atmosphere

Article

Spatiotemporal Air Pollution Forecasting in Houston-TX: A Case Study for Ozone Using Deep Graph Neural Networks

Victor Oliveira Santos [1,*], Paulo Alexandre Costa Rocha [1,2], John Scott [3], Jesse Van Griensven Thé [1,3] and Bahram Gharabaghi [1,*]

[1] School of Engineering, University of Guelph, 50 Stone Rd E, Guelph, ON N1G 2W1, Canada
[2] Mechanical Engineering Department, Technology Center, Federal University of Ceará, Fortaleza 60020-181, CE, Brazil
[3] Lakes Environmental, 170 Columbia St. W, Waterloo, ON N2L 3L3, Canada
* Correspondence: volive04@uoguelph.ca (V.O.S.); bgharaba@uoguelph.ca (B.G.)

Abstract: The presence of pollutants in our atmosphere has become one of humanity's greatest challenges. These pollutants, produced primarily by burning fossil fuels, are detrimental to human health, our climate and agriculture. This work proposes the use of a spatiotemporal graph neural network, designed to forecast ozone concentration based on the GraphSAGE paradigm, to aid in our understanding of the dynamic nature of these pollutants' production and proliferation in urban areas. This model was trained and tested using data from Houston, Texas, the United States, with varying numbers of time-lags, forecast horizons (1, 3, 6 h ahead), input data and nearby stations. The results show that the proposed GNN-SAGE model successfully recognized spatiotemporal patterns underlying these data, bolstering its forecasting performance when compared with a benchmarking persistence model by 33.7%, 48.7% and 57.1% for 1, 3 and 6 h forecast horizons, respectively. The proposed model produces error levels lower than we could find in the existing literature. The conclusions drawn from variable importance SHAP analysis also revealed that when predicting ozone, solar radiation becomes relevant as the forecast time horizon is raised. According to EPA regulation, the model also determined nonattainment conditions for the reference station.

Keywords: air pollution; ozone; Houston; forecasting; machine learning; graph neural networks; SHAP analysis

Citation: Oliveira Santos, V.; Costa Rocha, P.A.; Scott, J.; Van Griensven Thé, J.; Gharabaghi, B. Spatiotemporal Air Pollution Forecasting in Houston-TX: A Case Study for Ozone Using Deep Graph Neural Networks. *Atmosphere* **2023**, *14*, 308. https://doi.org/10.3390/atmos14020308

Academic Editor: Jaroslaw Krzywanski

Received: 10 January 2023
Revised: 30 January 2023
Accepted: 1 February 2023
Published: 3 February 2023

1. Introduction

In recent years, it has become increasingly apparent that anthropogenic pollutant emissions pose one of the greatest threats to human society and the planet as we know it [1]. According to the Statistical Review of World Energy [2], even though total fossil fuel consumption decreased from 490.07 EJ in 2019 to 463.24 EJ in 2020, the harmful effects of gases emitted from fossil fuels still heavily impact human health, the condition of our climate and the economy. A wide body of existing research makes the negative influence of pollutants on the human body abundantly clear. This research highlights a significant contribution to the development of various illnesses across all pathology, be they neurological [1,3], cancerous [4,5], coronary [6], respiratory [7], or even psychological [8], amounting to countless premature deaths [9–11].

Hazardous air pollutants also impact the global environment, as previous studies suggest [12–14]. The accumulation of pollutants in the atmosphere due to human activity boosts the absorption and reemission of energy in the atmosphere, resulting in a trend of gradual increase in global temperature caused by the greenhouse effect [15,16]. The economic loss due to the impacts of pollutant gases have also been studied thoroughly, geographically spanning Africa [17], the United States [18], Europe [19,20], India [21], China [22] and beyond.

Clearly, the negative outcomes of pollutant emissions are a major obstacle for health and development across the globe. In order to mitigate the harmful impacts of this pollution, air quality regulations were developed, primarily targeted towards limiting the concentrations of hazardous materials in the atmosphere [23].

Different predictive models were developed to keep track of toxic gas concentration in the atmosphere. Chemical transport models (CTM) rely on meteorological and chemical processes occurring in the atmosphere to simulate pollutant concentrations, offering a broad spatiotemporal coverage [24–26]. This physics-based approach seeks to establish a mathematical model to assess pollutant concentration, a computationally demanding process [27,28]. Other deterministic models, such as Weather Research and Forecasting (WRF) coupled with Community Multi-Scale Air Quality (CMAQ) are viable options for estimating pollution levels. These models have been successfully used to assess pollutant concentrations in different studies [24,29,30]. Deterministic models combined with machine learning techniques have also proved to be viable tools for pollution forecasting [28,31,32].

Physics-based models, such as transport models, are computationally expensive, requiring long processing times [27,28,33] and relying on detailed spatial analysis for the correct modeling of urban areas [34]. To overcome these drawbacks, researchers investigated machine learning (ML) to model complex spatiotemporal processes. Machine learning can accurately represent non-linear processes and eliminate the need for the explicit programming of each physical process for its implementation [25,28,35,36]. Machine learning models have improved computational performance compared to physics-based models [37].

Previous research in this field [38] investigated three different machine learning (ML) models, namely support vector machines (SVM), decision trees (DT) and artificial neural networks (ANN), using both univariate and multivariate data input. The authors proposed forecasting pollutant concentration up to 24 h ahead in Qatar. The research results indicated that the multivariate approach led to better results, reaching the best value for Normalized Root Mean Squared Error (nRMSE) of 6.1% for 1 h in advance ozone estimation. In [39], the authors proposed to use regression trees (RT) to forecast the daily maximum 8 h average ozone concentrations over China. The RT model successfully captured spatiotemporal concentrations of ozone, being more accurate than CTM models and requiring less computational effort, with a coefficient of determination (R^2) and Root Mean Squared Error (RMSE) of 0.69 and 26 $\mu g/m^3$, respectively. In another research effort [40], the authors compared eight different ML models to predict future ozone levels in India for 24 h. The models were trained and tested using two distinct approaches: a whole year of data and seasonal parted data. The results showed that amongst the assessed models, XGBoost reached the best results for the yearly training with an R^2 equal to 0.61. The overall models' predictive capacity increased for the seasonal-trained approach, where XGBoost was the best model again, with R^2 equal to 0.75 for forecasting during the winter season, an improvement of 18%.

More recently, deep learning (DL) models were applied to estimate pollutant concentration levels [33]. Seng et al. [41] used the long short-term memory (LSTM) model with a multi-output and multi-index of supervised learning to estimate air quality over Beijing, considering the data from the present monitoring station and the ones on its surroundings. The results proved that their proposed model achieved the best results compared to their baseline DL models. Deep learning techniques can also be used with satellite data to estimate pollution levels in the atmosphere, as found in [42]. The authors used the Learning Surface Ozone (LESO) framework and deep forests to estimate ozone levels over the United States, Europe and India. The results showed improvement for R^2 of over 30%. In the US and Europe, R^2 achieved values greater than 0.9, while in India, the model had a worse performance reaching an R^2 of 0.39. Combinations of DL models have also been used to forecast atmospheric conditions regarding pollutants. In [43], authors propose using convolutional neural networks (CNN) and LSTM models to predict air quality for Barcelona, Kocaeli and Istanbul. This hybrid approach combines the best characteristics of each model to analyze temporal data by the LSTM and the spatial understanding of

data by CNN. Compared with results found in the literature, their CNN-LSTM architecture improved up to 53% prediction for particulate matter, 31% for ozone forecasting, 47% for oxides of nitrogen and 46% for sulphur dioxide, using a multivariate approach. Deep learning models were also implemented along attention mechanisms, allowing the model to focus on important information from the dataset [44]. In study [45], a DL model based on CNN and spatiotemporal attention was developed to forecast atmospheric particulate matter (PM_{10} and $PM_{2.5}$) up to 4 h ahead over the Yangtze River Delta airshed in China. The proposed model's performance was superior over the assessed state-of-the-art baseline models, reaching mean improvement up to 15.25% for $PM_{2.5}$ and 16.91% for PM_{10}, proving to be a reliable framework for pollutant prediction.

Another promising tool for estimating air pollution is based on graph theory. Graph-based models supplement traditional DL models by directly taking into account spatiotemporal characteristics underlying the used information [46,47]. In [48], authors combined graph convolutional neural networks (GC) and LSTM models to forecast $PM_{2.5}$ concentration in China. The authors compared the GC-LSTM proposed model with other state-of-the-art approaches, showing that their model outperformed the baselines for all forecast horizons from 1 h up to 72 h ahead. Graph attention networks (GAT) are another possible paradigm to predict future pollutant concentration. In work [49], the authors developed a convolutional graph sequence-to-sequence with an attention model to assess future ozone concentrations over a long time series. The proposed model proved realiable and had superior performance over the benchmarking models considered. An application of GAT together with a recurrent network can be found in research [50]. Their authors developed a GAT model that learns the spatial dependencies of the forecast pollutant for different locations. The application of graph-based models proved to achieve better results than previous ML baseline models, improving RMSE over 3.5% compared with the best benchmarking result. The reviewed literature points to the importance of ML, DL and, more recently, graph-based models in developing precise pollutant forecasting tools.

The present work proposes a novel graph-structured approach based on GraphSAGE [51] to predict ozone concentration over the urban area of Houston, Texas. A different number of time-lags and forecast horizons were considered during this study to further understand how these factors impact the model's prediction. To accomplish this, four different approaches were proposed and compared:

1. In the first approach, the input data consisting only of past ozone concentration data and hour of the day (HoD) and day of the year (DoY) were used to estimate ozone concentration. This limitation yields the highest number of available stations to spatially feed the model.
2. For the second approach, ozone, NO_2/NO_X and weather information (wind speed, wind direction, outdoor temperature, relative humidity, solar radiation) and HoD/DoY were used to estimate ozone concentration. This is the maximum weather data that we can use, which yields a smaller number of stations but favours some important ozone precursors (chemical and physical).
3. The third approach used past ozone data, NO_2/NO_X, Volatile organic compounds (VOCs) (propane, isobutane, benzene, toluene, ethylbenzene), weather (wind speed, wind direction, outdoor temperature) and HoD/DoY information to estimate ozone concentration. The VOCs information limited the other input variables and the number of stations, resulting in the smallest set of stations.
4. Finally, the last approach used past ozone data, NO_2/NO_X, weather and HoD/DoY information, aiming for the maximum number of stations aggregated with chemical/physical information.

To improve on the research that was introduced above, this work's main contributions aim at deepening the scientific understanding of ozone forecasting using advanced machine learning methods by:

1. Developing a new model to forecast air pollution concentrations.
2. Analyzing the inclusion of spatial information to deep learning air quality forecasts.

3. The importance analysis of input variables for different data configurations over distinct deep learning forecasting horizons.
4. Producing an accurate and reliable model for ozone prediction based on DL and graph theory.

The remainder of this work is organized as follows: the applied methodology is presented in Section 2, followed by the presentation and discussion of results in Section 3. In Section 4, the found results are discussed. In Section 5 a conclusion closes the work.

2. Materials and Methods

2.1. Houston Database

The data used for training and testing the proposed model were acquired from the Houston metropolitan area, provided by the Texas Commission on Environmental Quality (data acquired from https://www17.tceq.texas.gov/tamis/index.cfm, accessed on 1 December 2022). This site was chosen since it is a major urban area with an important industrial and economic center, hosting large oil refineries and petrochemical industries in the North American region [52,53]. Due to its highly populated area and economic activities, this region is often classified as non-attained by the Environmental Protection Agency (EPA) for the US. This means that Houston does not comply with the National Ambient Air Quality Standards (NAAQS) for air pollutant concentrations [52–54]. For the present work, data from monitoring stations were used. The historical data comprise the period from 2011 to 2019 and has hourly information for concentrations of ozone, NO_2/NO_x, the VOCs propane, isobutane, benzene, toluene and ethylbenzene. This research supplemented the pollutant information with weather data such as wind speed, wind direction, outdoor temperature, relative humidity and solar radiation. A map of the studied region and its stations is presented in Figure 1.

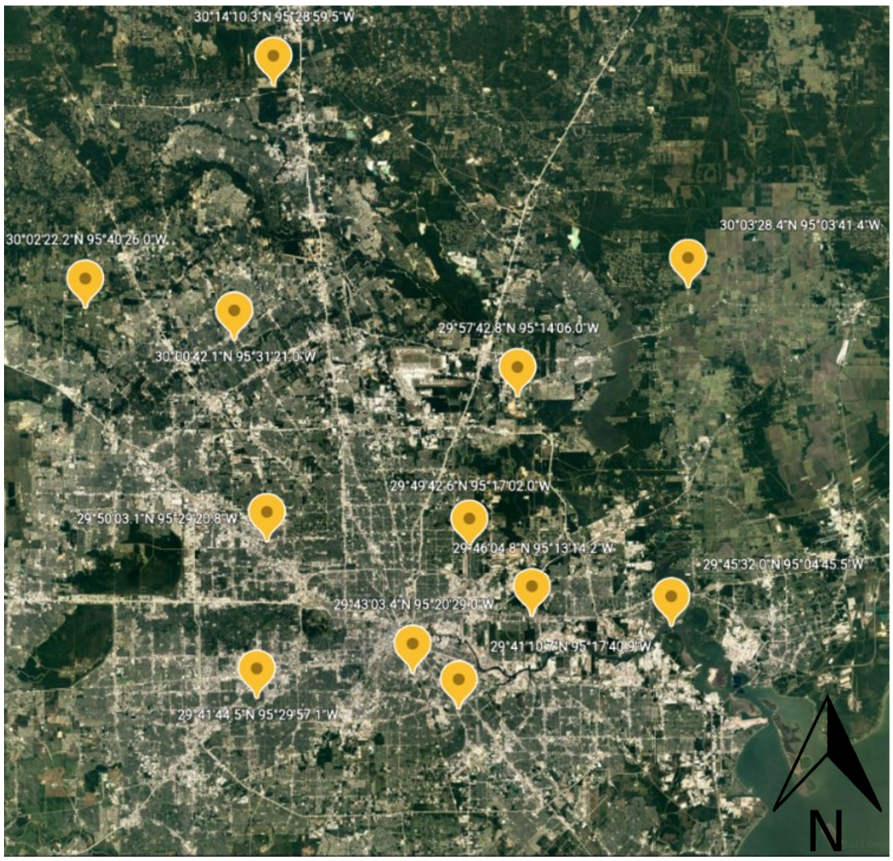

Figure 1. Map of the Houston area and weather stations.

The topographic maps containing the resulting stations and their height for each proposed approach are depicted in panes (a) to (d) of Figure 2. It is worth mentioning that most stations are on flat, low-lying terrain.

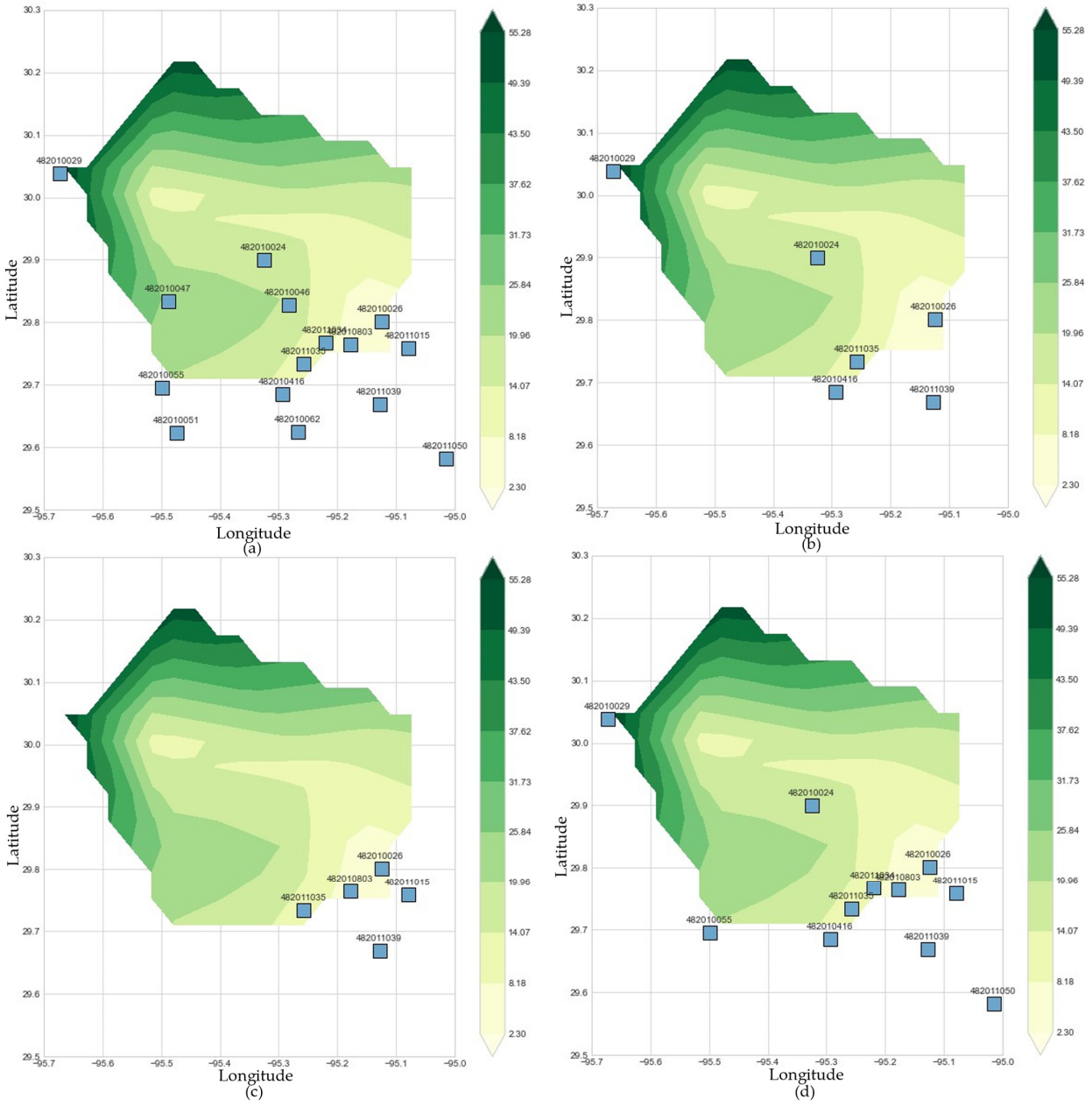

Figure 2. Topographic maps showing the spatial distribution for the used stations for approaches (**a**) 1, (**b**) 2, (**c**) 3 and (**d**) 4.

In Figure 2, the right-side bar indicates the terrain's height, in meters, for the Houston area, whilst the blue squares represent the location of each weather station. The varying number of stations in each pane is caused by the lack of data from some stations, resulting in stations being filtered out to keep the measurements uniform on each approach, keeping the ones containing enough valid information for the model assessment. Based on that, the reference station used for assessing all the proposed forecasting approaches is station 482011035, located in the southern part of Houston, as shown in Figure 2.

Wind speed and direction are the weather variables with the greatest influence on ozone concentration in the atmosphere [55]. Thus, it is of interest to understand the wind

behavior at the weather stations during the assessed period. The wind characterization for the stations showed that the wind essentially had a similar trend for each station, being more prominent from a region between east and south. Similar behavior is found for the stations of all the different approaches but number 1, since it does not consider any weather information.

2.2. Persistence Model

The persistence model was selected as a baseline model to assess the proposed model's performance. The persistence model is often used in forecasting tasks as benchmarking to compare with novel proposed models and states that the predicted attribute will be the same as the one measured at present [56–58]. The persistence model performs well for short forecast horizons [56,59,60], offering good baseline comparison.

2.3. LASSO Model

The LASSO (Least Absolute Shrinkage and Selection Operator) model [61] performs attribute selection and normalization of linear problems using L1 penalization. For the present work, LASSO was used as a benchmarking model to assess the size of the historical dataset.

2.4. GraphSAGE Model

The GraphSAGE (sample and aggregate) model is a framework where nodes are equally aggregated to learn embedding functions to generalize unseen information by training a set of aggregator functions using samples of a constant size set of neighboring nodes [51,62,63]. Its architecture allows for the topological structure of different nodes to be learnt, as well as their distribution along the dataset, ultimately leading to the model's generalization [51]. The main advantage of graph-based models over the traditional ML models is that they naturally consider spatial relationships between the assessed stations. This is essential in forecasting pollutant concentration, such as ozone, since it relies on both temporal and spatial features, also facilitating the understanding of its base structure [46,47]. Figure 3 displays a logic representation for the graph paradigm used in this work, while the applied GraphSAGE structure is depicted in Figure 4.

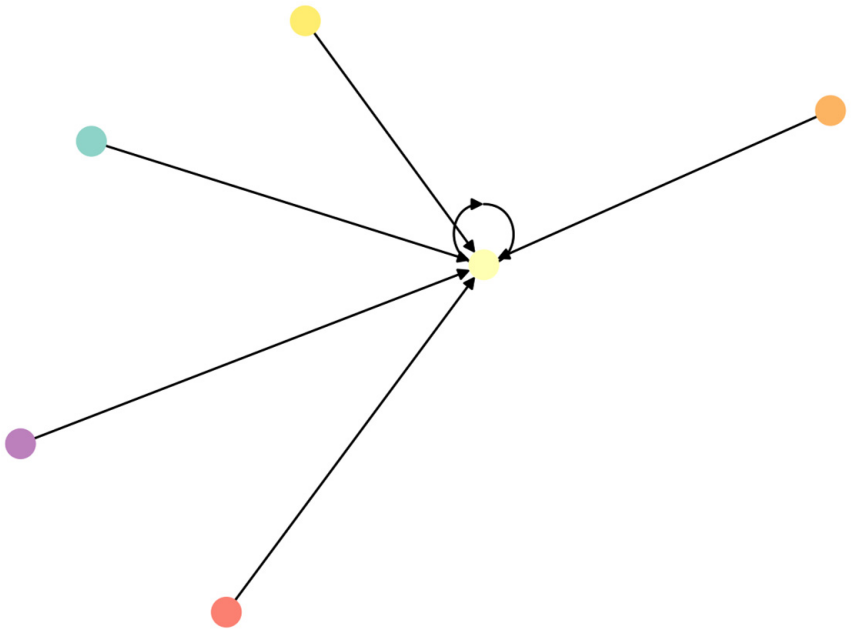

Figure 3. Example of logic representation for the graph topology.

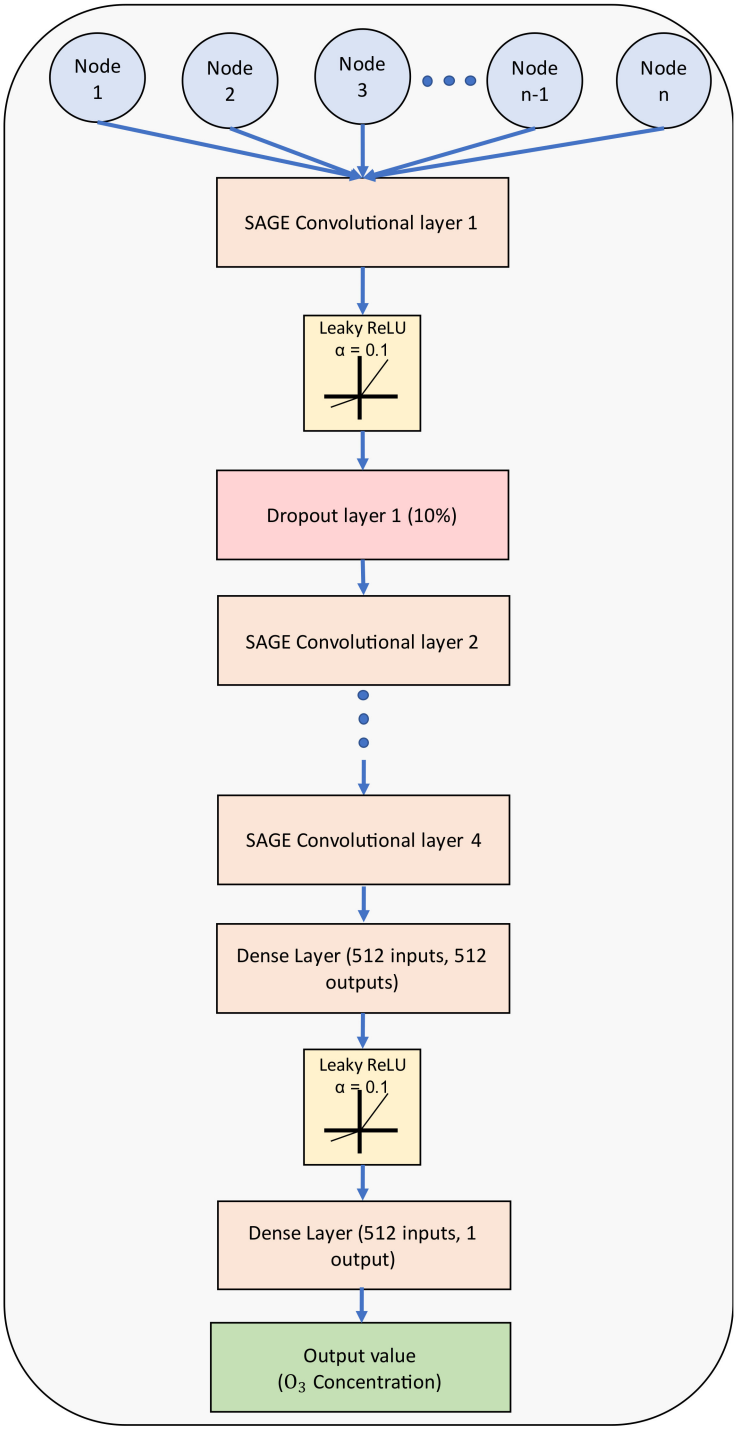

Figure 4. Proposed GraphSAGE structure used.

Figure 3 depicts the graph approach using five nearby stations, represented as circles as an example. From the figure, nearby stations send information to the reference study station, located in the middle of the image at the endpoint of the arrow's heads, also sending data to itself.

Figure 4 shows how the input spatiotemporal data from the nodes are fed to the model, passing through successive convolutional, dropout layers and leaky ReLU activation functions. After four repetitions, the output data are then sent to be processed by a dense layer, where they will finally return the predicted value for ozone concentration by its output layer.

2.5. Evaluation Metrics

To compare the proposed model performance against baseline models, the metrics RMSE, nRMSE, Forecast Skill and R^2 were used. Their formulation can be found in references [57] for RMSE, nRMSE and Forecast Skill and reference [64] for R^2.

2.6. SHAP Analysis

Although ML models lead to state-of-the-art results in many scientific fields, these models lack explainability, making them hard to be interpreted beyond their results. Addressing this matter, Shapley Additive Explanations (SHAP) was implemented after model training using the SHAP library for Python language (the SHAP library documentation can be read at https://shap.readthedocs.io/en/latest/, accessed on 26 January 2023). The SHAP approach offers a way to locally explain ML models through game theory, by assessing a model's results for each case when a variable is not considered, finding relationships between such variables and ranking them from the most to the least important [65]. The SHAP approach has proved to be a versatile tool with multidisciplinary applications, being successfully implemented for regression problems [66], model optimization [67,68], parameter selection [69] and even in the medical field [70,71], helping researchers to fully analyze their results.

3. Results

3.1. Dataset Size

For the model to return meaningful and reliable results, it is necessary to verify if the information contained in the dataset is sufficient for training. To do so, different dataset sizes for training were tested. The results achieved by the GraphSAGE model were compared to the baseline models LASSO and persistence, as shown in Figure 5 (a) for RMSE and (b) for Forecast Skill, respectively.

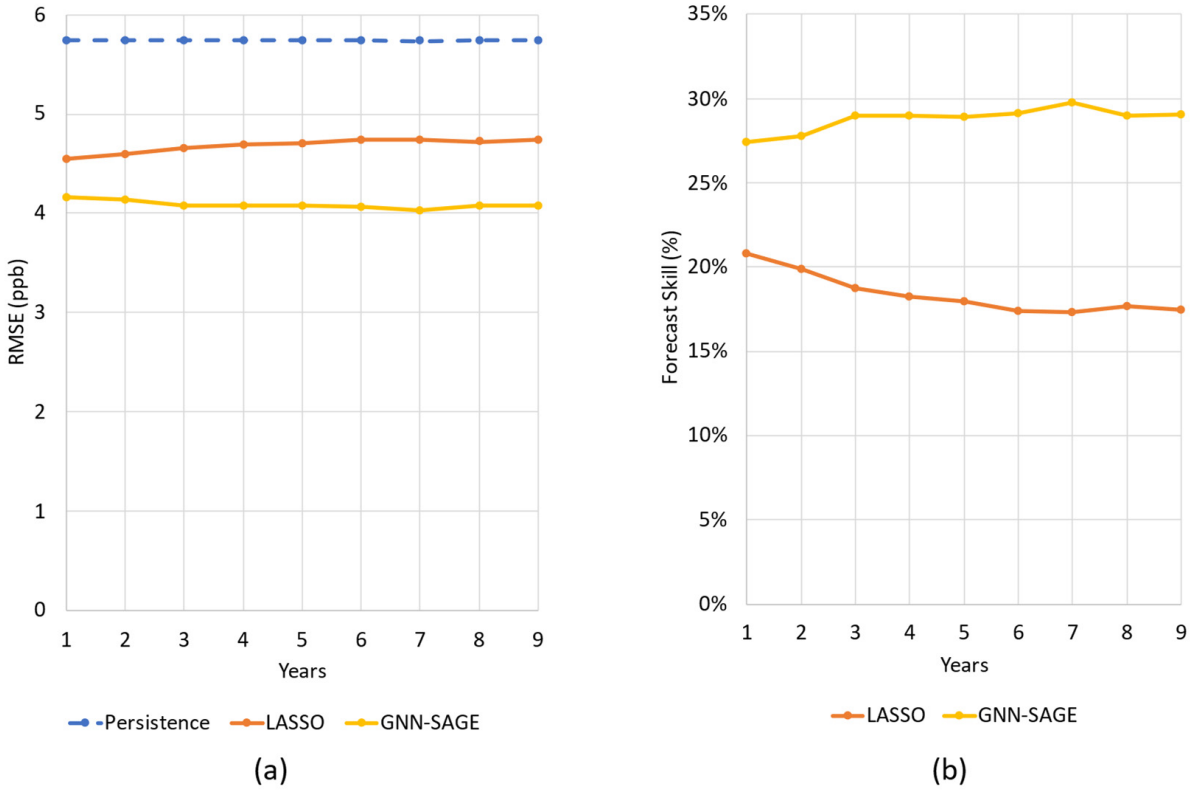

Figure 5. Influence of different sizes for the training dataset for GraphSAGE model compared with benchmarking models LASSO and persistence in terms of RMSE (**a**) and Forecast Skill (**b**).

In Figure 5, the x-axis represents the number of years used to train the model, ranging from one to nine years. For panel (a), it is visible that increasing the number of training years has a positive influence over the GraphSAGE model, reaching the optimum value of 4.03 ppb for seven years. The baseline model LASSO showed worse performance, with a larger RMSE of 5.74 ppb for the same training data size. Figure 5b leads to similar results in terms of Forecast Skill, also indicating that both models were able to beat persistence. Once again, the graph-based model outperforms the baseline LASSO model for all assessed configurations, reaching peak results for seven years and Forecast Skill of 30% compared to persistence, whilst LASSO achieved 17%.

From Figure 5, it is also possible to conclude that the data behavior for GraphSAGE shows convergence on its results for eight years of training data. Thus, it is possible to indicate that this dataset size has enough information to provide meaningful results when used by the graph-based model.

3.2. Ozone Concentration Prediction for 1 h Forecast Horizon

The results summarized in Figure 6 were achieved for the prediction of ozone level 1 h in advance for different approaches using GraphSAGE.

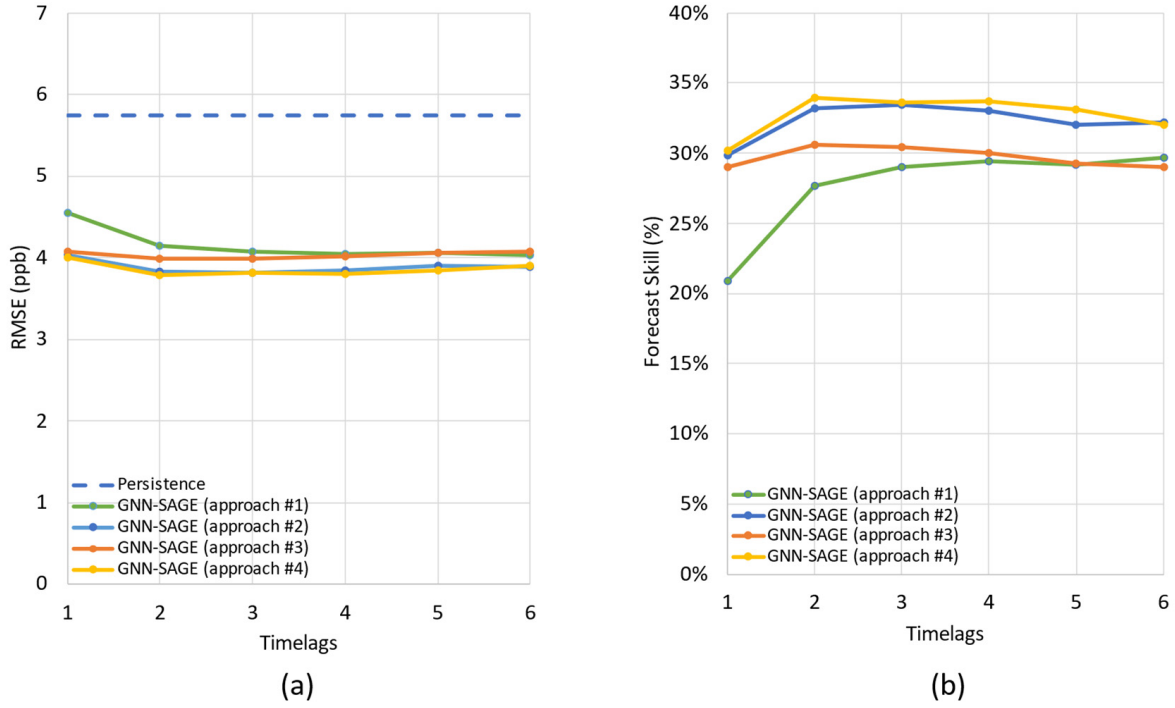

Figure 6. Effect of different number of time-lags on the model's prediction for 1 h forecast horizon in terms of RMSE (**a**) and Forecast Skill (**b**).

In Figure 6a, the worst performance for all time-lags but 6 h was reached using approach 1 (which uses only previous ozone concentration information). On the other hand, approach 4 (focused on using the max number of stations) led to more accurate results for ozone prediction, reaching the lowest RMSE value of 3.80 ppb for 4 h time-lags. Approach 2 (similar to approach 4 but using fewer stations and more weather variables) managed to reach very similar results to approach 4, with slightly worse performance with the use of 1 up to 5 h time-lags and the same result for 6 h time-lags. Interestingly, for approach 3 (which uses VOCs information to predict future ozone concentration), this variation showed little improvement over the model's forecasting capacity, surpassing the number of time-lags assessed by approach 1 for 6 h time-lags.

In pane (b), the Forecast Skill values for each approach show that approach 4 indeed leads to more significant improvement for ozone forecasting. Such an approach improved

the model's accuracy over the persistence model to a maximum of 34%. Visualizing Figure 6b, it is possible to see that, once again, GraphSAGE using approach 1 had the worst performance and approach 3 worsened for further forecasting horizons.

The results provided in Figure 6 indicate that the prediction of future ozone values should not consider only previous values of its concentrations but rather consider external (exogenous) variables such as wind speed, wind direction and nitrogen oxides. However, considering VOCs information led to no significant change in the model's forecasting capacity, but it deteriorated its performance for a longer time-lag value. Figure 7 shows the ozone concentration values forecast by GraphSAGE using approach 4, 4 h time-lags and a validation dataset along an arbitrary time window of the validation dataset. It is also important to point out that each date represents the start of the respective day at 00:00 h.

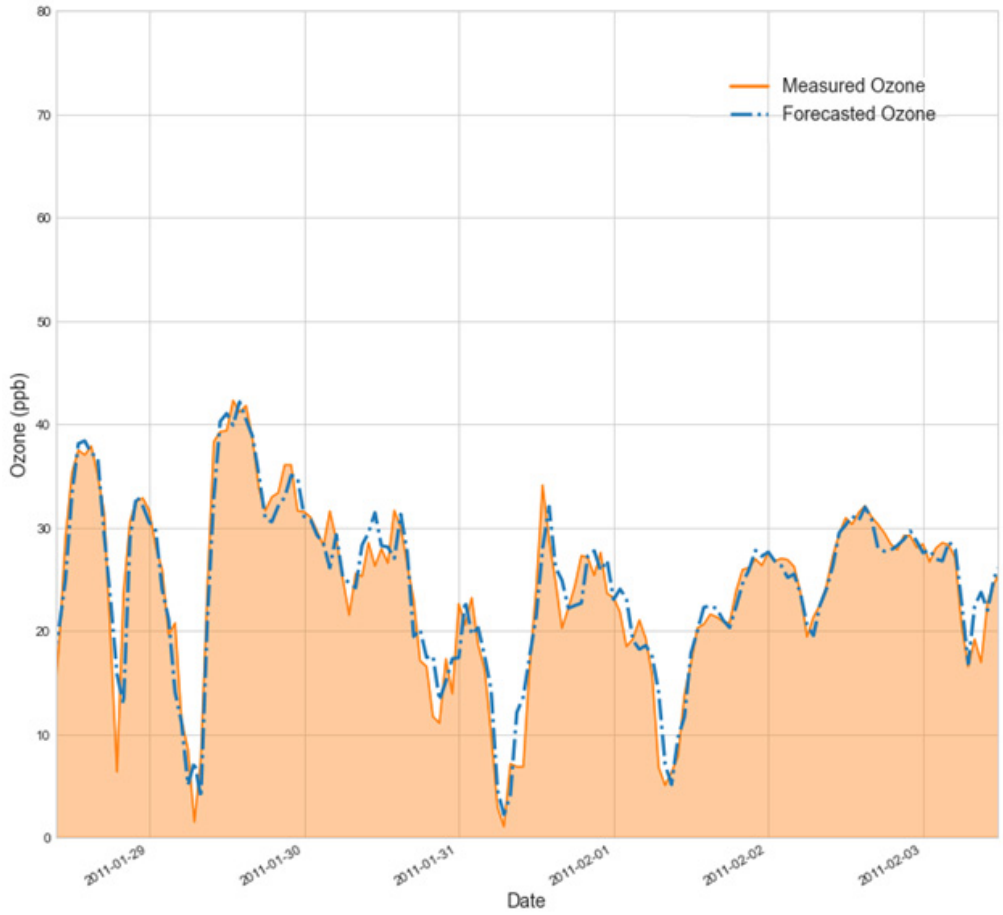

Figure 7. Comparison between forecast ozone values using GraphSAGE and validation dataset and actual measured concentrations for 1 h forecasting horizon.

From Figure 7, it is shown that the graph model can identify ozone peaks during the assessed period, an important feature in checking EPA compliance. It also followed closely the actual data trend over the period, indicating good overall performance. This can be better visualized in Figure 8, where the measured and forecast ozone levels, in ppb, are compared with each other. The histograms in Figure 8 represent the frequency of the statistical probabilistic distribution for the assessed values of ozone, in ppb unit.

From Figure 8, it is easier to observe the good agreement between forecast and predicted ozone values using the proposed GraphSAGE DNN model. The aggregated points around the regression line indicate good agreement between forecast and measured values of ozone level, with RMSE and R^2 being 3.8 ppb and 0.95, respectively.

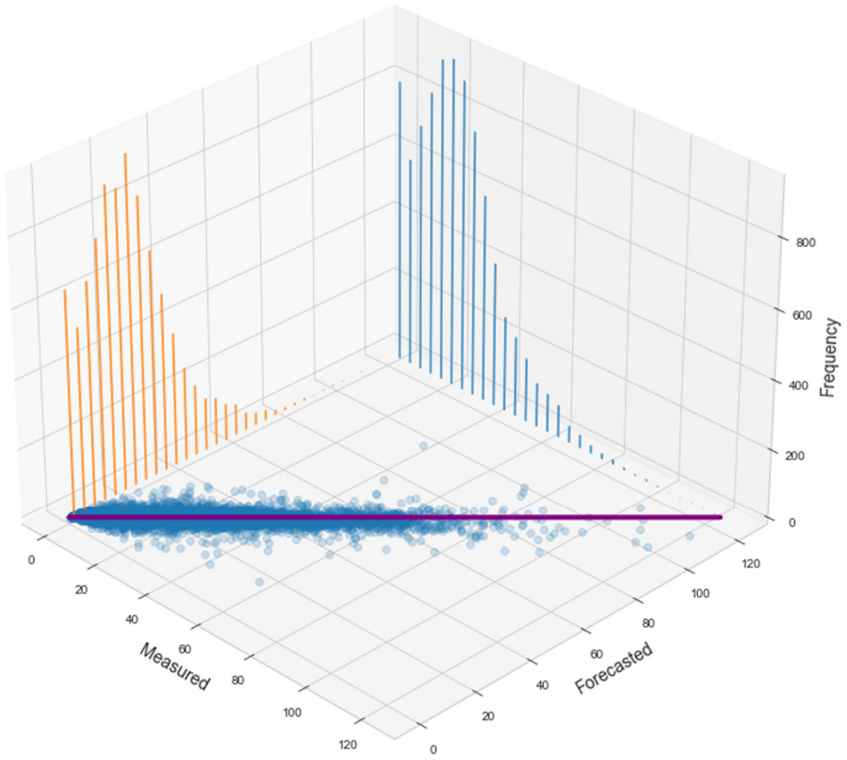

Figure 8. Scatter plot with marginal distribution for GraphSAGE model forecasting ozone 1 h ahead.

3.3. Ozone Concentration Prediction for 3 h Forecast Horizon

The results for ozone estimation considering a 3 h ahead forecasting horizon are presented next. Figure 9 depicts the results achieved using the proposed GraphSAGE model for the four proposed approaches.

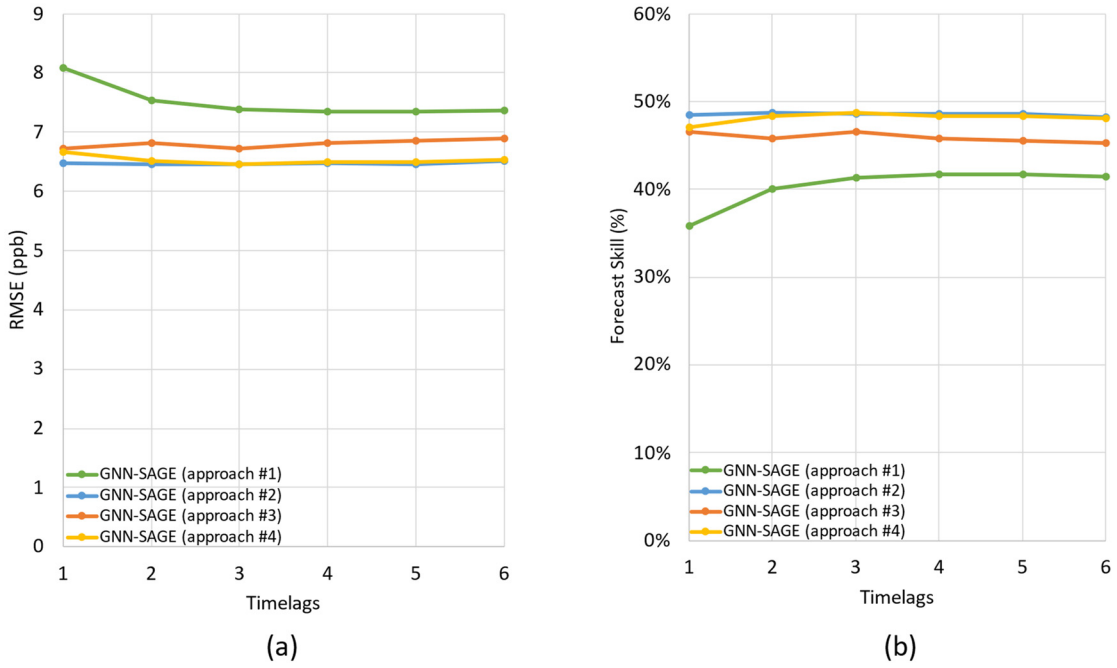

Figure 9. Effect of the different number of time-lags on the model's prediction for 3 h forecast horizon in terms of RMSE (**a**) and Forecast Skill (**b**).

Figure 9, for all time-lags and approaches 2, 3 and 4, demonstrates that increasing the number of time-lags insignificantly improves the proposed GraphSAGE's performance. Again, approach 1 reached the worst RMSE and Forecast Skill values. Approach 1 improved the model's performance from 1 to 3 h time-lags but stabilized for higher values. Approaches 2 and 4 again led to very similar results, approach 2 being the best one with the lowest RMSE and Forecast Skill values of 6.45 ppb and 49% for 2 time-lags, respectively. Approach 3 once more did not show significative variation for the model's performance, but instead showed a degradation in the model's performance starting from 3 h time-lags. This may indicate that information from VOC data adds noise to the model using this specific approach. Figure 10 presents the regression values using the graph model, comparing them with actual measurements over an arbitrary reference period of the validation dataset.

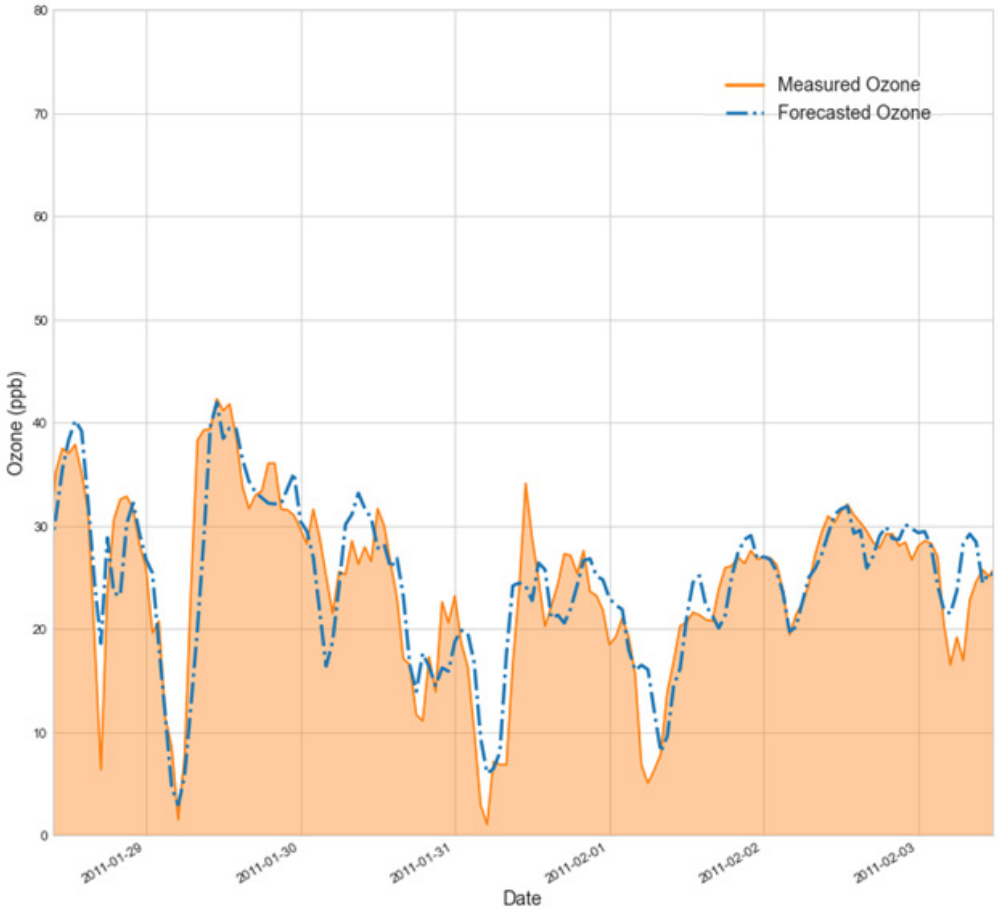

Figure 10. Comparison between forecasted ozone values using GraphSAGE for the validation dataset and actual measured concentrations for 3 h forecasting horizon.

From Figure 10, it is possible to see that the model can still follow the trends on the actual data, but less accurately when compared with Figure 7, for a 1 h ahead forecasting horizon. For this case, some peaks are not properly reproduced, being underestimated by the GraphSAGE model; this can be seen for the period between 31 January and 1 February. This is also reflected in the scatter plot depicted in Figure 11.

As previously stated, as expected, the model's performance was inferior to the one for the 1 h forecast horizon. Figure 11 shows that for the 3 h ahead horizon, there is still good agreement between real and forecast ozone concentrations, but the points are more dispersed over the plot area. This sparser distribution leads to higher RMSE and lower R^2, which reached the respective values of 6.45 ppb and 0.84.

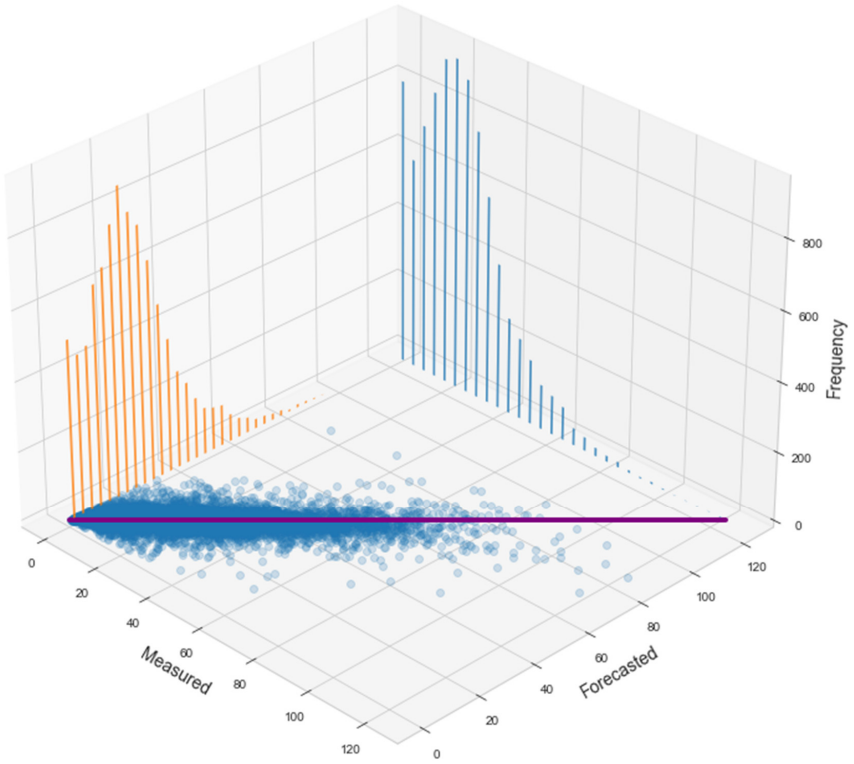

Figure 11. Scatter plot with marginal distribution for GraphSAGE model forecasting ozone 3 h ahead.

3.4. Ozone Concentration Prediction for 6 h Forecast Horizon

The following results regard the case for predicting ozone concentration for 6 h in advance. The effects of the different number of time-lags on the model's prediction for the 6 h forecast horizon are presented in Figure 12 in terms of RMSE (a) and Forecast Skill (b).

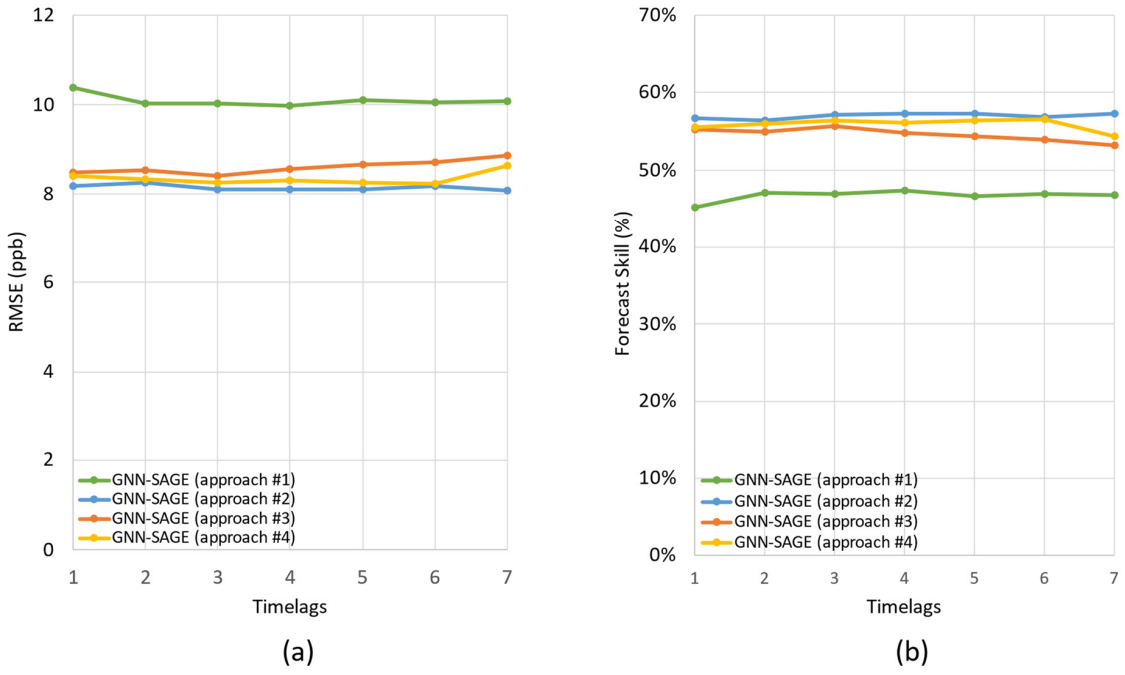

Figure 12. Effect of the different number of time-lags on the model's prediction for 6 h forecast horizon in terms of RMSE (**a**) and Forecast Skill (**b**).

From Figure 12, predicting ozone levels 6 h into the future proved challenging for the model. All approaches, besides number 1, return values for RMSE and Forecast Skill very close together, indicating that the model is not relying on the different features used by each approach, as well as their past values (the model is quite insensible to the number of time-lags). Approach number 2 generated the best results for both assessed metrics, reaching the best values for RMSE and Forecast Skill of 8.09 ppb and 57% for 7 h time-lags, respectively. Approaches 3 and 4 could not provide better results; they degraded the model's accuracy for time-lags greater than 3 h for approach 3 and 6 h for approach 4. The results from approach 1 indicate stable behavior after 3 h time-lags. Figure 13 shows how GraphSAGE forecasting followed the real measured ozone concentration trend.

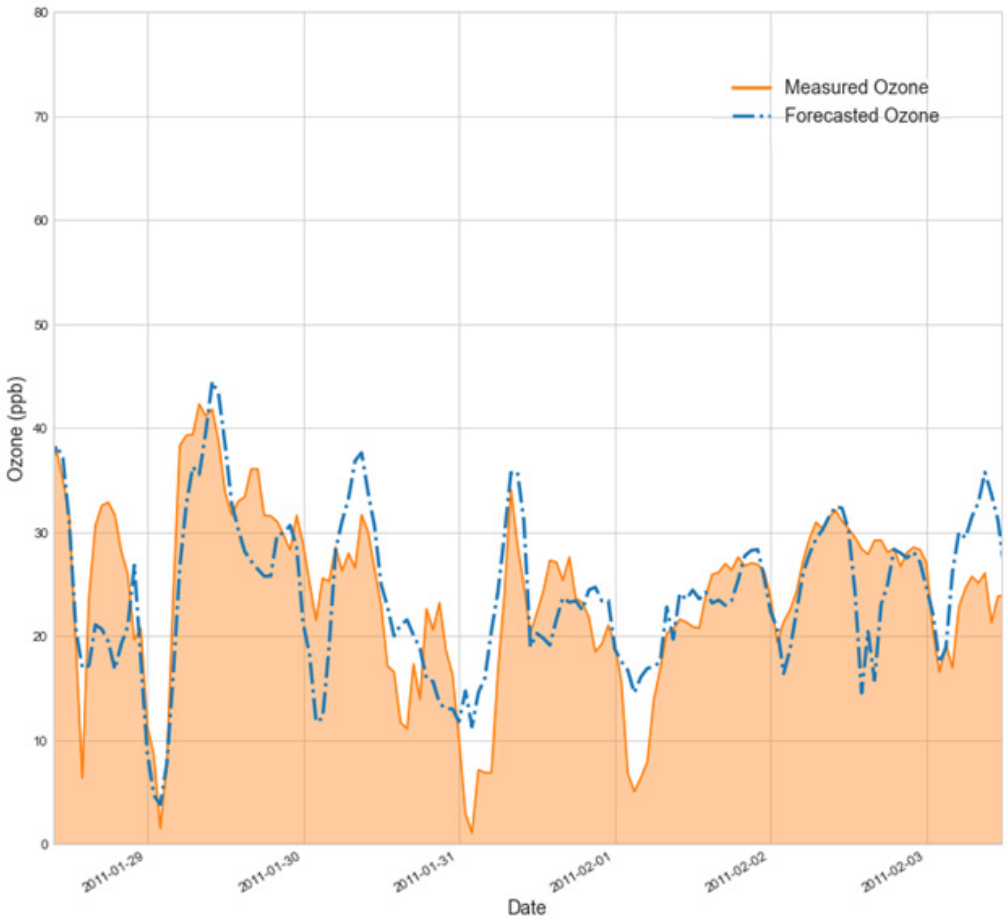

Figure 13. Comparison between forecast ozone values using GraphSAGE and validation dataset and actual measured concentrations for 6 h forecasting horizon.

Again, comparing previous results for 1 and 3 h forecasting horizons, the present case has the weakest performance among the three. For 6 h forecasting horizons, the model mismatches the actual measurements, occasionally underestimating or overestimating peaks. Figure 14 displays the scatter plot for predicted values compared with actual measurements concerning the 6 h forecasting horizon.

The worst model's performance for 6 h forecasting horizon results in dispersed points in Figure 14, and the maximum forecast value is essentially 80 ppb. This 'peak shaving' phenomenon is common for longer time horizon forecasting once the model loses the time connection of the data and starts tending to give average values as output. Again, this leads to increased RMSE, 8.08 ppb and reduced R^2, 0.75.

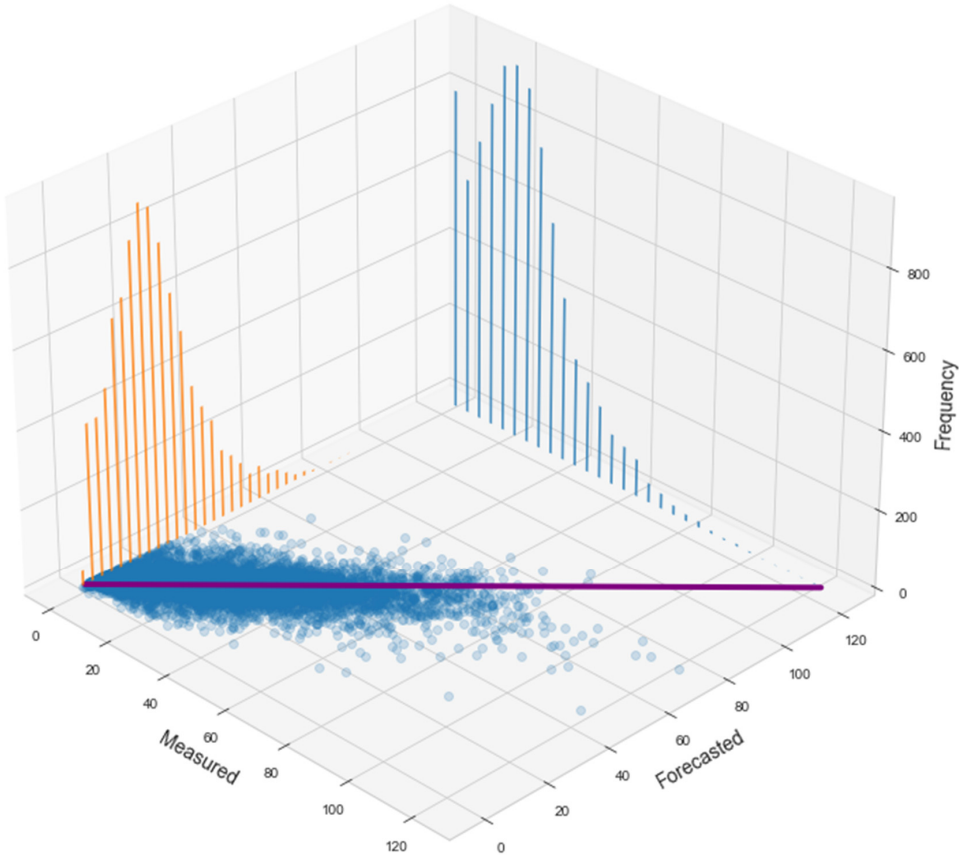

Figure 14. Scatter plot with marginal distribution for GraphSAGE model forecasting ozone 6 h ahead.

3.5. Importance of Different Input Attributes to Ozone Forecasting

Figure 15a–c shows for every studied forecasting horizon how important the attributes are for the GraphSAGE model for each case, accordingly to SHAP analysis considering the best results setup for each forecast horizon.

In Figure 15, the variables are ordered in a descending order from the most influential to the least influential attribute to predict future ozone levels. The right-side bar represents the feature value regarding its correlation to the model's output, where the higher the value the higher is the correlation between the variable and the forecast ozone level. Positive SHAP values indicate that positive values for the assessed variable contribute positively for the model's prediction, while negative values have negative influence over the model's output.

From Figure 15, the most influential attribute for (a) and (b) is the past hour's information of ozone data from the reference station, indicated by "Ozone_L(1)", while in pane (c) the variable "sin(HoD)" takes the lead as the most influential. The importance of the HoD variable for ozone prediction is due to its cyclic nature, reaching maximum values during daytime due to photochemical reactions, and minimum values during nighttime due to deposit and elimination processes [72,73]. Data from the previous 1 h information for ozone concentration, e.g., "Ozone_L(1)", have a strong influence over the model's performance, indicating that information from further in the past adds little to future estimations due to its volatile behavior. It is also visible that past ozone information coming from stations other than the reference one, e.g., "Ozone_L(1)_Station(1)", also plays an important role in the model's predictive performance. For all forecasting horizons, ozone concentration from surrounding stations is always present on the top part of the SHAP plot, indicating a strong influence over the model.

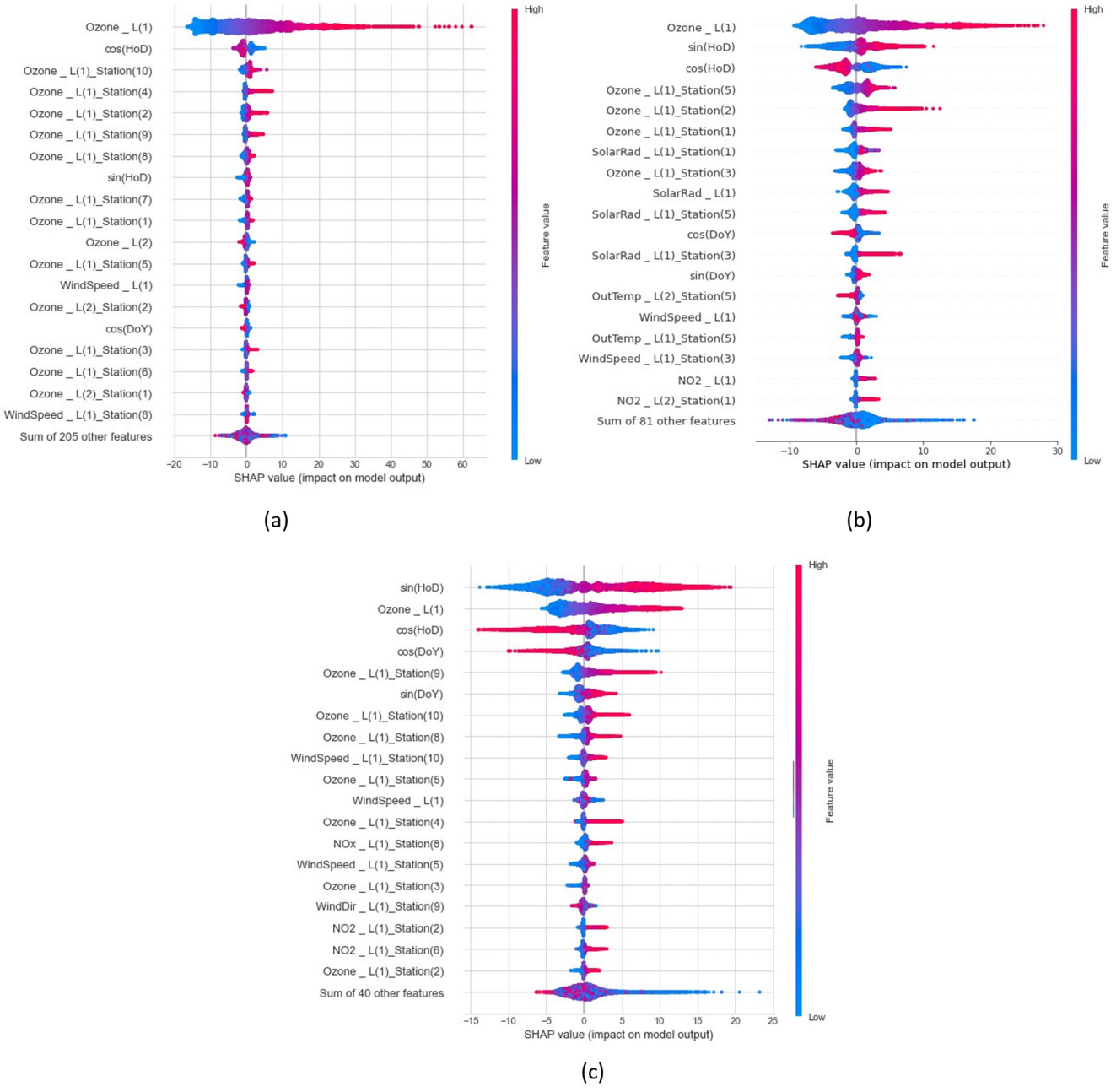

Figure 15. Influence of different attributes for 1 h (**a**), 3 h (**b**) and 6 h (**c**) forecasting horizons.

Still, in Figure 15, wind speed is another influential variable on ozone prediction by the GraphSAGE model for all forecasting horizons. For 1, 3 and 6 h horizons, the best results were achieved by approach 4, approach 2 and again approach 2, respectively. For the 1 h forecast, wind speed is the most influential weather variable, according to SHAP analysis. For 3 and 6 h, solar radiation is the most relevant weather variable, indicating that iterations between the Sun's radiation and ozone are more relevant to the model's predictive performance due to the daily photochemical reactions between ozone and solar radiation. This may explain the superior performance of approach 2 over the remaining ones for longer forecasting windows.

3.6. Model's Performance for the Determination of Compliance with EPA Regulations

To further assess the proposed GraphSAGE performance, the present work also considered EPA regulations concerning ozone levels in the atmosphere. The EPA states that

a region is non-attained when the fourth-highest daily value for the daily maximum 8 h average of a 3 consecutive year average surpasses the threshold of 70 ppb. To verify the station's compliance to EPA regulations through GraphSAGE assessment, one year of data was used instead of the three consecutive years. The motivation behind this change was that more precise and detailed results from the graph model were desired and the three-year period average would lead to the same results as if measured yearly. The results for this analysis are presented in Figure 16, where the dotted red line represents the 70 ppb limit.

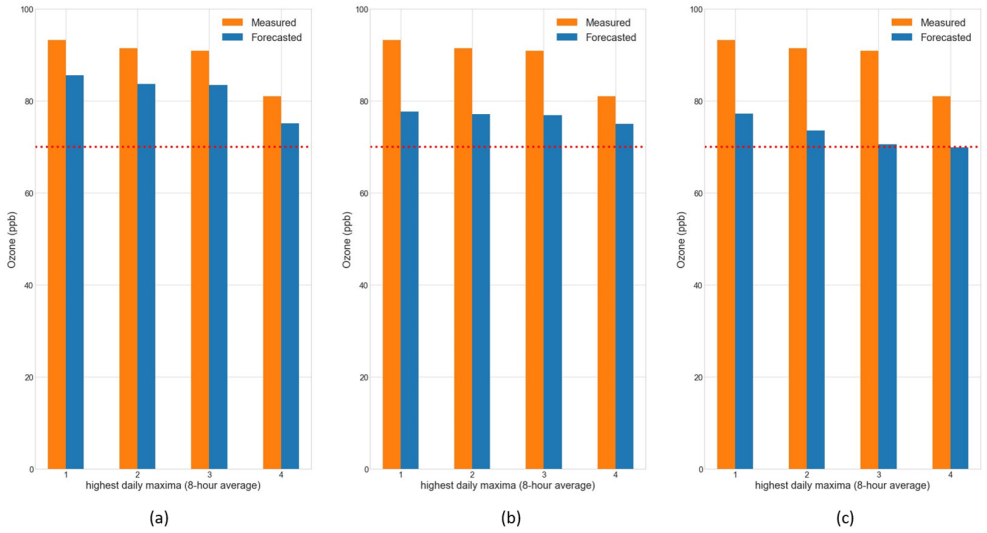

Figure 16. Determination of nonattainment status for reference station for 1 h (**a**), 3 h (**b**) and 6 h (**c**) forecasting horizons.

In Figure 16a, the proposed GraphSAGE model successfully classified the reference station as non-attained, even in the presence of a slight underestimation of ozone concentrations for every daily maximum. The same result was achieved for the 3 h forecasting horizon in panel (b), where the applied graph model was able to determine the nonattainment condition for the reference station. Lastly, in Figure 16c, the model failed by a tiny amount to correctly attribute the studied station as non-attained for the fourth-highest daily maxima, achieving the value of 69.90 ppb. However, this marginal difference does not mitigate the overall good model performance, since it was able to correctly predict the station's non-attained condition for 1 and 3 h ahead.

4. Discussion

The GraphSAGE model, developed in this research, proved to be a viable tool for assessing future ozone concentrations in urban areas. The graph structure allows better modeling for ozone, considering both its spatial and temporal characteristics. For the 1 h forecasting horizon, the model achieved the best overall performance. For this case of short-term prediction, approach 4 provided the best results. For such a configuration, the regression line comparing predicted and estimated values showed excellent agreement, with RMSE and R^2 values of 3.8 ppb and 0.95, respectively. Using time-lags greater than 4 h did not improve the models' performance.

The SHAP analysis showed that for this forecast horizon, wind speed acts as an essential variable for the model's performance, alongside spatiotemporal data from the surrounding stations and time/seasonal data such as HoD and DoY. The results achieved from SHAP analysis are in accordance with the ones found in the literature. In work [74], SHAP analysis identified that for ozone estimation DoY played a major role as a variable over their model prediction. However, results for horizons greater than 1 h were improved when approach number 2 was used, indicating that solar radiation starts to play an essential role in the model's prediction, and variables DoY and HoD are not enough to satisfactorily

interpret the cyclic photochemical reactions between Sun and ozone [72,73]. Concerning weather variables, they have been proved to exert an important influence over ozone levels [75–79], although they vary accordingly to each study. The importance of weather variables may indicate why approach 1 did not achieve satisfactory results when compared against the others, since it lacked weather data. Work [80] attested the importance of VOCs over ozone levels. However, approach 3, which uses VOC data, did not achieve results as good as approaches 2 and 4. This can be explained by data availability for this approach, since VOCs information limited the other input variables and the number of stations, resulting in the smallest set of stations.

According to EPA regulation, the nonattainment condition for the reference station was also tested using the proposed GraphSAGE model. In this sense, one year of data was evaluated considering 1, 3 and 6 h forecasting horizons. For 1 and 3 h ahead, GraphSAGE successfully attested the nonattainment condition over the reference station. For the remaining horizon, it did not identify the nonattainment state for the reference location. Nevertheless, GraphSAGE's performance is not eclipsed by this result alone since the difference between the forecasted and actual values was minimal.

The developed GraphSAGE models proved to be able to reach good predictive results and can determine the future condition of the reference station in advance. To verify their performance within the ozone forecasting field, results drawn from the literature were compiled to be compared. Table 1 summarizes the model's results for RMSE, Forecast Skill and R^2, while Table 2 compiles results for ozone prediction studies from the literature.

Table 1. Summary of results for ozone prediction using GraphSAGE model.

Forecast Horizon	RMSE	Forecast Skill	R^2
1 h	3.80 ppb	33.7%	0.95
3 h	6.45 ppb	48.7%	0.84
6 h	8.09 ppb	57.1%	0.75

Table 2. Literature values for ozone prediction.

Model	Metric Value	Author
Double attention recurrent neural network	RMSE (R^2) 7.71 ppb (0.96) for 1 h horizon 10.95 ppb (0.91) for 3 h horizon 14.11 ppb (0.86) for 6 h horizon	(Zhang et al., 2023) [81]
Attention-based sequence to sequence model	RMSE 12.40 ppb for 1 h horizon 22.87 ppb for 3 h horizon 30.62 ppb for 6 h horizon	(Jia et al., 2021) [82]
Diffusion convolutional recurrent neural network	RMSE 9.35 ppb for 1 h horizon during Winter and Spring seasons 11.03 ppb for 1 h horizon during Summer and autumn seasons	(Wang et al., 2022) [83]
Combination between WRF, CMAQ and LSTM models	RMSE 7.09 ppb for 6 h horizon	(Sun et al., 2021) [84]
Model uses multiple linear regression-based XGBoost	RMSE 12.92 ppb for 1 h horizon	(Nabavi et al., 2021) [85]

Compared with the attention-based models proposed in [81,82], GraphSAGE has proved to reach better RMSE values for all the forecasting horizons, achieving a mean improvement of 39% for the former and over 72% for the later. Still, concerning work [81], the graph-based model did not surpass the reference one considering the coefficient of determination R^2, which indicates better modeled values by the reference model. This, however, does not mitigate the superior performance for GraphSAGE regarding RMSE.

Comparing this study's results against deep learning approaches in [83–85], once again, the graph model proved to be of competitive or superior performance. Although it is not possible to compare GraphSAGE results accordingly to specific seasons, the included attribute of DoY should consider, in part, seasonal variations over ozone concentration. This way, in [83], GraphSAGE proved to capture spatiotemporal relations underlying ozone modeling better, improving ozone prediction by 59.36% over the reference model for winter and spring, and was 65.55% superior for summer and autumn for the 1 h horizon, indicating better generalization by the GraphSAGE model. Comparing our model with reference [84], which is based on LSTM and CTM, for the 6 h forecasting horizon GraphSAGE did not surpass this reference's result. However, the values are still in the same range, proving that the graph model offers competitive results for ozone forecasting. Lastly, comparing the XGBoost-based reference model in [85], RMSE improved by 71% for the assessed horizon.

5. Conclusions

In the present work, a graph-based model named GraphSAGE was developed to assess future ozone concentration. Using historical data from Houston, Texas, ranging from 2011 to 2019, the model was trained and tested for different dataset configurations, variables, time-lags and forecast horizons. Four different approaches were tested to obtain the best prediction for ozone: the first approach considered past ozone concentration; approach 2 considered NO_2/NO_x and weather information (wind speed, wind direction, outdoor temperature, relative humidity, solar radiation); approach 3 considered VOCs information to assess future ozone concentration; and approach 4 considered the greatest number of weather stations containing NO_2/NO_X, weather and HoD/DoY information.

The resultant values for predicted ozone were assessed using the persistence model as benchmarking, alongside RMSE, R^2 and Forecast Skill metrics. The results showed that for the 1 h horizon, approach 4 led to better results for 4 h time-lags, indicating that further increasing this parameter would rather add noise to the model. For 3 and 6 h horizons, approach 2 returned the best metric values. The respective time-lags for these horizons were 2 and 7 h. The RMSE and R^2 values were 3.80 ppb and 0.95 for the 1 h forecasting horizon, 6.45 ppb and 0.84 for the 3 h horizon and 8.09 ppb and 0.75 for the 6 h forecasting horizon. Compared with the persistence benchmarking model, Forecast Skill showed improvement for GraphSAGE of 33.7%, 48.7% and 57.1% for 1, 3 and 6 h horizons, in that order.

The model's performance was also evaluated for its capacity to attest to the nonattainment condition for the reference station. Concerning 1 and 3 h horizons, GraphSAGE successfully predicted the nonattainment state for the reference location. For 6 h ahead forecasting, it failed to correctly predict such a state by a minimal difference, which does not mitigate the overall good model performance. Lastly, the graph-based model also showed good performance when compared with the literature results. It improved ozone forecasting by 72% when compared with attention-based models and 71% when compared with XGBoost-based models. The results place GraphSAGE as the state-of-the-art model for pollutant assessment.

Author Contributions: Conceptualization, J.V.G.T. and B.G; methodology, P.A.C.R., J.V.G.T. and B.G.; software, P.A.C.R.; validation, P.A.C.R., J.V.G.T. and B.G; formal analysis, P.A.C.R.; investigation, P.A.C.R., J.V.G.T. and B.G.; resources, J.V.G.T. and B.G.; data curation, J.V.G.T. and B.G.; writing—original draft preparation, V.O.S., P.A.C.R. and J.S.; writing—review and editing, V.O.S., P.A.C.R., J.S., J.V.G.T. and B.G.; visualization, V.O.S. and P.A.C.R.; supervision, J.V.G.T. and B.G.; project administration, J.V.G.T. and B.G.; funding acquisition, B.G. and J.V.G.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) Alliance, Grant No. 401643, in association with Lakes Environmental Software Inc., and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico—Brasil (CNPq), Grant No. 305456/2019-9.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The used data were acquired from https://www17.tceq.texas.gov/tamis/index.cfm (accessed on 26 January 2023). The used algorithm can be downloaded from https://drive.google.com/drive/folders/1mJL3i8iVOXmuvaxt_T8ValLIK31RLVHm?usp=sharing (accessed on 26 January 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, Z.; Li, G.; Huang, J.; Wang, Z.; Pan, X. Impact of Air Pollution Waves on the Burden of Stroke in a Megacity in China. *Atmos. Environ.* **2019**, *202*, 142–148. [CrossRef]
2. Statistical Review of World Energy—Energy Economics—Home. Available online: https://www.bp.com/en/global/corporate/energy-economics/statistical-review-of-world-energy.html (accessed on 7 December 2022).
3. Croze, M.L.; Zimmer, L. Ozone Atmospheric Pollution and Alzheimer's Disease: From Epidemiological Facts to Molecular Mechanisms. *JAD* **2018**, *62*, 503–522. [CrossRef] [PubMed]
4. Lin, C.-C.; Chiu, C.-C.; Lee, P.-Y.; Chen, K.-J.; He, C.-X.; Hsu, S.-K.; Cheng, K.-C. The Adverse Effects of Air Pollution on the Eye: A Review. *Int. J. Environ. Res. Public Health* **2022**, *19*, 1186. [CrossRef]
5. Sivarethinamohan, R. Impact of Air Pollution in Health and Socio-Economic Aspects: Review on Future Approach. *Mater. Today Proc.* **2021**, *37*, 2725–2729. [CrossRef]
6. Li, G.; Lu, H.; Hu, W.; Liu, J.; Hu, M.; He, J.; Huang, F. Outdoor Air Pollution Enhanced the Association between Indoor Air Pollution Exposure and Hypertension in Rural Areas of Eastern China. *Env. Sci. Pollut. Res.* **2022**, *29*, 74909–74920. [CrossRef]
7. Nazar, W.; Niedoszytko, M. Air Pollution in Poland: A 2022 Narrative Review with Focus on Respiratory Diseases. *Int. J. Environ. Res. Public Health* **2022**, *19*, 895. [CrossRef]
8. Tian, X.; Zhang, C.; Xu, B. The Impact of Air Pollution on Residents' Happiness: A Study on the Moderating Effect Based on Pollution Sensitivity. *Int. J. Environ. Res. Public Health* **2022**, *19*, 7536. [CrossRef]
9. Vohra, K.; Vodonos, A.; Schwartz, J.; Marais, E.A.; Sulprizio, M.P.; Mickley, L.J. Global Mortality from Outdoor Fine Particle Pollution Generated by Fossil Fuel Combustion: Results from GEOS-Chem. *Environ. Res.* **2021**, *195*, 110754. [CrossRef]
10. Huang, J.; Pan, X.; Guo, X.; Li, G. Impacts of Air Pollution Wave on Years of Life Lost: A Crucial Way to Communicate the Health Risks of Air Pollution to the Public. *Environ. Int.* **2018**, *113*, 42–49. [CrossRef]
11. Lelieveld, J.; Evans, J.S.; Fnais, M.; Giannadaki, D.; Pozzer, A. The Contribution of Outdoor Air Pollution Sources to Premature Mortality on a Global Scale. *Nature* **2015**, *525*, 367–371. [CrossRef]
12. Perera, F.; Nadeau, K. Climate Change, Fossil-Fuel Pollution, and Children's Health. *N. Engl. J. Med.* **2022**, *386*, 2303–2314. [CrossRef]
13. Balogun, A.-L.; Tella, A.; Baloo, L.; Adebisi, N. A Review of the Inter-Correlation of Climate Change, Air Pollution and Urban Sustainability Using Novel Machine Learning Algorithms and Spatial Information Science. *Urban Clim.* **2021**, *40*, 100989. [CrossRef]
14. IPCC. *Global Warming of 1.5 °C: IPCC Special Report on Impacts of Global Warming of 1.5 °C above Pre-Industrial Levels in Context of Strengthening Response to Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*, 1st ed.; Cambridge University Press: Cambridge, UK, 2022; ISBN 978-1-00-915794-0.
15. Li, Y.; Shang, J.; Zhang, C.; Zhang, W.; Niu, L.; Wang, L.; Zhang, H. The Role of Freshwater Eutrophication in Greenhouse Gas Emissions: A Review. *Sci. Total Environ.* **2021**, *768*, 144582. [CrossRef]
16. Mikhaylov, A.; Moiseev, N.; Aleshin, K.; Burkhardt, T. Global Climate Change and Greenhouse Effect. *Entrep. Sustain. Issues* **2020**, *7*, 2897–2913. [CrossRef]
17. Fisher, S.; Bellinger, D.C.; Cropper, M.L.; Kumar, P.; Binagwaho, A.; Koudenoukpo, J.B.; Park, Y.; Taghian, G.; Landrigan, P.J. Air Pollution and Development in Africa: Impacts on Health, the Economy, and Human Capital. *Lancet Planet. Health* **2021**, *5*, e681–e688. [CrossRef]
18. Errigo, I.M.; Abbott, B.W.; Mendoza, D.L.; Mitchell, L.; Sayedi, S.S.; Glenn, J.; Kelly, K.E.; Beard, J.D.; Bratsman, S.; Carter, T.; et al. Human Health and Economic Costs of Air Pollution in Utah: An Expert Assessment. *Atmosphere* **2020**, *11*, 1238. [CrossRef]
19. Jakubowska, A.; Rabe, M. Air Pollution and Limitations in Health: Identification of Inequalities in the Burdens of the Economies of the "Old" and "New" EU. *Energies* **2022**, *15*, 6225. [CrossRef]
20. Dechezleprêtre, A.; Rivers, N.; Stadler, B. *The Economic Cost of Air Pollution: Evidence from Europe*; OECD Economics Department Working Papers: Paris, France, 2019; Volume 1584.
21. Pandey, A.; Brauer, M.; Cropper, M.L.; Balakrishnan, K.; Mathur, P.; Dey, S.; Turkgulu, B.; Kumar, G.A.; Khare, M.; Beig, G.; et al. Health and Economic Impact of Air Pollution in the States of India: The Global Burden of Disease Study. *Lancet Planet Health* **2021**, *5*, e25–e38. [CrossRef]
22. Chen, Z.; Wang, F.; Liu, B.; Zhang, B. Short-Term and Long-Term Impacts of Air Pollution Control on China's Economy. *Environ. Manag.* **2022**, *70*, 536–547. [CrossRef]

23. Steinebach, Y. Instrument Choice, Implementation Structures, and the Effectiveness of Environmental Policies: A Cross-National Analysis. *Regul. Gov.* **2022**, *16*, 225–242. [CrossRef]
24. Senthilkumar, N.; Gilfether, M.; Metcalf, F.; Russell, A.G.; Mulholland, J.A.; Chang, H.H. Application of a Fusion Method for Gas and Particle Air Pollutants between Observational Data and Chemical Transport Model Simulations Over the Contiguous United States for 2005. *Int. J. Environ. Res. Public Health* **2019**, *16*, 3314. [CrossRef] [PubMed]
25. Liu, H.; Yan, G.; Duan, Z.; Chen, C. Intelligent Modeling Strategies for Forecasting Air Quality Time Series: A Review. *Appl. Soft Comput.* **2021**, *102*, 106957. [CrossRef]
26. Wang, L.; Zhao, Y.; Shi, J.; Ma, J.; Liu, X.; Han, D.; Gao, H.; Huang, T. Predicting Ozone Formation in Petrochemical Industrialized Lanzhou City by Interpretable Ensemble Machine Learning. *Environ. Pollut.* **2023**, *318*, 120798. [CrossRef] [PubMed]
27. Friberg, M.D.; Zhai, X.; Holmes, H.A.; Chang, H.H.; Strickland, M.J.; Sarnat, S.E.; Tolbert, P.E.; Russell, A.G.; Mulholland, J.A. Method for Fusing Observational Data and Chemical Transport Model Simulations To Estimate Spatiotemporally Resolved Ambient Air Pollution. *Environ. Sci. Technol.* **2016**, *50*, 3695–3705. [CrossRef] [PubMed]
28. Sayeed, A.; Choi, Y.; Jung, J.; Lops, Y.; Eslami, E.; Salman, A.K. A Deep Convolutional Neural Network Model for Improving WRF Simulations. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–11. [CrossRef]
29. Ballesteros-González, K.; Sullivan, A.P.; Morales-Betancourt, R. Estimating the Air Quality and Health Impacts of Biomass Burning in Northern South America Using a Chemical Transport Model. *Sci. Total Environ.* **2020**, *739*, 139755. [CrossRef]
30. López-Noreña, A.I.; Berná, L.; Tames, M.F.; Millán, E.N.; Puliafito, S.E.; Fernandez, R.P. Influence of Emission Inventory Resolution on the Modeled Spatio-Temporal Distribution of Air Pollutants in Buenos Aires, Argentina, Using WRF-Chem. *Atmos. Environ.* **2022**, *269*, 118839. [CrossRef]
31. Mazzeo, A.; Zhong, J.; Hood, C.; Smith, S.; Stocker, J.; Cai, X.; Bloss, W.J. Modelling the Impact of National vs. Local Emission Reduction on PM2.5 in the West Midlands, UK Using WRF-CMAQ. *Atmosphere* **2022**, *13*, 377. [CrossRef]
32. Sayeed, A.; Lops, Y.; Choi, Y.; Jung, J.; Salman, A.K. Bias Correcting and Extending the PM Forecast by CMAQ up to 7 Days Using Deep Convolutional Neural Networks. *Atmos. Environ.* **2021**, *253*, 118376. [CrossRef]
33. Cabaneros, S.M.; Calautit, J.K.; Hughes, B.R. A Review of Artificial Neural Network Models for Ambient Air Pollution Prediction. *Environ. Model. Softw.* **2019**, *119*, 285–304. [CrossRef]
34. Kašpar, V.; Zapletal, M.; Samec, P.; Komárek, J.; Bílek, J.; Juráň, S. Unmanned Aerial Systems for Modelling Air Pollution Removal by Urban Greenery. *Urban For. Urban Green.* **2022**, *78*, 127757. [CrossRef]
35. Tian, C.; Niu, T.; Wei, W. Developing a Wind Power Forecasting System Based on Deep Learning with Attention Mechanism. *Energy* **2022**, *257*, 124750. [CrossRef]
36. Rocha, P.A.C.; Santos, V.O. Global Horizontal and Direct Normal Solar Irradiance Modeling by the Machine Learning Methods XGBoost and Deep Neural Networks with CNN-LSTM Layers: A Case Study Using the GOES-16 Satellite Imagery. *Int. J. Energy Environ. Eng.* **2022**, *13*, 1271–1286. [CrossRef]
37. O'Donncha, F.; Hu, Y.; Palmes, P.; Burke, M.; Filgueira, R.; Grant, J. A Spatio-Temporal LSTM Model to Forecast across Multiple Temporal and Spatial Scales. *Ecol. Inform.* **2022**, *69*, 101687. [CrossRef]
38. Bashir Shaban, K.; Kadri, A.; Rezk, E. Urban Air Pollution Monitoring System With Forecasting Models. *IEEE Sens. J.* **2016**, *16*, 2598–2606. [CrossRef]
39. Zhan, Y.; Luo, Y.; Deng, X.; Grieneisen, M.L.; Zhang, M.; Di, B. Spatiotemporal Prediction of Daily Ambient Ozone Levels across China Using Random Forest for Human Exposure Assessment. *Environ. Pollut.* **2018**, *233*, 464–473. [CrossRef]
40. Juarez, E.K.; Petersen, M.R. A Comparison of Machine Learning Methods to Forecast Tropospheric Ozone Levels in Delhi. *Atmosphere* **2021**, *13*, 46. [CrossRef]
41. Seng, D.; Zhang, Q.; Zhang, X.; Chen, G.; Chen, X. Spatiotemporal Prediction of Air Quality Based on LSTM Neural Network. *Alex. Eng. J.* **2021**, *60*, 2021–2032. [CrossRef]
42. Zhu, S.; Xu, J.; Zeng, J.; Yu, C.; Wang, Y.; Yan, H. Satellite-Derived Estimates of Surface Ozone by LESO: Extended Application and Performance Evaluation. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *113*, 103008. [CrossRef]
43. Gilik, A.; Ogrenci, A.S.; Ozmen, A. Air Quality Prediction Using CNN+LSTM-Based Hybrid Deep Learning Architecture. *Env. Sci. Pollut. Res.* **2022**, *29*, 11920–11938. [CrossRef]
44. Veličković, P.; Cucurull, G.; Casanova, A.; Romero, A.; Liò, P.; Bengio, Y. Graph Attention Networks. In Proceedings of the International Conference on Learning (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
45. Zhang, K.; Yang, X.; Cao, H.; Thé, J.; Tan, Z.; Yu, H. Multi-Step Forecast of PM2.5 and PM10 Concentrations Using Convolutional Neural Network Integrated with Spatial–Temporal Attention and Residual Learning. *Environ. Int.* **2023**, *171*, 107691. [CrossRef] [PubMed]
46. Wilson, T.; Tan, P.-N.; Luo, L. A Low Rank Weighted Graph Convolutional Approach to Weather Prediction. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018; pp. 627–636.
47. Zhang, S.; Tong, H.; Xu, J.; Maciejewski, R. Graph Convolutional Networks: A Comprehensive Review. *Comput. Soc. Netw.* **2019**, *6*, 11. [CrossRef]
48. Qi, Y.; Li, Q.; Karimian, H.; Liu, D. A Hybrid Model for Spatiotemporal Forecasting of PM2.5 Based on Graph Convolutional Neural Network and Long Short-Term Memory. *Sci. Total Environ.* **2019**, *664*, 1–10. [CrossRef] [PubMed]
49. Mao, W.; Jiao, L.; Wang, W. Long Time Series Ozone Prediction in China: A Novel Dynamic Spatiotemporal Deep Learning Approach. *Build. Environ.* **2022**, *218*, 109087. [CrossRef]

50. Wang, S.; Qiao, L.; Fang, W.; Jing, G.; Sheng, V.; Zhang, Y. Air Pollution Prediction Via Graph Attention Network and Gated Recurrent Unit. *Comput. Mater. Contin.* **2022**, *73*, 673–687. [CrossRef]
51. Hamilton, W.L.; Ying, R.; Leskovec, J. Inductive Representation Learning on Large Graphs. In Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, 4–9 December 2018; pp. 1024–1034.
52. Pan, S.; Roy, A.; Choi, Y.; Eslami, E.; Thomas, S.; Jiang, X.; Gao, H.O. Potential Impacts of Electric Vehicles on Air Quality and Health Endpoints in the Greater Houston Area in 2040. *Atmos. Environ.* **2019**, *207*, 38–51. [CrossRef]
53. Sadeghi, B.; Choi, Y.; Yoon, S.; Flynn, J.; Kotsakis, A.; Lee, S. The Characterization of Fine Particulate Matter Downwind of Houston: Using Integrated Factor Analysis to Identify Anthropogenic and Natural Sources. *Environ. Pollut.* **2020**, *262*, 114345. [CrossRef]
54. EPA, U. Green Book | US EPA. Available online: https://www3.epa.gov/airquality/greenbook/jnc.html (accessed on 16 December 2022).
55. Vizuete, W.; Nielsen-Gammon, J.; Dickey, J.; Couzo, E.; Blanchard, C.; Breitenbach, P.; Rasool, Q.Z.; Byun, D. Meteorological Based Parameters and Ozone Exceedances in Houston and Other Cities in Texas. *J. Air Waste Manag. Assoc.* **2022**, *72*, 969–984. [CrossRef]
56. Baïle, R.; Muzy, J.-F. Leveraging Data from Nearby Stations to Improve Short-Term Wind Speed Forecasts. *Energy* **2023**, *263*, 125644. [CrossRef]
57. Yang, D.; Kleissl, J.; Gueymard, C.A.; Pedro, H.T.C.; Coimbra, C.F.M. History and Trends in Solar Irradiance and PV Power Forecasting: A Preliminary Assessment and Review Using Text Mining. *Sol. Energy* **2018**, *168*, 60–101. [CrossRef]
58. Hanifi, S.; Liu, X.; Lin, Z.; Lotfian, S. A Critical Review of Wind Power Forecasting Methods—Past, Present and Future. *Energies* **2020**, *13*, 3764. [CrossRef]
59. Soman, S.S.; Zareipour, H.; Malik, O.; Mandal, P. A Review of Wind Power and Wind Speed Forecasting Methods with Different Time Horizons. In Proceedings of the North American Power Symposium, Arlington, TX, USA, 26–28 September 2010; pp. 1–8.
60. Baïle, R.; Muzy, J.F.; Poggi, P. Short-Term Forecasting of Surface Layer Wind Speed Using a Continuous Random Cascade Model. *Wind Energy* **2011**, *14*, 719–734. [CrossRef]
61. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc. B* **1996**, *58*, 267–288. [CrossRef]
62. Zhou, J.; Cui, G.; Hu, S.; Zhang, Z.; Yang, C.; Liu, Z.; Wang, L.; Li, C.; Sun, M. Graph Neural Networks: A Review of Methods and Applications. *AI Open* **2020**, *1*, 57–81. [CrossRef]
63. Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; Yu, P.S. A Comprehensive Survey on Graph Neural Networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4–24. [CrossRef]
64. Weisberg, S. *Applied Linear Regression*; John Wiley & Sons: New York, NY, USA, 2005; ISBN 978-0-471-70408-9.
65. Lundberg, S.M.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
66. Iban, M.C. An Explainable Model for the Mass Appraisal of Residences: The Application of Tree-Based Machine Learning Algorithms and Interpretation of Value Determinants. *Habitat Int.* **2022**, *128*, 102660. [CrossRef]
67. Fatahi, R.; Nasiri, H.; Dadfar, E.; Chehreh Chelgani, S. Modeling of Energy Consumption Factors for an Industrial Cement Vertical Roller Mill by SHAP-XGBoost: A "Conscious Lab" Approach. *Sci. Rep.* **2022**, *12*, 7543. [CrossRef]
68. Cheng, Y.; Huang, X.-F.; Peng, Y.; Tang, M.-X.; Zhu, B.; Xia, S.-Y.; He, L.-Y. A Novel Machine Learning Method for Evaluating the Impact of Emission Sources on Ozone Formation. *Environ. Pollut.* **2023**, *316*, 120685. [CrossRef]
69. Walia, S.; Kumar, K.; Agarwal, S.; Kim, H. Using XAI for Deep Learning-Based Image Manipulation Detection with Shapley Additive Explanation. *Symmetry* **2022**, *14*, 1611. [CrossRef]
70. Nohara, Y.; Matsumoto, K.; Soejima, H.; Nakashima, N. Explanation of Machine Learning Models Using Shapley Additive Explanation and Application for Real Data in Hospital. *Comput. Methods Programs Biomed.* **2022**, *214*, 106584. [CrossRef]
71. Mun, S.-K.; Chang, M. Development of Prediction Models for the Incidence of Pediatric Acute Otitis Media Using Poisson Regression Analysis and XGBoost. *Environ. Sci. Pollut. Res.* **2022**, *29*, 18629–18640. [CrossRef] [PubMed]
72. Xia, N.; Du, E.; Guo, Z.; de Vries, W. The Diurnal Cycle of Summer Tropospheric Ozone Concentrations across Chinese Cities: Spatial Patterns and Main Drivers. *Environ. Pollut.* **2021**, *286*, 117547. [CrossRef] [PubMed]
73. Chen, L.; Pang, X.; Li, J.; Xing, B.; An, T.; Yuan, K.; Dai, S.; Wu, Z.; Wang, S.; Wang, Q.; et al. Vertical Profiles of O3, NO2 and PM in a Major Fine Chemical Industry Park in the Yangtze River Delta of China Detected by a Sensor Package on an Unmanned Aerial Vehicle. *Sci. Total Environ.* **2022**, *845*, 157113. [CrossRef] [PubMed]
74. Kang, Y.; Choi, H.; Im, J.; Park, S.; Shin, M.; Song, C.-K.; Kim, S. Estimation of Surface-Level NO2 and O3 Concentrations Using TROPOMI Data and Machine Learning over East Asia. *Environ. Pollut.* **2021**, *288*, 117711. [CrossRef] [PubMed]
75. Chen, Z.; Li, R.; Chen, D.; Zhuang, Y.; Gao, B.; Yang, L.; Li, M. Understanding the Causal Influence of Major Meteorological Factors on Ground Ozone Concentrations across China. *J. Clean. Prod.* **2020**, *242*, 118498. [CrossRef]
76. Wang, Z.; Li, J.; Liang, L. Spatio-Temporal Evolution of Ozone Pollution and Its Influencing Factors in the Beijing-Tianjin-Hebei Urban Agglomeration. *Environ. Pollut.* **2020**, *256*, 113419. [CrossRef]
77. Gagliardi, R.V.; Andenna, C. A Machine Learning Approach to Investigate the Surface Ozone Behavior. *Atmosphere* **2020**, *11*, 1173. [CrossRef]
78. Du, J.; Qiao, F.; Lu, P.; Yu, L. Forecasting Ground-Level Ozone Concentration Levels Using Machine Learning. *Resour. Conserv. Recycl.* **2022**, *184*, 106380. [CrossRef]

79. Sadeghi, B.; Ghahremanloo, M.; Mousavinezhad, S.; Lops, Y.; Pouyaei, A.; Choi, Y. Contributions of Meteorology to Ozone Variations: Application of Deep Learning and the Kolmogorov-Zurbenko Filter. *Environ. Pollut.* **2022**, *310*, 119863. [CrossRef]
80. Ma, M.; Yao, G.; Guo, J.; Bai, K. Distinct Spatiotemporal Variation Patterns of Surface Ozone in China Due to Diverse Influential Factors. *J. Environ. Manag.* **2021**, *288*, 112368. [CrossRef]
81. Zhang, Y.; Li, F.; Ni, C.; Gao, S.; Zhang, S.; Xue, J.; Ning, Z.; Wei, C.; Fang, F.; Nie, Y.; et al. Prediction and Cause Investigation of Ozone Based on a Double-Stage Attention Mechanism Recurrent Neural Network. *Front. Environ. Sci. Eng.* **2023**, *17*, 21. [CrossRef]
82. Jia, P.; Cao, N.; Yang, S. Real-Time Hourly Ozone Prediction System for Yangtze River Delta Area Using Attention Based on a Sequence to Sequence Model. *Atmos. Environ.* **2021**, *244*, 117917. [CrossRef]
83. Wang, D.; Wang, H.-W.; Lu, K.-F.; Peng, Z.-R.; Zhao, J. Regional Prediction of Ozone and Fine Particulate Matter Using Diffusion Convolutional Recurrent Neural Network. *Int. J. Environ. Res. Public Health* **2022**, *19*, 3988. [CrossRef]
84. Sun, H.; Fung, J.C.H.; Chen, Y.; Chen, W.; Li, Z.; Huang, Y.; Lin, C.; Hu, M.; Lu, X. Improvement of PM2.5 and O3 Forecasting by Integration of 3D Numerical Simulation with Deep Learning Techniques. *Sustain. Cities Soc.* **2021**, *75*, 103372. [CrossRef]
85. Nabavi, S.O.; Nölscher, A.C.; Samimi, C.; Thomas, C.; Haimberger, L.; Lüers, J.; Held, A. Site-Scale Modeling of Surface Ozone in Northern Bavaria Using Machine Learning Algorithms, Regional Dynamic Models, and a Hybrid Model. *Environ. Pollut.* **2021**, *268*, 115736. [CrossRef]