npj Climate and Atmospheric Science https://doi.org/10.1038/s41612-026-01371-2

Article in Press

High-resolution aerosol liquid water content in the contiguous United States using machine learning

Received: 27 July 2025

Accepted: 25 February 2026

Cite this article as: Zhang, B., Yin, L., Yang, Y. *et al.* High-resolution aerosol liquid water content in the contiguous United States using machine learning. *npj Clim Atmos Sci* (2026). https://doi.org/10.1038/s41612-026-01371-2

Bingqing Zhang, Lifei Yin, Yuhan Yang, Hongyu Guo, Lu Xu, Qian Di, Yaguang Wei, Jing Wei, Da Pan, Joel Schwartz, Nga L. Ng, Rodney J. Weber & Pengfei Liu

High-Resolution Aerosol Liquid Water Content in the Contiguous United States using Machine Learning

Bingqing Zhang (1), Lifei Yin (1), Yuhan Yang (1), Hongyu Guo (2,3), Lu Xu (4), Qian Di (5), Yaguang Wei (6), Jing Wei (7), Da Pan (8), Joel Schwartz (9), Nga L. Ng (1, 8, 10), Rodney J. Weber (1), Pengfei Liu* (1)

(1) School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia 30332, USA
(2) School of Environmental Science and Engineering, Sun Yat-sen University, Guangzhou, Guangdong 510275, China
(3) Guangdong Provincial Key Laboratory of Environmental Pollution Control and Remediation Technology, Guangzhou 510006, China
(4) Department of Energy, Environmental and Chemical Engineering, Washington University in St. Louis, St. Louis, Missouri 63130, USA
(5) Vanke School of Public Health, Tsinghua University, Beijing 100190, China
(6) Department of Environmental Medicine and Climate Science, Icahn School of Medicine at Mount Sinai, New York, NY 10029, USA
(7) Department of Atmospheric and Oceanic Science, Earth System Science Interdisciplinary Center, University of Maryland, College Park, Maryland 20742, USA
(8) School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA
(9) Department of Environmental Health, Harvard TH Chan School of Public Health, Boston, Massachusetts 02115, USA
(10) School of Chemical and Biomolecular Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA
(11) School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332, USA

E-mail: pengfei.liu@eas.gatech.edu

Submitted: February 25, 2026

npj Climate and Atmospheric Sciences

*To Whom Correspondence Should be Addressed

Abstract

Aerosol liquid water content (ALWC) plays an important role in climate and public health by influencing aerosol formation, chemical composition, and toxicity. However, ALWC remains sparsely measured and poorly constrained across space and time, despite its large variability. In this study, we derived a high resolution (1km × 1km, daily) ALWC dataset for the contiguous US from 2000 to 2019. The dataset was generated by training machine learning (ML) models on outputs from a chemical transport model (GEOS-Chem) to capture the thermodynamic relationships between ALWC and relevant predictors, then applying these relationships to high-resolution, biased-corrected input datasets. Compared with GEOS-Chem simulations, the ML-based dataset better captures daily variations and spatial heterogeneity in ALWC. The predicted ALWC levels are highest in the Midwest US and lowest in the Western US, largely driven by regional differences in $PM_{2.5}$ concentration, chemical composition, temperature, and relative humidity. Over the study period, ALWC declined significantly across most regions, driven primarily by the reduction in sulfate. We further demonstrate that ALWC provides a physically meaningful constraint for interpreting variability in water-soluble iron, a health-relevant fraction of aerosol metals, highlighting the potential value of this dataset for future studies of aerosol toxicity and epidemiological exposure.

Introduction

Water can be absorbed by aerosol particles composed of soluble species, contributing to a substantial fraction of aerosol volume, particularly at relative humidity (RH) above 60%[1]. Aerosol water influences aerosol mass and composition by affecting the gas-particle partitioning of semi-volatile species[2,3], promoting multiphase reactions that lead to secondary aerosol formation[4,5], and enhancing the solubility of trace metals[6]. Furthermore, aerosol water can largely amplify light scattering, enhancing aerosol cooling effects and reducing visibility[7,8].

Despite its abundance and importance, aerosol water is not routinely measured in field studies due to the lack of direct measurement techniques. Instead, several studies have estimated aerosol liquid water content (ALWC) indirectly based on aerosol hygroscopicity and particle number size distributions (PNSD) [1,9,10], or retrieve it from optical measurements such as light scattering enhancement[11,12] and lidar measurements[13]. Another indirect method involves using thermodynamic models such as ISORROPIA-II[14], which rely on the inputs of aerosol component concentrations and meteorological conditions. However, field-based studies have typically been limited to short-term and location-specific measurements, while ALWC levels can vary largely across times or locations[15,16]. To estimate ALWC at broader spatial and temporal scales, chemical transport models (CTMs), such as CMAQ or GEOS-Chem, have been employed[17,18]. However, biases in simulated aerosol composition could propagate into errors in simulated ALWC[19,20]. More recently, studies have utilized machine learning (ML) techniques to fuse model simulations, observations, and satellite data to generate aerosol mass[21,22], components[23,24] and meteorological data[25,26], with reduced bias and high spatial resolution. However, those datasets are often limited to daily (for meteorological data) or even annual temporal resolution (for aerosol components), which hinders their ability to directly estimate ALWC that is dependent on variability at finer temporal scales. For instance, Pan et al. [27] showed that neglecting diurnal variations in temperature and RH led to underestimation of particle-phase partitioning and thus ALWC (see their Figures S21 and S22).

In this study, we developed a high spatiotemporal resolution dataset of ALWC for the contiguous US (CONUS) at a 1 km × 1 km daily resolution spanning 2000 to 2019. This dataset was generated by integrating high-spatial-resolution datasets of $PM_{2.5}$ mass, composition, meteorological conditions, and GEOS-Chem simulations using ML approaches. The ML models were trained using ALWC calculated within GEOS-Chem under a thermodynamic equilibrium assumption, based on simulated aerosol composition and meteorological conditions. While biases in GEOS-Chem simulated ALWC primarily arise from biases in simulated aerosol inputs, the underlying thermodynamic relationships linking ALWC to aerosol composition and meteorology are physically based. By learning the physically based relationship between ALWC and related variables, the ML models were applied to higher-resolution and less-biased inputs to generate spatially continuous ALWC estimates.

We validated the ALWC dataset by comparing it with estimates derived from field-measured aerosol composition and benchmarking it against GEOS-Chem simulations. Subsequently, we analyzed the

spatial and temporal variations of ALWC and identified its key controlling factors. Finally, we illustrate the importance and potential applications of the ALWC dataset by examining its role in aerosol iron solubility from both theoretical and observational perspectives, and by discussing implications for future studies.

Results

Machine learning model performance

We selected daily median ALWC as the ML prediction output, rather than daily mean values (see Methods), because ALWC increases exponentially with RH, especially when RH exceeds 80%[1,28]. Consequently, daily mean values can be disproportionately influenced by short periods of very high RH, often occurring at night, whereas the median provides a more stable and representative measure of overall daily ALWC. However, daily median ALWC can still be influenced by sustained periods of high RH. For instance, GEOS-Chem simulates daily median ALWC values exceeding 100 $\mu g/m^3$ under certain conditions (Figure S1), primarily when RH remains above 95% for more than 12 hours within a day. Such extreme cases account for ~1% of the total dataset and therefore have limited impact on overall ALWC estimates.

For training, all input variables and training targets were derived from GEOS-Chem simulations at 0.5° × 0.625° resolution, as direct observations remain too sparse for direct ML training. The ML models were trained to learn the physically based thermodynamic relationship between ALWC, aerosol composition, and meteorology, as represented by ISORROPIA-II calculations in GEOS-Chem. For prediction, bias-corrected, high-spatial-resolution predictors with coarser temporal resolution (Figure 1a; Table S1) were used as inputs, with simulated daily median ALWC as the prediction target. During training, hourly GEOS-Chem outputs were temporally aggregated to match the resolution of the available high-resolution input datasets used in prediction, reflecting practical data limitations rather than a modeling choice. Although sub-daily meteorological variability and seasonal variability in $PM_{2.5}$ composition are not explicitly resolved in the inputs, their effects are implicitly encoded in the GEOS-Chem-derived targets. This framework enables the ML models to recover unresolved temporal variability from daily and annual predictors, while partially accounting for sub-daily temperature variability via daily maximum and minimum temperatures.

Three tree-based ML models including random forest (RF), extreme gradient boosting (XGB), and light gradient boosting machine (lightGBM) were trained and cross-validated using 80% of the GEOS-Chem simulations for the selected years (i.e., 2000, 2005, 2010, 2015, 2020), and additional out-of-sample validation was performed using the remaining 20% from each year as the testing set (Methods). Among all models tested, the RF model performed slightly better, achieving the highest correlation coefficient (R^2), lowest root mean square error (RMSE) and mean absolute error (MAE) in both cross-validation and out-of-sample validation (Table S2, Figure S1). The XGB model and the LightGBM model show slightly lower R^2 and higher RMSE and MAE. To examine the potential influence of data imbalance in

the training dataset, we further evaluate model performance across the ALWC distribution. Specifically, we divide the testing set into four percentile-based groups and calculate the normalized RMSE (nRMSE) and normalized MAE (nMAE) for each group. Similar levels of nRMSE and nMAE across groups in all ML model results suggest that, even if the target distribution is uneven, model performance remains consistent across different ALWC levels (Table S2). These results together demonstrate that ML models can reasonably capture the relationship between daily median ALWC and the selected input variables.

Feature importance calculated for each ML model indicates that daily mean $PM_{2.5}$ concentrations and RH are the primary contributors to ALWC predictions across all three ML models, followed by the daily maximum temperature (Figure S2). The relatively minor contribution of individual $PM_{2.5}$ components (i.e., SO_4^{2-}, NO_3^-, NH_4^+ and OC) is likely attributable to their coarse temporal resolution, as only annual high-resolution data were available for these species. Nevertheless, their influence is indirectly reflected in the overall importance of total $PM_{2.5}$ levels. While other features contributed less, they still played a role in the predictions. In this study, we primarily focus on the RF model results due to its better performance.

Once trained, the ML model offers a computationally efficient alternative that produces results in near real-time given the required inputs. We applied the trained RF model to the best-available high-spatial-resolution datasets at 1 km × 1 km resolution (Figure 1b) to generate the final prediction. These input datasets, largely developed in previous studies (Table S1), incorporate extensive observational constraints to reduce biases and achieve higher spatial resolution, resulting in less-biased and finer-scale ALWC predictions that are largely independent of the model-driven and biased GEOS-Chem-simulated ALWC. The final output (i.e., a 1 km × 1 km, daily ALWC dataset) was further validated using the "observed" ALWC, derived from ISORROPIA-II for the inorganic fractions and κ-Köhler theory for the organic fractions with measured aerosol component concentrations as inputs (see Methods).

To evaluate our ALWC dataset, we compared it with three datasets of "observed ALWC" (Methods). The first dataset is from the Southeastern Aerosol Research and Characterization (SEARCH) network[29-31], which provides daily ALWC estimates at eight sites in the southeastern US from 2001 to 2016. As a benchmark, we also compared GEOS-Chem simulated ALWC with observed ALWC for the years when GEOS-Chem simulations were available (i.e., 2005, 2010, 2015). Although the correlation with observations on daily basis remains relatively week, both GEOS-Chem simulations and ML predictions captured the overall ALWC variations, with scatter points broadly distributed around the 1:1 line (Figure S3). The ML model achieved lower errors, as indicated by smaller RMSE values. For years without GEOS-Chem simulations, the ML model maintained similar performance, with stable R^2 and RMSE values (Figure S4). While correlations at the daily scale remain relatively weak, broader agreement becomes more apparent in long-term trends. Observed annual mean ALWC values show a declining trend after 2005, which is captured by both GEOS-Chem simulations and ML model predictions (Figure S5). From 2000 to 2005, both models suggest an increasing trend, although the sparse observational data during this period (fewer than 100 available days out of 365 annually) make it challenging to

calculate annual averages and confirm this trend. Throughout the study period, GEOS-Chem tends to underestimate ALWC levels at most sites, especially in the early years, whereas the ML model consistently predicts higher ALWC levels that are closer to observations (Figure S5). However, the SEARCH sites exhibit limited spatial variability in ALWC, with no substantial contrast observed between urban-influenced locations (e.g., JST, BHM) and more remote sites (e.g., CTR, OAK, OLF). This may reflect the large contribution of regionally homogeneous biogenic aerosols in the southeastern US[32]. Consequently, this dataset provides limited opportunity to assess the benefits of the ML model's higher spatial resolution.

To extend the validation to a broader spatial scale, we compared our ML results with a second dataset consisting of 23 field campaign observations distributed across the CONUS (Table S3, Figure S6b). Since most campaigns lasted only one to five weeks, we focused on comparing daily variations and overall distributions. Compared with GEOS-Chem simulations, the ALWC distributions predicted by ML were closer to observed values in 17 out of 23 campaigns (Figure s6a). In addition, while GEOS-Chem simulations showed moderate correlations with observed daily ALWC at three campaigns in the west coastal and southern US, ML predictions achieved higher correlations and lower RMSE values for the same campaign. Moreover, ML results demonstrated moderate to high correlation at all sites across all periods, suggesting the capability in capturing the daily variations across different regions (Figure S6 c-d). Note that on several days during the campaigns, observations indicate extremely high ALWC levels while ML predictions suggest more moderate levels, or vice versa. These discrepancies are primarily due to differences in the RH datasets used: MERRA2 (used for calculating observed ALWC, as it provides hourly-scale RH; see Methods) versus PRISM (used for ML predictions). Such mismatches are particularly pronounced when one dataset reports RH > 95% while the other indicates more moderate values. ALWC estimates on these high-RH days are subject to greater uncertainty, and caution is warranted when interpreting them. Nevertheless, these instances are relatively rare, and both MERRA2 and PRISM show overall good agreement with RH measurements from ground-based weather monitoring stations across the contiguous United States (Table S4).

The second dataset provided validation for sites at different locations; however, its spatial coverage remained sparse, and the campaigns only captured intermittent periods within our study timeframe. To further evaluate ALWC over broader spatial and temporal scales, we utilized a third dataset consisting of aerosol composition measurements compiled from multiple national air quality monitoring networks[27]. Although this dataset has coarser spatial (~50km) and temporal (biweekly) resolution than the previous two datasets, it is well suited for validating long-term ALWC trends across a wider geographic area. As shown in Figure S7, sites in all regions except Western US exhibit good correlations with observations. More importantly, ML predictions demonstrate substantially reduced biases compared with GEOS-Chem simulations, particularly at sites in the Midwestern and eastern US where ALWC levels are high, as indicated by lower RMSE values. While this dataset captures ALWC contributions from inorganic species only, whereas the ML predictions account for both inorganic and

organic species, the organic contribution is relatively minor and insufficient to explain the discrepancy between GEOS-Chem results and observations (Figure S7b). Instead, the large discrepancies are primarily driven by biases in GEOS-Chem-simulated aerosol species, which are effectively reduced in ML predictions (Figure S7c). ML predictions in the western US exhibit lower R^2 and higher RMSE, suggesting greater uncertainty at lower ALWC level cases. Despite this, ML results successfully capture the long-term ALWC trend when compared with observations, as shown in Figure S7e.

The reduced bias in the ML results relative to GEOS-Chem simulations could be attributed to the use of less-biased SNA aerosol (i.e., SO_4^{2-}, NO_3^-, NH_4^+) and total $PM_{2.5}$ mass in ML models inputs (Figures S8-S9). When GEOS-Chem overestimates SNA aerosols, leading to overestimates of ALWC, the ML results typically show lower levels of ALWC than GEOS-Chem simulations (i.e., more points falling into the lower left quadrant compared with upper left quadrant in Figure S10). Conversely, when GEOS-Chem underestimates SNA aerosols and thus underestimates ALWC, ML results tend to produce higher ALWC (more points in the upper right quadrant compared with the lower right in Figure S10). These results imply the ML results can potentially correct bias in GEOS-Chem simulated ALWC that were caused by inaccuracies in the simulated aerosol components, through the use of observation-constrained, less-biased datasets as inputs (Figures S8-S9).

Spatial distributions

Figure 2(a) shows the spatial distribution of predicted ALWC averaged from 2000 to 2019, with spatial distributions for individual years presented in Figures S11-S13. Overall, ALWC in the CONUS exhibits a consistent east-middle-west gradient through the study period. The highest ALWC values are observed in the midwestern US, with elevated levels also occurring in the northeastern and southern US. In contrast, lower levels are observed in the west coastal US, with localized hotspots in California and the Pacific Northwest, and the lowest levels occurring in the interior western US. This spatial pattern largely reflects aerosol mass concentrations, resulting in similar spatial distributions between ALWC and $PM_{2.5}$ (Figure S14). However, while $PM_{2.5}$ levels are comparable across eastern US, ALWC levels are higher in the north than in the south. This discrepancy is primarily driven by the differences in RH, with regional-averaged RH differing by approximately 4% annually and larger on a daily scale (Figure S15). Aerosol composition also plays a role by affecting the aerosol hygroscopicity, as the Midwest and northern regions contain a higher fraction of hygroscopic sulfate and nitrate aerosols, whereas the southern regions has a greater proportion of less-hygroscopic organic aerosols (Figure S16).

To further examine the spatial distributions of ALWC on a finer scale, we plotted ALWC and its driving factors in six metropolitan statistical areas (MSAs; Figure 2b-g). The high-resolution dataset reveals heterogeneities between densely populated urban cores and the less populated surrounding areas, with differences reaching as large as 4-5 $\mu g/m^3$ within the selected MSAs. These contrasts are partially driven by elevated $PM_{2.5}$ concentrations in urban cores due to higher population density and emission intensity. However, ALWC spatial distributions within MSAs do not always mirror $PM_{2.5}$ patterns, as they are also strongly modulated by RH and temperature. For example, in addition to urban cores, coastal regions

such as Seattle (Figure 2b), Houston (Figure 2f), and Los Angeles (Figure 2g) also exhibit higher ALWC levels due to proximity to the ocean and associated higher RH. The high-resolution data also capture urban heat island (UHI) effects, where urban cores experience higher temperatures and consequently lower RH. For example, in the Pittsburgh MSA (Figure 2d) and St. Louis MSA (Figure 2e), urban temperatures are approximately 1 K higher than surrounding areas, resulting in a 2% difference in annual RH. As a result, the highest ALWC levels within an MSA do not always occur in urban cores but may instead be found in suburban areas, where $PM_{2.5}$ levels are slightly lower, but temperatures are cooler and RH is higher. This pattern aligns with previous studies of UHI effects, which have reported similar contrasts between urban and suburban ALWC levels [33]. Overall, ALWC distribution is jointly shaped by aerosol mass concentrations, hygroscopicity, and meteorological conditions, with influences spanning regional to local scales.

Long-term trends and their driving factors

ALWC levels across CONUS showed an overall decreasing trend during the study period, dropping from 6.02 $\mu g/m^3$ in 2000 to 3.37 $\mu g/m^3$ in 2019. Urban areas exhibit a more pronounced decline compared to non-urban areas. The decreasing rate remains relatively stable annually and across different seasons, though with greater interannual variability in winter during the first decade of the study (Figure 3). The large decrease in ALWC is likely driven by the decreasing trend of sulfate, which is highly hygroscopic and a major contributor to both aerosol dry mass[34] and ALWC[12]. Mean sulfate levels across the CONUS decreased from 2.23 $\mu g/m^3$ in 2000 to 0.73 $\mu g/m^3$ in 2019, similar to the decreasing rate of ALWC, while other key aerosol components (i.e., NO_3^- and OC) showed relatively smaller declines (Figure 3 c-d). Temperature and RH during the study period showed slightly increasing trends of ~1K and ~2% over the 21 years, respectively, which likely had limited effects on the ALWC levels (Figure 3e-f).

To further analyze the long-term trend of ALWC and its driving factors, we performed grid-cell-wise linear regressions over the study period and characterized trends using the regression slope, which is less sensitive to year-to-year variabilities than simple differences. For aerosol species (SO_4^{2-}, NO_3^- and OC) available only as annual means, regression was applied to annual values (20 data points). For variables available at daily resolution (ALWC, temperature, and RH), monthly mean anomalies were used. Monthly climatologies were calculated over 2000–2019 and subtracted from monthly means prior to regression, thereby removing the seasonal cycle and isolating long-term trends. As shown in Figure S17, ALWC levels show a robust decreasing trend in the Midwestern US, eastern, and southern US, and some regions in California, with decline rates reaching up to ~0.6 $\mu g/m^3/a$ in some areas such as Indiana, Ohio and southern Michigan. ALWC shows weak increasing trends in some regions of the northwestern and western US; however, these trends are not statistically significant ($R^2<0.4$, $p\sim0.05$), indicating large interannual variability.

When averaged across the CONUS, RH and temperature exhibit only minor increasing trends (Figure 3), despite being statistically significance at the grid scale (Figure S17). However, the small magnitude

of these changes, together with the observation that larger meteorological trends tend to occur in regions with less pronounced ALWC decreases, suggests that meteorological conditions are unlikely to be the primary drivers of the long-term decline in ALWC. In contrast, SNA aerosols show robust decreasing trends, especially in regions with pronounced ALWC declines (Figures S17). SO_4^{2-} decreased significantly across most of the CONUS, with the largest reduction in the eastern US, consistent with large SO_2 emission reduction from power plants[35]. NO_3^- exhibits less pronounced decline, reflecting the competing effects of reduced NO_x emissions versus lower aerosol acidity and higher NH_3 emission, which favor the partition of nitrate into the particle phase[36]. OC exhibits a moderate decline in the southeastern US, driven primarily by the decrease of anthropogenic emissions such as vehicle emissions and residential fuel burning, while emissions from natural sources such as wildfire and biogenic processes remain relatively stable[37]. Overall, the largest reductions in ALWC occurred in regions with large decreases in SO_4^{2-} and NO_3^-, particularly in the eastern US and California. These results highlight the critical role of inorganic species in driving ALWC declines in these areas, consistent with findings from previous studies[38].

Implications on aerosol iron solubility

Metal-containing particles, such as iron (Fe), have been linked to adverse health outcomes[39-41]. Across the CONUS, mineral dust is the dominant source of Fe, although contributions from fossil fuel combustion and biomass burning are non-negligible in certain regions[42,43]. Previous studies have reconstructed high-spatial-resolution Fe datasets based on fusing ground measurements, satellite data, reanalysis data and other relating variables using machine learning models[44], enabling detailed assessments of Fe exposure in epidemiological studies. However, subsequent epidemiological analyses have not identified clear or consistent associations between total Fe and health outcomes such as cardiovascular or respiratory diseases[45,46]. This lack of robust associations does not necessarily imply low toxicity of Fe. One plausible explanation is that existing datasets represent total Fe mass, whereas the health-relevant fraction (i.e., soluble or bioavailable Fe) typically constitutes only a small proportion (often less than 10%).

Extensive studies have demonstrated that Fe solubility is strongly influenced by aerosol acidity [6,47-49], suggesting that acid-promoted dissolution is an important pathway for converting total Fe into soluble Fe. However, there are currently no observationally constrained datasets that characterize soluble Fe variability across large spatial or temporal scales. Moreover, several studies have suggested that even when acid-promoted dissolution dominates, aerosol pH alone may be insufficient to explain observed variability in Fe solubility, and that ALWC may play an additional and important role [6]. Here we illustrate the importance of ALWC and the potential application of our dataset through additional analysis. We first present a qualitative discussion on how ALWC influences Fe solubility from a theoretical perspective, and then exam the potential impact of ALWC using a previously published observational dataset[50]. This dataset represents the only high-quality observational dataset with sufficient detail available for this purpose. Additional information about the observational campaign is

provided in Text S1, and full methodological details can be found in the original publication[50]. Finally, we discuss how the ALWC dataset developed in this study can be applied in future investigations of Fe solubility and related health impact studies.

For practical purposes in large-scale modeling or interpretation of observational data, a simplified first-order approximation is commonly used to estimate water-soluble Fe (WS-Fe) formation rate through the acid-promoted dissolution pathway.

$$\frac{d}{dt}[WS-Fe] \propto R_{Fe} \times [Fe_{insol}] \propto K(T) \times \alpha[H^+]^m \times A \times [Fe_{insol}] \quad (1)$$

where R_{Fe} represents the dissolution rate, which depends approximately linearly on the temperature-dependent dissolution rate coefficient $K(T)$, proton activity $\alpha[H^+]^m$ and aerosol surface area A; Proton activity is typically approximated as 10^{-pH} by assuming unity of activity coefficient [51]. The empirical reaction order m generally ranges from 0.1 to 1, depending on the mineralogy of the Fe-containing particles[51]. $[Fe_{insol}]$ represents the concentration of insoluble Fe in aerosol particles. Because the soluble fraction of aerosol Fe is typically less than ~10%, the depletion of insoluble Fe during dissolution is small, and $[\mathrm{Fe_{insol}}]$ can be treated as approximately constant.

The atmospheric concentration of WS-Fe produced through this pathway can therefore be estimated as the time-integrated formation minus removal by deposition:

$$[WS-Fe] = \int R_{Fe} \times [Fe_{insol}]dt - Loss \quad (2)$$

In this framework, ALWC only has a secondary influence on WS-Fe production through its effect on aerosol pH, since pH depends logarithmically on ALWC and is often buffered by semi-volatile species such as NH_3-NH_4^+ system. A more important role of ALWC is to provide the aqueous medium required for acid-promoted Fe dissolution, thereby regulating whether and how long the reaction can proceed. As a result, variations in ALWC can significantly modulate WS-Fe formation even under similar aerosol pH conditions[6].

Observational data further indicates that aerosol pH alone, although well-established control on Fe solubility, cannot fully explain the large variability observed in ambient measurements. For example, when aerosol pH falls in the range of 1-2, measured Fe solubility spans from ~0.04 to 0.4. The solubility variability decreases with increasing pH (Figure S18). In contrast, ALWC shows a relatively strong linear correlation with Fe solubility, particularly within the pH 1–2 regime, and the correlation remains evident, albeit with a lower slope, at pH 2-3 (Figure S19). Since aerosol pH <3 is representative of most regions in the contiguous US[52], this suggests that ALWC can be a major driver of the variability in Fe solubility under prevailing atmospheric conditions. This behavior is consistent with the role of ALWC in providing the aqueous medium required for acid-promoted dissolution, while pH, which reflects a buffered thermodynamic state, varies relatively weakly on short timescales.

While the above analysis mainly focuses on the acid-promoted pathway, ligand-promoted dissolution

pathway represents another important mechanism[53]. Laboratory and field studies, particularly in marine environments where aerosol pH is relatively higher (i.e., 3–6), have shown that ligands (e.g., oxalate) can enhance Fe solubility [53-55]. However, over continental regions such as the contiguous US, where aerosol pH tends to be lower, acid-promoted dissolution is likely the dominant pathway. Observational studies over land (e.g., in Canada) that simultaneously measured pH, oxalate, and Fe concentrations indicate that the two pathways may operate together in complex ways, but current data are insufficient to disentangle their individual contributions at scale[56].

Our ALWC dataset provides an important missing constraint for understanding the spatial and temporal variability in Fe solubility. For example, if ALWC is accounted for, soluble Fe levels may differ between regions with different ALWC levels, even when total Fe concentrations are similar. Likewise, long-term declines in soluble Fe observed in some regions may be driven by reductions in ALWC, despite relatively stable aerosol pH conditions[57]. These examples suggest that relying solely on aerosol pH may underestimate the true spatiotemporal variability in Fe solubility. Additional observations across a wider range of environmental conditions are needed to further investigate these relationships and to improve predictive models.

We emphasize that current observations remain too limited to fully resolve the mechanistic relationship between aerosol pH, ALWC, and Fe solubility. A robust quantitative model linking these variables is not yet available. However, the high-resolution ALWC dataset developed here is a valuable step forward. It enables more physical informed interpretation of observed Fe solubility patterns and offers a mechanistically relevant variable that can be incorporated into epidemiological studies to examine the interactions between metal solubility and health outcomes. The dataset can also serve as input for future numerical modeling and machine learning efforts aimed at estimating soluble Fe concentrations. Similar reasoning may apply to other redox-active metals, such as copper (Cu), which also undergo aqueous-phase processing and may respond to variations in ALWC. While ligand-promoted dissolution pathways may also be important, further coordinated observations are needed to constrain their role alongside acid-promoted dissolution.

Discussion

To our knowledge, this study represents the first effort to construct high-resolution ALWC dataset across the CONUS at 1 km × 1 km daily resolution. Although we leveraged all available observations to validate our dataset both directly (by comparing with "observed ALWC") and indirectly (by comparing with input datasets such as $PM_{2.5}$ and aerosol components, such as Figure S8), the overall observational coverage remains limited. Expanding ALWC measurements across a wider range of regions and seasons would strengthen the validation process and improve understanding of large-scale spatial variability, urban-suburban contrasts, and diurnal to seasonal patterns.

Beyond the dataset itself, this study demonstrates a generalizable framework for leveraging ML models

to learn non-linear physical and chemical processes, as well as unresolved temporal variations from CTM output. By incorporating high-spatial-resolution, observation-constrained input data, this method allows for effective bias correction and enhanced spatial resolution in the predicted ALWC. In addition, once trained, the ML model can be readily applied to generate ALWC under small perturbations of the input data at minimal computational cost, making it suitable for rapid sensitivity analysis. The method can be readily extended to reconstruct other sparsely observed variables, such as aerosol pH, for which direct ML training is not feasible. Future work to develop high-resolution aerosol pH datasets can further improve predictions of WS-Fe and other trace metals in aerosols. Ultimately, these studies will help better understand the solubility and bioavailability of trace metals and enable robust assessments of their health effects on human populations. More broadly, the implications of this work extend beyond iron solubility, as ALWC plays a central role in governing aerosol scattering and aqueous-phase reactions in fine particulate matter, with relevance to aerosol composition, toxicity, and climate effects.

Methods

Method overview

The methodological framework is designed to learn the relationship between ALWC and its related variables from chemical transport model simulations and then apply the relationships to higher-resolution and less-biased input data to generate final ALWC predictions. The key assumption underlying this approach is that biases in simulated ALWC are primarily driven by biases in simulated aerosol species concentrations. Meteorological variables are taken from reanalysis products that assimilate extensive climate observations and therefore generally agree well with observations, allowing them to be treated as relatively unbiased.

The relationship between ALWC, aerosol composition, and meteorology is physically based, as ALWC is calculated using the thermodynamic equilibrium model ISORROPIA-II[14]. This model is widely applied both within GEOS-Chem and in observational analyses based on field measurements to estimate ALWC from aerosol composition and meteorological conditions. We therefore assume that this thermodynamic relationship remains unchanged across GEOS-Chem simulations, observational applications, and the trained ML model. Consequently, during the prediction process, improvements in ALWC estimates arise from the use of improved input datasets with higher spatial resolution and reduced biases relative to chemical transport model simulations.

ML models

The training dataset is from simulations performed with the GEOS-Chem model (version 12.9.3, 0.5° × 0.625° spatial resolution; https://zenodo.org/records/3974569, last access: June 21, 2024). The target variable is daily median ALWC (calculated based on hourly data). We trained three tree-based ML models, including random forest (RF), extreme gradient boosting (XGB), and light gradient boosting machine (lightGBM) due to their capability of handling large dataset efficiently and delivering robust performance. A schematic diagram of the model training and prediction processes can be found in Figure 1.

The input features for the ML models include variables related to time and location, daily meteorological conditions, daily or annual species concentrations (listed in Figure 1a and Table S1). While our goal was to predict daily ALWC, some input features were only available on an annual scale. The ML models learned daily ALWC variations from the features with daily resolution, while also using the annual-scale variables to capture spatial differences. The target of the ML models was the sum of daily median ALWC contributed by inorganic species ($ALWC_{inorg}$, calculated by ISORROPIA-II[14]) and organic species ($ALWC_{org}$, calculated by Equation 3 based on κ-Köhler theory [58]).

$$ALWC_{org} = \frac{M_{org}\rho_w}{\rho_{org}} \times \frac{\kappa_{org}}{(\frac{1}{RH}-1)} \quad (3)$$

Where M_{org} is the mass concentration of organic matter (OM); ρ_w and ρ_{org} are the density of water

or OM, assumed to be 1g/cm^3 or 1.3g/cm^3, respectively; RH is the relative humidity, and κ_{org} is the hygroscopicity parameter of OM. In this study, we assumed a value of 0.1, consistent with that used in the GEOS-Chem model[59], and supported by chamber experiments [58,60,61] as well as field campaigns conducted across various regions worldwide[62-64]. We expect that the choice of κ_{org} has limited influence on ALWC simulations in GEOS-Chem and ML models, given the relatively low hygroscopicity of organic matter compared to inorganic components and its generally low mass fractions across most of the contiguous US. Sensitivity tests using GEOS-Chem simulations also suggest that organic compounds contribute only marginally to total ALWC (Figure S7b).

We ran the GEOS-Chem nested grid simulation over North America for five years covering the study period (2000, 2005, 2010, 2015, and 2020). We randomly selected 80% of the dataset for each year and combined them as the training dataset. We performed 10-fold cross-validation to test the robustness of the ML models with the selected hyperparameters. The remaining 20% of the data for each year were used separately to perform an out-of-sample test to further evaluate the performance of the final model trained on the entire training dataset. We report the average and the standard deviation of R^2, RMSE, and MAE as evaluation metrics (Table S2). To assess the impact of data imbalance on model performance, we evaluated the trained ML models across different ranges of ALWC values. Specifically, we divided the test dataset into four quantile-based ranges (i.e., 0%-25%, 25%-50%, 50%-75%, 75%-100%), and calculate the normalized RMSE and MAE for each range. Comparing performance across ranges allows us to assess model behavior at different ALWC levels.

We implemented the RF algorithm using the *RandomForestRegressor* function from the python package *sci-kit learn*, the XGB algorithm with *XGBRegressor* function from the python package *xgboost*, and the lightGBM algorithm with *LGBMRegressor* function from the python package *lightgbm*. Hyperparameters for each model were optimized through grid search using the function *GridSearchCV* from *sci-kit learn*. The optimal hyperparameters were as follows: for the RF model, the best configuration was a number of trees =10 with no limitations on the tree depth, further increase the number of trees result in an exponential increase in runtime with nearly the same model performance; for the XGB model, the best configuration was a number of trees=100, tree depth=5, and learning rate=0.1; for the LGBM model, the best configuration was a number of trees=500, tree depth=5, learning rate=0.1. Feature importance for each algorithm was calculated using the build-in functions provided by the respective Python packages.

ALWC predictions

High-resolution input variables from multiple datasets were applied to the trained ML models to predict ALWC at a 1km × 1km resolution for the period 2000-2019 (Figure 1b). Daily maximum temperature (T_{max}) and daily maximum temperature difference (T_{diff}, calculated as the difference between T_{max} and daily minimum temperature T_{min}) from 2000 to 2019 were obtained from Daymet Version 4R1 at a 1km × 1km spatial resolution[25,26], derived primarily by interpolating and extrapolating ground-based observations using statistical modeling techniques. Both T_{max} and T_{min} from Daymet show good

agreement with observations (Tables S5-S6). Although empirical equations are available to derive high-resolution daily mean RH (RH_{mean}) for the Daymet dataset (Equations 4-5), this method did not yield reliable estimates when compared with observations, resulting in correlation coefficient (r) values of 0.43-0.72 and slopes of 0.74-0.83 (Table S4). Instead, we used RH_{mean} calculated from the Parameter-elevation Relationships on Independent Slope Model (PRISM) [65], which provides daily mean dew point temperature (TD_{mean}) and daily mean temperature (T_{mean}) at a 4km × 4km resolution. PRISM derives these data by spatially interpolating meteorological observations through a regression model that applies spatial weighting to account for climatically important landscape features[66]. RH_{mean} calculated from PRISM using Equation 6 exhibited improved agreement with observations compared to Daymet, with r values of 0.71-0.91 and slopes of 0.87-0.99 (Table S4). We linearly interpolated the PRISM data from its original 4 km × 4 km resolution to 1 km × 1 km resolution to match the spatial scale of the other input features.

$$T_{mean(Daymet)} = 0.606 \times T_{max} + 0.394 T_{min} \quad (4)$$

$$RH_{mean\ (Daymet)} = 100\% \times \frac{VP}{b_1 \times e^{\frac{b_2 \times T_{mean}}{b_3 + T_{mean}}}} \quad (5)$$

$$RH_{mean\ (PRISM)} = 100\% \times \frac{a_1 \times e^{\frac{a_2 \times TD_{mean}}{a_3 + TD_{mean}}}}{a_1 \times e^{\frac{a_2 \times T_{mean}}{a_3 + T_{mean}}}} \quad (6)$$

Where VP represents the water vapor pressure in Daymet (pa); empirical constants for calculating Daymet RH_{mean} are b_1=610.78Pa, b_2=17.269, b_3=237.3 °C; the empirical constants for calculating PRISM RH_{mean} are a_1=610.94Pa, a_2=17.625, a_3=243.04°C[67].

Daily mean concentrations of $PM_{2.5}$[21,68] from 2000 to 2019, NO_2[69,70], and daily maximum 8-hour O_3[71,72] from 2000 to 2016 at 1 km × 1 km spatial resolution were from ensemble ML model predictions from previous studies. In short, independent ML models were trained with various input features including satellite and ground-based measurements, land-use terms, chemical transport model simulations, and meteorological variables, and the results were combined by a geographically weighted generalized additive model to obtain the final predictions. Due to lack of more recent data, NO_2 and O_3 data from 2016 were used as inputs for ALWC predictions in 2017 to 2019. Annual mean concentrations of $PM_{2.5}$ components (SO_4^{2-}, NO_3^-, NH_4^+ and OC) from 2000-2019 were predicted using super learning and ensemble weighted averaging of ML models at a spatial resolution of 50m in urban areas and 1km in non-urban area from previous studies[23,73]. The model fused $PM_{2.5}$ component measurements from 987 monitoring sites and hundreds of other predictors, including satellite-derived measurements, chemical transport model simulations, meteorological conditions, land-use data, and other variables. We regridded the dataset to 1 km × 1 km resolution as the input of our ML models.

Observational records

We compiled the best available aerosol component measurements from various ground-based

observations, including monitoring networks and intensive field campaigns, to estimate ALWC and validate our datasets. ALWC contributed by inorganic species were estimated using ISORROPIA-II in reverse mode, due to lack of gas-phase measurements at most sites. ALWC contributed by organics (if measurements of organic species were available) was estimated using Equation 3, assuming κ=0.1 and ρ_{org} =1.3g/cm^3. If RH and temperature data were not available at the measurements of aerosol components, we used values from the MERRA2 reanalysis. We defined the calculated ALWC from measured aerosol components and measured or MERRA2-based meteorology as "observed ALWC" and compared them with ML predictions at each site for validation, although direct measurement techniques for ALWC are currently not available.

The first dataset consists of long-term continuous measurements of $PM_{2.5}$ components at an hourly scale obtained from the SEARCH network, covering the period from 2001 to 2016. This network contains data from eight stations across the southeastern US, including Jefferson Street in Atlanta, Georgia (JST, 33.776°N, 84.413°W), Yorkville in Georgia (YRK, 33.931°N, 85.046°W), North Birmingham in Alabama (BHM, 33.553°N, 86.815°W), Centreville in Alabama (CTR, 32.902°N, 87.250°W), Gulfport in Mississippi (GRP, 30.391°N, 89.050°W), Oak Grove in Mississippi (OAK, 30.985°N, 88.932°W), Pensacola in Florida (PNS, 30.437°N, 87.256°W), and Outlying Landing Field #8 in Florida (OLF, 30.551°N, 87.376°W). The details of each site's condition are discussed in Hansen et al[31] and the details of the continuous measurements of $PM_{2.5}$ components are discussed in Edgerton [30]. We calculated hourly ALWC values and converted them to daily medians, referred to as the "observed ALWC". The primary strength of this dataset is its long duration and high temporal resolution, enabling validation of both daily ALWC variations and long-term trends. However, its main limitation is its restricted spatial coverage, as it is limited to the southeastern US.

The second dataset includes hourly measurements of major aerosol components (i.e., SO_4^{2-}, NO_3^-, NH_4^+, and OC) from 23 field campaigns measured by Aerosol Mass Spectrometers (AMS) and Aerosol Chemical Speciation Monitors (ACSM). This includes 15 campaigns compiled by the Aerosol Mass Spectrometer Global Database[74], one from the Southern Oxidant and Aerosol Study (SOAS), and seven from the Southeastern Center for Air Pollution and Epidemiology (SCAPE) [75,76]. The details of these campaigns are provided in Table S3 and the corresponding references. For most of the campaigns, only aerosol species concentrations were measured, with limited or no simultaneous measurements of RH and temperature. To address this, we used hourly temperature and RH data from MERRA2 in the ALWC calculations. This dataset provides high temporal resolution due to hourly measurements, making it useful for evaluating daily ALWC variations. Additionally, the sites are not limited to the southeastern US, allowing for insights into spatial variations in ALWC. However, most campaigns lasted only a few weeks, limiting its ability to validate the long-term trends.

The third dataset is a compilation of measurements of gaseous and aerosol composition from monitoring networks across the US[27]. This dataset integrates observations from multiple networks including the Clean Air Status and Trends Network (CASTNET), the Interagency Monitoring of Protected Visual

Environments (IMPROVE) network, the US EPA's $PM_{2.5}$ Chemical Speciation Monitoring Network (CSN) and the Ammonia Monitoring Network (AMoN). Because some sites from different networks are located in close proximity, observations within a 50 km radius were averaged to minimize inconsistencies arising from differences in sample collection and measurement methods. The dataset includes measurements of gaseous species (HNO_3 and NH_3) and aerosol species (SO_4^{2-}, NO_3^-, NH_4^+, Cl^-, and other non-volatile cations). To ensure consistency in temporal resolution, all observations were averaged to a biweekly timescale to match the lowest sampling frequency from AMoN. ALWC levels were calculated at hourly scale, with temperature and RH from MERRA2 while keeping species concentrations fixed at the biweekly averages. A previous study has demonstrated that ISORROPIA-II provides reliable estimates of aerosol phase partitioning, supporting its application for ALWC estimation[27]. This dataset provides the broadest spatial coverage among the three, making it valuable for assessing ALWC levels across different regions as well as long-term trends. However, its spatial resolution is lower, as observations within 50 km are averaged, and its temporal resolution is also reduced, as biweekly-averaged aerosol composition was used, meaning that diurnal variations in ALWC are driven solely by meteorological conditions rather than changes in aerosol composition. In addition, this dataset does not include measurements of organic species due to limited measurements available, thus the calculated ALWC reflects contributions from inorganic species only. For comparison with this dataset, data from the exact grid of GEOS-Chem simulations are used, as the spatial resolution (0.5° × 0.625°) is similar to a 50 km window, while ML predictions are averaged over a 25 km radius.

Data availability

Daily mean PM2.5, 8-hour maximum ozone, and NO_2 datasets at 1km×1km resolution are publicly available from NASA Earthdata ($PM_{2.5}$: https://doi.org/10.7927/g2n9-ca10; O_3: https://doi.org/10.7927/5tht-jg22; NO_2: https://doi.org/10.7927/rz28-p167). Annual mean aerosol composition data, including SO_4^{2-}, NO_3^-, NH_4^+, and OC, are publicly available at https://dx.doi.org/10.7927/7wj3-en73; Daily maximum and minimum temperatures are from Daymet (Daily Surface Weather Data on a 1-km Grid for North America, Version 4 R1): https://doi.org/10.3334/ORNLDAAC/2129. Daily mean RH on a 4km×4km grid are from PRISM (the Parameter-elevation Relationships on Independent Slope Model, https://prism.oregonstate.edu); The 1km daily ALWC data generated in this study will be made publicly available upon publication of the work.

Acknowledgements

Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under Award Number R01AG074357 and RF1 AG079487, and the National Science Foundation Division of Atmospheric and Geospace Sciences under AGS-2307151. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

P.L. and B.Z. designed the study, B.Z. performed the ML model, generated the dataset and wrote the initial manuscript. L.Y. performed the GEOS-Chem simulations. Y.Y, H.G., L.X., D.P., N.L.N., R.J.W. provided observational data. Q.D., Y.W., J.W., J.S. provide the high-resolution datasets of the ML model inputs. All the co-authors commented on data analysis and contributed to the writing of the manuscript.

Competing Interests

The authors declare no competing financial or non-financial interests.

Reference

1 Bian, Y. X., Zhao, C. S., Ma, N., Chen, J. & Xu, W. Y. A study of aerosol liquid water content based on hygroscopicity measurements at high relative humidity in the North China Plain. *Atmos. Chem. Phys.* **14**, 6417-6426 (2014). https://doi.org:10.5194/acp-14-6417-2014

2 Pye, H. O. T. *et al.* On the implications of aerosol liquid water and phase separation for organic aerosol mass. *Atmos. Chem. Phys.* **17**, 343-369 (2017). https://doi.org:10.5194/acp-17-343-2017

3 Faust, J. A., Wong, J. P. S., Lee, A. K. Y. & Abbatt, J. P. D. Role of Aerosol Liquid Water in Secondary Organic Aerosol Formation from Volatile Organic Compounds. *Environmental Science & Technology* **51**, 1405-1413 (2017). https://doi.org:10.1021/acs.est.6b04700

4 Wang, X. *et al.* The secondary formation of inorganic aerosols in the droplet mode through heterogeneous aqueous reactions under haze conditions. *Atmospheric Environment* **63**, 68-76 (2012). https://doi.org:https://doi.org/10.1016/j.atmosenv.2012.09.029

5 Ravishankara, A. R. Heterogeneous and Multiphase Chemistry in the Troposphere. *Science* **276**, 1058-1065 (1997). https://doi.org:10.1126/science.276.5315.1058

6 Wong, J. P. S. *et al.* Fine Particle Iron in Soils and Road Dust Is Modulated by Coal-Fired Power Plant Sulfur. *Environmental Science & Technology* **54**, 7088-7096 (2020). https://doi.org:10.1021/acs.est.0c00483

7 Wang, Y. *et al.* Mutual promotion between aerosol particle liquid water and particulate nitrate enhancement leads to severe nitrate-dominated particulate matter pollution and low visibility. *Atmos. Chem. Phys.* **20**, 2161-2175 (2020). https://doi.org:10.5194/acp-20-2161-2020

8 Elias, T. *et al.* Enhanced extinction of visible radiation due to hydrated aerosols in mist and fog. *Atmos. Chem. Phys.* **15**, 6605-6623 (2015). https://doi.org:10.5194/acp-15-6605-2015

9 Wu, Z. J. *et al.* Particle hygroscopicity and its link to chemical composition in the urban atmosphere of Beijing, China, during summertime. *Atmos. Chem. Phys.* **16**, 1123-1138 (2016). https://doi.org:10.5194/acp-16-1123-2016

10 Jin, X. *et al.* Significant contribution of organics to aerosol liquid water content in winter in Beijing, China. *Atmos. Chem. Phys.* **20**, 901-914 (2020). https://doi.org:10.5194/acp-20-901-2020

11 Kuang, Y. *et al.* A novel method for calculating ambient aerosol liquid water content based on measurements of a humidified nephelometer system. *Atmos. Meas. Tech.* **11**, 2967-2982 (2018). https://doi.org:10.5194/amt-11-2967-2018

12 Guo, H. *et al.* Fine-particle water and pH in the southeastern United States. *Atmos. Chem. Phys.* **15**, 5211-5228 (2015). https://doi.org:10.5194/acp-15-5211-2015

13 Tan, W. *et al.* Profiling Aerosol Liquid Water Content Using a Polarization Lidar. *Environmental Science & Technology* **54**, 3129-3137 (2020). https://doi.org:10.1021/acs.est.9b07502

14 Fountoukis, C. & Nenes, A. ISORROPIA II: a computationally efficient thermodynamic equilibrium model for K+—Ca2+—Mg2+—NH4+—Na+—SO42-—NO3-—Cl-—H2O aerosols. *Atmos. Chem. Phys.* **7**, 4639-4659 (2007). https://doi.org:10.5194/acp-7-4639-2007

15 Nguyen, T. K. V., Zhang, Q., Jimenez, J. L., Pike, M. & Carlton, A. G. Liquid Water: Ubiquitous Contributor to Aerosol Mass. *Environmental Science & Technology Letters* **3**, 257-263 (2016). https://doi.org:10.1021/acs.estlett.6b00167

16 Nenes, A. *et al.* Aerosol acidity and liquid water content regulate the dry deposition of inorganic

reactive nitrogen. *Atmos. Chem. Phys.* **21**, 6023-6033 (2021). https://doi.org:10.5194/acp-21-6023-2021

17 Carlton, A. G., Christiansen, A. E., Flesch, M. M., Hennigan, C. J. & Sareen, N. Multiphase Atmospheric Chemistry in Liquid Water: Impacts and Controllability of Organic Aerosol. *Accounts of Chemical Research* **53**, 1715-1723 (2020). https://doi.org:10.1021/acs.accounts.0c00301

18 Sareen, N., Waxman, E. M., Turpin, B. J., Volkamer, R. & Carlton, A. G. Potential of Aerosol Liquid Water to Facilitate Organic Aerosol Formation: Assessing Knowledge Gaps about Precursors and Partitioning. *Environmental Science & Technology* **51**, 3327-3335 (2017). https://doi.org:10.1021/acs.est.6b04540

19 Miao, R. *et al.* Model bias in simulating major chemical components of PM2.5 in China. *Atmos. Chem. Phys.* **20**, 12265-12284 (2020). https://doi.org:10.5194/acp-20-12265-2020

20 Wyat Appel, K., Bhave, P. V., Gilliland, A. B., Sarwar, G. & Roselle, S. J. Evaluation of the community multiscale air quality (CMAQ) model version 4.5: Sensitivities impacting model performance; Part II—particulate matter. *Atmospheric Environment* **42**, 6057-6066 (2008). https://doi.org:https://doi.org/10.1016/j.atmosenv.2008.03.036

21 Di, Q. *et al.* (NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, New York, 2021).

22 Wei, J. *et al.* Long-term mortality burden trends attributed to black carbon and PM2·5 from wildfire emissions across the continental USA from 2000 to 2020: a deep learning modelling study. *The Lancet Planetary Health* **7**, e963-e975 (2023). https://doi.org:https://doi.org/10.1016/S2542-5196(23)00235-8

23 Amini, H. *et al.* Hyperlocal super-learned PM2.5 components across the contiguous US. *Research Square* **PREPRINT (Version 2)** (2022). https://doi.org:https://doi.org/10.21203/rs.3.rs-1745433/v2

24 Meng, X., Hand, J. L., Schichtel, B. A. & Liu, Y. Space-time trends of PM2.5 constituents in the conterminous United States estimated by a machine learning approach, 2005–2015. *Environment International* **121**, 1137-1147 (2018). https://doi.org:https://doi.org/10.1016/j.envint.2018.10.029

25 Thornton, M. M., Shrestha, R., Wei, Y., Thornton, P. E. & Kao, S. C. (ORNL Distributed Active Archive Center, 2020).

26 Thornton, P. E. *et al.* Gridded daily weather data for North America with comprehensive uncertainty quantification. *Scientific Data* **8**, 190 (2021). https://doi.org:10.1038/s41597-021-00973-0

27 Pan, D. *et al.* Regime shift in secondary inorganic aerosol formation and nitrogen deposition in the rural United States. *Nature Geoscience* **17**, 617-623 (2024). https://doi.org:10.1038/s41561-024-01455-9

28 Pilinis, C., Seinfeld, J. H. & Grosjean, D. Water content of atmospheric aerosols. *Atmospheric Environment (1967)* **23**, 1601-1606 (1989). https://doi.org:https://doi.org/10.1016/0004-6981(89)90419-8

29 Edgerton, E. S. *et al.* The Southeastern Aerosol Research and Characterization Study: Part II. Filter-Based Measurements of Fine and Coarse Particulate Matter Mass and Composition. *Journal of the Air & Waste Management Association* **55**, 1527-1542 (2005). https://doi.org:10.1080/10473289.2005.10464744

30 Edgerton, E. S. *et al.* The Southeastern Aerosol Research and Characterization Study, Part 3: Continuous Measurements of Fine Particulate Matter Mass and Composition. *Journal of the Air & Waste Management Association* **56**, 1325-1341 (2006). https://doi.org:10.1080/10473289.2006.10464585

31 Hansen, D. A. *et al.* The Southeastern Aerosol Research and Characterization Study: Part 1—Overview. *Journal of the Air & Waste Management Association* **53**, 1460-1471 (2003). https://doi.org:10.1080/10473289.2003.10466318

32 Kim, P. S. *et al.* Sources, seasonality, and trends of southeast US aerosol: an integrated analysis of surface, aircraft, and satellite observations with the GEOS-Chem chemical transport model. *Atmos. Chem. Phys.* **15**, 10411-10433 (2015). https://doi.org:10.5194/acp-15-10411-2015

33 Battaglia, M. A., Jr., Douglas, S. & Hennigan, C. J. Effect of the Urban Heat Island on Aerosol pH. *Environmental Science & Technology* **51**, 13095-13103 (2017). https://doi.org:10.1021/acs.est.7b02786

34 Zhang, Q. *et al.* Ubiquity and dominance of oxygenated species in organic aerosols in anthropogenically-influenced Northern Hemisphere midlatitudes. *Geophysical Research Letters* **34** (2007). https://doi.org:https://doi.org/10.1029/2007GL029979

35 Hand, J. L., Schichtel, B. A., Malm, W. C. & Pitchford, M. L. Particulate sulfate ion concentration and SO₂ emission trends in the United States from the early 1990s through 2010. *Atmos. Chem. Phys.* **12**, 10353-10365 (2012). https://doi.org:10.5194/acp-12-10353-2012

36 Shah, V. *et al.* Chemical feedbacks weaken the wintertime response of particulate sulfate and nitrate to emissions reductions over the eastern United States. *Proceedings of the National Academy of Sciences* **115**, 8110-8115 (2018). https://doi.org:doi:10.1073/pnas.1803295115

37 Ridley, D. A., Heald, C. L., Ridley, K. J. & Kroll, J. H. Causes and consequences of decreasing atmospheric organic aerosol in the United States. *Proceedings of the National Academy of Sciences* **115**, 290-295 (2018). https://doi.org:doi:10.1073/pnas.1700387115

38 Guo, H. *et al.* Fine particle pH and gas–particle phase partitioning of inorganic species in Pasadena, California, during the 2010 CalNex campaign. *Atmos. Chem. Phys.* **17**, 5703-5719 (2017). https://doi.org:10.5194/acp-17-5703-2017

39 Fang, T. *et al.* Highly Acidic Ambient Particles, Soluble Metals, and Oxidative Potential: A Link between Sulfate and Aerosol Toxicity. *Environmental Science & Technology* **51**, 2611-2620 (2017). https://doi.org:10.1021/acs.est.6b06151

40 Chen, L. C. & Lippmann, M. Effects of Metals within Ambient Air Particulate Matter (PM) on Human Health. *Inhalation Toxicology* **21**, 1-31 (2009). https://doi.org:10.1080/08958370802105405

41 Abbaspour, N., Hurrell, R. & Kelishadi, R. Review on iron and its importance for human health. *J Res Med Sci* **19**, 164-174 (2014).

42 Wang, R. *et al.* Sources, transport and deposition of iron in the global atmosphere. *Atmos. Chem. Phys.* **15**, 6247-6270 (2015). https://doi.org:10.5194/acp-15-6247-2015

43 Ito, A. Atmospheric Processing of Combustion Aerosols as a Source of Bioavailable Iron. *Environmental Science & Technology Letters* **2**, 70-75 (2015). https://doi.org:10.1021/acs.estlett.5b00007

44 Amini, H. *et al.* (Research Square, 2022).

45 Vu, B. N. *et al.* Association of Annual Exposure to Air Pollution Mixture on Asthma

Hospitalizations in the United States. *American Journal of Respiratory and Critical Care Medicine* **211**, 1636-1643 (2025). https://doi.org:10.1164/rccm.202409-1853OC

46 Danesh Yazdi, M. *et al.* Long-term exposure to PM2.5 components and cardiovascular admissions in medicare patients. *Environmental Research* **286**, 122779 (2025). https://doi.org:https://doi.org/10.1016/j.envres.2025.122779

47 Oakes, M. *et al.* Iron Solubility Related to Particle Sulfur Content in Source Emission and Ambient Fine Particles. *Environmental Science & Technology* **46**, 6637-6644 (2012). https://doi.org:10.1021/es300701c

48 Zhu, Y. *et al.* Iron solubility in fine particles associated with secondary acidic aerosols in east China. *Environmental Pollution* **264**, 114769 (2020). https://doi.org:https://doi.org/10.1016/j.envpol.2020.114769

49 Shi, Z. *et al.* Impacts on iron solubility in the mineral dust by processes in the source region and the atmosphere: A review. *Aeolian Research* **5**, 21-42 (2012). https://doi.org:https://doi.org/10.1016/j.aeolia.2012.03.001

50 Yang, Y. & Weber, R. J. Ultrafiltration to characterize PM2.5 water-soluble iron and its sources in an urban environment. *Atmospheric Environment* **286**, 119246 (2022). https://doi.org:https://doi.org/10.1016/j.atmosenv.2022.119246

51 Myriokefalitakis, S. *et al.* Changes in dissolved iron deposition to the oceans driven by human activity: a 3-D global modelling study. *Biogeosciences* **12**, 3973-3992 (2015). https://doi.org:10.5194/bg-12-3973-2015

52 Zhang, B. *et al.* Significant contrasts in aerosol acidity between China and the United States. *Atmos. Chem. Phys.* **21**, 8341-8356 (2021). https://doi.org:10.5194/acp-21-8341-2021

53 Sakata, K. *et al.* Iron (Fe) speciation in size-fractionated aerosol particles in the Pacific Ocean: The role of organic complexation of Fe with humic-like substances in controlling Fe solubility. *Atmos. Chem. Phys.* **22**, 9461-9482 (2022). https://doi.org:10.5194/acp-22-9461-2022

54 Paris, R. & Desboeufs, K. V. Effect of atmospheric organic complexation on iron-bearing dust solubility. *Atmos. Chem. Phys.* **13**, 4895-4905 (2013). https://doi.org:10.5194/acp-13-4895-2013

55 Paris, R., Desboeufs, K. V. & Journet, E. Variability of dust iron solubility in atmospheric waters: Investigation of the role of oxalate organic complexation. *Atmospheric Environment* **45**, 6510-6517 (2011). https://doi.org:https://doi.org/10.1016/j.atmosenv.2011.08.068

56 Tao, Y. & Murphy, J. G. The Mechanisms Responsible for the Interactions among Oxalate, pH, and Fe Dissolution in PM2.5. *ACS Earth and Space Chemistry* **3**, 2259-2265 (2019). https://doi.org:10.1021/acsearthspacechem.9b00172

57 Weber, R. J., Guo, H., Russell, A. G. & Nenes, A. High aerosol acidity despite declining atmospheric sulfate concentrations over the past 15 years. *Nature Geoscience* **9**, 282-285 (2016). https://doi.org:10.1038/ngeo2665

58 Petters, M. D. & Kreidenweis, S. M. A single parameter representation of hygroscopic growth and cloud condensation nucleus activity. *Atmos. Chem. Phys.* **7**, 1961-1971 (2007). https://doi.org:10.5194/acp-7-1961-2007

59 Martin, R. V., Jacob, D. J., Yantosca, R. M., Chin, M. & Ginoux, P. Global and regional decreases in tropospheric oxidants from photochemical effects of aerosols. *Journal of Geophysical Research: Atmospheres* **108** (2003). https://doi.org:https://doi.org/10.1029/2002JD002622

60 King, S. M. *et al.* Cloud droplet activation of mixed organic-sulfate particles produced by the photooxidation of isoprene. *Atmos. Chem. Phys.* **10**, 3953-3964 (2010). https://doi.org:10.5194/acp-10-3953-2010

61 Rickards, A. M. J., Miles, R. E. H., Davies, J. F., Marshall, F. H. & Reid, J. P. Measurements of the Sensitivity of Aerosol Hygroscopicity and the κ Parameter to the O/C Ratio. *The Journal of Physical Chemistry A* **117**, 14120-14131 (2013). https://doi.org:10.1021/jp407991n

62 Ervens, B. *et al.* CCN predictions using simplified assumptions of organic aerosol composition and mixing state: a synthesis from six different locations. *Atmos. Chem. Phys.* **10**, 4795-4807 (2010). https://doi.org:10.5194/acp-10-4795-2010

63 Dusek, U. *et al.* Enhanced organic mass fraction and decreased hygroscopicity of cloud condensation nuclei (CCN) during new particle formation events. *Geophysical Research Letters* **37** (2010). https://doi.org:https://doi.org/10.1029/2009GL040930

64 Gunthe, S. S. *et al.* Cloud condensation nuclei in pristine tropical rainforest air of Amazonia: size-resolved measurements and modeling of atmospheric aerosol composition and CCN activity. *Atmos. Chem. Phys.* **9**, 7551-7575 (2009). https://doi.org:10.5194/acp-9-7551-2009

65 PRISM Climate Group. (Oregon State University,).

66 Daly, C. *et al.* Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. *International Journal of Climatology* **28**, 2031-2064 (2008). https://doi.org:https://doi.org/10.1002/joc.1688

67 Spangler, K. R., Weinberger, K. R. & Wellenius, G. A. Suitability of gridded climate datasets for use in environmental epidemiology. *Journal of Exposure Science & Environmental Epidemiology* **29**, 777-789 (2019). https://doi.org:10.1038/s41370-018-0105-2

68 Di, Q. *et al.* An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. *Environment International* **130**, 104909 (2019). https://doi.org:https://doi.org/10.1016/j.envint.2019.104909

69 Di, Q. *et al.* (NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, New York, 2022).

70 Di, Q. *et al.* Assessing NO2 Concentration and Model Uncertainty with High Spatiotemporal Resolution across the Contiguous United States Using Ensemble Model Averaging. *Environmental Science & Technology* **54**, 1372-1384 (2020). https://doi.org:10.1021/acs.est.9b03358

71 Requia, W. J. *et al.* An Ensemble Learning Approach for Estimating High Spatiotemporal Resolution of Ground-Level Ozone in the Contiguous United States. *Environmental Science & Technology* **54**, 11037-11047 (2020). https://doi.org:10.1021/acs.est.0c01791

72 Requia, W. J. *et al.* (NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, New York, 2021).

73 Amini, H. *et al.* (NASA Socioeconomic Data and Applications Center (SEDAC), Palisades, New York, 2023).

74 Qi Zhang, Caroline Parworth, Michael Lechner & Jimenez, J. L. (https://sites.google.com/site/amsglobaldatabase).

75 Xu, L., Suresh, S., Guo, H., Weber, R. J. & Ng, N. L. Aerosol characterization over the southeastern United States using high-resolution aerosol mass spectrometry: spatial and seasonal variation of aerosol composition and sources with a focus on organic nitrates. *Atmos. Chem. Phys.* **15**, 7307-7336 (2015). https://doi.org:10.5194/acp-15-7307-2015

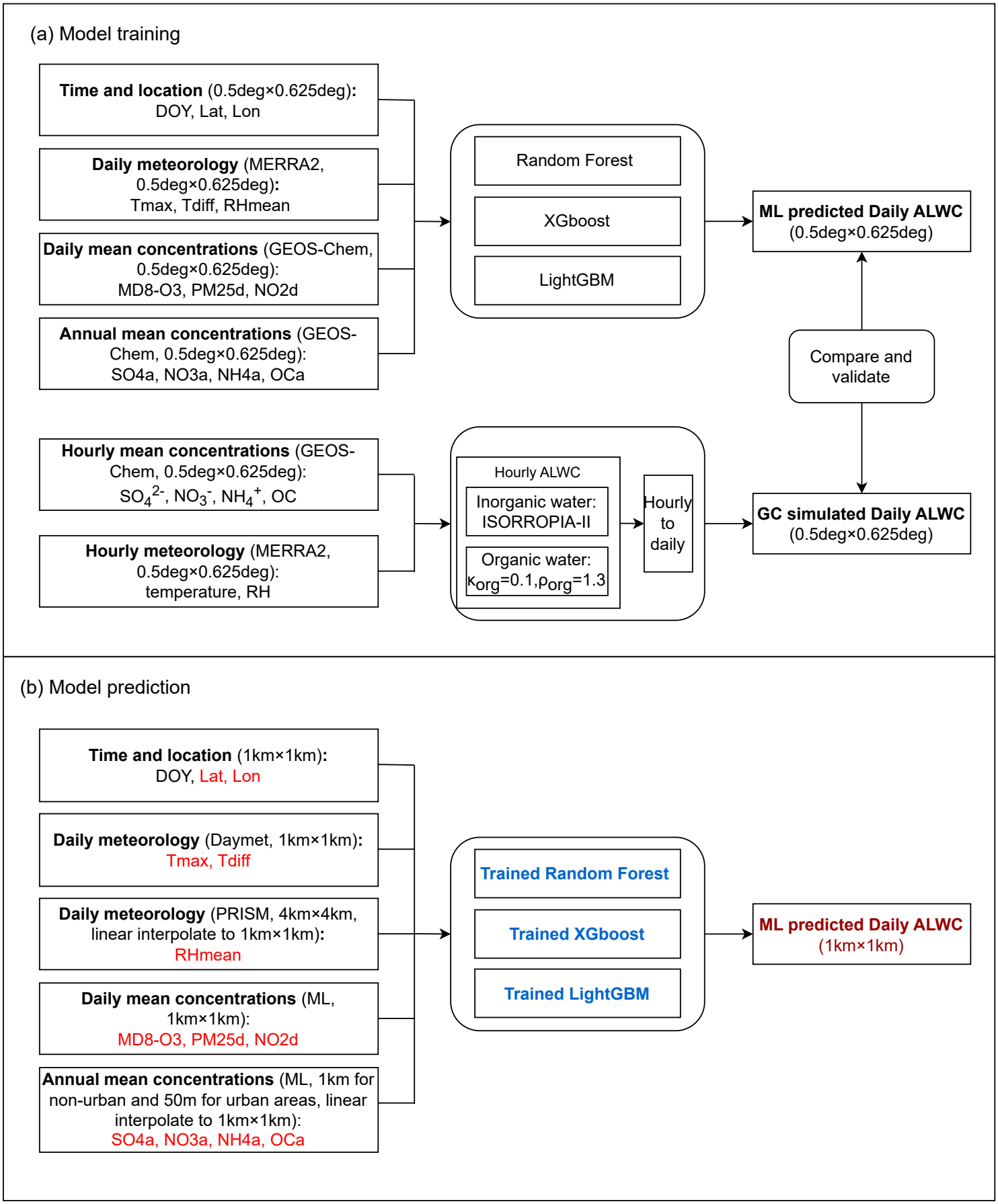
76 Xu, L. *et al.* Effects of anthropogenic emissions on aerosol formation from isoprene and monoterpenes in the southeastern United States. *Proceedings of the National Academy of Sciences* **112**, 37-42 (2015). https://doi.org:doi:10.1073/pnas.1417609112

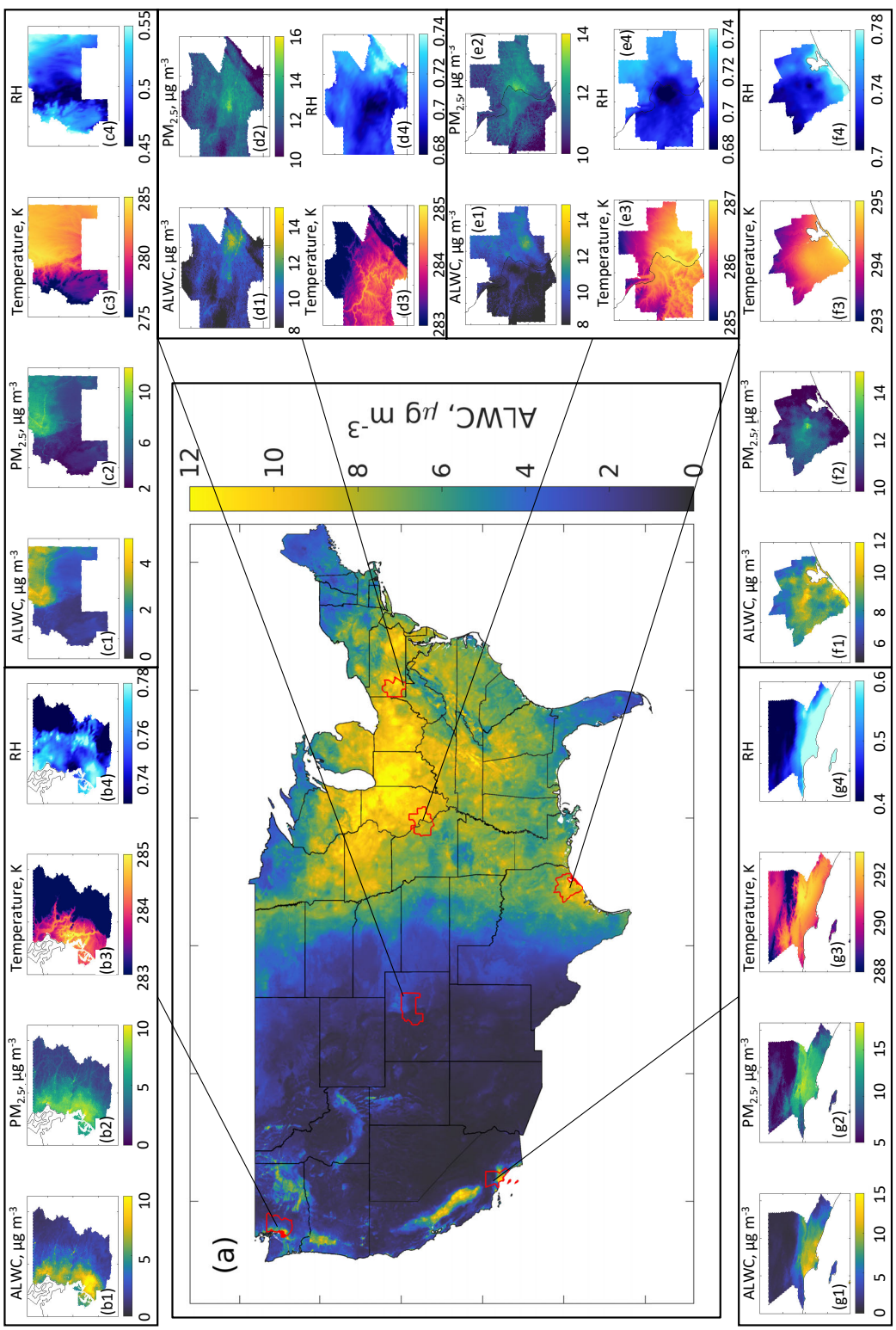
Figures

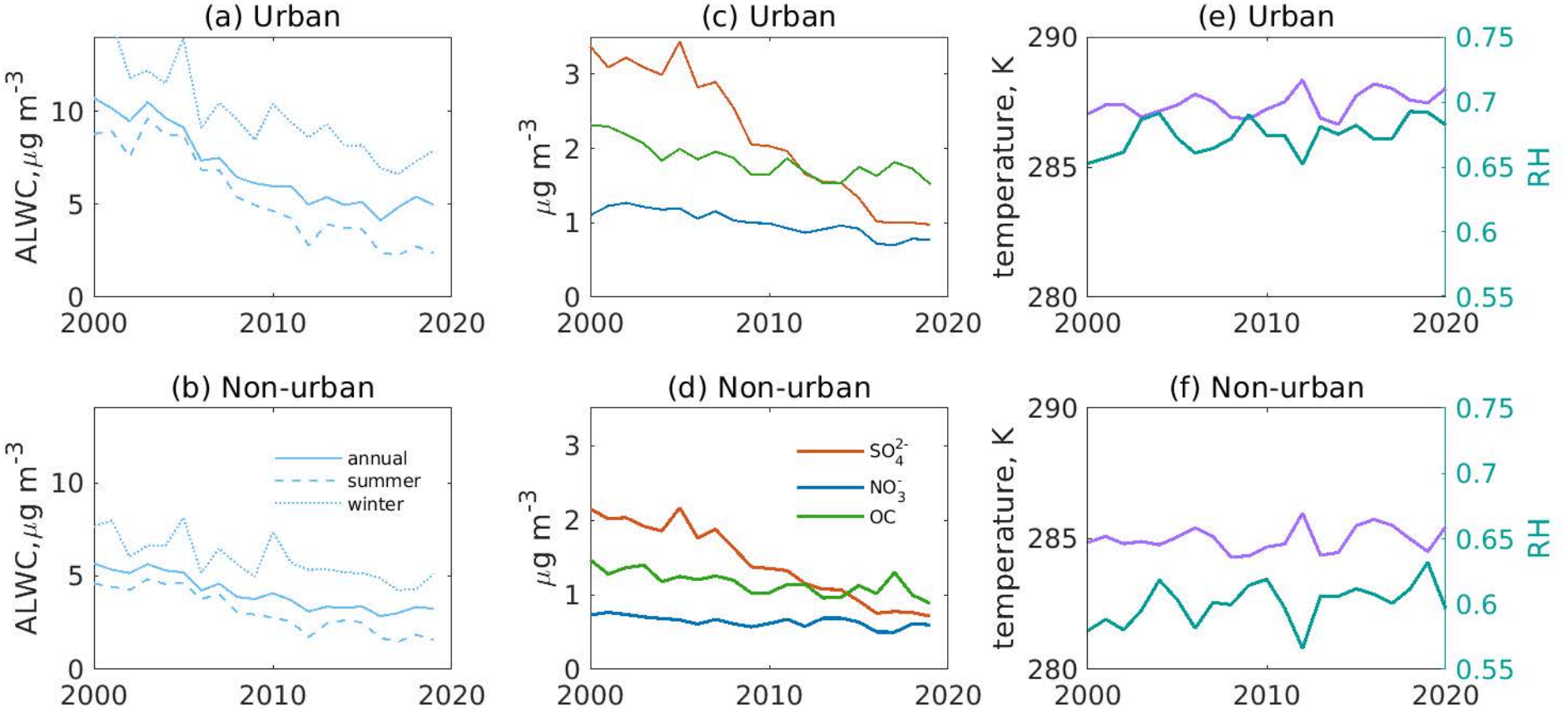
Figure 1. Schematic diagram of the ML model training (a) and prediction (b) processes used in this study. The full names and data sources of each input variable are listed in Table S1.

Figure 2. Maps of ALWC and its driving forces from 2000 to 2019. (a) Averaged ALWC across CONUS. (b1-g4) Averaged ALWC, $PM_{2.5}$, temperature, and relative humidity (RH) in six metropolitan statistical areas (MSAs). b1-b4: Seattle-Tacoma-Bellevue, WA; c1-c4: Denver-Aurora-Lakewood, CO; d1-d4: Pittsburgh, PA; e1-e4: St. Louis, MO-IL; f1-f4: Houston-The Woodlands-Sugar Land, TX; g1-g4: Los Angeles-Long Beach-Anaheim, CA. Note that color scales differ across subplots.

Figure 3. Time series of annual mean ALWC, sulfate (SO_4^{2-}), nitrate (NO_3^-), organic carbon (OC), temperature and RH in 2000-2019 in urban and non-urban areas.


(a) Model training
Time and location (0.5deg×0.625deg): DOY, Lat, Lon
Daily meteorology (MERRA2, 0.5deg×0.625deg): Tmax, Tdiff, RHmean
Daily mean concentrations (GEOS-Chem, 0.5deg×0.625deg): MD8-O3, PM25d, NO2d
Annual mean concentrations (GEOS-Chem, 0.5deg×0.625deg): SO4a, NO3a, NH4a, OCa
Random Forest
XGboost
LightGBM
ML predicted Daily ALWC (0.5deg×0.625deg)
Compare and validate
Hourly mean concentrations (GEOS-Chem, 0.5deg×0.625deg): SO_4^{2-}, NO_3^-, NH_4^+, OC
Hourly meteorology (MERRA2, 0.5deg×0.625deg): temperature, RH
Hourly ALWC
Inorganic water: ISORROPIA-II
Organic water: κ_{org}=0.1,ρ_{org}=1.3
Hourly to daily
GC simulated Daily ALWC (0.5deg×0.625deg)
(b) Model prediction
Time and location (1km×1km): DOY, Lat, Lon
Daily meteorology (Daymet, 1km×1km): Tmax, Tdiff
Daily meteorology (PRISM, 4km×4km, linear interpolate to 1km×1km): RHmean
Daily mean concentrations (ML, 1km×1km): MD8-O3, PM25d, NO2d
Annual mean concentrations (ML, 1km for non-urban and 50m for urban areas, linear interpolate to 1km×1km): SO4a, NO3a, NH4a, OCa
Trained Random Forest
Trained XGboost
Trained LightGBM
ML predicted Daily ALWC (1km×1km)


(a)
ALWC, μg m-3
0 2 4 6 8 10 12
(b1) (b2) (b3) (b4)
(c1) (c2) (c3) (c4)
(d1) (d2) (d3) (d4)
(e1) (e2) (e3) (e4)
(f1) (f2) (f3) (f4)
(g1) (g2) (g3) (g4)
PM2.5, μg m-3
Temperature, K
RH


(a) Urban
ALWC,μg m-3
0
5
10
2000
2010
2020
(c) Urban
μg m-3
0
1
2
3
2000
2010
2020
(e) Urban
temperature, K
280
285
290
RH
0.55
0.6
0.65
0.7
0.75
2000
2010
2020
(b) Non-urban
ALWC,μg m-3
annual
summer
winter
0
5
10
2000
2010
2020
(d) Non-urban
μg m-3
SO4 2-
NO3 -
OC
0
1
2
3
2000
2010
2020
(f) Non-urban
temperature, K
280
285
290
RH
0.55
0.6
0.65
0.7
0.75
2000
2010
2020