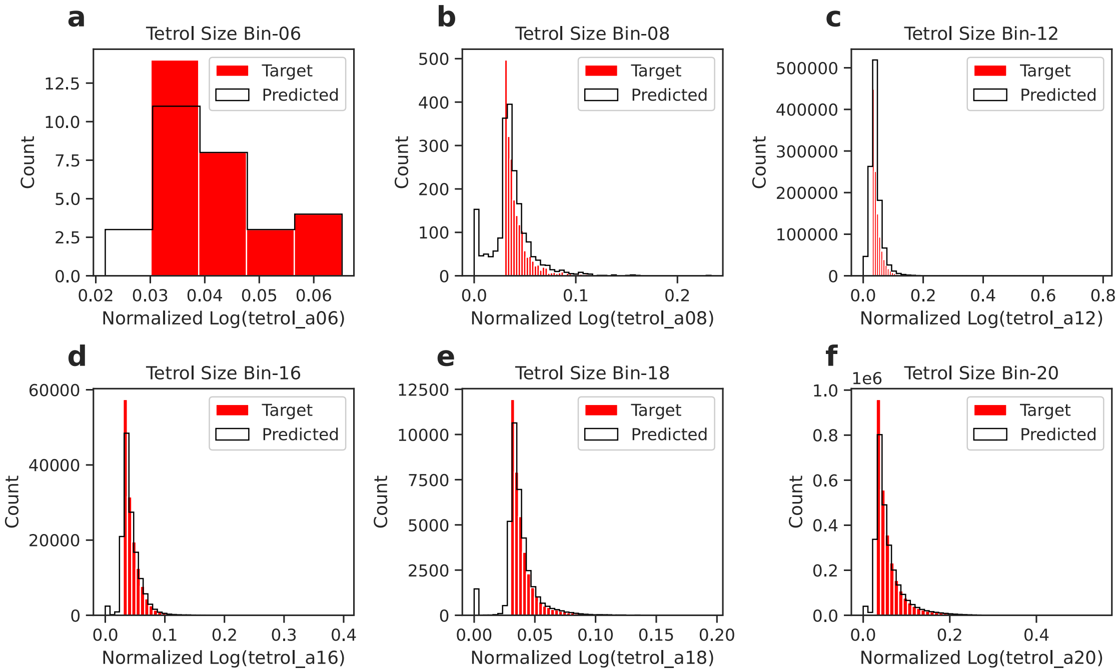

Fig. 6 WRF-Chem default (target) and WRF-Chem DNN (predicted) particle-phase tetrol distributions. a–f Selected Tetrol size bins 6, 8, 12, 16, 18, and 20 correspond to particles with diameter ranges of 9.8–15.5 nm, 24.6–39.1 nm, 156.2–248.0 nm, 1.0–1.6 μm, 2.5–4.0 μm, and 6.3–10.0 μm, respectively. Here the distribution is for the full 3D WRF-Chem domain during the 6-day simulation period.

domain, leading to skewed distribution. For visual clarity, in Fig. 6, we only plot the distributions of target tetrol concentrations with magnitudes greater than 3e-3 μgm^{-3}, but we have shown that the WRF-Chem DNN predictions capture well both zero and non-zero target concentrations simulated by the Default WRF-Chem model (Figs. 3 and 4). These results indicate that the WRF-Chem DNN captures the size distribution of IEPOX-SOA concentrations. This result is important given that simulating size distributions over several sections (20 bins in our study) is challenging. Most regional and global models only discretize the particle distributions using a few size sections or modes (3–7), whereas our simulations discretize them over much more size bins (20 size sections) to capture details of particle formation and growth. In our simulations, particle concentrations over individual size bins are much sparser (i.e., contain more values approaching zero), compared to previous models that use fewer size sections to represent particle-size distributions. Therefore simulating the size distribution of IEPOX-SOA over so many size sections was a difficult task for our WRF-Chem DNN model.

Computational performance of WRF-Chem DNN simulations

In addition to accurately simulating the 3D size distributions of IEPOX-SOA, the WRF-Chem DNN model is significantly faster and displays a computational gain of a factor of 2 compared to the WRF-Chem Default model. We compared different simulations with WRF-Chem DNN and WRF-Chem Default models varying the simulated periods and initial conditions and recorded their computational times as documented in Supplementary Table 2. In all simulations, WRF-Chem DNN is two times faster than the WRF-Chem Default simulations. Note that the actual WRF-Chem simulation time is significantly higher than the cost of training a DNN model (7 h on 8 GPU's, as described in "Methods"). Therefore, relative to the runtime of WRF-Chem model, the computational cost of training a DNN model is almost negligible. For all simulations with WRF-Chem Default and WRF-Chem DNN models, we use the high-performance computing (HPC) cluster 2 nodes with 36 Intel© Xeon™ Gold CPU on each node to run the parallel simulation.

DISCUSSION

In this work, we trained 20 different DNN models with different weights and biases to represent the formation of IEPOX-SOA within each of the 20 particle-size sections in WRF-Chem. The computational speed of running 20 embedded models in WRF-Chem was similar to running a single embedded DNN model. We found that training a single DNN model to represent IEPOX-SOA formation over all the 20 size bins concurrently was challenging due to the high dimensionality of input and output data corresponding the WRF-Chem Default model predictions. Such a high-dimensional dataset with many features requires a more complex DNN architecture than that used in this study. The training time for such a complex DNN architecture and hyperparameter tuning is large. We found that a better approach is to train different DNN models for different size bins, thus reducing both: (1) the complexity of the DNN architecture and (2) the dimensionality of the training data. Our tests showed that computational costs of embedding a single DNN in WRF-Chem was similar to embedding 20 DNNs, but the benefits of this approach are manifested by simpler DNN architecture for each of the 20 DNNs, reduced training time and high accuracy over all the size sections. The DNN can successfully simulate the complex functional dependence of the formation of IEPOX-SOA on the composition of the inorganic aerosol (sulfate, ammonium, nitrate), acidity, particle water, and diffusion limitations in the organic shell. These parameters vary spatially and temporally as a function of meteorology (temperature, relative humidity) and chemistry. By training the DNN to predict fluxes of IEPOX-SOA to particles rather than concentrations, the model could maintain the mass balance that has been a challenge in simulating concentrations of chemical species. A unique feature of our work is the ability to simulate IEPOX-SOA formation in each of the 20 size bins of particles. To the best of our knowledge, our study represents the first application of a DNN to simulate IEPOX-SOA in many size bins, since most previous models have worked with simpler aerosol size distribution treatments (often with less than eight modes or size sections). We overcame another obstacle related to training the DNN with very skewed and sparse input and output distributions in the training data. The delta distributions were difficult to simulate with a DNN, and we discovered that applying an inverse hyperbolic sine transform to data distributions worked best. Our work provides several methodological advances in training the next generation of DNN models to simulate challenging atmospheric aerosol chemistry and size distribution datasets. We have demonstrated that our trained WRF-Chem DNN model is generalizable to a wide range of weather and chemical regimes, including distributions that were not seen by the model during training and over a much longer time period of 6 days compared to the 7-h training time over the Amazon, and also over several days during the wet season of the Amazon. Our results highlight that the model temporal accuracy did not deteriorate significantly with longer temporal extrapolation. We provide a clear proof of concept for successful implementations of machine-learning algorithms that speed up complex physics and chemistry calculations over sparse output distributions while maintaining mass balance between gas and particle phases within a 3D regional chemical transport model. Applicability of this approach to other regions and chemical and meteorological regimes that are largely out of distributions (OOD) compared to the Amazon is plausible, especially by leveraging the advancements like transfer learning in the area of machine learning[20,21].

METHODS

In this section, we describe the development of our data-driven deep-learning model. In subsequent sections, we describe the WRF-Chem Default modeling approach that simulates the target IEPOX-SOA data, and the details of our deep-learning model architecture and its training procedure. We also describe the integration and embedding of our trained DNN model within WRF-Chem to replace the default IEPOX-SOA solver.

WRF-Chem and DNN model training data

We use the regional Weather Research and Forecasting Model coupled to chemistry (WRF-Chem v 4.2) model[22] at moderately high resolution with 10 km grid spacing to simulate atmospheric chemistry and SOA formation over the Amazon. The modeling domain encompasses a region of 1500 × 1000 km around the city of Manaus in Brazil. Details of model configuration, emissions, and chemistry are presented in ref. [11]. Aerosols are simulated with 20 size sections ranging from 1 nm to 10 μm, as described in ref. [23] within the Model for Simulating Aerosol Interactions and Chemistry (MOSAIC)[16]. Aerosols are assumed to be mixed internally, and both particle number and mass are simulated in each bin. Aerosol species in MOSAIC include sulfate, nitrate, ammonium, sodium, chloride, calcium, carbonate, other inorganics (OIN), elemental carbon (EC), organic matter, and aerosol water. Each chemical component of the particle phase is represented by 20 size sections as both interstitial and cloud-borne aerosols. The model advects a large number of species, greatly increasing the computational cost compared to chemistry packages without SOA. Trace gases, aerosols, and clouds are simulated simultaneously with meteorology, therefore the model is computationally expensive. SOA includes several components, including biogenic SOA from isoprene, monoterpenes, and sesquiterpenes, and anthropogenic and biomass-burning SOA. Pure gas-phase chemistry of SOA is represented by the volatility basis set (VBS) approach, while multiphase chemistry of IEPOX-SOA is simulated explicitly. In this study, we focus on the submodule in WRF-Chem related to IEPOX-SOA formation in aqueous aerosols. Particle sulfate is one of the key nucleophiles that is needed for IEPOX-SOA formation. In addition, total organic aerosol (OA) modulates IEPOX-SOA since it forms a viscous shell around the inorganic core limiting IEPOX-SOA formation. Both sulfate and total organic aerosols predicted by our WRF-Chem model were also evaluated with aircraft measurements in our previous WRF-Chem study[11]. Based on known chemistry of IEPOX-SOA from previous studies, we ensure that we have evaluated all key intermediate variables (like sulfate, total organic aerosol, RH, temperature, isoprene) to IEPOX-SOA formation that were measured by the aircraft.

To generate WRF-Chem training data, we ran the WRF-Chem simulation for 7 h from 2014-09-28 12:00 UTC to 2014-09-28 20:00 UTC and wrote 3D WRF-Chem inputs and outputs to the IEPOX-SOA solver within the aqueous chemistry module every 5 min (equal to the WRF-Chem chemistry timestep of 5 min). The choice of days Sep 21–28, 2014 were motivated by our recent study over the Amazon[11] when aircraft-based field measurements were available for model evaluation of IEPOX-SOA. During this time, the model has been extensively evaluated with aircraft-based field measurements over the Amazon. The model was initialized with a previous 12-day WRF-Chem simulation[11] (Sep 18–Oct 1 2014) on 2014-09-28 12:00 UTC. The physical and chemical processes in surface and upper air are indeed different and therefore for this study, we rely on the training data produced from the WRF-Chem high-fidelity simulations. To develop the model for predicting IEPOX-SOA we train our model using the full WRF-Chem 3D Spatio-temporal data to ensure that the model generalizes to physical processes in different environment. This is a unique strength of our proposed model that it is trained to predict the IEPOX-SOA accurately irrespective of the altitude in the environment. The various input and output features at the beginning and the end of the IEPOX-SOA submodule are explicitly written on the WRF-Chem outputs. Instead of writing concentrations of simulated IEPOX-SOA components in each of the 20 size bins, we write out their fluxes, i.e., changes in their concentrations calculated as a difference between their concentrations at the beginning and end of the aqueous chemistry submodule within WRF-Chem per unit time. The DNN model is tasked with predicting these change in concentrations in each size bin. When the 20 trained DNN models (one for each size bin) are used to make predictions by embedding them in WRF-Chem (replacing the aqueous chemistry submodule), we add the DNN output fluxes of IEPOX-SOA components to the incoming IEPOX-SOA concentrations to derive the output concentrations of 2-methyltetrols (tetrols) and IEPOX organosulfates (IEPOXOS). Predicting fluxes instead of absolute concentrations ensures the preservation of mass balance between gas- and particle phases of IEPOX-SOA and prevents the model from deviating far from the outputs over longer time periods. Consistently, Strum et al.[18] showed that by training the emulator to predict the flux information, nonphysical predictions by the emulators can be avoided.

In addition to the flux data we also output additional variables necessary for calculating the IEPOX-SOA concentration at the beginning of the IEPOX-SOA solver (aqsoagamma). The variables are used as input for model training, and described in Table 1. Both input and target variables have a large variability in the 3D domain; hence, before using them for training a DNN model, normalization is important. In addition, the data distribution of the output fluxes is sparse and almost close to a delta-distribution with many values approaching zero. Learning a delta-distribution as a regression task is challenging for machine-learning models. Performing transformations of such sparse datasets before feeding them for training can significantly improve model learning. Therefore, in this work the full 3D WRF-Chem simulated dataset for 7 h duration is first transformed using the Inverse Hyperbolic Sine (IHS) transform and then normalized using min-max normalization. The IHS transform is described in Eq. (1), and

Table 1. Input and output variables used for training the 20 DNNs corresponding to each of the 20 particulate size bins.

	Full name	Acronym
13 Inputs	Temperature (K)	tk
	Relative humidity (%)	rh
	Ambient pressure (atm)	p
	Isoprene epoxydiol gas concentration (nmol/m^3)	iepoxgas
	Organic aerosol concentration [Bin:1-20]	TOTOA
	Particle water (kg/m^3-air)	water [Bin:1-20]
	Particle sulfate (nmol/m^3)	so4 [Bin:1-20]
	Particle nitrate (nmol/m^3)	no3 [Bin:1-20]
	Particle ammonium (nmol/m^3)	nh4 [Bin:1-20]
	Diffusion Coefficient (cm^2/s)	DORGCOAT [Bin:1-20]
	H + ion concentration (mol/kg-water)	PHW_BIN [Bin:1-20]
	Total particle radius (cm)	RTOTAL [Bin:1-20]
	Radius of the aqueous core (cm)	RAQ [Bin:1-20]
2 Outputs (nmol/m^{3}s^{-1})	particle Iepox organosulfate Flux [Bin:1-20]	iepoxosFluxes
	particle Tetrol Flux [Bin:1-20]	tetrolFluxes

The variables without "BIN" appended are common variables for each model.

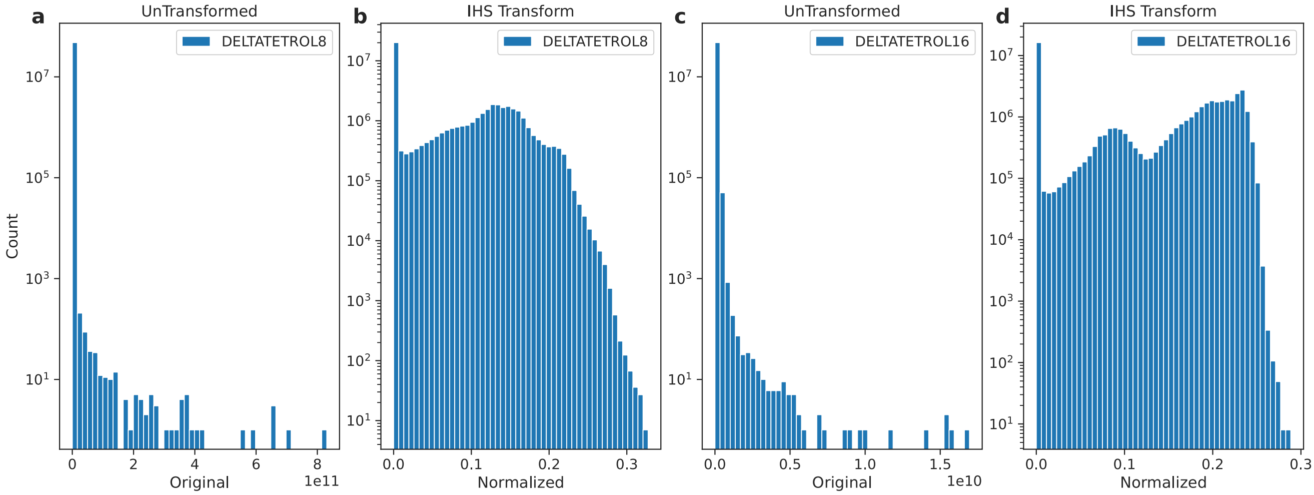

Fig. 7 Training data transformation using IHS. The inverse hyperbolic sine (IHS) transform with the min-max normalization of the output IEPOX-SOA fluxes greatly decreases the number of zeroes in the training dataset, changes the target distribution, and makes the training of DNN model feasible. **a** Original distribution of Tetrol flux of bin 8. **b** IHS-transformed distributions of Tetrol flux bin 8. **c** Original distribution of Tetrol flux Bin 16. **d** IHS-transformed Tetrol flux Bin 16 distribution.

the transformation works with data defined on the entire real number line, including zero and negative. The IHS transform behaves similar to a log transform when y values are large. Here, $\theta > 0$ and for any value of θ zero maps to zero.

$$f(y, \theta) = sinh^{-1}(\theta y)/\theta \quad (1)$$

In Fig. 7, we compare the actual data distribution and IHS-transformed normalized distribution for a target flux of tetrol particles (DELTATETROL) in 2 selected size bins. The transformed distributions show significant change, with a more visible tail. The value of θ was set to 100. Other transformed and normalized distributions are shown in Supplementary Figs. 1 and 2. During our model training experiments, we saw a significant improvement in model performance with respect to predicting output fluxes for the IHS-transformed outputs compared to the untransformed outputs. For each size bin, first we split the data into train (75%), validation (15%), and test (10%) sets. We then transformed and normalized the training set and used the same normalization scale to transform validation and test sets to avoid data leakage during model training. We prepared individual size-bin data using the above procedure. Next, we describe the DNN architecture for each model and additional model training details.

Deep neural network architecture and training details

We chose a DNN model over conventional machine-learning approaches such as ridge regression, gradient boosting trees etc. The choice of the DNN emulator over other approaches was due to the big high-dimensional training data (165 input features, 40 output features in three dimensions spatially and over multi-day timescales) generated by WRF-Chem. In the machine-learning literature, it has been shown that conventional approaches scale poorly in large data with high dimensions[24]. Also, our training data are highly nonlinear and represent complex physics and chemistry. DNNs are artificial neural networks and are widely used in the scientific community in several disciplines, including chemistry, material science, climate science, physical science applications, astronomy, etc.[25–27]. Details on DNN are documented in ref. [28]. Recently, the Earth system sciences community has also leveraged DNNs to address a variety of challenging problems[17,27,29–31].

For the dataset described in section-emulating aqueous aerosol chemistry involves many input variables. In our first attempt to develop an emulator for the two IEPOX-SOA components (IEPOX organosulfate: Iepoxos and 2-methyltetrols: Tetrol), we developed a fully connected deep neural network of six layers with a batch normalization layer model to take 165 inputs (size-bin dependent variables $(8 \times 20) = 160$, exogenous variable $= 5$) and predict 40 outputs (two outputs per size bin $(2 \times 20) = 40$). We encountered two major limitations with this approach. First, due to the high input dimensions (165 input variables) and the output dimensions (40 output), the simplified network architecture did not show good accuracy with the test data. A search of more complex architectures could have addressed this limitation, but embedding the complex architecture in the WRF-Chem is challenging. The second limitation was that even with the best possible simplified network trained in a high-dimensional setting; when embedded with WRF-Chem as an emulator for the aqueous aerosol module, the emulator faced significant challenges in ensuring the mass balance to provide a converged solution. Note that in this first attempt, we use the actual concentrations of IEPOXOS and Tetrol as our prediction targets for each bin.

To address the challenges encountered with our first approach, we trained 20 individual DNN models for each of the 20 individual size-bin IEPOX-SOA dataset. This addressed the dimensional explosion problem, since each of the 20 DNN models requires a smaller number of input (13) and output (2) features corresponding to a bin size, compared to a single DNN model that is tasked with predicting IEPOX-SOA in all the 20 size bins concurrently. To address the challenges associated with mass conservation, we trained our model to predict the fluxes (i.e., the change in the concentrations of target variables per unit time) of Iepoxos and Tertrol rather than actual concentration. The work of ref. [18] showed that emulating fluxes rather than absolute concentrations obeys mass conservation much better than emulating the concentrations. A schematic of the emulator architecture used in this study for the WRF-Chem DNN-embedded model is shown in Supplementary Fig. 3. Each DNN model is presented with the individual bin input and target data, listed in Table 1. Once the models were trained on their respective training data, we embedded all 20 models (one for each bin) in WRF-Chem to replace the aqueous SOA calculation routine.

We use the TensorFlow-Keras Python library to implement and train the DNN model used in this work[32,33]. All DNNs used feed-forward neural networks, with each densely connected layer and batch normalization layers. The DNN model in a supervised learning setup is used to describe the relationship between input and output variables. We construct our DNN models for all the bins with 3 hidden layers, 64 neurons in each layer with the rectified linear (relu) activation. In our training procedure for all the model we use the SGD optimizer[34,35] with a learning rate of 0.0058, momentum of 0.9 and batch size of 256. We use the mean square error (mse) as a choice of loss function for training all the DNN models. To avoid over-fitting and other DNN training issues we chose our DNN hyper-parameters, including the number of neurons, activation function, learning rate, and batch size, using the ray-tune[36] hyperparameter optimization python library. We trained all the size-bin DNN models concurrently using 8 NVIDIA A100 GPUs using the ray library[37]. It took 6.8 h for training all the size-bin models each of which were trained for 30 epochs.

Integrating machine learning with WRF-Chem

The ultimate objective in this work is to develop a surrogate model for SOA (iepoxos, tetrol) predictions to bypass the computationally expensive aqsoagamma solver in WRF-Chem Default. As described in the previous sections, we generated the training data using WRF-Chem predictions. Once the model is trained and evaluated offline, we integrated the DNN model into the WRF-Chem code. WRF-Chem is mostly written using FORTRAN90 but machine-learning models are developed using Python, therefore, we translated the deep-learning models into an ascii text format which is then read into the WRF-Chem FORTRAN code easily. This translation and embedding is done using the FotranToKeras bridge[38] library. We integrate the DNN models into the WRF-Chem module *module_mosaic_therm.F* as *aqsoa_gamma_DNN* subroutine.

At runtime of the WRF-Chem DNN the *module_mosaic_therm.F* module is called at every grid cell and timestep to invoke the DNN-based subroutine, similar to the reference aqsoagamma solver within the WRF-Chem Default model. The subroutine creates the desired input vectors shape and normalizes them before feeding them as inputs to the embedded deep-learning model. To ensure the stability of the coupled run, the inputs are normalized and bound between 0 and 1. The subroutine then receives the predictions for the flux of Iepoxos and Tetrol by invoking the trained DNN. These outputs are then denormalized to obtain predictions for the Iepoxos and Tetrol fluxes in the computational domain. The subroutine also integrates these fluxes across the chemistry timestep and adds them as tendencies to the incoming concentrations of the IEPOX-SOA components (IEPOXOS and Tetrols) in each of the 20 size bins.

DATA AVAILABILITY

All WRF-Chem simulation data used to develop the study are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

All codes used for analyses are available upon reasonable request.

Received: 11 May 2022; Accepted: 9 March 2023;
Published online: 31 March 2023

REFERENCES

1. Seinfeld, J. H. et al. Improving our fundamental understanding of the role of aerosol-cloud interactions in the climate system. *Proc. Natl Acad. Sci. USA* **113**, 5781–5790 (2016).
2. Shrivastava, M. et al. Recent advances in understanding secondary organic aerosol: implications for global climate forcing. *Rev. Geophys.* **55**, 509–559 (2017).
3. Shiraiwa, M. et al. Aerosol health effects from molecular to global scales. *Environ. Sci. Technol.* **51**, 13545–13567 (2017).
4. Shrivastava, M. et al. Global long-range transport and lung cancer risk from polycyclic aromatic hydrocarbons shielded by coatings of organic aerosol. *Proc. Natl Acad. Sci. USA* **114**, 1246–1251 (2017).
5. Zhu, C.-S. et al. The rural carbonaceous aerosols in coarse, fine, and ultrafine particles during haze pollution in northwestern china. *Environ. Sci. Pollut. Res.* **23**, 4569–4575 (2016).
6. Guenther, A. B. et al. The model of emissions of gases and aerosols from nature version 2.1 (megan2.1): an extended and updated framework for modeling biogenic emissions. *Geosci. Model. Dev.* **5**, 1471–1492 (2012).
7. Surratt, J. D. et al. Reactive intermediates revealed in secondary organic aerosol formation from isoprene. *Proc. Natl Acad. Sci. USA* **107**, 6640–6645 (2010).
8. Claeys, M. et al. Formation of secondary organic aerosols through photooxidation of isoprene. *Science* **303**, 1173–1176 (2004).
9. Gaston, C. J. et al. Reactive uptake of an isoprene-derived epoxydiol to submicron aerosol particles. *Environ. Sci. Technol.* **48**, 11178–11186 (2014).
10. Zhang, Y. et al. Joint impacts of acidity and viscosity on the formation of secondary organic aerosol from isoprene epoxydiols (iepox) in phase separated particles. *ACS Earth Space Chem.* **3**, 2646–2658 (2019).

11. Shrivastava, M. et al. Tight coupling of surface and in-plant biochemistry and convection governs key fine particulate components over the amazon rainforest. *ACS Earth Space Chem*. **6**, 380–390 (2022).
12. Octaviani, M. et al. Modeling the size distribution and chemical composition of secondary organic aerosols during the reactive uptake of isoprene-derived epoxydiols under low-humidity condition. *ACS Earth Space Chem.* **5**, 3247–3257 (2021).
13. Keller, C. A. & Evans, M. J. Application of random forest regression to the calculation of gas-phase chemistry within the geos-chem chemistry model v10. *Geosci. Model. Dev.* **12**, 1209–1225 (2019).
14. Kelp, M. M., Jacob, D. J., Kutz, J. N., Marshall, J. D. & Tessum, C. W. Toward stable, general machine-learned models of the atmospheric chemical system. *J. Geophys. Res. Atmos.* **125**, e2020JD032759 (2020).
15. Zaveri, R. A. & Peters, L. K. A new lumped structure photochemical mechanism for large-scale applications. *J. Geophys. Res. Atmos.* **104**, 30387–30415 (1999).
16. Zaveri, R. A., Easter, R. C., Fast, J. D. & Peters, L. K. Model for simulating aerosol interactions and chemistry (mosaic). *J. Geophys. Res. Atmos.* **113**, 1–29 (2008).
17. Reichstein, M. et al. Deep learning and process understanding for data-driven earth system science. *Nature* **566**, 195–204 (2019).
18. Sturm, P. O. & Wexler, A. S. Conservation laws in a neural network architecture: enforcing the atom balance of a Julia-based photochemical model (v0. 2.0). *Geosci. Model. Dev. Discuss*. **15**, 3417–3431 (2022).
19. Shrivastava, M. et al. Urban pollution greatly enhances formation of natural aerosols over the amazon rainforest. *Nat. Commun.* **10**, 1–12 (2019).
20. Zhuang, F. et al. A comprehensive survey on transfer learning. *Proc. IEEE* **109**, 43–76 (2020).
21. Kirkpatrick, J. et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl Acad. Sci. USA* **114**, 3521–3526 (2017).
22. Grell, G. A. et al. Fully coupled "online" chemistry within the wrf model. *Atmos. Environ*. **39**, 6957–6975 (2005).
23. Zhao, B. et al. High concentration of ultrafine particles in the amazon free troposphere produced by organic new particle formation. *Proc. Natl Acad. Sci. USA* **117**, 25344–25351 (2020).
24. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *nature* **521**, 436–444 (2015).
25. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547 (2018).
26. Mjolsness, E. & DeCoste, D. Machine learning for science: state of the art and future prospects. *science* **293**, 2051–2055 (2001).
27. Runge, J. et al. Inferring causation from time series in earth system sciences. *Nat. Commun*. **10**, 1–13 (2019).
28. Goodfellow, I., Bengio, Y. & Courville, A. *Deep Learning* (MIT Press, 2016).
29. Ham, Y.-G., Kim, J.-H. & Luo, J.-J. Deep learning for multi-year enso forecasts. *Nature* **573**, 568–572 (2019).
30. Yuan, Q. et al. Deep learning in environmental remote sensing: achievements and challenges. *Remote. Sens. Environ*. **241**, 111716 (2020).
31. Camps-Valls, G., Tuia, D., Zhu, X. X. & Reichstein, M. *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences* (John Wiley & Sons, 2021).
32. Chollet, F. keras, GitHub. https://github.com/fchollet/keras (2015).
33. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous systems. www.tensorflow.org (2015).
34. Robbins, H. & Monro, S. A stochastic approximation method. *Annal. Math. Stat* **22**, 400–407 (1951).
35. Bottou, L., Curtis, F. E. & Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Rev.* **60**, 223–311 (2018).
36. Liaw, R. et al. Tune: a research platform for distributed model selection and training. Preprint at https://arxiv.org/abs/1807.05118 (2018).
37. Moritz, P., Nishihara, R., Wang, S., Tumanov, A., Liaw, R., Liang, E., Elibol, M., Yang, Z., Paul, W., Jordan, M. I. & Stoica, I. Ray: A distributed framework for emerging {AI} applications. In 13th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 2018) (pp. 561–577).
38. Ott, J. et al. A Fortran-Keras deep learning bridge for scientific computing. *Sci. Program* **2020**, 1–13 (2020).

ACKNOWLEDGEMENTS

This research was supported by the U.S. Department of Energy (DOE) Office of Science and Office of Biological and Environmental Research (BER) through the Early Career Research Program. The Pacific Northwest National Laboratory (PNNL) is operated for the DOE by Battelle Memorial Institute under contract DE-AC06-76RL01830. Computational resources for the simulations were provided by PNNL Institutional Computing and EMSL (a DOE Office of Science User Facility sponsored by the Office of Biological and Environmental Research located at PNNL). This research used resources from the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

AUTHOR CONTRIBUTIONS

M.S. designed the study and performed WRF-Chem simulations for IEPOX-SOA, H.S. trained the DNN model, performed evaluations between WRF-Chem outputs and the embedded WRF-Chem-DNN model and generated visualizations, and B.S. contributed to embedding the DNN model within WRF-Chem. M.S. and H.S. wrote the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41612-023-00353-y.

Correspondence and requests for materials should be addressed to Manish Shrivastava.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.