Integrated Multiple Directed Attention-based Deep Learning for Improved Air Pollution Forecasting

Integrated Multiple Directed Attention-based Deep Learning for Improved Air Pollution Forecasting

Abdelkader Dairi, Fouzi Harrou, *Member, IEEE,*, Sofiane Khadraoui and Ying Sun

***Abstract*—In recent years, human health across the world is becoming concerned by a constant threat of air pollution, which causes many chronic diseases and premature mortalities. Poor air quality does not have only serious adverse effects on human health and vegetation, but also some major negative political, societal, and economic impacts. Hence, it is essential investing more effort on accurate forecasting of ambient air pollution to provide practical and relevant solutions, achieve acceptable air quality, and plan for prevention. In this work, we propose a flexible and efficient deep learning-driven model to forecast concentrations of ambient pollutants. The paper introduces first the traditional Variational AutoEncoder (VAE) and the attention mechanism to develop the forecasting modeling strategy based on the innovative Integrated Multiple Directed Attention Deep Learning architecture (IMDA). To assess the performance of the proposed forecasting methodology, experimental validation is then performed using air pollution data from four US states. Six statistical indicators have been used to evaluate the forecasting accuracy. A discussion of the results obtained finally demonstrates the satisfying performance of IMDA-VAE methods to forecast different pollutants in different locations. Furthermore, results indicate that the proposed IMDA-VAE model can effectively improve air pollution forecasting performance and outperforms the deep learning models, namely VAE, Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), bidirectional LSTM, bidirectional GRU, and ConvLSTM. We also showed that the forecasting results of the proposed model surpass the performance of LSTM and GRU with the attention mechanism.**

***Index Terms*—Air pollution concentrations forecasting, Multiple Directed Attention, deep learning, IMDA-VAE, time series.**

I. Introduction

CONCERNS for the environment, health, and safety have been attracting considerable attention worldwide due to the new environmental challenges that threaten the planet. Air pollution is becoming a critical problem in urban areas and industrialized countries and one of the principal factors for global warming. Mitigating air pollution is a paramount issue in developing countries, notably in larger urban areas with a high concentration of emission sources, including vehicles and industrial activities. Many epidemiological studies showed the effect of certain chemical compounds such as sulfur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), or dust particles in the air on the health of the general population, and particularly noticeable on sensitive people such as asthmatics, children, and elderly [1]. Today, air quality is a multidisciplinary problem that mobilizes epidemiological specialists, specialists in transport modeling, emissions and transformation of pollutants, geographical systems, forecasting, and local authorities and industrialists. Thus, monitoring the ambient air quality is essential to achieve acceptable air quality [2]. Over the past few decades, much effort has been made to enhance air quality [3], [4]. For instance, air quality networks (composed of numerous measurement stations) were installed across almost all countries to monitor numerous air pollutants' concentration levels. This study attempts to design an efficient deep learning data-driven model for forecasting concentrations of carbon monoxide (CO), nitrogen monoxide (NO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2 and ozone (O_3).

Reliable forecasting of air pollutants gives valuable information to help people take the necessary precautions to avoid undesirable consequences. Also, it permits taking more adequate countermeasures for preventing air pollution crisis and protecting public health. The need for flexible forecasting techniques that can accurately forecast air pollutants has considerably drawn researchers' and engineers' attention [5]. Various model-based and data-based methods have been designed to improve air pollution modeling and forecasting over the past four decades [1], [6], [7]. Conventional time-series-based models are among the widely utilized methods for air pollution forecasting in the literature [8]–[10]. These models comprise autoregressive integrated moving average (ARIMA) and its alternatives, like seasonal-ARIMA [11], and Holt-Winters models [12], [13]. Nevertheless, the forecast error in these methods is apparent when the concentration levels of pollutants exhibit irregular variations [14]. Also, the linear nature of statistical models (e.g., AR and ARIMA) does not enable forecasting the non-linear and non-stationary of ambient air pollution accurately [15].

To mitigate the weakness mentioned above, machine learning models, which are more flexible, such as neural network forecasting and support vector machine, have been widely employed in improving air pollution forecasting [16]–[18]. Machine learning models showed a suitable ability to model the complicated relationship between process variables without the need for an explicit model formulation to be specified. Over the last two decades, several machine learning-based

A. Dairi is with the University of Science and Technology of Oran-Mohamed Boudiaf (USTO-MB), Computer Science department Signal, image and speech laboratory (SIMPA) laboratory, El Mnaouar, BP 1505, Bir El Djir 31000, Oran, Algeria

F. Harrou, and Y. Sun are with Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, KAUST, Thuwal, 23955-6900, Saudi Arabia, ✉ fouzi.harrou@kaust.edu.sa

S. Khadraoui is with the University of Sharjah, Department of Electrical Engineering, Sharjah, United Arab Emirates

This work was supported by funding from King Abdullah University of Science and Technology (KAUST), Office of Sponsored Research (OSR) under Award No: OSR-2019-CRG7-3800.

methods have been applied for air pollution forecasting. For instance, in [16], a wavelet-based neural network approach has been proposed for forecasting one step-ahead hourly, daily mean, and daily maximum concentrations of ozone (O_3), sulfur dioxide (SO_2), carbon monoxide (CO), nitrogen oxides (NO), nitrogen dioxide (NO_2) and dust particles (PM2.5). Particularly, by using maximum overlap wavelet transform (MODWT), this approach decomposes each time series of each air pollutant into different time-scale components. Then, the Elman network is applied to these time-scale components. In this study, the wavelet network approach is used for one-step-ahead forecasting. In [19], a hybrid method merging the advantages of the empirical wavelet transform (EWT), multi-agent evolutionary genetic algorithm (MEGA), and nonlinear autoregressive network with exogenous inputs (NARX) is introduced for enhancing multi-step air pollutant concentrations forecasting. Specifically, the air pollutant series are decomposed via the EWT, and then the optimized NARX neural by the MAEGA model are applied to forecasting air pollutant concentrations. Results showed the outperformance of this model compared to conventional models when applied for forecasting PM2.5, SO_2, NO_2, and CO concentrations levels in Beijing, China. However, these methods are not suited for real-time forecasting because wavelet transform needs batch data. In [20], an ANN model has been combined with numerical models to improve the prediction of daily concentrations of air pollutants, including SO2, NO2, and PM10, using meteorological variables. Similarly, in [17], the authors introduced a hybrid forecasting model to forecast hourly pollutant levels (i.e., NO_2, NO, O_3, CO, PM_{10}) based on artificial neural networks (ANN) combined with uncertainty analysis by Monte Carlo simulations (MCS). This hybrid model uses several selected input meteorological variables for improved forecasting. It has been shown that the combination of ANN with MCS offers a promising tool for air pollution predictions. In [18], at first, a genetic algorithm and a linear method of stepwise fit are applied to select the relevant features from the prediction point of view. Then the random forest (RF), the multilayer perception (MLP), the radial basis function (RBF), and the support vector machine (SVM)are applied to the selected features for daily forecasting of atmospheric pollutants NO_2, O_3, SO_2, PM_{10}. It has been shown that the features selection step with an ensemble of predictors enables improved forecasting quality of atmospheric pollution. In [21], an approach to predict air concentration levels of air pollutants (i.e., NO_2, NOX, SO_2, O_3, and PM2.5) is proposed using based on sparse response back-propagation training feedforward neural networks (called FFANN-SRBP). This approach outperformed the multiple linear regression (MLR) and FFANN based on back-propagation (FFANN-BP) in terms of prediction precision.

Precise air pollution forecasting provides relevant information about the future pollution evolution, which is crucial for effective air pollution monitoring and assists in planning for prevention. Deep learning has recently emerged as a promising research line in modeling and forecasting time series data, both in academia and industry [22]–[26]. Various deep techniques have been applied in the literature to improve ambient air pollution forecasting. In [27], the authors considered an Aggregated Long Short-Term Memory model (ALSTM) deep learning approach to improve air pollution forecasting. Essentially, this forecasting approach consists of aggregating three LSTM models into a predictive model using information obtained from nearby industrial air quality stations and external sources of pollution. Results proved the outperformance of this aggregated deep learning method compared to LSTM, SVR (Support Vector Machine based Regression), and GBTR (Gradient Boosted Tree Regression) in predicting PM2.5 concentrations. In [28], the convolutional-based bidirectional gated recurrent unit (CBGRU) method is applied to forecast PM2.5 concentration levels. This approach combines both one-dimensional (1D) convolutional neural networks' desirable features and bidirectional gated recurrent unit neural networks. This approach showed good forecasting performance of PM2.5 concentration compared to SVR, Gradient Boosting Regressor (GBR), LSTM, GRU, and bidirectional GRU. In [29], LSTM optimized using a particle swarm optimization algorithm is applied for predicting ambient air pollutants concentrations (PM2.5, PM10 ,NO_2, CO, O_3, and SO_2). In [10], an approach combining RNN models with LSTM (RNN + LSTM) is proposed to predict PM10 particles in different places in the city Skopje. Results show that using both meteorological and air pollution measurements enhances LSTM and RNN + LSTM models' forecasting accuracy. Also, it has been shown that the combined RNN + LSTM models consistently outperform the ARIMA approach.

The authors in [30] and [31] proposed a method for air pollution prediction from historical time series pollutant data and meteorological data using RNN and LSTM models. Another related work [32] utilizes an LSTM model to forecast the 8 hours moving average concentrations of ozone, where results obtained showed forecasts with low error. Deep learning models with RNN, LSTM and GRU architectures are applied for forecasting air quality based on AirNet dataset that includes both meteorological time series and air quality data [33]. The analysis examined in [33] showed the GRU model slightly outperforms the RNN and LSTM architectures for PM_{10} concentration prediction. A deep learning approach is proposed in [34] and [35] for the air pollution prediction in South Korea, where a Stacked Autoencoders model (SAE) is utilized for learning and training data. An LSTM model-based approach is proposed in [36] to predict air pollutant concentrations. This forecasting approach presented in [36] is capable to automatically extract useful features such as the spatiotemporal correlations within air pollutant concentrations. The spatiotemporal deep learning (STDL) architectures have also been used for spatiotemporal prediction of air quality. A combination of a convolutional neural network and LSTM model is proposed in [37] to forecast air quality up to 48 hours, and extract the spatial-temporal relations. The authors in [38] presented a comparative study of the performance of CNN-LSTM model with traditional machine learning models in terms of their ability to forecast $PM_{2.5}$ concentration, where the obtained results showed that the CNN-LSTM model provides the lowest root mean square error and mean absolute

error. A CNN-GRU model proposed in [39] is applied to forecast three air pollutants ($PM_{2.5}$, PM_{10}, and O_3) of monitoring stations over 48 hours.

In this paper, we propose an innovative deep learning model to improve the forecasting quality of concentrations levels of ambient air pollutants (NO_2, O_3, SO_2, and CO) based on a wind attention mechanism and variational autoencoder (VAE) deep model. The contribution of this paper is threefold.

- The first contribution is mainly related to the development of an innovative Integrated Multiple Directed Attention Deep Learning architecture (called IMDA-VAE) based on the traditional VAE and the attention mechanism. To the best of the authors' knowledge, this is the first study introducing the traditional VAE and the attention mechanism to accurately forecast various air pollutants concentrations and leverage air pollution complexity. Essentially, the proposed method extends the capability of the traditional VAE to model temporal dependencies by introducing the self-attention mechanism at a multi-level of the VAE model. The proposed flexible modeling framework permits exploiting the suitable performance of VAE in time-series modeling and flexible nonlinear approximation and the focus on the relevant features via the attention-based mechanism. Exploiting all these sophisticated statistical tools is advantageous in the sense that it has the potential to improve short-term forecasting of ambient air pollution time-series data.
- The second contribution lies in the validation of the proposed method through three different forecasting experimentations: uni-variate, multi-variate with a single output, and multi-variate with multiple outputs.
- Finally, the third contribution is the comparative study of the proposed forecasting model and some powerful recurrent neural network models performed in this work to show and evaluate their performance and capabilities in forecasting air ambient pollutants. Air pollution data from four US states were used in the experiments and assessment of the outputs' deep learning-driven forecasting methods. The results demonstrate that the designed IMDA-VAE method offers satisfying forecasting performance of different air ambient pollutants and outperformed other deep learning models including GRUs, LSTM, BiGRU, Bidirectional LSTM, ConvLSTM, as well as LSTM and GRU with the attention mechanism (termed LSTM-A and GRU-A).

The following section presents the preliminary material needed in this study and briefly introduces the IMDA-VAE forecasting methodology. In Section II, the performances of the considered methods are illustrated using air pollution data from four US states. Finally, in Section IV we conclude this study and sheds light on potential future research lines.

II. Methodology

In this section, we briefly describe the basic concept of vanilla autoencoders and variational autoencoders. Afterward, we briefly describe how self-attention mechanisms work. We then introduce the proposed IMDA-VAE deep learning-driven method. The overall schematic of the proposed forecasting strategy is presented in Figure 1.

A. Variational autoencoder

To better understand the VAE, we first present a short description of the autoencoder (AE). The AE comprises three principal elements, namely the encoder, the latent space, and the decoder (Figure 2). The encoder receives input data, pollutants' concentration time series and projects it into the latent space employing neural networks. Crucially, the pertinent information is stored in the latent space, and the aim is to optimize how data can be scattered in this latent space. On the other hand, the decoder represents a mirror of the encoder because it attempts to reconstruct the compressed latent space's input data. In short, we can summarize the AE procedure using this chain of events.

$$\mathbf{x} \overset{\text{encoder}(\mathbf{x};\theta_e)}{\Longrightarrow} \mathbf{z} \overset{\text{decoder}(\mathbf{z};\theta_d)}{\Longrightarrow} \hat{\mathbf{x}}$$

where, θ_e, and θ_d respectively denote parametrizations of the encoder and decoder networks.

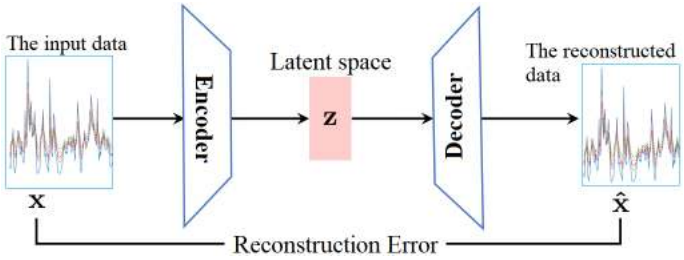

Fig. 2: Basic illustration of the autoencoder.

The AE aims to give a dimensionality reduction procedure, which learns to encode the data such that a distance metric loss is optimized. Essentially, the more effective the encoder is learning data compression, the more accurate the reconstructed data will be. A traditional and straightforward loss function utilized for the AE is the mean squared error, which consists of the squared deviation between the input, X, and output data, $\hat{X}$.

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{X} - X)^2 \tag{1}$$

VAE is a generative model having similar architecture to the vanilla AE. Note that an AE cannot generate new data because it is principally based on encoding input data into discrete values in the latent space. Accordingly, this procedure limits the AE only on memorizing the input data without the possibility for data generation. VAE is a generative model having similar architecture to the vanilla AE. Note that an AE cannot generate new data because it is principally based on encoding input data into discrete values in the latent space. Accordingly, this procedure limits the AE only on memorizing the input data without the possibility for data generation. To alleviate this limitation, the VAE enables encoding the input data into a sampling layer, making the latent space continuous. In this way, the decoder of the VAE can generate new and

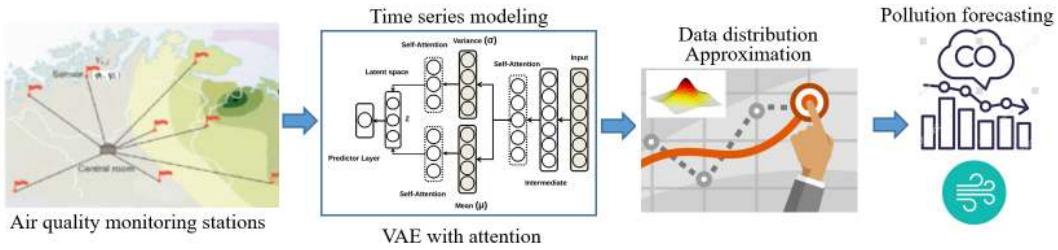

Fig. 1: A Schematic diagram of the overall forecasting stractegy.

realistic data based on the learned features from the latent space.

VAE integrates the desirable features of the variational inference and autoencoder, which enable efficient extraction of releveant low dimensional and hidden features in raw data. VAEs are one of the most powerful and effective class of deep generative models [40]. Basically, this is mainly due to their important advantages and ability of extracting low-dimensional and relevant features from raw data in an unsupervised way [41]. That is, VAE models are trained to learns representations from complex without the need for data labeling. It is very worth mentioning that VAEs offer fundamental properties of dimensionality reduction, which makes them useful for transforming high-dimensional input data into compressed while preserving the essential properties of the original representation. VAEs can be utilized for complex probability distribution approximation using the main results of stochastic gradient descent [40]. Unlike the traditional autoencoders, VAEs overcome the commonly known overfitting problem by introducing a regulation mechanism in the training process. The regularization term improves the capability of the generative models to sample data points using learned data distribution represented in the latent space. Figure 3 illustrates the variational autoencoder architecture, which shows that the VAE is built using a structure consisting of an encoder and a decoder (i.e., two neural networks).

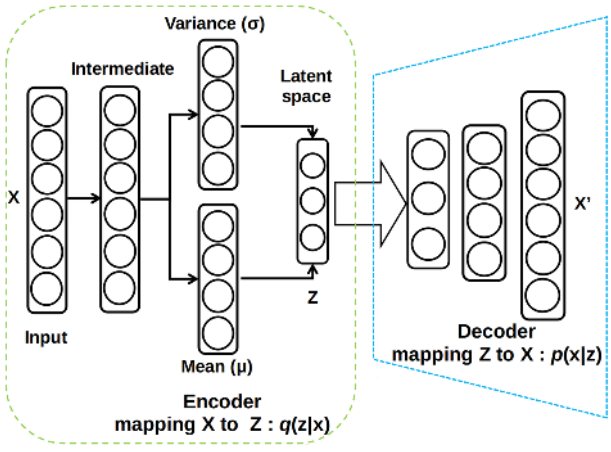

Fig. 3: Variational Autoencoder architecture.

The VAE encoder aims at learning the latent variable $\mathbf{z}$ through measured data acquired by sensors, while the decoder utilized the latent variable obtained $\mathbf{z}$ to ultimately reconstruct the input measurements. The residuals between the original input measurements and the reconstructed data obtained from the decoder should always be as much as possible close to zero. The latent variable $\mathbf{z}$ provided by the encoder is utilized as a feature extractor applied to the input measurements, which allows us to reduce the dimensionality of original data and obtain low dimensional data. The encoder is generally derived by a posterior approximation of $\mathbf{q}_\theta(\mathbf{z}|\mathbf{x})$, while the decoder is obtained with a likelihood $\mathbf{p}_\phi(\mathbf{x}|\mathbf{z})$, where θ and ϕ are the parameters of the encoder and the decoder, respectively.

It should be mentioned that the loss function has a significant effect on the feature extraction for training VAE. Suppose $\mathbf{X}_t = [x_{11}, x_{2t}, \ldots, x_{Nt}]$ is the vector of the input data points of the VAE at time instant t, and $\mathbf{X}_t^{'}$ is the data reconstructed by the VAE model. For maximum likelihood learning of parameters, we may write [41]:

$$log p_\phi(\mathbf{x}^{'}) = D_{KL}\left[q_\theta(\mathbf{z}|\mathbf{x})\middle\|p_\phi(\mathbf{x})\right] + \mathcal{L}(\theta, \phi; \mathbf{x}), \quad (2)$$

where $D_{KL}[.]$ is the Kulback-Leibler divergence, while $\mathcal{L}$ denotes the likelihood of the encoder and decoder parameters θ and ϕ. Hence, the loss function is expressed as:

$$\mathcal{L}(\theta, \phi) = \underbrace{\mathbb{E}_{\mathbf{z}\sim \mathrm{q}_\theta(\mathbf{z}|\mathbf{x})}\left(log \mathrm{p}(\mathbf{x}^{'}|\mathbf{z})\right)}_{\text{Reconstruction term}} - \underbrace{D_{KL}\left(\mathrm{q}_\theta(\mathbf{z}|\mathbf{x})||\mathrm{p}_\phi(\mathbf{z})\right)}_{\text{Regularization term}}. \quad (3)$$

The main objective is the design of a suitable VAE model for which the reconstruction loss function converges to zero. The strength of the decoder's ability to learn data reconstruction is achieved by the reconstruction term given in (3). The second term of regularization in (3), that defines the Kulback-Leibler divergence, aims at separating the encoder distribution function ($\mathrm{q}_\theta(\mathbf{z}|\mathbf{x})$) and the latent variable ($\mathbf{z}$, $|\mathrm{p}_\phi(\mathbf{z})$). To minimize the loss function with respect to the encoder and decoder parameters, one can use the gradient descent procedure in the training stage. The main reason behind minimizing the loss function is to obtain a regular latent space, z, as well as a satisfactory sampling of the new data points using $\mathbf{z} \sim \mathrm{p}_\phi(\mathbf{z})$ [41].

Now, consider $p_\phi(\mathbf{z}) = \mathcal{N}(z; 0, I)$, thus $q_\theta(\mathbf{z}|\mathbf{x})$ can be expressed as

$$log q_\theta(\mathbf{z}|\mathbf{x}) = log \mathcal{N}(\mathbf{z}; \mu, \sigma^2 I), \quad (4)$$

where μ and σ denote the mean and standard deviation of the approximate posterior, respectively. The latent space z is built based on a deterministic function g parameterized by ϕ and

an auxiliary noise variable $\varepsilon \sim p(\varepsilon)$; i.e., $\varepsilon \sim \mathcal{N}(0, I)$.

$$z = g_\phi(x, \varepsilon) = \mu + \sigma \odot \varepsilon \tag{5}$$

The reconstruction error can be written out as,

$$\begin{aligned} \mathcal{L}(\theta, \phi, \mathbf{x}) &= \frac{1}{2} \sum_i \left(1 + log((\sigma_i)^2) - (\mu_i)^2 - (\sigma_i)^2\right) \\ &+ \frac{1}{L} \sum_{l=1}^{L} log(p_\phi(\mathbf{x}|\mathbf{z}^{(l)})) \end{aligned} \tag{6}$$

where the $\odot$ denotes the element-wise product.

The VAE parameters are optimized iteratively based on a gradient-descent algorithm with the Adam optimizer [42]. For more details about VAE, see [43]. The VAE is trained via the procedure given in Algorithm 1.

Algorithm 1: Variational autoencoder training algorithm

Input: : Training dataset $X = \{x^1, \ldots, x^k\}$
Output: : $\{\theta, \phi\}$
θ: Encoder parameters;
ϕ: Decoder parameters;
M: number of mini-batch (drawn from full dataset);
$\{\theta, \phi\} \longleftarrow$ Initialize model parameters randomly;
O_l: is the output of the l variable layer;
$UpdateModelPrameters$: Admam optimizer has been used [42];
repeat
  $X_m \longleftarrow RandomMinibatch(X, M)$;
  $O_{intermediate} \longleftarrow Layer_{intermediate}(X_m)$;
  $O_\sigma = Layer_\sigma(X_m)$;
  $O_\mu = Layer_\mu(X_m)$;
  Draw samples from $\epsilon \sim \mathcal{N}(0, 1)$;
  $O_z = f_\phi(X_m, \varepsilon) = O_\mu + O_\sigma \odot \varepsilon$ (Equation (5));
  $\mathcal{G} = \mathcal{L}(\theta, \phi, X_m)$ (Equation (6));
  $\{\theta, \phi\} \longleftarrow$ Update parameters using gradients via a gradient-descent algorithm;
until *convergence of parameters* $\{\theta, \phi\}$;

It is worth noting that VAE has been used recently to forecast time-series data in many applications, including solar power production forecasting [44], traffic flow forecasting [45], stock market prediction [46], forecasting of COVID-19 spread [47], [48]. It has also demonstrated good performance in anomaly detection and classification [49]–[51].

B. Attention mechanism

The attention mechanism's key idea is to try mimicking human behavior by focusing on some particular areas. For example, when looking at an image, the brain pays more attention to a specific region of interest. The attention mechanism was inspired from the human visual process and adapted to neural machine translation first [52] and images processing [53]. Specifically, during the training phase, the primary purpose is to concentrate on specific features through a weighted sum approach represented by a context vector. The context vector $\mathbf{V}$ is expressed at time t as follows:

$$\mathbf{V}_t = \sum_t \alpha_\mathbf{t} \mathbf{h_t}, \tag{7}$$

where h_t represents the hidden states provided by the model that feeds the attention model, a recurrent network is usually utilized. The term α_t denotes the normalized attention model weights computed as follows:

$$\alpha_\mathbf{t} = softmax(\mathbf{e_t}), \tag{8}$$

The e_t refers to the attention model weights (also called alignment score), which is computed via a feed-forward neural network [52], conditioned on the previous hidden state h_{t-1}:

$$\mathbf{e_t} = \sigma(\mathbf{W_a h_{t-1}} + \mathbf{b_a}), \tag{9}$$

With (W_a,b_a) are weights matrix and bias vector of the attention model computed during the training. Indeed, the attention model context vector $\mathcal{V}$ is a dynamic representation of the relevant part of the time-series input at time t. It is viewed as a weighted sum that highlights the importance of some data position in the sequence using normalized the attention model weights that can be interpreted as a probability. Attention enables the model to focus on essential pieces of the feature space; in other words, it allows the learning to emphasize a particular area of the data sequence through reviewing its memory during the prediction time.

It is worth pointing out that there are two types or modes of Attention: additive [52] (Equation 9) and multiplicative [54]. The main difference between them is how to compute the alignment score:

$$\mathbf{e_t} = \mathbf{W_a} \cdot \mathbf{h_t}. \tag{10}$$

Additive Attention utilizes a linear combination of hidden states through one neural network layer, specifically a feed-forward with by default hyperbolic tangent (tanh) nonlinearity as an activation function. On the other hand, multiplicative attention computes the attention scores by reducing the hidden states using matrix multiplications; other activation functions with learnable parameters can be applied to the dot product.

C. Self-attention mechanism

The self-attention mechanism improves the attention mechanism by reducing external information dependence (source to target) and capturing the internal correlation of input data [55]. An essential characteristic of self-attention is its flexibility to be applied to any layer representing a data sequence like a time-series, which improves the internal input structure's learning by focusing on the relationship between elements of the same input (sequence). Furthermore, self-attention is based on the same principle of the attention mechanism. Specifically, the model generates a new representation of the features space through a weighted sum of extracted features using only one data sequence as input compared to the attention mechanism. Indeed, the self-attention data processing start by computing the weights (called the score) between data point in position

i and j for a given input data sequence $\mathcal{X}$, as follow [55]:

$$\mathcal{E}_{ij} = \frac{(\mathbf{W_a}\mathcal{X}_\mathbf{i})^T(\mathbf{W_a}\mathcal{X}_\mathbf{j})}{\sqrt{d}}, \tag{11}$$

where $\mathbf{W_a}$ is the weight matrix of the self-attention model computed during the training, and d is the dimension of $(\mathbf{W_a}\mathcal{X}_i)$, the division by d makes the convergence faster. Normalization of the weights $\mathcal{E}_{ij}$ is performed to represent them as a probability (sum of all weights values equals to 1), using a *softmax* transformation:

$$\mathcal{A}_{ij} = softmax(\mathcal{E}_{\mathbf{ij}}) = \frac{\exp(\mathcal{E}_{ij})}{\sum_j \exp(\mathcal{E}_{ij})}. \tag{12}$$

The final self-attention model output is expressed as:

$$O_i = \sum_{j=1}^{n} \mathcal{A}_{\mathbf{ij}}(\mathbf{W_a}\mathcal{X}_\mathbf{i}). \tag{13}$$

This output enhances the quality of the extracted features effectively and describes, more specifically, the internal correlation of the input elements.

D. Integrated Multiple Directed Attention-based Variational Autoencoder

This paper introduces an integrated multiple directed attention-based deep learning (called IMDA-VAE) based on the traditional VAE and the attention mechanism for improved ambient air pollution forecasting. The proposed IMDA-VAE approach extends the VAE model's capability and improves the forecasting accuracy compared to the uni-directional and bidirectional recurrent neural network models. Specifically, we introduced the self-attention mechanism at multi-levels to the encoder part of the traditional VAE (Figure 4). Indeed, attention as non-linear transformation improves modeling and forecasting quality using weighted features vector. Moreover, the attention's technique was usually incorporated in the decoder side [52], [53], to map the sequence of images to their text description sequence (as caption). Indeed, the forecasting problem can be viewed as a sequence to sequence mapping; in the univariate case, the models learning aims to map a sequence of pollutant measurements to the next concentration value of a given pollutant. While multivariate cases, especially with multiple outputs, maps long sequences, including all pollutants measurements, to forecast short sequences. This principle of recurrent neural networks is based on one-step supervised learning. However, VAE is a composite model composed of an encoder and decoder, where the main is to learn the approximation of the training dataset probability distribution in an unsupervised approach. It is expected that the incorporation of the robust variation inference approach, an efficient regularization with an enforced attention mechanism, will improve the univariate and multivariate forecasting performances.

The historical pollutants measurements are processed first through a non-linear transformation based on a dense layer (called intermediate) into a continuous representation. Furthermore, this new representation is passed through a self-attention layer that emphasizes the internal correlation within

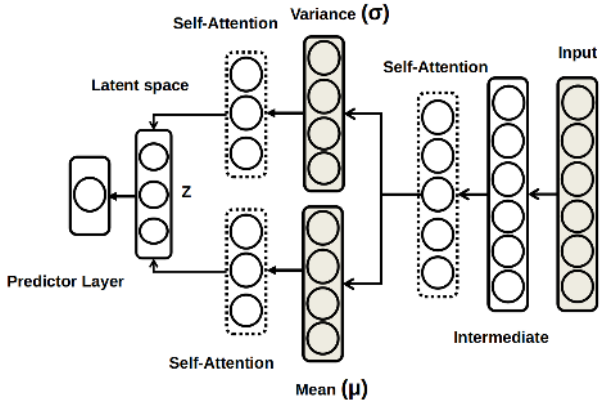

Fig. 4: The proposed approach: Variational autoencoder with Attention

the pollutants' time-series via computing a context vector, representing a weighted sum of features. Moreover, regularization aims to prevent the over-fitting problem during the learning of pollutants dynamics by adding information and improving the learning quality by reducing error and penalizing the loss function.

Note that the $L2$ *regularization* is used for weights normalization, while *bias-regularizer:* $L1$ bias extenuation [56]. Moreover, the regularization process is used to guarantee the weighted sum diversification. The continuous representation and the context vector (weighted sum) feed the covariance matrix σ and the mean μ of the regularized data distributions. Furthermore, the regularization is applied to encouraging the distributions to be closest to a standard normal distribution and enforce covariance matrix to be close to the identity. Finally, the latent space is constructed after duel self-attention transformation of the regularized mean and variance, a set of data points is sampled from the latent space (see Equation 3) to be reconstructed by the decoder model. In the proposed approach, the decoder is a deep fully connected neural network; it represents the reverse path used to train the encoder. Here, the loss is measured via Kullback-Leibler (KL), representing the divergence between the training data and learned probability distribution. Furthermore, KL is key to monitor the model parameters convergence, where usually once its value decrease (close to zero) and stabilizes, the training can be stopped. During the training, the reconstruction error is back-propagated, and the model parameters are updated accordingly. Once the training is done, the latent space is used to forecast the next values (measurement) of the pollutants in question depending on the configuration used (univariate or multivariate)

In contrast to the recurrent models, the proposed approach training is performed in an unsupervised manner. Specifically, the model learns first the data distribution of the considered pollutant and approximates it using the model parameters to sample new data points from a features space (also called the latent space); the data points shared the same features of real data points. Moreover, the proposed approach model parameters are adjusted and optimized through a fine-tuning process based on supervised learning, aiming to learn the

mapping of a given data sequence of pollutant measurement to its next value. Notably, the forecasting is done at the latent space level (encoding space). In other words, the last layer of the IMDA-VAE acts as a forecasting layer.

III. Experiments and Results

This section describes data used in the present study, with full details about experimental settings and execution procedure. Moreover, we provide an analysis of obtained results and discussion.

A. Data description

The dataset used in this study was collected by the United States Environmental Protection Agency, in which the concentration level of several ambient pollutants is recorded daily in different states. The datasets are publicly accessible on the website "https://www.epa.gov/outdoor-air-quality-data/download-daily-data". The pollutants measurements were collected during a period of 16 years exactly (2000-2016). This study focuses on four significant pollutants, namely Nitrogen Dioxide (NO_2), Sulphur Dioxide (SO_2), Carbon Monoxide (CO), and Ozone (O_3). In our study, we picked measurements from four locations, namely California, Arizona, Texas, and Pennsylvania. It should be mentioned that each state contains numerous air quality monitoring stations; however, we selected one station per state for measurement collection.

The descriptive statistics of the Arizona ambient air pollution datasets are listed in Table I. The skewness of the normal distribution (or any perfectly symmetric distribution) is zero, and the kurtosis of the normal distribution is 3. We can conclude from Table I that the Arizona ambient air pollution time-series datasets are non-Gaussian distributed with positive support and exhibit different intervals of variability.

TABLE I: Statistics summary of the Arizona ambient air pollution datasets.

metric	NO2	O3	SO2	CO
mean	55.736	0.111	2.909	1.249
std	26.193	0.042	2.935	0.635
min	3.339	0.013	-0.117	0.025
25%	35.500	0.079	0.752	0.809
50%	52.167	0.112	2.153	1.123
75%	72.667	0.142	4.133	1.550
max	211.933	0.283	30.475	5.025
kurtosis	1.159	-0.455	7.845	2.901
skew	0.828	0.142	2.193	1.385

Monthly distributions of concentration levels of the four ambient pollutants from California are illustrated in Figure 5(a-d). Figure 5(c) shows that the highest O_3 concentration levels can be seen in the summer season due to the local photochemical production. Also, Figure 5 (a) and (c) show that NO2 is negatively correlated with the variation of O_3 concentration levels. This fact is because O_3 is formed through the photochemical destruction of nitrogen dioxide (NO_2) under sunlight. Similarly, we can see the presence of a negative correlation between O3 and CO because of photochemical production of O_3 principally from the oxidation of natural and anthropogenic hydrocarbons, carbon monoxide (CO), and methane (CH4) by hydroxyl (OH) radical in the availability of enough quantity of NOx. Of course, O_3 is adversely correlated with O_3 precursors (i.e., NO_2 and CO) [57].

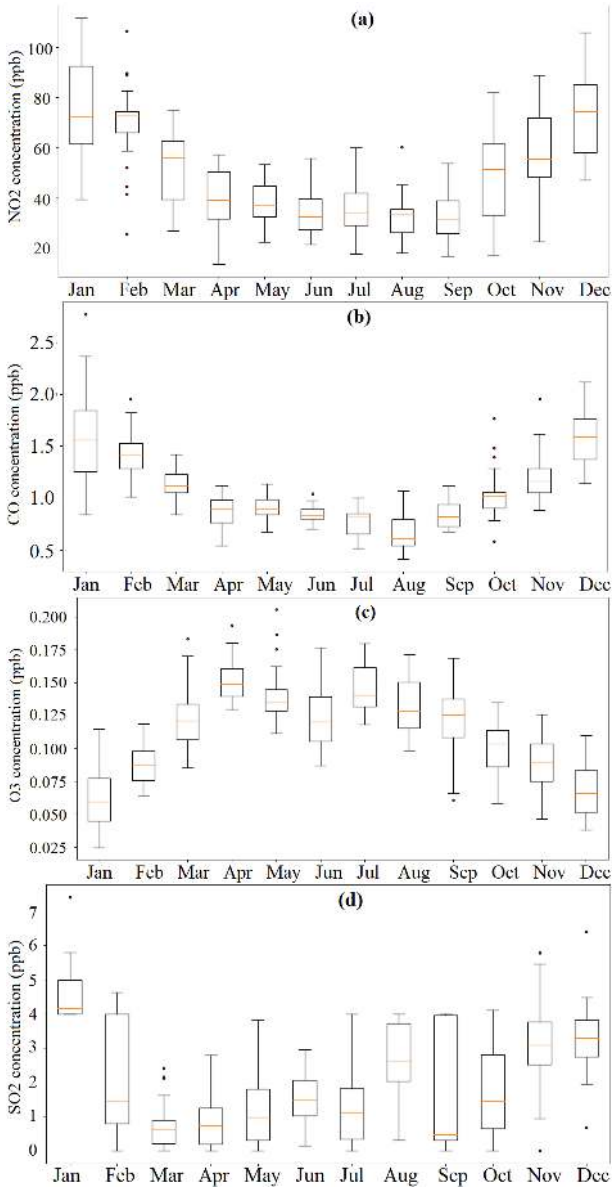

Fig. 5: Monthly distribution of concentration levels of (a) NO_2, (b) CO, (c) O_3, and (d) SO_2 in California.

Figure 6 presents the correlation coefficients between the NO_2, CO, O_3, and SO_2 recorded respectively in Arizona, California, Pennsylvania, and Texas. Figure 6 shows the presence of a weak negative correlation between O_3 and its precursors (NO_2 and CO), and a moderate positive correlation between precursors, NO_2 and CO. Besides, distinct patterns regarding the SO_2 and the other pollutants are obtained for each station. Interpretation of the relation between pollutants is relatively difficult because of their dependence on several factors, including meteorological variables and the transportation phenomena. Specifically, sometimes pollution concentration can be high due to transported pollution produced elsewhere in

the region. Various studies have been reported in the literature to investigate the correlation between different ambient air pollutants [57], [58].

B. Experiments settings

To evaluate the performance of the proposed approach in forecasting air ambient pollutants, several experiments are conducted to show its superiority and efficiency through a comparative study including powerful recurrent neural network models, namely GRU, LSTM, BiGRU, BiLSTM, and ConvLSTM. We also compared the performance of the proposed IMDA-VAE approach to that GRU and LSTM with the attention mechanism (LSTM-A and GRU-A). Importantly, these models with attention are designed by stacking three components. For instance, in LSTM-A, LSTM is used to model time dependencies and extract temporal features, followed by the attention module that aims to improve and highlight relevant extracted features, and finally, we added a forecaster layer represented by a fully connected layer. In a similar way to LSTM-A, we designed GRU-A. Here, both LSTM-A and GRU-A are trained through supervised learning. We used the same settings (hyper-parameters) used for all models: 300 epochs, a learning rate of 0.0001, batch size of 250, 32 hidden units, Rmsprop optimizer, and Binary cross-entropy as a lost function. To demonstrate the advantage of the proposed IMDA-VAE compared to the traditional deep learning models, we consider the following experimentations in this study.

- **Univariate forecasting**: In this experiment, each pollutant is modeled and forecasted individually, which means that five models are trained to forecast the next values based on only the measurement of the pollutant in question.
- **Multivariate forecasting with one output (prediction)**: The forecasting of each pollutant is done using all other pollutants, which means that the training aims to learn how to predict the next value of a given pollutant based on a sequence of measurements of the four other pollutants. In the end, a forecasting model dedicated to a pollutant with multiple inputs is obtained.
- **Multi-variates forecasting with multi-output**: This experimentation can be seen as a one-shots task, where all pollutants are forecasted based on all historical measurements of all pollutants. In the end, only one forecasting model for all pollutants with multiple inputs to forecast all pollutants next values is derived.

C. Measurements of effectiveness

To assess the forecast precision and compare the models, some validation metrics like Coefficient of determination (R^2), Root Mean Square Error (RMSE), mean absolute error (MAE), explained variance (EV), mean absolute percentage error (MAPE), Mean bias error (MBE), and Relative Mean bias error (rMBE) are used (Table II); where y_t is concentration level of a pollutant, $\hat{y}_t$ is its corresponding forecasted values, and n is the number of data points [59], [60]. The more precise forecasting is, the lower RMSE, MAE, MBE and rMBE values and high R^2, EV, and MAPE values are.

TABLE II: Definition of measurements of effectiveness.

Metric	Definition
R^2	$\frac{\sum_{t=1}^{n}[(y_{t,}-\bar{y})\cdot(\hat{y}_t-\bar{y})]^2}{\sqrt{\sum_{t=1}^{n}(y_t-\bar{y})^2}\cdot\sqrt{\sum_{t=1}^{n}(\hat{y}_t-\bar{y})^2}}$
RMSE	$\sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t-\hat{y}_t)^2}$
MAE	$\frac{\sum_{t=1}^{n}\lvert y_t-\hat{y}_t\rvert}{n}$
MAPE	$\frac{100}{n}\sum_{t=1}^{n}\left\lvert\frac{y_t-\hat{y}_t}{y_t}\right\rvert\%$
EV	$1-\frac{\mathrm{Var}(\hat{\mathbf{y}}-\mathbf{y})}{\mathrm{Var}(\mathbf{y})}$
MBE	$\frac{1}{n}\sum_{t=1}^{n}(\hat{y}_t-y_t)$
rMBE	$\frac{\sum_{t=1}^{n}(\hat{y}_t-y_t)}{\sum_{t=1}^{n}y_t}.100$

D. Results Analysis

The first set of experiments aims to analyze the performance of the investigated models in forecasting the concentration levels of the four pollutants separately. Towards this end, in this univariate forecasting, each model is trained in a supervised way to learn temporal-dependencies included in time-series data measurement of each pollutant. Indeed, the objective is to learn the prediction of the next value of a given measurement sequence. This study has been implemented using a fast algorithm-based CPU Intel i7 with 12Go RAM. We used Python 2.7.0 with Keras 2.3And TensorFlow 2.0 under Ubuntu 18 LTS. The training set consists of daily concentration levels of five ambient pollutants (NO_2, CO, O_3, and SO_2) from January 2000 through December 2014. At first, we normalize the training data, $\mathbf{e}$, by min-max normalization within the interval [0, 1] and then used it for models construction. The normalization is performed as follows,

$$\widetilde{y} = \frac{(y - y_{min})}{(y_{max} - y_{min})} \tag{14}$$

where y_{min} and y_{max} denotes the minimum and maximum of the original data, respectively. This procedure is reversed after the forecasting process.

Here, the k-fold cross-validation technique with $K = 10$, as recommended in [61], [62], is adopted in this study to build the forecasting models. Using the training datasets in these experimentations, the set of hyperparameters are fixed for all models used, namely optimizer='rmsprop', loss function='Cross Entropy', batch size=250, epochs=200, and learning rate = 0.001. For RNNs hidden units, we used hidden units=32, while in the IMDA-VAE model for all layers (intermediate, mean, variance), 12 is used as the number of hidden units. The proposed model contains nine layers, its configuration in terms of hidden units per layer $< 3, 12, 12, 15, 15, 15, 15, 3, 1 >$. This configuration was determined using grid search method; we also used for the training a batch size = 250, learning rate = 10^{-3}, number of epochs = 200, optimizer: Rmsprop, and loss

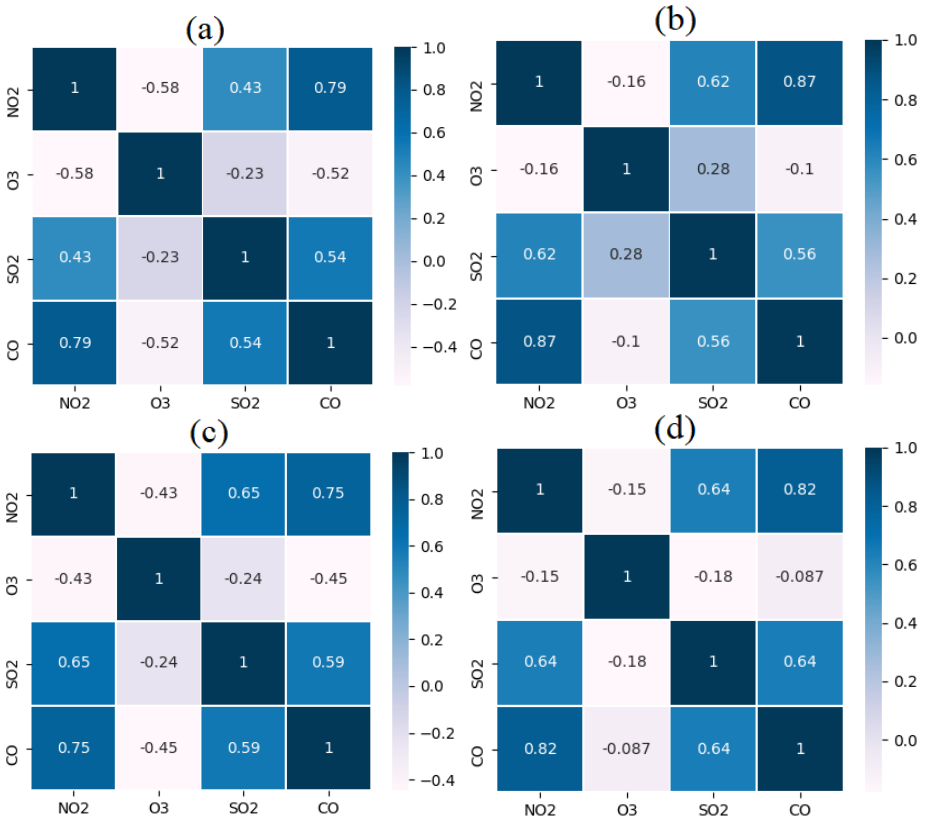

Fig. 6: Pairwise correlation matrix between ambient air pollutants (NO_2, CO, O_3, and SO_2) obtained in (a) Arizona, (b) California, (c) Pennsylvania, and (d) Texas.

function: binary cross-entropy. For the fine-tuning step, the standard backpropagation algorithm has been used to adjust model parameters [63]–[65].

The testing period is from January 1st to November 30th, 2015. For a visual illustration, the observed test set together with model forecasts using the five studied models are charted in Figure 7 for NO_2, CO, O_3, and SO_2 from Arizona. Results from the other stations are omitted because they give relatively similar results. Figure 7 shows relatively narrower bands around the actual observations, which indicates a good forecast quality. One exception is from SO_2 time-series data, which indicates the deviation of deep learning model forecasts. From Figure 7, the proposed IMDA-VAE approach shows a good ability in forecasting the future trends of SO_2 concentration dynamics.

Tables III, IV, V, and VI list the obtained validation metrics of testing data from California, Arizona, Texas, and Pennsylvania, respectively. In terms of all metrics calculated, IMDA-VAE is the best approach for this univariate time series forecasting problem with high efficiency and satisfying accuracy. The proposed IMDA-VAE approach outperforms the recurrent models (i.e., GRU, LSTM, BiGRU, ConvLSTM, LSTM-A and GRU-A) on forecasting all investigated pollutants (i.e., NO_2, O_3, SO_2, and CO) measured from four locations, namely California, Arizona, Texas, and Pennsylvania. As expected, the proposed IMDA-VAE approach achieved the lowest forecasting errors (RMSE, MAE, MBE and RMBE) and the highest score of R2 and EV. It could be attributed to its capacity to model time-dependencies and select relevant features using the attention mechanism. We should highlight that the performance of Bidirectional recurrent networks, namely the BiLSTM and the BiGRU, are superior to the uni-directional models represented by LSTM and GRU. This is mainly due to the ability of Bidirectional models in processing data in two directions, forward and backward.

Table III shows the performance comparison of the considered models based on California pollutants measurements. The proposed approach has accounted for more than 95% of the variability of *NO2* which is the best R2 and EV, while lowest RMSE=9.484, MEA=7.153, MBE=0.122, RMBE=-2.763 were recorded. Also, it can be seen that the Bi-directional data processing performed by BiLSTM and BiGRU model is better time-dependencies when forecasting *NO2* (i.e., R2=93% and EV=94%) for (R2, and EV) compared to unidirectional recurrent models like LSTM and GRU (R2=89% and EV=91%).

Similar conclusions hold true also for forecasting concentration levels of O_3, around 95% of the variability was accounted for by using the proposed approach and followed by BiLSTM with 94%, while the remaining models scored 93%. Here also MPAE=7% was recorded by the proposed approach demonstrating its high performance, while the other models recorded MPAE=8%. Note that similar performance was recorded by the considered models when forecasting *CO* concentration levels. On the other hand, the validation

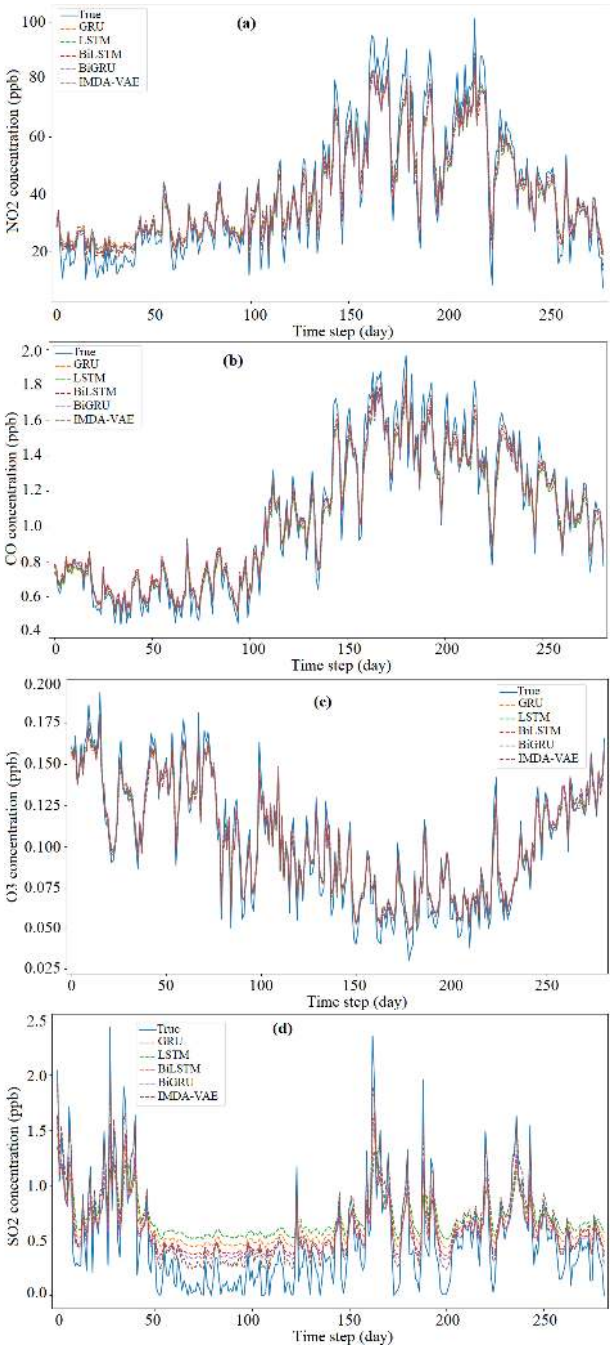

Fig. 7: Records and forecasts using the five models of (a) NO_2, (b) CO, (c) O_3, and (d) SO_2 from Arizona.

TABLE III: Performance comparison of the proposed using pollution testing datasets from California.

Polluant	Model	RMSE	MAE	R2	EV	MAPE (%)	MBE	RMBE(%)
NO2	GRU	13.617	11.026	0.902	0.913	22	-4.661	-5.875
NO2	LSTM	13.775	11.452	0.899	0.915	24	-5.526	-6.89
NO2	BiGRU	11.421	9.334	0.931	0.943	19	-4.868	-6.119
NO2	BiLSTM	10.997	8.985	0.936	0.946	18	-4.393	-5.555
NO2	ConvLSTM	16.005	12.755	0.865	0.871	24	-3.502	-4.487
NO2	VAE	11.04798	8.69401	0.93518	0.94905	17	-0.12203	-2.76251
NO2	GRU-A	15.727	13.004	0.869	0.909	28	-8.733	-10.469
NO2	LSTM-A	17.941	15.484	0.829	0.879	34	-9.739	-11.536
NO2	IMDA-VAE	9.484	7.153	0.952	0.955	13	-2.47718	-3.21039
O3	GRU	0.025	0.019	0.938	0.941	8	0.006	1.739
O3	LSTM	0.025	0.020	0.936	0.941	8	0.008	2.412
O3	BiGRU	0.025	0.019	0.938	0.942	8	0.006	1.798
O3	BiLSTM	0.023	0.018	0.945	0.945	8	-0.001	-0.366
O3	ConvLSTM	0.028	0.022	0.921	0.921	9	-0.001	-0.374
O3	VAE	0.02191	0.01656	0.95153	0.95371	7	-5.11047	-6.40452
O3	GRU-A	0.024	0.018	0.942	0.944	8	0.004	1.259
O3	LSTM-A	0.03	0.024	0.911	0.922	9	0.011	3.352
O3	IMDA-VAE	0.021	0.017	0.954	0.954	7	0.00171	0.52843
SO2	GRU	1.126	0.894	0.861	0.868	13	-0.261	-2.971
SO2	LSTM	1.242	0.971	0.830	0.835	14	-0.208	-2.383
SO2	BiGRU	0.919	0.734	0.907	0.924	11	-0.388	-4.347
SO2	BiLSTM	0.817	0.645	0.927	0.928	9	-0.129	-1.491
SO2	ConvLSTM	1.207	0.973	0.841	0.844	14	-0.156	-1.801
SO2	VAE	1.48731	1.41898	0.75645	0.97814	20	-0.00467	-1.41316
SO2	GRU-A	1.853	1.529	0.622	0.635	23	-0.34	-3.829
SO2	LSTM-A	1.95	1.602	0.581	0.592	23	-0.306	-3.462
SO2	IMDA-VAE	0.590	0.470	0.962	0.962	6	-0.07146	-0.83079
CO	GRU	0.248	0.189	0.942	0.945	5	-0.063	-1.438
CO	LSTM	0.259	0.195	0.936	0.937	5	-0.036	-0.832
CO	BiGRU	0.262	0.225	0.935	0.956	6	-0.152	-3.41
CO	BiLSTM	0.235	0.197	0.947	0.968	5	-0.147	-3.302
CO	ConvLSTM	0.289	0.208	0.920	0.921	5	-0.021	-0.483
CO	VAE	0.2141	0.1709	0.95637	0.97054	4	-1.41898	-14.26337
CO	GRU-A	0.39	0.296	0.855	0.861	7	-0.078	-1.781
CO	LSTM-A	0.47	0.361	0.79	0.801	8	-0.107	-2.437
CO	IMDA-VAE	0.183	0.133	0.968	0.969	3	0.01262	0.29467

TABLE IV: Performance comparison of the proposed using pollution testing datasets from Texas.

Polluant	Model	RMSE	MAE	R2	EV	MAPE (%)	MBE	RMBE (%)
NO2	GRU	14.402	12.016	0.771	0.826	33	-7.065	-10.622
NO2	LSTM	14.896	12.345	0.755	0.812	35	-7.178	-10.773
NO2	BiGRU	13.522	11.535	0.798	0.848	31	-6.711	-10.144
NO2	BiLSTM	13.268	11.34	0.806	0.858	31	-6.854	-10.338
NO2	ConvLSTM	14.175	11.667	0.78	0.802	31	-4.45	-6.948
NO2	VAE	14.33501	12.07114	0.77316	0.85095	34	-0.01113	-0.86821
NO2	GRU-A	15.254	12.736	0.743	0.782	35	-5.963	-9.117
NO2	LSTM-A	15.756	13.2	0.726	0.772	37	-6.443	-9.778
NO2	IMDA-VAE	12.373	10.370	0.831	0.86717	28	-5.4587	-8.41001
O3	GRU	0.017	0.014	0.839	0.898	20	-0.01	-8.599
O3	LSTM	0.017	0.013	0.847	0.895	19	-0.009	-7.928
O3	BiGRU	0.016	0.013	0.868	0.909	18	-0.009	-7.251
O3	BiLSTM	0.015	0.012	0.88	0.921	17	-0.009	-7.326
O3	ConvLSTM	0.015	0.012	0.879	0.893	16	-0.005	-4.422
O3	VAE	0.01304	0.01063	0.90873	0.94324	15	-8.39428	-12.37314
O3	GRU-A	0.016	0.013	0.862	0.881	18	-0.006	-5.123
O3	LSTM-A	0.017	0.013	0.851	0.868	19	-0.006	-4.884
O3	IMDA-VAE	0.013	0.010	0.91133	0.93874	15	-0.00709	-6.0628
SO2	GRU	1.198	0.867	0.76	0.78	30	-0.344	-14.275
SO2	LSTM	1.407	1.078	0.669	0.705	36	-0.462	-18.28
SO2	BiGRU	1.123	0.786	0.789	0.8	20	-0.255	-10.976
SO2	BiLSTM	1.161	0.818	0.775	0.785	21	-0.251	-10.847
SO2	ConvLSTM	1.353	0.967	0.697	0.704	30	-0.215	-9.367
SO2	VAE	1.25142	0.93551	0.73823	0.7601	29	-0.00795	-6.75241
SO2	GRU-A	1.689	1.324	0.523	0.559	42	-0.461	-18.25
SO2	LSTM-A	1.797	1.354	0.46	0.493	42	-0.442	-17.626
SO2	IMDA-VAE	1.087	0.701	0.806	0.810	19	-0.12927	-5.8888
CO	GRU	0.296	0.213	0.867	0.87	22	-0.047	-3.577
CO	LSTM	0.317	0.241	0.847	0.85	25	-0.047	-3.553
CO	BiGRU	0.286	0.207	0.876	0.88	20	-0.054	-4.101
CO	BiLSTM	0.275	0.203	0.885	0.889	20	-0.051	-3.876
CO	ConvLSTM	0.321	0.237	0.844	0.847	23	-0.041	-3.135
CO	VAE	0.2849	0.20845	0.87659	0.87678	21	-0.36169	-14.90242
CO	GRU-A	0.433	0.326	0.715	0.717	33	-0.043	-3.307
CO	LSTM-A	0.411	0.315	0.743	0.746	33	-0.046	-3.496
CO	IMDA-VAE	0.241	0.169	0.911	0.918	17	0.03558	-2.72316

metrics of SO_2 forecast values from the proposed indicate a good forecast quality by achieving an R2 of 0.96 and low forecasting error (i.e., MAPE=6%, RMSE=0.59, and MAE =0.47). Bidirectional (LSTM and GRU) models perform better than unidirectional models by reaching an MPAE of 9% for BiLSTM and 11% for BiGRU. Similar conclusions are obtained when using data from the other states.

The following experiments are devoted to a comparative assessment for predicting concentration levels of a single pollutant by using the remaining pollutants as multiple inputs as presented in Section III-B. Results in terms of R2, EV and MAPE metrics are presented in Figures 8, 9, 10, and 11. Results confirm the superior performance of the proposed IMDA-VAE approach again compared to the traditional VAE without attention and the other investigated deep learning models by achieving the highest (R2, EV) and the lowest mean error on all experimentations. It is also interesting to see better outcomes from the Bidirectional models (BiLSTM and BiGRU) than uni-directional models (LSTM and GRU). Results confirm that the IMDA-VAE approach clearly outperforms the LSTM-A and GRU-A models. Overall, results

TABLE V: Performance comparison of the proposed using pollution testing datasets from Pennsylvania.

Polluant	Model	RMSE	MAE	R2	EV	MAPE (%)	MBE	RMBE(%)
NO2	GRU	17.82	14.00	0.769	0.777	21	3.483	4.652
NO2	LSTM	18.03	14.41	0.763	0.768	22	2.47	3.255
NO2	BiGRU	17.34	13.35	0.781	0.797	19	4.745	6.446
NO2	BiLSTM	18.10	13.72	0.762	0.778	19	4.755	6.461
NO2	ConvLSTM	19.44	15.13	0.728	0.744	22	4.739	6.426
NO2	VAE	15.95685	13.07987	0.81454	0.81504	22	-0.0555	-4.11631
NO2	GRU-A	21.368	17.128	0.667	0.714	23	8	11.371
NO2	LSTM-A	28.854	23.033	0.394	0.416	34	5	7.698
NO2	IMDA-VAE	15.97	12.86	0.814	0.815	19	0.09919	0.12675
O3	GRU	0.014	0.012	0.882	0.883	29	0.001	1.797
O3	LSTM	0.014	0.011	0.882	0.883	29	0.001	1.697
O3	BiGRU	0.014	0.011	0.888	0.892	27	0.003	3.343
O3	BiLSTM	0.014	0.011	0.884	0.888	27	0.002	3.15
O3	ConvLSTM	0.016	0.013	0.850	0.860	32	0.004	5.625
O3	VAE	0.01304	0.01066	0.89803	0.90148	23	-0.82783	-1.0455
O3	GRU-A	0.015	0.012	0.869	0.885	28	0.005	7.105
O3	LSTM-A	0.017	0.013	0.832	0.839	31	0.004	4.857
O3	IMDA-VAE	0.012	0.010	0.916	0.916	26	0.00054	0.699
SO2	GRU	2.86	2.17	0.677	0.782	7	1.624	21.939
SO2	LSTM	2.88	2.10	0.672	0.745	38	1.359	17.728
SO2	BiGRU	2.53	1.96	0.747	0.881	25	1.843	25.646
SO2	BiLSTM	2.49	1.89	0.754	0.866	27	1.682	22.904
SO2	ConvLSTM	3.37	2.56	0.556	0.721	40	2.05	29.255
SO2	VAE	1.75778	1.49462	0.87788	0.92761	40	0.00242	3.1831
SO2	GRU-A	4.372	3.234	0.244	0.485	51	2.466	37.584
SO2	LSTM-A	4.615	3.373	0.158	0.399	54	2.47	37.662
SO2	IMDA-VAE	1.54	1.11	0.906	0.906	25	-0.07924	-0.8701
CO	GRU	0.22	0.14	0.886	0.898	11	0.071	5.784
CO	LSTM	0.24	0.16	0.866	0.876	13	0.066	5.372
CO	BiGRU	0.22	0.13	0.889	0.899	9	0.068	5.561
CO	BiLSTM	0.22	0.13	0.887	0.917	8	0.113	9.565
CO	ConvLSTM	0.29	0.17	0.804	0.829	11	0.103	8.618
CO	VAE	0.21237	0.16243	0.89473	0.90192	16	-1.12169	-11.05221
CO	GRU-A	0.395	0.276	0.636	0.705	19	0.172	15.305
CO	LSTM-A	0.448	0.311	0.531	0.608	22	0.182	16.349
CO	IMDA-VAE	0.18	0.12	0.920	0.920	9	-0.00072	-0.0555

TABLE VI: Performance comparison of the proposed using pollution testing datasets from Arizona.

Polluant	Model	RMSE	MAE	R2	EV	MAPE (%)	MBE	RMBE (%)
NO2	GRU	6.49	5.449	0.914	0.914	20	-0.248	-0.609
NO2	LSTM	6.298	5.221	0.919	0.919	18	0.214	0.533
NO2	BiGRU	6.028	4.949	0.926	0.926	7	0.348	0.867
NO2	BiLSTM	5.535	4.432	0.937	0.938	15	0.496	1.241
NO2	ConvLSTM	7.375	5.756	0.889	0.895	18	1.728	4.45
NO2	VAE	5.94335	5.05048	0.92761	0.94932	21	-0.02125	-1.97247
NO2	GRU-A	7.908	6.809	0.872	0.873	23	0.663	1.668
NO2	LSTM-A	9.022	7.298	0.833	0.834	25	0.631	1.587
NO2	IMDA-VAE	5.361	4.415	0.941	0.942	16	-0.84549	-2.04861
O3	GRU	0.008	0.006	0.957	0.958	8	-0.001	-1.047
O3	LSTM	0.008	0.006	0.958	0.959	8	-0.001	-1.201
O3	BiGRU	0.008	0.006	0.959	0.961	8	-0.002	-1.555
O3	BiLSTM	0.008	0.006	0.957	0.958	8	-0.001	-1.373
O3	ConvLSTM	0.01	0.008	0.929	0.933	10	-0.002	-2.206
O3	VAE	0.00707	0.00568	0.96564	0.96647	7	-3.2545	-7.45071
O3	GRU-A	0.008	0.006	0.959	0.964	8	-0.003	-2.588
O3	LSTM-A	0.007	0.006	0.963	0.966	7	-0.002	-1.911
O3	IMDA-VAE	0.007	0.006	0.965	0.965	7	0.00014	0.13453
SO2	GRU	0.338	0.286	0.465	0.574	34	-0.153	-23.546
SO2	LSTM	0.387	0.334	0.301	0.497	41	-0.205	-29.22
SO2	BiGRU	0.235	0.196	0.742	0.778	24	-0.088	-15.079
SO2	BiLSTM	0.252	0.214	0.703	0.756	27	-0.106	-17.654
SO2	ConvLSTM	0.312	0.27	0.522	0.664	34	-0.17	-25.915
SO2	VAE	0.51242	0.50263	0.9013	0.93812	22	-0.00108	-1.03356
SO2	GRU-A	0.405	0.339	0.233	0.366	39	-0.169	-25.433
SO2	LSTM-A	0.422	0.356	0.166	0.327	42	-0.186	-27.253
SO2	IMDA-VAE	0.201	0.179	0.811	0.921	20	-0.153	-23.58239
CO	GRU	0.082	0.063	0.959	0.96	6	0.016	1.569
CO	LSTM	0.083	0.064	0.957	0.961	6	0.023	2.241
CO	BiGRU	0.071	0.055	0.969	0.969	5	0.009	0.85
CO	BiLSTM	0.064	0.051	0.975	0.975	5	-0.004	-0.356
CO	ConvLSTM	0.086	0.067	0.954	0.955	7	0.008	0.776
CO	VAE	0.07616	0.06187	0.96387	0.96669	7	-0.49933	-50.1772
CO	GRU-A	0.135	0.108	0.887	0.887	11	0	-0.047
CO	LSTM-A	0.112	0.091	0.922	0.923	10	-0.014	-1.331
CO	IMDA-VAE	0.061	0.048	0.977	0.977	5	0.00043	0.04082

testify the superior performance of the IMDA-VAE approach (see Tables VII, VIII, IX, and X).

The final experiment aims to assess the potentials of the IMDA-VAE technique in forecasting several pollutants simultaneously (i.e., multivariate forecasting) by using historical data from all pollutants. Here, we used only ambient pollution data from California, and datasets from other stations are omitted because they give relatively similar results. The major advantage of multivariate forecasting is that by using only one model, forecasting several pollutants can be obtained simultaneously compared to univariate forecasting that requires a model for each time-series data. However, multivariate forecasting is relatively challenging because the cross-correlation among variables and time dependencies of multiple variables need to be modeled. Here, an important variable that can impact the forecasting accuracy is the *timestep*, which consists of the amount of data (in days) used to look back from the past to predict the following values. In this experiment, we evaluate the forecasting performance using different *timestep* values: 3, 6, 9, 12, and 15. Table XI illustrates the comparative forecasting results of the considered models. We can see that our model recorded the best score with the highest (R2, EV) for all experiments, especially for *timestep*=15, it records 0.9, which is a good performance for multivariate forecasting of all pollutants.

Furthermore, in this study, we compared the time cost between the proposed method with attention and without

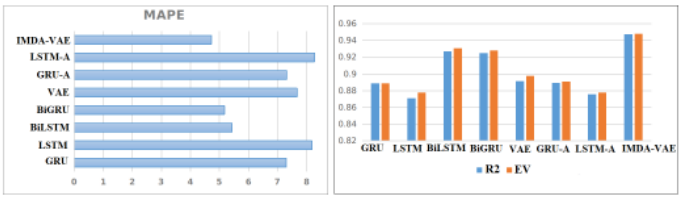

Fig. 8: Validation metrics for multivariate forecasting of CO concentrations.

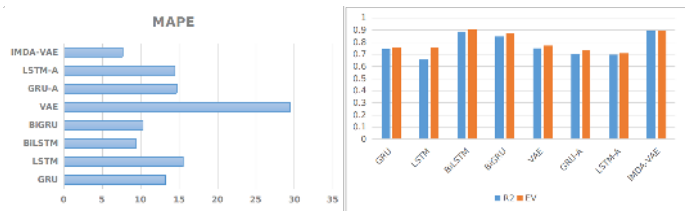

Fig. 9: Validation metrics for multivariate forecasting of SO_2 concentrations.

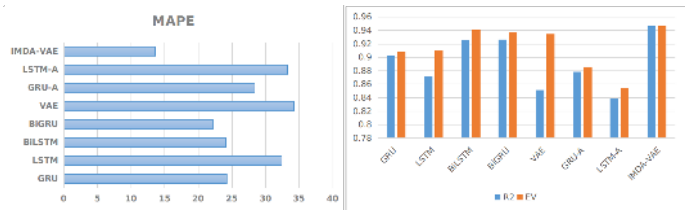

Fig. 10: Validation metrics for multivariate forecasting of NO_2 concentrations.

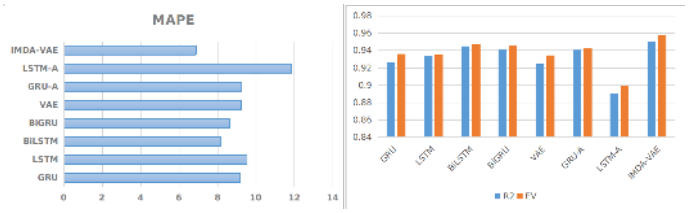

Fig. 11: Validation metrics for multivariate forecasting of O_3 concentrations.

TABLE VII: Performance comparison of the proposed approach, using CO pollutant.

State	Model	RMSE	MAE	R2	EV	MAPE
Arizona	GRU	2.000E-02	1.560E-02	0.9404	0.9404	8.4787
Arizona	LSTM	2.000E-02	1.520E-02	0.9413	0.9417	8.1507
Arizona	BiLSTM	1.730E-02	1.420E-02	0.9478	0.9479	7.3555
Arizona	BiGRU	2.000E-02	1.490E-02	0.9423	0.9428	7.8493
Arizona	VAE	1.702E-02	1.510E-02	0.94081	0.94292	8.78567
Arizona	GRU-A	2.000E-02	0.0152	0.9447	0.9457	8.9539
Arizona	LSTM-A	2.240E-02	0.0182	0.919	0.919	10.8221
Arizona	IMDA-VAE	1.612E-02	1.296E-02	0.95877	0.95896	7.13433
California	GRU	1.72E-02	1.10E-02	0.889	0.889	7.293
California	LSTM	1.76E-02	1.20E-02	0.871	0.878	8.179
California	BiLSTM	1.20E-02	9.00E-03	0.927	0.931	5.436
California	BiGRU	1.20E-02	9.00E-03	0.925	0.928	5.168
California	VAE	1.42E-02	1.00E-02	0.8912	0.89784	7.66479
California	GRU-A	1.73E-02	0.0113	0.8895	0.8908	7.3162
California	LSTM-A	1.73E-02	0.0125	0.8756	0.8777	8.2712
California	IMDA-VAE	1.10E-02	7.70E-03	0.94766	0.94782	4.71853
Pennsylvania	GRU	3.610E-02	2.540E-02	0.7818	0.8372	12.5766
Pennsylvania	LSTM	3.460E-02	2.400E-02	0.8057	0.8504	12.15
Pennsylvania	BiLSTM	3.460E-02	2.300E-02	0.7984	0.8602	10.8596
Pennsylvania	BiGRU	3.460E-02	2.290E-02	0.8063	0.8633	10.8826
Pennsylvania	VAE	3.380E-02	2.100E-02	0.82191	0.82218	14.0252
Pennsylvania	GRU-A	4.240E-02	0.0313	0.703	0.8221	15.2904
Pennsylvania	LSTM-A	4.120E-02	0.0302	0.7268	0.8361	14.7612
Pennsylvania	IMDA-VAE	2.881E-02	1.940E-02	0.86364	0.86393	10.60016
Texas	GRU	2.450E-02	1.840E-02	0.8093	0.8219	23.5827
Texas	LSTM	2.450E-02	1.840E-02	0.7999	0.8089	23.6235
Texas	BiLSTM	2.240E-02	1.720E-02	0.8325	0.8502	21.5842
Texas	BiGRU	2.450E-02	1.780E-02	0.8047	0.8181	21.1139
Texas	VAE	2.500E-02	1.850E-02	0.78396	0.79411	27.21827
Texas	GRU-A	2.830E-02	0.0207	0.7542	0.7708	28.3301
Texas	LSTM-A	2.830E-02	0.0213	0.7482	0.7568	28.2549
Texas	IMDA-VAE	2.191E-02	1.564E-02	0.84519	0.86233	19.62706

TABLE VIII: Performance comparison of the proposed approach, using NO_2 pollutant.

State	Model	RMSE	MAE	R2	EV	MAPE
Arizona	GRU	3.320E-02	2.610E-02	0.9054	0.9151	27.6786
Arizona	LSTM	3.000E-02	2.440E-02	0.9159	0.928	27.2443
Arizona	BiLSTM	2.830E-02	2.360E-02	0.9264	0.9271	22.873
Arizona	BiGRU	3.000E-02	2.350E-02	0.9243	0.9244	22.2422
Arizona	VAE	3.000E-02	2.550E-02	0.92009	0.94457	25.19873
Arizona	GRU-A	3.870E-02	0.0324	0.8663	0.8675	32.7772
Arizona	LSTM-A	4.240E-02	0.034	0.8431	0.8438	32.1599
Arizona	IMDA-VAE	2.720E-02	2.143E-02	0.93457	0.93674	19.74715
California	GRU	2.240E-02	1.750E-02	0.9022	0.9085	24.2111
California	LSTM	2.450E-02	2.080E-02	0.8717	0.911	32.3879
California	BiLSTM	2.000E-02	1.590E-02	0.9252	0.9403	24.0743
California	BiGRU	2.000E-02	1.530E-02	0.9262	0.9375	22.2067
California	VAE	2.480E-02	1.910E-02	0.85128	0.93422	34.20658
California	GRU-A	2.450E-02	0.02	0.8788	0.8859	28.39
California	LSTM-A	2.830E-02	0.0228	0.8395	0.8549	33.2576
California	IMDA-VAE	1.612E-02	1.200E-02	0.94634	0.94668	13.59851
Pennsylvania	GRU	6.630E-02	5.170E-02	0.7242	0.7407	34.7455
Pennsylvania	LSTM	6.630E-02	5.270E-02	0.7209	0.7384	35.6146
Pennsylvania	BiLSTM	6.320E-02	4.640E-02	0.7473	0.7796	27.3349
Pennsylvania	BiGRU	6.080E-02	4.590E-02	0.7633	0.7916	27.9394
Pennsylvania	VAE	5.895E-02	4.290E-02	0.78842	0.79404	38.3737
Pennsylvania	GRU-A	7.870E-02	0.0624	0.6052	0.6739	37.659
Pennsylvania	LSTM-A	8.000E-02	0.0626	0.5961	0.6737	36.7711
Pennsylvania	IMDA-VAE	5.891E-02	4.286E-02	0.78047	0.80233	24.68344
Texas	GRU	4.120E-02	3.460E-02	0.7623	0.8136	35.2952
Texas	LSTM	4.360E-02	3.690E-02	0.7355	0.8012	37.5332
Texas	BiLSTM	3.740E-02	3.130E-02	0.8059	0.8345	30.0722
Texas	BiGRU	3.740E-02	3.050E-02	0.8126	0.8319	29.1011
Texas	VAE	4.720E-02	3.800E-02	0.69557	0.81882	40.75016
Texas	GRU-A	4.240E-02	0.0354	0.7447	0.7864	36.165
Texas	LSTM-A	4.360E-02	0.0363	0.7322	0.7595	36.8032
Texas	IMDA-VAE	3.391E-02	2.726E-02	0.84087	0.8599	24.97254

TABLE IX: Performance comparison of the proposed approach, using SO_2 pollutant.

State	Model	RMSE	MAE	R2	EV	MAPE
Arizona	GRU	1.000E-02	1.020E-02	0.3849	0.5466	102.308
Arizona	LSTM	1.000E-02	1.040E-02	0.3486	0.5076	99.3034
Arizona	BiLSTM	1.000E-02	7.200E-03	0.6789	0.7276	69.6978
Arizona	BiGRU	1.000E-02	7.800E-03	0.6362	0.7162	79.7304
Arizona	VAE	1.000E-02	1.080E-02	0.35603	0.83327	101.53512
Arizona	GRU-A	1.410E-02	1.180E-02	0.1735	0.3909	120.3919
Arizona	LSTM-A	1.410E-02	1.270E-02	0.0553	0.3321	131.2864
Arizona	IMDA-VAE	7.750E-03	6.500E-03	0.73399	0.83608	58.86316
California	GRU	2.400E-02	1.100E-02	0.742	0.757	13.208
California	LSTM	2.500E-02	1.300E-02	0.656	0.759	15.524
California	BiLSTM	1.900E-02	8.000E-03	0.884	0.905	9.419
California	BiGRU	2.100E-02	9.000E-03	0.846	0.873	10.295
California	VAE	2.400E-02	1.500E-02	0.01251	0.77078	29.45029
California	GRU-A	1.410E-02	1.240E-02	0.7031	0.7321	14.6651
California	LSTM-A	1.410E-02	1.250E-02	0.6966	0.7089	14.466
California	IMDA-VAE	9.490E-03	7.430E-03	0.893	0.895	7.653
Pennsylvania	GRU	2.000E-02	1.600E-02	0.6345	0.7881	37.9877
Pennsylvania	LSTM	2.240E-02	1.700E-02	0.5737	0.7245	42.3646
Pennsylvania	BiLSTM	1.730E-02	1.270E-02	0.7277	0.8298	26.8206
Pennsylvania	BiGRU	2.000E-02	1.600E-02	0.612	0.7988	30.7536
Pennsylvania	VAE	1.900E-02	1.550E-02	0.69652	0.75061	64.53344
Pennsylvania	GRU-A	2.830E-02	2.260E-02	0.3054	0.657	48.9787
Pennsylvania	LSTM-A	3.000E-02	2.370E-02	0.2325	0.623	50.985
Pennsylvania	IMDA-VAE	1.265E-02	9.680E-03	0.86125	0.86258	27.89178
Texas	GRU	2.450E-02	1.730E-02	0.6975	0.7201	69.1408
Texas	LSTM	2.450E-02	1.960E-02	0.6487	0.6838	79.1535
Texas	BiLSTM	2.000E-02	1.350E-02	0.7784	0.7906	50.0694
Texas	BiGRU	2.000E-02	1.320E-02	0.765	0.7747	48.0348
Texas	VAE	2.220E-02	1.610E-02	0.73183	0.73997	42.91426
Texas	GRU-A	2.450E-02	1.850E-02	0.6639	0.6914	75.4354
Texas	LSTM-A	2.650E-02	1.840E-02	0.6403	0.6599	72.9253
Texas	IMDA-VAE	1.975E-02	1.071E-02	0.78816	0.79334	37.78046

TABLE X: Performance comparison of the proposed approach, using O_3 pollutant.

State	Model	RMSE	MAE	R2	EV	MAPE
Arizona	GRU	3.000E-02	2.400E-02	0.9518	0.9518	10.0178
Arizona	LSTM	3.320E-02	2.570E-02	0.9424	0.9424	10.8426
Arizona	BiLSTM	2.650E-02	2.080E-02	0.9641	0.9665	9.2641
Arizona	BiGRU	2.650E-02	2.150E-02	0.9619	0.9639	9.376
Arizona	VAE	2.980E-02	2.090E-02	0.95581	0.95748	8.48824
Arizona	GRU-A	3.160E-02	0.0253	0.9461	0.9564	11.9185
Arizona	LSTM-A	3.610E-02	0.0286	0.932	0.9489	13.1987
Arizona	IMDA-VAE	2.646E-02	2.111E-02	0.96372	0.96418	8.19743
California	GRU	3.200E-02	2.400E-02	0.926	0.936	9.19
California	LSTM	3.200E-02	2.300E-02	0.934	0.935	9.535
California	BiLSTM	3.200E-02	2.100E-02	0.945	0.947	8.157
California	BiGRU	3.200E-02	2.200E-02	0.942	0.946	8.613
California	VAE	3.200E-02	2.400E-02	0.92453	0.93472	9.23233
California	GRU-A	2.830E-02	0.0206	0.9408	0.9421	9.253
California	LSTM-A	3.740E-02	0.0303	0.8902	0.9001	11.8519
California	IMDA-VAE	2.530E-02	1.996E-02	0.95063	0.95742	6.87696
Pennsylvania	GRU	5.660E-02	4.650E-02	0.8505	0.8527	32.6143
Pennsylvania	LSTM	5.390E-02	4.410E-02	0.864	0.8672	30.7883
Pennsylvania	BiLSTM	5.570E-02	4.530E-02	0.8574	0.8608	30.9349
Pennsylvania	BiGRU	5.660E-02	4.610E-02	0.8513	0.8532	33.1447
Pennsylvania	VAE	5.390E-02	4.410E-02	0.86707	0.8676	32.10249
Pennsylvania	GRU-A	5.740E-02	0.0468	0.8482	0.8678	30.3479
Pennsylvania	LSTM-A	5.660E-02	0.0451	0.8543	0.8727	30.0304
Pennsylvania	IMDA-VAE	5.050E-02	4.166E-02	0.88242	0.88535	28.86358
Texas	GRU	3.610E-02	2.850E-02	0.8478	0.8614	18.154
Texas	LSTM	3.460E-02	2.770E-02	0.8562	0.8763	18.13
Texas	BiLSTM	3.160E-02	2.550E-02	0.8858	0.8868	15.0679
Texas	BiGRU	3.320E-02	2.590E-02	0.8759	0.8814	16.2527
Texas	VAE	2.590E-02	2.080E-02	0.92111	0.92493	12.72924
Texas	GRU-A	3.610E-02	0.0295	0.8498	0.8502	17.6332
Texas	LSTM-A	3.870E-02	0.0316	0.83	0.8329	19.5214
Texas	IMDA-VAE	2.550E-02	2.078E-02	0.92438	0.92464	11.7666

attention. As expected, the traditional approach without attention is less time-consuming than the approach with attention due to computation cost related to the attention mechanism. More specifically, when conducting the experiments using an ordinary laptop (CPU Intel i3), the average execution time in second for the VAE and IMDA-VAE is 0.0076 and 0.0227, respectively. Thus, the average time allocated to the attention mechanism is approximately around 15 milliseconds in this case. On the other hand, when conducting the experiment using a PC with an Intel i7 CPU and equipped with a GPU, both approaches have an average execution time of less than 10^{-4} seconds. In short, when using time-series, the attention mechanism is not time-consuming.

In summary, the overall forecasting results demonstrate the

TABLE XI: Multivariate with multi-output Performance results using pollution datasets from California.

timesteps	model	RMSE	MAE	R2	EV	MAPE
3	GRU	15.11	6.563	0.841	0.849	26
3	LSTM	15.50	6.875	0.832	0.841	27
3	BiLSTM	14.78	6.406	0.847	0.854	26
3	BiGRU	14.12	6.08	0.861	0.863	24
3	VAE	15.779	7.113	0.826	0.839	28
3	GRU-A	14.066	6.066	0.861	0.864	24
3	LSTM-A	13.808	5.876	0.866	0.867	23
3	IMDA-VAE	13.44	5.620	0.87369	0.87526	21
6	GRU	14.37	6.137	0.856	0.858	25
6	LSTM	15.21	6.644	0.838	0.845	27
6	BiLSTM	13.59	5.741	0.871	0.871	23
6	BiGRU	13.55	5.691	0.872	0.872	23
6	VAE	14.763	6.500	0.848	0.855	26
6	GRU-A	14.259	6.129	0.857	0.859	24
6	LSTM-A	13.811	5.828	0.866	0.867	24
6	IMDA-VAE	12.91	5.251	0.884	0.884	21
9	GRU	14.70	6.491	0.849	0.856	28
9	LSTM	13.89	5.905	0.865	0.865	25
9	BiLSTM	13.68	5.742	0.869	0.869	23
9	BiGRU	13.18	5.56	0.878	0.878	22
9	VAE	14.427	6.289	0.854	0.862	26
9	GRU-A	13.683	5.877	0.869	0.869	24
9	LSTM-A	13.886	5.975	0.865	0.866	26
9	IMDA-VAE	12.54	5.174	0.888	0.890	20
12	GRU	14.92	6.545	0.845	0.854	27
12	LSTM	13.63	5.808	0.871	0.872	23
12	BiLSTM	13.98	6.106	0.864	0.869	28
12	BiGRU	13.34	5.654	0.876	0.876	22
12	VAE	13.105	5.532	0.881	0.882	24
12	GRU-A	13.516	5.896	0.873	0.873	26
12	LSTM-A	14.392	5.930	0.856	0.86	21
12	IMDA-VAE	12.53	5.073	0.891	0.891	20
15	GRU	13.42	5.764	0.875	0.877	25
15	LSTM	13.47	5.726	0.874	0.875	24
15	BiLSTM	13.34	5.751	0.876	0.877	24
15	BiGRU	13.59	5.968	0.872	0.875	26
15	VAE	12.644	5.416	0.889	0.889	25
15	GRU-A	13.374	5.873	0.875	0.876	27
15	LSTM-A	14.225	5.912	0.859	0.863	24
15	IMDA-VAE	11.96	4.838	0.901	0.901	20

high ability of the VAE based on robust variational inferences to approximate data probability distribution of a given pollution time-series, with a self-attention mechanism incorporated at multi-level to highlight and emphasize the correlation between data points of a given sequence. It has been shown that the variational inference with attention units improves time-dependencies modeling and univariate and multivariate forecasting without recurrent connections or memory cells. It should be highlighted that the best forecasting performance of the proposed IMDA-VAE approach has been obtained for univariate forecasting. Also, it has been shown that univariate forecasting outperformed multivariate forecasting in this setting. This is mainly due to the absence of high cross-correlation between the considered pollutants. This may be improved by considering meteorological variables, such as temperature and pressure, which facilitate the description of pollution dynamics, particularly for SO_2. The proposed IMDA-VAE approach can be used for online forecasting due to its simple architecture; only the encoder part is used for the forecasting, and data are processed in only one direction.

IV. Conclusion

Air pollution is a global issue, with most regions of the globe affected by concentrations known to have adverse health outcomes. The fast evolution of industrial technology has induced various adverse environmental impacts. Monitoring the ambient air quality is essential to achieve acceptable air quality. This work presented a novel deep hybrid model by introducing an attention mechanism to the variational autoencoder (called IMDA-VAE) to improve air pollution forecasting. In this study, we proved the efficiency of the proposed approach for univariate and multivariate forecasting of air pollution time-series data. Results showed that the proposed IMDA-VAE model provides more accurate forecasting of concentrations of four principal pollutants (i.e., NO_2, O_3, SO_2, and CO) than uni-directional and bi-directional recurrent networks, namely VAE, Gated GRUs, LSTM, BiGRU, BiLSTM, ConvLSTM, LSTM-A, and GRU-A. The forecasting accuracy has been evaluated by six statistical indicators, including R2, RMSE, MAE, MAPE, EV, MBE, and RMBE. Metrics demonstrated the high ability of deep hybrid model with attention to model temporal-dependencies in unsupervised learning without complex recurrent networks gating and memory mechanism; variational inference approximation exhibits promising performance in time-dependent modeling. Besides, univariate forecasts showed better accuracy than multivariate forecasting in this setting. This mainly due to the absence of high cross-correlation between the four studied pollutants.

Despite the adequate forecasting results of ambient air pollution obtained using the proposed IMDA-VAE model, the work presented in this study guides future works. Since pollution measurements may contain noisy features with time and frequency contributions, we plan to enhance the proposed IMDA-VAE-based forecasting model by developing a multi-scale IMDA-VAE model that combines IMDA-VAE techniques with wavelet-based multi-resolution representation. Another direction for future improvement is to incorporate explanatory variables, such as meteorological measurements, in constructing the deep learning models. Further, it will be interesting to design an early detection system of abnormal pollution to foster reactive control, enabling avoiding exposition to abnormal pollution with high concentrations.

References

[1] F. Harrou, F. Kadri, S. Khadraoui, and Y. Sun, "Ozone measurements monitoring using data-based approach," *Process Safety and Environmental Protection*, vol. 100, pp. 220–231, 2016.

[2] S. Ali, T. Glass, B. Parr, J. Potgieter, and F. Alam, "Low Cost Sensor With IoT LoRaWAN Connectivity and Machine Learning-Based Calibration for Air Pollution Monitoring," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–11, 2020.

[3] W. Hernandez, A. Mendez, R. Zalakeviciute, and A. M. Diaz-Marquez, "Analysis of the information obtained from PM 2.5 concentration measurements in an urban park," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 9, pp. 6296–6311, 2020.

[4] K. Gu, Z. Xia, and J. Qiao, "Stacked selective ensemble for PM 2.5 forecast," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 3, pp. 660–671, 2019.

[5] D. Antanasijević, V. Pocajt, A. Perić-Grujić, and M. Ristić, "Multiple-input–multiple-output general regression neural networks model for the simultaneous estimation of traffic-related air pollutant emissions," *Atmospheric Pollution Research*, vol. 9, no. 2, pp. 388–397, 2018.

[6] F. Harrou, L. Fillatre, M. Bobbia, and I. Nikiforov, "Statistical detection of abnormal ozone measurements based on constrained generalized likelihood ratio test," in *52nd IEEE Conference on Decision and Control*. IEEE, 2013, pp. 4997–5002.

[7] F. Harrou, A. Dairi, Y. Sun, and F. Kadri, "Detecting abnormal ozone measurements with a deep learning-based strategy," *IEEE Sensors Journal*, vol. 18, no. 17, pp. 7222–7232, 2018.

[8] M. Abhilash, A. Thakur, D. Gupta, and B. Sreevidya, "Time series analysis of air pollution in Bengaluru using ARIMA model," in *Ambient Communications and Computer Systems*. Springer, 2018, pp. 413–426.

[9] U. Kumar and V. Jain, "ARIMA forecasting of ambient air pollutants (O3, NO, NO2 and CO)," *Stochastic Environmental Research and Risk Assessment*, vol. 24, no. 5, pp. 751–760, 2010.

[10] M. Arsov, E. Zdravevski, P. Lameski, R. Corizzo, N. Koteli, K. Mitreski, and V. Trajkovik, "Short-term air pollution forecasting based on environmental factors and deep learning models," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*. IEEE, 2020, pp. 15–22.

[11] M. H. Lee, N. H. Rahman, M. T. Latif, M. E. Nor, and N. A. Kamisan, "Seasonal ARIMA for forecasting air pollution index: A case study," *American Journal of Applied Sciences*, vol. 9, no. 4, p. 570, 2012.

[12] L. M. B. Ventura, F. de Oliveira Pinto, L. M. Soares, A. S. Luna, and A. Gioda, "Forecast of daily PM 2.5 concentrations applying artificial neural networks and Holt–Winters models," *Air Quality, Atmosphere & Health*, vol. 12, no. 3, pp. 317–325, 2019.

[13] L. Wu, X. Gao, Y. Xiao, S. Liu, and Y. Yang, "Using grey Holt–Winters model to predict the air quality index for cities in China," *Natural Hazards*, vol. 88, no. 2, pp. 1003–1012, 2017.

[14] Z. Zhao, W. Chen, X. Wu, P. C. Chen, and J. Liu, "LSTM network: a deep learning approach for short-term traffic forecast," *IET Intelligent Transport Systems*, vol. 11, no. 2, pp. 68–75, 2017.

[15] P. Jiang, C. Li, R. Li, and H. Yang, "An innovative hybrid air pollution early-warning system based on pollutants forecasting and extenics evaluation," *Knowledge-Based Systems*, vol. 164, pp. 174–192, 2019.

[16] A. Prakash, U. Kumar, K. Kumar, and V. Jain, "A wavelet-based neural network model to predict ambient air pollutants' concentration," *Environmental Modeling & Assessment*, vol. 16, no. 5, pp. 503–517, 2011.

[17] M. Arhami, N. Kamali, and M. M. Rajabi, "Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by monte carlo simulations," *Environmental Science and Pollution Research*, vol. 20, no. 7, pp. 4777–4789, 2013.

[18] K. Siwek and S. Osowski, "Data mining methods for prediction of air pollution," *International Journal of Applied Mathematics and Computer Science*, vol. 26, no. 2, pp. 467–478, 2016.

[19] H. Liu, H. Wu, X. Lv, Z. Ren, M. Liu, Y. Li, and H. Shi, "An intelligent hybrid model for air pollutant concentrations forecasting: Case of beijing in china," *Sustainable Cities and Society*, vol. 47, p. 101471, 2019.

[20] J. He, Y. Yu, Y. Xie, H. Mao, L. Wu, N. Liu, and S. Zhao, "Numerical model-based artificial neural network model and its application for quantifying impact factors of urban air quality," *Water, Air, & Soil Pollution*, vol. 227, no. 7, p. 235, 2016.

[21] W. Ding, J. Zhang, and Y. Leung, "Prediction of air pollutant concentration based on sparse response back-propagation training feedforward neural networks," *Environmental Science and Pollution Research*, vol. 23, no. 19, pp. 19 481–19 494, 2016.

[22] F. Harrou, Y. Sun, A. S. Hering, M. Madakyaru *et al.*, *Statistical process monitoring using advanced data-driven and deep learning approaches: theory and practical applications*. Elsevier, 2020.

[23] W. Mao, J. He, and M. J. Zuo, "Predicting remaining useful life of rolling bearings based on deep feature representation and transfer learning," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 4, pp. 1594–1608, 2019.

[24] X. Yuan, S. Qi, and Y. Wang, "Stacked enhanced auto-encoder for data-driven soft sensing of quality variable," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 10, pp. 7953–7961, 2020.

[25] F. Harrou, T. Cheng, Y. Sun, T. O. Leiknes, and N. Ghaffour, "A data-driven soft sensor to forecast energy consumption in wastewater treatment plants: A case study," *IEEE Sensors Journal*, 2020.

[26] F. Harrou, F. Kadri, and Y. Sun, "Forecasting of photovoltaic solar power production using LSTM approach," in *Advanced Statistical Modeling, Forecasting, and Fault Detection in Renewable Energy Systems*. IntechOpen, 2020, p. 3.

[27] Y.-S. Chang, H.-T. Chiao, S. Abimannan, Y.-P. Huang, Y.-T. Tsai, and K.-M. Lin, "An LSTM-based aggregated model for air pollution forecasting," *Atmospheric Pollution Research*, vol. 11, no. 8, pp. 1451–1463, 2020.

[28] Q. Tao, F. Liu, Y. Li, and D. Sidorov, "Air pollution forecasting using a deep learning model based on 1d convnets and bidirectional gru," *IEEE access*, vol. 7, pp. 76 690–76 698, 2019.

[29] S. Al-Janabi, M. Mohammad, and A. Al-Sultan, "A new method for prediction of air pollution based on intelligent computation," *Soft Computing*, vol. 24, no. 1, pp. 661–680, 2020.

[30] T.-C. Bui, V.-D. Le, and S.-K. Cha, "A deep learning approach for forecasting air pollution in south korea using lstm," *Machine Learning*, 2018. [Online]. Available: https://arxiv.org/abs/1804.07891v3

[31] S. Kim, J. M. Lee, J. Lee, and J. Seo, "Deep-dust: predicting concentrations of fine dust in seoul using lstm," *Climate Informatics*, 2019. [Online]. Available: http://arxiv.org/abs/1901.10106

[32] B. Freeman, G. Taylor, B. Gharabaghi, and J. Thé, "Forecasting air quality time series using deep learning," *Journal of the Air & Waste Management Association*, vol. 68, pp. 866–886, 2018.

[33] V. Athira, P. Geetha, R. Vinayakumar, and K. Soman, "Deepairnet: applying recurrent networks for air quality prediction," *Procedia Computer Science*, vol. 132, pp. 1394–1403, 2018. [Online]. Available: https://doi.org/10.1016/j.procs.2018.05.068

[34] T. Xayasouk and H. Lee, "Air pollution prediction system using deep learning," *WIT Transactions on Ecology and the Environment*, vol. 230, pp. 71–79, 2018.

[35] X. Li, L. Peng, Y. Hu, J. Shao, and T. Chi, "Deep learning architecture for air quality predictions," *Environmental Science and Pollution Research*, vol. 23, pp. 22 408–22 417, 2016. [Online]. Available: https://doi.org/10.1007/s11356-016-7812-9

[36] X. Li, L. Peng, X. Yao, S. Cui, Y. Hu, C. You, and T. Chi, "Long short-term memory neural network for air pollutant concentration predictions: method development and evaluation," *Environmental Pollution*, vol. 231, pp. 997–1004, 2017. [Online]. Available: https://doi.org/10.1016/j.envpol.2017.08.114

[37] P.-W. Soh, J.-W. Chang, and J.-W. Huang, "Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations," *IEEE Access*, vol. 6, pp. 38 186–38 199, 2018.

[38] H. Chiou-Jye and K. Ping-Huan, "A deep cnn-lstm model for particulate matter (pm2.5) forecasting in smart cities," *Sensors*, vol. 18, 2018. [Online]. Available: https://doi.org/10.3390/s18072220

[39] H. Wang, B. Zhuang, Y. Chen, N. Li, and D. Wei, "Deep inferential spatial-temporal network for forecasting air pollution concentrations," *Machine Learning*, 2018. [Online]. Available: https://arxiv.org/abs/1809.03964v1

[40] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *stat*, vol. 1050, p. 1, 2014.

[41] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[43] J. An and S. Cho, "Variational autoencoder based anomaly detection using reconstruction probability," *Special Lecture on IE*, vol. 2, no. 1, pp. 1–18, 2015.

[44] A. Dairi, F. Harrou, Y. Sun, and S. Khadraoui, "Short-term forecasting of photovoltaic solar power production using variational auto-encoder driven deep learning approach," *Applied Sciences*, vol. 10, no. 23, p. 8400, 2020.

[45] G. Boquet, A. Morell, J. Serrano, and J. L. Vicario, "A variational autoencoder solution for road traffic forecasting systems: Missing data imputation, dimension reduction, model selection and anomaly detection," *Transportation Research Part C: Emerging Technologies*, vol. 115, p. 102622, 2020.

[46] H. Gunduz, "An efficient stock market prediction model using hybrid feature reduction method based on variational autoencoders and recursive feature elimination," *Financial Innovation*, vol. 7, no. 1, pp. 1–24, 2021.

[47] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun, "Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study," *Chaos, Solitons & Fractals*, vol. 140, p. 110121, 2020.

[48] M. R. Ibrahim, J. Haworth, A. Lipani, N. Aslam, T. Cheng, and N. Christie, "Variational-lstm autoencoder to forecast the spread of coronavirus across the globe," *PloS one*, vol. 16, no. 1, p. e0246120, 2021.

[49] Y. Zerrouki, F. Harrou, N. Zerrouki, A. Dairi, and Y. Sun, "Desertification Detection using an Improved Variational AutoEncoder-Based Approach through ETM-Landsat Satellite Data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 202–213, 2020.

[50] L. Li, J. Yan, H. Wang, and Y. Jin, "Anomaly detection of time series with smoothness-inducing sequential variational auto-encoder," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1177–1191, 2021.

[51] C. Zhang and Y. Chen, "Time series anomaly detection with variational autoencoders," *arXiv preprint arXiv:1907.01702*, 2019.

[52] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[53] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.

[54] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[55] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[56] C. Cortes, M. Mohri, and A. Rostamizadeh, "L2 regularization for learning kernels," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2009, pp. 109–116.

[57] J. Ngarambe, S. J. Joen, C.-H. Han, and G. Y. Yun, "Exploring the relationship between particulate matter, CO, SO2, NO2, O3 and urban heat island in Seoul, Korea," *Journal of Hazardous Materials*, vol. 403, p. 123615, 2021.

[58] Ş. Ç. Doğruparmak and B. Özbay, "Investigating correlations and variations of air pollutant concentrations under conditions of rapid industrialization–kocaeli (1987–2009)," *CLEAN–Soil, Air, Water*, vol. 39, no. 7, pp. 597–604, 2011.

[59] R. A. Rajagukguk, R. A. Ramadhan, and H.-J. Lee, "A review on deep learning models for forecasting time series data of solar irradiance and photovoltaic power," *Energies*, vol. 13, no. 24, p. 6623, 2020.

[60] J. Zhang, A. Florita, B.-M. Hodge, S. Lu, H. F. Hamann, V. Banunarayanan, and A. M. Brockway, "A suite of metrics for assessing the performance of solar power forecasting," *Solar Energy*, vol. 111, pp. 157–175, 2015.

[61] M. Kuhn, K. Johnson *et al.*, *Applied predictive modeling*. Springer, 2013, vol. 26.

[62] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013, vol. 112.

[63] G. E. Hinton, "A practical guide to training restricted boltzmann machines," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 599–619.

[64] R. Sarikaya, G. E. Hinton, and A. Deoras, "Application of deep belief networks for natural language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 778–784, 2014.

[65] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.