Air Quality, Atmosphere and Health
https://doi.org/10.1007/s11869-021-01126-3

Predicting PM2.5 atmospheric air pollution using deep learning with meteorological data and ground-based observations and remote-sensing satellite big data

Pratyush Muthukumar[1] · Emmanuel Cocom[1] · Kabir Nagrecha[1] · Dawn Comer[2] · Irene Burga[2] · Jeremy Taub[3] · Chisato Fukuda Calvert[3] · Jeanne Holm[2] · Mohammad Pourhomayoun[1]

Received: 10 August 2021 / Accepted: 2 November 2021

Abstract

Air pollution is one of the world's leading factors for early deaths. Every 5 s, someone around the world dies from the adverse health effects of air pollution. In order to mitigate the effects of air pollution, we must first understand it, find its patterns and correlations, and predict it in advance. Air pollution prediction requires highly complex predictive models to solve this spatiotemporal problem. We use advanced deep learning models including the Graph Convolutional Network (GCN) and Convolutional Long Short-Term Memory (ConvLSTM) to learn patterns of particulate matter 2.5 (PM 2.5) over spatial and temporal correlations. We model meteorological features with a time-series set of multidimensional weighted directed graphs and interpolate dense meteorological graphs using the GCN architecture. We also use remote-sensing satellite imagery of various atmospheric pollutant matters. We utilize government maintained ground-based PM2.5 sensor data along with remote sensing satellite imagery using a ConvLSTM to predict PM2.5 over the greater Los Angeles county area roughly 10 days in the future using 10 days of data from the past in 46-h increments. Our error results on the PM2.5 predictions over time and along each sensor location show significant improvement over existing research in the field utilizing spatiotemporal deep predictive algorithms.

Keywords Air pollution prediction · Spatiotemporal forecasting · Deep convolutional LSTM · Remote-sensing satellite imagery · Ground-based air quality sensors

Introduction

Air pollution is a deadly and growing global threat. According to the WHO (2018), around 92% of the world's population breathes in polluted air resulting in 7 million deaths annually. Air pollution is the cause of many adverse health effects including aggravated cardiovascular and respiratory illness, asthma, and emphysema. Due to ambient air pollution, the global life span has been shortened by an average of 1.8 years. In addition to adverse health effects, global air pollution costs an estimated $5 trillion annually in deaths, healthcare costs, and lost labor, according to the World Bank (WorldBank 2016). By 2050, the number of premature deaths from the exposure to particulate matter (PM), a category of air pollutants, is expected to more than double worldwide (Marchal et al. 2011). Clearly, it is paramount to the safety of the global population to find an effective and accurate solution to the complex task of mitigating ambient air pollution.

To mitigate the deadly effects of air pollution, we must first be able to understand it, discover its causes and patterns, and predict it in advance. In this paper, we apply predictive models including deep neural networks and advanced machine learning algorithms to learn correlations of spatiotemporal air pollution in various locations over time and predict for the future. When developing these state-of-the-art models, we utilized both the spatial and temporal patterns in the data. Air pollution prediction is inherently a spatiotemporal task: air pollutants travel in the air and thus affect surrounding areas (spatial correlation); air pollution concentrations in the future depend on prior concentrations (temporal correlation).

✉ Pratyush Muthukumar
pmuthuk2@calstatela.edu

Extended author information available on the last page of the article.

Published online: 23 November 2021

Air pollution prediction has been a topic of interest for decades, with the most recent approaches focusing on using the predictive capabilities of deep neural networks; see the survey paper Bellinger et al. (2017) and the references therein. Current deep learning research in the field seeks to utilize these predictive models to learn and predict either spatial correlations or temporal correlations in ambient air pollution, but we seldom see models capable of both (Abrahamsen 2018; Grover et al. 2015; Weyn et al. 2020; Narejo and Pasero 2017). This paper proposes a model capable of learning the spatial and temporal correlations of air pollution measured through both remote sensing satellite imagery and ground-based sensors.

In order to do so, we employ a two-stage model to combine the learned representations of the numerous features that we use to predict spatiotemporal particulate matter 2.5 (PM2.5), or particulate matter pollutants with a diameter of less than 2.5 μm, in various areas of Los Angeles county over time. The first stage of our model utilizes the cutting-edge, highly accurate, and effective Graph Convolutional Network (GCN) to learn and predict patterns in meteorological and spatial correlations in our ground-based sensor data.

The Graph Convolutional Network is an advanced deep learning architecture utilizing the properties of graphs. Graphs prove to be a valuable and effective method of modeling air pollution and weather forecasting, as many of the methods of collecting and recording the values of these features are in the form of ground-based sensors. Thus, we can model these sensors as nodes in a weighted directed graph to preserve the spatial and distance-based correlations among sensors. The goal of the Graph Convolutional Network is to learn the feature embeddings and patterns of nodes and edges in a graph. The GCN learns the features of an input graph $G(V, E)$ typically expressed with an adjacency matrix A as well as a feature vector x_i for every node i in the graph expressed in a matrix of size $V \times D$ where V is the number of vertices in the graph and D is the number of input features for each vertex. The output of the GCN is an $V \times F$ matrix where F is the number of output features for each vertex. We can then construct a deep neural network with an initial layer embedding of $h_v^0 = x_i$ to perform convolution neighborhoods of nodes, similar to a Convolutional Neural Network (CNN). Then, the k-th layer of the neural network's embedding on vertices h_v^k is

$$h_v^k = \sigma\left(W_k \sum_{u \in N(v) \cup v} \frac{h_v^{k-1}}{\sqrt{|N(u)||N(v)|}}\right), \forall k > 0,$$

where σ is some non-linear activation function, h_v^{k-1} is the previous layer embedding of v, W_k is a transformation matrix for self and neighbor embeddings, and $\sum_{u \in N(v)} \frac{h_u^{k-1}}{|N(v)|}$ is the average of a neighbor's previous layer embeddings. The neural network can be trained efficiently through sparse batch operations on a layer-wise propagation rule

$$H^{(k+1)} = \sigma(D^{-\frac{1}{2}} \tilde{A} D^{-\frac{1}{2}} H^{(k)} W_k),$$

where I is the identity matrix, $\tilde{A} = A + I$, and D is the diagonal node degree matrix defined as $D_{ii} = \sum_j A_{i,j}$ (Kipf and Welling 2016). In this way, the GCN is able to train the neural network to output a graph with output feature vectors for each node in the graph. In our implementation, we extend the GCN's capabilities further by providing a feature matrix constructed of feature vectors for each edge in the graph such that the GCN outputs a graph with an output feature matrix for all nodes and edges in the graph.

Our second stage of the model utilizes a highly accurate and effective deep learning architecture that learns and predicts for data considering both spatial and temporal correlations. We utilize the cutting-edge Convolutional Long-Short Term Memory (ConvLSTM) model architecture to predict spatiotemporal air pollution using input data of remote-sensing satellite imagery, ground-based sensor data, and the output of the GCN model. The ConvLSTM model is a variant of the traditional Long Short-Term Memory (LSTM) model, a time-series Recurrent Neural Network.

Traditional LSTM models rely on a single-dimensional input vector parameterized by time. The structure of the LSTM model relies on a time series of gates and cells that retain and propagate information from previous cells and time from the model. For a traditional FC-LSTM (Fully Connected Long Short-Term Memory), the time parameterized input gates i_t, forget gates f_t, cell states c_t, output gates o_t, and hidden gates h_t are defined as

$$\begin{aligned}
i_t &= \sigma(W_i x_t + W_i h_{t-1} + W_i \circ c_{t-1} + b_i) \\
f_t &= \sigma(W_f x_t + W_f h_{t-1} + W_f \circ c_{t-1} + b_f) \\
c_t &= f_t \circ c_{t-1} + i_t \circ \tanh(W_x x_t + W_h h_{t-1} + b_c) \\
o_t &= \sigma(W_x x_t + W_h x_{h-1} + W_c \circ c_t + b_o) \\
h_t &= o_t \circ \tanh(c_t),
\end{aligned}$$

where W denotes the weight matrix and $\circ$ denotes the Hadamard matrix multiplication product (Hochreiter and Schmidhuber 1997). In a traditional FC-LSTM, both the inputs and outputs are 1-dimensional time-series vectors. As a result, LSTM models do not allow for or utilize spatial correlations in data.

The ConvLSTM model improves upon the FC-LSTM by applying convolution within the cells and gates of the LSTM to allow for multidimensional video-like inputs and outputs. This can be achieved rather simply by replacing the Hadamard products used to define the key equations for the FC-LSTM with the convolution operation. Note that

there are two methods to induce convolution in a traditional LSTM model. The key equations for the ConvLSTM are

$$
\begin{aligned}
i_t &= \sigma(W_i x_t + W_i h_{t-1} + W_i * c_{t-1} + b_i) \\
f_t &= \sigma(W_f x_t + W_f h_{t-1} + W_f * c_{t-1} + b_f) \\
c_t &= f_t * c_{t-1} + i_t * \tanh(W_x x_t + W_h h_{t-1} + b_c) \\
o_t &= \sigma(W_x x_t + W_h x_{h-1} + W_c * c_t + b_o) \\
h_t &= o_t * \tanh(c_t),
\end{aligned}
$$

where $*$ denotes the convolution operation (Shi et al. 2015). One such method is described as the ConvLSTM model by using the convolution operation within the cells and gates of the LSTM, thus directly allowing the inputs and outputs of the ConvLSTM to be time-series multidimensional data. Another method of inducing convolution is to perform convolution prior to the LSTM model. By modularizing the convolution operation and training a CNN to transform video-like inputs to 1-dimensional time-parameterized output vectors and using the output in a traditional FC-LSTM, we can achieve a similar level of learning and prediction based on spatial and temporal correlations. Recent research into this approach has resulted in the model denoted the Convolutional Neural Network - Long Short-Term Memory (CNN-LSTM), which, as the name suggests, utilizes a CNN and LSTM run in succession to utilize and predict video-like inputs. In this paper, we perform spatiotemporal prediction using the ConvLSTM model; however, there is prior research on alternatively utilizing the CNN-LSTM model to predict spatiotemporal air pollution (Li et al. 2020a, b; Yan et al. 2021; Guo et al. 2019).

Methodology

In this paper, we propose a two-stage model capable of learning spatiotemporal trends based on remote-sensing satellite imagery of air pollution and data of ground-based sensors monitoring air pollutants and meteorological features. We find that including meteorological features are essential to an accurate prediction of ambient air pollution. Air pollutants are closely correlated to meteorological data. Liu et al. (2020) found that of the 896 government-monitored air pollution sensors in China, 675 ground-based sensors reported an increase in carbon monoxide (CO), sulfur dioxide (SO_2), nitrogen dioxide (NO_2), and PM2.5 when there was a greater than 10% increase in wind speed at the same location. In addition to including meteorological data, we find that including a mixture of both remote-sensing satellite imagery of air pollution and ground-based sensor air pollution data is necessary for a robust and multifaceted approach to spatiotemporal air pollution prediction. Remote-sensing satellite imagery provides information on atmospheric air pollution, while ground-based sensors provide finer-grained information on air pollution at sea level or within cities. Since the level of air pollution may vary greatly with respect to altitude, we utilize both remote-sensing satellite imagery and ground-based sensor data as input to our model in order to fully understand and predict air pollution. Finally, we find that data from other air pollutants prove to be beneficial when predicting for a particular pollutant—in our case PM2.5. In our dataset, we utilize remote-sensing satellite imagery of nitrogen dioxide as an input feature when predicting for PM2.5. Nitrogen dioxide is an adverse air pollutant that is highly correlated to PM2.5 since a large portion of ambient PM2.5 is generated through the chemical reactions of atmospheric nitrogen dioxide (Brook 2008).

Model architecture

We propose a two-stage model to learn and predict spatiotemporal PM2.5 using meteorological data, ground-based sensor data, and remote-sensing satellite imagery. The goals of our approach include learning spatial correlations of meteorological data through the GCN architecture, utilizing both spatial and temporal correlations in satellite and remote-sensing data through the ConvLSTM model, and combining the GCN and ConvLSTM models sequentially.

The first stage of our model utilizes the GCN architecture to learn patterns of meteorological data through a graph representation. To do so, we first construct a weighted directed graph representation with the meteorological data described in "Implementation." The goal of the GCN architecture is to interpolate and predict a denser meteorological graph than the input graph. The task of interpolation is inherently an effective task to obtain high-level learned feature embeddings. By utilizing a GCN for spatial interpolation, we can train a model to predict meteorological trends in areas not provided by the input graph; and thus, we can later use these learned correlations as input to construct a video-like sequence of spatially continuous predicted meteorological features over time in a geographical area. For our model, we adapted previous work by Wu et al. (2020) on spatiotemporal kriging with Graph Convolutional Networks to interpolate our nodes and edges of the meteorological graph. We train the GCN for this interpolation task by systematically "hiding" a small percentage of node and edge attributes. The model then learns to predict for the hidden meteorological feature values at these nodes and edges based on the disparity between the predicted hidden meteorological features for a time period against the ground truth meteorological features. Once the training is complete, the GCN is capable of interpolating a sparse meteorological graph into a dense graph containing various meteorological features. The GCN

will create one such dense meteorological graph for each sample parameterized by time.

An intermediate step in our model is to convert the dense meteorological graph into an image-based format and concatenate many time-series samples into a video-like input to the ConvLSTM model. We utilize a pre-built model from the StellarGraph Python library to collect the high-level embeddings into an image-based format for the ConvLSTM model. The StellarGraph package allows for an unsupervised learning graph representation learning approach to create a matrix of high-level weights corresponding to the representations of nodes and edges in the meteorological graph. This set of weights is bounded by the geographic area we have defined, and as a result, the high-level embedding weight array is calculated for each time step of the meteorological dataset. By converting the dense meteorological graphs into spatiotemporal embeddings of video-like input, we can pass the learned meteorological information as input to the second stage of our model.

The second stage of our model utilizes the ConvLSTM architecture to predict spatiotemporal PM2.5. The inputs to the ConvLSTM model are all video-like in shape: all input data is in the form of sets of images or arrays parameterized over time. The inputs to the ConvLSTM model are the learned meteorological information outputs from the first stage of the model, the remote-sensing satellite imagery of air pollutants, and the ground-based sensor data of air pollutants. The output of the ConvLSTM model is a set of predicted ground-based PM2.5 sensor values around Los Angeles county for multiple days in the future. Figure 1 displays a visualization of our model architecture.

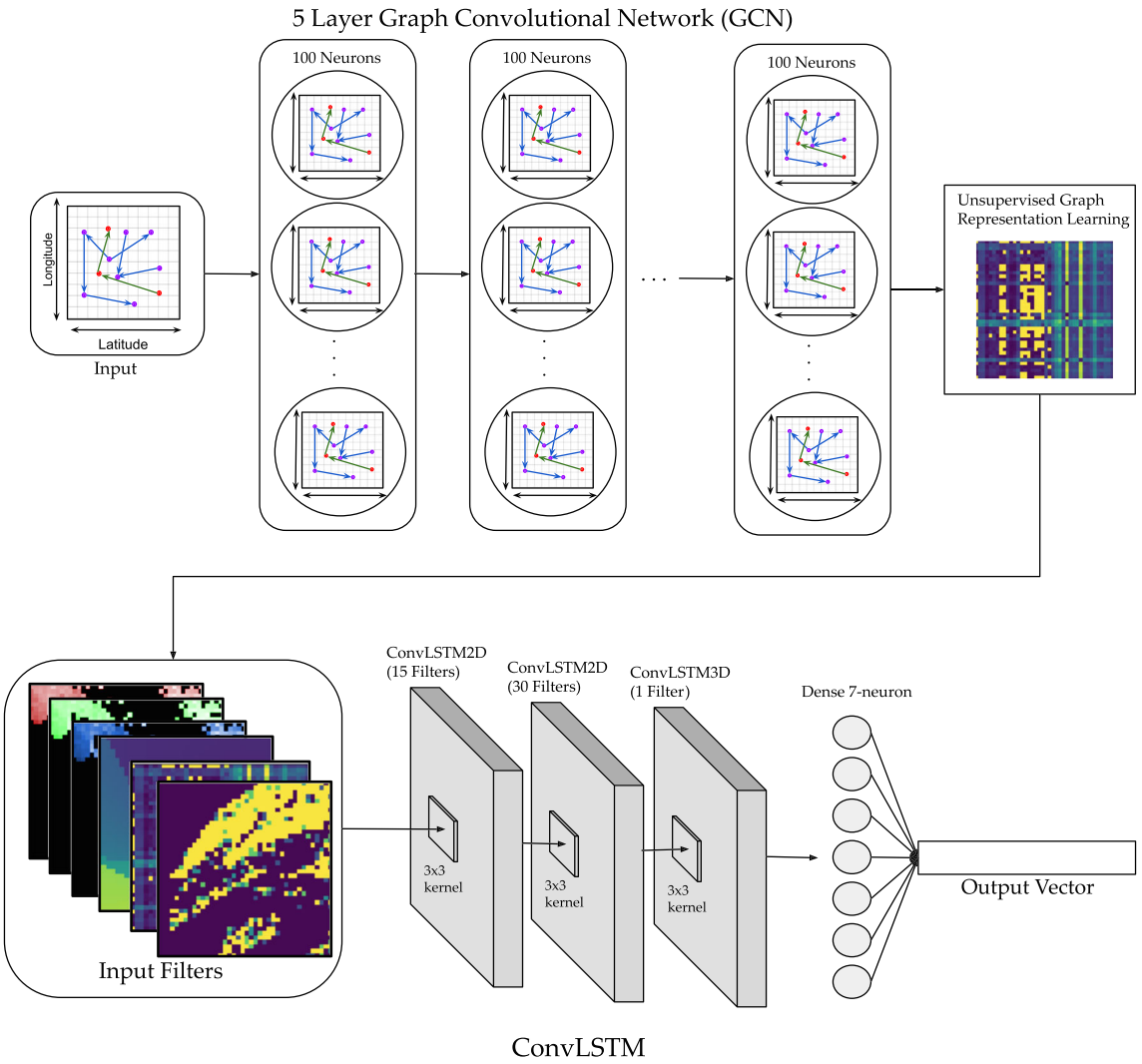

Fig. 1 Model architecture

Dataset

Our geographical area of interest for prediction is the greater Los Angeles county. For all data sources in our dataset, we select a region of roughly 2500 mi^2, or a 50 mile × 50 mile square region of northwest Los Angeles county. For remote-sensing satellite imagery in our dataset, we crop the satellite images to fit the geographic boundaries we defined. For the ground-based sensors, we use the data from all sensors within the latitude and longitude range of our geographic boundary.

Our temporal area of interest for prediction is the roughly 5 years worth of data from August 3, 2015, to March 19, 2020. Each sample of our dataset is selected to be 46 h apart from each other. This 46-h frequency is chosen based on the longest temporal frequency of all data sources from our model, and we find that the remote-sensing satellite imagery of nitrogen dioxide in our area of interest is produced every 46 h. However, some of the other data sources including the ground-based sensor data is recorded hourly, but in order to normalize our dataset, we select a time frequency of 46 h between samples for all data sources in our dataset. For each of our data sources, we collect 882 samples corresponding to the 1642 days of data from August 3, 2015, to March 19, 2020. Note that due to our input data's temporal frequency being out of daily cycle, we utilize data from various hours of the day including nighttime and daytime imagery. In collecting all remote-sensing satellite imagery for our deep learning model, we carefully consider the physical effects of sunlight and other temporal confounders on our imaged pollutant data. Thus, we selected data from sources where the imagery was ensured to be isolated from the physical effects of sunlight such that the differences between imagery temporally spaced apart are uniquely the true differences in pollutant concentrations.

Our meteorological data was collected from the Iowa State University Environmental Mesonet database (Todey et al. 2002). The Environmental Mesonet database collects and records hourly Meteorological Aerodrome (METAR) Reports from Automated Surface Observing Systems (ASOS) located near various airports and municipal airstrips within the continental United States. ASOS is primarily used by airlines and air traffic controllers to monitor meteorological features near and around airport runways. The METAR data provides a full hourly report of 17 ground-level meteorological features including wind speed, wind direction, relative humidity, dew point, precipitation, Air Quality Index (AQI), air pressure, air temperature, etc. The complete list of meteorological features collected from each site is presented in Table 1. Within our geographic boundaries, there are 24 ASOS sensors providing full METAR reports. In order to use these meteorological features in combination with the model and create a meteorological graph structure, we needed to normalize the various units of these meteorological features. We did this by calculating each data point's percentile value. The percentile value is calculated daily and essentially is the current hour's raw value divided by the metric's maximum daily value. In this way, we normalize the units so that we retain the important meteorological information, but we do not need the domain-specific units it is associated with. Figure 2 describes the geographical area of interest and site locations for the raw meteorological features we collected.

Table 1 METAR meteorological features for each of the 24 ASOS sites within Los Angeles county collected from Mesonet

Meteorological feature	Unit	Stationary/non-stationary
Air temperature	F	Stationary
Dew point	F	Stationary
Relative humidity	%	Stationary
Heat index/wind chill	F	Stationary
Wind direction	∘	Non-Stationary
Wind speed	mph	Non-stationary
Altimeter	in	Stationary
Sea level pressure	mb	Stationary
1-h precipitation	in	Stationary
Visibility	mi	Stationary
Wind gust	mph	Stationary
AQI	N/A	Stationary
Peak wind gust	mph	Non-Stationary
Peak wind direction	∘	Non-Stationary
Cloud height level 1	ft	Stationary
Cloud height level 2	ft	Stationary
Cloud height level 3	ft	Stationary

Our ground-based air pollution dataset was collected from the Southern California Air Resources Board AQMIS2 API. We collect ground-based sensor data on PM2.5 which is our prediction target. For the geographic range we have defined, there are seven PM2.5 sensors collecting hourly data in the following locations: Lancaster, Santa Clarita, Reseda, Glendora, Los Angeles - Main St, Long Beach, and Long Beach - Rt 710. These seven PM2.5 sensors are the only government-maintained PM2.5 sensors in the geographical bounds; however, there are various low-cost individually maintained sensors we chose not to use for evaluation of our model as we are unable to estimate the uncertainty error of such sensors.

Our remote-sensing satellite imagery was collected from the NASA Multi-Angle Implementation of Atmospheric Correction (MAIAC) algorithm data source (Lyapustin and Wang 2007). The MAIAC algorithm is a preprocessing algorithm performed on imagery collected by the NASA Moderate Resolution Imaging Spectroradiometer (MODIS) instrument onboard the NASA Terra and Aqua satellites.

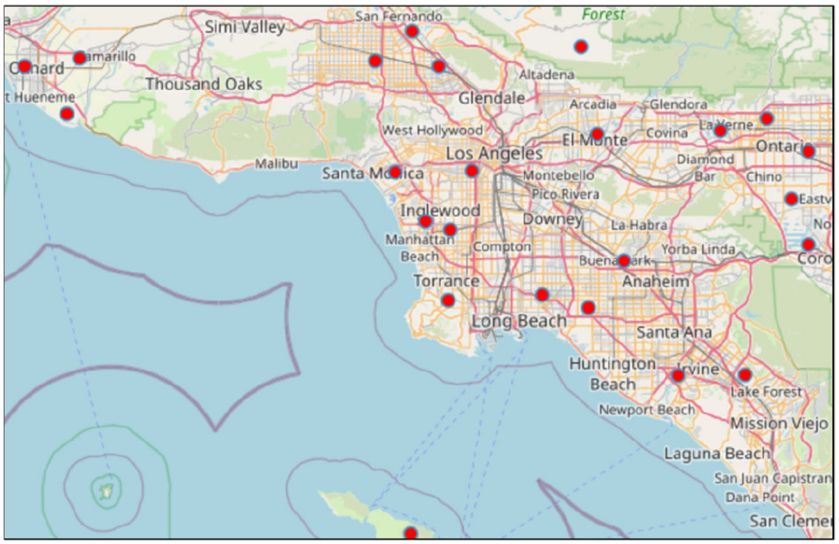

Fig. 2 METAR ASOS observations of Mesonet database: 24 sensor locations in Los Angeles, where each sensor records 17 meteorological attributes hourly

The Terra and Aqua satellites orbit the Earth every 1–2 days and provide imagery over 36 spectral bands utilizing the MODIS imaging instrument. The MAIAC algorithm is a highly advanced preprocessing algorithm that converts raw MODIS imagery to data analytics-ready images by retrieving atmospheric aerosol and air pollutant data from MODIS images, normalizing pixel values, and removing cloud cover masks. For our model, we use the MAIAC MODIS/Terra+Aqua Daily AOD dataset. AOD or Aerosol Optical Depth is a measure of the direct amount of sunlight being blocked by atmospheric aerosols and air pollutants. AOD is perhaps the most comprehensive measure of ambient air pollution and years of research has shown a strong correlation between AOD readings and PM2.5 concentrations in both atmospheric and ground-level settings (Li et al. 2015; Xiao et al. 2017). The MAIAC MODIS AOD dataset we utilize as input to our model records the blue-band Aerosol Optical Depth at a central wavelength of 0.47 µm. The raw MAIAC MODIS AOD dataset provides a spatial resolution of 1-km/pixel for an area of 1200km × 1200km. However, for our implementation, we crop the imagery in order to fit our defined geographic bounds.

Figure 3 describes a sample image of NASA MAIAC AOD data for our desired geographic bounds. Note that the figure provides a visualization of the raw grid-like data of the MAIAC AOD imagery; and thus, the color values of the visualization correspond to AOD values, not raw RGB imagery. The brighter colored pixels in the visualization correspond to higher AOD values. Figure 3 visualizes the downsampled 40 pixel × 40 pixel MAIAC AOD imagery, as shown in the axis labels along the visualization.

We also utilized remote-sensing satellite imagery on nitrogen dioxide (NO_2) data from the U.S. Geological Survey's (USGS) Earth Explorer database (Faundeen et al. 2002). The Earth Explorer database collects remote-sensing satellite imagery from the European Space Agency's Sentinel-2 satellite. The Sentinel-2 satellite was launched in March 2015 to image and record terrain and atmospheric data using 13 spectral bands along a 290-km orbital swath. Our model utilizes data imaging NO_2 on one such imaging band with a central wavelength of 945.1 nm. Due to the orbital swath of the Sentinel-2 satellite, the images are collected with a temporal frequency of 46 h. Since this is the largest temporal frequency of all our data sources, we normalized all other data sources and set the temporal frequency of our total dataset to 46 h. The nitrogen dioxide

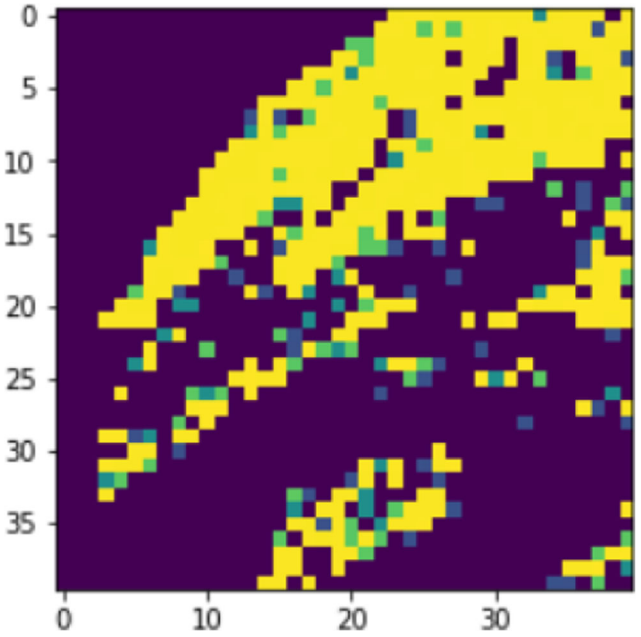

Fig. 3 Sample MAIAC Satellite AOD Imagery (April 29, 2019, Los Angeles AOD NASA MAIAC Imagery)

in the satellite imagery is shown as the light blue cloud-like structures. It is important to note that the Sentinel-2 satellite imagery product of NO_2 that we use layers the NO_2 structures at the top of the image such that the background imagery of the terrain and ocean will not interfere with the imagery of the NO_2 structures. The goal of our data preprocessing for this imagery is then to isolate the NO_2 layer from the remainder of the image through pixel masking. Figure 4 provides a sample raw NO_2 imagery visualization collected from Sentinel-2 for our desired geographic bounds.

For the remote-sensing satellite imagery of the nitrogen dioxide data, we remove the pixel colors from the image that do not represent the nitrogen dioxide structures. These pixel values include the pixels for the terrain and ocean, as including them as input to the model will likely introduce noise, thus reducing the accuracy of the model. To isolate the light-blue structures that represent the nitrogen dioxide imagery, we apply a pixel mask to select pixels within the light blue color range. We set the RGB values of all other non light-blue color values to (0,0,0).

Implementation

In order to use the meteorological data with our model architecture, we must create a weighted directed graph bounded by a geographical grid of our specified area of interest using the meteorological features. For each time step of the meteorological dataset, we create a weighted directed graph denoting the nodes of the graph as static meteorological features pertaining to a sensor location and the edges denoting non-static meteorological features. We define static features as scalar measurements of individual meteorological features at a sensor location. For example, the node attributes for our meteorological graph include relative humidity, AQI, temperature, air pressure, dew point, heat index, etc. Edge attributes consist of non-static meteorological features that rely on or connect multiple sensors. For example, the edge attributes consist of the physical distance in miles from meteorological sensor locations, the wind speed, and the wind direction. For each time step, we can create a multidimensional weighted directed graph containing the spatial and distance-based information of all meteorological sensors and their recorded features. We then repeat this process to create these multidimensional weighted directed graphs for each time step of 46 h in the dataset. Figure 5 describes the weighted directed meteorological graph construction process. Algorithm 1 describes a step-by-step procedure of creating these weighted directed meteorological graphs for a single time step.

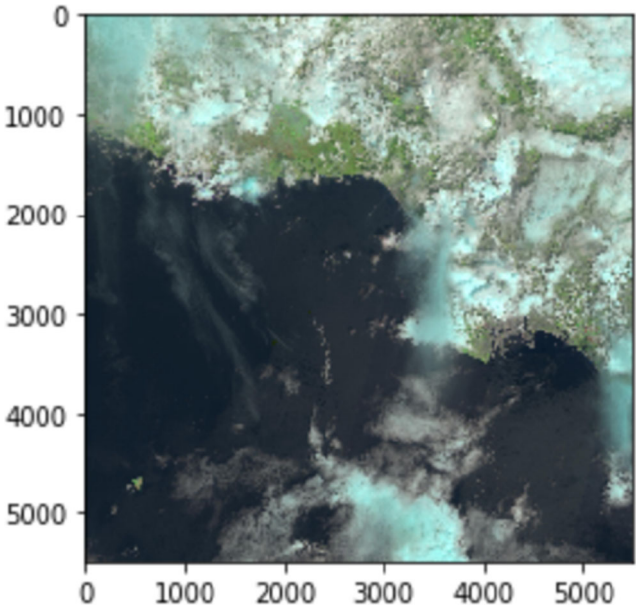

Fig. 4 Sample raw data (Source: USGS EarthExplorer database of satellite imagery of Los Angeles taken on April 29, 2019, by ESA's Sentinel-2 satellite)

Algorithm 1 Meteorological graph construction.

Input: Meteorological site features $f_i \in F$, where each f_i contains site coordinates x_i, y_i and a set of site-specific static $s_i \in S$ and non-static $n_i \in N$ feature values. Boundary latitude values $\text{lat}_{\max}$, $\text{lat}_{\min}$. Boundary longitude values $\text{long}_{\max}$, $\text{long}_{\min}$.

Initialize 40x40 array grid A.
Initialize weighted directed graph $G = (V, E)$
for $f_i \in F$ **do**
  $\text{grid}_x, \text{grid}_y = \left\lfloor \frac{x_i \cdot 40}{\text{long}_{\max} - \text{long}_{\min}} \right\rfloor, \left\lfloor \frac{y_i \cdot 40}{\text{lat}_{\max} - \text{lat}_{\min}} \right\rfloor$
  $A[\text{grid}_x][\text{grid}_y]$ = vector of site-specific static values s_i
  Set $A[\text{grid}_x][\text{grid}_y]$ as vertex of G
end for
for $f_i \in F$ **do**
  for $n_i \in N$ **do**
    Let start_x, start_y be the starting coordinates of a weighted directed edge in G
    $\text{start}_x, \text{start}_y = \text{grid}_x, \text{grid}_y$
    Recover end_x, end_y from site-specific non-static value n_i.
    Create weighted directed edge in G starting from vertex located at $(\text{start}_x, \text{start}_y)$ and ending at vertex located at $(\text{end}_x, \text{end}_y)$ with weight of $|n_i|$.
  end for
end for

Output: Geographically-bound graph feature matrix grid A, Weighted Directed Graph G

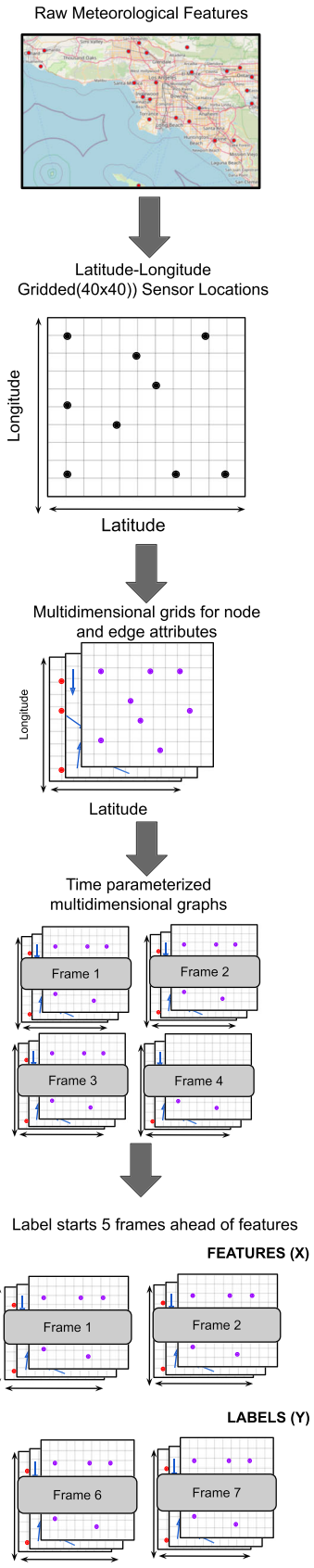

Fig. 5 Visualization of meteorological weighted directed graph creation process

We use the Keras ConvLSTM layer to implement our model. This implementation requires the input data to be in the form of a 5-dimensional tensor with dimensions (sample, frame, row, column, filter). For the remote-sensing satellite imagery in our dataset, we set the row, column, and filter dimensions as the 2D image along with the RGB color values as the filter. We downsample the satellite imagery into a 40px × 40px image (or 40 row × 40 column array) for the 5D tensor input. While downsampling, we continue to preserve the geographic boundaries we have defined.

We train the GCN on the multidimensional weighted directed meteorological graphs created by hiding a set of the attributes and training for an interpolation of the hidden values, as described in "Model architecture." We provide a visualization of this interpolation training process in Fig. 6. We visualize two frames of the interpolation training process on the meteorological graph structure for a single static attribute of AQI.

We similarly downsample the output of the GCN-learned meteorological graph representations into a 40 × 40 pixel image. For the ground-based sensor air pollution data, we create a grid bounded by our geographical area of interest and translate the latitude and longitude coordinates to the 40 × 40 grid. For the 33 grid locations that do not contain values, we set to 0, as the data is normalized; and thus, the null value of 0 does not affect predictions. We have now constructed a set of 3D input "images." However, to construct a 5D tensor, we must bundle all input frames over time into many samples. We bundle five consecutive frames into a single sample, where each frame represents information at a time step of 46 h. Each bundle of frames then represents roughly 10 days or 230 h of remote-sensing satellite imagery and ground-based sensor data. Note that the input data bundles are staggered such that for example the first sample consists of data from frames 1–5, the second sample consists of data from frames 2–6, and so on. In this way, we continue to preserve a continuous flow of temporal correlations among samples. By constructing this 5D tensor, we can transform the 880 3D input "images" into a 5D tensor of shape (880, 5, 40, 40, 6). In the 5D tensor, we have a 5D filter where 3 of the dimensions come from the RGB channels from the remote-sensing satellite imagery of the nitrogen dioxide data, 1 of the dimensions come from the RGB channels of the remote-sensing satellite imagery of the MAIAC MODIS AOD data, 1 of the dimensions comes from the output of the meteorological GCN, and 1 of the dimensions comes from the ground-based sensor grid of data values. Figure 7 provides a visualization of these input filters. Note that because we utilize data of the raw values of PM2.5 and NO_2 in $\frac{\mu g}{m^3}$ as well as the AQI values, the ConvLSTM model can derive the raw concentration of PM2.5 and NO_2, as AQI can be directly calculated through a

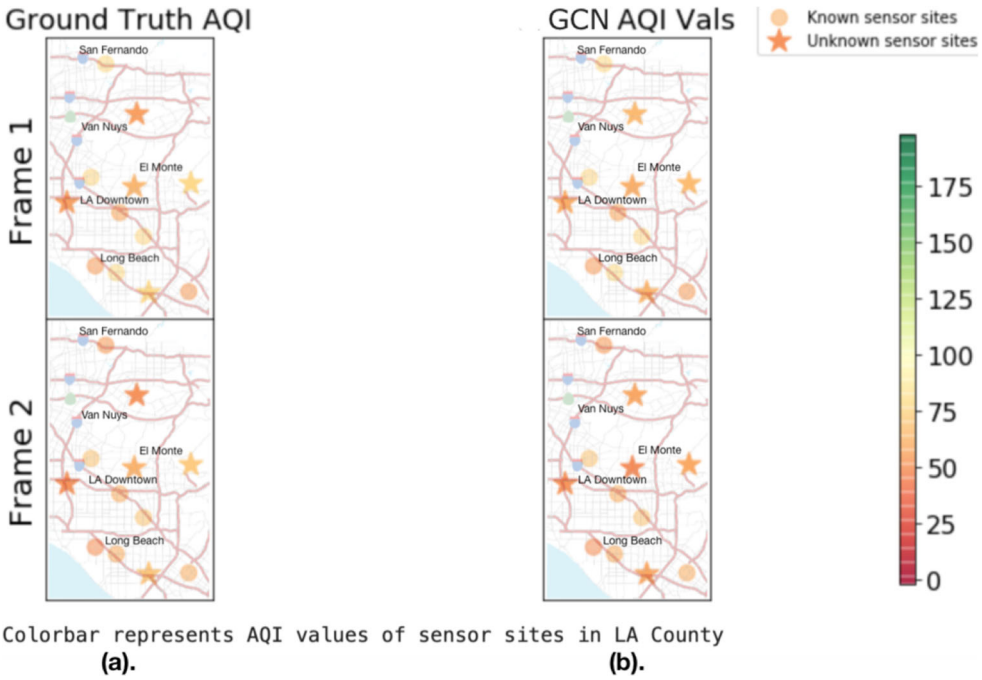

Fig. 6 AQI node attribute training prediction visualization. **a** shows the ground-truth AQI node attribute values over 2 frames separated by 46 h; **b** shows the GCN predicted AQI node attribute values over 2 frames separated by 46 h

linear relationship between the raw value and concentration of an air pollutant.

The output of the ConvLSTM model that uses data of roughly 10 days or a sample of 5 frames in the past will be the predicted PM2.5 values for 5 frames in the future at an interval of 46 h. In order to evaluate and test our model, we added a final Dense Keras layer with 7 neurons to give a prediction of only the 7 PM2.5 sensor locations instead of a spatially continuous prediction of a 40 × 40 grid over Los Angeles county. We have the capability to produce spatially continuous predictions of PM2.5 with our current model, but in order to evaluate against existing ground truth values with little to no measurement error or uncertainty, we restricted the prediction to sensor locations available in the Southern California ARB AQMIS2 API.

Results

Our model predicts spatiotemporal PM2.5 in terms of micrograms per cubic meter ($\frac{\mu g}{m^3}$) at seven sensor locations in Los Angeles county every 46 h at roughly 10 days in the future intervals using meteorological and air pollution data of remote-sensing satellite imagery and ground-level sensors from roughly 10 days in the past. We use 880 days of data from August 3, 2015, to December 3, 2019, as training data and evaluate our prediction on a test dataset of 55 samples or 105 days of data from December 5, 2019, to March 19, 2020. Figure 8 provides a visualization of the distribution and variance of the ground truth PM2.5 values for each sensor location.

To measure the accuracy of our model, we use the Root Mean Square Error (RMSE) and Normalized Root

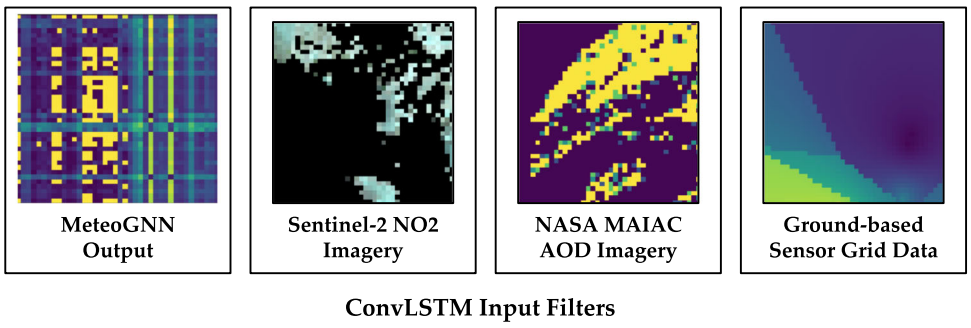

Fig. 7 Visualization of various ground-based sensor data and satellite imagery input filters to our ConvLSTM model

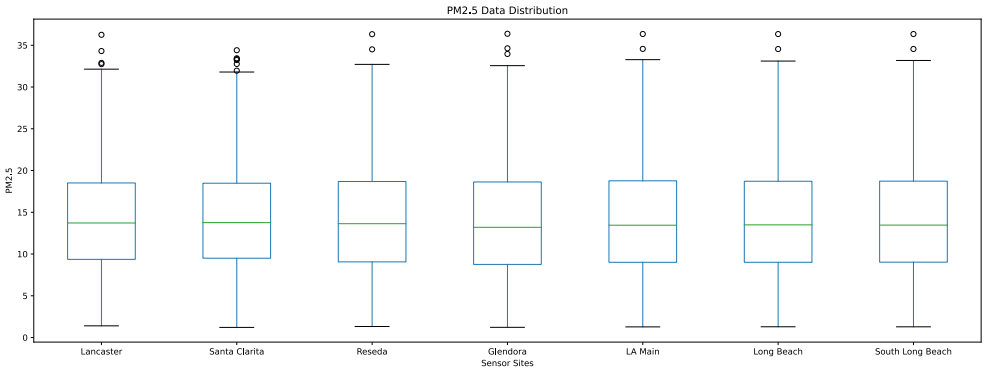

Fig. 8 Data distribution plot of PM2.5 ground-truth sensors in LA County during testing timeframe (Dec 5, 2019–March 19, 2020)

Mean Square Error (NRMSE) error. RMSE and NRMSE is calculated as

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

$$\text{NRMSE} = \frac{\text{RMSE}}{\bar{y}}$$

where n is the number of observations, $\hat{y}$ is the predicted value, y is the ground truth, and $\bar{y}$ is the mean of the test data.

Table 2 displays the prediction RMSE and NRMSE metric results on the first five frames or roughly 10 days of the test set.

Table 3 displays the prediction RMSE and NRMSE metric results for the first frame average and fifth frame average for each sensor location throughout the test set. Note that first frame average error denotes the average error of the immediate next frame predicted using the previous five frames, while the fifth frame average error denotes the average error of the fifth of five frames using ten frames earlier than the fifth frame. Since the first frame predictions use more recent data to predict, the average first frame error is significantly lower than the average fifth frame error.

We provide a visualization of our predicted raw PM2.5 values against the ground truth for each sensor location in Fig. 9.

Our results show significant improvement over current state-of-the-art models on predicting spatiotemporal PM2.5 air pollution using both remote-sensing satellite imagery and ground-based sensor data. Our first frame prediction's percent accuracy is 91.24% which is a 30.1% decrease in hourly error from Shi et al. (2015), one of the earliest and highest cited implementations of the ConvLSTM model for PM2.5 prediction. Moreover, our results show an 85% decrease in the first frame error compared to our previous models using solely the ConvLSTM model on Sentinel-2 satellite imagery (Muthukumar et al. 2020a, b, 2020c, 2021; Cocom et al. 2020; Nagrecha et al. 2020). The averaged RMSE and NRMSE decrease over time with later frames, but this is expected as the nature of PM2.5 results in concentrations 5 days in the future being more correlated to 5 days in the past as compared to concentrations 10 days in the future. We also describe the trends of our predicted PM2.5 values against the ground-truth PM2.5 values and the testing set mean for a single sensor location, Lancaster, over the test set in Fig. 10. Similarly, we provide a visualization for the Santa Clarita site in Fig. 11. Note that in a practical interpretation of our predictions, we can expect our model to predict trends in Los Angeles PM2.5 values within 46 h

Table 2 RMSE and NRMSE error values in terms of parts per million (ppm) for first 5 frames of test set or roughly 10 days of data (December 5, 2019–December 11, 2019)

Metric	Frame	Value
RMSE	1 (46 h ahead)	0.000751
	2 (92 h ahead)	0.000938
	3 (138 h ahead)	0.001223
	4 (184 h ahead)	0.001759
	5 (230 h ahead)	0.002823
NRMSE	1 (46 h ahead)	0.0876
	2 (92 h ahead)	0.1402
	3 (138 h ahead)	0.1608
	4 (184 h ahead)	0.2103
	5 (230 h ahead)	0.2510

Table 3 RMSE and NRMSE error values in terms of parts per million (ppm) averaged over 5 frame bundles (first frame averages and fifth frame averages) of test set for each sensor location

Metric	Sensor location	Average value	
		1st frame	5th frame
RMSE	Lancaster	0.001451	0.003932
	Glendora	0.001233	0.003841
	Santa Clarita	0.001028	0.003405
	Reseda	0.001115	0.003639
	LA - Main St	0.000834	0.003213
	Long Beach	0.000750	0.003069
	Long Beach - RT 710	0.000901	0.003118
NRMSE	Lancaster	0.1148	0.2866
	Glendora	0.1065	0.2705
	Santa Clarita	0.0890	0.2519
	Reseda	0.0907	0.2666
	LA - Main St	0.0647	0.2409
	Long Beach	0.0541	0.2261
v	Long Beach - RT 710	0.0702	0.2370

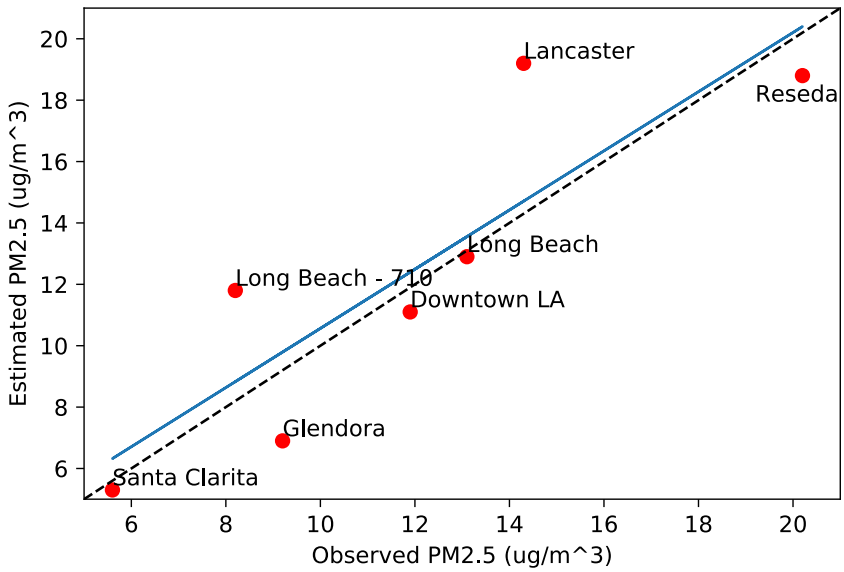

Fig. 9 Predicted vs actual average raw PM2.5 values for each sensor location

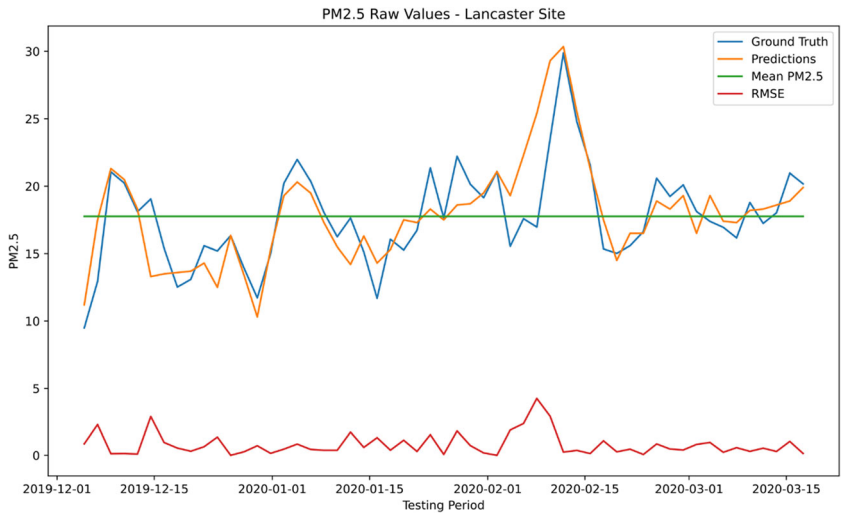

Fig. 10 Predicted vs actual RMSE plot of raw PM2.5 during the testing timeframe (Dec 5, 2019–March 19, 2020) for the Lancaster sensor

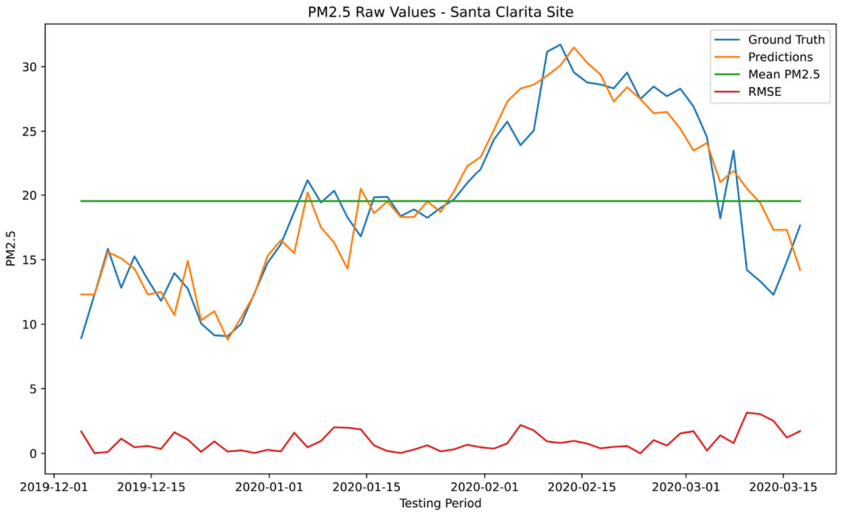

Fig. 11 Predicted vs actual RMSE plot of raw PM2.5 during the testing timeframe (Dec 5, 2019–March 19, 2020) for the Santa Clarita sensor

prior or after its true occurrence and with a predicted value $\pm$ 3 to 5 $\frac{\mu g}{m^3}$ of the true value.

Conclusion

In this paper, we use complex deep learning models to accurately predict spatiotemporal PM2.5 in Los Angeles county over time in 46-h temporal frequencies using meteorological and air pollution remote-sensing satellite imagery and ground-based sensor data. In designing our model, we include information on spatial and temporal correlations as well as meteorological features and related air pollutant matter data to understand, learn, and predict spatiotemporal PM2.5 air pollution.

We utilized various state-of-the-art predictive models including the Graph Convolutional Network (GCN) and the Convolutional Long Short-Term Memory (ConvLSTM). We created a time parameterized set of multidimensional weighted directed graphs to represent 17 meteorological features in 24 sensor locations within the greater Los Angeles county area. We then utilized the GCN architecture to perform convolution on neighborhoods of nodes in order to interpolate dense meteorological graphs using spatiotemporal kriging. We also used unsupervised graph representation learning algorithms to create high-level embedding "images" of the dense meteorological graphs and used these high-level embeddings as input to the ConvLSTM model. In addition to the outputs from the GCN, we also supplied government-monitored ground-based PM2.5 sensor data in grid form, NASA MODIS MAIAC AOD remote-sensing satellite imagery, and ESA Sentinel-2 nitrogen dioxide remote-sensing satellite imagery as input to the ConvLSTM. We then bundled the input data into samples consisting of five consecutive frames of data or roughly 10 days of data. We calculate the RMSE and NRMSE error values of the predicted PM2.5 values over the first 5 frames as well as the averaged RMSE and NRMSE error values of the predicted sample for each sensor location. We find that our results show significant improvement upon current research in the field utilizing spatiotemporal deep predictive algorithms.

We also find that the results of our model can match to explain various real-world events, chemical processes, and physical processes of ground-based PM2.5 in Los Angeles county. For example, as described in Figs. 10 and 11, we can see a significant and drastic drop in the predicted PM2.5 values across all site locations in Los Angeles county around the beginning of March 2020, which corresponds to the advancement of the worldwide COVID-19 pandemic and the start of the stay-at-home lockdown issued within Los Angeles county.

This work can be used to inform and assist researchers in various disciplines on the movement of PM2.5 along temporal and spatial coordinates.

Future work

In the future, we hope to calculate and account for the data fusion under uncertainty error for ground-based sensor measurements to ensure the validity of recorded values. Doing this will allow us include low-cost individually maintained ground-level sensor data as inputs and predictive targets in order to increase the spatial resolution of predictions. We hope to include additional meteorological features such as insolation and solar irradiance as these features are key to the photochemical production of atmospheric aerosols. We also hope to include wildfire and smoke data as features to our model, as various studies have found a significant correlation between wildfires and rising air pollution levels (Liu et al. 2016; Reid et al. 2015).

This research can also extend further than Los Angeles county and predict an array of pollutants including carbon monoxide, ozone, and sulfur dioxide. We hope to utilize community-maintained site monitoring stations in order to collect fine-grained concentration data of additional air pollutants including NO_2, SO_2, CO, and O_3.

Funding This research is supported by NASA and the City of Los Angeles through the Predicting What We Breathe research project.

Availability of data and material The authors will make data sources available as supplementary material soon.

Code availability The authors will make code available as supplementary material soon.

Declarations

Conflict of interest The authors declare no competing interests.

References

Abrahamsen EB (2018) Ole Magnus Brastein, and Bernt Lie Machine Learning in Python for Weather Forecast based on Freely Available Weather Data

Bellinger C, Jabbar MSM, Zaïane O, Osornio-Vargas A (2017) A Systematic Review of Data Mining and Machine Learning for Air Pollution Epidemiology. BMC Publ Health 17(1):1–19

Brook RD (2008) Cardiovascular Effects of Air Pollution. Clin Sci 115(6):175–187

Cocom E, Muthukumar P, Holm J, Comer D, Lyons A, Burga I, Hasenkopf CA, Calvert C, Pourhomayoun M (2020) Particulate Matter Forecasting in Los Angeles County with Sparse Ground-based Sensor Data Analytics. In: AGU Fall Meeting Abstracts, vol 2020, p A061–0004

Faundeen JL, Kanengieter RL, Buswell MD (2002) US geological survey spatial data access. J Geospatial Eng 4(2):145–145

Grover A, Kapoor A, Horvitz E (2015) A Deep Hybrid Model for Weather Forecasting. In: Proceedings of the 21th ACM, SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 379–386

Guo G, Guo W, Chen C-H, Wang X, Liu G (2019) The air quality prediction based on a convolutional LSTM network. In: International Conference on Web Information Systems and Applications. Springer, pp 98–109

Hochreiter S, Schmidhuber J (1997) Long Short-term Memory. Neural Comput 9(8):1735–1780

Kipf ThomasN, Welling Max (2016) Semi-supervised classification with graph convolutional networks. arXiv:1609.02907

Li J, Carlson BE, Lacis AA (2015) How well do satellite AOD observations represent the spatial and temporal variability of PM2.5 concentration for the United States? Atmosp Environ 102:260–273

Li S, Xie G, Ren J, Guo L, Yang Y, Xu X (2020a) Urban pm2. 5 concentration prediction via attention-based CNN–LSTM. Appl Sci 10(6):1953

Li T, Hua M, Wu X (2020b) A hybrid CNN-LSTM model for forecasting particulate matter (PM2.5). IEEE Access 8:26933–26940

Liu JC, Mickley LJ, Sulprizio MP, Dominici F, Xu Y, Ebisu K, Anderson GB, Khan RFA, Bravo MA, Bell ML (2016) Particulate air pollution from wildfires in the western US under climate change. Clim Change 138(3):655–666

Liu Y, Zhou Y, Lu J (2020) Exploring the Relationship Between Air Pollution and Meteorological Conditions in China under Environmental Governance. Sci Rep 10(1):1–11

Lyapustin A, Wang Y (2007) MAIAC-multi-angle implementation of atmospheric correction for MODIS. In: AGU Spring Meeting Abstracts, vol 2007, pp A51B–05

Marchal V, Dellink R, Van Vuuren D, Clapp C, Chateau J, Magné B, Van Vliet J (2011) OECD Environmental Outlook to 2050. Organ Econ Co-oper Dev 8:397–413

Muthukumar P, Cocom E, Holm J, Comer D, Lyons A, Burga I, Hasenkopf C, Pourhomayoun M (2020a) Real-time Spatiotemporal Air Pollution Prediction with Deep Convolutional LSTM through Satellite Image Analysis. In: 16th International Conference on Data Science (ICDATA '20). Springer Nature, pp 317–328

Muthukumar P, Cocom E, Holm J, Comer D, Lyons A, Burga I, Hasenkopf C, Pourhomayoun M (2020b) Real-time Spatiotemporal NO2 Air Pollution Prediction with Deep Convolutional LSTM through Satellite Image Analytics. In: AGU, Fall Meeting Abstracts, vol 2020

Muthukumar P, Cocom E, Nagrecha K, Holm J, Comer D, Lyons A, Burga I, Calvert CF, Pourhomayoun M (2020c) Satellite Image Atmospheric Air Pollution Prediction through Meteorological Graph Convolutional Network with Deep Convolutional LSTM. In: 7th Annual Conference on Computational Science and Computational Intelligence (CSCI-ISAI '20). IEEE CPS

Muthukumar P, Nagrecha K, Cocom E, Comer D, Burga I, Taub J, Calvert C, Holm J, Pourhomayoun M (2021) Predicting PM2.5 Air Pollution using Deep Learning with Multisource Satellite and Ground-based Observations and Meteorological and Wildfire Big Data. In: AGU, Fall Meeting Abstracts, vol 2021

Nagrecha K, Muthukumar P, Cocom E, Holm J, Comer D, Burga I, Pourhomayoun M (2020) Sensor-Based Air Pollution Prediction using Deep CNN-LSTM. In: 2020 International Conference on Computational Science and Computational Intelligence (CSCI). IEEE, pp 694–696

Narejo S, Pasero E (2017) Meteonowcasting using Deep Learning Architecture. (IJACSA) Int JAdv Comput Sci Appl 8(8)

Reid CE, Jerrett M, Petersen ML, Pfister GG, Morefield PE, Tager IB, Raffuse SM, Balmes JR (2015) Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. Environ Sci Technol 49(6):3887–3896

Shi X, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-c (2015) Convolutional LSTM Network: A machine learning approach for precipitation nowcasting. arXiv:1506.04214

Todey DP, Herzmann DE, Takle ES (2002) The iowa environmental mesonetcombining observing systems into a single network. In: Sixth Symposium on Integrated Observing Systems
s

Weyn JA, Durran DR, Caruana R (2020) Improving Data-Driven Global Weather Prediction using Deep Convolutional Neural Networks on a Cubed Sphere. J Adv Model Earth Syst 12(9):e2020MS002109

WHO (2018) Air Pollution and Child Health: Prescribing Clean Air: Summary. Technical report, World Health Organization

WorldBank (2016) The Cost of Air Pollution: Strengthening the Economic Case for Action. Washington: World Bank Group

Wu Y, Zhuang D, Labbe A, Sun L (2020) Inductive Graph Neural Networks for Spatiotemporal Kriging. arXiv:2006.07527

Xiao Q, Wang Y, Chang HH, Meng X, Geng G, Lyapustin A, Liu Y (2017) Full-coverage high-resolution daily PM2. 5 estimation using MAIAC AOD in the Yangtze River Delta of China. Remote Sens Environ 199:437–446

Yan R, Liao J, Yang J, Sun W, Nong M, Li F (2021) Multi-hour and multi-site air quality index forecasting in beijing using CNN, LSTM, CNN-LSTM, and spatiotemporal clustering. Expert Syst Appl 169:114513

Affiliations

Pratyush Muthukumar[1] · Emmanuel Cocom[1] · Kabir Nagrecha[1] · Dawn Comer[2] · Irene Burga[2] · Jeremy Taub[3] · Chisato Fukuda Calvert[3] · Jeanne Holm[2] · Mohammad Pourhomayoun[1]

Emmanuel Cocom
ecocom@calstatela.edu

Kabir Nagrecha
knagrec2@calstatela.edu

Dawn Comer
dawn.comer@lacity.org

Irene Burga
irene.burga@lacity.org

Jeremy Taub
jeremy@openaq.org

Chisato Fukuda Calvert
chisato@openaq.org

Jeanne Holm
jeanne.holm@lacity.org

Mohammad Pourhomayoun
mpourho@calstatela.edu

1 Department of Computer Science, California State University Los Angeles, Los Angeles, CA, USA

2 City of Los Angeles, Los Angeles, CA, USA

3 OpenAQ, Washington, DC, USA