Multi-source data fusion for filling gaps in satellite Aerosol Optical Depth (AOD) using generative models

Anusha Srirenganathan Malarvizhi†
Earth Systems and Geoinformation Science
George Mason University
Fairfax VA USA
asrireng@gmu.edu

Phoebe Pan
Thomas Jefferson High School for Science and Technology
Alexandria VA USA
phoebe.m.pan.@gmail.com

ABSTRACT

Aerosol Optical Depth (AOD) is a crucial parameter for monitoring air quality, but satellite-based measurements often suffer from significant gaps due to cloud cover and other obstructions. These missing data, usually categorized as Missing Not At Random (MNAR), pose challenges for accurate air quality assessments. This study applies a Generative Adversarial Imputation Network (GAIN) to impute missing AOD data from the MODIS MAIAC dataset across the Northeast United States, addressing the MNAR challenge by leveraging relevant meteorological covariates, such as cloud cover, relative humidity, and temperature.

The GAIN model was trained using data from 2021 to 2022, with hyperparameter tuning conducted to optimize performance. The tuning process revealed that a low learning rate and minimal weight decay yielded the most stable and accurate results. The model was validated against AERONET data, achieving a correlation coefficient (R) of 0.89, demonstrating strong alignment between imputed and observed AOD values. The GAIN model also demonstrated strong predictive accuracy, achieving an average R^2 of 0.94, MSE of 0.0046, and RMSE of 0.0676. Cross-validation confirmed the robustness and generalizability of the model across various datasets. The model's performance was compared with traditional imputation methods like MICE and MissForest. GAIN outperformed both models, superiorly handling MNAR data and minimizing error across all metrics. This comparative analysis emphasizes the GAIN model's ability to capture complex spatial and temporal dependencies in the dataset effectively. In addition to filling data gaps, the GAIN model preserved the spatial distribution of AOD, showing higher concentrations in urban areas and regions with elevated pollution. During the 2023 Canadian wildfire event, the model successfully imputed AOD levels, capturing the sharp rise in aerosol concentrations. This study demonstrates the effectiveness of GAIN in handling complex MNAR scenarios, offering a reliable solution for improving AOD data coverage and enhancing the accuracy of air quality assessments.

GeoIndustry '24, October 29-November 1, 2024, Atlanta, Georgia USA

979-8-4007-1145-9/24/10
https://doi.org/10.1145/3681766.3699607

CCS CONCEPTS

• Machine Learning • Neural Networks • Generative Adversarial Networks • Air Quality Modeling • Imputation

KEYWORDS

Aerosol Optical Depth (AOD), MODIS MAIAC, Satellite Remote Sensing, Generative Adversarial Imputation Networks (GAIN), GAN, Hyperparameter tuning, Air Quality, Missing Not At Random (MNAR), AERONET

ACM Reference format:

Anusha Srirenganathan Malarvizhi and Phoebe Pan. 2034. Multi-source data fusion for filling gaps in satellite Aerosol Optical Depth (AOD) using generative models: An adversarial approach to the missing data problem. In *Proceedings of ACM SIGSPATIAL conference (SIGSPATIAL'24). ACM, Georgia, GA, USA, 2 pages.* https://doi.org/10.1145/3681766.3699607

1 Introduction

In recent years, there has been an increasing focus on generative models, driven by the limitations of traditional supervised learning [11]. Supervised learning requires large quantities of labeled data and often demands substantial human effort to annotate millions of examples to achieve high performance [4]. In response, researchers have focused on unsupervised or semi-supervised learning methods to minimize the need for human supervision and the number of training examples, frequently utilizing generative models. Among these, Generative Adversarial Networks (GANs), proposed by Ian Goodfellow, have emerged as a compelling approach [4]. GANs consist of two neural networks - the generator and the discriminator - that adopt an adversarial framework to learn complex data distributions effectively, making them an ideal choice for tasks such as image generation [22], image inpainting [34], imputation [12], and more.

One significant challenge in many fields is the issue of missing data, which can hinder the accuracy and effectiveness of data-driven studies [39]. For example, in the field of optical remote sensing, satellite observations have gaps due to various factors, such as cloud cover obstructing the sensors [27]. This is particularly problematic when measuring Aerosol Optical Depth (AOD), an essential parameter for monitoring air quality and assessing the impacts of climate change. However, these gaps in satellite

observations can lead to missing data points that significantly impact the accuracy and reliability of studies related to air quality and climate monitoring.

Previous studies have found that more than 70% of AOD data can be missing due to persistent cloud cover and other obstructions, creating substantial gaps in satellite datasets [3]. This missing data falls under the category of Missing Not At Random (MNAR), where the likelihood of data being missing is related to the unobserved values themselves [39]. In the case of AOD measurements, the missingness is influenced by atmospheric conditions that also affect aerosol concentrations. For instance, conditions like high humidity and pollution levels, which lead to increased cloud cover, can also contribute to higher AOD values [7, 28]. As a result, when satellites cannot observe AOD due to cloud cover, the missing data is more likely to occur in areas or times with elevated aerosol levels, illustrating an MNAR scenario. These situations pose significant challenges because they introduce biases that traditional imputation methods may not adequately address [40]. Advanced imputation methods, such as the Generative Adversarial Imputation Network (GAIN), can address these challenges. [35] has shown that GAIN can outperform other state-of-the-art imputation methods in MNAR settings, demonstrating its ability to handle the complexities associated with missing data that depend on unobserved values [35]. To our knowledge, this is the first study that applied GAIN for AOD imputation to address the challenge of non-random missingness in MAIAC AOD data, particularly in scenarios where atmospheric conditions like cloud cover frequently obscure measurements. The motivation to impute AOD data stems from the necessity of having continuous, reliable datasets for air quality studies, particularly for predicting ground-level $PM_{2.5}$. Imputation of AOD is, therefore, a crucial step to ensure complete data coverage, which is essential for accurate public health analysis and environmental policymaking. To address the concern about the accuracy of using imputed AOD data for $PM_{2.5}$ predictions, we validate the imputed AOD against ground-based measurements, such as AERONET. This minimizes potential biases introduced by the imputation process, providing a reliable foundation for $PM_{2.5}$ predictions. Various studies show that adequately validated imputed AOD can significantly enhance air quality estimates, particularly in areas with limited ground-based monitoring [14, 20]. We chose GAIN for this study because it particularly excelled at handling Missing Not At Random (MNAR) data. While newer models exist, GAIN's adversarial approach balances accuracy and computational efficiency, making it suitable for our large-scale AOD imputation task.

The rest of this paper is organized as follows: Section 2 provides a comprehensive literature review, covering previous work on AOD imputation and related methodologies. Section 3 details the study area, discusses the use case and outlines the data used in this research. Section 4 elaborates on the workflow, including the model architecture, training process, hyperparameter tuning, and accuracy assessment. Section 5 presents the results and provides a detailed discussion of the findings, and Section 6 concludes the paper.

2 Literature Review

Several studies have addressed the issue of missing AOD using various imputation methods. **Traditional approaches include statistical methods**, such as maximum likelihood estimation, and interpolation techniques, like ordinary kriging, to fill gaps and minimize the prediction error [19, 33, 41]. To overcome these limitations, **advanced spatial statistical techniques** like Spatial Statistical Data Fusion (SSDF) and spatiotemporal kriging have been developed, offering improved accuracy by accounting for spatial variability and efficiently managing large-scale data [8, 18, 33]. **Multiple Imputation (MI) models** address missing values by generating plausible estimates based on the dataset's distributions and relationships of the observed variables. Studies have demonstrated MI's effectiveness, with R^2 values ranging from 0.77 to 0.86, aligning with AERONET observations [31, 32]. AI and machine learning (ML) models have proven effective for AOD imputation due to their ability to handle large, complex datasets and capture intricate non-linear relationships between variables [40]. Among these models, **gradient-based models** optimize predictions by iteratively adjusting parameters to minimize errors, making them highly effective in imputation tasks. A comparative study [2] reported that an XGBoost model validation with AERONET achieved a correlation coefficient of 0.83 and RMSE of 0.06 [19], while a LightGBM model showed strong agreement with AERONET, yielding an R^2 of 0.8 and RMSE of 0.15 [37], and another achieving a correlation coefficient of 0.84, RMSE of 0.19 [36]. **Deep learning models** excel at capturing complex spatial-temporal patterns, offering better generalization and more accurate simulation of spatial variations in AOD data. [14] developed an autoencoder-based deep residual network with the imputed MAIAC AOD strongly correlating with AERONET AOD, showing a correlation coefficient of 0.83, an R^2 of 0.69, and an RMSE of 0.04. [13] used a full residual deep network with improved generalization, showing a correlation of $R^2 = 0.78$ when validated against AERONET data.

While reviewed studies demonstrate strong imputation performance using various methods, several limitations should be considered. Traditional interpolation methods, including kriging, often struggle with large datasets and complex spatial correlations, leading to less precise AOD predictions [19, 33, 41]. Spatiotemporal kriging depends heavily on the availability of satellite data, and when data is insufficient, it may generate unrealistic features or result in over-smoothed outputs [18, 33]. Some studies fail to achieve full spatial coverage, leaving gaps in AOD estimates [2, 8, 33]. Other potential weaknesses include limited validation sites and daily-level models, which restrict the generalizability and scalability of the findings. Other possible limitations include limited validation sites [13, 21, 36].

GAIN provides a robust solution to many challenges observed in previous studies. It ensures complete spatial coverage, eliminates gaps that hinder comprehensive analysis, and can incorporate relevant meteorological covariates to capture complex spatial and temporal patterns. GAIN's flexibility allows it to generalize effectively across diverse regions and timeframes, addressing the scalability issues seen in daily-level models. Its ability to accurately

model extreme and localized aerosol events, such as wildfires, makes it ideal for AOD imputation in dynamic and challenging atmospheric conditions.

3 Study Area and Data

3.1 Study region and use case

The study focuses on the Northeast region of the United States, as designated by the National Centers for Environmental Information (NCEI) [10]. This region includes 11 states: Connecticut, Delaware, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, and Washington D.C. Figure 1 illustrates the study area, showing the location of AERONET stations across these states. The Northeast is chosen due to its relevance to current air quality research interests, particularly urban air quality dynamics and trends.

This region includes major metropolitan areas such as New York City, Boston, and Washington D.C., which is crucial for studying air quality in densely populated areas. The Northeast region was notably affected by the 2023 Canadian wildfires [29, 30], during early June of 2023, when dense smoke plumes significantly deteriorated air quality across the region.

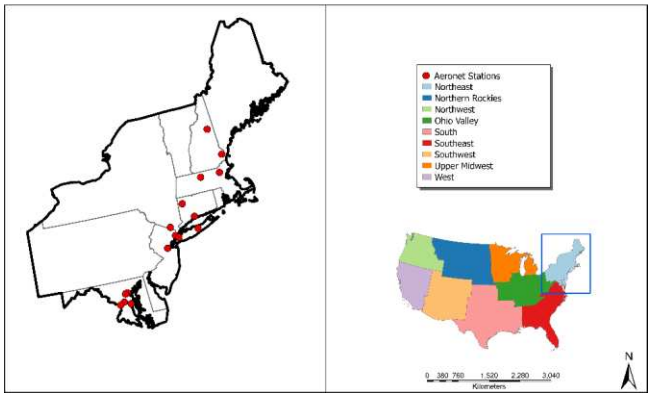

Figure 1: Study area showing the location of AERONET stations in the Northeast region of the United States

3.2 Data

In this study, we utilized satellite-derived aerosol data from MODIS MAIAC, aerosol species data from MERRA-2, and meteorological variables from the ERA5 dataset, complemented by geographical variables. These datasets provided critical inputs for aerosol optical depth (AOD) imputation and subsequent model validation. Full details on data sources and processing can be found in Appendix A.

4 Methods

4.1 Pre-processing of MODIS MAIAC AOD

In this study, daily MAIAC AOD images at 550 nm were collected, consisting of 5 tiles per day that provided full coverage of the study area. Pre-processing involved rigorous QA/QC procedures and reprojection to maintain data quality and consistency. To ensure high data reliability, only AOD values with precise cloud masking and best quality were selected [15]. The AOD data, initially in sinusoidal projection, was reprojected to the USA Contiguous Lambert Conformal Conic projection and interpolated to a 1 km spatial resolution. Finally, the five tiles were mosaicked into a single image, comprehensively covering the entire study area.

For consistent analysis, covariates must be spatially and temporally matched with the MAIAC AOD data. To achieve this, meteorological variables from ERA-5 ECMWF and aerosol species data from MERRA-2 were aligned with the Terra and Aqua MODIS satellite overpasses, which occur daily at approximately 10:30 and 13:30 local time. Daily averages of the hourly data were calculated within the satellite overpass windows. These variables were then reprojected to the USA Contiguous Lambert Conformal Conic projection and resampled to a 1 km × 1 km resolution to match the MAIAC AOD data for seamless integration.

4.2 GAIN model architecture diagram

The GAIN, adapted from its original framework [35], is modified here for aerosol optical depth (AOD) imputation, specifically supporting spatial data. GAIN operates with two key components: the generator and the discriminator. The generator takes three inputs: a satellite AOD data with gaps $\tilde{X}$, a mask matrix M indicates the locations of missing values, and a noise vector Z introduces randomness into the imputation process. The generator is defined as:

$$\bar{X} = G(\tilde{X}, M, (1 - M) \odot Z)$$

where $\odot$ denotes element-wise multiplication, and (1 - M) masks the noise vector Z, ensuring that noise is applied only to missing values. The imputed data vector $\bar{X}$ contains the generator's predictions for both observed and missing components. The final imputed data vector $\hat{X}$ is given by:

$$\hat{X} = M \odot \tilde{X} + (1 - M) \odot \tilde{X}$$

On the other hand, the discriminator evaluates the generator's output by attempting to distinguish between observed and imputed values. It takes the completed data $\hat{X}$ and a hint matrix H as inputs, where the hint matrix provides partial information about the mask. The discriminator aims to improve imputation quality by guiding the generator to predict missing values that resemble observed data. The adversarial loss function is fundamental to the GAIN framework, facilitating the interaction between the generator and discriminator. This loss function allows the generator to improve its imputation process by minimizing the discrepancy between observed and imputed values. The adversarial loss for the discriminator D is expressed as:

$$\mathcal{L}_D(m, \hat{m}, b) = \sum_{i:b_i=0} [m_i \log(\hat{m}_i) + (1 - m_i) \log(1 - \hat{m}_i)]$$

where m_i represents the observed mask, and $\widehat{m_i}$ represents the discriminator's predicted probability that a value is observed. The loss measures the discriminator's ability to distinguish between observed (real) and imputed (fake) values. The discriminator is trained to maximize this loss, while the generator is trained to minimize it. The generator is penalized when the discriminator successfully identifies imputed values, encouraging it to generate more plausible imputations over time.

The total loss function used to train the generator is a weighted combination of the adversarial loss, $\mathcal{L}_{\mathcal{G}}$, and the reconstruction loss, $\mathcal{L}_{\mathcal{M}}$. The adversarial loss, $\mathcal{L}_{\mathcal{G}}$, becomes smaller when the generator successfully fools the discriminator into classifying imputed values as observed. On the other hand, the reconstruction loss, $\mathcal{L}_{\mathcal{M}}$, ensures that the generator's output for observed features remains close to the observed values. This term, typically expressed as a Mean Squared Error (MSE) between the observed features and the generator's output for those features, is minimized when the imputed data for the observed entries is close to the true values.

The generator, G, is trained to minimize the weighted sum of these two loss terms:

$$\min_{G} \sum_{j=1}^{k_G} \mathcal{L}_{\mathcal{G}}\big(m(j), \hat{m}(j), b(j)\big) + \alpha \mathcal{L}_{\mathcal{M}}\big(\tilde{x}(j), \hat{x}(j)\big)$$

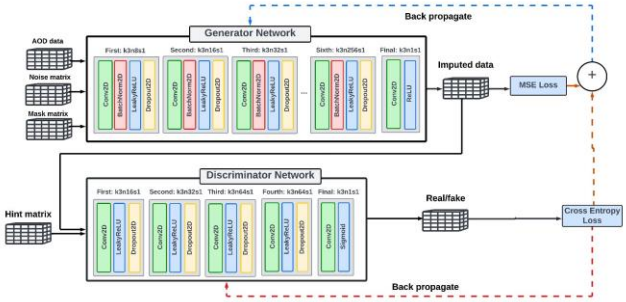

Figure 2: GAIN model architecture diagram

Figure 2 showcases the detailed architecture of the generator and discriminator networks, highlighting the convolutional layers and their specifications. In the generator network, each convolutional layer is followed by batch normalization, LeakyReLU activation, and dropout layers to maintain robust training. The first layer uses a kernel size 3x3 with 8 output channels (k3n8s1), meaning the filter is 3x3 with a stride of 1, producing eight feature maps. As the layers progress, the output channels increase to 16, with the kernel and stride remaining consistent at 3x3 and stride 1, respectively. Similarly, the discriminator network employs convolutional layers with varying numbers of output channels, starting with eight and increasing to 64 in the final layers. Each convolution operation applies a 3x3 kernel, maintains spatial features, and uses stride 1 to preserve resolution. This architecture ensures that both networks learn spatial features effectively while maintaining the integrity of the input data during the adversarial training process.

4.3 Baseline models for imputation comparison

Two baseline models, MissForest and MICE, were used to compare the performance of GAIN. MissForest is a non-parametric imputation method that uses random forests to handle both continuous and categorical data [26]. It iteratively builds random forests to predict missing values based on observed data. It is particularly effective for complex datasets with non-linear relationships and mixed variable types. MICE, on the other hand, is a widely used statistical method that imputes missing data by performing multiple iterations of regressions [1, 24]. Each variable with missing values is treated as the dependent variable in a regression model, while the other variables act as predictors. This process is repeated for each incomplete variable until the imputations stabilize.

4.4 Training and Hyperparameter Tuning

The input data for this study comprised daily MAIAC AOD observations along with meteorological variables from ERA-5 ECMWF and aerosol species data from MERRA-2, covering the years 2021 and 2022. The data was randomly split into three subsets: training (65%), validation (15%), and test (20%). The randomization ensured that each subset maintained representative coverage of the spatiotemporal variability in the dataset, avoiding any potential biases. The training set was used for model fitting, the validation set was reserved for hyperparameter tuning, and the completely independent test set was used for final evaluation to assess the model's generalization capabilities. The year 2023 dataset was designated to assess the model's effectiveness in handling a real-world scenario by imputing AOD values during the 2023 Canadian wildfire, which impacted air quality in the Northeast region, offering a critical use case for the model's ability to manage extreme air quality events.

Table 1: Hyperparameter search space and their search algorithms for the GAIN model.

Hyperparameter	Search Space	Algorithm
learning rate	1e-4 - 1e-2	Log-uniform distribution
Batch size	8, 16	Grid Search
Optimizers	Adam, AdamW, SGD	Choice
Activation function	ReLU, LeakyReLU	Choice
Alpha	0.1, 0.01, 0.05	Choice
Weight decay	1e-6 - 1e-4	Log-uniform distribution

In this study, hyperparameter tuning was conducted to optimize the performance of the Generative Adversarial Imputation Network (GAIN) model by systematically searching for the best combination of hyperparameters. Table 1 summarizes the key hyperparameters such as learning rate, batch size, optimizers, activation functions, alpha and weight decay, their search space, and their search algorithm. The tuning process explored continuous and discrete hyperparameter spaces using a combination of log-uniform distribution, grid search, and choice-based methods.

4.5 Accuracy assessment

The performance of the AOD imputation models was evaluated using several metrics, including the coefficient of determination (R^2), mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE). Each metric provides a distinct insight into the model's accuracy and reliability, offering a well-rounded assessment of the imputation results. The formulas for these metrics are presented below.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

In these formulas, $\hat{y}_i$ represents the predicted AOD values, y_i denotes the observed (true) AOD values and $\bar{y}$ is the mean of the observed AOD values across the dataset. The difference between the predicted and true values $\hat{y}_i - y_i$ quantifies the error in each prediction, which is then aggregated across all samples to assess the overall model performance.

The AOD imputation models were further validated using ground-based AERONET measurements after the evaluation metrics. AOD retrievals from the MAIAC product were matched to the AERONET observations to ensure spatial and temporal consistency. For spatial matching, a 3.5 km radius was applied around each AERONET site, and the satellite pixels within this radius were averaged to generate a single value representing the AOD at that location. AERONET measurements within a 30-minute window of the MODIS satellite overpass were averaged for temporal matching. The correlation between the imputed AOD values and the AERONET AOD was then calculated, providing a robust measure of the model's accuracy in reproducing ground-truth observations.

5 Experiments and Results

5.1 Descriptive Statistics

Table 2 shows the seasonal variation in mean AOD, standard deviation of AOD, and the mean missing rate for 2021, 2022, and 2023. One significant observation from the data is the seasonal variation in the missing rates, particularly the consistently high ones during winter. Winter 2021 exhibits the highest missing rate at 85%, followed closely by Winter 2022 at 83%. These elevated rates reflect the challenges in satellite observation during winter, as cloud cover, snow, and fog are prevalent in colder climates, obstructing the collection of clear satellite data. This pattern is consistent with previous research, which has shown that missing rates tend to be higher during the winter season due to persistent cloud cover and adverse weather conditions [3, 21]. In contrast, the summer months generally show lower missing rates, with Summer 2022 recording the lowest missing rate at 60%. This trend is likely due to clearer skies and fewer atmospheric disturbances in warmer months, which allow for more consistent satellite data collection. Meanwhile, summer also tends to have higher mean AOD values, such as 0.255 in Summer 2021 and 0.31 in Summer 2023, likely due to increased atmospheric activity like wildfires and human emissions. Winter months, by comparison, display lower mean AOD values, possibly due to reduced aerosol activity. These seasonal variations in AOD values and missing rates demonstrate the significant influence of atmospheric conditions on satellite-derived data. The frequency of missing data during winter underscores the challenges posed by harsh weather conditions, while the more precise conditions in summer enable more reliable data collection. These patterns highlight the need to consider seasonal effects when analyzing AOD data carefully.

Table 2: Descriptive statistics on MAIAC AOD and missing rate

	2021				2022				2023			
	Spring	Summer	Fall	Winter	Spring	Summer	Fall	Winter	Spring	Summer	Fall	Winter
Mean AOD	0.15	0.255	0.104	0.079	0.110	0.142	0.088	0.105	0.17	0.31	0.12	0.10
SD AOD	0.020	0.075	0.022	0.039	0.028	0.014	0.029	0.017	0.10	0.15	0.08	0.02
Mean missing rate	67%	71%	65%	85%	71%	60%	0%	83%	63%	70%	61%	83%

5.2 Accuracy assessment

We implemented a 10-fold cross-validation to ensure robust model evaluation and to guard against overfitting. Throughout the training, the learning process remained stable, and the model parameters converged effectively, with the generator and discriminator showing balanced performance, indicated by an average generator loss of 0.0377 and discriminator loss of 0.0342. The histograms in Figure 3, comparing imputed AOD data (left) and original AOD data (right), demonstrate the GAIN model's effectiveness in maintaining the original data distribution, even under conditions where the missingness is likely MNAR. The similarity between the distributions suggests that the model effectively accounted for systematic missingness, such as cloud cover, and captured the underlying aerosol patterns. By accurately replicating the original AOD distribution, the model demonstrates its capability to handle

the challenges posed by MNAR data linked to specific environmental conditions. The comparison further reveals that the range of values remains consistent across both the imputed and original datasets, with most AOD values clustered between 0.0 and 0.4 and extending to 1.4 in both cases. Additionally, while the number of pixels increases due to the imputation process, the relative proportion of pixels within each range remains comparable to the original data. This indicates that the imputed data not only aligns with the overall distribution but also preserves the pattern of AOD values, ensuring that the imputation process accurately reflects the spatial and statistical characteristics of the original dataset.

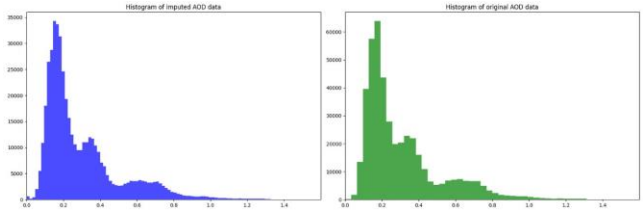

Figure 3: Comparison of the histograms of imputed AOD data (left) and original AOD data (right), demonstrating the distribution of values before and after imputation.

The GAIN model performed well in terms of accuracy, achieving an average test MSE of 0.0046, an MAE of 0.0458, and an RMSE of 0.0676. The model also demonstrated strong predictive power, with an R^2 value of 0.94, indicating a high correlation between observed and imputed AOD values. These results remained consistent across all cross-validation folds, demonstrating stability and generalizability. The imputation increased data coverage to 100% by filling all missing AOD values, and the post-imputation AOD distribution maintained consistent spatial trends, with higher AOD values in urban regions compared to rural areas. As shown in Figures 4(a) to 4(d), the results highlight that AOD levels remain higher in urban areas across all seasons, especially during spring and summer, when aerosol concentrations are generally more elevated. Regions with high cloud cover exhibited lower imputation accuracy due to large proportions of missing data, leading to slightly lower R^2 values in these areas.

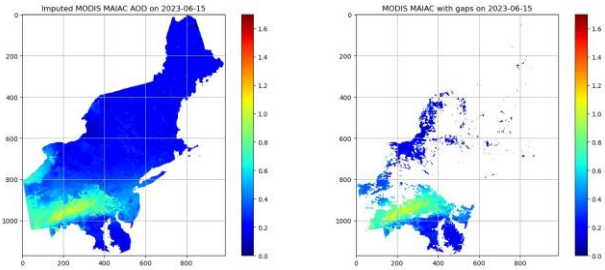

Figure 4(a): Imputed vs. original MODIS MAIAC AOD in spring (June 15, 2023)

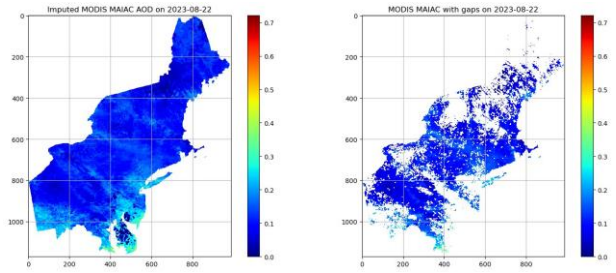

Figure 4(b): Imputed vs. original MODIS MAIAC AOD in summer (August 22, 2023).

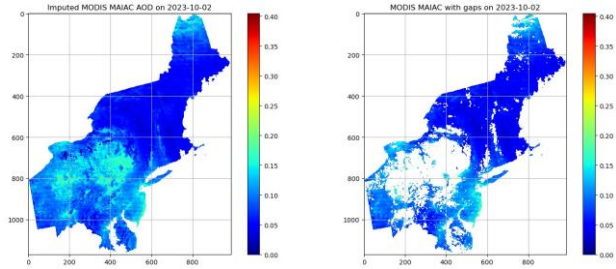

Figure 4(c): Imputed vs. original MODIS MAIAC AOD in fall (October 2, 2023).

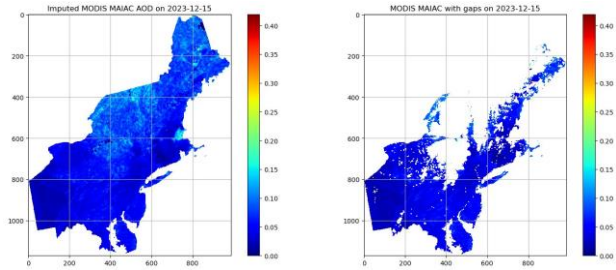

Figure 4(d): Imputed vs. original MODIS MAIAC AOD in winter (December 15, 2023).

Figure 5 shows the scatterplot that illustrates the relationship between AERONET AOD (x-axis) and imputed MODIS MAIAC AOD (y-axis) values. The correlation between the two datasets is strong, as indicated by the R^2 value of 0.899, demonstrating that the imputed MODIS values generally follow the trend of the AERONET observations. However, the slope of the regression line, 0.735, indicates a systematic underestimation of AOD values by MODIS. Most of the data points fall below the 1:1 line, which would represent a perfect agreement between the two datasets, emphasizing this bias. The underestimation becomes more pronounced at higher AOD values, where the gap between the regression and 1:1 line widens.

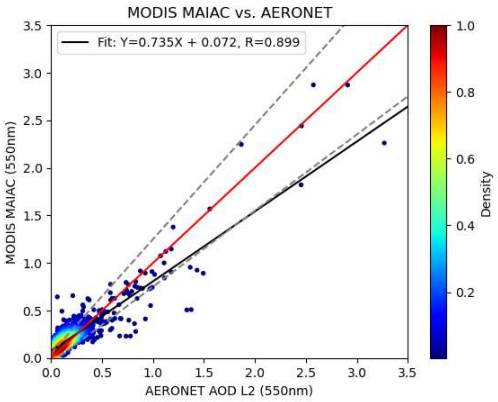

Figure 5: Scatter plot of MODIS MAIAC AOD vs. AERONET AOD with a linear fit (black) and 1:1 line (red), showing R^2 = 0.899.

The highest density of points is observed at lower AOD values (below 0.5), where the imputed MODIS AOD values align more closely with AERONET. However, as the AOD values increase beyond 1.5, the data points have greater spread and variability, with some points deviating significantly from both the regression and 1:1 lines. This suggests that while the imputation is more reliable for lower AOD values, it struggles with accuracy at higher values. The overall trend indicates that the imputation model performs well at capturing the general relationship but may need further adjustment to reduce bias and improve performance at higher AOD levels.

5.3 Comparison across various imputation models

In this experiment, we simulated varying levels of missingness in the MAIAC AOD dataset, ranging from 10% to 80%, to evaluate how different imputation models handled increasing amounts of missing data. Two baseline imputation methods, MissForest and MICE, were included for comparison against the Generative Adversarial Imputation Networks (GAIN) model. Each model's performance was assessed using Mean Squared Error (MSE) and the coefficient of determination (R^2). These metrics provided a comprehensive evaluation of each model's ability to accurately attribute missing data and maintain predictive reliability as the level of missingness increased.

Figure 6, which plotted the MSE mean vs. missingness level, revealed that GAIN consistently delivered superior performance, maintaining the lowest MSE across all simulated missingness levels. Even as the percentage of missing data reached 80%, GAIN effectively controlled the error, indicating its robustness in dealing with high levels of missingness. MissForest, while competitive at lower levels of missingness, began to struggle as missingness exceeded 60%, resulting in a sharp increase in MSE, particularly at 70% and 80%. This suggested that MissForest had a threshold beyond which its imputation quality declined significantly. Conversely, MICE showed the least effective performance across the entire range of missingness levels. Its MSE increased steadily as missingness rose, indicating its difficulty in managing high levels of missing data. This behavior demonstrated that traditional imputation models like MICE were less suited for handling extensive missingness in datasets like MAIAC AOD.

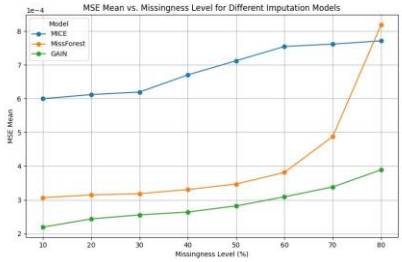

Figure 6: MSE trends across varying missingness levels for MICE, MissForest, and GAIN models.

Figure 7, which presented R^2 vs. missingness level, further illustrated the superiority of GAIN in these experiments. GAIN maintained the highest R^2 values across all levels of missingness, demonstrating that its imputation results remained closely aligned with the original data, even when large portions of the data were missing. MissForest showed stable performance at lower missingness levels, but its R^2 declined sharply as missingness increased beyond 60%, reflecting a loss of accuracy in its imputations. MICE, consistent with its performance in the MSE plot, exhibited the lowest R^2 values, with a noticeable drop-off as the missingness level increased. This indicated that MICE's imputations deviated more significantly from the original data as the proportion of missing data grew, making it less reliable for handling high levels of missingness.

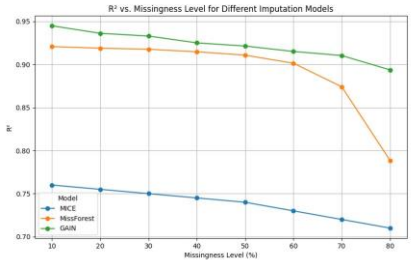

Figure 7: R^2 trends for MICE, MissForest, and GAIN models across varying missingness levels.

Overall, the results from these figures highlighted GAIN's ability to effectively impute missing data across a wide range of missingness levels, outperforming both MissForest and MICE. While MissForest performed reasonably well up to moderate levels of missingness, it became less effective as the percentage of missing data increased. MICE, on the other hand, consistently struggled with missing data, particularly at higher levels. These findings underscored the importance of selecting robust imputation methods like GAIN, especially when dealing with datasets prone to high levels of missingness, such as satellite-derived AOD data.

5.5 Use Case Evaluation

The Canada wildfire event in June 2023 provides a valuable use case for applying the Generative Adversarial Imputation Network (GAIN) model in filling gaps in satellite-derived aerosol optical depth (AOD) data. Both MODIS MAIAC satellite data and ground-

based AERONET measurements captured a significant rise in aerosol concentrations during the event, with AOD values exceeding 3.5 around June 6 - 7, 2023. This rapid increase in AOD levels, far above the typical range for the region, underscores the severity of the wildfire's impact on air quality. However, satellite observations are often limited by missing data, particularly during extreme events like wildfires, when cloud cover, smoke, and other atmospheric conditions can obstruct accurate readings.

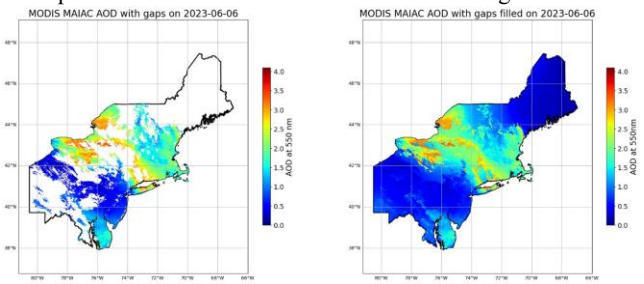

Figure 8: MODIS MAIAC AOD on June 6, 2023, during the air quality impact caused by the Canadian wildfire, showing the original data with gaps (left) and the imputed data after gap filling (right).

In this context, the GAIN model proves particularly useful by imputing missing AOD values and ensuring complete datasets for such critical air quality events. The left panel of Figure 8 shows the MODIS MAIAC AOD data with significant gaps due to cloud cover and other interference during the wildfire. These missing values are problematic when attempting to monitor the full extent of the event's impact. After applying the GAIN model, the right panel demonstrates the results, effectively filling in the gaps and providing a more comprehensive and continuous spatial representation of AOD levels across the region.

The ability to impute missing data during a critical event like this wildfire is crucial for accurate air quality assessments and public health responses. By generating a complete AOD map, the GAIN model allows researchers to better understand the distribution and concentration of aerosols, even in areas where direct satellite observations were unavailable. This, in turn, improves the reliability of the overall dataset and ensures that no significant data is lost, particularly during periods of heightened pollution.

Conclusion

In this study, we applied the Generative Adversarial Imputation Network (GAIN) for imputing missing aerosol optical depth (AOD) values from MODIS MAIAC satellite data, addressing the limitations of traditional imputation techniques, which assume data is Missing Completely at Random (MCAR) or Missing at Random (MAR). The GAIN model, designed to handle Missing Not at Random (MNAR) data, effectively filled significant gaps in AOD data, especially during critical air quality events like the 2023 Canada wildfire. Our model consistently demonstrated strong performance, with an average test generator loss of 0.0377 and discriminator loss of 0.0342. It achieved a Mean Squared Error (MSE) of 0.0046, Mean Absolute Error (MAE) of 0.0458, Root Mean Squared Error (RMSE) of 0.0676, and an R^2 of 0.94. Furthermore, the imputed AOD data aligned well with ground-based AERONET measurements, achieving a validation result of $R = 0.89$.

Hyperparameter tuning was critical to maximizing model performance, involving optimizers, learning rates, and weight decay parameters. In Appendix B, we discuss the results of hyperparameter tuning. The tuning results showed that the RMSprop optimizer, with a low learning rate and minimal weight decay, provided the most stable and lowest error results. This configuration minimized fluctuations in MAE and MSE across all iterations. In contrast, the Adam optimizer with a higher learning rate exhibited higher variability, highlighting the model's sensitivity to learning rates. The trials reinforced the importance of careful hyperparameter selection to ensure optimal model performance and stability.

Overall, this research highlights the effectiveness of GAIN in filling missing AOD values during periods of severe pollution events, demonstrating its capability to preserve the integrity of satellite data. Compared to traditional imputation methods, GAIN provided a more reliable representation of aerosol concentrations, supporting better monitoring and public health interventions during extreme air quality events.

ACKNOWLEDGMENTS

The authors gratefully acknowledge NASA for providing the MODIS MAIAC AOD data used in this study. We also extend our gratitude to the Global Modeling and Assimilation Office (GMAO) for providing the MERRA-2 aerosol data (MERRA-2 tavg1_2d_aer_Nx: 2d, 1-Hourly, Time-averaged, Single-Level, Assimilation, Aerosol Diagnostics V5.12.4).

REFERENCES

[1] Van Buuren, S. and Groothuis-Oudshoorn, K. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 45, 3 (Dec. 2011), 1–67. DOI:https://doi.org/10.18637/JSS.V045.I03.

[2] Chen, Z.Y. et al. 2020. Comparison of different missing-imputation methods for MAIAC (multiangle implementation of atmospheric correction) AOD in estimating daily PM2.5 levels. *Remote Sensing*. 12, 18 (Sep. 2020). DOI:https://doi.org/10.3390/RS12183008.

[3] Christopher, S.A. and Gupta, P. 2010. Satellite remote sensing of particulate matter air quality: the cloud-cover problem. *Journal of the Air & Waste Management Association (1995)*. 60, 5 (2010), 596–602. DOI:https://doi.org/10.3155/1047-3289.60.5.596.

[4] Goodfellow, I. et al. 2020. Generative Adversarial Networks. *COMMUNICATIONS OF THE ACM*. 63, 11 (2020). DOI:https://doi.org/10.1145/3422622.

[5] Hersbach, H. et al. 2020. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*. 146, 730 (Jul. 2020), 1999–2049. DOI:https://doi.org/10.1002/QJ.3803.

[6] Holben, B.N. et al. 1998. AERONET - A federated instrument network and data archive for aerosol

characterization. *Remote Sensing of Environment*. 66, 1 (Oct. 1998), 1–16. DOI:https://doi.org/10.1016/S0034-4257(98)00031-5.

[7] Jin, X. et al. 2022. The different sensitivities of aerosol optical properties to particle concentration, humidity, and hygroscopicity between the surface level and the upper boundary layer in Guangzhou, China. *Science of The Total Environment*. 803, (Jan. 2022), 150010. DOI:https://doi.org/10.1016/J.SCITOTENV.2021.150010.

[8] Jinnagara Puttaswamy, S. et al. 2014. Statistical data fusion of multi-sensor AOD over the Continental United States. *Geocarto International*. 29, 1 (2014), 48–64. DOI:https://doi.org/10.1080/10106049.2013.827750.

[9] Kahn, R.A. et al. 2007. Satellite-derived aerosol optical depth over dark water from MISR and MODIS: Comparisons with AERONET and implications for climatological studies. *Journal of Geophysical Research: Atmospheres*. 112, D18 (Sep. 2007), 18205. DOI:https://doi.org/10.1029/2006JD008175.

[10] Karl, T. and Koss, W. 1984. Regional and national monthly, seasonal, and annual temperature weighted by area, 1895-1983. (1984).

[11] Lecun, Y. et al. 2015. Deep learning. *Nature 2015 521:7553*. 521, 7553 (May 2015), 436–444. DOI:https://doi.org/10.1038/nature14539.

[12] Li, H. et al. 2023. Multistate time series imputation using generative adversarial network with applications to traffic data. *Neural Computing and Applications*. 35, 9 (Mar. 2023), 6545–6567. DOI:https://doi.org/10.1007/S00521-022-07961-4.

[13] Li, L. 2021. High-resolution mapping of aerosol optical depth and ground aerosol coefficients for mainland china. *Remote Sensing*. 13, 12 (Jun. 2021). DOI:https://doi.org/10.3390/RS13122324.

[14] Li, L. et al. 2020. Spatiotemporal imputation of MAIAC AOD using deep learning with downscaling. *Remote Sensing of Environment*. 237, (Feb. 2020), 111584. DOI:https://doi.org/10.1016/J.RSE.2019.111584.

[15] Lyapustin, A. et al. 2018. MODIS Collection 6 MAIAC algorithm. *Atmospheric Measurement Techniques*. 11, 10 (Oct. 2018), 5741–5765. DOI:https://doi.org/10.5194/AMT-11-5741-2018.

[16] Lyapustin, A. et al. 2011. Multiangle implementation of atmospheric correction (MAIAC): 1. Radiative transfer basis and look-up tables. *Journal of Geophysical Research: Atmospheres*. 116, D3 (Feb. 2011), 3210. DOI:https://doi.org/10.1029/2010JD014985.

[17] Lyapustin, A. et al. 2011. Multiangle implementation of atmospheric correction (MAIAC): 2. Aerosol algorithm. *Journal of Geophysical Research Atmospheres*. 116, 3 (2011). DOI:https://doi.org/10.1029/2010JD014986.

[18] Nguyen, H. et al. 2012. Spatial Statistical Data Fusion for Remote Sensing Applications. *Journal of the American Statistical Association*. 107, 499 (2012), 1004–1018. DOI:https://doi.org/10.1080/01621459.2012.694717.

[19] Nirala, M. 2008. Technical Note: Multi-sensor data fusion of aerosol optical thickness. *International Journal of Remote Sensing*. 29, 7 (Apr. 2008), 2127–2136. DOI:https://doi.org/10.1080/01431160701395336.

[20] Pu, Q. and Yoo, E.-H. 2022. A gap-filling hybrid approach for hourly PM2.5 prediction at high spatial resolution from multi-sourced AOD data. *Environmental Pollution*. 315, (2022), 120419. DOI:https://doi.org/https://doi.org/10.1016/j.envpol.2022.120419.

[21] Pu, Q. and Yoo, E.H. 2021. Ground PM2.5 prediction using imputed MAIAC AOD with uncertainty quantification. *Environmental Pollution*. 274, (Apr. 2021), 116574. DOI:https://doi.org/10.1016/J.ENVPOL.2021.116574.

[22] Radford, A. et al. UNSUPERVISED REPRESENTATION LEARNING WITH DEEP CONVOLUTIONAL GENERATIVE ADVERSARIAL NETWORKS.

[23] Randles, C.A. et al. 2017. The MERRA-2 Aerosol Reanalysis, 1980 Onward. Part I: System Description and Data Assimilation Evaluation. *Journal of Climate*. 30, 17 (Sep. 2017), 6823–6850. DOI:https://doi.org/10.1175/JCLI-D-16-0609.1.

[24] Rubin, D.B. 1987. Multiple Imputation for Nonresponse in Surveys. (Jun. 1987). DOI:https://doi.org/10.1002/9780470316696.

[25] Sayer, A.M. et al. 2013. Validation and uncertainty estimates for MODIS Collection 6 "deep Blue" aerosol data. *Journal of Geophysical Research Atmospheres*. 118, 14 (Jul. 2013), 7864–7872. DOI:https://doi.org/10.1002/JGRD.50600.

[26] Stekhoven, D.J. and Bühlmann, P. 2012. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 28, 1 (Jan. 2012), 112–118. DOI:https://doi.org/10.1093/BIOINFORMATICS/BTR597.

[27] Tayebi, A. et al. 2023. Contributions from experimental geostatistical analyses for solving the cloud-cover problem in remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*. 118, (Apr. 2023), 103236. DOI:https://doi.org/10.1016/J.JAG.2023.103236.

[28] Walcek, C.J. 1994. Cloud Cover and Its Relationship to Relative Humidity during a Springtime Midlatitude Cyclone. *Monthly Weather Review*. 122, 6 (Jun. 1994), 1021–1035. DOI:https://doi.org/10.1175/1520-0493(1994)122.

[29] Wang, Z. et al. 2024. Severe Global Environmental Issues Caused by Canada's Record-Breaking Wildfires in 2023. *Advances in Atmospheric Sciences*. 41, 4 (Apr. 2024), 565–571. DOI:https://doi.org/10.1007/S00376-023-3241-0/METRICS.

[30] Wang, Z. et al. 2024. Severe Global Environmental Issues Caused by Canada's Record-Breaking Wildfires in 2023. *Advances in Atmospheric Sciences*. Science Press.

[31] Xiao, Q. et al. 2021. Evaluation of gap-filling approaches in satellite-based daily PM2.5 prediction models. *Atmospheric Environment*. 244, (Jan. 2021). DOI:https://doi.org/10.1016/J.ATMOSENV.2020.117921.

[32] Xiao, Q. et al. 2017. Full-coverage high-resolution daily PM2.5 estimation using MAIAC AOD in the Yangtze River Delta of China. *Remote Sensing of Environment*. 199, (Sep. 2017), 437–446. DOI:https://doi.org/10.1016/J.RSE.2017.07.023.

[33] Yang, J. and Hu, M. 2018. Filling the missing data gaps of daily MODIS AOD using spatiotemporal interpolation. *Science of The Total Environment*. 633, (Aug. 2018), 677–683. DOI:https://doi.org/10.1016/J.SCITOTENV.2018.03.202.

[34] Yeh, R.A. et al. Semantic Image Inpainting with Deep Generative Models.

[35] Yoon, J. et al. 2018. GAIN: Missing Data Imputation using Generative Adversarial Nets. *35th International Conference on Machine Learning, ICML 2018*. 13, (Jun. 2018), 9042–9051.

[36] Yu, X. et al. 2024. A novel algorithm for full-coverage daily aerosol optical depth retrievals using machine learning-based reconstruction technique. *Atmospheric Environment*. 318, (2024), 120216. DOI:https://doi.org/https://doi.org/10.1016/j.atmosenv.2023.120216.

[37] Zeng, Q. et al. 2023. Full-coverage estimation of PM2.5 in the Beijing-Tianjin-Hebei region by using a two-stage model. *Atmospheric Environment*. 309, (Sep. 2023), 119956. DOI:https://doi.org/10.1016/J.ATMOSENV.2023.119956.

[38] Zhang, J. and Reid, J.S. 2006. MODIS aerosol product analysis for data assimilation: Assessment of over-ocean level 2 aerosol optical thickness retrievals. *Journal of Geophysical Research: Atmospheres*. 111, D22 (Nov. 2006). DOI:https://doi.org/10.1029/2005JD006898.

[39] Zhang, Y. et al. 2023. A systematic review of generative adversarial imputation network in missing data imputation. *Neural Computing and Applications*. 35, 27 (Sep. 2023), 19685–19705. DOI:https://doi.org/10.1007/S00521-023-08840-2/TABLES/3.

[40] Zhou, Y. et al. 2024. Missing Data Imputation: Do Advanced ML/DL Techniques Outperform Traditional Approaches? (2024), 100–115. DOI:https://doi.org/10.1007/978-3-031-70381-2.

[41] Zubko, V. et al. 2010. Study of data-merging and interpolation methods for use in an interactive online analysis system: MODIS terra and aqua daily aerosol case. *IEEE Transactions on Geoscience and Remote Sensing*. 48, 12 (Dec. 2010), 4219–4235. DOI:https://doi.org/10.1109/TGRS.2010.2050893.

APPENDIX

A Data Description

1. *MAIAC AOD:* The satellite Aerosol Optical Depth (AOD) data for this research is obtained from the MODIS MAIAC (Multi-Angle Implementation of Atmospheric Correction) algorithm. Developed by [16, 17], MAIAC is designed to retrieve AOD values from Terra and Aqua MODIS products over both bright and dark surfaces [38]. Terra and Aqua are polar-orbiting satellites, providing daily AOD products at a 1 km × 1 km spatial resolution at approximately 10:30 and 13:30 local time, respectively. MAIAC primarily relies on cloud-free pixels for quality control to ensure accurate aerosol-surface retrievals.
2. *AERONET AOD:* This research utilizes ground-based measurements from AERONET (Aerosol Robotic Network). AERONET is a globally distributed ground-based sun photometer designed to measure aerosol optical properties and atmospheric constituents precisely [6, 9]. In this study, we used level 2.0 data to validate imputed AOD values. Data was collected from 16 AERONET stations dispersed throughout the NE region (Figure 1). It should be noted that AERONET did not make direct observations at 550 nm. To address this, the study employed the Ångström index corresponding to the 440 nm – 675 nm wavelength, enabling the interpolation of the AOD Ångström index at 550nm (α) [25].
3. *Aerosol species*: In this study, we utilized the M2T1NXAER (or tavg1_2d_aer_Nx) dataset from MERRA-2 [23], an hourly collection of aerosol data provided at a spatial resolution of 0.5° × 0.625°. This product provides detailed diagnostics on aerosols, including the surface mass concentrations for various components such as black carbon, dust, sea salt, sulfate, and organic carbon. Additionally, it offers the total aerosol optical thickness (AOT) at 550 nm, capturing the extinction properties of aerosols.
4. *Meteorological variables*: The meteorological variables required for this study were obtained from the ERA5 dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). This dataset offered comprehensive meteorological data with global coverage, utilizing a regular latitude-longitude grid projection [5]. The data was provided in GRIB file format, with a spatial resolution of 0.25° x 0.25° for atmosphere reanalysis. The meteorological variables used in this study included boundary layer height, total cloud cover (TCC), relative humidity, temperature at 2m, and wind speed.

5. *Geographical variables.* For this study, the Global Multi-resolution Terrain Elevation Data 2010 (GMTED2010) elevation data, with a spatial resolution of 30-arc seconds (approximately 1 km), was downloaded from the GMTED USGS (2023). Likewise, the land use and land cover (LULC) information was obtained from the National Land Cover Database (NLCD), available at MRLC (2023). The NLCD, developed with a 30 m resolution and utilizing a 16-class legend, adhered to a modified Anderson Level II classification system.

B Hyperparameter Tuning

Figure 9 shows the hyperparameter tuning results by tracking the Mean Absolute Error (MAE) and Mean Squared Error (MSE) over several training iterations for different combinations of optimizers, learning rates, and weight decay. The figure highlights the impact of these hyperparameters on model performance and training stability. The key findings from the tuning experiments are outlined below:

1. The green line demonstrates the combination of RMSprop with a learning rate of 0.000064 and weight decay of 0.000001, consistently producing the lowest MAE and MSE across all training iterations. This configuration showed high model stability, with minimal fluctuations in error, indicating that the model responds well to small variations during training and maintains consistent performance with this set of hyperparameters.
2. The orange line reveals that the Adam optimizer with a higher learning rate of 0.001842 introduced significant variability in MAE and MSE, particularly in the initial iterations. The high sensitivity to this learning rate caused considerable instability in the training process, leading to erratic behavior and slower convergence. This suggests that the model is highly sensitive to a learning rate that is too high.
3. The red line for the SGD optimizer with a learning rate of 0.000016 and weight decay of 0.000002 exhibited high initial error and failed to show effective error reduction. The model's poor performance and inability to reduce MAE and MSE significantly indicate that this configuration leads to unstable and ineffective optimization, with sensitivity to low learning rates causing the model to stagnate.
4. The blue line shows the Adam optimizer with a lower learning rate of 0.000058 and weight decay of 0.000021, which provided steady and relatively low MAE and MSE throughout the training iterations. However, while this configuration improved model stability, it did not achieve the same level of error minimization as RMSprop, indicating that the model remains sensitive to optimizer choice, even when using stable learning rates.

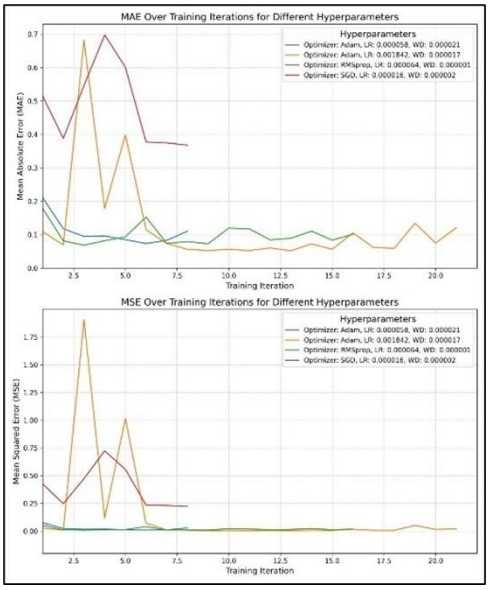

Figure 9: Comparison of MAE and MSE over training iterations for different hyperparameter configurations.