Science of the Total Environment 860 (2023) 160446

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Predicting hourly $PM_{2.5}$ concentrations in wildfire-prone areas using a SpatioTemporal Transformer model

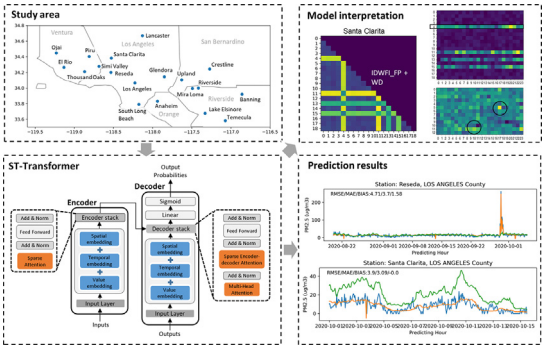
Manzhu Yu [a,*], Arif Masrur [b], Christopher Blaszczak-Boxe [c,d]

[a] Department of Geography, The Pennsylvania State University, United States of America
[b] Environmental Systems Research Institute, United States of America
[c] Department of Geosciences, The Pennsylvania State University, United States of America
[d] Department of Interdisciplinary Studies, Howard University, United States of America

HIGHLIGHTS

- Conduct time series forecasting on $PM_{2.5}$ concentrations at 19 EPA's AQS station locations in the greater Los Angeles area
- ST-Transformer model achieved an average RMSE of 6.92 ($\pm$2.93) μg m^{-3} and outperformed benchmark models.
- Sparse attention attends only to the useful and relevant information, preventing overestimation of $PM_{2.5}$ concentration.
- ST-Transformer can predict an abrupt, significant increase in $PM_{2.5}$ concentration during wildfires.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: Jianmin Chen

Keywords:
Wildfire
Air pollution
$PM_{2.5}$
Spatiotemporal prediction
Sparse self-attention
Transformer neural network

ABSTRACT

Globally, wildfires are becoming more frequent and destructive, generating a significant amount of smoke that can transport thousands of miles. Therefore, improving air pollution forecasts from wildfires is essential and informing citizens of more frequent, accurate, and interpretable updates related to localized air pollution events. This research proposes a multi-head attention-based deep learning architecture, SpatioTemporal (ST)-Transformer, to improve spatiotemporal predictions of $PM_{2.5}$ concentrations in wildfire-prone areas. The ST-Transformer model employed a sparse attention mechanism that concentrates on the most useful contextual information across spatial, temporal, and variable-wise dimensions. The model includes critical driving factors of $PM_{2.5}$ concentrations as predicting factors, including wildfire perimeter and intensity, meteorological factors, road traffic, $PM_{2.5}$, and temporal indicators from the past 24 h. The model is trained to conduct time series forecasting on $PM_{2.5}$ concentrations at EPA's air quality stations in the greater Los Angeles area. Prediction results were compared with other existing time series forecasting methods and exhibited better performance, especially in capturing abrupt changes or spikes in $PM_{2.5}$ concentrations during wildfire situations. The attention matrix learned by the proposed model enabled interpretation of the complex spatial, temporal, and variable-wise dependencies, indicating that the model can differentiate between wildfires and non-wildfires. The ST-Transformer model's accurate predictability and interpretation capacity can help effectively monitor and predict the impacts of wildfire smoke and be applicable to other complex spatiotemporal prediction problems.

* Corresponding author.
E-mail address: mqy5198@psu.edu (M. Yu).

http://dx.doi.org/10.1016/j.scitotenv.2022.160446
Received 18 August 2022; Received in revised form 18 November 2022; Accepted 19 November 2022
Available online 25 November 2022

1. Introduction

Climate change has aggravated heatwaves and droughts worldwide, resulting in wildfires increasing in frequency, size, and intensity and longer smoke seasons (Mazdiyasni and AghaKouchak, 2015a, b; Schiermeier, 2018a, b; Woodward et al., 2014). Climate predictions show that significant fire events, including mega-forest fires, will continue and become less predictable (Natole et al., 2021a, b). Smoke from wildfires is more harmful to human health than other forms of air pollution (Aguilera et al., 2021a, b) because it contains toxic particles, such as soot, and chemicals, such as carbon monoxide. One of the main concerns is $PM_{2.5}$, which can penetrate the lungs and circulatory system, often causing adverse health effects (Crouse et al., 2015; Makar et al., 2017). Acknowledging that wildfire activities may likely continue to rise, it is an urgent research priority to accurately predict air pollutant concentration induced by wildfire smoke, especially in wildfire-prone areas.

Air quality predictions for fire-prone areas can significantly help emergency managers and public health officials mitigate potentially adverse environmental and public health impacts. However, accurate prediction of air pollutant concentration, especially those that are wildfire-sourced, is challenging as such scenarios are highly related to wildfire characteristics, such as the state of the atmosphere, topography, fuel, and moisture (Jaffe et al., 2020). First, wildfire smoke is highly dynamic in space and time, resulting in air quality measurements drastically changing in spatiotemporal dimensions (Khaykin et al., 2020). Regulatory air quality sensors typically have long distances between stations so that fire smoke atmospheric dynamics and chemistry and environmental and public health impacts can be missed by sparsely distributed stations (Lu et al., 2021). Given limited air quality measurement platforms, it is often difficult to grasp the full impact of a specific fire event over a geographic area by relying on a single source. Second, it is challenging to accurately predict the location of the smoke in the air column – whether it is floating along the surface or aloft – owing to dynamic perimeters and behaviors of wildfires producing smoke and local weather conditions (e.g., wind speed and direction, etc.) (Liu et al., 2019b). There is an urgent need to incorporate dynamic fire behavior and the complex multi-dimensional (horizontal, vertical, and time) and multivariate (e.g., meteorological conditions, land use, and terrain) interactions for smoke emission into the prediction methods.

Traditionally, numerical models for smoke forecasts have been used to produce short-term air quality predictions in wildfire-prone areas. The increased resolution of numerical weather prediction and computational power recently opened new avenues for developing integrated systems in a more coupled way. WRF-SFIRE-CHEM couples fire progression, plume rise, smoke dispersion, and chemical transformations (Kochanski et al., 2016). High-Resolution Rapid Refresh-Smoke (HRRR-Smoke) (Ahmadov et al., 2017) leverages WRF-CHEM and satellite-derived fire radiative power to simulate biomass burning emissions, plume rise, and smoke transport but does not couple fire progression. Wildland-urban interface Fire Dynamics Simulator (WFDS) (Mell et al., 2007, 2009; Mueller et al., 2014) and FIRETEC (Linn et al., 2002, 2005; Pimont et al., 2011) resolve combustion and small-scale plume dynamics but without atmospheric chemistry. Community Multiscale Air Quality (CMAQ) model represents chemical transport models that focus on chemical smoke transformations and rely on external parameterizations for plume height and emission computations (Appel et al., 2017); concomitantly, CMAQ does not resolve fire-atmosphere interactions and plume dynamics of fire progression. Satellite and ground-observed smoke emissions have been utilized to evaluate these physical models, help reduce model bias, optimize smoke exposure estimates, and study the spatiotemporal variability of wildfire smoke (Mallia et al., 2020; Zou et al., 2019). However, several improvements for fire and smoke models were suggested (Liu et al., 2019a), including 1) parameterization of lateral fire spread; 2) profiles of vertical plume concentrations; and 3) estimates or measurements of near-event and downwind O_3, $PM_{2.5}$, their precursors, and intermediate chemical species (e.g., organic matter, SO_4^{2-}, mineral dust, elemental carbon, etc.). However, uncertainty exists with varying physical schemes and model configurations, and these numerical models are not readily available and not interpreted in a way that is helpful for the public.

Emerging trends leverage data-driven methods (e.g., time-series prediction, space-time interpolation, convolutional neural network (CNN), and deep fully convolutional neural network (FCNN)) based on historical air pollutant concentrations and meteorological conditions to predict air quality impacts by smoke emissions. CNN and deep FCNN have been applied to detect smoke plumes from high-resolution geostationary satellite imagery (Larsen et al., 2021; Ramasubramanian et al., 2019), with the limitation that satellite imagery can be skewed by cloud cover, terrain, and light and heat emitted by factories (Brey et al., 2018; Buysse et al., 2019). Research has also been conducted to fill in the gaps between ground-level sensors and provide a map distribution of air quality using data-driven methods (Cheng et al., 2018; Lin et al., 2018; Yi et al., 2018). However, these models might suffer from overgeneralizing the temporal trends of air pollutant concentrations, especially regarding the air quality being abruptly impacted by a nearby fire, leading to significant underestimations of air pollutant concentrations during wildfire situations. There is also a lack of thorough examination of different impacting variables regarding smoke emissions from fires, the relationship between air pollutants within the smoke, and how they drift in space.

The challenges expressed above will be addressed to provide more accurate predictions and examinations of the multi-dimensional and multivariate interactions that influence $PM_{2.5}$ concentration. This research, therefore, proposes a novel deep learning framework that learns the spatiotemporal trends of wildfire, smoke, and air pollutants by considering their spatial, temporal, and variable-wise dependencies. Specifically, this research aims to provide improved hourly predictions of $PM_{2.5}$ concentration by learning through the spatiotemporal dependencies among the input variables, including fire and smoke observations, meteorological conditions, and traffic conditions. Section 2 will review the literature on spatiotemporal prediction using machine learning and data-driven predictive methods for air pollution from wildfires. Section 3 will describe the problem statement and background of proposing a novel deep learning framework for spatiotemporal air quality prediction, especially in wildfire-prone areas. Section 4 will introduce the proposed model in detail. Section 5 will describe the data used for the experiments, comparative results, and model interpretation, followed by discussions and conclusions.

2. Literature review

2.1. Machine learning for spatiotemporal prediction

Spatiotemporal prediction is challenging due to complex variable heterogeneity, non-stationarity, spatial and long-range temporal dependencies (Li and Moura, 2020). Spatial dependency is complex; for example, air pollutant concentrations at a monitoring station may correlate more with far away stations than closer ones. Temporal dependencies might be long-range; for instance, measurements at a certain hour may be correlated with ones at closer time points but may also be correlated with the measurements from the same hour a day prior. Spatiotemporal predictions may rely on supporting relevant factors that are generally heterogeneous, e.g., wildfire progression influenced by wind speed and directions, and might be non-stationary due to expected or unexpected incidents, such as air quality impacted by fireworks or wildfires. These challenges hinder accurate forecasting when using traditional statistical time series forecasting methods, such as persistence and autoregressive integrated moving average (ARIMA) (Lim and Zohren, 2021). More recently, deep learning approaches for time series forecasting generally rely on a sequence-to-sequence framework that can map a context of the recent past to a target in the future, such as recurrent neural networks (RNNs), graph neural networks (GNNs), and attention mechanisms (Graves, 2012; Li et al., 2019; Wu et al., 2020; Zhou et al., 2021).

RNNs consist of feed-forward neural networks in which the hidden nodes are connected in series, but they are limited by the autoregressive problem that causes difficulty in interpreting long-term patterns (Wu

et al., 2020). Using RNNs to predict time series at multiple locations for the same targeting variable (such as air temperature, air quality, and precipitation intensity), the models are trained with input predictors at multiple locations (Han et al., 2021; Yu et al., 2021). Training RNNs in this way enables models to sufficiently capture temporal patterns for a particular location but lack intrinsic structures that handle spatial patterns and potentially the spatiotemporal intercorrelated dependencies between nearby locations.

GNNs can address this issue by explicitly representing the spatial structure using the adjacency matrix, which can be constructed prior to training using the distance between node pairs (Yu et al., 2018). In GNNs, information associated with a particular location is embedded in the nodes and passed to neighboring nodes along the edges, so spatial and spatiotemporal dependencies are learnable during training. One potential limitation is that the extracted spatial dependencies are generally local (between neighboring nodes) and might not be capable of predicting scenarios with long-range spatial relationships between locations (such as locations that are far away but have high semantic similarity) (Liu et al., 2019a, b, c). In addition, these spatial dependencies are modeled with a fixed Laplacian matrix that might not capture the dynamics when the targeting variable has a changing spatial dependency (Li et al., 2018). Instead of constructing the adjacency matrix prior to training, a Multi-Task Graph Neural Network Framework (MTGNN) was proposed to learn the adjacency matrix before applying the convolutions in the model structure (Wu et al., 2020). MTGNN improves the capability of GNN in learning dynamic spatial dependencies and has the potential of capturing variable spatiotemporal dependencies but still considers spatial and temporal dependencies as separate features to learn.

Recent studies leverage multi-head attention mechanisms to explicitly or implicitly capture the spatial and temporal dependencies for spatiotemporal predictions (Zhou et al., 2021; Li and Moura, 2020; Grigsby et al., 2021). In multi-head attention mechanisms, each time series is compared with all previous ones to propagate information over long sequences, and important information is retained through self-attentions (Vaswani et al., 2017). Based on this original work of using a multi-head attention mechanism in an encoder-decoder architecture called Transformer, Zhou et al. (2021) developed the Informer with improved temporal embeddings that learn the non-stationary and long-range temporal dependencies. Although Informer acknowledges the value of multiple variables per timestep as a single input token, the model only focuses on learning “temporal attention” among timesteps and ignores the different spatial relationships between variables. To explicitly capture the joint spatial and temporal dependencies in attention-based mechanisms, Li and Moura (2020) developed a graph transformer to account for the dynamic strengths of spatial dependency using a sparse Transformer, where the edges between less-dependent nodes are trimmed before training. To capture long-range temporal and complex spatial dependencies, Grigsby et al. (2021) developed the Spacetimeformer to flatten multivariate time series into extended sequences to separate the influences of different input predictors. However, the model will only consider spatial and temporal dependencies without variable-wise dependencies, because it uses one variable at each station as both the input and target variable instead of taking other relevant variables as predictors. In many cases of spatiotemporal prediction, the future conditions of a targeting variable depend on multiple relevant factors, which have non-linear and complex dependencies in spatial, temporal, and variable dimensions.

2.2. Machine learning for spatiotemporal prediction of air pollution related to wildfires

Studies have incorporated various observations or reanalysis datasets for wildfire occurrence, behavior, and smoke dispersion to account for wildfire impacts on air quality. Reid et al. (2015) utilized a generalized boosting model to select an optimal prediction model for daily $PM_{2.5}$ concentration predictions during the 2008 northern California wildfires from a set of 11 statistical algorithms and 29 predictor variables. They leveraged the simulation output of chemical transport models, satellite observation of aerosol optical depth (AOD), distance to nearest fires, meteorological conditions, land-use, traffic, and spatial and temporal characteristics. They concluded that the most important predictors were satellite retrieved AOD, simulation output, and distance to the nearest fire cluster. Yao et al. (2018) utilized a random forest model to estimate hourly $PM_{2.5}$ concentration at 5-km resolution during wildfire seasons from 2010 to 2015 in British Columbia, Canada. A similar set of predictors are used, including ecozone classification, fire activity, meteorology, and elevation. Two case studies were chosen during wildfire seasons in certain areas within the study region to demonstrate the capability of the proposed model. Sensitivity analysis was conducted to explore the sufficiency of predictors, leading to the conclusion that $PM_{2.5}$ concentrations from 24 h prior and the satellite retrieved AOD values, contributed little to the model.

More recently, deep learning approaches and emerging datasets, such as ensemble-based deep learning and low-cost sensor networks, have been leveraged in the spatiotemporal prediction of air quality impacted by wildfire smoke. Li et al. (2020) utilized ensemble-based deep learning to estimate $PM_{2.5}$ concentrations in California at a 1 km × 1 km weekly resolution, using satellite AOD retrieval, meteorological conditions, traffic, and wildfire smoke dispersion factors. Using this model to predict a long temporal range from 2008 to 2017, they concluded that integrating multi-source heterogeneous impact factors and the non-linear modeling of deep learning improved the spatiotemporal $PM_{2.5}$ estimation over a regional area. Lu et al. (2021) developed a random forest model that integrates low-cost sensor networks to estimate the hourly $PM_{2.5}$ concentrations at a 500-m resolution in Los Angeles County from 2018 to 2019. They used the meteorological conditions, land-use variables, traffic conditions, and temporal and spatial trends to account for the environmental context. Acknowledging the fact that wildfire smoke may significantly contribute to the improved predictions of $PM_{2.5}$ concentrations in nearby areas, Hung et al. (2021) separated smoke and non-smoke days based on satellite measurements and aerosol reanalysis products to analyze the wildfire smoke's impact on air quality in New York State from 2012 to 2019. They used an artificial neural network to estimate the ground-level $PM_{2.5}$ concentrations at air quality monitoring stations and found that the smoke inflow from fires and the vertical transport mechanisms of smoke generally improved prediction accuracy.

The validation process of the reviewed studies majorly focused on the average prediction errors, but the individual prediction might vary significantly from the observation, especially during abrupt changes in air pollutant concentrations. A challenge to predicting air quality in areas with potential impacts from various sources (including wildfire events) is capturing abrupt changes in air pollutant concentrations. This research will leverage a multi-head spatiotemporal attention mechanism that captures abrupt changes in $PM_{2.5}$ concentrations from wildfire smoke. The attention evolves across spatial and temporal dimensions and adjusts to the variable-wise dependencies over space and time.

3. Problem statement and background

3.1. Spatiotemporal air quality prediction problem

Predicting future air quality values given previous observations from stations within a particular area of interest is of prime importance. The complexity involves four axes: the prediction sequence's duration L_{out}, the previous observations' duration L_{in}, the number of input variables considered at each timestep N_v, and the number of monitoring stations N_s. The prediction model will use the input to predict air quality for future timesteps: $X_{N_s * N_v * L_{in}} \rightarrow Y_{N_s * L_{out}}$. As the numbers along the four axes grow, modeling the spatiotemporal relationships between stations and variables becomes increasingly complex. In the case of air quality monitoring, stations might experience different meteorological patterns due to their geographic location. Measurements from monitoring stations also show complex spatial relationships. For example, far away stations might be impacted by wildfire smoke simultaneously, thus having a short-time high

correlation between air quality measurements, but this correlation will not be sustained (e.g., when wind speed changes or wildfire progresses away from the stations).

3.2. Transformer for time series forecasting

To capture the spatiotemporal relationships between air quality measurements and predictor variables across stations, the foundation of our predictive model is the Transformer and its variant in time-series forecasting. The Transformer (Vaswani et al., 2017) is a sequence-to-sequence prediction model, which includes encoder and decoder layers (Fig. 1) and is widely used for natural language processing tasks.

The encoder consists of an input layer, a positional encoding, and four encoders. The input layer converts the input time series to a vector of dimension d through a fully connected layer, which will be used in the following multi-head attention mechanism. An adjusted layer replaces the original positional encoding with sine functions to match the predictand's diurnal patterns with a tunable period of time steps to repeat the pattern. The adjusted positional encoding encodes the time series data by element-wise addition of the input vector with a positional encoding vector. The resulting vector is fed into four encoder layers. Each encoder layer consists of two sub-layers: a multi-head self-attention sub-layer and a fully connected feed-forward sub-layer.

The multi-head self-attention sub-layer transforms the input vector through linear projections into H distinct query matrices Q, key matrices K, and value matrices V, where H represents the number of heads. Then the scaled dot-product attention computes a sequence of vector outputs for each head using Softmax (Goodfellow et al., 2016a, b) and matrix multiplication. The self-attention is computed as:

$$Attention(\mathrm{Q},\mathrm{K},\mathrm{V}) = softmax\left(\frac{\mathrm{Q}\mathrm{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)V \tag{1}$$

Specifically, the value matrices V is affected by the attention weights $a = softmax\left(\frac{\mathrm{Q}\mathrm{K}^{\mathrm{T}}}{\sqrt{d_{\mathrm{k}}}}\right)$, which is defined by how each value in the input time series (denoted by Q) is influenced by all other values in the time series (denoted by K). The Softmax function is applied to the attention weights a to have a probability distribution between 0 and 1. All weights are then applied to all the values in the value matrices V. The multi-head mechanism allows parallel linear projections of Q, K, and V for H times to have different representations of the input data, so that the model can account for different temporal dependence. The single attention output are further concatenated and projected again to produce a final result.

A residual sum and normalization follow each sub-layer. The encoder produces a d-dimensional vector to feed to the decoder. The decoder also comprises an input layer, a positional encoding layer, four identical decoder layers, and an output layer. The decoder input begins with the last data point of the encoder input, denoted as "Outputs" that are fed to the bottom decoder in the next time step. The decoder input layer maps the decoder input to a d-dimensional vector. Besides the two sub-layers in each encoder, the decoder inserts a third sub-layer (an encoder-decoder attention layer) to apply self-attention mechanisms over the encoder output. A residual sum and normalization follow each sub-layer. Finally, an output fully connected layer maps the output of the last decoder layer to the target time sequence.

The idea behind the Transformer network in machine translation is to allow encoding both short- and long-range correlations between words in the sentence. For that, the multi-head self-attention allows modeling the correlation of elements in sequences regardless of their distance, resulting in an effective global receptive field. Thus, this multi-head attention approach provides a flexible way to capture the complex correlation dynamics for applications with multiple monitoring locations in the case of air quality monitoring. However, while its global receptive field can effectively process the discrete tokens (e.g., words), it can fail to consider the local trend information inherent in continuous data. This failure of capturing local trends will be magnified when the model tries to learn multivariate dependencies across space and time, such as in the case of forecasting air quality, which typically manifests spatiotemporally dynamic distributions, depending on the sources of air pollutants (e.g., fireworks, other anthropogenic emissions, and extreme events, such as wildfires), wind speed, and wind directions. In addition, using the full attention mechanism may also overgeneralize the temporal trends, leading to inaccurate predictions, especially

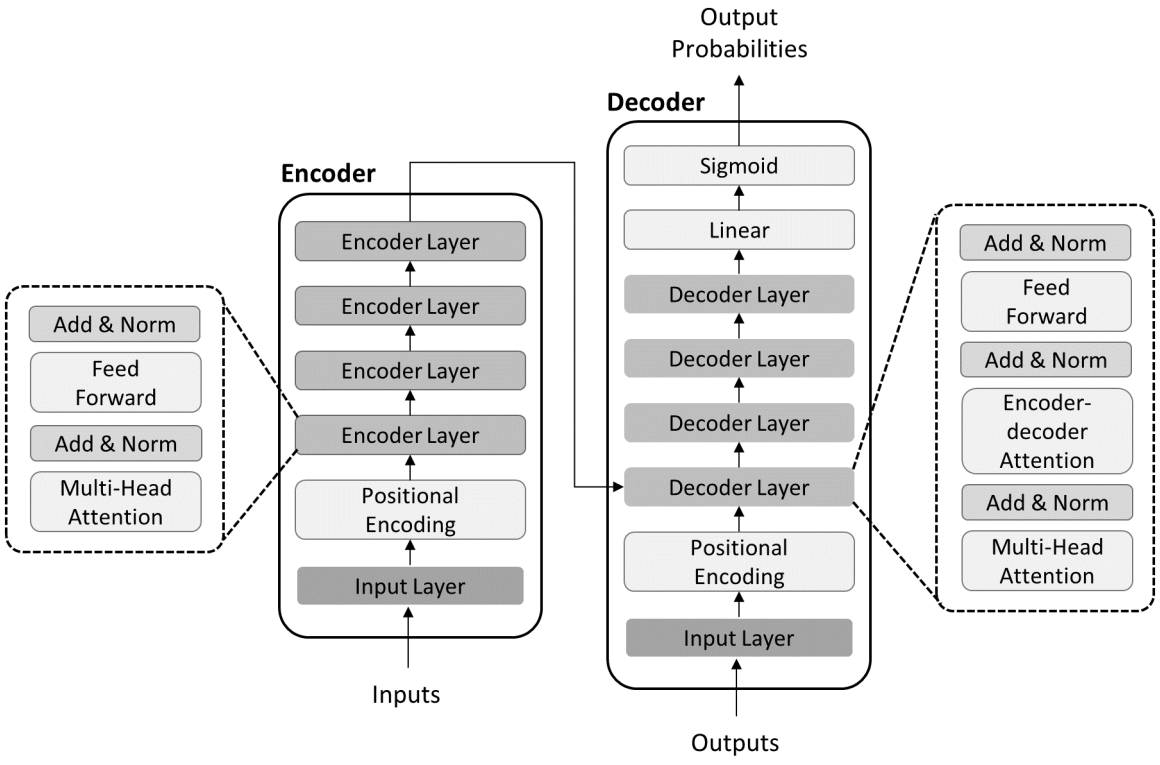


Fig. 1. Architecture of the generic Transformer model.

when there are abrupt changes in the environmental domain. Therefore, applying the generic Transformer model to air quality forecasting is likely to yield unreliable predictions in space and time.

4. SpatioTemporal Transformer

For an accurate times-series forecasting of air quality on heterogeneously distributed monitoring stations, the Transformer network needs to capture the dynamic (and abruptly changing) spatial and temporal correlations among those stations. Instead of using any predefined station distance structure, we utilize the SpatioTemporal Transformer model (Fig. 2), which improves the generic Transformer model to include a spatio-temporally dynamic and sparse dependency module. This enables the model to capture the spatial, temporal, and variable-wise interactions or dependencies. Thus, the SpatioTemporal Transformer replaces the regular multi-head self-attention with a sparse attention mechanism. The spatial and temporal dependencies among stations are embedded as weights in the projection operations on the queries and keys. The dependencies are captured separately by the spatial, temporal, and value embeddings.

4.1. Spatial, temporal, and value embeddings

4.1.1. Time embedding

In the SpatioTemporal Transformer model, the temporal dynamics are entirely modeled by the time embedding of the multi-head self-attention mechanism. Since attention builds the relationship between input and target with the weighted sum function (Eq. 1), the attention mechanism is irrelevant to the order of the input time series. Although temporal dependencies might be complex (short-range or long-range), the order of the input time series does matter – i.e., measurements may be correlated with the ones at closer time points or with the measurements from the same hour a day before. We explicitly embed the input order along with the input time series to capture the complex temporal dependencies, assuming that the model will be trained to learn the complex dependencies according to different situations.

The temporal features are extracted separately for each station from the date and time of the input timestamps: the hour of the day, day of the month, the month of the year, and day of the week. Each temporal feature is first embedded using a positional encoding component provided by the original Transformer paper (Vaswani et al., 2017):

$$\begin{cases} PE(pos, 2i) = \sin\left(pos/10,000^{\frac{2i}{d_{model}}}\right) \\ PE(pos, 2i+1) = \cos\left(\dfrac{pos}{10,000^{\frac{2i}{d_{model}}}}\right), \end{cases} \tag{2}$$

where pos is the position of X_t in the input time series, and i represents the ith dimension in the input vector. The positional encoding converted the order of the timestamp within the input time series as a vector. The resulting PE is then used as the input of a standard embedding layer that maps the integer index of each series to a higher-dimensional representation. The standard embedding layer will serve as a look-up table that stores the embeddings of fixed hours, days, months, and weekdays. The final time embedding Emb_T is generated by flattening the separate temporal features into a vectorized index for each new timestamp.

4.1.2. Spatial embedding

The spatial dynamics are also captured by the multi-head self-attention mechanism to represent the correlation strengths among stations. The spatial embedding Emb_S indicates at which station the time series originate. In addition to having dynamically changing air quality values, each monitoring station is also associated with other varying spatial characteristics, including the local geography of the monitoring station, the meteorological conditions, the nearby wildfire influences, and the emissions from nearby road traffic. Some of these spatial characteristics (e.g., location, elevation, distance to traffic) are invariant over time but vary over space, thereby, warrants accounting for spatial heterogeneity. However, other characteristics, such as air humidity and wind speed, simultaneously vary over space and time, requiring accounting for spatiotemporal variability. Therefore, this spatial embedding Emb_S will capture not only static but also dynamic spatial dependencies.

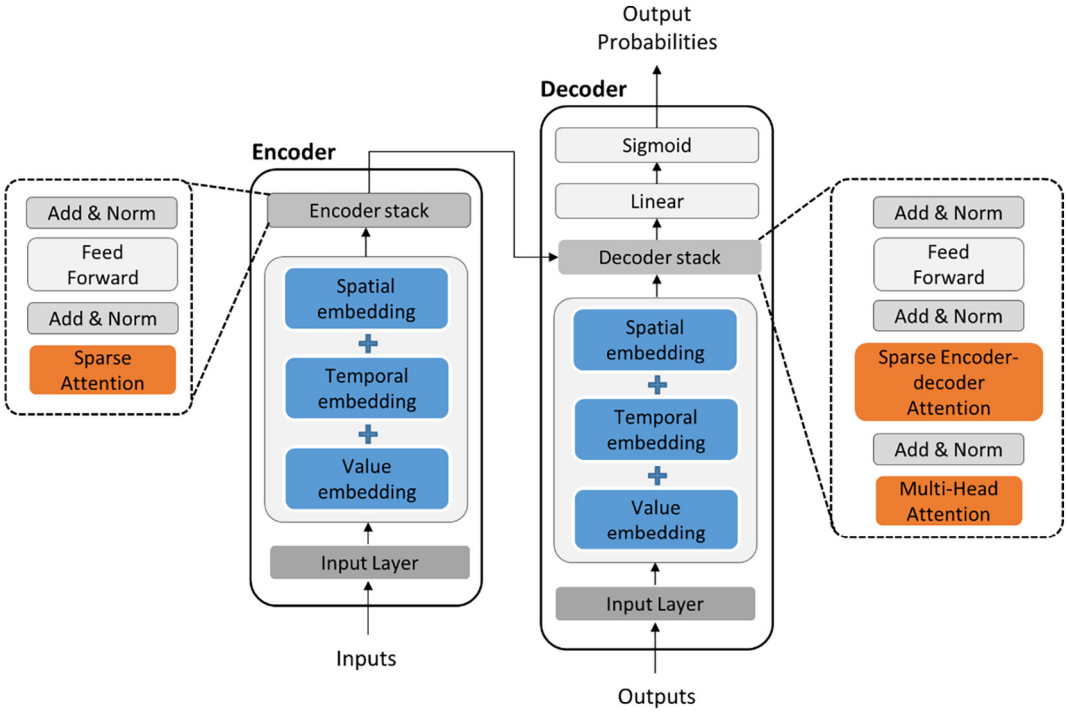


Fig. 2. Architecture of the proposed SpatioTemporal Transformer model.

The spatial embedding will first project each time stamp's input values ($N_s \times N_v$) into high-dimensional latent subspaces, using the feed-forward neural networks:

$$Z_t = w * X_t + b, \tag{3}$$

where Z_t is the high-dimensional representation of the input variables at timestamp t. Based on Z_t, the correlation strength $S_{t,\,ij}$ between station i and station j can be calculated using a cross-correlation function:

$$S_{t,ij} = xcorr\left(Z_{t,i}, Z_{t,j}\right), \tag{4}$$

and the resulting S_t is a $N_s \times N_s$ square matrix represents the correlation strengths of any two stations at a particular timestamp t. The final spatial embedding at a particular timestamp is then calculated as:

$$Emb_S = w * S_t * Z_t + b. \tag{5}$$

4.1.3. Value embedding

To account for the intercorrelations among multiple variables, the value embedding module projects the input time series ($N_s \times N_v \times L_{in}$) into high-dimensional representations $Emb_{V,}$ using the one-dimensional convolutions with a hidden dimension of d_{model}. The value embedding is different from spatial embedding, which separates different timestamps to calculate the correlation between stations and uses the correlations as weights to adjust the high-dimensional representation of input variables from all stations. Instead of separating timestamps, value embedding uses the entire input time series and transforms them into a single high-dimensional representation matrix, so that the complex interdependencies among variables and their dynamics can be highlighted in the model. The spatial, temporal, and value embeddings are then concatenated to represent a comprehensive embedded feature:

$$X_{ST} = Emb_T \oplus Emb_S \oplus Emb_V. \tag{6}$$

4.2. Sparse attention

To prevent overgeneralization of the temporal trends and capture the abrupt changes in air pollutant concentration (and not allow the multi-head attention mechanism to attend to all information), we utilized a sparse attention mechanism to concentrate on relevant information based on the final embedding X_{ST}. The sparse attention enables the focus on a few contextual information through the selection; this follows the idea of ProbAttention in Informer (Zhou et al., 2021) but in a simplified way. The most critical components for the attention are reserved, and other irrelevant information is removed.

In each self-attention head, the embedded feature X_{ST} are projected using feed-forward layers into three key components: namely the query Q_{ST}, the key K_{ST}, and the value V_{ST}, where $Q_{ST} = X_{ST}W^q_{ST}$, $K_{ST} = X_{ST}W^k_{ST}$, and $V_{ST} = X_{ST}W^v_{ST}$. Here, W^q_{ST}, W^k_{ST}, and W^v_{ST} represent the weight matrices for Q_{ST}, K_{ST}, and V_{ST}, respectively. An attention score S is calculated based on the query and key components:

$$S = \frac{Q_{ST}{K_{ST}}^T}{\sqrt{d_{model}}}. \tag{7}$$

The higher attention scores are selected based on a tunable threshold *thr* along each row of the attention score matrix S. This threshold *thr* is initially tuned to be the 80th percentile value. Each position in S can be referred to as S_{ij}, where i and j represents the row and column index of S. If S_{ij} is lower than the thr_i, then S_{ij} is adjusted to be $-\infty$. With the adjusted S, the sparse attention is calculated in the same way as the original Transformer model:

$$Sparse\ Attention = softmax\ (S)V_{ST}. \tag{8}$$

The model can obtain more focused attention related to contexts by combining the spatial, temporal, and value embeddings with the sparse attention mechanism.

In a nutshell, the differences between the proposed model with existing Transformer variants are three-fold. *First*, each monitoring station has the target variable and the impacting factors related to those target variables. Therefore, now, the model can distinguish between timestamps and tell the difference between the variables and stations. *Second*, the sparse connections are no longer along the spatial dimension alone; instead, they support joint connections along spatial, temporal, and variable dimensions. By attending to the most important information across multiple dimensions, the model can improve the predictive accuracy by identifying emerging hotspots among the environmental contexts. *Third*, the sparse attention mechanism can provide physically interpretable spatiotemporal attentions that explain the emerging hotspots learned from the model.

5. Experiments

We present our experimental results on the air quality measurements from 2017 to 2020 from 19 EPA AQS stations in the greater Los Angeles (LA) area. We also collect hourly information for each station's meteorological, wildfire, and traffic conditions. The proposed model predicts the $PM_{2.5}$ concentrations for the next 12 h based on the input of multiple variables of the previous 24 h. We removed the time series with missing values and applied the min-max normalization per variable.

5.1. Data

To evaluate the model performance, we divide the data into the training, validation, and testing set based on date, with the corresponding proportions being 70 %, 15 %, and 15 %, respectively. The predictors include the past 24 h of $PM_{2.5}$ concentrations, meteorological, wildfire-derived, and traffic flow variables. The predictand is the $PM_{2.5}$ concentrations for the next 12 h.

5.1.1. Predictand: $PM_{2.5}$ concentrations from EPA air quality system

Hourly $PM_{2.5}$ concentrations measured at U.S. Environmental Protection Agency (EPA) federal reference monitors (FRM) sites were used in the greater LA area for 2017–2020. $PM_{2.5}$ measurements and station locations were downloaded from the EPA's Air Quality System (AQS) Technology Transfer Network.[1] The $PM_{2.5}$ values from monitoring station data represent fine particulate matter from all sources, including ambient levels and wildfire smoke. $PM_{2.5}$ stations with >25 % missing data were excluded from the analysis, and missing values are imputed based on linear interpolation for each station. Ultimately, 19 $PM_{2.5}$ stations are included in this study (Fig. 3).

5.1.2. Predictors

5.1.2.1. Meteorology. Meteorology data were obtained from ERA-5 (Hersbach et al., 2020), a reanalysis dataset produced by ECMWF that contains hourly estimates of atmospheric, land, and oceanic climate variables. ERA5 combines vast amounts of historical observations into global estimates using advanced modeling and data assimilation systems. The meteorological features include boundary layer height (m), 2 m dewpoint temperature (K), surface pressure (Pa), 2 m temperature (K), 10 m u-component of wind (m s^{-1}), and 10 m v-component of wind (m s^{-1}). The u-component and v-components of wind are further computed into wind direction and speed. All meteorological parameters were obtained at the hourly temporal resolution and 0.25° × 0.25° spatial resolution.

5.1.2.2. Temporal trending variables. We included the "hour of the day" and "day of the year" to account for daily, monthly, and seasonal temporal

[1] http://www.epa.gov/ttn/airs/airsaqs.

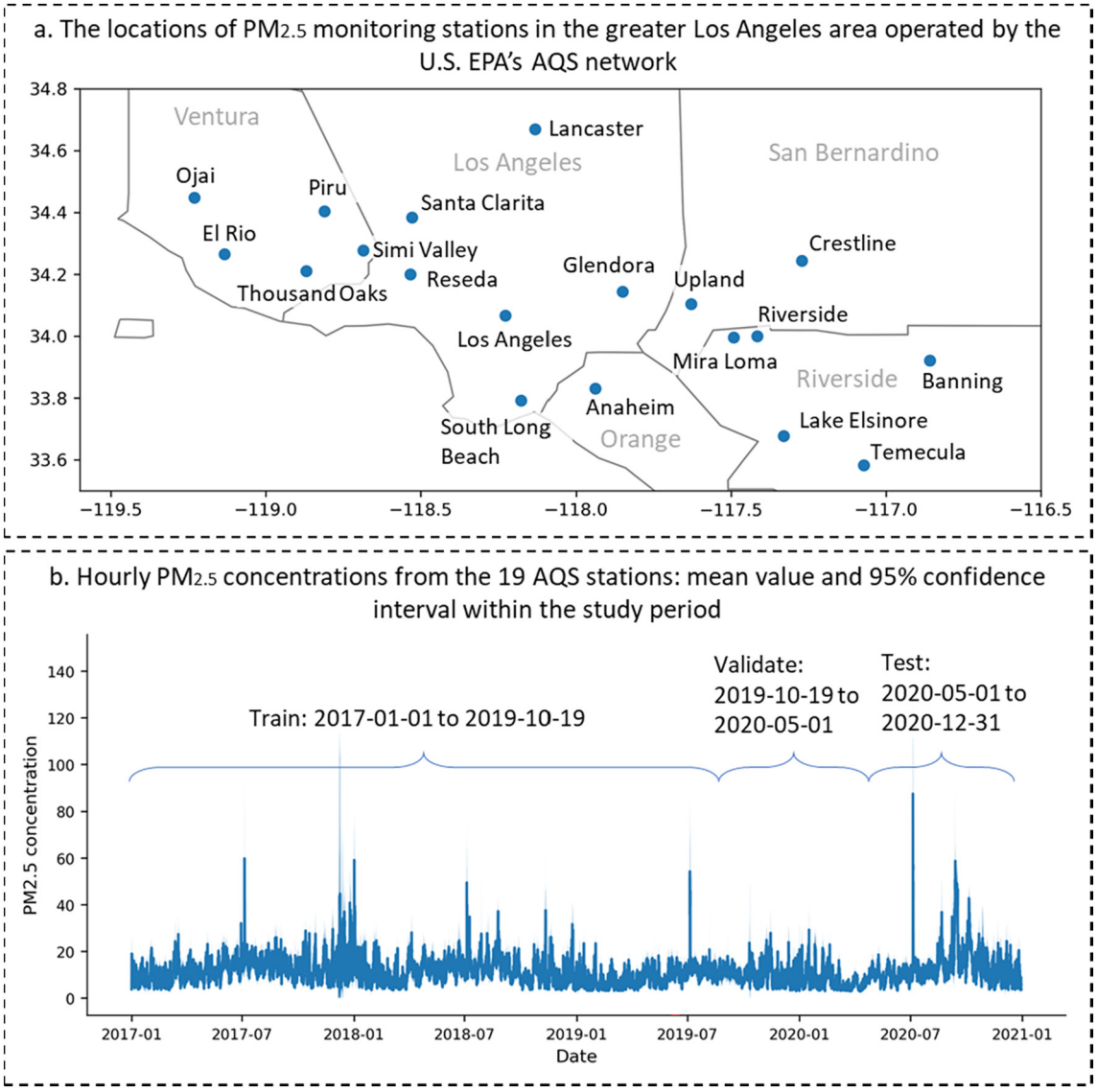


Fig. 3. (a) The locations of $PM_{2.5}$ monitoring stations in the greater Los Angeles area operated by the U.S. EPA's AQS network. (b) Hourly $PM_{2.5}$ concentrations from the 19 AQS stations: mean value and 95 % confidence interval.

variations. In our study period, extremely high $PM_{2.5}$ concentrations (>100 $\mu g/m^3$) were observed on July 4 and 5, primarily due to fireworks.

5.1.2.3. Road traffic. Traffic emissions are considered a significant source of $PM_{2.5}$ (Habre et al., 2021). We downloaded hourly traffic flow data from sensors deployed on the freeway system from the Caltrans Performance Measurement System (PeMS)'s clearinghouse.[2] For each AQS station, we computed the total hourly traffic flow from PeMS sensors within 5 km.

5.1.2.4. Wildfire perimeter. Smoke caused by wildfire is a crucial ambient $PM_{2.5}$ source that contributes to the health burden (for cardiovascular disease, cancer, respiratory diseases, etc.) in many regions, especially in California, where wildfire frequency and intensity have increased over recent decades (Gupta et al., 2018; Reid et al., 2015). The fire perimeter data was downloaded from CAL FIRE.[3] The alarm date, containment date, and the polygon representing the fire perimeter were recorded for each fire event. For each AQS station, at a particular hour, the inverse distance of the nearest fire (when available) is calculated.

5.1.2.5. Wildfire intensity. The fire intensity data was derived from the NASA Fire Information for Resource Management System (FIRMS) Moderate Resolution Imaging Spectroradiometer (MODIS) or Visible Infrared Imaging Radiometer Suite (VIIRS) Fire/Hotspot Data.[4] The fire intensity data is aggregated from the two sources: 1) MODIS Thermal Anomalies, and 2) VIIRS Active Fire and Thermal Anomalies. The data is composed of points or grids with the same resolution as the source satellite (MODIS: 1 km, VIIRS: 375 m), and each grid represents a hot spot of an area with high brightness temperature (BT) at a specific date and time. The high temperature can be considered high fire intensity for the observed date and time.

Two predictors were derived from this data: 1) IDWFI_FP (K) as the inverse distance weighted sum of brightness temperature from grids within any CAL FIRE fire perimeter at the current hour; 2) IDWFI (K) as the inverse distance weighted sum of brightness temperature from all grids available at the current hour. IDWFI is calculated when multiple wildfire hotspots occur at the hour of interest, and a more nearby wildfire hotspot has a higher weight in calculating the overall brightness temperature at a particular AQS station. With n as the number of wildfire hotspots, $dist_i$, $i \in (1, \ldots, n)$

[2] https://pems.dot.ca.gov/?dnode=Clearinghouse.
[3] https://frap.fire.ca.gov/frap-projects/fire-perimeters/.
[4] https://firms.modaps.eosdis.nasa.gov/active_fire/.

as the distance from each hotspot to the AQS station of interest, IDWFI is calculated as:

$$\text{IDWFI} = \Sigma_i^n \frac{dist_i}{\Sigma_1^n dist_i} \times \text{BT}. \quad (9)$$

IDWFI_FP uses the same equation for calculation, but the number of wildfire hotspots is smaller as wildfire hotspots that are not intersected with any fire perimeter provided by CAL FIRE are eliminated. Table 1 provides a full list of datasets used as predictors in this study.

5.2. Comparison and analysis of results on $PM_{2.5}$ concentration prediction

We compared the performance of the proposed ST-Transformer model with baseline models, including the Auto Regressive Integrated Moving Average (ARIMA; Box et al., 2015), Feedforward Neural Networks (FNN), Long Short Term Memory (LSTM), Multivariate Time Series Forecasting with Graph Neural Networks (MTGNN, (Wu et al., 2020) and Time Series Transformer (Li et al., 2019; Vaswani et al., 2017).

1) ARIMA uses the relationship between the current and historical values within the time series to predict future values. It assumes that data has an autoregression relationship with its past values. We used the Auto ARIMA model without tuning the required parameters (i.e., the p, d, and q values) of ARIMA to provide optimal forecasting.
2) Feedforward Neural Networks (FNN) is a multi-layer perceptron with additional hidden nodes between the input and the output layers. Data moves in the only forward direction in this network without any cycles or loops (Schmidhuber, 2015a, b). This research aims to produce multi-horizon time-series predictions so designing an FNN with multiple outputs is essential. In the output layer, each neuron focuses on predicting the target variable at a different predicting hour so the model does not consider the outputs to be sequential (i.e., the same variable at different time steps). The problem with the FNN architecture is that the outputs are essentially predicted from different systems simultaneously.
3) The Stacked-LSTM model has a stack of two LSTM layers and a final fully connected layer to predict the multi-step time series. Stacking LSTM allows for greater model complexity (Graves, 2012). The LSTM layers encode sequential information from input through the recurrent network. The densely connected layer takes the final output from the second LSTM layer for predictions.
4) The MTGNN model is a graph neural network that learns to connect multiple variables for time series prediction. The model architecture consists of three major components: graph structure learning, graph convolution, and temporal convolution. Instead of leveraging a predefined graph structure, MTGNN is adaptive to input data to learn the evolving adjacency matrices representing sparse neighboring dependencies. The temporal convolution is responsible for identifying temporal patterns with multiple frequencies and can handle long time series.
5) The Time Series Transformer model is described in Section 3.2. We used a latent dimension d of 96, a query size of 48, a value size of 48, a head number of 8, 4 layers of encoder and decoder to stack, and 48 backward elements to apply attention. We apply dropout techniques for the three types of sub-layers in the encoder and decoder: the self-attention sub-layer, the feed-forward sub-layer, and the normalization sub-layer. A dropout rate of 0.2 is used for each sub-layer.

In the experiments, we measure the accuracy of the models using mean absolute error (MAE), root mean squared error (RMSE), and mean bias:

$$\text{Mean absolute error} : \text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\widehat{y_i} - y_i| \quad (10)$$

$$\text{Root mean squared error} : \text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\widehat{y_i} - y_i)^2} \quad (11)$$

$$\text{Bias error} : \text{Bias} = \widehat{y_i} - y_i, \quad (12)$$

where $\widehat{y_i}$ is the prediction and y_i is the ground truth for data sample i.

Table 2 shows that all models considered achieve a mean RMSE within the range of 7–10 μg m^{-3}. The ST-Transformer model achieves the best performance of all models considered while the Stacked-LSTM model achieves the second-best performance. The Time Series Transformer has a similar RMSE and MAE as the Stacked-LSTM but exhibits a wider range of errors. Higher errors generally occur during wildfire events, especially when smoke observations are not considered important in the models. The RMSE, MAE, and BIAS values for each predicting hour during the test period are illustrated in Fig. 4. Most models, except for MTGNN, show the decaying performance of RMSE and MAE with the predicting future hour. As the best performing model, ST-Transformer has an increasing RMSE value from 7 to 8 μg m^{-3}, while the worst performing model, ARIMA, has an increasing RMSE from 7.5 to 12 μg m^{-3}. Except for ST-Transformer, the other models considered show a significant

Table 2
Performance comparison among the six considered models on the same testing dataset.

Models	RMSE (μg m^{-3})	MAE (μg m^{-3})	BIAS (μg m^{-3})
ARIMA	10.0 ± 3.28	6.05 ± 1.6	2.51 ± 2.97
FNN	8.83 ± 2.77	6.01 ± 1.93	3.26 ± 3.44
Stacked-LSTM	8.38 ± 2.92	5.73 ± 2.01	2.8 ± 3.55
MTGNN	9.78 ± 2.98	6.65 ± 1.1	0.23 ± 4.05
Transformer	8.48 ± 3.86	5.42 ± 3.0	1.76 ± 4.24
ST-Transformer	6.92 ± 2.93	4.0 ± 1.36	−0.51 ± 0.87

Table 1
Datasets used in this study.

Source	Shortname	Variable	Resolutions
EPA Air Quality System (AQS)	$PM_{2.5}$	$PM_{2.5}$ concentration (μg m^{-3})	Hourly
ECMWF ERA-5	BLH	Boundary layer height (m)	0.25° × 0.25°, hourly
	D2M	2 m dewpoint temperature (K)	
	SP	Surface pressure (Pa)	
	T2M	2 m temperature (K)	
	WD	10 m wind direction (°)	
	WS	10 m wind speed (ms^{-1})	
California Department of Forestry and Fire Protection (CAL FIRE)	IDWF	Inverse distance of the nearest fire (km^{-1})	–
Temporal trends	Day	Day of year	–
	Hour	Hour of day	
Caltrans Performance Measurement System (PeMS)	TRF	Total traffic flow within 5 km (count)	Hourly
NASA Fire Information for Resource Management System (FIRMS)	IDWFI_FP	Inverse distance weighted sum of brightness temperature of fire hotspots within CALFIRE fire perimeters (K)	Sub-daily, 1 km grids
	IDWFI	Inverse distance weighted sum of brightness temperature of all fire hotspots (K)	

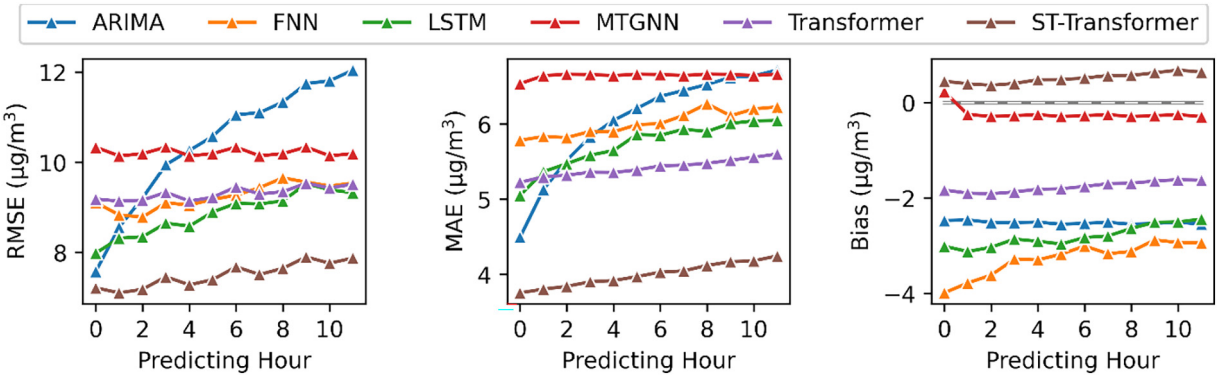


Fig. 4. Performance changes of different models on the test dataset as the predicting hour increases.

underestimation of $PM_{2.5}$ concentrations due to the failure to predict high $PM_{2.5}$ concentration values after wildfires.

To examine the spatial distribution of the prediction errors, we compared the average RMSE values for each station between the considered models (Fig. 5). The averaged RMSEs of ARIMA are large in Los Angeles County, where $PM_{2.5}$ concentrations vary significantly from hour to hour. FNN and LSTM showed a similar spatial distribution of RMSE values to ARIMA, but with lower RMSE values in general. The other three models, MTGNN, Transformer, and ST-Transformer show similar spatial distributions, where the Glendora station in Los Angeles County has the highest mean RMSE. In the ST-Transformer model, most stations (except for the Glendora station) benefited from leveraging neighboring stations' information with evolving attention. Without spatiotemporal attention, the Transformer model failed to predict accurately for stations in Ventura County, even with the information from neighboring stations. The RMSE spatial distribution of the MTGNN prediction showed that the learned adjacency

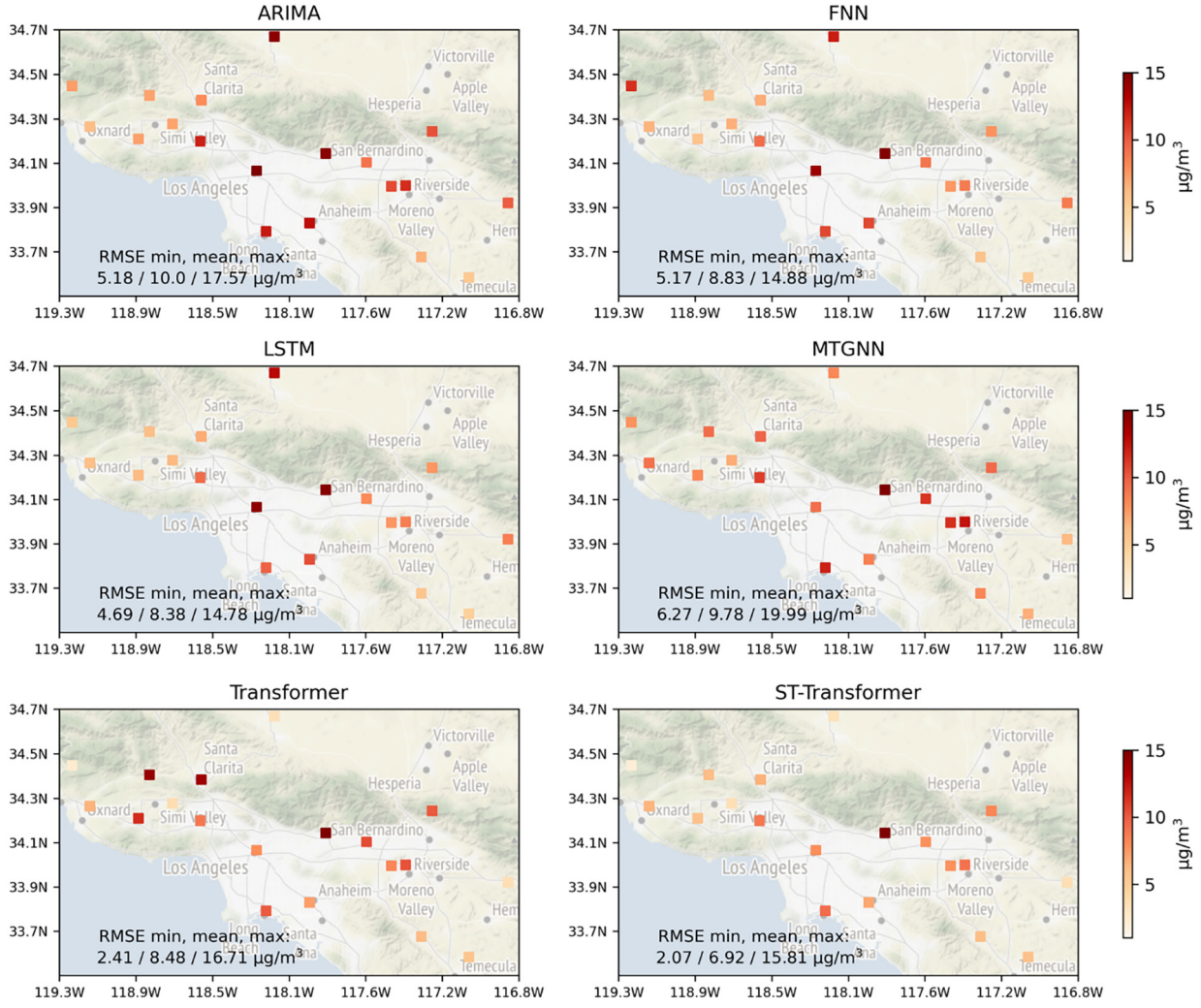


Fig. 5. Visualization of spatial RMSEs of the considered model predictions compared with the ground truth.

matrix might be adjusted with data, but in an area impacted by occasional wildfire smoke, the learned adjacency matrix is insufficient for predicting highly variable $PM_{2.5}$ concentrations.

5.3. Time series prediction

To demonstrate the effectiveness of the proposed ST-Transformer model, we showcase the predictions for days with and without nearby wildfires during the test period, along with the observed time series and the predictions from Transformer. Fire incidents are selected from the CAL FIRE database of 2020 wildfire incidents when the incident caused a burn area larger than 1000 acres (Table 3), whereas the non-wildfire period is selected when no major wildfires are on record.

The characteristics of air quality time series during wildfire events show rapidly increased air pollutant concentrations, significantly different from the days without wildfire impacts. It is challenging to model air quality for areas with potential wildfire impacts, because air quality time series are not easily separated into wildfire-impacted or non-wildfire days. When training a time series forecasting model, using the days without wildfire impacts, the predictions are more likely to follow the usual scenario of a daily or weekly recurrent pattern. When training the model using days with wildfire impacts, the predictions will be higher values with greater variance. We trained our models using all of the time series, regardless of distance to an existing wildfire, so that the models can predict $PM_{2.5}$ concentrations as a function of distance from wildfires.

The ST-Transformer predictions, Transformer predictions, and observations are illustrated in Fig. 6, in orange, green, and blue lines. In each panel, the blue curve is the observed $PM_{2.5}$ concentrations, and the orange and green curves are the predicted concentrations by ST-Transformer and Transformer models. The ST-Transformer predictions can capture the peaked $PM_{2.5}$ concentrations during wildfire events, while the Transformer predictions significantly underestimate the high $PM_{2.5}$ concentrations. For example, during the Lake Fire (Aug 12–Sep 6, 2020) and the co-occurring Ranch 2 Fire (Aug 13–Oct 5, 2020) in Los Angeles County, spiking high $PM_{2.5}$ concentration values were observed on Sep 29–31, 2020, for most monitoring stations in Los Angeles County (Fig. 6a–f). While the Transformer model can qualitatively predict an increase in $PM_{2.5}$ concentration, it cannot predict the same magnitude of the observation, which is ~200 μg m^{-3}. However, the ST-Transformer can predict an abrupt increase in $PM_{2.5}$ concentration. Such improvement of the ST-Transformer over the Transformer should be expected because the embedded spatial attention mechanism can utilize the interaction among stations by adjusting the attention weights. The attention mechanism captures the flow of spatial, temporal, and variable dependencies about the environmental contexts, including wildfire hotspots, wildfire smoke detection, and wind speed and direction (mode details are provided in Section 5.4).

In addition, the Transformer model may overestimate the $PM_{2.5}$ concentrations while taking information from neighboring stations. For example, during the Lake Fire (Aug 12–Sep 6, 2020) and the co-occurring Ranch 2 Fire (Aug 13–Oct 5, 2020) in Los Angeles County, most stations had spiking $PM_{2.5}$ concentrations of ~200 μg m^{-3}, whereas the Santa Clarita station was less impacted by the nearby wildfires, resulting in a lower $PM_{2.5}$ concentration increase to ~60 μg m^{-3} (Fig. 6e). The Transformer model uses a full attention mechanism that leverages the nearby stations' information equally, thus resulting in an overestimation during days without wildfire impacts (Fig. 6e, green line representing Transformer results). A similar example can be observed at the Riverside–Rubidoux station (Fig. 6h) during Airport Fire (Dec 1–12, 2020) and the co-occurring Sanderson Fire (Dec 12–14, 2020). The full attention inside the Transformer model leads to an overestimation of $PM_{2.5}$ concentration based on the nearby stations' context, whereas the ST-Transformer accurately attends only to the useful and relevant information.

Such overestimations are also observed during days without major co-occurring wildfires. During non-wildfire days, observations of $PM_{2.5}$ concentrations generally show weekly or daily recurrent patterns. ST-Transformer generally follows the temporal patterns of the observed $PM_{2.5}$ concentration, whereas Transformer may overestimate due to the full attention mechanism (Fig. 7). For example, during Oct 1–15, 2020, no major wildfires were happening in the study area, but Transformer overestimated the $PM_{2.5}$ concentration at Santa Clarita, Crestline–Lake Gregory, Upland, Thousand Oak–Moorpark Road, and Piru–Pacific stations (Fig. 7a, i, k, m, n). These overestimations indicated that the full attention of the Transformer model attends equally to the information from nearby positions in the high dimensional subspaces, so the trained model may not differentiate non-wildfire situations from wildfire situations well. However, the ST-Transformer uses evolving sparse attention to actively select meaningful information from explicitly separated spatial, temporal, and variable dimensions, resulting in better differentiation between wildfire and non-wildfire situations.

5.4. Model interpretation

The attention weights learned by the ST-Transformer to account for the evolving spatiotemporal information of multiple factors can provide hints on the importance of different stations for predictions. Fig. 8 demonstrates the spatial attention weights on six time series chosen from the test dataset, in which three of them are under the influence of wildfire and the other three without fire. Each panel corresponds to an attention weight matrix with its (i, j)th element being the weight measuring how much information is learned from station j for forecasting the air quality at station i. The color of the blocks in each panel represents the magnitude of attention weight. Wildfire influences can be associated with a single station (Fig. 8a: nearby fire), multiple stations (Fig. 8b: nearby fire and smoke transported by wind), or all stations within the area (Fig. 8c: large or long-term fire). Comparing attention weights with and without wildfire influence, wildfire signals are more outstanding than the ones in no-fire situations, where the latter show lower attention weights and relatively similar influences from other stations. The more important variables are also not related to wildfires, no matter the prediction relies on information from one station (Fig. 8d: previous $PM_{2.5}$), multiple stations (Fig. 8e: temporal indicator), or all stations in the area (Fig. 8f: traffic).

These attention weights can be decomposed to understand the attention learned from other stations at different past time steps. To predict the air quality at station j, the ST-Transformer learns to attend to the past information from station j itself, multiple stations, or all available stations in the region (Figs. 9). For predictions that rely on past information from its own station (Fig. 9a and d), temporal dependency generally becomes more important as the time approaches the predicted future time steps (Fig. 9a1 and d1). The variable-temporal attention plots (Fig. 9a2 and d2) generally show one identifiable hotspot, indicating the variables' joint influence on the prediction. Under wildfire influence, Fig. 9(a2) shows an attention hotspot from the past 16th–22nd time steps in variables: 2 m temperature (T2M), 10 m wind direction (WD), 10 m wind speed (WS), day of year, and hour of day. In contrast, without wildfire influence, variables that are not relevant to wildfire show higher attention weights. For example,

Table 3
Fire incidents in the study area during the test period (CAL FIRE: https://www.fire.ca.gov/incidents/2020/).

Fire event	Start date	Containment date	County	Acres
Ranch 2 Fire	8/13/2020	10/5/2020	Los Angeles	4237
Lake Fire	8/12/2020	9/28/2020	Los Angeles	31089
Soledad Fire	7/5/2020	7/10/2020	Los Angeles	1525
Bond Fire	12/2/2020	12/10/2020	Orange	6686
Blue Ridge Fire	10/26/2020	11/7/2020	Orange	13964
Silverado Fire	10/26/2020	11/7/2020	Orange	12466
Sanderson Fire	12/12/2020	12/14/2020	Riverside	1933
Airport Fire	12/1/2020	12/12/2020	Riverside	1087
El Dorado Fire	9/5/2020	11/16/2020	San Bernardino and Riverside	22744
Holser Fire	8/17/2020	8/30/2020	Ventura	3000

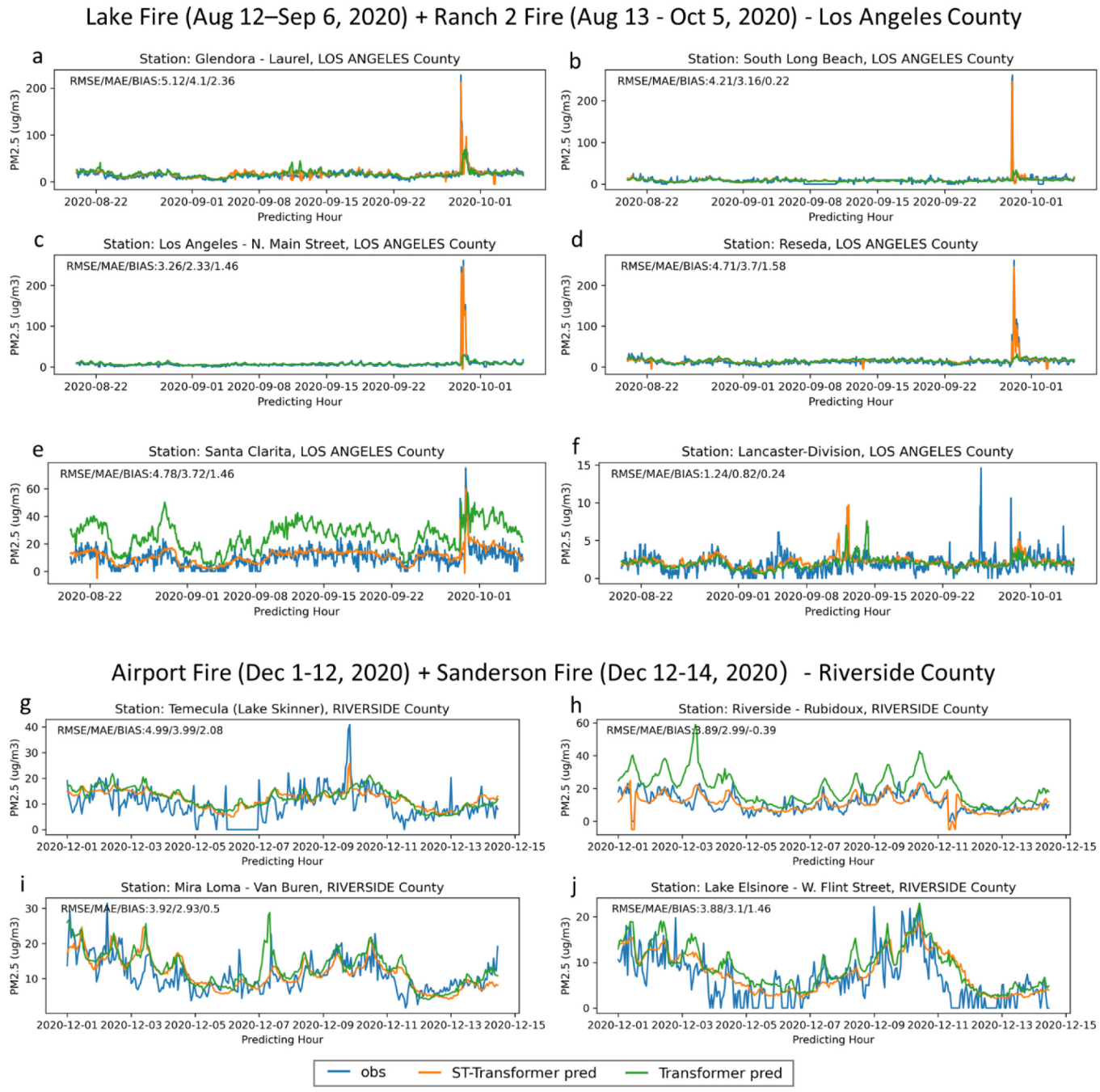


Fig. 6. The observation, Transformer prediction, and ST-Transformer prediction of $PM_{2.5}$ concentrations during wildfire events. (Note: the y-axis ranges may vary.) Models are initiated every 3 h, and predictions are averaged from model results in a rolling update manner.

Fig. 9(d2) shows a more focused hotspot from the past 11th time step of $PM_{2.5}$, indicating that the prediction relies majorly on that single variable from the past.

For predictions that rely on past information from multiple stations in the region (Fig. 9b and e), the stations used to predict at a targeting station are not necessarily nearby, indicating that the complex and long-range spatial dependencies are being captured in the model. The scattered hotspots are easy to be identified. Under the influence of wildfire, Fig. 9(b2) shows great attention weights for the inverse distance of the fire intensity (IDWFI) and wind direction (WD), indicating the high impact of nearby wildfires on multiple stations. Without wildfire influences, Fig. 9(e2) shows the single hotspot of attention from the variable 'hour of day' at Hour 13.

For predictions that rely on past information from all stations in the region (Fig. 9c and f), the temporal dependency shows scattered hotspots across the past hours and different stations (Fig. 9c1 and f1), but the temporal attentions are more complex than in previous cases, with more rapidly increasing and decreasing attentions The variable-temporal attention plots (Fig. 9c2 and f2) also show more than one identifiable hotspots, indicating that the variables' influences on the prediction are complex and dynamic. Notably, $PM_{2.5}$ concentrations from past hours are not necessarily the most important variable (Fig. 9c2), and the temporal dependency between predicting variables and predictions is not necessarily more important over time. Under wildfire influences, the temporal attentions from wildfire variables, especially IDWFI_FP (code: 11) and IDWFI (code: 12), show scattered hotspots at different past hours. Without wildfire influences,

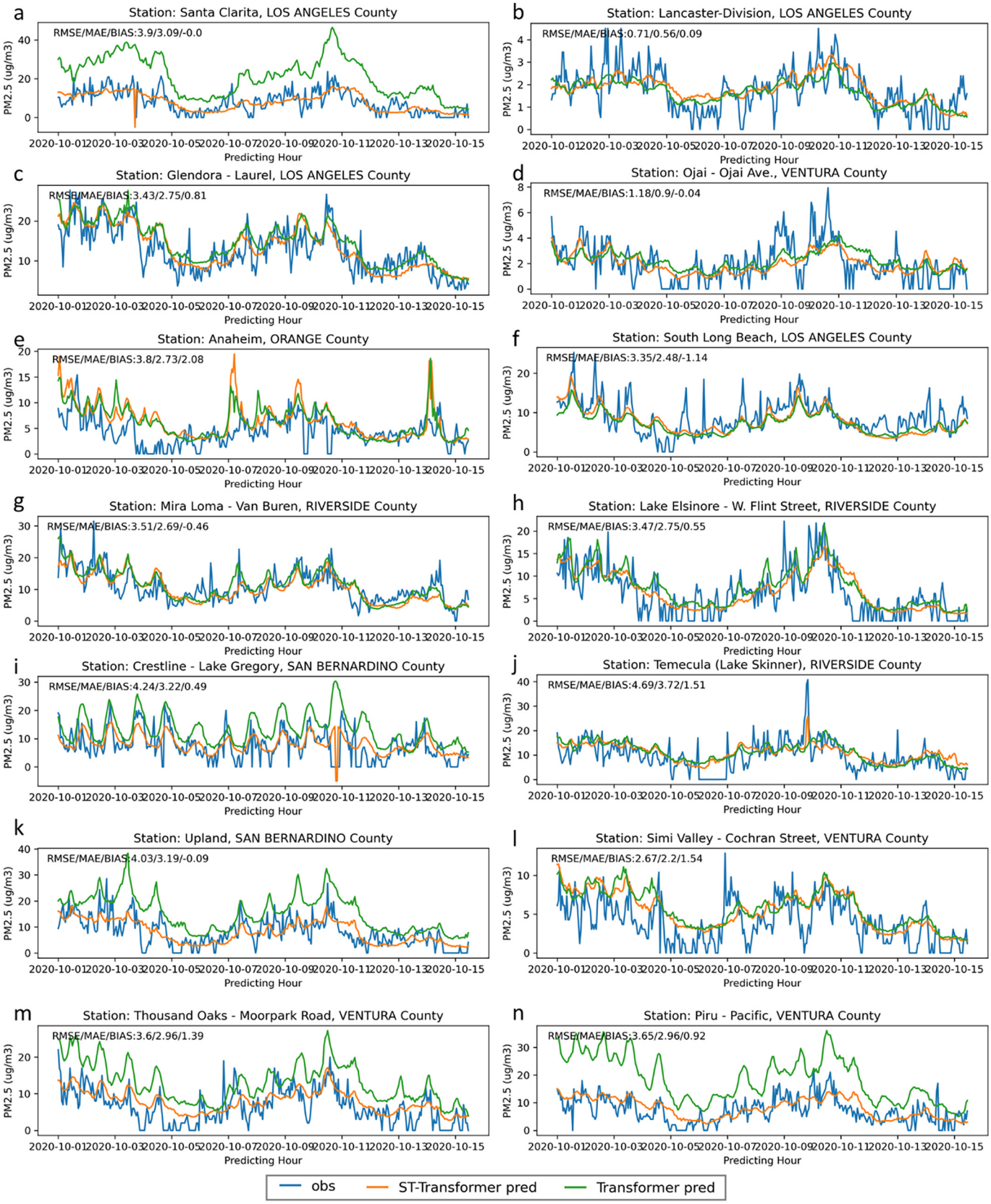


Fig. 7. The observation, Transformer prediction, and ST-Transformer prediction of $PM_{2.5}$ concentrations during non-wildfire situations. (Note: the y-axis ranges may vary.) Models are initiated every three hours, and predictions are averaged from model results in a rolling update manner.

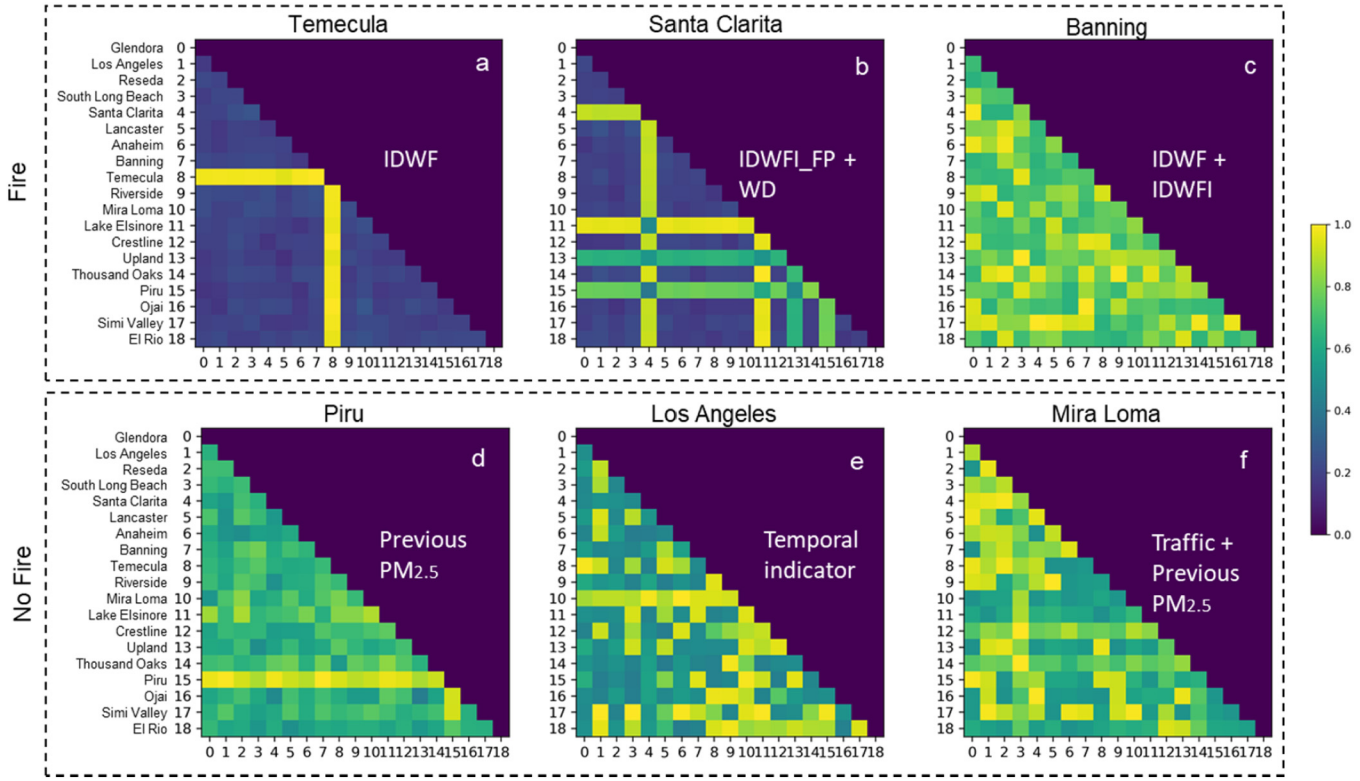


Fig. 8. Visualization of the attention weights learned by the ST-Transformer. Each panel corresponds to the spatial attention weights of one of the six chosen days (three under the influence of wildfire, three without fire) from the testing data set. Each panel corresponds to an attention weight matrix with its (i, j)th element of the weight measuring how much information is learned from station j for forecasting the air quality at station i.

Fig. 9(f2) shows multiple hotspots, but one of them indicated great attention to traffic flow at Hours 2–5.

6. Discussions

This research aimed to address the challenge of a deep learning-based approach to accurately predicting and physically interpreting air quality dynamics in space and time, particularly the concentration of $PM_{2.5}$ amidst ongoing wildfire activity. Specifically, we address two challenges: 1) accurately capture dynamic multivariate dependencies across space and time; and 2) prevent overgeneralizing the temporal trends of $PM_{2.5}$ concentration when wildfires cause abrupt environmental changes. The proposed spatiotemporal Transformer (ST-Transformer) network addresses these challenges by dynamically scaling the multi-head attention's receptive field to capture multivariate dependencies via learning spatial, temporal, and value embeddings and by introducing a sparse attention mechanism to capture abrupt changes in data. This study contributes to the existing literature on using attention mechanisms for multivariate time series forecasting, with a novelty that the multivariate represents spatial locations, temporal changes, and contextual variables used in the model. Adding multiple contextual variables at a particular station location complicates the problem into another dimension; thus, the interactions between the input variables and the target variable are more complex. As our experiments suggest, the ST-Transformer model captures the complex dependencies across multiple dimensions using the sparse attention mechanism and outperforms the other existing spatiotemporal forecasting models.

This research particularly focused on applying the proposed ST-Transformer model to accurately predict $PM_{2.5}$ concentrations in a wildfire-prone area, where a significant source of $PM_{2.5}$ is the emission from wildfires. The model considers contextual variables that account for the wildfire behavior and smoke coverage to accurately predict $PM_{2.5}$ concentration in such areas. Although wildfire behavior and smoke transport are dynamic and challenging to quantify, the ST-transformer model is designed to learn the environmental changes and their impacts on regulating $PM_{2.5}$ concentration in local regions. The other benefit of using ST-Transformer lies in the availability of model interpretation, as attention-based mechanisms provide attention maps learned from the model. Visualizing the attention matrices demonstrated the complex dependencies along spatial, temporal, and variable-wise dimensions. In the attention visualizations, one or more high attention clusters, or hotspots, can be observed and identify the spatial, temporal, and variable factors that contribute to the change of $PM_{2.5}$ concentration in the prediction. These attention visualizations demonstrate the feasibility of using ST-Transformer for wildfire smoke-induced $PM_{2.5}$ concentration mapping. When nearby wildfires are the major sources of the increase of $PM_{2.5}$ concentration, the attention mechanisms generally show high attention values in those variables related to wildfire occurrence. However, these high attention values are not necessarily associated with just nearby stations; instead, they might be associated with stations far away from each other but simultaneously impacted by the dynamic transport of wildfire smoke. Moreover, in non-fire situations, the attention visualizations may show spikes in surface temperature, wind speed and direction, past $PM_{2.5}$ concentration, or traffic flow. Therefore, the ST-Transformer model can be used to identify complex dependencies among wildfires, other emission sources, and air pollution over space and time.

Another advantage of the ST-Transformer model is its capability to predict abrupt changes in $PM_{2.5}$ concentrations based on contextual information. Compared to other existing time series forecasting models (e.g., ARIMA, FNN, LSTM, MTGNN, and Time Series Transformer), the ST-Transformer model predicts more accurately when there are spikes in the observed $PM_{2.5}$ concentrations due to nearby wildfires, while other models tend to underestimate the $PM_{2.5}$ concentrations significantly. With the sparse attention mechanisms, the ST-Transformer model can better differentiate the situations of wildfire and non-wildfire and choose to leverage the contextual information more intelligently. The accurate predictions from the ST-transformer model of $PM_{2.5}$ during wildfire seasons

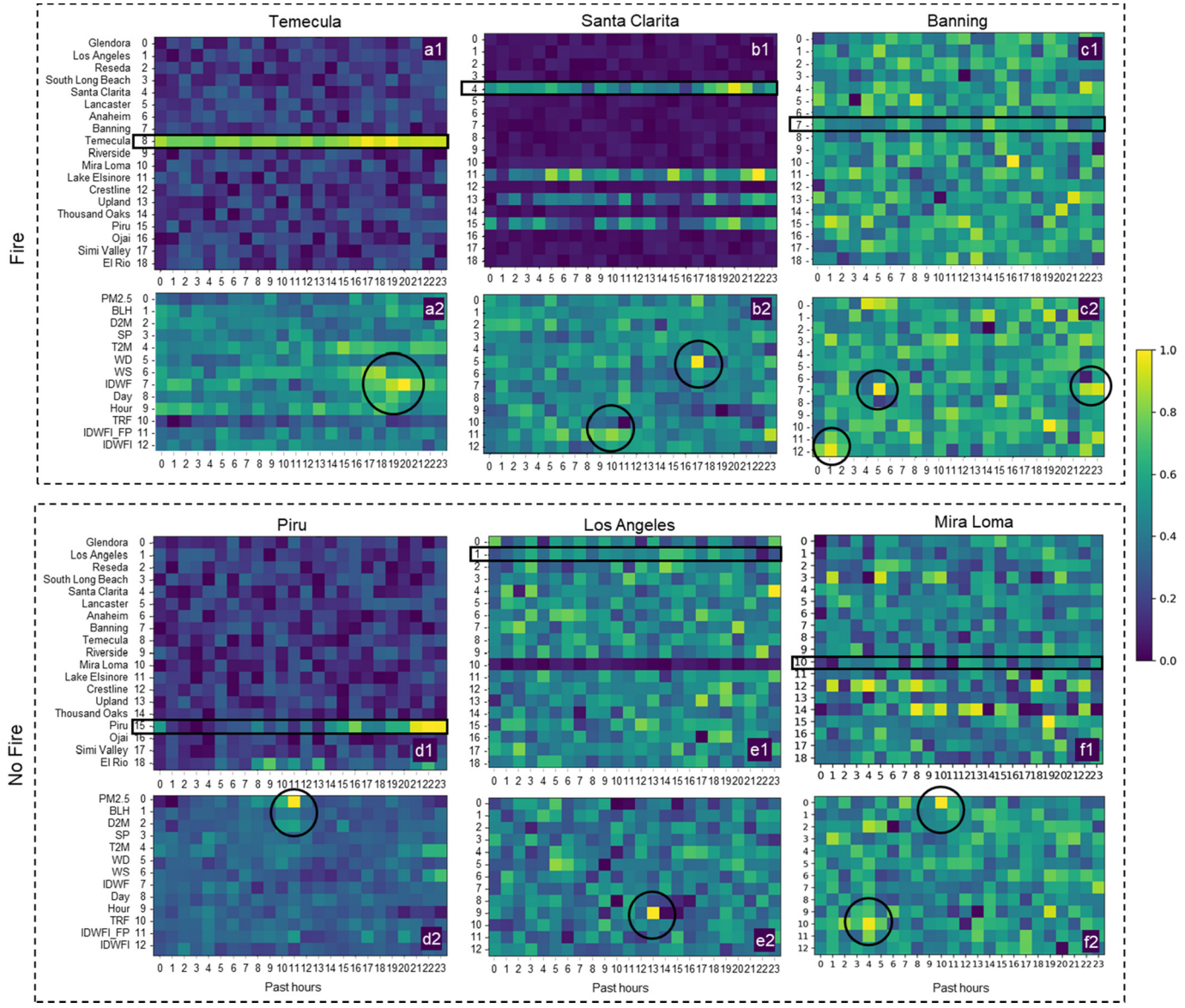


Fig. 9. Visualization of the attention weights learned by the ST-Transformer, where attention weights are learned only from the to-be-predicted station.

will be helpful to anyone as a citizen or those in the Emergency Services Sector who need to monitor and predict the impacts of wildfire smoke effectively. The predictions will also help public health officials provide appropriate guidelines to vulnerable populations more sensitive to the increase of $PM_{2.5}$ amounts in the atmosphere.

Future improvements for this research are threefold. *First*, the generalizability of the model can be tested across different regions within the wildfire-prone area, and the interpretation of feature spatiotemporal importance can be used to understand the local context for air pollution predictions. *Second*, the sensitivity of historical periods and predicting periods can be tested across a series of experiments to investigate the model's capability of predicting longer-term time series. *Third*, while this research predicts $PM_{2.5}$ concentrations, the components of $PM_{2.5}$ can be further decomposed as the output of the predictions, providing a more comprehensive estimation for public health guidance. *Fourth*, the model conducts time series forecasting for locations of monitoring stations, but does not provide estimates for places without a monitoring station. Future developments of spatiotemporal forecasting models that can estimate air quality for places both with and without monitoring stations will be more beneficial for air pollution exposure studies.

7. Conclusion

This research presented a spatiotemporal Transformer (ST-Transformer) model to address the challenge of accurately predicting and physically interpreting air quality, particularly $PM_{2.5}$ concentration. The model supports separate embedding components that isolate spatial, temporal, and value embeddings to facilitate a more focused and sparse attention mechanism that helps learn the complex and dynamically changing spatial, temporal, and variable-wise dependencies for estimating air quality during wildfires. The proposed model is applied and investigated for an hourly $PM_{2.5}$ concentration prediction in the greater LA area from 2017 to 2020. The input variables include wildfire perimeter (CAL FIRE), wildfire intensity (NASA Active Fire), meteorological factors, traffic, $PM_{2.5}$ concentration, and temporal indicators from the past 24 h, and the targeting output variable is the $PM_{2.5}$ concentrations in the future 12 h. The ST-Transformer model achieved an average RMSE of 6.92 ($\pm$2.93) $\mu g\ m^{-3}$ and outperformed other considered models, including the ARIMA, FNN, LSTM, MTGNN, and Time Series Transformer. The ST-Transformer model also showed significant improvements from other considered models during wildfire events, where the time series of $PM_{2.5}$ concentration shows an identifiable spike. The attention matrix learned by the model also

interpreted the complex spatial, temporal, and variable-wise dependencies, indicating that the model can differentiate situations between wildfires and non-wildfires. The ST-Transformer model can be easily adopted for other spatiotemporal prediction problems, such as predicting water quality, precipitation, and solar radiation, at multiple monitoring sites where each observation location is associated with multiple environmental variables. The model is also suitable for predicting variables with complex temporal dependencies, such as when observations generally show spikes or pits without clear reasons. The ST-Transformer model supports generating attention matrices that might help understand the reasons for spikes and pits through the decomposed visualization of attention clusters across the spatial, temporal, and variable-wise dimensions. The accurate predictability and interpretation capacity of the ST-transformer can help effectively monitor and predict the impacts of wildfire smoke and can be applicable to other complex spatiotemporal prediction problems, including water quality, precipitation, and solar radiation mapping.

CRediT authorship contribution statement

Manzhu Yu: Data curation, Conceptualization, Methodology, Code implementation, Experiment, Result analysis, Paper Writing.

Arif Masrur: Result analysis, Paper review and editing.

Christopher Blaszczak-Boxe: Result analysis, Paper review and editing.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Aguilera, R., Corringham, T., Gershunov, A., Benmarhnia, T., 2021. Wildfire smoke impacts respiratory health more than fine particles from other sources: observational evidence from Southern California. Nat. Commun. 12, 1493. https://doi.org/10.1038/s41467-021-21708-0.

Ahmadov, R., Grell, G., James, E., Csiszar, I., Tsidulko, M., Pierce, B., McKeen, S., Benjamin, S., Alexander, C., Pereira, G., Freitas, S., Goldberg, M., 2017. Using VIIRS fire radiative power data to simulate biomass burning emissions, plume rise and smoke transport in a real-time air quality modeling system. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Presented at the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 2806–2808 https://doi.org/10.1109/IGARSS.2017.8127581.

Appel, K.W., Napelenok, S.L., Foley, K.M., Pye, H.O.T., Hogrefe, C., Luecken, D.J., Bash, J.O., Roselle, S.J., Pleim, J.E., Foroutan, H., Hutzell, W.T., Pouliot, G.A., Sarwar, G., Fahey, K.M., Gantt, B., Gilliam, R.C., Heath, N.K., Kang, D., Mathur, R., Schwede, D.B., Spero, T.L., Wong, D.C., Young, J.O., 2017. Description and evaluation of the community multiscale air quality (CMAQ) modeling system version 5.1. Geosci. Model Dev. 10, 1703–1732. https://doi.org/10.5194/gmd-10-1703-2017.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series Analysis: Forecasting and Control. John Wiley & Sons 978-1-118-67502-1.

Brey, S.J., Ruminski, M., Atwood, S.A., Fischer, E.V., 2018. Connecting smoke plumes to sources using Hazard Mapping System (HMS) smoke and fire location data over North America. Atmos. Chem. Phys. 18, 1745–1761. https://doi.org/10.5194/acp-18-1745-2018.

Buysse, C.E., Kaulfus, A., Nair, U., Jaffe, D.A., 2019. Relationships between particulate matter, ozone, and nitrogen oxides during urban smoke events in the Western US. Environ. Sci. Technol. 53, 12519–12528. https://doi.org/10.1021/acs.est.9b05241.

Cheng, W., Shen, Y., Zhu, Y., Huang, L., 2018. A neural attention model for urban air quality inference: learning the weights of monitoring stations. Proc. AAAI Conf. Artif. Intell. 32.

Crouse, D.L., Peters, P.A., Hystad, P., Brook, J.R., van Donkelaar, A., Martin, R.V., Villeneuve, P.J., Jerrett, M., Goldberg, M.S., Pope, C.A., Brauer, M., Brook, R.D., Robichaud, A., Menard, R., Burnett, R.T., 2015. Ambient PM2.5, O_3, and NO_2 exposures and associations with mortality over 16 years of follow-up in the Canadian Census Health and Environment Cohort (CanCHEC). Environ. Health Perspect. 123, 1180–1186. https://doi.org/10.1289/ehp.1409276.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press 9780262035613.

Graves, A., 2012. Long short-term memory. In: Graves, A. (Ed.), Supervised Sequence Labelling With Recurrent Neural Networks, Studies in Computational Intelligence. Springer, Berlin, Heidelberg, pp. 37–45 https://doi.org/10.1007/978-3-642-24797-2_4.

Grigsby, J., Wang, Z., Qi, Y., 2021. Long-Range Transformers for Dynamic Spatiotemporal Forecasting ArXiv210912218 Cs Stat.

Gupta, P., Doraiswamy, P., Levy, R., Pikelnaya, O., Maibach, J., Feenstra, B., Polidori, A., Kiros, F., Mills, K.C., 2018. Impact of California fires on local and regional air quality: the role of a low-cost sensor network and satellite observations. GeoHealth 2, 172–181. https://doi.org/10.1029/2018GH000136.

Habre, R., Girguis, M., Urman, R., Fruin, S., Lurmann, F., Shafer, M., Gorski, P., Franklin, M., McConnell, R., Avol, E., Gilliland, F., 2021. Contribution of tailpipe and non-tailpipe traffic sources to quasi-ultrafine, fine and coarse particulate matter in southern California. J. Air Waste Manag. Assoc. 71 (2), 209–230.

Han, J.M., Ang, Y.Q., Malkawi, A., Samuelson, H.W., 2021. Using recurrent neural networks for localized weather prediction with combined use of public airport data and on-site measurements. Build. Environ. 192, 107601. https://doi.org/10.1016/j.buildenv.2021.107601.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G.D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R.J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., Thépaut, J.-N., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146, 1999–2049. https://doi.org/10.1002/qj.3803.

Hung, W.-T., Lu, C.-H.(Sarah), Alessandrini, S., Kumar, R., Lin, C.-A., 2021. The impacts of transported wildfire smoke aerosols on surface air quality in New York State: a multi-year study using machine learning. Atmos. Environ. 259, 118513. https://doi.org/10.1016/j.atmosenv.2021.118513.

Jaffe, D.A., O'Neill, S.M., Larkin, N.K., Holder, A.L., Peterson, D.L., Halofsky, J.E., Rappold, A.G., 2020. Wildfire and prescribed burning impacts on air quality in the United States. J. Air Waste Manag. Assoc. 70, 583–615. https://doi.org/10.1080/10962247.2020.1749731.

Khaykin, S., Legras, B., Bucci, S., Sellitto, P., Isaksen, L., Tencé, F., Bekki, S., Bourassa, A., Rieger, L., Zawada, D., Jumelet, J., Godin-Beekmann, S., 2020. The 2019/20 Australian wildfires generated a persistent smoke-charged vortex rising up to 35 km altitude. Commun. Earth Environ. 1, 1–12. https://doi.org/10.1038/s43247-020-00022-5.

Kochanski, A.K., Jenkins, M.A., Yedinak, K., Mandel, J., Beezley, J., Lamb, B., 2016. Toward an integrated system for fire, smoke and air quality simulations. Int. J. Wildland Fire 25, 534. https://doi.org/10.1071/WF14074.

Larsen, A., Hanigan, I., Reich, B.J., Qin, Y., Cope, M., Morgan, G., Rappold, A.G., 2021. A deep learning approach to identify smoke plumes in satellite imagery in near-real time for health risk communication. J. Expo. Sci. Environ. Epidemiol. 31, 170–176. https://doi.org/10.1038/s41370-020-0246-y.

Li, L., Girguis, M., Lurmann, F., Pavlovic, N., McClure, C., Franklin, M., Wu, J., Oman, L.D., Breton, C., Gilliland, F., Habre, R., 2020. Ensemble-based deep learning for estimating PM2.5 over California with multisource big data including wildfire smoke. Environ. Int. 145, 106143. https://doi.org/10.1016/j.envint.2020.106143.

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.-X., Yan, X., 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. Advances in Neural Information Processing Systems. Curran Associates, Inc.

Li, Y., Moura, J.M.F., 2020. Forecaster: A Graph Transformer for Forecasting Spatial and Time-Dependent Data. https://doi.org/10.3233/FAIA200231.

Li, Y., Yu, R., Shahabi, C., Liu, Y., 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting ArXiv170701926 Cs Stat.

Lim, B., Zohren, S., 2021. Time-series forecasting with deep learning: a survey. Philos. Trans. R. Soc. Math. Phys. Eng. Sci. 379, 20200209. https://doi.org/10.1098/rsta.2020.0209.

Lin, Y., Mago, N., Gao, Y., Li, Y., Chiang, Y.-Y., Shahabi, C., Ambite, J.L., 2018. Exploiting spatiotemporal patterns for accurate air quality forecasting using deep learning. Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, SIGSPATIAL '18. Association for Computing Machinery, New York, NY, USA, pp. 359–368 https://doi.org/10.1145/3274895.3274907.

Linn, R., Reisner, J., Colman, J.J., Winterkamp, J., 2002. Studying wildfire behavior using FIRETEC. Int. J. Wildland Fire 11, 233–246. https://doi.org/10.1071/wf02007.

Linn, R., Winterkamp, J., Colman, J.J., Edminster, C., Bailey, J.D., 2005. Modeling interactions between fire and atmosphere in discrete element fuel beds. Int. J. Wildland Fire 14, 37–48. https://doi.org/10.1071/WF04043.

Liu, P., Chang, S., Huang, X., Tang, J., Cheung, J.C.K., 2019. Contextualized non-local neural networks for sequence learning. Proc. AAAI Conf. Artif. Intell. 33, 6762–6769. https://doi.org/10.1609/aaai.v33i01.33016762.

Liu, Y., Kochanski, A., Baker, K.R., Mell, W., Linn, R., Paugam, R., Mandel, J., Fournier, A., Jenkins, M.A., Goodrick, S., Achtemeier, G., Zhao, F., Ottmar, R., French, N.H., Larkin, N., Brown, T., Hudak, A., Dickinson, M., Potter, B., Clements, C., Urbanski, S., Prichard, S., Watts, A., McNamara, D., 2019a. Fire behavior and smoke modeling: model improvement and measurement needs for next-generation smoke research and forecasting systems. Int. J. Wildland Fire 28, 570. https://doi.org/10.1071/wf18204.

Liu, Y., Kochanski, A., Baker, K.R., Mell, W., Linn, R., Paugam, R., Mandel, J., Fournier, A., Jenkins, M.A., Goodrick, S., Achtemeier, G., Zhao, F., Ottmar, R., French, N.H.F., Larkin, N., Brown, T., Hudak, A., Dickinson, M., Potter, B., Clements, C., Urbanski, S., Prichard, S., Watts, A., McNamara, D., Liu, Y., Kochanski, A., Baker, K.R., Mell, W., Linn, R., Paugam, R., Mandel, J., Fournier, A., Jenkins, M.A., Goodrick, S., Achtemeier, G., Zhao, F., Ottmar, R., French, N.H.F., Larkin, N., Brown, T., Hudak, A., Dickinson, M., Potter, B., Clements, C., Urbanski, S., Prichard, S., Watts, A., McNamara, D., 2019b. Fire behaviour and smoke modelling: model improvement and measurement needs for next-generation smoke research and forecasting systems. Int. J. Wildland Fire 28, 570–588. https://doi.org/10.1071/WF18204.

Lu, Y., Giuliano, G., Habre, R., 2021. Estimating hourly PM2.5 concentrations at the neighborhood scale using a low-cost air sensor network: a Los Angeles case study. Environ. Res. 195, 110653. https://doi.org/10.1016/j.envres.2020.110653.

Makar, M., Antonelli, J., Di, Q., Cutler, D., Schwartz, J., Dominici, F., 2017. Estimating the causal effect of low levels of fine particulate matter on hospitalization. Epidemiol. Camb. Mass 28, 627–634. https://doi.org/10.1097/EDE.0000000000000690.

Mallia, D.V., Kochanski, A.K., Kelly, K.E., Whitaker, R., Xing, W., Mitchell, L.E., Jacques, A., Farguell, A., Mandel, J., Gaillardon, P.-E., Becnel, T., Krueger, S.K., 2020. Evaluating wildfire smoke transport within a coupled fire-atmosphere model using a high-density observation network for an episodic smoke event along Utah's Wasatch Front. J. Geophys. Res. Atmos. 125, e2020JD032712. https://doi.org/10.1029/2020JD032712.

Mazdiyasni, O., AghaKouchak, A., 2015. Substantial increase in concurrent droughts and heatwaves in the United States. Proc. Natl. Acad. Sci. 112, 11484–11489. https://doi.org/10.1073/pnas.1422945112.

Mell, W., Jenkins, M.A., Gould, J., Cheney, P., 2007. A physics-based approach to modelling grassland fires. Int. J. Wildland Fire 16, 1–22. https://doi.org/10.1071/WF06002.

Mell, W., Maranghides, A., McDermott, R., Manzello, S.L., 2009. Numerical simulation and experiments of burning Douglas fir trees. Combust. Flame 156, 2023–2041. https://doi.org/10.1016/j.combustflame.2009.06.015.

Mueller, E., Mell, W., Simeoni, A., 2014. Large eddy simulation of forest canopy flow for wildland fire modeling. Can. J. For. Res. https://doi.org/10.1139/cjfr-2014-0184.

Natole, M., Ying, Y., Buyantuev, A., Stessin, M., Buyantuev, V., Lapenis, A., 2021. Patterns of mega-forest fires in east Siberia will become less predictable with climate warming. Environ. Adv. 4, 100041. https://doi.org/10.1016/j.envadv.2021.100041.

Pimont, F., Dupuy, J.-L., Linn, R.R., Dupont, S., 2011. Impacts of tree canopy structure on wind flows and fire propagation simulated with FIRETEC. Ann. For. Sci. 68, 523. https://doi.org/10.1007/s13595-011-0061-7.

Ramasubramanian, M., Kaulfus, A., Maskey, M., Ramachandran, R., Gurung, I., Freitag, B., Christopher, S., 2019. Pixel level smoke detection model with deep neural network. Image and Signal Processing for Remote Sensing XXV. Presented at the Image and Signal Processing for Remote Sensing XXV. International Society for Optics and Photonics, p. 1115515 https://doi.org/10.1117/12.2532562.

Reid, C.E., Jerrett, M., Petersen, M.L., Pfister, G.G., Morefield, P.E., Tager, I.B., Raffuse, S.M., Balmes, J.R., 2015. Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. Environ. Sci. Technol. 49, 3887–3896. https://doi.org/10.1021/es505846r.

Schiermeier, Q., 2018. Droughts, heatwaves and floods: how to tell when climate change is to blame. Nature 560, 20–23.

Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Netw. 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

Aguilera, R., Corringham, T., Gershunov, A., Benmarhnia, T., 2021. Wildfire smoke impacts respiratory health more than fine particles from other sources: observational evidence from Southern California. Nat. Commun. 12, 1493. https://doi.org/10.1038/s41467-021-21708-0.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762 ArXiv170603762 Cs.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press.

Mazdiyasni, O., AghaKouchak, A., 2015. Substantial increase in concurrent droughts and heatwaves in the United States. Proc. Natl. Acad. Sci. 112, 11484–11489. https://doi.org/10.1073/pnas.1422945112.

Natole, M., Ying, Y., Buyantuev, A., Stessin, M., Buyantuev, V., Lapenis, A., 2021. Patterns of mega-forest fires in east Siberia will become less predictable with climate warming. Environ. Adv. 4, 100041. https://doi.org/10.1016/j.envadv.2021.100041.

Schiermeier, Q., 2018. Droughts, heatwaves and floods: how to tell when climate change is to blame. Nature 560, 20–23.

Schmidhuber, J., 2015. Deep learning in neural networks: an overview. Neural Netw. 61, 85–117. https://doi.org/10.1016/j.neunet.2014.09.003.

Woodward, A., Smith, K.R., Campbell-Lendrum, D., Chadee, D.D., Honda, Y., Liu, Q., Olwoch, J., Revich, B., Sauerborn, R., Chafe, Z., Confalonieri, U., Haines, A., 2014. Climate change and health: on the latest IPCC report. Lancet 383, 1185–1189. https://doi.org/10.1016/S0140-6736(14)60576-6.

Wu, Z., Pan, S., Long, G., Jiang, J., Chang, X., Zhang, C., 2020. Connecting the Dots: Multivariate Time Series Forecasting With Graph Neural Networks ArXiv200511650 Cs Stat.

Yao, J., Brauer, M., Raffuse, S., Henderson, S.B., 2018. Machine learning approach to estimate hourly exposure to fine particulate matter for urban, rural, and remote populations during wildfire seasons. Environ. Sci. Technol. 52, 13239–13249. https://doi.org/10.1021/acs.est.8b01921.

Yi, X., Zhang, J., Wang, Z., Li, T., Zheng, Y., 2018. Deep distributed fusion network for air quality prediction. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. Association for Computing Machinery, New York, NY, USA, pp. 965–973.

Yu, B., Yin, H., Zhu, Z., 2018. Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting. Proc. Twenty-Seventh Int. Jt. Conf. Artif. Intell, pp. 3634–3640 https://doi.org/10.24963/ijcai.2018/505.

Yu, M., Xu, F., Hu, W., Sun, J., Cervone, G., 2021. Using Long Short-Term Memory (LSTM) and Internet of Things (IoT) for Localized Surface Temperature Forecasting in an Urban Environment.

Zhou, H., Zhang, Shanghang, Peng, J., Zhang, Shuai, Li, J., Xiong, H., Zhang, W., 2021. Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting ArXiv201207436 Cs.

Zou, Y., O'Neill, S.M., Larkin, N.K., Alvarado, E.C., Solomon, R., Mass, C., Liu, Y., Odman, M.T., Shen, H., 2019. Machine learning-based integration of high-resolution wildfire smoke simulations and observations for regional health impact assessment. Int. J. Environ. Res. Public Health 16. https://doi.org/10.3390/ijerph16122137.