A 1D CNN-Based Emulator of CMAQ: Predicting NO\textsubscript{2} Concentration over the Most Populated Urban Regions in Texas

MAHSA PAYAMI,\textsuperscript{a} YUNSOO CHOI,\textsuperscript{a} AHMED KHAN SALMAN,\textsuperscript{a} SEYEDALI MOUSAVINEZHAD,\textsuperscript{a} JINCHEOL PARK,\textsuperscript{a} AND ARMAN POUYAEI\textsuperscript{a,b}

\textsuperscript{a} Department of Earth and Atmospheric Sciences, University of Houston, Houston, Texas
\textsuperscript{b} Program in Atmospheric and Oceanic Sciences, Princeton University, Princeton, New Jersey

(Manuscript received 3 July 2023, in final form 11 January 2024, accepted 9 February 2024)

ABSTRACT: In this study, we developed an emulator of the Community Multiscale Air Quality (CMAQ) model by employing a one-dimensional (1D) convolutional neural network (CNN) algorithm to predict hourly surface nitrogen dioxide (NO\textsubscript{2}) concentrations over the most densely populated urban regions in Texas. The inputs for the emulator were the same as those for the CMAQ model, which includes emission, meteorological, and land-use/land-cover data. We trained the model over June, July, and August (JJA) of 2011 and 2014 and then tested it on JJA of 2017, achieving an index of agreement (IOA) of 0.95 and a correlation of 0.90. We also employed temporal threefold cross validation to evaluate the model’s performance, ensuring the robustness and generalizability of the results. To gain deeper insights and understand the factors influencing the model’s surface NO\textsubscript{2} predictions, we conducted a Shapley additive explanations analysis. The results revealed solar radiation reaching the surface, planetary boundary layer height, and NO\textsubscript{x} (NO + NO\textsubscript{2}) emissions as key variables driving the model’s predictions. These findings highlight the emulator’s ability to capture the individual impact of each variable on the model’s NO\textsubscript{2} predictions. Furthermore, our emulator outperformed the CMAQ model in terms of computational efficiency, being more than 900 times as fast in predicting NO\textsubscript{2} concentrations, enabling the rapid assessment of various pollution management scenarios. This work offers a valuable resource for air pollution mitigation efforts, not just in Texas, but with appropriate regional data training, its utility could be extended to other regions and pollutants as well.

SIGNIFICANCE STATEMENT: This work develops an emulator of the Community Multiscale Air Quality model, using a one-dimensional convolutional neural network to predict hourly surface NO\textsubscript{2} concentrations across densely populated regions in Texas. Our emulator is capable of providing rapid and highly accurate NO\textsubscript{2} estimates, enabling it to model diverse scenarios and facilitating informed decision-making to improve public health outcomes. Notably, this model outperforms traditional methods in computational efficiency, making it a robust, time-efficient tool for air pollution mitigation efforts. The findings suggest that key variables like solar radiation, planetary boundary layer height, and NO\textsubscript{x} (NO + NO\textsubscript{2}) emissions significantly influence the model’s NO\textsubscript{2} predictions. By adding appropriate training data, this work can be extended to other regions and other pollutants such as O\textsubscript{3}, PM\textsubscript{2.5}, and PM\textsubscript{10}, offering a powerful tool for pollution mitigation and public health improvement efforts worldwide.

KEYWORDS: Air pollution; Air quality; Atmospheric composition; Artificial intelligence; Deep learning; Other artificial intelligence/machine learning

1. Introduction

Nitrogen dioxide (NO\textsubscript{2}), classified by the U.S. Environmental Protection Agency (EPA) as one of the six major types of air pollutants known as criteria pollutants, is closely monitored and regulated by multiple countries and international organizations due to its detrimental effects on human health and public safety (EPA 2023). NO\textsubscript{2} mainly originates from fuel combustion, particularly emissions from vehicles (e.g., cars, trucks, and buses), power plants, and off-road equipment (EPA 2023), posing a significant concern in the ambient environment. Many studies have established the link between NO\textsubscript{2} exposure and cardiovascular and respiratory diseases (Wright et al. 2023; Chen et al. 2007; Faustini et al. 2014), emphasizing the increasing need for effective management strategies due to the heightened awareness and understanding of NO\textsubscript{2}’s health impacts (Huangfu and Atkinson 2020). Moreover, in the presence of sunlight, NO\textsubscript{2} undergoes photolysis and breaks down into NO and a free oxygen atom. This free oxygen atom is a key player in the formation of ozone, as it reacts with molecular oxygen (Seinfeld and Pandis 2016). Therefore, it also serves as a crucial precursor to forming ground-level ozone, which further amplifies health risks (Sillman 1999; Nelson et al. 2023).

To address such increasing concern in air quality management and the complex nature of air quality modeling, chemical transport models (CTMs) have played a pivotal role in understanding and simulating the complicated chemical and physical interactions of pollutants within the atmosphere. For
example, the Community Multiscale Air Quality (CMAQ) model, developed by the EPA, has been widely used as a comprehensive air quality modeling system that integrates meteorological information, emissions inventories, and chemical reaction schemes (Byun and Schere 2006). By leveraging these factors, CMAQ simulates the behaviors of air pollutants across different spatial and temporal scales ranging from short-term local episodes to long-term regional trends, making it a valuable tool for air quality management, policy evaluation, and health impact assessments (Appel et al. 2017). CMAQ has been extensively utilized in various studies, such as assessing the health impacts and economic consequences from elevated PM$_{2.5}$ loadings originating from wildfires (Pan et al. 2023) and estimating the health benefits associated with air quality improvement strategies (Fann et al. 2012). Although CMAQ has significantly contributed to our understanding of air pollution and facilitated decision-making processes, its application is not without challenges. The model’s reliance on solving complex systems of partial differential equations (PDEs) for transport, ordinary differential equations (ODEs) for chemistry, and algebraic equations for partitioning, culminates in substantial computational costs (Salman et al. 2024). This complexity is especially evident during high-resolution simulations or comprehensive air quality assessments over extended periods, which can restrict the model’s applicability to broader geographical scopes or longer-duration pollution events (Jiang and Yoo 2018).

In response to the computational challenges presented by traditional CTMs, there has been an increasing shift toward the adoption of emulators to expedite the simulation process (Kelp et al. 2020). These emulators are created to mirror the key dynamics of the numerical models, utilizing the same inputs to produce simulations that are both precise and computationally efficient. In particular, emulators based on machine learning (ML) and deep learning (DL) algorithms have been recognized for their excessive capabilities in capturing complex patterns and relationships within large numerical datasets (Nonnenmacher and Greenberg 2021; Kelp et al. 2020). Furthermore, the capability of DL-based models to utilize graphics processing units (GPUs) offers a stark advantage over traditional CTMs, including CMAQ, which are typically written in Fortran and designed for use only with central processing units (CPUs) (Do et al. 2023). Liu et al. (2021) successfully integrated a residual network (ResNet) emulator into CMAQ’s gas-phase chemistry solver, enhancing computation speeds by up to 85.2 times for 194 chemical species using GPU (Liu et al. 2021). Also, another study conducted by Wang et al. (2022) utilized a ResNet model in the Global Nested Air Quality Prediction Modeling System for the carbon bond mechanism Z (CBM-Z), accelerating the process by 300–750 times for 47 species (Wang et al. 2022). These developments underscore the significant computational advantages of deep learning methods in atmospheric chemistry simulations.

Deep learning models are frequently characterized by their “black box” nature, where the reasoning behind their outputs is not readily apparent, presenting a considerable challenge in understanding the process behind their decision-making (Sadeghi et al. 2022; Houdou et al. 2024). However, the advancement of explainable artificial intelligence (XAI) techniques, such as Shapley additive explanations (SHAP; Lundberg and Lee 2017), has offered a significant breakthrough in explaining the decision-making process of these complex models. SHAP values provide a granular understanding of how each input feature influences the model’s predictions, thus helping to explain the behavior of models that would otherwise be opaque. This explanatory power is crucial for validating the model’s outputs, ensuring that the important features influencing predictions are understood and the model’s decisions are trustworthy (Lundberg and Lee 2017). The application of SHAP in deep learning opens up new avenues for researchers to understand the inner workings of their models, leading to more reliable applications in different fields such as air pollution modeling (Vega García and Aznarte 2020).

The primary focus of this study is to develop a DL-based emulator of CMAQ and successfully replicate its estimation of surface NO$_2$ concentrations with comparable accuracy while achieving improved computational efficiency. Following the earlier success of our previous studies in utilizing one-dimensional (1D) convolutional neural network (CNN) algorithms for diverse applications (Sayeed et al. 2022, 2023; 2021; Ghahremanloo et al. 2023, 2021; Sadeghi et al. 2022), we developed an emulator of the CMAQ model. This model is capable of simulating hourly surface NO$_2$ concentrations across the most densely populated urban regions in the state of Texas. The focus of our study was on the summer months (June, July, and August) of 2017. We have chosen summertime because ozone levels are typically at their peak during this season, and since NO$_2$ is one the most important precursors of ozone, it makes it the most critical season for analyzing the impact and trends of NO$_2$ concentrations (Mousavinezhad et al. 2023). It is worth emphasizing that the inputs for the emulator coincided with those used in the CMAQ model. In addition, we conducted a SHAP analysis, supported by a series of feature engineering practices, in order to gain deeper insights into the emulator’s inner workings as well as to better understand the factors influencing its simulations of surface NO$_2$ concentrations. This emulator’s ability to accurately and rapidly simulate NO$_2$ concentrations lays the groundwork for assessing the effectiveness of various emission reduction strategies and informing air quality management decisions. Moreover, with appropriate regional data training, its utility could reach beyond Texas. Also, by integrating real-time data and leveraging its rapid processing capabilities, the emulator stands to enhance data assimilation techniques, leading to more dynamic and responsive air quality modeling and predictive analytics. In addition, there is potential for this emulator to be adapted for modeling other key pollutants, including O$_3$, PM$_{10}$, and PM$_{2.5}$, which we aim to explore in our forthcoming studies.

2. Method

a. Study area

Texas, the second-largest state in the United States, covers an area of 695,660 km$^2$ and is home to over 30 million people, making it a vital economic and cultural hub (U.S. Census Bureau 2023). However, it also faces significant air pollution
challenges (Pinakana et al. 2023; Li et al. 2023). For this research, the study area focuses on the most densely populated cores in Texas, including the Dallas–Fort Worth metroplex, Houston metropolitan, San Antonio, Austin, El Paso, Corpus Christi, Lubbock, Killeen, Laredo, Amarillo, and Brownsville. The spatial resolution of the study's data is 12 km with a temporal resolution of 1 h, and the detailed number of pixels for each region can be found in Table S1 in the online supplemental material. These areas, which account for more than 50% of the state's population, have high concentrations of emissions from transportation, industry, and other sources that contribute to NO2 and its adverse effects on human health. The significant population and economic activities in these urban regions highlight the need for effective air pollution modeling and improved overall air quality to protect public health and the environment. Figure 1 illustrates the urban regions selected for this study.

**b. Data preparation**

To train our 1D CNN-based emulator of CMAQ (hereinafter referred to as the emulator) to replicate the estimation of surface NO2 concentration as closely as possible, we first prepared a series of input data required by CMAQ, which included meteorological, land-use and land-cover, and emissions variables. To choose the input variables, we applied recursive feature elimination (RFE) to our deep learning model for optimal selection from nine meteorological and five land-use variables. RFE, a process of iteratively removing the least impactful features, helped identify key variables guided by the index of agreement during initial training (Salman et al. 2024).

To prepare the required meteorological data and land surface information for training the emulator, we used the Weather Research and Forecasting (WRF) Model, version 4.0, developed by the National Center for Atmospheric Research (NCAR; Skamarock and Klemp 2008). This is a widely used numerical weather prediction model. Following the same model configuration as used in our previous study over the contiguous United States (Jung et al. 2022), we simulated hourly meteorological fields and land surface conditions over the study area at a 12-km spatial resolution for June, July, and August of 2011, 2014, and 2017. We converted these WRF simulation outputs into CMAQ-compatible data structures by using the Meteorology–Chemistry Interface Processor (MCIP) before proceeding with training the emulator.

To prepare the emissions inventory data for training the emulator, we employed the Sparse Matrix Operator Kernel Emission (SMOKE) modeling system, version 4.7, which generates CMAQ-ready emissions input data from bottom-up estimates of air pollutant emissions. Using the U.S. National Emission Inventory (NEI) 2011, 2014, and 2017 datasets (Eyth et al. 2019; Eyth and Vukovich 2016) supported by the SMOKE modeling system, we obtained estimates of anthropogenic emissions over the study area for June, July, and August of 2011, 2014, and 2017. We specifically selected these years to align with the NEI’s triennial release schedule, ensuring the most accurate emission data for our model. For the same period, we prepared biogenic emissions and biomass burning emissions over the study area using the Biogenic Emission Inventory System (BEIS), version 3.61, built within SMOKE and the Fire Inventory from the National Center for...
Predictor variables Emissions NO NO mol s⁻¹ NO₂ NO2 mol s⁻¹ HONO HONO mol s⁻¹ VOC VOC_INV g s⁻¹ Meteorological variables Surface pressure PRSFC Pa Temperature at 2 m TEMP2 K Wind speed at 10 m WSPD10 m s⁻¹ Wind direction at 10 m WDIR10 ° Solar radiation reaching surface RGRND W m⁻² Mixing ratio at 2 m Q2 kg kg⁻² Planetary boundary layer PBL m Rain from nonconvective cell R.N. cm Rain from convective cell R.C. cm Land surface property Surface elevation HGT_M m Urban development Urban_Development — Surface albedo ALBEDO12M — Green fraction GREENFRAC — Landmask LANDMASK 0 or 1 (land or water) Target variable Concentration Surface NO₂ concentration ACONC_NO2 ppm

Table 1. Predictor and target variables used for training the emulator.

Atmospheric Research (FINN), version 1.5 (Wiedinmyer et al. 2011), respectively. We then merged anthropogenic, biogenic, and fire emissions to derive the bottom-up estimates of nitric oxide (NO), NO₂, nitrous acid (HONO), and nonmethane volatile organic compounds (NMVOCs) (hereinafter referred to as VOCs) across the study area at a 12-km spatial resolution. Further details about the emission modeling procedure used in this study can be found in Jung et al. (2022).

Next, to prepare the target variable for the emulator, we used the meteorological fields, land surface properties, and emissions described above as input for standard CMAQ simulations. Using CMAQ, version 5.2, with the same model configuration and initial conditions as those used in our previous study (Jung et al. 2022), we simulated hourly surface NO₂ concentrations over the study area at 12-km spatial resolution for June, July, and August of 2011, 2014, and 2017. Table 1 lists all the predictor and target variables used for training the emulator, along with the corresponding acronyms and units assigned to each variable.

Next, we combined the meteorological, emissions, land-use and land-cover data, and surface NO₂ concentrations for the years 2011 and 2014 to prepare the training set for the emulator. Meanwhile, the data from 2017 served as an evaluation dataset. This approach enables an assessment of the model’s performance on unseen data, ensuring a rigorous evaluation of its accuracy and generalizability. By using two distinct periods for training and one for evaluation, the study aims to capture the temporal variability in the emission sources and atmospheric conditions, leading to a more robust and reliable emulator model. The results obtained from this evaluation process will provide valuable insights into the model’s ability to predict surface NO₂ concentrations accurately, offering a potential alternative to the traditional CMAQ model for air quality management and decision-making.

c. Deep CNN architecture

The model’s architecture was inspired by previous research studies (Sayeed et al. 2022, 2023, 2021; Ghahremanloo et al. 2023, 2021; Sadeghi et al. 2022) that had been used for air quality modeling purposes, for example, estimation of daily ground-level NO₂ concentrations using deep learning (Ghahremanloo et al. 2021). We refined the architecture for this study’s specific spatial domain by assessing its performance with various filter sizes, kernel sizes, numbers of dense layers and neurons, and train–test splits. Our model consisted of five one-dimensional CNN layers, with 32 filters in the first layer and 64 filters in the subsequent layers, along with a kernel size of two. We also included four dense layers, each with 32 neurons, after the convolution layers. To add nonlinearity, we used the rectified linear unit (ReLU) as an activation function for all the layers. The deep learning algorithm was implemented using the Keras framework with TensorFlow as the backend and the Adam optimization function (Kingma and Ba 2014; Abadi et al. 2016; Chollet 2015). The loss function used for model training was the IOA, which aims to maximize the alignment between the emulator and actual CMAQ values by minimizing the total squared difference relative to the potential error. We utilized a cross-validation technique to track the neural network model’s training progress, with 30% of the dataset randomly reserved for testing. The model was trained for 30 epochs with a batch size of 128. The structure of the model used in this study is depicted schematically in Fig. 2, which provides a clear overview of the different layers and their connections.

d. Temporal K-fold cross validation

K-fold cross validation is a commonly used method in machine learning for evaluating the performance and generalization capabilities of a model (Bickel et al. 2009). The main idea behind k-fold cross validation is to divide the dataset into k equally sized subsets (folds), where k – 1 folds are used for

Unauthenticated | Downloaded 06/26/24 09:50 PM UTC
training the model, and the remaining fold is used for testing. This process is repeated $k$ times, with each fold being used exactly once as the test set. The model’s performance is then averaged over the $k$ iterations to obtain a more robust and reliable estimate of its accuracy and generalizability (James et al. 2021). This technique helps to mitigate the risk of overfitting as it assesses the model’s ability to generalize to unseen data and provides insights into the model’s stability across different data samples (Cawley and Talbot 2010).

In this study, we employed two distinct cross-validation strategies: threefold yearly temporal cross validation and threefold monthly temporal cross validation. As mentioned, the dataset consists of data for the months of June, July, and August from 2011, 2014, and 2017. For yearly threefold cross validation, each year was considered a fold, and the model was trained for two years and tested on the remaining year. For the monthly threefold cross validation, data from all June, July, and August months were grouped together, with each month serving as a fold. The model was trained using data from two of these months and tested on the remaining month. Utilizing these cross-validation approaches allowed us to rigorously evaluate the model’s performance across different temporal scales, ensuring a more accurate assessment of its ability to capture the complex dynamics of air quality.

e. Evaluation of model performances

To evaluate the performance of the developed emulator model, we used three key metrics: Pearson’s correlation coefficient $R$, index of agreement (IOA), and mean absolute error (MAE). The Pearson correlation coefficient is a statistical measure that captures the linear relationship between two datasets, providing a value between $-1$ and $1$. A value of 1 signifies a perfect positive correlation, $-1$ indicates a perfect negative correlation, and 0 indicates no linear correlation between the datasets (Benesty et al. 2008; Ghahremanloo et al. 2021). In addition, we employed the IOA, proposed by Willmott (Willmott 1981), which is a more robust and comprehensive measure of model performance that overcomes some of the limitations associated with correlation (Pouyaei et al. 2023). The IOA compares the mean-square error and the potential error, and its values range from 0 to 1, with 1 representing a perfect match between actual and predicted values. In addition, the MAE is a useful metric that calculates an average of the absolute differences between predicted and actual values. It provides a straightforward measure of average error magnitude and is especially helpful in assessing the scale of the prediction errors.

Utilizing these three metrics will offer a comprehensive understanding of the model’s performance and its ability to predict NO$_2$ concentrations accurately. Equations (1)–(3) present the formulas for $R$, IOA, and MAE, respectively:

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

(1)

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

and (2)

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |E_i - O_i|.$$
\[ \text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |E_i - O_i|, \]  

where \( E_i \) denotes the estimated samples produced by the emulator, and \( O_i \) represents the observed samples from the CMAQ model. Further, \( E \) corresponds to the mean value of the estimated samples produced by the emulator, and \( O \) signifies the mean value of the observed samples from the CMAQ; \( N \) is the total number of samples.

In addition, we conducted comparisons of computational times between the emulator and the original CMAQ. Because of limited access to industrially controlled environments, a direct comparison under identical CPU configurations was not feasible. Instead, an approximate measure of computational time was obtained by running both the emulator and CMAQ using the same number of CPU cores. While this approach may not provide an exact comparison, it can offer a general guideline for understanding the relative computational efficiency of the DL-based emulator. It is important to note that factors such as hardware specifications, software optimizations, and parallel processing capabilities can affect computational times. Despite these limitations, we aimed to provide end users with an indication of the computational performance of the emulator approach in comparison with the conventional CMAQ model. The simulations were conducted using available high-performance computing resources to ensure optimal performance within the given constraints.

f. Quantified assessment of the deep CNN-based emulator’s performances

**SHAP analysis**

The SHAP analysis stands out as an innovative approach that improves the interpretability of machine learning models. By leveraging principles from cooperative game theory, SHAP values shed light on how each input feature distinctly contributes to the model’s predictions, thus enabling a more nuanced understanding of the model’s decision-making processes. The insights gained from SHAP analyses can guide the validation of model outputs, ensuring that the critical features that drive predictions are well understood, which in turn aids in making informed decisions based on those models. Deep SHAP is a specialized adaptation of the SHAP framework, explicitly tailored for deep learning models (Lundberg and Lee 2017; Grahmanloo et al. 2021; Mousavinezhad et al. 2023). For this study, we have opted to utilize deep SHAP.

Also, it is crucial to eliminate collinear variables prior to conducting SHAP analysis to ensure accurate and reliable estimation of feature importances. Multicollinearity may result in biased and unstable estimates, ultimately affecting the model’s interpretability (Lundberg and Lee 2017). The presence of multicollinearity can be systematically detected by utilizing the variance inflation factor (VIF). The VIF serves as a diagnostic metric that quantifies the degree of inflation in the variances of the estimated regression coefficients due to the linear relationships among the independent variables (Neter et al. 1983):

\[ \text{VIF}_j = \frac{1}{1 - R^2_j}, \]

As demonstrated in Eq. (4), VIF\(_j\) corresponds to the \( j \)th predictor, whose values range from 1 to infinity (Thompson et al. 2017). While there is no universal consensus on a specific VIF threshold value to indicate the presence of multicollinearity, some researchers have suggested 10 as a threshold (Mukundamago et al. 2023; Abegaz et al. 2023; Thompson et al. 2017). In this study, we have also chosen 10 as the VIF threshold.

3. Results and discussion

The results presented in the subsequent sections are predominantly based on the summertime data (June, July, and August) from 2017, with the training set comprising data from the same period in 2011 and 2014. The only exception to this is the threefold temporal cross-validation section, which employs a different approach detailed in section 2.

a. Comparative analysis of CMAQ and emulator NO\(_2\) predictions

In this section, we have generated a series of hexbin plots to visually represent the relationship between the CMAQ NO\(_2\) estimates and the emulator NO\(_2\) predictions for each of the study areas (Fig. 3). These areas include the Dallas–Fort Worth metroplex, Houston metropolitan, San Antonio, Austin, El Paso, Corpus Christi, Lubbock, Killeen, Laredo, Amarillo, and Brownsville, as studied during the summer of 2017. We have also included an hourly time series for June, July, and August of 2017 for one pixel located at the center of the Dallas–Fort Worth metroplex region in the online supplemental material (Fig. S1).

Across all study areas, the plots exhibit a strong correlation of 0.90 between the CMAQ NO\(_2\) estimates and the emulator NO\(_2\) predictions. The strong correlation indicates the effectiveness of the emulator model in capturing the patterns and variations of surface NO\(_2\) levels. As an example, Fig. S1 in the online supplemental material shows the hourly time series for June, July, and August of 2017 for one grid of 12 km × 12 km in Dallas. The robust performance of this model in mimicking the CMAQ values and recognizing the patterns, underlies the potential of the emulator approach as a complementary tool to traditional CMAQ simulations, providing efficient and precise predictions promptly. While it excels in certain areas such as computational efficiency, it is crucial to acknowledge its limitations. These can include a dependency on the quality and quantity of training data and a potential inability to effectively capture rare or unexpected events that may be under-represented in the training dataset. Nevertheless, by enabling rapid execution of numerous scenarios, our emulator can significantly aid decision-making in air quality management for densely populated urban areas. Furthermore, the consistency of the emulator model’s performance across various cities, diverse emission sources, and different conditions demonstrates its adaptability and generalizability. This adaptability
and generalizability, enhanced by appropriate regional data training, can further expand its application for other regions. The highest correlation value was observed at 0.91 in the Dallas–Fort Worth metroplex, Lubbock, and Laredo regions. This strong correlation is mainly attributable to the substantial training set available for the Dallas–Fort Worth metroplex, which allowed the model to effectively learn and capture the relationships between the predictor variables and surface...
NO$_2$ concentrations. It is noteworthy that despite smaller datasets for Lubbock and Laredo, the model still performed admirably, indicating robust learning. However, a larger and more diverse dataset would enable better generalization when predicting unseen data. This outcome underscores the importance of comprehensive datasets for training machine learning models in environmental applications, given the inherently complex and multidimensional nature of environmental data patterns.

Conversely, the Killeen region, with its relatively smaller geographical area leading to a reduced size of the training dataset (Table S1 in the online supplemental material), displayed the lowest correlation of 0.88. The smaller training set size may have limited the model’s capacity to understand and capture the full complexity of the interactions between the predictor variables and surface NO$_2$ concentrations in this area. This is consistent with other studies suggesting that larger training datasets generally lead to improved model performance due to more effective learning and generalization, emphasizing the challenges posed by limited data availability when applying deep learning models (Gütter et al. 2022; Hestness et al. 2017).

During the model evaluation, the MAE was also investigated. The Houston metropolex and El Paso exhibited the highest MAE values of 1.25 and 1.34 ppb, respectively. These values indicate a larger deviation between the emulator’s predictions and CMAQ NO$_2$ values in these regions when compared with others. Several potential factors could contribute to these larger discrepancies. One consideration could be the specific atmospheric conditions prevailing in these regions. For instance, the Houston metropolex, being one of the largest metropolitan areas in the United States with significant industrial activity, may exhibit complex air quality dynamics due to variations in emissions from various sources (Souri et al. 2016). Conversely, El Paso is located in a unique geographical area that might influence the diffusion and distribution of pollutants.

**b. Threefold temporal cross validation**

In this section, we present the findings from the two cross-validation strategies employed in this study: monthly threefold temporal cross validation and yearly threefold cross validation. For the monthly threefold cross validation, the model demonstrated consistent performance across all three folds. This consistency indicates its ability to effectively capture the relationships between input variables and surface NO$_2$ concentrations on a monthly time scale. The average correlation and IOA are 0.92 and 0.96, respectively, further validating the model’s robustness in predicting surface NO$_2$ concentrations for different months. Detailed correlation and IOA results for each fold are provided in Table 2 to illustrate the model’s performance.

Similarly, for the yearly threefold cross validation, the model exhibited stable performance across the three years, showcasing its competence in predicting surface NO$_2$ concentrations on an annual time scale. The average performance metrics for the yearly cross validation, including correlation and IOA, were 0.90 and 0.95, respectively. These metrics illustrate the model’s ability to generalize across different years. Interestingly, our model demonstrated better performance in the monthly cross validation than in the yearly cross validation. This observation can be attributed to the closer temporal proximity of the datasets within a month, enabling the model to accurately capture the short-term dynamics and seasonal trends of surface NO$_2$ levels.

**c. SHAP analysis for noncollinear variables**

1) **VIF analysis and multicollinearity reduction**

We conducted an analysis of the VIF to identify multicollinearity among the predictor variables. Based on the VIF values, we detected high multicollinearity in some of the predictor variables, for example, surface pressure, surface temperature, and total VOC. This high multicollinearity can negatively impact the interpretability of the SHAP results. However, removing highly correlated variables from models, often due to multicollinearity concerns, is not always advisable. Such variables, despite their collinearity, can be crucial for model accuracy and should not be routinely excluded (O’Brien 2017). Indeed, in our study, removing variables with high multicollinearity resulted in a 2% decrease in correlation and a 1.5% decrease in IOA. Therefore, we retained all variables in the main model and provided a similar model that excluded the highly collinear variables for the SHAP analysis to ensure interpretability without compromising the model’s integrity.

Since NO and NO$_2$ emissions are extremely collinear, we defined a new variable, NO, which is the sum of NO and NO$_2$ emissions. After eliminating the highly collinear variables, the remaining predictor variables were wind speed (WSPD10), wind direction (WDIR10), solar radiation reaching the surface (RGRND), planetary boundary layer (PBL), rain from a convective cell (R.C.), rain from a nonconvective cell (R.N.), green fraction (GREENFRAC), urban development (Urban_Development), surface elevation (HGT_M), and NO$_2$ emission (NO$_2$). These predictor variables, along with their respective VIF values, were presented in Fig. 4. These variables, now with reduced multicollinearity, were then used as input for the SHAP analysis, which will help us better understand the contributions of each predictor variable.
to the model’s predictions and improve the model’s interpretability.

2) SHAP VALUES FOR EACH AREA

This study generated SHAP plots and absolute SHAP values to evaluate the importance of various predictor variables on surface NO₂ levels for all major urban areas in Texas. Among the cities analyzed, four representative cities—Austin, Corpus Christi, Houston, and Dallas–Fort Worth—were selected for a more detailed discussion in the study. Figure 5 showcases the SHAP values for these areas. SHAP plots for all the other urban areas are available in the online supplemental material (Fig. S2). These cities were chosen based on their diverse geographical locations, emission sources, and atmospheric conditions, offering a comprehensive understanding of the key factors driving surface NO₂ variations in different regions.

Interpreting SHAP plots is critical for understanding how the features impact the model’s predictions. In a SHAP summary plot, each dot represents the impact of a feature’s value on a model prediction for that particular sample. The position of the dot along the x axis shows the magnitude and direction of this impact, referred to as the SHAP value. A feature’s SHAP value indicates how much it contributes to or detracts from the model’s prediction for that sample. A feature’s impact on the change in the prediction is proportional to the distance from the plot’s vertical centerline. Points to the right of this line indicate a positive impact on the prediction, while points to the left signify a negative impact. The color scale denotes the feature value (red being high and blue being low).

Hence, this plot provides insight into each input feature’s (emission, meteorological, and land-use and land-change variables) contribution to the model’s predictions. For instance, a cluster of red points on the right side for a feature indicates that high values of this feature increase the prediction value (Lundberg et al. 2020).

To identify the most important features, the absolute SHAP values plots can be used. This plot displays the average of absolute SHAP values for each feature, ordered by their importance. The feature with the highest mean absolute SHAP value is considered the most important since it contributes the most to the model’s predictions across all variables. Lower-ranked features also contribute to the model’s predictions but to a lesser extent. In other words, features with higher SHAP values (whether positive or negative) are more influential in the model’s decision-making process (Lundberg et al. 2020; Ghahremanloo et al. 2021).

The SHAP analysis results for all the regions in this study signified that the solar radiation reaching the surface, PBL height, and NOₓ emissions emerged as the most influential variables shaping surface NO₂ predictions. An increase in solar radiation reaching the surface usually results in a decrease in NO₂ concentration due to the acceleration of photolysis rates, which facilitates the conversion of NO₂ into NO, subsequently reducing the overall concentration of NO₂ (Seinfeld and Pandis 2016; Vega García and Aznarte 2020). Similarly, an increase in PBL height is associated with a reduction in NO₂ concentrations. The PBL height signifies the atmospheric layer where the surface directly impacts the vertical distribution of pollutants. As
PBL height expands, pollutants like NO₂ are dispersed over a larger volume, leading to a decrease in their concentrations near the surface (Stull 1988). Contrastingly, an increase in NOₓ emissions directly boosts NO₂ concentrations. NOₓ, a combination of NO and NO₂, directly contributes to the NO₂ levels in the atmosphere. Therefore, areas with high NOₓ emissions are expected to have elevated NO₂ concentrations (Li et al. 2022).

Fig. 5. SHAP plots for (a) Dallas–Fort Worth metroplex, (b) Corpus Christi, (c) Houston metropolitan, and (d) Austin.
These factors play a critical role in the prediction of the surface NO2 concentrations and the SHAP results highlight the emulator’s ability to correctly capture each input feature’s (emission, meteorological, and land-use and land-change variables) contribution to the model’s predictions. In the case of Corpus Christi, the wind direction was also identified as an important predictor variable. This finding reflects the city’s unique coastal location, where wind patterns can significantly impact the transport and dispersion of air pollutants. The wind rose and polar plot diagrams for Corpus Christi in Fig. S3 of the online supplemental material demonstrate how prevailing southeast winds affect the area’s air quality by dispersing pollutants and influencing NO2 concentrations. By leveraging the SHAP analysis, this study uncovers the critical variables that drive surface NO2 concentrations, offering a detailed understanding of the factors that significantly influence CMAQ model outcomes in the selected urban areas.

d. Computational time

This section presents a comparison of the computational time between the CMAQ model and the proposed deep CNN emulator. This comparison is crucial to demonstrate the efficiency of the deep CNN model as an alternative to the numerical CMAQ model for air quality simulations. The extensive computational requirements of the CMAQ model, resulting from its complex simulation of various chemical and physical processes, are well-documented in the literature (Appel et al. 2017; Byun and Schere 2006). The development of computationally efficient alternatives, such as our emulator, significantly streamlines large-scale, high-resolution air quality simulations. Its notable efficiency is particularly important when conducting decision-making tasks that involve testing hundreds to thousands of scenarios. Traditionally, such tasks could require hours or even days of simulations with conventional CTMs. In contrast, the emulator enables rapid processing, transforming the speed of decision-making in air quality management.

The use of GPUs accelerates computational time since GPUs are designed to handle multiple parallel computations, making them well suited for computationally intensive tasks like deep learning and large-scale simulations (Molnár et al. 2010). The deep CNN model, specifically designed to run on GPUs, exhibited significantly reduced computational time relative to the CMAQ model. Table 3 presents a comparison of the computational time between the CMAQ and emulator models for a 1-h simulation on a 120 x 120 grid. The speedup factor is calculated based on the computational time of the emulator model relative to the CMAQ model for each configuration. As shown in Table 3, the emulator is more than 900 times as fast as CMAQ when using 1 CPU and 1 GPU and is more than 600 times as fast when using only 1 CPU. We utilized the Intel Xeon Gold 6252 CPU to measure the computational time for CMAQ, whereas the NVIDIA A30 and Intel Xeon Gold 6242 CPU were employed to gauge the processing time for the emulator.

The use of deep CNNs for air quality simulations, particularly in conjunction with advanced computing technologies like GPUs, has the potential to alleviate the computational demands, allowing for more comprehensive, detailed, and time-sensitive analyses. The emulator’s ability to precisely and rapidly simulate NO2 concentrations lays the groundwork for assessing various emission reduction strategies and improving decision-making processes in air quality management.

4. Conclusions

This study successfully demonstrates the applicability of a one-dimensional CNN-based emulator model in simulating surface NO2 concentrations across major urban regions in Texas. The emulator utilizes the same input (emission, meteorological, and land-use and land-cover variables) as the CMAQ model. When trained on data from the summer months of 2011 and 2014 and tested on summer 2017 data, our emulator effectively predicts hourly surface NO2 concentrations with an IOA of 0.95 and a correlation of 0.90. The model’s robust performance is further validated through monthly and yearly threefold cross validations, which consistently yield high correlation and IOA value. The SHAP analysis emphasizes the emulator’s ability to capture the individual impact of each variable on the model’s NO2 predictions, highlighting the significance of PBL height, solar radiation reaching the surface, and NOx emissions as influential variables shaping surface NO2 predictions across all studied regions. Notably, the study also emphasizes the unique influence of wind direction on NO2 concentrations in coastal regions, such as Corpus Christi.

In terms of computational efficiency, the emulator model surpasses numerical air quality simulation models like CMAQ. By utilizing GPUs, the CNN-based model drastically reduces computational time while maintaining predictive accuracy, enabling large-scale, high-resolution air quality simulations. The proposed emulator achieves almost 900 times as fast computation as CMAQ when utilizing 1 CPU and 1 GPU. This finding has wide-ranging implications, potentially facilitating more extensive, detailed, and time-sensitive air quality analyses, ultimately contributing to improved decision-making in air quality management and policy. The ability to model various scenarios in a short time frame is key to understanding and responding to air quality challenges more effectively.
Last, the use of 1D CNN in developing an emulator for air quality simulation proves to be a promising approach. It offers a robust and interpretable model that effectively captures and simulates the surface NO2 concentration levels comparable to those produced by the CMAQ model, while also achieving a significant reduction in the computational time. Future research can further extend this approach to broaden the analysis to encompass additional geographical areas and applications by training the emulator with data specific to other regions. Moreover, there is an opportunity to adapt this emulator for tracking other important air pollutants, including O3, PM10, and PM2.5. Exploring the use of higher dimensional CNNs, such as 2D and 3D CNNs, could also present interesting avenues for future work. In addition, the interpretability and transparency of the model can be further enhanced by incorporating more advanced techniques for feature importance analysis.

Acknowledgments. The research conducted in this study received partial financial support from the NASA Aura Science Team Grant (NNH19ZDA001N-AURAST). We acknowledge the support of the Research Computing Data Core at the University of Houston for their assistance with high-performance computing in this work.

Data availability statement. Individual researchers interested in accessing these materials can request permission by contacting the corresponding author. The details on model configurations and the procedure for obtaining the input data used in the data preparation phase can be found in Jung et al. (2022).

REFERENCES


