

Final Report

Project Name: Estimating Surface PM_{2.5} from Satellite Aerosols

Proposal for Grant Activities/PCR Number: 582-23-41321-025

Grant Number: 582-18-81339

Grant Name: University of Houston

Grantee Project Manager: Yunsoo Choi

Address: 426-F SR1, 3507 Cullen Blvd, Houston, Texas 77004

Phone: 713.743.9748

E-mail: ychoi6@uh.edu

May 2024

Abstract

This report presents the outcomes of the project supported by the Texas Commission on Environmental Quality (TCEQ), focusing on the estimation of daily surface PM_{2.5} concentrations at 4 km spatial resolution across Texas from 2018 to 2022 using an advanced deep learning (DL) model. A key innovation of this project was the creation of continuous, accurate, and gap-free grids of surface PM_{2.5} estimates for each day, which effectively addressed the challenges posed by missing data in satellite imagery. Monitoring PM_{2.5}, a critical pollutant known for its detrimental impacts on human health and the environment, typically employs various methods, such as satellite remote sensing and ground-based monitoring stations. However, these approaches have notable limitations. For instance, satellite data often have substantial gaps and rely on proxies like Aerosol Optical Depth (AOD) for PM_{2.5} monitoring. AOD, however, does not consistently correlate well with surface PM_{2.5} levels, thus compromising reliability. To mitigate these limitations, the UH-AQF Artificial Intelligence (AI) group at the University of Houston utilized sophisticated DL techniques to create high-resolution, gap-free grids of surface PM_{2.5} levels, markedly improving the accuracy of PM_{2.5} monitoring in all regions.

Our research employed a two-phase DL model combining a deep convolutional neural network (DeepCNN) with a depthwise partial convolutional neural network (PCNN). In the first phase, we used predictor variables such as satellite AOD, surface observations, and outputs from chemical transport models to generate gap-free daily PM_{2.5} grids. The second phase refined these estimates using additional features, producing highly accurate daily PM_{2.5} maps. The performance of our DL model was rigorously evaluated against ground station measurements using metrics like the correlation coefficient (R), index of agreement (IOA), mean absolute bias (MAB), and root mean square error (RMSE). From 2018 to 2022, our PCNN-DeepCNN model demonstrated robust capability in estimating surface PM_{2.5}, achieving R and IOA scores between 0.89-0.93 and 0.94-0.96, respectively. Additionally, the model maintained a low bias with MAB ranging from 1.3 to 1.79 $\mu\text{g}/\text{m}^3$ and RMSE from 1.87 to 3.25 $\mu\text{g}/\text{m}^3$ in different years. Spatial cross-validation results further confirmed the model's advanced spatial accuracy, exhibiting R values from 0.81 to 0.87 and IOA from 0.89 to 0.93. In addition to its high accuracy, a significant advantage of our model over previous studies is its ability to handle missing values in satellite data. This ensures that our estimated PM_{2.5} maps are complete, providing PM_{2.5} values for all regions, including those without monitoring stations and areas affected by dense cloud cover.

The project also aimed to enhance policy-making by providing high-resolution PM_{2.5} coverage across Texas, supporting the implementation of the State Implementation Plan (SIP). Our team developed a method to calculate design values (DVs) for all regions and identify non-attainment areas using the DL-estimated PM_{2.5} grids. Traditional methods, relying on sparse EPA monitoring stations, often fail to provide DVs in many counties due to a lack of monitoring infrastructure. In contrast, our DL model delivered comprehensive coverage for all 254 Texas counties. A comparative analysis revealed that our DL-derived DVs (DL-DV) detected significantly more counties violating the National Ambient Air Quality Standards (NAAQS) for PM_{2.5} than EPA-derived DVs (EPA-DV). For instance, the EPA-DV identified violations in 8, 8, and 10 counties for the years 2020, 2021, and 2022, respectively, whereas the DL-DV indicated 94, 76, and 71 non-compliant counties for the same periods. For these analyses, the annual PM_{2.5} NAAQS was set at 9 $\mu\text{g}/\text{m}^3$. This discrepancy highlights potential underestimations of air quality issues by traditional EPA

methods and underscores the necessity for advanced monitoring techniques like ours to achieve more accurate air quality data.

Our team also developed a web portal with various tools for data analysis and visualization, enhancing the accessibility and utility of the data for researchers, policymakers, and the public. This report underscores the potential of advanced DL models to significantly enhance our understanding and monitoring of air quality. These models are particularly valuable in regions where traditional monitoring infrastructure is absent or where high cloud cover impedes PM_{2.5} monitoring using satellite data. The methodologies and findings of this project are expected to inform future air quality monitoring and management strategies, promoting better health and environmental outcomes.

1. Introduction

Particulate matter with an aerodynamic diameter of less than 2.5 μm , known as PM_{2.5}, poses a significant threat to human health (Feng et al., 2016; Pascal et al., 2014) and the environment (Ma et al., 2021; Manisalidis et al., 2020). Numerous studies have established a link between PM_{2.5} exposure and adverse health outcomes, including cardiovascular disease (Hayes et al., 2020), respiratory illnesses (Pun et al., 2017; Xing et al., 2016), myocardial infarction (Madrigano et al., 2013), and increased mortality rates (Schwartz et al., 1996). Therefore, monitoring surface PM_{2.5} concentrations is crucial for a range of applications, including public health initiatives and environmental research.

Surface PM_{2.5} monitoring is conventionally conducted using three primary methods, one of which includes measurements at ground stations (Ginzburg et al., 2015). While these surface monitoring stations offer precision and accuracy, their limited number restricts spatial coverage. Consequently, estimating PM_{2.5} exposures based on these stations is likely to be fraught with errors due to their sparse distribution, uneven placement, and variability in measurement frequencies (Ghahremanloo et al., 2021). Another strategy for monitoring surface PM_{2.5} involves the use of chemical transport models (CTMs), such as the Community Multiscale Air Quality (CMAQ) modeling system (Zhang et al., 2019). Despite their widespread use and relevance, these models often exhibit significant biases that can limit their utility across various scientific domains. For instance, Ghahremanloo et al. (2022) evaluated the accuracy of daily PM_{2.5} simulations by CMAQ across the contiguous United States (CONUS) from February to June 2019, finding a correlation coefficient (R) of 0.50 and a mean absolute bias (MAB) of 2.79 $\mu\text{g}/\text{m}^3$, indicating the model's limited accuracy in PM_{2.5} simulations.

Satellite remote sensing also represents an alternative method for monitoring PM_{2.5} levels (Kloog et al., 2011). However, this approach is hindered by the presence of missing data in satellite imagery due to factors such as cloud cover, snow, and high surface reflectivity. For instance, Xiao et al. (2017) reported an average of nearly 60% missing data in aerosol optical depth (AOD) images of the Moderate Resolution Imaging Spectroradiometer (MODIS) from 2013 to 2014 in the Yangtze River Delta, China, primarily due to these conditions. Additionally, satellites do not directly measure surface PM_{2.5} concentrations but instead rely on proxies like AOD. The relationship between AOD and surface PM_{2.5} is not consistently strong, which complicates this method's reliability. Fu et al. (2022) investigated the correlation between AOD and PM_{2.5} using ground-based observations from 19 stations across China during 2017-2019 and found correlations ranging from a low of 0.03 to a medium of 0.60, indicating significant variability. They

attributed this variability to meteorological conditions and aerosol composition differences, which often decouple AOD from $PM_{2.5}$ levels. Furthermore, AOD represents the extinction of solar radiation by aerosols throughout the entire atmospheric column, which may not accurately reflect ground-level $PM_{2.5}$ levels, especially in the presence of elevated aerosol layers (Ghahremanloo et al., 2021). Consequently, reliance on AOD alone for $PM_{2.5}$ monitoring can introduce significant biases. Thus, developing more accurate methodologies that address the limitations of current $PM_{2.5}$ monitoring tools is crucial. Estimating daily $PM_{2.5}$ grids at high accuracy can significantly improve the quality of various applications, such as public health studies and governmental decision-making.

One advanced approach to estimating surface concentrations of $PM_{2.5}$ is the application of machine learning (ML) and deep learning (DL) models. Recent advancements in satellite-derived data products have further propelled the prominence of data-driven ML approaches. These approaches are particularly valued for their capacity to develop accurate estimation models through the use of extensive supervised training datasets (Park et al., 2020). In the realms of atmospheric sciences, climate, oceanography, and chemistry, numerous studies have applied ML/DL algorithms for purposes such as forecasting (Gong et al., 2022; Grigsby et al., 2023; Wu et al., 2020), image imputation (Lops et al., 2022; Singh et al., 2024), and the creation of digital twins (Payami et al., 2024; Salman et al., 2024). These applications leverage surface $PM_{2.5}$ data as target regression labels and training models to map input predictors to $PM_{2.5}$ levels through supervised learning. Once trained, these models can accurately estimate $PM_{2.5}$ concentrations in areas underserved by ground-based monitoring, demonstrating significant generalization capabilities (Ghahremanloo et al., 2023; Sayeed et al., 2022).

Early research in this field predominantly utilized simpler linear or generalized linear regression models to estimate $PM_{2.5}$ levels (Gupta & Christopher, 2009; Zhao et al., 2018). However, recent advancements have shifted focus towards more sophisticated nonlinear methods to achieve higher accuracy. For instance, Hu et al. (2017) employed a random forest (RF) model that combined satellite and model AOD data with meteorological factors and land-use information to generate daily $PM_{2.5}$ estimates across the CONUS with a 12 km spatial resolution. Similarly, Ghahremanloo et al. (2021) utilized the RF algorithm, incorporating Multiangle Implementation of Atmospheric Correction (MAIAC) AOD and various predictor variables, to estimate daily ground-level $PM_{2.5}$ concentrations at a finer 1 km spatial resolution in Texas. This study integrated aerosol-related parameters, including daily Ångström Exponent and column densities of dust, sea salt, sulfur dioxide (SO_2), sulfate (SO_4), and organic carbon (OC). Expanding the toolkit further, Park et al. (2020) implemented a convolutional neural network (CNN) that processed 27 input parameters, such as AOD from MODIS and global models, alongside meteorological data and land-use characteristics. Their model focused on estimating daily mean $PM_{2.5}$ concentrations in the CONUS, demonstrating that finer spatial resolution in AOD data could significantly enhance the accuracy of $PM_{2.5}$ predictions compared to coarser resolutions. Additionally, Di et al. (2016) developed a hybrid neural network model that integrated satellite data, land-use regression, and CTMs outputs to estimate $PM_{2.5}$ levels in the United States from 2000 to 2022. The inclusion of convolutional layers in their model facilitated the effective processing of diverse data types, thus enabling accurate mapping of the complex spatial and temporal variations of $PM_{2.5}$.

This project, funded by the Texas Commission on Environmental Quality (TCEQ), introduced an advanced DL model to estimate daily surface concentrations of $PM_{2.5}$ across Texas from 2018 to 2022 with a 4 km

spatial resolution. A notable innovation of this project, distinguishing it from previous efforts, is the creation of continuous, gap-free grids of estimated surface $PM_{2.5}$ for each day, effectively eliminating the impact of missing values from satellite imagery in the final outputs. The DL approach utilized in this study also demonstrated high accuracy in estimating surface $PM_{2.5}$ levels. The UH-AQF artificial intelligence (AI) group also aimed to improve policy-making by offering detailed $PM_{2.5}$ coverage across Texas, aiding the State Implementation Plan (SIP). Our team devised a method using DL to calculate design values (DVs) in all regions and identify non-attainment areas statewide. Additionally, we have developed an accessible data portal equipped with a range of analytical tools designed to facilitate various analyses, including temporal and spatial averaging, calculation of DVs, and data downloading, all tailored to work with the DL-estimated $PM_{2.5}$ grids.

2. Study Area and Data

Our research project delved into Texas, the second-largest state nestled in the south-central regions of the United States, boasting 254 counties and prominent urban centers such as Dallas, Houston, Austin, and San Antonio. Spanning an expansive area of 696,200 square kilometers and home to a populace of 31.9 million (source: <https://www.census.gov>), Texas served as the focal point of our investigation. The study, conducted from January 1, 2018, to December 31, 2022, involved the utilization of a variety of aerosol and meteorology-related predictor variables as inputs for training our DL model. This model was devised to estimate daily $PM_{2.5}$ concentrations at a spatial resolution of 4 km. Further elucidation regarding the specific target and predictor variables is provided below. Figure 1 illustrates the study area alongside the $PM_{2.5}$ monitoring stations established by the Environmental Protection Agency (EPA).

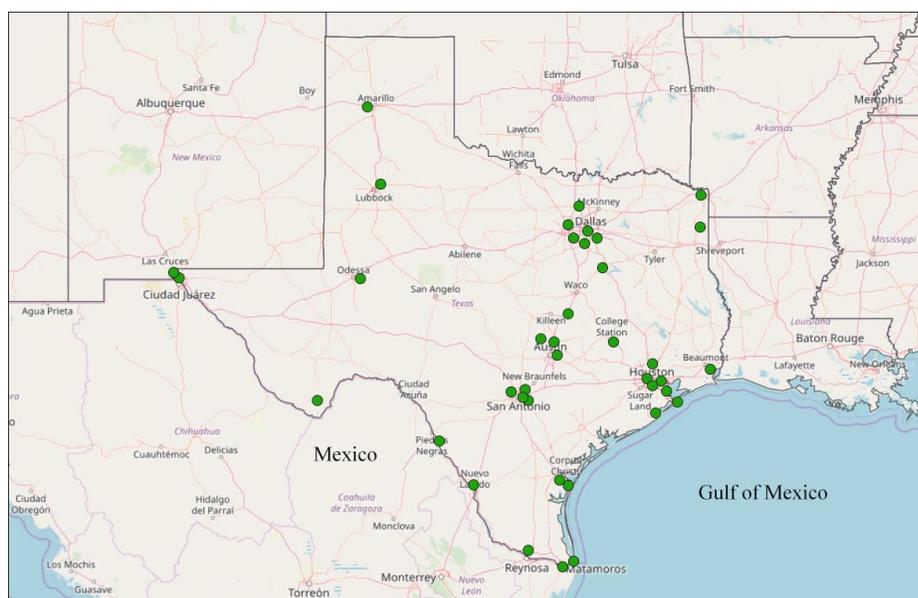


Figure 1. The study area encompasses the state of Texas. Green circles denote EPA ground stations monitoring hourly $PM_{2.5}$ concentrations throughout Texas.

2.1. Target Variable

The target variable in our DL model was surface $PM_{2.5}$ observations. We sourced hourly surface $PM_{2.5}$ concentrations recorded at air quality monitoring stations throughout the CONUS from January 2018 to December 2022 via the EPA Air Quality System (AQS). While the first phase of our DL model employs EPA stations across the CONUS for training, the second phase of the model focuses solely on stations in Texas. The downloaded hourly $PM_{2.5}$ observations were aggregated to derive daily mean concentrations, subsequently subject to a filtration process to exclude data of low quality based on the quality assurance information furnished by the EPA. Below are detailed explanations of the predictor variables utilized in our DL model. All variables were prepared for Texas from 2018 to 2022, with the exception of three features: MODIS AOD, CMAQ $PM_{2.5}$, and the percentage of urban space (PUS). These were prepared for the first phase of the DL model over the CONUS to augment the training dataset for this phase.

2.2. Predictor Variables

2.2.1. MODIS Products

The MODIS instrument aboard NASA's Terra and Aqua satellites is used to measure various Earth system variables, including AOD. Several algorithms, such as the MAIAC, have been developed to retrieve AOD data from MODIS measurements. MAIAC stands out for its ability to retrieve AOD at a high spatial resolution of 1 km (Lyapustin et al., 2011). This algorithm is particularly adept at correcting for atmospheric effects and retrieving aerosol information over both dark vegetated regions and bright surfaces, although it currently excludes snow and bright salt pans. To obtain the MAIAC AOD data, we utilized the Level-1 Atmosphere Archive and Distribution System (LAADS) (<https://ladsweb.modaps.eosdis.nasa.gov/>). The MAIAC AOD was used only in the first phase (i.e., partial convolutional neural network (PCNN)) of our DL model. We selected only pixels classified as superior quality, as indicated by the quality assurance flags included in the dataset (Liu et al., 2009). After excluding images containing clouds and invalid pixels, our dataset consisted of 7,837 swath images from 2018 to 2022. In our study, we prepared AOD datasets over the CONUS by combining multiple swath images retrieved per day. Our DL model requires input data without missing values. To address missing and invalid data, we averaged images from multiple days. Additionally, we prepared daily Enhanced Vegetation Index (EVI) data sourced from the MODIS MOD13C1 product, which provides global vegetated surface data at a spatial resolution of 0.05° (Didan, 2015). Figure 2 displays the monthly mean MAIAC AOD from MODIS for June 2018 over Texas, while Figure 3 presents the MODIS EVI for the entire globe in 2022.

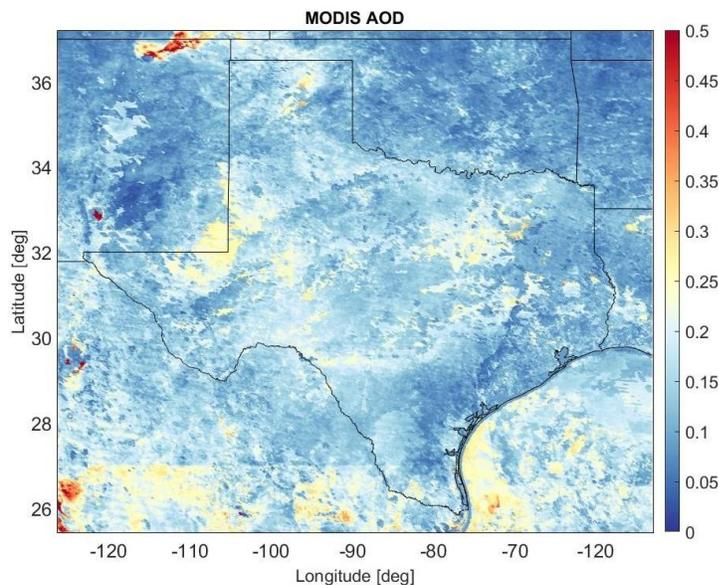


Figure 2. Spatial distribution of MODIS AOD over Texas in June 2018.

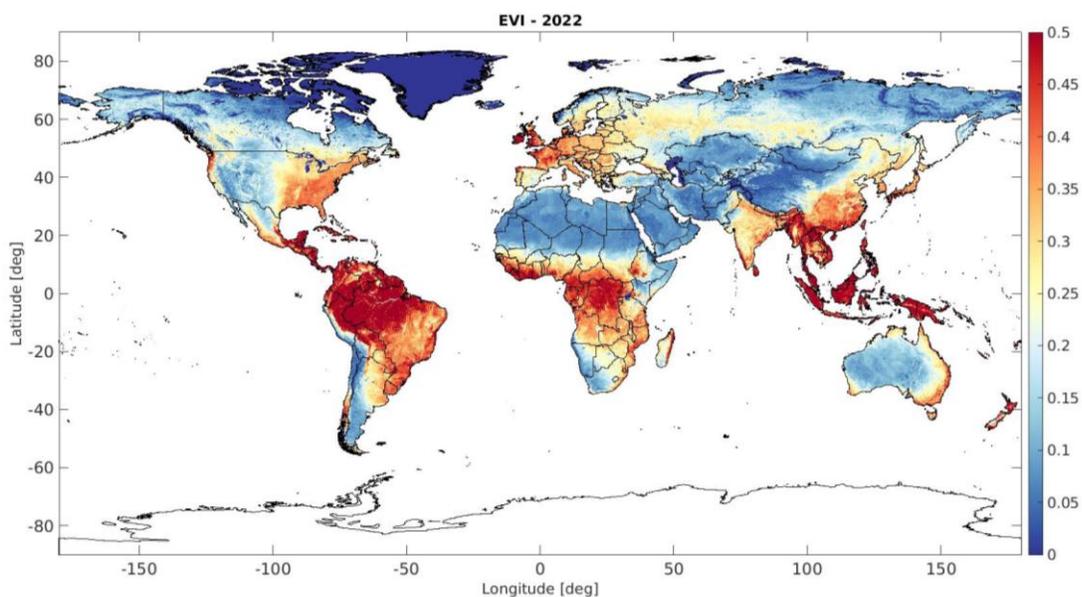


Figure 3. Spatial distribution of MODIS EVI over the globe in 2022.

2.2.2. Chemical Transport Model Simulated $PM_{2.5}$

We used the CMAQ model version 5.2 (Byun & Schere, 2006) to simulate surface $PM_{2.5}$ concentrations over the CONUS at a 12 km spatial resolution from February to June 2019 and 2020. The CMAQ $PM_{2.5}$ data was exclusively utilized during the initial phase, known as depthwise PCNN, of the DL model. The key aspect of depthwise PCNN in estimating surface $PM_{2.5}$ is capturing the spatial pattern of $PM_{2.5}$ over the study area, derived from CMAQ $PM_{2.5}$ and AOD data. Given this focus, the duration of the CMAQ dataset, even if it does not encompass full calendar years, is sufficient for training the depthwise PCNN because

the key information lies in the spatial distribution patterns rather than the long-term temporal completeness of the data (Ghahremanloo et al., 2023). The emissions input for CMAQ simulations included anthropogenic and biogenic emissions sourced from the 2017 EPA National Emissions Inventory (NEI) (Eyth & Vukovich, 2016; Eyth et al., 2016), lightning-induced nitric oxide (NO) emissions from the National Lightning Detection Network (NLDN) (Orville, 2008) and biomass burning emissions from the Fire INventory from National Center for Atmospheric Research (FINN) version 1.5 (Wiedinmyer et al., 2006, 2011, 2014). We prepared the meteorology input for CMAQ through simulations using the Weather Research and Forecasting (WRF) model version 4.0. We used the National Centers for Environmental Prediction (NCEP) North American Regional Reanalysis (NARR) data as the initial and boundary conditions and implemented the indirect soil moisture and temperature nudging technique (Gilliam & Pleim, 2010; Pleim & Xiu, 2003) with a four-dimensional data assimilation option (Hogrefe et al., 2015). We then averaged the CMAQ-simulated hourly surface PM_{2.5} concentrations into the daily mean. The cubic convolution interpolation method (Ghahremanloo et al., 2023; Keys, 1981) was used to resample the spatial resolution of CMAQ PM_{2.5} data from 12 km to 4 km. This approach interpolates values by fitting a smooth curve through the 16 nearest cell centers, employing a symmetric kernel of piecewise cubic polynomials that ensures continuity and a continuous first derivative across unit subintervals from -2 to +2. Subsequently, the resampled CMAQ datasets were subjected to validation against surface observations of PM_{2.5} from EPA stations in 2019 and 2020. We also applied the cubic convolution interpolation method to resample all variables in the data products described in Sections 2.2.3 and 2.2.4 to a spatial resolution of 4 km. The mean CMAQ PM_{2.5} concentrations during the February-June 2019 over the CONUS is shown in Figure 4.

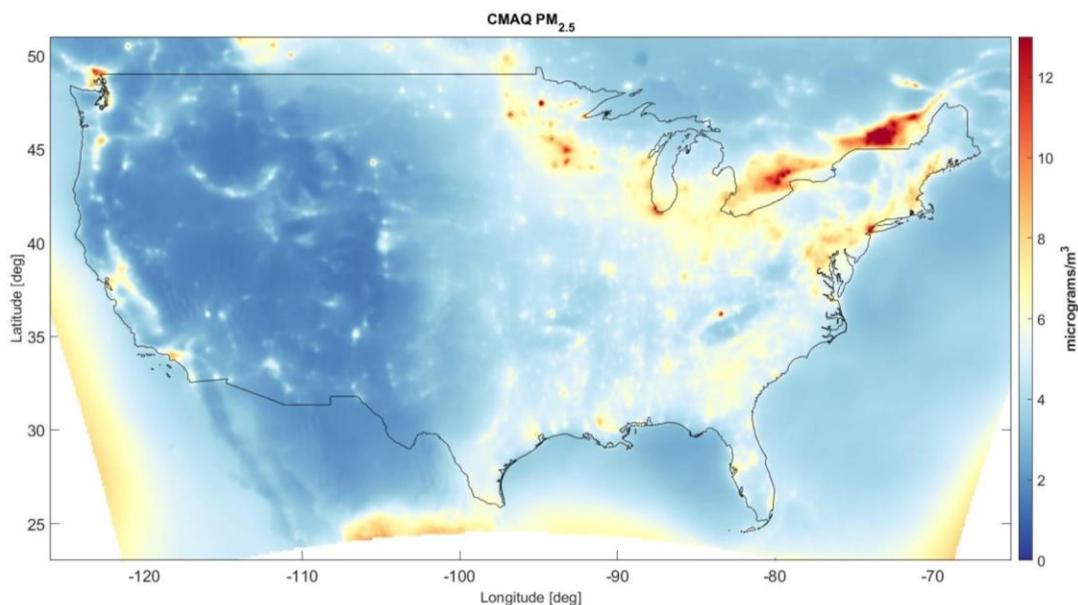


Figure 4. Mean PM_{2.5} concentrations generated using CMAQ model during February-June 2019 over the CONUS.

2.2.3. North American Land Data Assimilation System (NLDAS)

NLDAS is a project that combines satellite and ground-based observational data with advanced land surface models to provide comprehensive and accurate reanalysis data of land surface conditions in North

America at a 0.125° spatial resolution. The required meteorological variables over Texas for the 2018-2022 period were downloaded from the LAADS & Distribution System Distributed Active Archive Center (DAAC) (<https://ladsweb.modaps.eosdis.nasa.gov/>). We prepared daily meteorological fields, including air temperature, surface pressure, specific humidity, U and V components of wind speed (UWind and VWind, respectively), downward shortwave/longwave radiation flux (SRad and LRad, respectively), and convective available potential energy (CAPE) at 4 km resolution from NLDAS, using the resampling process described earlier. Figure 5 illustrates an example of NLDAS datasets, specifically displaying NLDAS air temperature data in 2019 that has been resampled to a 4 km spatial resolution across the CONUS.

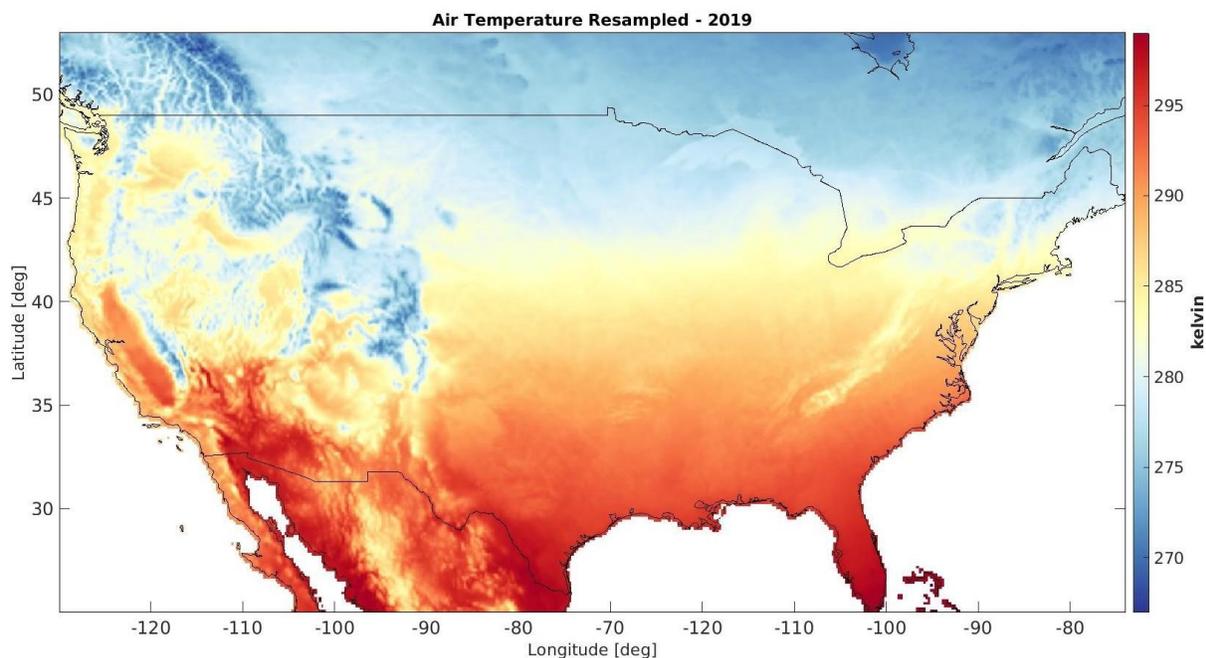


Figure 5. NLDAS air temperature data resampled to 4 km spatial resolution over the CONUS in 2019.

2.2.4. Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2)

The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) products provide hourly estimates of atmospheric, oceanic, and land surface conditions spanning from 1980 to the present at a resolution of 0.5°×0.625°. The MERRA-2 datasets for the 2018-2022 period were downloaded from the LAADS DAAC. We prepared daily planetary boundary layer height (PBLH), surface layer height (SLH), and surface concentrations of SO₄, organic carbon (OC), black carbon (BC), dust, and sea salt from MERRA-2. As MERRA-2 does not offer direct PM_{2.5} data, we calculated surface PM_{2.5} concentrations using the below equation from Provençal et al. (2017). Notably, we excluded gridded nitrate concentrations from the calculation, as the MERRA-2 dataset does not provide this data across the study area. The Spatial distribution of MERRA-2 PM_{2.5} resampled to 4 km spatial resolution in 2019 is shown in Figure 6.

$$\text{MERRA PM}_{2.5} = (1.375 \times \text{SO}_4) + (1.8 \times \text{OC}) + \text{BC} + \text{Dust} + \text{Sea salt} \quad (1)$$

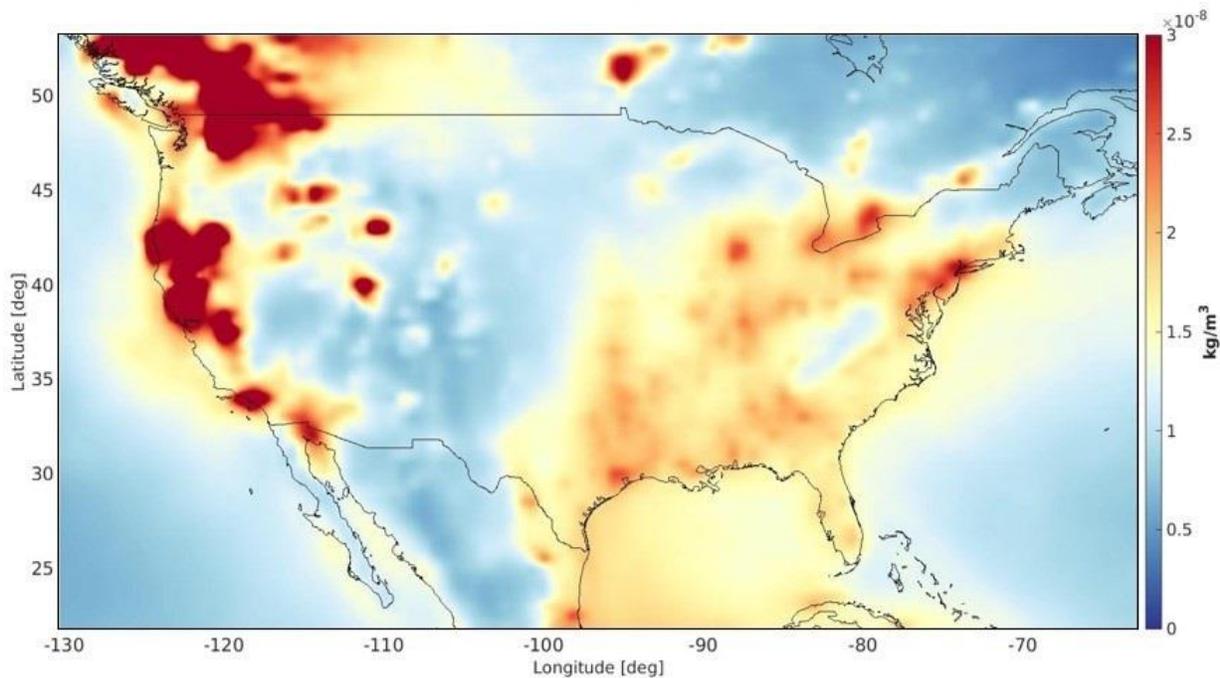


Figure 6. MERRA-2 PM_{2.5} concentrations resampled to 4 km spatial resolution over the CONUS in 2019.

2.2.5. Other Variables

The road density (RD) data, expressed in kilometers per square kilometer (km/km²) and shown in Figure 7, was sourced from the Global Roads Inventory Project (Meijer et al., 2018). This dataset, which has a spatial resolution of 8 km, exclusively represents the density of highways in each pixel, omitting other road types. Figure 8 showcases population density (PD) data, expressed as the number of people per square kilometer, which was obtained from the NASA Socio-economic Data and Applications Center (SEDAC) (<https://sedac.ciesin.columbia.edu/>) and derived at a spatial resolution of 4 km. Additionally, Figure 9 features the PUS data, also at a 4 km resolution, compiled from the U.S. National Land Cover Database (NLCD) (<https://www.mrlc.gov/data/nlcd-2019-land-cover-conus>). This database details the extent of land development across the CONUS. Surface elevation data, presented in Figure 10, was acquired from the GTOPO30 dataset by the United States Geological Survey (USGS) (<https://earthexplorer.usgs.gov/>), offering a global digital elevation model at a spatial resolution of 1 km, covering the entire Earth.

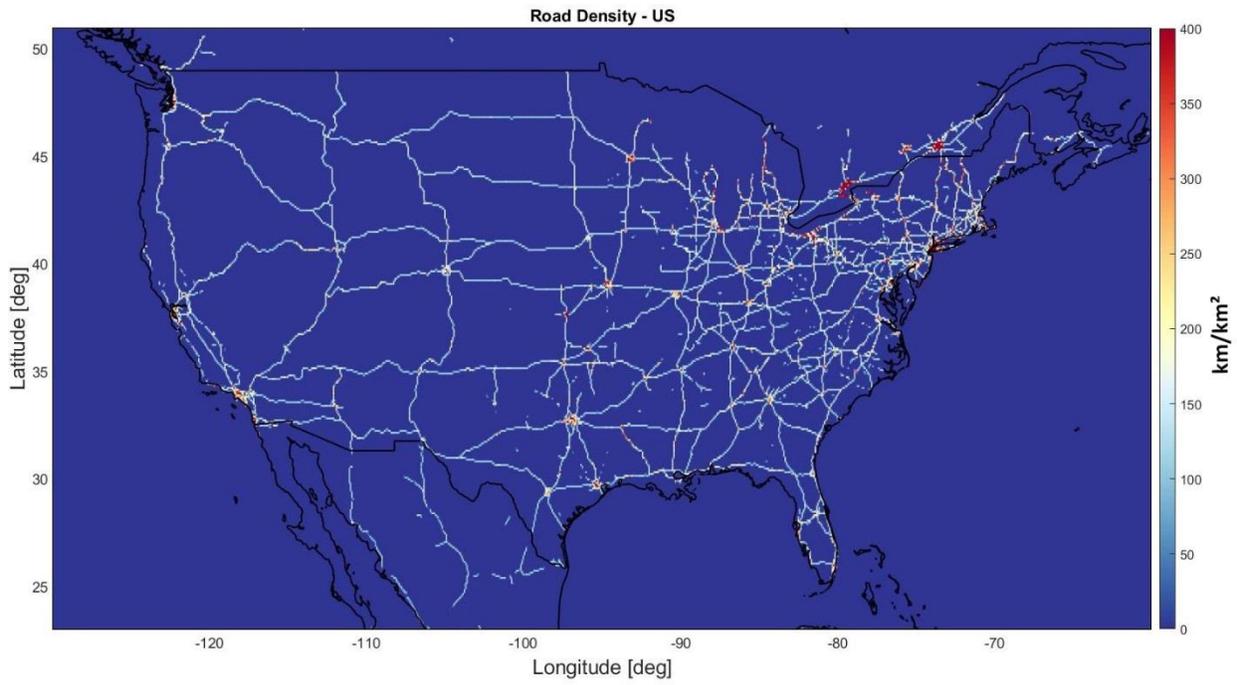


Figure 7. Road density (RD) data showing the density of highways (km/km²) in each pixel over the CONUS.

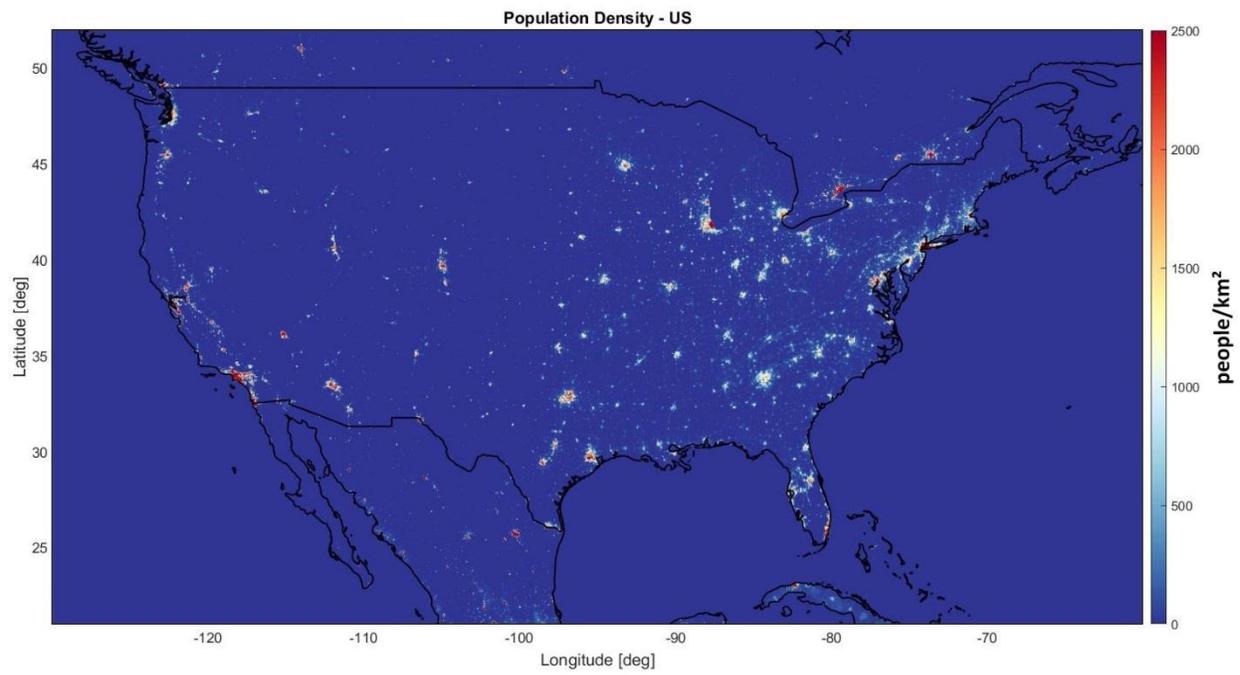


Figure 8. Population density (PD) (number of people/km²) over the CONUS in 2020.

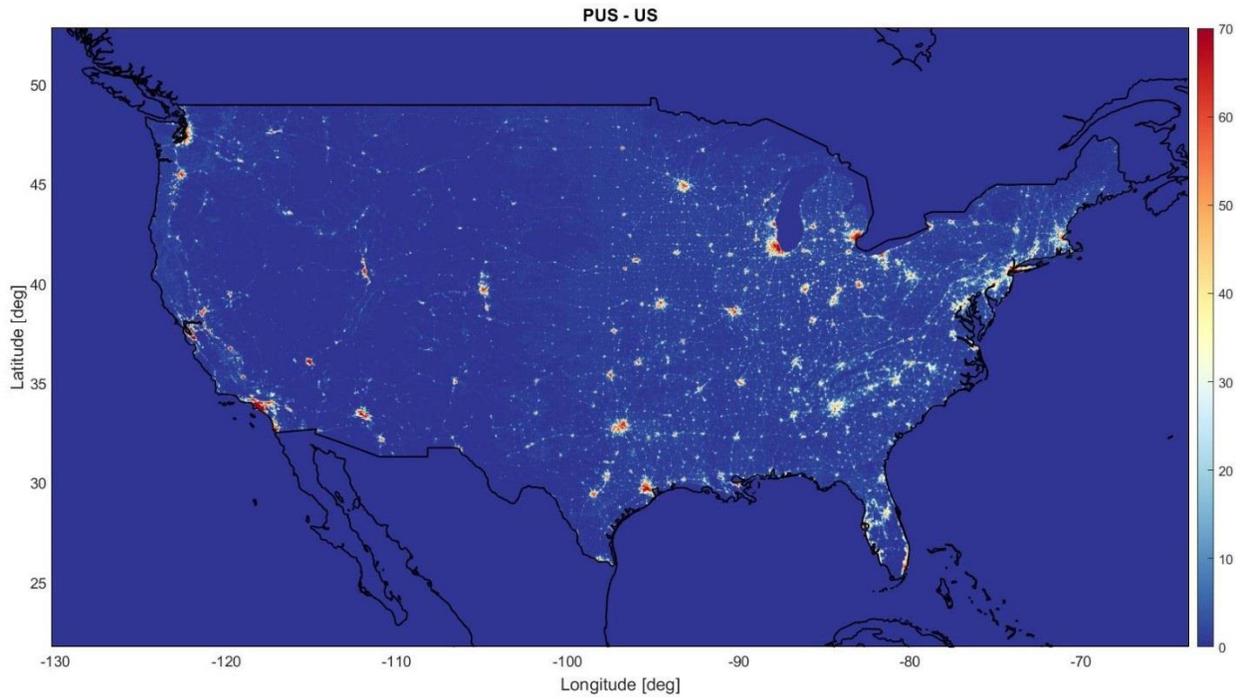


Figure 9. Percentage of urban space (PUS) over the CONUS in 2019.

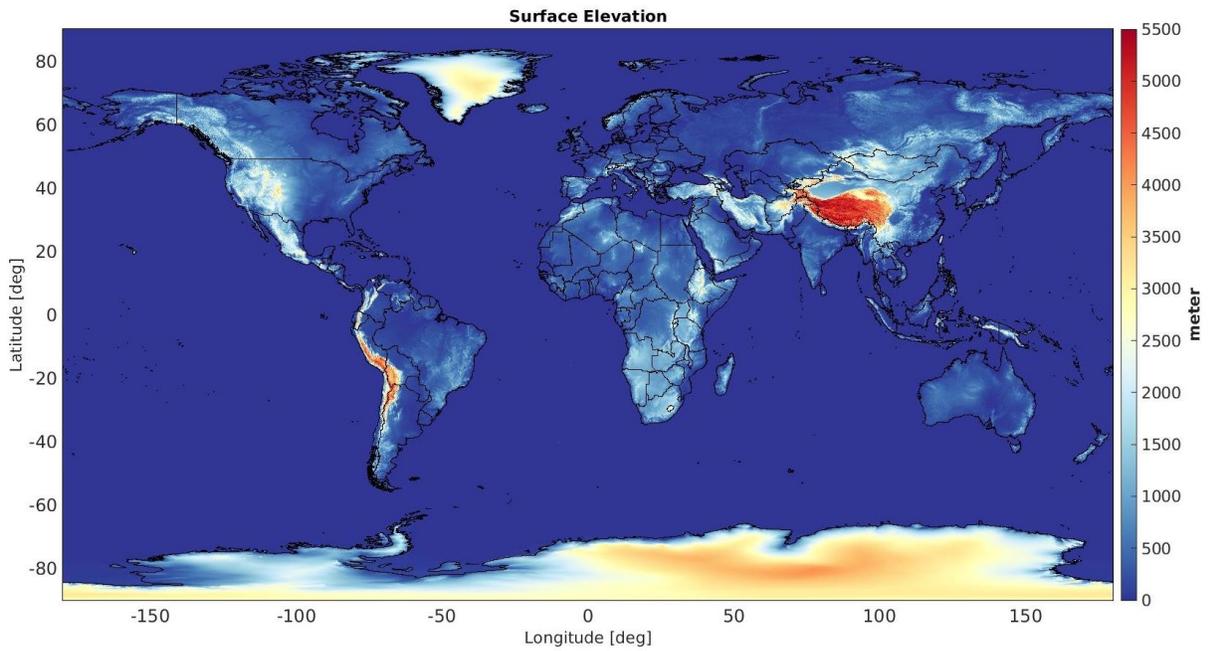


Figure 10. Surface elevation data over the globe.

3. Methodology

The UH-AQF AI group utilized a two-phase DL framework to estimate daily surface $PM_{2.5}$ concentrations at 4 km spatial resolution over Texas from 2018 to 2022. The initial phase employs a depthwise PCNN and the variables of MODIS AOD, CMAQ $PM_{2.5}$, and PUS to fill the gaps between EPA $PM_{2.5}$ stations, creating gap-free daily $PM_{2.5}$ grids (hereafter referred to as PCNN- $PM_{2.5}$) from 2018 to 2022. Subsequently, a deep convolutional neural network (DeepCNN) was applied in the second phase. This DeepCNN utilized additional predictor variables such as surface elevation, MERRA $PM_{2.5}$, MODIS EVI, PD, PUS, RD, and various meteorological parameters to refine and bias-correct the output from the first phase (i.e., PCNN- $PM_{2.5}$). The result was a set of highly accurate, gap-free daily grids of surface $PM_{2.5}$ concentrations from 2018 to 2022. Figure 11 depicts the schematic structure of the PCNN-DeepCNN model employed in this study. Furthermore, subsequent sections offer comprehensive details on the programming platforms, virtual environments, the PCNN-DeepCNN model, and other exercises implemented in this project.

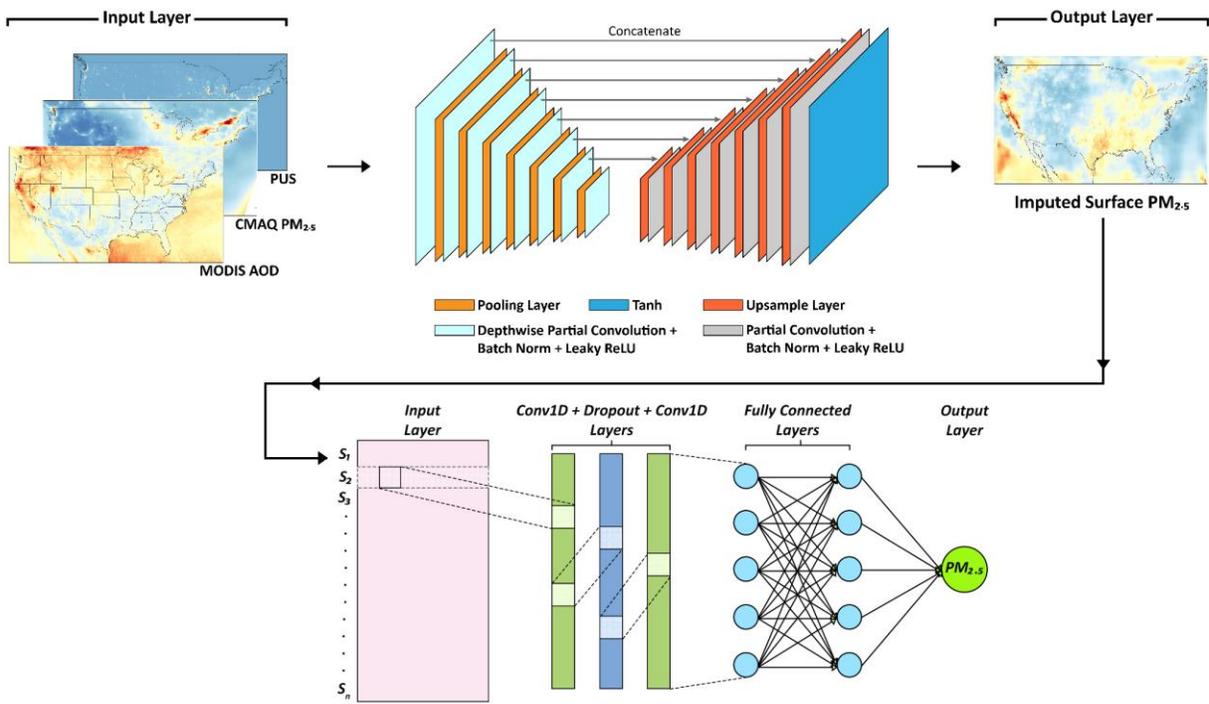


Figure 11. The schematic structure of the PCNN-DeepCNN model used for surface $PM_{2.5}$ estimation. The $S_1, S_2, S_3 \dots S_n$ refer to the predictor variables used in the second phase (DeepCNN) of the model.

3.1. Software

For this project, we employed both MATLAB and Python programming languages to optimize our workflow. MATLAB was primarily utilized for preprocessing and preparing the datasets, ensuring they were ready for analysis. In contrast, Python was the main tool for developing the DL models and generating outputs. This dual-language approach leveraged the strengths of each platform to efficiently

handle different aspects of our computational tasks. We also established a dedicated virtual environment on our local servers to facilitate the development of Python codes and DL models. Typically, developing a DL model in Python involves a few basic steps: installing Python, adding required libraries via the terminal, coding in a single “.py” file or a notebook, and executing the program from the terminal. This approach suffices for straightforward Python scripting. However, more complex DL projects, such as the estimation of surface PM_{2.5} levels using DL, often require managing multiple files, packages, and dependencies. To effectively handle these complexities, it becomes essential to develop and isolate an environment specific to the project. Thus, the creation of a unique virtual environment was imperative for our work. A virtual environment in Python is a self-contained directory that houses a specific version of the Python interpreter along with various additional libraries, independent of those installed system-wide. This setup prevents conflicts between project dependencies and system-installed libraries.

Within the dedicated Python virtual environment we established for this project, we installed several essential libraries. A library is a curated collection of classes, methods, and modules that application codes can leverage to perform specific tasks without needing to develop these functionalities from scratch. Libraries are often specialized (e.g., handling strings, input/output operations, and network connections), making their Application Programming Interfaces (APIs) more focused and reducing dependencies. Essentially, libraries offer a repository of class definitions that enhance code reusability. The term "Code Reusability" refers to the practice of using pre-written code to achieve desired functionalities in new projects, thus allowing developers to focus more on solving specific problems rather than reinventing basic routines. For example, some libraries feature a `findLastIndex()` function that can locate the last occurrence of a character in a string, which developers can readily use by calling `findLastIndex(charToFind)` with the character of interest as a parameter. Below are examples of key libraries installed in our virtual environment, which support a range of data manipulation, analysis, and ML tasks:

- **Pandas:** Developed under the Berkeley Software Distribution, this open-source library is crucial in data science for data manipulation, analysis, and cleaning, often serving as an alternative to using languages like R for data tasks.
- **NumPy:** Known for its robust capabilities in scientific computation, NumPy supports extensive matrix operations and mathematical computations, making it indispensable for linear algebra and serving as a multidimensional container for large data arrays.
- **SciPy:** This library extends NumPy's capabilities by adding advanced mathematical functions, optimization, regression, and more. It's designed for efficient high-level computations integral to scientific computing.
- **Scikit-learn:** As a comprehensive ML toolkit, this library offers a wide array of algorithms for classification, regression, and clustering, including support-vector machines and RFs, and is designed to interoperate with NumPy and SciPy.
- **Keras:** Developed by Google, Keras facilitates DL by providing an intuitive API for constructing and training neural networks, and it's known for its user-friendliness and modularity.
- **TensorFlow:** Created by Google Brain, this library is a cornerstone in performing high-performance numerical computations used extensively in ML and DL, capable of conducting complex calculations essential for researchers in mathematics, physics, and AI.

These libraries not only streamline the development process but also significantly enhance our project's capabilities to tackle complex data-driven challenges.

3.2. First Phase of the PCNN-DeepCNN Model: Depthwise PCNN

The standard version of the PCNN model was developed based on a U-Net CNN architecture (Ronneberger et al., 2015) which has been widely used for tasks where input and output have a one-to-one correspondence, such as image inpainting. U-Net models offer several benefits relevant to the research: (a) have a straightforward structure, (b) allow global localization and context, and (c) preserve the full context of input images (Alom et al., 2018). The vanilla PCNN model utilizes a U-Net-like encoder-decoder architecture but replaces convolutional layers with partial convolutional layers, allowing only valid pixel processing. It also incorporates skip connections similar to U-Net to combat the loss of spatial information caused by downsampling in the encoding layers. These connections transmit the output of an encoding layer to the corresponding decoding layer, enabling the network to upsample encoded information efficiently (Liu et al., 2018). In place of padding, the model uses partial convolution padding (Liu et al., 2018) with the appropriate masking at the edges of the image. This ensures that values outside the boundaries are treated as missing or holes. It also guarantees that incorrect values outside the image do not affect the inpainted content at the edges (Liu et al., 2018). Depthwise PCNN, a more advanced version of standard PCNN, introduces depthwise convolution (Liu et al., 2021; Lops et al., 2022). Unlike standard PCNNs, where the convolution operation is performed across all input channels for each filter, in the depthwise PCNNs, each filter operates on a single input channel, which reduces the parameter count compared to conventional PCNN, thereby streamlining the network's complexity, and enhancing parameter efficiency. This is particularly beneficial when training data is limited or computational resources are constrained (Liu et al., 2021). Furthermore, the depthwise PCNN can learn more channel-specific features due to the individual operation of each filter on a single input channel, potentially leading to better performance when different input channels contain diverse types of information (Lops et al., 2022).

Before training our depthwise PCNN, we standardized our three main input variables (CMAQ PM_{2.5}, MODIS AOD, and PUS) to a uniform pixel resolution of 778×1456 to ensure consistency across the datasets. We applied data augmentation to enhance the model's generalization ability across various spatial contexts. We extracted 256×256 pixels patches from the larger matrices of 778×1456 pixels, creating overlapping patches with varied strides to diversify training data. The resulting 256×256 images served as the input to our depthwise PCNN, with corresponding masks applied to simulate the actual distribution of air quality monitoring stations. These masks were randomly augmented to retain a subset of stations, mirroring the actual sparse and irregular distribution of air quality sensors. This approach allowed us to create a training dataset that closely mimics the partial availability of PM_{2.5} data due to the uneven distribution of monitoring infrastructure. By training the depthwise PCNN with this augmented dataset, the model is expected to be better equipped for estimating PM_{2.5} levels, accurately reflecting the variability and spatially incomplete nature of observed air quality data. The trained depthwise PCNN model was utilized to fill the gaps between EPA PM_{2.5} monitoring stations on a daily basis. Since the depthwise PCNN utilizes the U-Net architecture, it can be trained on input resolutions of 256×256 pixels

and then applied to images of varying resolutions during the final estimation. Thus, we perform the final estimation at a 778×1456-pixel resolution in order to avoid irregularities that could occur when combining overlapping patches. Consistent with the training phase, the imputation process has three input predictor variables: MODIS AOD, PUS, and grid-based station PM_{2.5} measurements, which replace CMAQ PM_{2.5} used in the training phase. Unlike previous studies that relied on more straightforward methods like Kriging, inverse distance weighting (IDW), and RF, for air pollutant interpolation (Di et al., 2016; Ghahremanloo et al., 2021; Hu et al., 2017) the depthwise PCNN approach offers increased accuracy in imputation tasks due to its effectiveness in capturing spatial pollutant patterns, as highlighted by Lops et al. (2021).

3.3. Second Phase of the PCNN-DeepCNN Model: DeepCNN

After filling the gaps between EPA PM_{2.5} stations and preparing the PCNN-PM_{2.5} explained in the previous section, we fed PCNN-PM_{2.5} along with other predictor variables into a DeepCNN model to bias correct daily PCNN-PM_{2.5} grids. The DeepCNN integrates convolutional and dense layers, starting with a 1D convolution layer (16 filters with ReLU activation function), a dropout layer to prevent overfitting, a flatten layer, and two dense layers (500 and 265 neurons with Swish activation function). The regression-focused output layer has a single neuron. The hyper-parameters of the DeepCNN model (e.g., number of neurons, filters, and dropout percentage) were obtained using the Optuna hyperparameter optimization framework (Akiba et al., 2019). We used RMSprop with an adaptive learning rate and momentum to optimize the model and used a custom loss function based on the index of agreement (IOA). The optimum hyperparameters selected by Optuna for the DeepCNN model are shown in Table 1.

Table 1. Hyperparameters of the DeepCNN model selected by the Optuna framework.

| Hyperparameter | Number/Description |
|--|--|
| Number of convolutional layers | 1 |
| Activation function in convolutional layer | Relu |
| Dropout layer in convolutional layer | One layer with 0.065 drop rate |
| Number of filters | 16 |
| Kernel size | 2 |
| Number of dense layers | 2 |
| Number of neurons in dense layers | 500 and 265 neurons in the first and second layers |
| Activation functions in dense layers | Relu and Swish in the first and second layers |
| Dropout layer in dense layers | One dropout (0.16) layer in the first layer |
| Learning rate | 0.000255 |
| Optimizer | RMSprop |
| Batch size | 952 |

After training the PCNN-DeepCNN model, our team used it to create the daily mean surface PM_{2.5} maps at 4 km spatial resolution over Texas in the 2018-2021 period. The main code for the map creation phase was developed in Python programming platform and the application of parallel computing in our powerful

servers significantly reduced the running time for the map creation phase. The code uses the DL models trained specifically for each year from 2018 to 2021 and incorporates the same predictor variables used in the training phase, but in a 2D grid format, to estimate daily PM_{2.5} levels over the study area. Parallel computing involves breaking down significant problems into smaller, autonomous, and usually similar segments that can be concurrently processed by multiple processors. These processors communicate through shared memory, and upon finishing their tasks, their outcomes are merged as part of a comprehensive algorithm. The primary objective of parallel computing is to enhance the accessible computational capability, leading to quicker application processing and more efficient problem-solving. In order to implement parallel computing for PM_{2.5} map creation, our team used the *multiprocessing* package in Python. This module is designed to facilitate the creation of processes using an interface comparable to the threading module. The *multiprocessing* package provides the capability for both local and remote concurrency, bypassing the Global Interpreter Lock by employing sub-processes in lieu of threads. As a result, programmers can make optimal use of multiple processors available on a specific machine. The multiprocessing module is compatible with both UNIX and Windows operating systems.

3.4. Feature Engineering

We employed a two-step feature engineering approach to address the multicollinearity and select the best combination of predictor variables for our DL model. In the initial step, we identified highly correlated variables by calculating the variance inflation factor (VIF) of each predictor. We excluded the predictor variables with VIFs greater than ten, the threshold value of which was set based on the premise that the variables with a VIF higher than ten were considered to exhibit significant multicollinearity with other variables (Ghahremanloo et al., 2023; Kline, 2015). Eliminating these highly correlated predictors ensures that each variable included contributes uniquely to the estimation of PM_{2.5} concentrations without being redundantly influenced by correlations with other predictors, thereby enhancing our model's accuracy and reliability. We then utilized SHapley Additive exPlanations (SHAP) feature importance (Lundberg & Lee, 2017) to determine the optimal combination of predictor variables for our model. SHAP evaluates feature importance by comparing model estimations with and without each feature. As its estimates, however, are influenced by the order in which the model captures each feature, this process is implemented in every possible order so that the features can be fairly compared (Lundberg and Lee, 2017). Subsequently, we excluded less important predictor variables based on their SHAP feature importance scores of the selected parameters in pre-trained models. Specifically, four predictor variables—air temperature, specific humidity, SLH, and CAPE—were identified as redundant and were consequently eliminated from the model during the feature engineering process. The final predictor variables employed in the DeepCNN phase comprise PCNN-PM_{2.5} (the first phase's output), SRad, LRad, UWind, VWind, surface pressure, PBLH, Precipitation, MERRA PM_{2.5}, MODIS EVI, surface elevation, PD, PUS, and RD.

3.5. Model Evaluation

We evaluated the accuracy of our PCNN-DeepCNN model for surface PM_{2.5} estimation using both ten-fold cross-validation (10-CV) and spatial cross-validation (Spatial-CV) techniques. In 10-CV, the dataset is divided randomly into ten non-overlapping clusters, with the model trained on nine clusters and tested

on the remaining one. This process is repeated ten times, allowing each cluster to serve as the test set once. This comprehensive approach ensures that every data point is used for validation at least once, providing a thorough assessment of the model's generalization ability. The schematic workflow of the 10-CV approach is depicted in Figure 12. In contrast, Spatial-CV is specifically employed to evaluate the spatial performance of our model. This method divides samples based on their monitoring stations and is crucial for assessing how well the model estimates $PM_{2.5}$ concentrations across different spatial areas, considering the spatial autocorrelation inherent in environmental data. Model performance is evaluated using several metrics, including R, IOA, MAB, and root mean square error (RMSE), which compare the surface $PM_{2.5}$ estimated by our model against station observations over Texas. High values of R and IOA indicate strong agreement between the estimated and observed $PM_{2.5}$ concentrations, indicating high accuracy. Meanwhile, MAB and RMSE quantify the average magnitude and variability of errors. Additionally, we assessed the performance of the DeepCNN model across four seasons to evaluate its performance under seasonal variations. Below are the formulas used to calculate the evaluation metrics in this project, where E_i is estimated samples, O_i the observed samples, \bar{E} the mean estimated samples, \bar{O} the mean observed samples, and N is a number of samples.

$$R = \frac{\sum_{i=1}^n (E_i - \bar{E}) \cdot (O_i - \bar{O})}{\sqrt{\sum_{i=1}^n (E_i - \bar{E})^2} \cdot \sqrt{\sum_{i=1}^n (O_i - \bar{O})^2}} \quad (2)$$

$$IOA = 1 - \frac{\sum_{i=1}^n (E_i - O_i)^2}{\sum_{i=1}^n (|E_i - \bar{E}| + |O_i - \bar{O}|)^2} \quad (3)$$

$$MAB = \frac{1}{N} \sum_{i=1}^n |E_i - O_i| \quad (4)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^n (E_i - O_i)^2} \quad (5)$$

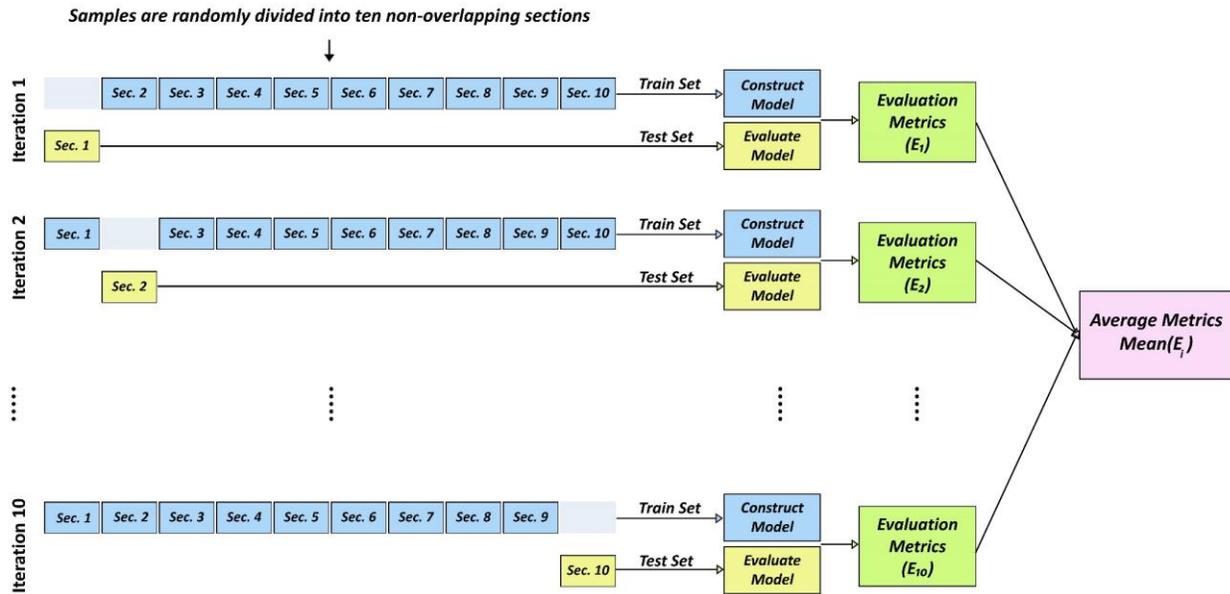


Figure 12. The schematic structure of the ten-fold cross-validation (10-CV) approach for evaluating the performance of the Deep Convolutional Neural Network (Deep-CNN) model at surface PM_{2.5} estimation.

4. Results and Discussions

4.1. Performance of the First Phase of the PCNN-DeepCNN Model

Table 2 details the performance of the PCNN model in filling the gaps between EPA PM_{2.5} stations and creating the daily PCNN-PM_{2.5} grids across the CONUS from 2018 to 2022. The number of valid pixels with station observations inside accounted for less than 0.1% of the total number of pixels in the input, indicating the scarcity of in-situ measurement data. Despite the limited number of stations, the PCNN model demonstrated reliable accuracy, achieving R ranging from 0.35 to 0.60, MAB between 2.00 and 12.58 $\mu\text{g}/\text{m}^3$, and an RMSE from 3.11 to 15.08 $\mu\text{g}/\text{m}^3$. Moreover, the IOA ranged from 0.59 to 0.86, underscoring the model estimates' strong agreement with observed data. The lower accuracy for 2020 compared to other years can be attributed to the high number of extreme events like the 2020 California wildfires and COVID-19 lockdown, which the PCNN model could not capture properly. The overall results suggest that, even with limited spatial coverage of input data, the PCNN could robustly estimate PM_{2.5} concentrations with fair accuracy, highlighting its potential for broad applications in air quality monitoring and environmental research. It is important to note that the results in Table 2 are not the final accuracy of the model and they only show the accuracy of the depthwise PCNN, the first phase of the PCNN-DeepCNN model.

Table 2: Validation Results of the output of the first phase of the model, the imputed surface PM_{2.5} Levels (PCNN-PM_{2.5}), over the CONUS from 2018 to 2022. R, IOA, MAB, and RMSE refer to the correlation coefficient, the index of

agreement, the mean absolute bias, and the root mean square error, respectively. MAB and RMSE are in the unit of $\mu\text{g}/\text{m}^3$.

| Year | R | IOA | MAB | RMSE |
|------|------|------|-------|-------|
| 2018 | 0.59 | 0.82 | 4.12 | 5.65 |
| 2019 | 0.60 | 0.86 | 2.00 | 3.11 |
| 2020 | 0.35 | 0.59 | 12.58 | 15.08 |
| 2021 | 0.52 | 0.73 | 6.24 | 7.76 |
| 2022 | 0.40 | 0.78 | 2.74 | 4.83 |

4.2. Performance of the Second Phase of the PCNN-DeepCNN Model

Table 3 presents the results of 10-CV and Spatial-CV for the DeepCNN model, which demonstrates the final accuracy of the PCNN-DeepCNN model in estimating daily surface $\text{PM}_{2.5}$ concentrations across Texas from 2018 to 2022. The 10-CV data indicate a substantial improvement by the DeepCNN in reducing bias compared to the initial phase of the model. Specifically, the MAB was reduced from a range of 2.0-12.58 $\mu\text{g}/\text{m}^3$ to 1.3-1.79 $\mu\text{g}/\text{m}^3$, and the RMSE decreased from 3.11-15.08 $\mu\text{g}/\text{m}^3$ to 1.87-3.25 $\mu\text{g}/\text{m}^3$. Additionally, R and IOA saw significant enhancements, improving from 0.35-0.60 to 0.89-0.93 for R, and from 0.59-0.86 to 0.94-0.96 for IOA in the transition from the PCNN (first phase) to the DeepCNN (second phase) model. Despite the relatively lower performance of the PCNN model in 2020, the incorporation of DeepCNN in the second phase markedly improved its accuracy, effectively correcting biases. These improvements underscore the PCNN-DeepCNN model's capability to produce accurate, gap-free grids of daily surface $\text{PM}_{2.5}$ levels at a 4 km spatial resolution over Texas.

The Spatial-CV results also underscore the advanced spatial accuracy of our DL system in estimating surface $\text{PM}_{2.5}$ levels. In the Spatial-CV, the model exhibited consistent R values, ranging from 0.81 to 0.87. Similarly, IOA values were robust, varying from 0.89 to 0.93, which closely aligns with the results from the overall cross-validation. The notable improvement in performance from the PCNN model in the first phase to the DeepCNN model in the second phase can be attributed primarily to the bias correction achieved through the integration of additional predictor variables by DeepCNN. While the PCNN phase was aimed at imputing missing $\text{PM}_{2.5}$ data, it did not adequately address systematic biases or external factors that influence $\text{PM}_{2.5}$ levels. This was particularly evident in 2020, a year marked by significant disruptions such as widespread wildfires on the West Coast and the COVID-19 pandemic, which significantly altered air quality patterns by reducing emissions in some regions and changing patterns of human and industrial activity (Ghahremanloo et al., 2022). Figure 13 presents a scatter plot that compares estimated $\text{PM}_{2.5}$ concentrations from the PCNN-DeepCNN model with observed $\text{PM}_{2.5}$ levels across the study area throughout the period from 2018 to 2022, further highlighting the high accuracy of the model at $\text{PM}_{2.5}$ estimation.

Table 3: Validation Results of the PCNN-DeepCNN model at estimating surface $\text{PM}_{2.5}$ levels over Texas from 2018 to 2022. R, IOA, MAB, and RMSE refer to the correlation coefficient, the index of agreement, the mean absolute bias, and the root mean square error, respectively. MAB and RMSE are in the unit of $\mu\text{g}/\text{m}^3$.

| | 10-CV | 10-CV | 10-CV | 10-CV | Spatial-CV | Spatial-CV | Spatial-CV | Spatial-CV |
|------|-------|-------|-------|-------|------------|------------|------------|------------|
| Year | R | IOA | MAB | RMSE | R | IOA | MAB | RMSE |
| 2018 | 0.92 | 0.96 | 1.66 | 2.73 | 0.86 | 0.92 | 2.07 | 3.73 |
| 2019 | 0.91 | 0.95 | 1.30 | 1.87 | 0.84 | 0.91 | 1.72 | 2.61 |
| 2020 | 0.93 | 0.96 | 1.79 | 3.25 | 0.87 | 0.93 | 2.15 | 5.51 |
| 2021 | 0.91 | 0.95 | 1.79 | 2.68 | 0.87 | 0.93 | 2.06 | 3.78 |
| 2022 | 0.89 | 0.94 | 1.51 | 2.76 | 0.81 | 0.89 | 1.84 | 3.43 |

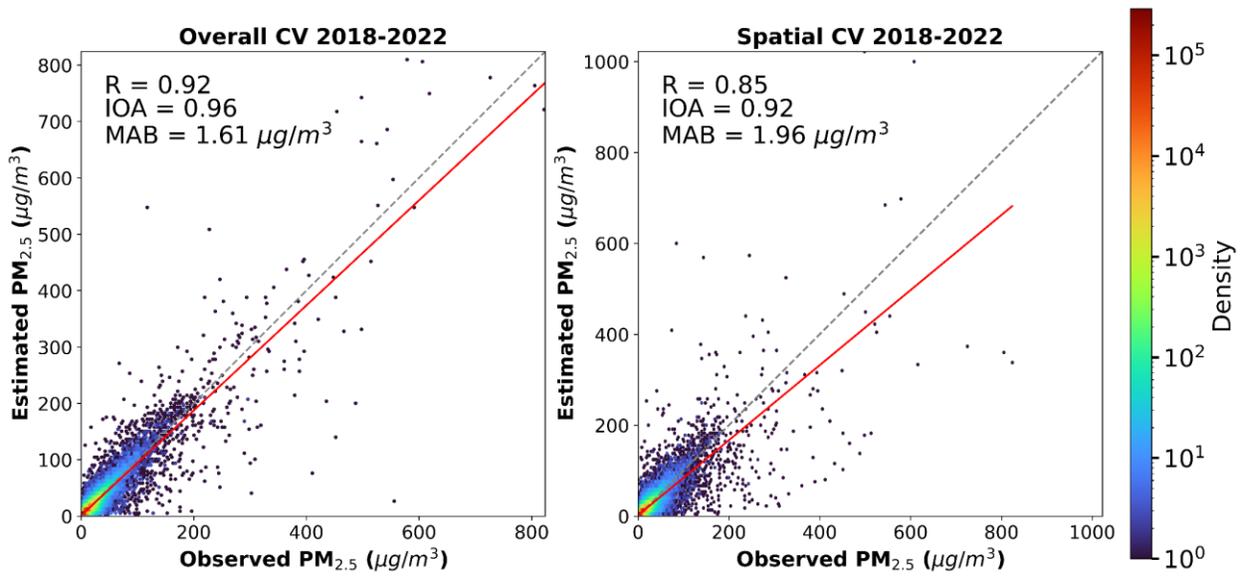


Figure 13. Scatterplots comparing estimated and observed surface $PM_{2.5}$ levels across the study area averaged over 2018-2022. The color bar represents the density of samples within the scatterplots, with warmer colors (red) indicating higher data point concentrations and cooler colors (blue) showing areas with fewer data points.

In evaluating the performance of our PCNN-DeepCNN model, a comparative analysis with existing studies reveals insightful contrasts. [Ghahremanloo et al. \(2021\)](#) employed an RF algorithm to estimate ground-level $PM_{2.5}$ concentrations across Texas for the period between 2014 and 2018, achieving 10-CV R ranging between 0.83 and 0.90, with an MAB of 1.47 to 1.77 $\mu g/m^3$. However, their model was unable to handle the missing values in the MAIAC AOD images, resulting in $PM_{2.5}$ maps that also contained gaps. Compared to [Park et al. \(2020\)](#), who recorded a 10-CV R of 0.90 at $PM_{2.5}$ estimation using an RF model, our DL model again demonstrated relatively superior performance. [Park et al. \(2020\)](#) also highlighted the effectiveness of CNN models over RF in specific scenarios and the incremental benefit of ensemble methods like CNN-RF in spatially stratified cross-validation. This comparison not only underscores the efficacy of our model but also emphasizes its potential applicability in diverse environmental and temporal settings. An important advantage of our model is its spatiotemporal consistency and ability to eliminate missing

values, enhancing data reliability. This feature, coupled with its high accuracy, sets it apart from other methods that are susceptible to data gaps.

Figure 14 presents a comparative seasonal performance of the PCNN-DeepCNN model for $PM_{2.5}$ estimation. The model achieved the best performance in summer with an IOA of 0.92. Winter had the highest MAB at $1.78 \mu\text{g}/\text{m}^3$, indicating some inconsistencies likely due to seasonal variability. The lowest MAB was in spring at $1.32 \mu\text{g}/\text{m}^3$, highlighting the model's better estimation accuracy during this season. Our analysis revealed that the winter season posed the most significant challenge for $PM_{2.5}$ estimation. This observation aligns with the results of [Di et al. \(2016\)](#) and [Park et al. \(2020\)](#), who also reported lower R values during winter. The lower winter performance in our study may be attributed to factors such as complex atmospheric conditions, including temperature inversions ([Di et al., 2016](#)), increased emission sources, and changes in the chemical composition of $PM_{2.5}$ ([Wei et al., 2019](#)).

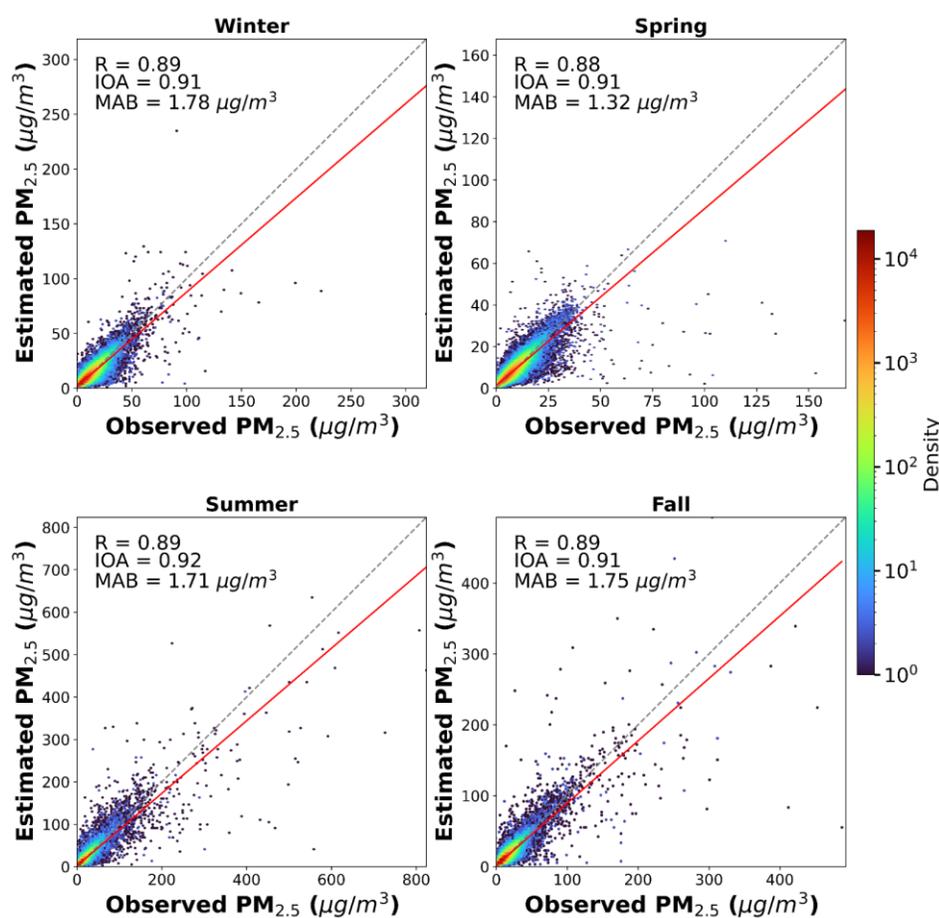


Figure 14. Scatterplots comparing estimated and observed surface $PM_{2.5}$ levels across the study area, segmented by seasons. The color bar represents the density of samples within the scatterplots, with warmer colors (red) indicating higher data point concentrations and cooler colors (blue) showing areas with fewer data points.

Figure 15 illustrates the comparative analysis of observed and estimated $PM_{2.5}$ concentrations over three sample stations in Texas during 2018, employing the PCNN-DeepCNN model. Our DL model, which was

not previously trained on data from these stations, effectively estimates $PM_{2.5}$ levels, capturing both the trends and actual concentrations with high accuracy. Values of R range from 0.85 to 0.91, indicating a strong agreement between the estimated and observed data. The MAB values, ranging from 1.5 to $2.4 \mu\text{g}/\text{m}^3$, further affirm the model's accuracy in quantifying air quality. This figure underscores the model's robust generalization capabilities, as it performs well even on previously unseen data, a testament to its potential applicability across different geographical locations without the need for retraining. The charts demonstrate how the model closely follows the peaks and troughs of $PM_{2.5}$ fluctuations, suggesting it could be a valuable tool for real-time air quality monitoring and public health advisories. The minimal divergence in MAB across the stations also highlights the consistent performance of the model across varied environmental conditions within Texas.

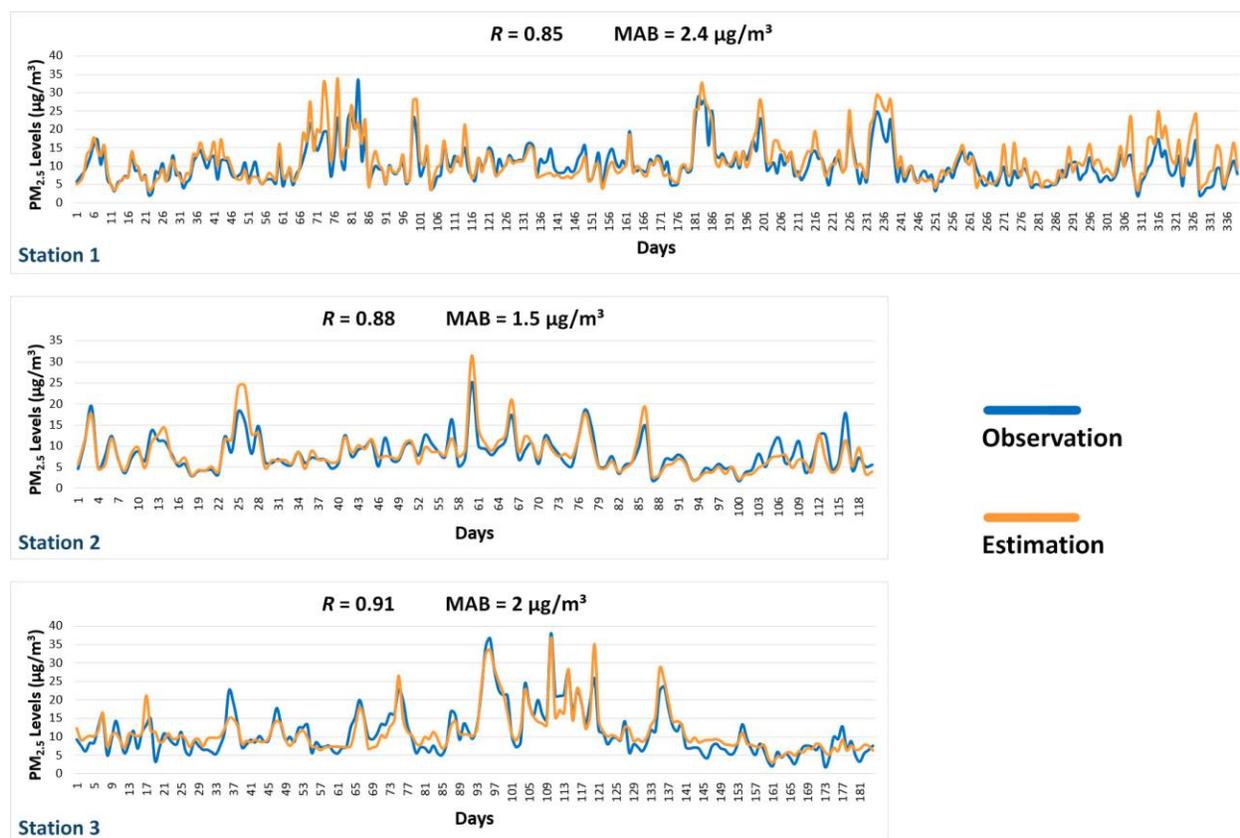


Figure 15. Comparison between the observed and estimated $PM_{2.5}$ concentrations in three sample stations over Texas in 2018. The number of observations at each station is different and stations do not measure $PM_{2.5}$ for the entire 365 days.

Figure 16 shows the estimated $PM_{2.5}$ concentrations over Texas in the 2018-2022 period. According to this figure, $PM_{2.5}$ levels are higher in urban environments, such as Houston, Dallas, and Austin. Eastern and southern regions of Texas also experience higher levels of $PM_{2.5}$, compared to other regions (e.g., western Texas). The higher levels of $PM_{2.5}$ in East Texas might be attributed to a large number of oil refineries, industrial sections, and power plants in this region (Ghahremanloo et al., 2021). Moreover, according to

[Ghahremanloo et al. \(2022\)](#), In 1994, the governments of the United States, Mexico, and Canada signed the North American Free Trade Agreement (NAFTA) to create a free trade zone. Since that time, NAFTA has fostered the growth and accumulation of industrial regions and rapid population growth in regions close to the U.S.-Mexico border ([Karnaes and John, 2019](#)). In light of these changes, the U.S. EPA predicted that air pollution would be a significant problem in regions close to the border ([Karnaes and John, 2019](#)). This can be one main reason for relatively higher $PM_{2.5}$ levels in southern Texas. Figure 16 also plots surface $PM_{2.5}$ observations (i.e., circles in Figure 16) measured in EPA stations on the DL-estimated $PM_{2.5}$ maps over Texas in the 2018-2022 period to evaluate performance of the $PM_{2.5}$ map creation phase. According to this figure, there is a high degree of agreement between observed and estimated $PM_{2.5}$ concentrations, highlighting promising performance of our DL models in the map creation phase. In order to make an apple to apple comparison between the annual mean $PM_{2.5}$ maps and $PM_{2.5}$ observations in Figure 16, we included only those EPA stations which measure $PM_{2.5}$ in at least 340 days at each year.

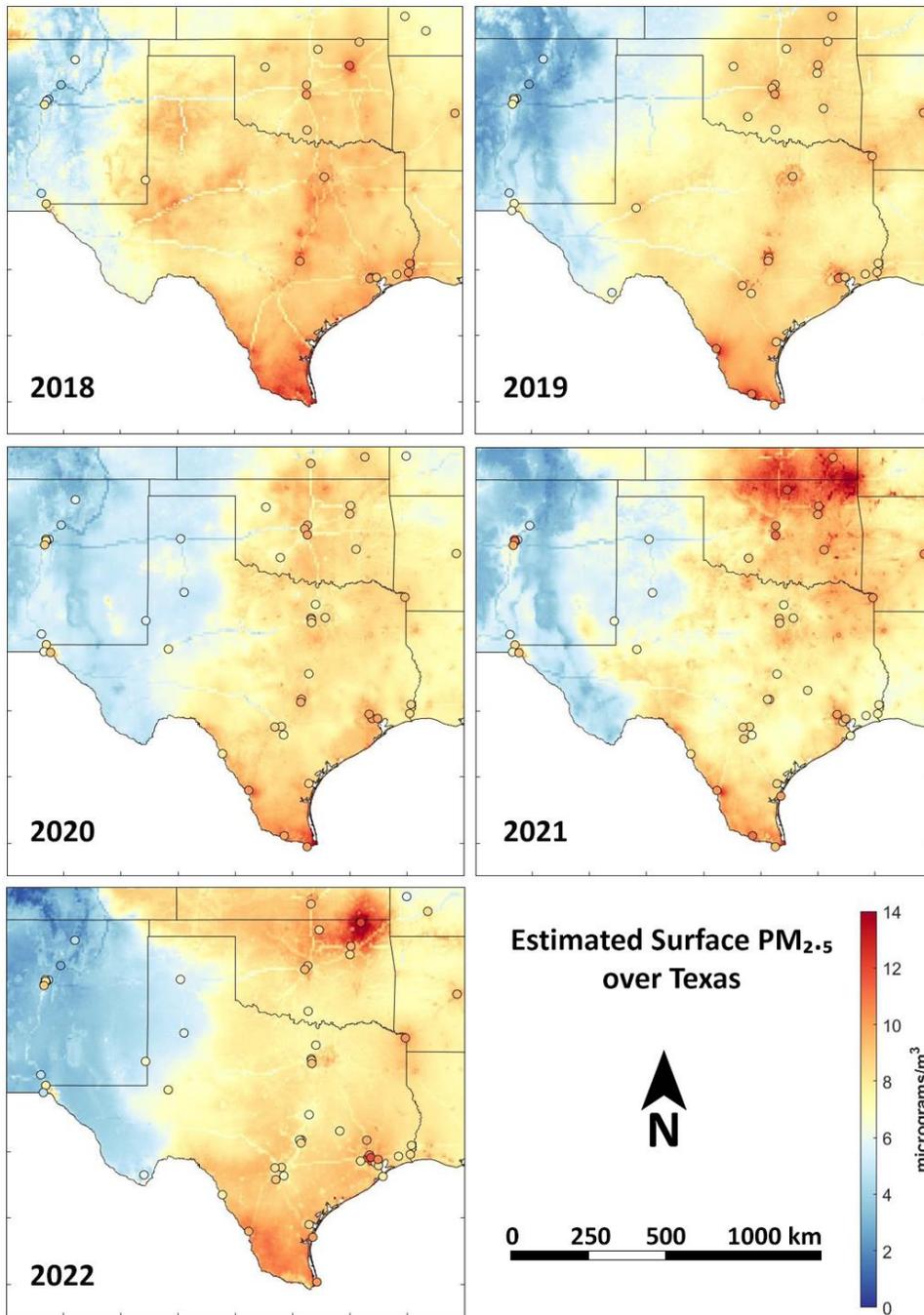


Figure 16. Maps showing the estimated PM_{2.5} levels over Texas from 2018 to 2022 along with the EPA PM_{2.5} observations showed in circles to validate the maps.

Figure 17 compares county-wise DVs calculated using our DL estimated PM_{2.5} grids (left column) (henceforth referred to as DL-DV) with the DVs calculated using EPA monitoring stations (right column) (henceforth referred to as EPA-DV). The National Ambient Air Quality Standards (NAAQS) for PM_{2.5}, established by the EPA, set the annual standard at 9 μg/m³. This standard is used to determine which regions are in compliance and which violate the NAAQS for PM_{2.5}. The EPA's methodology for calculating

DVs involves selecting the highest DV from all the stations in a county as the representative DV for that county. Our study adopts the same principle but utilizes the DL-estimated PM_{2.5} grids for this calculation. It means that each pixel from our estimated grids play the role of a monitoring station. Since the DL-estimated PM_{2.5} pixels are uniformly available across Texas, using them instead of monitoring stations to calculate DVs leads to significant improvements. Relying solely on monitoring stations results in many counties lacking DVs due to the absence of monitoring infrastructure. For instance, in the years 2020, 2021, and 2022, monitoring stations provided DVs for only 15, 19, and 22 out of 254 counties in Texas, respectively. In contrast, calculating DVs using our DL-estimated PM_{2.5} grids ensures coverage for all 254 counties in Texas, thereby enhancing the accuracy and completeness of DV calculations.

Our DL model's high-resolution grids capture the spatial variability of air quality within each county, offering a comprehensive view that surpasses the capabilities of point-based data from monitoring stations. In our comparative analysis of NAAQS violations across Texas, we observed notable discrepancies between the findings based on EPA-DV and DL-DV. Specifically, the EPA-DV identified eight, eight, and ten counties in 2020, 2021, and 2022, respectively, as violating the NAAQS for PM_{2.5} in Texas. In contrast, the DL-DV indicated a significantly higher number of non-compliant counties—94, 76, and 71 for the same years. This difference can be largely attributed to the incomplete coverage of EPA-DV across all counties in the CONUS, suggesting a potential underestimation of air quality issues. Furthermore, while all counties identified as violating the NAAQS based on EPA-DV were also found non-compliant in the DL-DV assessment, additional counties were flagged by the DL-DV. Specifically, four counties in 2020, seven in 2021, and four in 2022, considered compliant according to EPA-DV, were found to be non-compliant in the DL-DV analysis. This discrepancy underscores the critical need for enhanced monitoring techniques and suggests that our DL methodology could provide a more accurate and inclusive assessment of air quality, detecting violations that may be missed due to gaps in the existing EPA monitoring network.

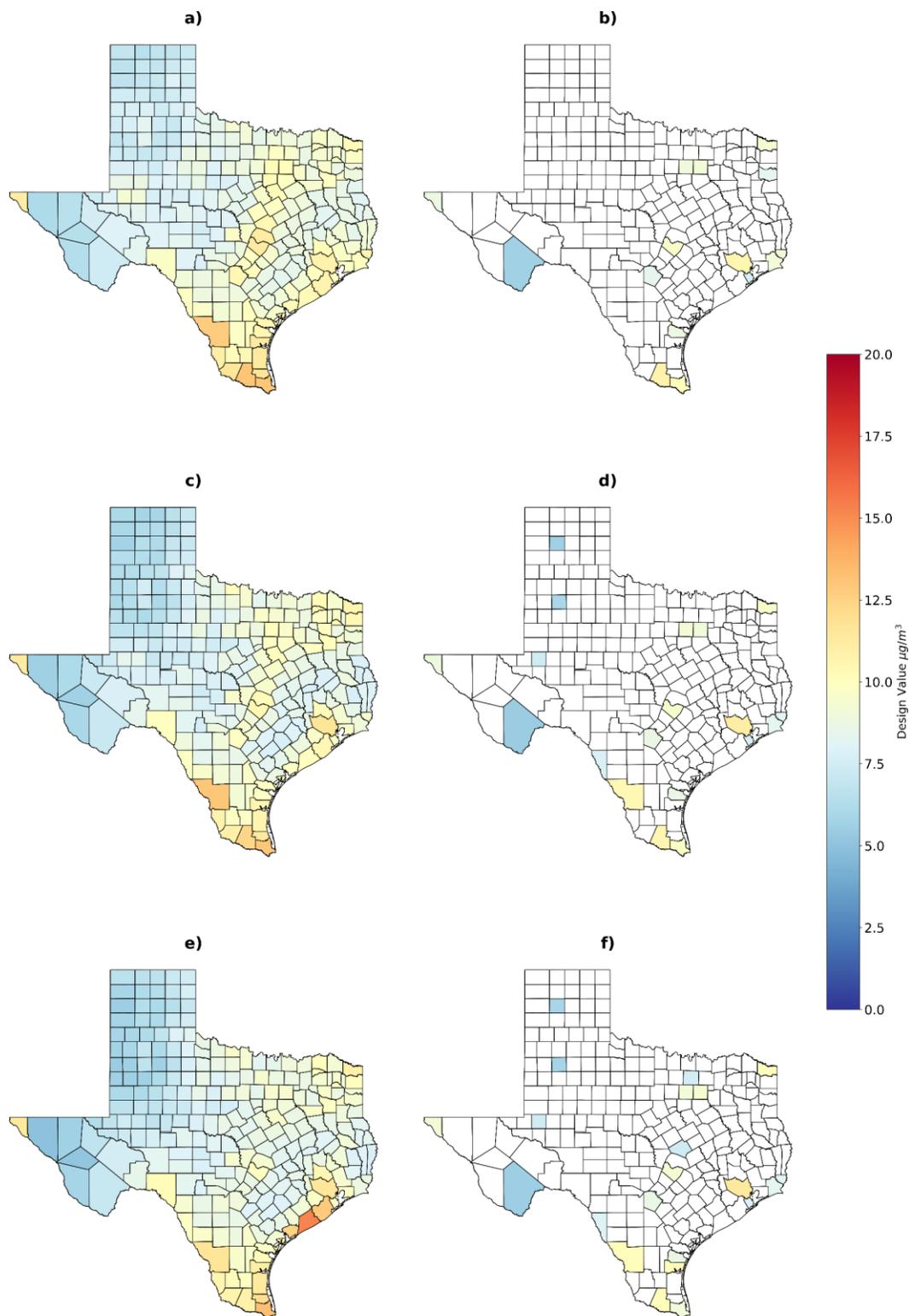


Figure 17. Comparative county-wise Visualization of PM_{2.5} design values (DVs): The left column illustrates PM_{2.5} DVs across Texas for (a) 2020, (c) 2021, and (e) 2022 as estimated by our model (DL-DV). The right column displays the EPA's reported PM_{2.5} DVs (EPA-DV) for (b) 2020, (d) 2021, and (f) 2022.

5. Web Portal for Supporting the State Implementation Plan (SIP)

The UH-AQF AI team has developed a data portal dedicated to data processing and visualization. The tools available on this portal provide substantial support for implementing SIP across Texas and analyzing PM_{2.5} levels. For instance, as noted in the previous section, generating DVs using our DL-estimated PM_{2.5} grids—a feature also available through this web portal—ensures comprehensive coverage for all 254 counties in Texas. This enhancement in accuracy and completeness of DV calculations aids decision-makers in developing more effective SIPs. To develop this portal, we leveraged the Streamlit library in Python. Streamlit is an open-source framework tailored for crafting interactive web applications for data science projects. It stands out for its user-friendliness, enabling swift prototyping with minimal coding, seamless integration with popular data science libraries, and effortless app sharing through various deployment options. Now, the portal is accessible via our local servers at (<http://spock.geosc.uh.edu:8501/>) and most of the processing is done online in the server upon request. Our portal boasts seven distinct tools meticulously engineered for data processing, visualization, and data retrieval. These tools comprise the Temporal Averaging Tool, Slider Tool, Temporal Data Comparison Tool, Data Download Tool, AOD-PM_{2.5} Comparison Tool, GIF Tool, and Design Value Tool. Each tool is meticulously designed to accept user inputs and execute specific functions tailored for processing PM_{2.5} data generated through DL. Below is further explanations about each tool included in the portal.

5.1. Temporal Averaging Tool

The tool allows users to define a date range of their choice and generates a spatial plot showcasing the average DL-estimated PM_{2.5} concentrations map within that specified timeframe. Users enjoy the flexibility of inputting date ranges ranging from a minimum of 2 days to a maximum of 4 full years. To illustrate, refer to Figure 18, which presents a sample plot for Temporal Averaging Tool. In this scenario, the user has opted for a date range spanning from January 1, 2022, to December 31, 2022. The tool then leverages this input to compute the average concentrations across the designated date range and subsequently generates the corresponding spatial plot. This tool operates in two distinct modes: a simple mode (depicted in Figure 18a), which generates a straightforward 2D plot, and an interactive mode (as shown in Figure 18b), where an interactive plot is produced. In the interactive mode, Longitude, Latitude, and Estimated PM_{2.5} levels are displayed at each point where the cursor is positioned, enhancing user engagement and data exploration.

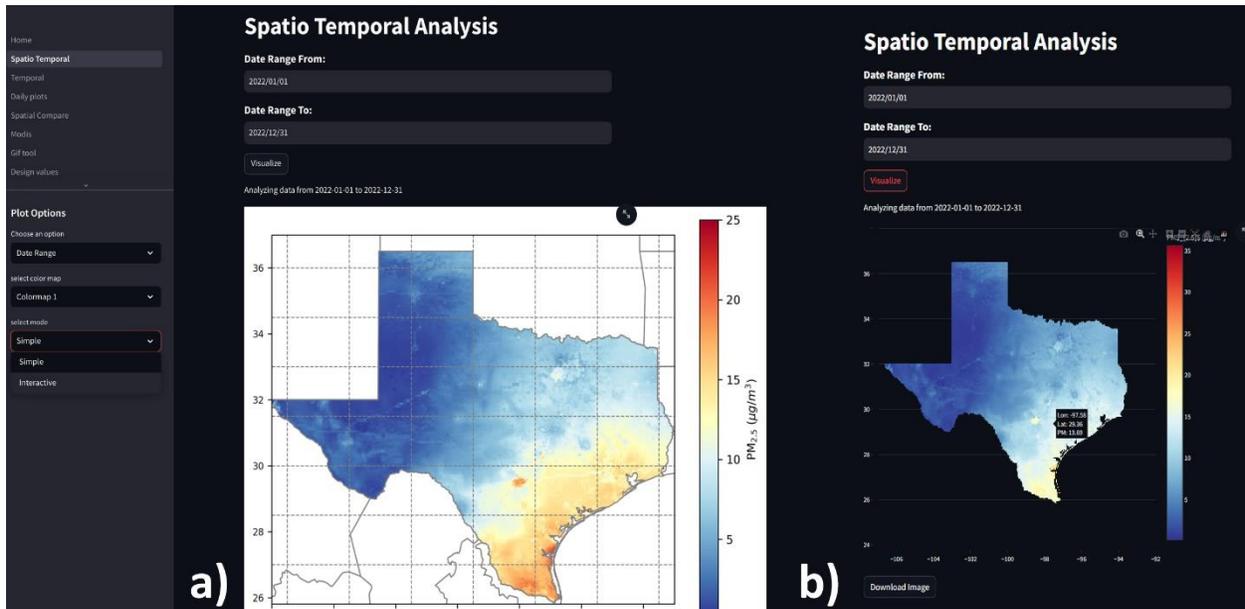


Figure 18. Temporal Averaging Tool that creates plots showing the mean PM_{2.5} levels during the desired period in a) simple mode and b) interactive mode.

5.2. Slider Tool to Compare Two Days of DL-Estimated PM_{2.5} Maps

This tool is specifically engineered for comparing the spatial distribution of DL-estimated PM_{2.5} concentrations for two specific dates. It features a slider functionality, enabling users to seamlessly navigate and visualize the spatial plots for both the selected day and the following day. An illustrative example of this spatial comparison tool is showcased in Figure 19. By simply sliding the tool to the left, users can observe the spatial plot for January 28, 2021, while sliding it to the right will unveil the plot for January 27, 2021. This interactive feature facilitates the straightforward comparison of data between two consecutive days, enhancing user-friendliness and aiding in data analysis.

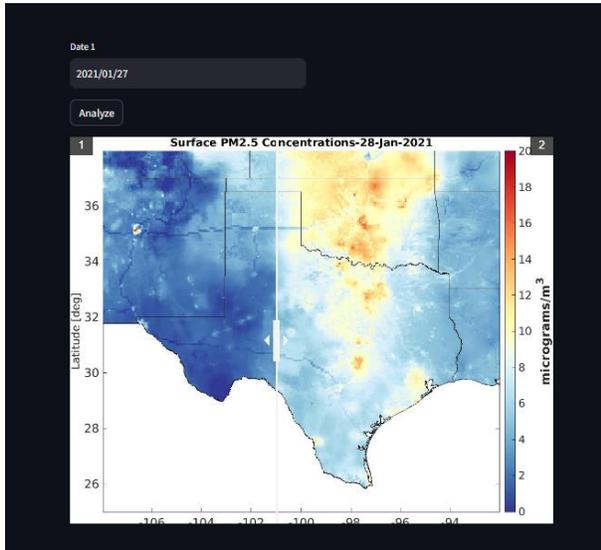


Figure 19. Slider Tool to compare the estimated PM_{2.5} maps from January 27, 2021, to January 28, 2021.

5.3. Temporal Comparison Tool

The tool necessitates input parameters, such as a designated date range and coordinate range values (latitude/longitude range), which can be specified for either a specific geographical area or a single point location. Users have the flexibility to input these parameters directly or via an interactive map interface, facilitating easy selection for a single point location or a specific geographical domain. Once these parameters are set, the tool computes the average PM_{2.5} concentration over the specified spatiotemporal region. Additionally, the tool allows for the selection of multiple locations or areas, generating an interactive time-series plot showcasing the PM_{2.5} concentration trends for all selected locations within a single figure. Furthermore, users are provided with the option to export the resulting plot as a CSV file for further analysis or documentation. For visual reference, refer to Figure 20, which demonstrates a sample execution of this process to compare PM_{2.5} trends within a geographic domain and two other point locations, encompassing the entirety of the year 2022.

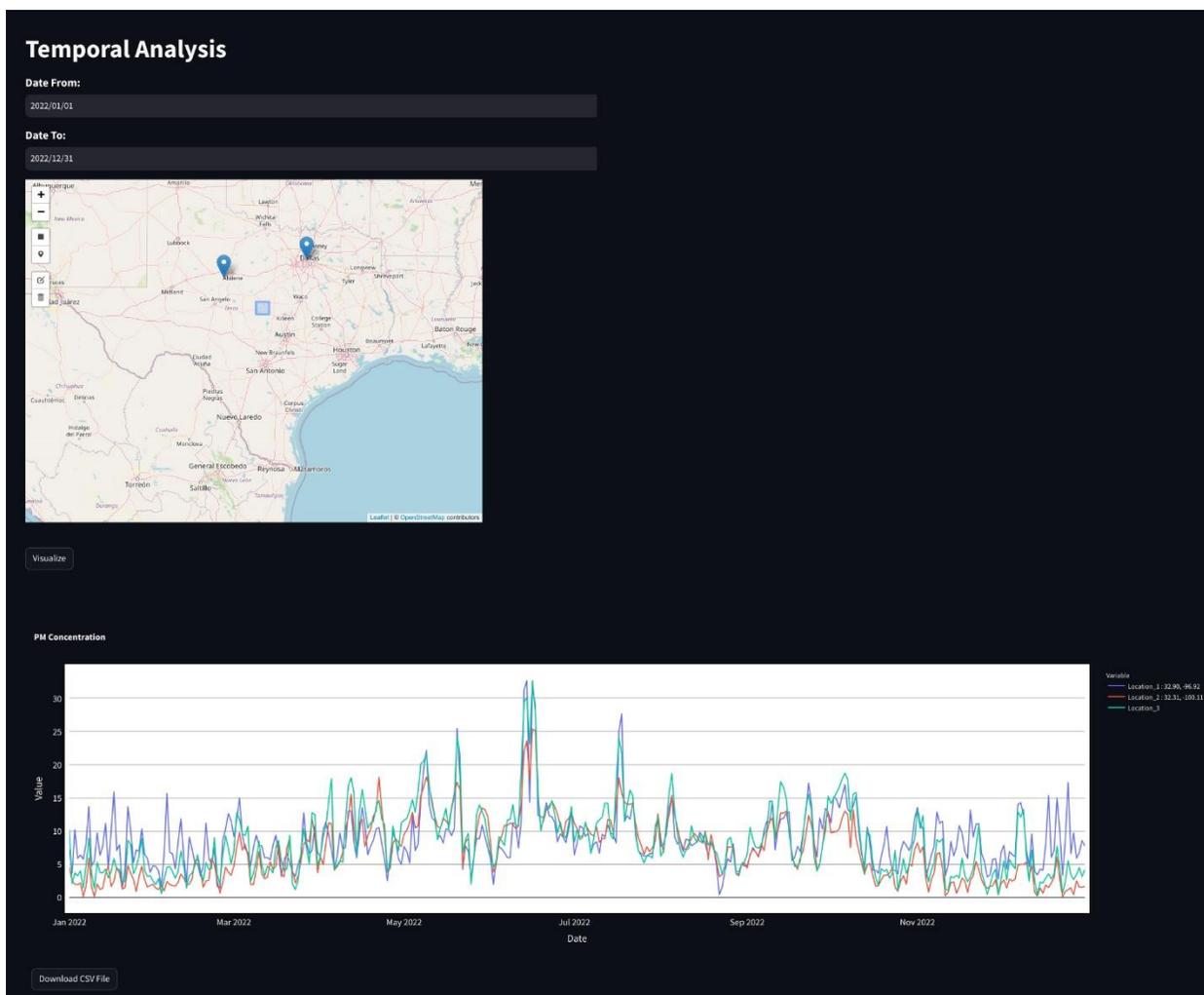


Figure 20. Temporal Comparison Tool to compare estimated PM_{2.5} levels in different locations.

5.4. Data Download Tool

This tool allows users to input a date and generates estimated PM_{2.5} levels over Texas for that specific day. It offers three distinct modes of operation: 1) Pre-prepared daily plots: Users can input a date, and the tool generates a pre-prepared PM_{2.5} plot for the specified date. 2) Simple plot tool: Users have the option to input a date along with additional parameters such as color map, color bar, min-max range, font size, font type, and title. The tool then produces a plot incorporating the specified arguments for the given date. 3) Interactive plot tool: Users input a date, and the tool generates an interactive spatial map illustrating daily estimates. Both pre-prepared and simple plots can be directly downloaded from the tool interface. For visual representation, refer to Figure 21, which illustrates this tool's interface.

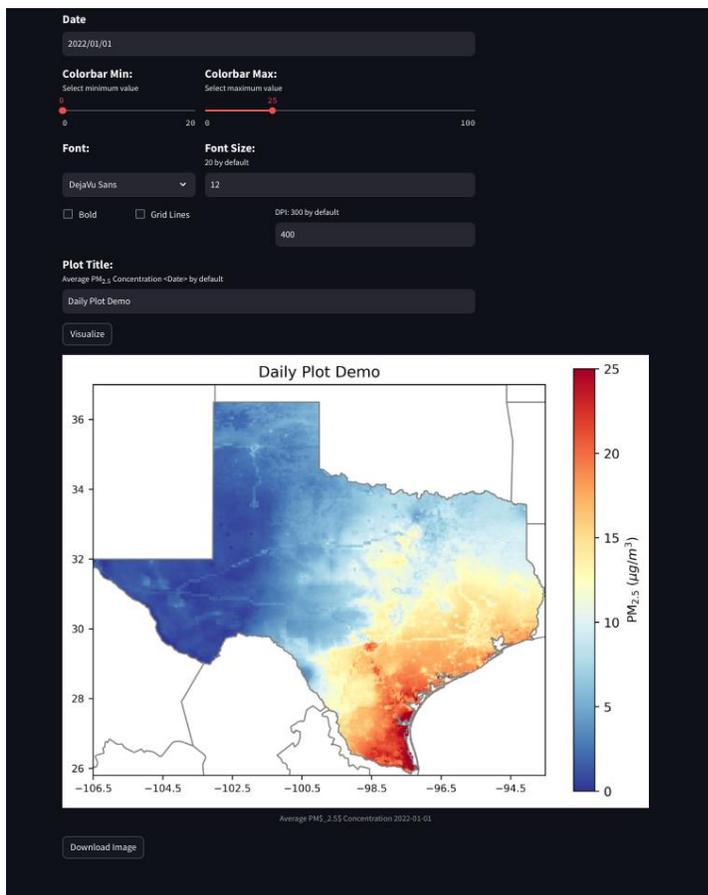


Figure 21. Data Download Tool.

5.5. AOD-vs-PM_{2.5} Comparison Tool

This tool is designed to display the DL-estimated PM_{2.5} alongside the AOD data sourced from MODIS for the same day. This tool allows for a rough comparison between PM_{2.5} and AOD levels on a given day. Notably, the MODIS AOD integrated into the data portal is the MODIS Terra and Aqua combined MAIAC product. It is essential to highlight that unlike the DL-estimated PM_{2.5}, the MODIS AOD image reflects AOD levels at the satellite's overpass time and is not a representation of daily mean values. Consequently, users should exercise caution when interpreting and comparing these two distinct products. Figure 22 shows that despite the temporal difference between DL-estimated PM_{2.5} and the MODIS AOD, the spatial distribution of PM_{2.5} and AOD reveals a noticeable similarity in the selected date.

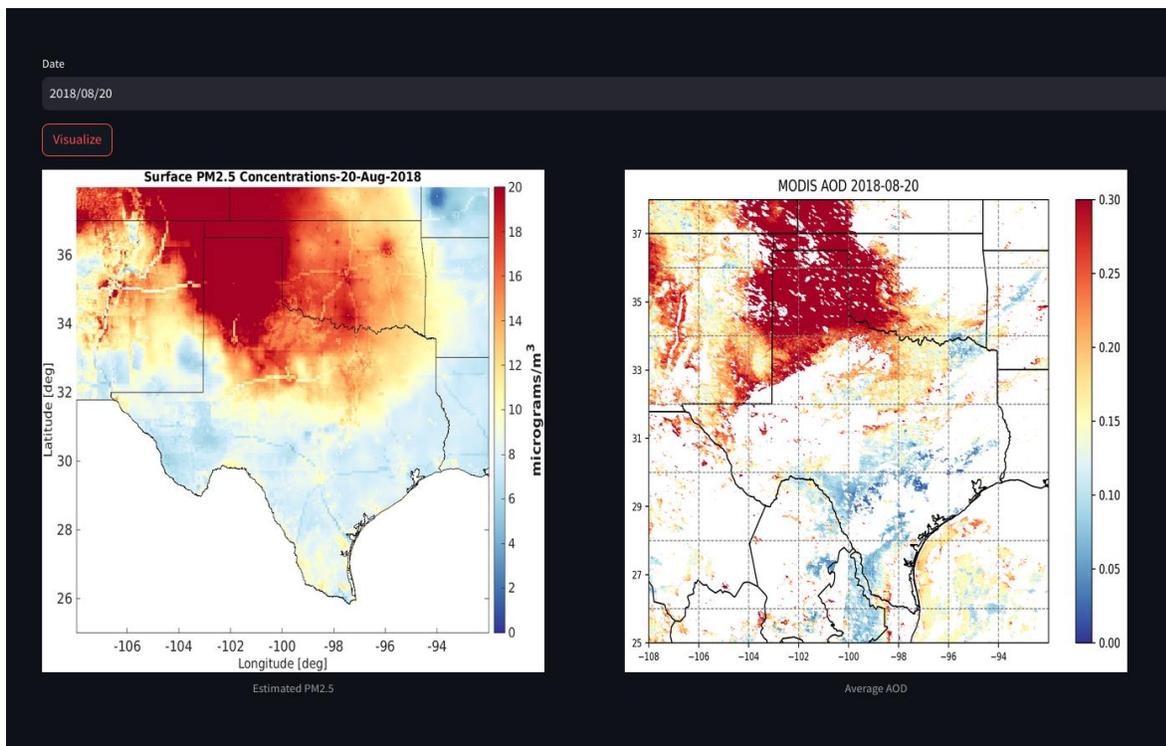


Figure 22. The AOD-vs-PM_{2.5} Comparison Tool to compare the DL-estimated PM_{2.5} and MODIS AOD for the same day.

5.6. GIF Tool

The GIF tool allows users to select a specific time period and generate a Graphics Interchange Format (GIF) file illustrating the spatiotemporal distribution changes of PM_{2.5} across Texas. This feature helps users in monitoring and analyzing changes in PM_{2.5} concentrations within the studied region. Figure 23 showcases the generated GIF, exhibiting PM_{2.5} changes from August 19 to August 25, 2018, and highlighting transportation of PM_{2.5} plume from western to eastern Texas during six days. We recommend utilizing the latest version of Microsoft Word to ensure proper display of the GIF in Figure 23. The GIF is also available at

“https://www.dropbox.com/scl/fi/jti3gn6nwyga2fmp8ex3l/GIF_For_TCEQ.gif?rlkey=vip2kg9ilix39o816r1k3t0xc&dl=0” at Drop Box.

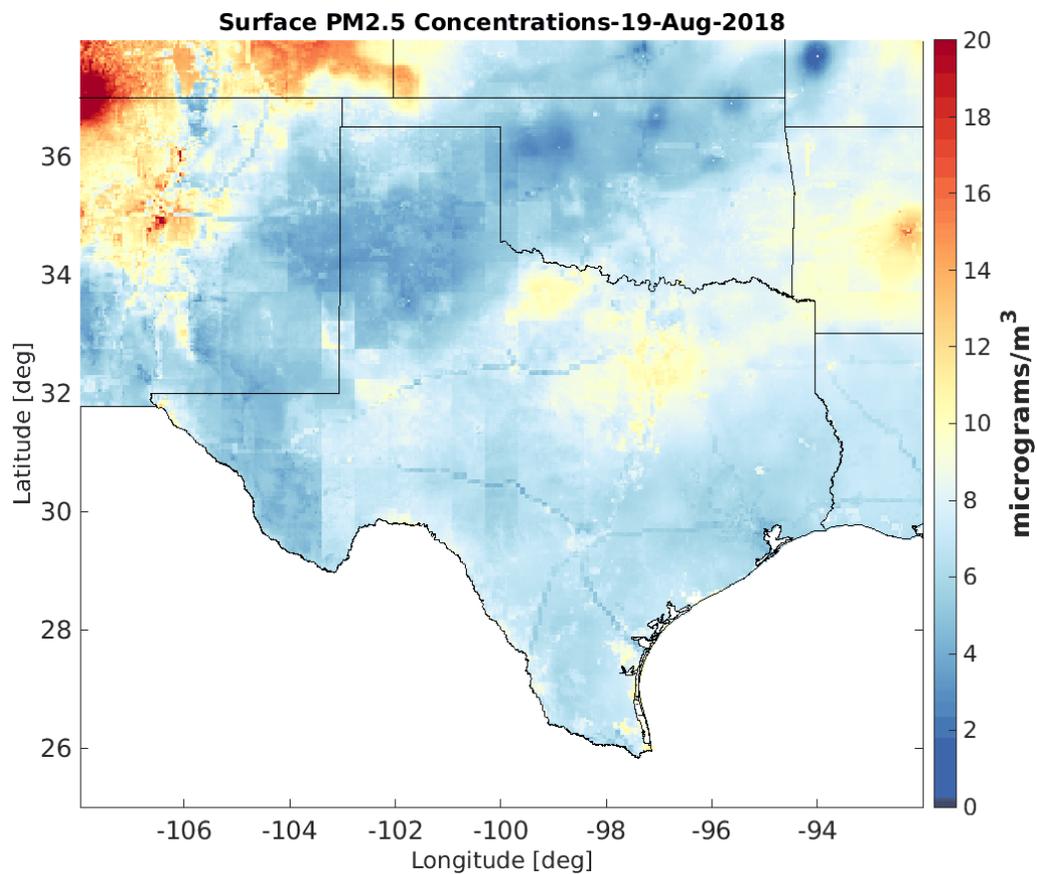


Figure 23. The Graphics Interchange Format (GIF) Tool showing the spatiotemporal changes in the DL-estimated $PM_{2.5}$ levels over Texas from August 19 to August 25, 2018.

5.7. Design Value Tool

The Design Value Tool displays the estimated DVs for all counties. The tool offers two modes: a) Simple plot (Figure 24a): Users can input the year, and the tool displays the DVs for the input year. The DV calculation adopts the EPA's principle of assigning the maximum value of DV in all stations in a county as the DV value of the county. Our tool, however, uses pixels from the estimated $PM_{2.5}$ grids instead of stations. b) Interactive plot (Figure 24b): Users can input the year, and the tool displays an interactive map displaying DV for the selected year for all counties. The elevation of each county in the Figure 24b is an indicator of the DV value for the given county, and the users can zoom, rotate, and tilt the output for better visualization.

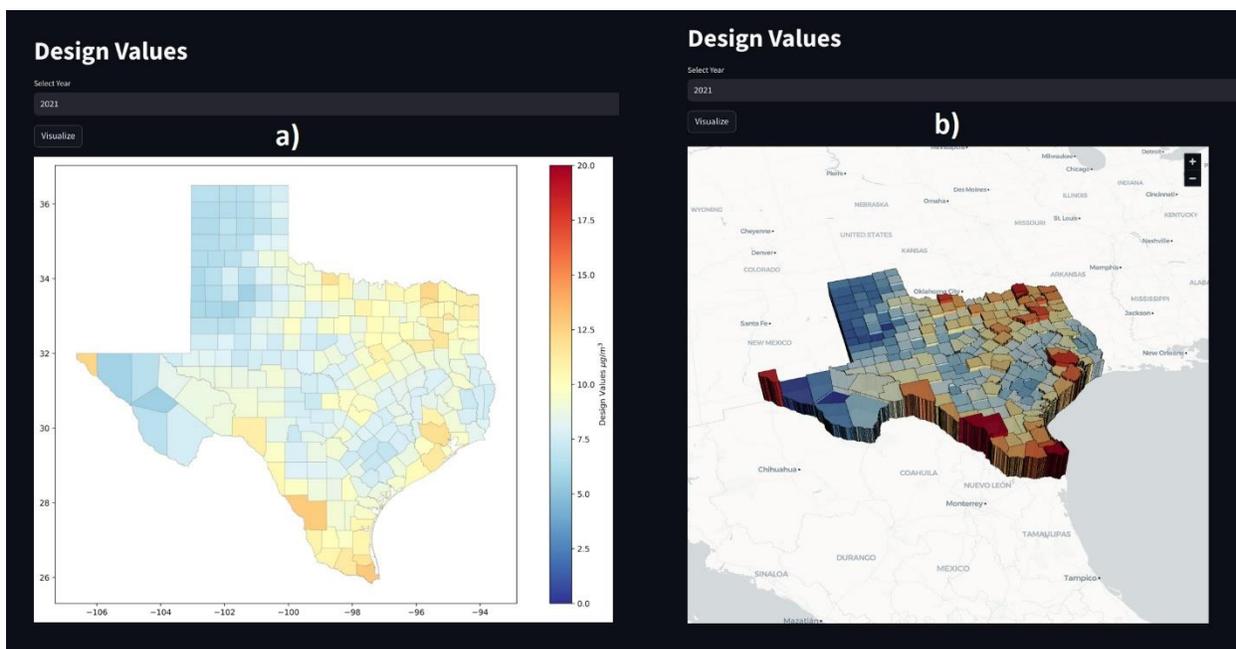


Figure 24. The Design Value Tool showing the design values (DVs) for each county over Texas in 2021.

6. Future Enhancement and Extension of the Project

The methodology developed under this grant can be further enhanced and extended in various ways. One promising approach is to leverage products derived from new geostationary Earth orbit (GEO) satellites, such as the Tropospheric Emissions: Monitoring of Pollution (TEMPO), which provides more consistent and frequent observations. The temporal resolution of GEO satellites typically spans from a few minutes to an hour, which is notably finer compared to those on the low Earth orbit (LEO) satellites that, at best, have a 12-hour orbiting cycle over specific geographic locations. TEMPO, designed to monitor aerosols and air pollutants across North America every daylight hour, offers unparalleled temporal resolution. Utilizing TEMPO AOD data, our PCNN-DeepCNN model can estimate surface $PM_{2.5}$ concentrations hourly across Texas and four neighboring states (henceforth referred to as Texas+4) of New Mexico, Oklahoma, Arkansas, and Louisiana. Incorporating four neighboring states into the $PM_{2.5}$ analysis could enhance our understanding of $PM_{2.5}$ transport into Texas. Additionally, the comprehensive spatial and temporal coverage of TEMPO enhances our model's capability to estimate surface $PM_{2.5}$ levels even during nighttime, showcasing the advanced potential of our PCNN-DeepCNN model.

Furthermore, this project can be expanded to include high-resolution estimations of other pollutants, such as nitrogen dioxide (NO_2) and ozone, across the Texas+4 domain over an extended period. While TEMPO data is only available from November 2023 onwards, our PCNN-DeepCNN model can utilize this alongside data from other satellite instruments—including the Tropospheric Monitoring Instrument (TROPOMI) on Sentinel-5p, the Ozone Monitoring Instrument (OMI) on Aura, and MODIS on Terra and Aqua satellites—to estimate surface concentrations of $PM_{2.5}$, NO_2 , and ozone from 2000 to 2024. Table 4 outlines the projected budget for expanding the project to daily estimate surface concentrations of $PM_{2.5}$, NO_2 , and ozone at 4 km resolution over the Texas+4 domain for this period. The estimated cost for monitoring each

pollutant is \$120,000. However, the bundled project cost—which includes all three pollutants—is only \$200,000. This reduction is due to partial overlapping processes and data products across the pollutants, which decreases both workload and cost. Additionally, the domain could be expanded to encompass the entire CONUS if TCEQ deems it advantageous. Table 5 mirrors Table 4 but differs in that it presents budgets for hourly estimations rather than daily. It should be noted that the budget and other details presented in Tables 4 and 5 are preliminary estimates. All aspects are open for negotiation to align with the requirements of both the TCEQ and the UH-AQF AI group at the University of Houston.

Table 4: Estimated budget for daily estimations of surface concentrations of PM_{2.5}, NO₂, and ozone at 4 km spatial resolution across the Texas+4 domain from 2000 to 2024. Texas+4 domain includes five states: Texas, New Mexico, Oklahoma, Arkansas, and Louisiana. The final row details the information for the bundled project, which encompasses the estimation of all three pollutants.

| Pollutant | Domain | Period | Resolution | Task Duration | Budget |
|-------------------------|---------------|---------------|-------------------|----------------------|---------------|
| PM_{2.5} | Texas+4 | 2000-2024 | 4km/Daily | 15 months | \$120,000 |
| NO₂ | Texas+4 | 2000-2024 | 4km/Daily | 15 months | \$120,000 |
| Ozone | Texas+4 | 2000-2024 | 4km/Daily | 15 months | \$120,000 |
| All Three | Texas+4 | 2000-2024 | 4km/Daily | 22 months | \$200,000 |

Table 5: Estimated budget for hourly estimations of surface concentrations of PM_{2.5}, NO₂, and ozone at 4 km spatial resolution across the Texas+4 domain from 2000 to 2024. Texas+4 domain includes five states: Texas, New Mexico, Oklahoma, Arkansas, and Louisiana. The final row details the information for the bundled project, which encompasses the estimation of all three pollutants.

| Pollutant | Domain | Period | Resolution | Task Duration | Budget |
|-------------------------|---------------|---------------|-------------------|----------------------|---------------|
| PM_{2.5} | Texas+4 | 2000-2024 | 4km/Hourly | 18 months | \$150,000 |
| NO₂ | Texas+4 | 2000-2024 | 4km/Hourly | 18 months | \$150,000 |
| Ozone | Texas+4 | 2000-2024 | 4km/Hourly | 18 months | \$150,000 |
| All Three | Texas+4 | 2000-2024 | 4km/Hourly | 24 months | \$260,000 |

References

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework (arXiv:1907.10902). arXiv. <https://doi.org/10.48550/arXiv.1907.10902>
- Alom, M. Z., Hasan, M., Yakopcic, C., Taha, T. M., & Asari, V. K. (2018). Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation (arXiv:1802.06955). arXiv. <https://doi.org/10.48550/arXiv.1802.06955>
- Di, Q., Kloog, I., Koutrakis, P., Lyapustin, A., Wang, Y., & Schwartz, J. (2016). Assessing PM_{2.5} Exposures with High Spatiotemporal Resolution across the Continental United States. *Environmental Science & Technology*, 50(9), 4712–4721. <https://doi.org/10.1021/acs.est.5b06121>
- Feng, S., Gao, D., Liao, F., Zhou, F., & Wang, X. (2016). The health effects of ambient PM_{2.5} and potential mechanisms. *Ecotoxicology and environmental safety*, 128, 67-74.
- Fu, W., Yue, X., Li, Z., Tian, C., Zhou, H., Li, K., ... & Hu, Y. (2022). Decoupling between PM_{2.5} concentrations and aerosol optical depth at ground stations in China. *Frontiers in Environmental Science*, 10, 979918.
- Ghahremanloo, M., Choi, Y., Sayeed, A., Salman, A. K., Pan, S., & Amani, M. (2021). Estimating daily high-resolution PM_{2.5} concentrations over Texas: Machine Learning approach. *Atmospheric Environment*, 247, 118209.
- Ghahremanloo, M., Lops, Y., Choi, Y., Jung, J., Mousavinezhad, S., & Hammond, D. (2022). A comprehensive study of the COVID-19 impact on PM_{2.5} levels over the contiguous United States: A deep learning approach. *Atmospheric Environment*, 272, 118944.
- Ghahremanloo, M., Lops, Y., Choi, Y., Mousavinezhad, S., & Jung, J. (2023). A Coupled Deep Learning Model for Estimating Surface NO₂ Levels From Remote Sensing Data: 15-Year Study Over the Contiguous United States. *Journal of Geophysical Research: Atmospheres*, 128(2), e2022JD037010. <https://doi.org/10.1029/2022JD037010>
- Ginzburg, H., Liu, X., Baker, M., Shreeve, R., Jayanty, R. K. M., Campbell, D., & Zielinska, B. (2015). Monitoring study of the near-road PM_{2.5} concentrations in Maryland. *Journal of the Air & Waste Management Association*, 65(9), 1062-1071.
- Gong, B., Langguth, M., Ji, Y., Mozaffari, A., Stadtler, S., Mache, K., & Schultz, M. G. (2022). Temperature forecasting by deep learning methods. *Geoscientific Model Development*, 15(23), 8931–8956. <https://doi.org/10.5194/gmd-15-8931-2022>
- Grigsby, J., Wang, Z., Nguyen, N., & Qi, Y. (2023). Long-Range Transformers for Dynamic Spatiotemporal Forecasting (arXiv:2109.12218). arXiv. <https://doi.org/10.48550/arXiv.2109.12218>
- Gupta, P., & Christopher, S. A. (2009). Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach. *Journal of Geophysical Research: Atmospheres*, 114(D14). <https://doi.org/10.1029/2008JD011496>
- Hayes, R. B., Lim, C., Zhang, Y., Cromar, K., Shao, Y., Reynolds, H. R., ... & Thurston, G. D. (2020). PM_{2.5} air pollution and cause-specific cardiovascular disease mortality. *International journal of epidemiology*, 49(1), 25-35.

- Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., & Liu, Y. (2017). Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. *Environmental Science & Technology*, 51(12), 6936–6944. <https://doi.org/10.1021/acs.est.7b01210>
- Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., & Liu, Y. (2017). Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. *Environmental Science & Technology*, 51(12), 6936–6944. <https://doi.org/10.1021/acs.est.7b01210>
- Karnaie, S., & John, K. (2019). Source apportionment of PM2.5 measured in South Texas near USA–Mexico border. *Atmospheric Pollution Research*, 10(5), 1663-1676.
- Kloog, I., Koutrakis, P., Coull, B. A., Lee, H. J., & Schwartz, J. (2011). Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmospheric environment*, 45(35), 6267-6275.
- Liu, G., Reda, F. A., Shih, K. J., Wang, T.-C., Tao, A., & Catanzaro, B. (2018). Image Inpainting for Irregular Holes Using Partial Convolutions (arXiv:1804.07723). arXiv. <https://doi.org/10.48550/arXiv.1804.07723>
- Liu, J., Li, M., Luo, Y., Yang, S., Li, W., & Bi, Y. (2021). Alzheimer’s disease detection using depthwise separable convolutional neural networks. *Computer Methods and Programs in Biomedicine*, 203, 106032. <https://doi.org/10.1016/j.cmpb.2021.106032>
- Lops, Y., Ghahremanloo, M., Pouyaei, A., Choi, Y., Jung, J., Mousavinezhad, S., Salman, A. K., & Hammond, D. (2022). Spatiotemporal Estimation of TROPOMI NO2 Column with Depthwise Partial Convolutional Neural Network (arXiv:2204.05917). arXiv. <https://doi.org/10.48550/arXiv.2204.05917>
- Lops, Y., Pouyaei, A., Choi, Y., Jung, J., Salman, A. K., & Sayeed, A. (2021). Application of a Partial Convolutional Neural Network for Estimating Geostationary Aerosol Optical Depth Data. *Geophysical Research Letters*, 48(15), e2021GL093096. <https://doi.org/10.1029/2021GL093096>
- Ma, D., Li, G., & He, F. (2021). Exploring PM2.5 environmental efficiency and its influencing factors in China. *International journal of environmental research and public health*, 18(22), 12218.
- Madrigano, J., Kloog, I., Goldberg, R., Coull, B. A., Mittleman, M. A., & Schwartz, J. (2013). Long-term exposure to PM2.5 and incidence of acute myocardial infarction. *Environmental health perspectives*, 121(2), 192-196.
- Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and health impacts of air pollution: a review. *Frontiers in public health*, 8, 505570.
- Park, Y., Kwon, B., Heo, J., Hu, X., Liu, Y., & Moon, T. (2020). Estimating PM2.5 concentration of the conterminous United States via interpretable convolutional neural networks. *Environmental Pollution*, 256, 113395. <https://doi.org/10.1016/j.envpol.2019.113395>
- Park, Y., Kwon, B., Heo, J., Hu, X., Liu, Y., & Moon, T. (2020). Estimating PM2.5 concentration of the conterminous United States via interpretable convolutional neural networks. *Environmental Pollution*, 256, 113395. <https://doi.org/10.1016/j.envpol.2019.113395>
- Pascal, M., Falq, G., Wagner, V., Chatignoux, E., Corso, M., Blanchard, M., ... & Larrieu, S. (2014). Short-term impacts of particulate matter (PM10, PM10–2.5, PM2.5) on mortality in nine French cities. *Atmospheric Environment*, 95, 175-184.

- Payami, M., Choi, Y., Salman, A. K., Mousavinezhad, S., Park, J., & Pouyaei, A. (2024). A 1D CNN-based emulator of CMAQ: Predicting NO₂ concentration over the most populated urban regions in Texas. *Artificial Intelligence for the Earth Systems*. <https://journals.ametsoc.org/view/journals/aies/aop/AIES-D-23-0055.1/AIES-D-23-0055.1.xml>
- Pun, V. C., Kazemiparkouhi, F., Manjourides, J., & Suh, H. H. (2017). Long-term PM_{2.5} exposure and respiratory, cancer, and cardiovascular mortality in older US adults. *American journal of epidemiology*, 186(8), 961-969.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation (arXiv:1505.04597). arXiv. <https://doi.org/10.48550/arXiv.1505.04597>
- Salman, A. K., Choi, Y., Park, J., Mousavinezhad, S., Payami, M., Momeni, M., & Ghahremanloo, M. (2024). Deep learning based emulator for simulating CMAQ surface NO₂ levels over the CONUS. *Atmospheric Environment*, 316, 120192.
- Sayeed, A., Choi, Y., Pouyaei, A., Lops, Y., Jung, J., & Salman, A. K. (2022). CNN-based model for the spatial imputation (CMSI version 1.0) of in-situ ozone and PM_{2.5} measurements. *Atmospheric Environment*, 289, 119348. <https://doi.org/10.1016/j.atmosenv.2022.119348>
- Schwartz, J., Dockery, D. W., & Neas, L. M. (1996). Is daily mortality associated specifically with fine particles?. *Journal of the Air & Waste Management Association*, 46(10), 927-939.
- Singh, D., Choi, Y., Park, J., Salman, A. K., Sayeed, A., & Song, C. H. (2024). Deep-BCSI: A deep learning-based framework for bias correction and spatial imputation of PM_{2.5} concentrations in South Korea. *Atmospheric Research*, 301, 107283. <https://doi.org/10.1016/j.atmosres.2024.107283>
- Wei, J., Huang, W., Li, Z., Xue, W., Peng, Y., Sun, L., & Cribb, M. (2019). Estimating 1-km-resolution PM_{2.5} concentrations across China using the space-time random forest approach. *Remote Sensing of Environment*, 231, 111221. <https://doi.org/10.1016/j.rse.2019.111221>
- Wu, N., Green, B., Ben, X., & O'Banion, S. (2020). Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case (arXiv:2001.08317). arXiv. <http://arxiv.org/abs/2001.08317>
- Xiao, Q., Wang, Y., Chang, H. H., Meng, X., Geng, G., Lyapustin, A., & Liu, Y. (2017). Full-coverage high-resolution daily PM_{2.5} estimation using MAIAC AOD in the Yangtze River Delta of China. *Remote Sensing of Environment*, 199, 437-446.
- Xing, Y. F., Xu, Y. H., Shi, M. H., & Lian, Y. X. (2016). The impact of PM_{2.5} on the human respiratory system. *Journal of thoracic disease*, 8(1), E69.
- Zhang, Q., Xue, D., Liu, X., Gong, X., & Gao, H. (2019). Process analysis of PM_{2.5} pollution events in a coastal city of China using CMAQ. *Journal of Environmental Sciences*, 79, 225-238.
- Zhao, R., Gu, X., Xue, B., Zhang, J., & Ren, W. (2018). Short period PM_{2.5} prediction based on multivariate linear regression model. *PLOS ONE*, 13(7), e0201011. <https://doi.org/10.1371/journal.pone.0201011>

Submitted to TCEQ by
Yunsoo Choi