applied sciences

MDPI

Article

Space-Time Prediction of $PM_{2.5}$ Concentrations in Santiago de Chile Using LSTM Networks

Billy Peralta [1,†], Tomás Sepúlveda [1,†], Orietta Nicolis [1,2,*,†] and Luis Caro [3]

[1] Facultad de Ingeniería, Universidad Andres Bello, Av. Antonio Varas 880, Providencia 7500971, Chile
[2] Research Center for Integrated Disaster Risk Management (CIGIDEN), Av. Vicuña Mackenna 4860, Macul 7820436, Chile
[3] Departamento de Ingeniería Informática, Universidad Católica de Temuco, Temuco 4781312, Chile
* Correspondence: orietta.nicolis@unab.cl
† These authors contributed equally to this work.

Abstract: Currently, air pollution is a highly important issue in society due to its harmful effects on human health and the environment. The prediction of pollutant concentrations in Santiago de Chile is typically based on statistical methods or classical neural networks. Existing methods often assume that historical values are known at a fixed geographic point, such that air pollution can be predicted at a future hour using time series analysis. However, these methods are inapplicable when it is necessary to know the pollutant concentrations at every point of the space. This work proposes a method that addresses the space-time prediction of $PM_{2.5}$ concentration in Santiago de Chile at any spatial points through the use of the LSTM recurrent network model. In particular, by considering historical values of air pollutants ($PM_{2.5}$, PM_{10} and nitrogen dioxide) and meteorological variables (temperature, wind speed and direction and relative humidity), measured at fixed monitoring stations, the proposed model can predict $PM_{2.5}$ concentrations for the next 24 h in a new location where measurements are not available. This work describes the experiments carried out, with particular emphasis on the pre-processing step, which constitutes an important factor for obtaining relatively good results. The proposed multilayer LSTM model obtained R^2 values equal to 0.74 and 0.38 in seven stations when considering forecasts of 1 and 24 h, respectively. As future work, we plan to include more input variables in the proposed model and to use attention-based networks.

Keywords: space-time prediction; pollution model; $PM_{2.5}$; recurrent neural networks

Citation: Peralta, B.; Sepúlveda, T.; Nicolis, O.; Caro, L. Space-Time Prediction of $PM_{2.5}$ Concentrations in Santiago de Chile Using LSTM Networks. *Appl. Sci.* **2022**, *12*, 11317. https://doi.org/10.3390/app122211317

Academic Editor: Yves Rybarczyk and Rasa Zalakeviciute

Received: 7 October 2022
Accepted: 4 November 2022
Published: 8 November 2022

1. Introduction

In recent years, environmental pollution has become a source of great concern for the countries of the world, since every year it causes the premature deaths of approximately 6.5 million people [1]. One of the most important components is particulate matter smaller in diameter than 2.5 microns ($PM_{2.5}$), which is composed of particles small enough to penetrate the lungs. This implies an increased risk of mortality due to harmful effects on the cardiovascular and respiratory systems [2–5]. Main emitters of $PM_{2.5}$ are vehicles, power plants, industrial factories, mining processing centers and houses that use wood or coal as a heating source [6]. Santiago de Chile is one of the most polluted cities in the South America. This is mainly due to its particular geographic position and to a not very restrictive policy on the emission of contaminants. In particular, Santiago is located in a valley with a smooth slope, surrounded by mountains (with altitudes between 1500 and 4000 m) which limit air circulation. In the winter period (from April to August), due to strong thermal inversions and weak winds, the dispersion of atmospheric pollutants is very poor, causing frequent episodes of high pollution [7,8] with negative effects on the health. Evidence of the relation between particular matter pollutants (PM_{10} and $PM_{2.5}$) and mortality is shown by Ostro et al. [9] and Valdés et al. [10], whereas [11] studied the correlation between $PM_{2.5}$ concentrations and children hospitalized for respiratory diseases

in Santiago. In the last few decades, the Ministry of Environment has been applying several restriction measures to emission sources in order to protect human health. Theses measures are based on the Chilean National Standards, which are not so stringent as the international guidelines proposed by WHO [12,13]. For example, for the $PM_{2.5}$, four ranges of increasing concentrations define air quality, starting from the first level (good to moderate) including a maximum 24 h average of $PM_{2.5}$ less than 80 $\mu g/m^3$, to the last level (hazardous) where the 24 h average is greater than 170 $\mu g/m^3$ [14,15]. Instead, the latest WHO guidelines recommended 25 $\mu g/m^3$ (24 h mean) and an annual level of 10 $\mu g/m^3$ [13]. However, some temporary measures implemented by the Chilean government allowed us to reduce the particular matter concentrations of 20%, with a consequent benefit on the the population health in the short run [16]. Measures that could reduce pollution in the short run refer, mainly, to driving restrictions and temporarily shutting down or reducing the usage of stationary emissions sources. A result of [16] was that after three days of an episode of a high-pollution announcement, there were approximately fifteen fewer (cumulative) deaths above the age of 64, and most of that reduction was due to decreases in deaths due to respiratory causes. The impacts of short term exposure to air pollutants were also demonstrated by [17,18] Hence, a good prediction of the pollution levels for the next few hours or days could be very useful for environmental management, since it allows one to alert the population of forthcoming high pollution episodes and to support decision makers to implement restriction measures [14,16]. For example, Catalano et al. [19] integrated two prediction pollution models in traditional traffic management support systems for a sustainable mobility of road vehicles in urban areas; and Liu and Gao [20] reviewed the evidence on greenhouse gas (GHG) mitigation measures and the related health co-benefits, and provided recommendations for further development and implementation of climate change response policies.

Several methods have been proposed in the literature for predicting the $PM_{2.5}$ concentrations; most of them are based on statistical models [6,21–23] and neural networks [24,25]. In general, the existing works for the prediction of $PM_{2.5}$ concentrations in Santiago de Chile using neural networks are based on temporal models [8,14,25–27].

The main difficulty in predicting space-time pollution is due to the fact that air pollutants are strictly correlated with meteorological variables (normally collected at the same spatial points and temporal lags) which are often not available when the prediction at new site has to be made. Additionally, the correlations of the pollutant concentrations with past meteorological variables are often weak, complex and show nonlinear behavior. For this reason, statistical spatio-temporal models which normally include linear relations with exogenous variables cannot be used if predictions or simulations of these variables are not provided (see, for example, [6,28]).

The long short term memory (LSTM) network proposed by Hochreiter and Schmidhuber [29] is a recurrent neural network which is typically used for temporal predictions. However, this model has the limitation that it does not allow predictions to be made at points other than those available during training. Given that in Santiago de Chile, few monitoring stations are available in the town, a space-time approach is necessary to predict pollution at new locations where measurements are not available.

The main aim of this work is considering predictions at spatial points other than those trained, by proposing a recurrent neural network that uses the historic information collected by monitoring stations located at some geographic points for predicting the pollution at new sites for $n-$ahead time steps. In particular, in this work we used the LSTM neural network [29] for the prediction of $PM_{2.5}$ concentrations in the city of Santiago de Chile, considering the measurements of multiple meteorological variables (such as speed and wind direction, temperature and relative humidity) and pollutants (nitrogen dioxide, $PM_{2.5}$ and PM_{10}) collected in the previous 24 h. In other words, the proposed model can be used for forecasting the $PM_{2.5}$ levels at any spatial point by generating a space-time prediction using past information and spatial locations.

The novelties of this work can be summarized in the following two points: (a) the proposal of a space-time machine learning model for predicting the pollution at arbitrary spatial points and at any future time; (b) the application to the pollution prediction in Santiago de Chile, one of the cities with the worst pollution in the world. About the point (a), the existing pollution prediction studies based on neural networks typically made temporal predictions for different spatial points, without considering spatial correlations, and other works used statistical space-time models where knowledge of future values of meteorological variables is assumed. These variables are normally predicted or simulated by complex deterministic models. In this proposal the prediction can be made at each point in the space without knowing the past values at the same point and without the need to run a different model for the prediction or simulation of the exogenous variables, since the past variables of meteorological data are considered as the input of the proposed model. About the point (b), few works have addressed space-time pollution prediction in Santiago de Chile, especially due to the poor availability of monitoring stations. We think that the proposed model could be used by the governmental decision makers for implementing temporary restriction measures in areas where there are not available data, and consequently reducing short term effects on human health due to $PM_{2.5}$ exposure.

2. Related Works

Spatio-temporal prediction is a challenging issue, given that spatial and temporal correlations of data have to be detected simultaneously. Many statistical and computational models have been proposed for dvarious fields, including crime, traffic and transportation, climate and environment monitoring, hydrology and epidemiology [30]. Most of the literature on spatio-temporal models presents statistical approaches based on separable or no separable spatial and temporal covariance structures. An extensive review of geostatistical space-time models for addressing environmental problems (monitoring acid deposition, forecasting precipitation, pollution, etc.) is presented in [31]. Hierarchical spatio-temporal modeling concepts and computational methods are described in [32], considering many issues, including environmental processes and climate trends, besides mapping public-health data and the spread of invasive species. Recently, especially due to the availability of big datasets, computational models based on deep neural networks have arisen for predicting space-time data. Amato et al. [33] introduced a new framework for the spatio-temporal prediction of climate and environmental data by decomposing time series into a basis function and stochastic spatial coefficients. In [34], deep learning (DL), and in particular, recurrent neural networks (RNNs), were implemented for wind speed forecasting, motivated by the use of renewable energy in northeast of the U.S. Bay et al. [35] proposed two adaptive modules for enhancing the graph convolutional network (GCN) for understanding traffic dynamics and predicting the future status of an evolving traffic system. A slightly different approach consists of splitting the whole dataset into several subsets in a hierarchical manner and training a local prediction model for each subset, as in Shang et al. [36], or considering a decomposition of the original series, as proposed by Wang et al. [37].

Regarding the prediction of $PM_{2.5}$ concentrations in Santiago de Chile, various models have been proposed based on statistical methods or classical neural networks [6,8,26]. Nicolis et al. [6] proposed an improved Bayesian spatio-temporal dynamic model for the prediction of $PM_{2.5}$ in the city of Santiago de Chile by calibrating meteorological variables derived from the Weather Research and Forecasting Model (WRF). From a neural-network perspective, Pérez et al. [26] used an approach based on multiperceptron neural networks, subsequently improving the quality of the model by considering meteorological variables such as wind direction [8]. In southern Chile, Diaz-Robles et al. [27] proposed a pollution prediction model based on a mixture of neural networks and ARIMA models.

Regarding the pollution study in other countries, Fan et al. [38] considered the use of various stacked models of LSTM recurrent networks and multilayer perceptrons for the prediction of pollution in North China, obtaining that the use of LSTM networks allows

better performance. Huang and Kuo [39] proposed a mixed model consisting of a sequence starting with a convolutional network whose output feeds a recurrent short-long memory (LSTM) network for $PM_{2.5}$ concentration prediction in Beijing and Shanghai. It achieved an average of 97.8% correct prediction. Kris et al. [40] used satellite imagery combined with ground-level measurements and deep convolutional neural networks, achieving a value of 0.75 for R^2. Yanlin Qi et al. [41] combined convolutional networks with recurrent networks, thereby achieving a mean square error value of 0.72 for the next 72 h in the Jing-Jin-Ji area, Beijing. Fabiana et al. [42] was able to forecast the concentrations of $PM_{2.5}$ and PM_{10} for the city of Bogotá (Colombia). They performed principal component analysis for the development of forecast models and artificial neural network and clustering methods. These models, although powerful, are highly complex, which requires careful management of the parameters of the component networks.

Recently, there has been an emphasis on spatio-temporal models using deep neural networks. For example, in [43–45], LSTM or attention-based networks were proposed to perform this prediction; however, they use the historical values of the geographical point to be predicted. They also use additional values such as $PM_{2.5}$ measurements given by cheap sensors. A similar work was provided by [46]; they proposed a novel hybrid deep learning model that combines convolutional neural networks (CNN) and long short term memory (LSTM) together to forecast air quality at high resolution.

In general, for the study of pollution in Santiago de Chile, there is not much literature on the use of modern architectures of neural networks. The majority of studies focus on air pollution in other countries, which is often characterized by different spatial and temporal features. For these reasons, we explored the use of LSTM networks for the space-time prediction of $PM_{2.5}$ in Santiago de Chile. We emphasize the data pre-processing because it is required for correct validation of models; the prediction is performed in a spatial point without historical data.

3. Methods and Materials

3.1. Space-Time Prediction

Let us denote by $p(\mathbf{s}_i, t_j)$ the $PM_{2.5}$ pollution at site $\mathbf{s}_i = (x_i, y_i)$, with $i = 1, \ldots, S$, where (x_i, y_i) represents the vector of spatial coordinates, and time t_j, with $j = 1, \ldots, T$. While in the temporal prediction the main goal is to forecast the level of pollution in a given station $\mathbf{s}_1$ for $n-$ahead time steps ($n \geq 1$), $\{p(\mathbf{s}_1, t_{T+1}), \ldots, p(\mathbf{s}_1, t_{T+n})\}$, considering the historical information of pollution in the same site $\mathbf{s}_1$, $\{p(\mathbf{s}_1, t_1), \ldots, p(\mathbf{s}_1, t_T)\}$, the spatio-temporal prediction consists of predicting the pollutant concentration at new site s_0 where there are not monitoring stations—that is, the past information is not available, but the past information of the nearest stations is used. Such prediction can be extended to every point on a regular grid for producing a prediction map. Figure 1 shows an example of the difference between temporal and spatio-temporal predictions. Mathematically, the space-time prediction can be expressed as a nonlinear transformation $f(\cdot)$ of the past pollution observations at different sites using no linear functions. Hence, the proposed model can be written as

$$\hat{p}(\mathbf{s}_0, t_{T+n}) = f(p(\mathbf{s}_k, t_{T-m}), \mathbf{X}(\mathbf{s}_k, t_{T-m})),$$

where $\hat{p}(\mathbf{s}_0, t_{T+n})$ is the prediction of the $PM_{2.5}$ concentration at the station $\mathbf{s}_0$ and time t_{T+n}; $p(\mathbf{s}_k, t_{T-m})$ represents the sequence of past observations of $PM_{2.5}$ at the sites $\{\mathbf{s}_k\}$, with $k = 1, \ldots, K$, and at the past times $\{t_{T-m}\}$, with $m = 0, 1, \ldots, M$—that is, $p(\mathbf{s}_k, t_{T-m}) = \{p(\mathbf{s}_1, t_{T-1}), \ldots, p(\mathbf{s}_n, t_{T-m})\}$; finally, $\mathbf{X}(\mathbf{s}_k, t_{T-m})$ is a sequence of covariates given by past meteorological and pollutant variables—that is, $\mathbf{X}(\mathbf{s}_k, t_{T-m}) = \{X(\mathbf{s}_1, t_{T-1}), \ldots, X(\mathbf{s}_n, t_{T-m})\}$.

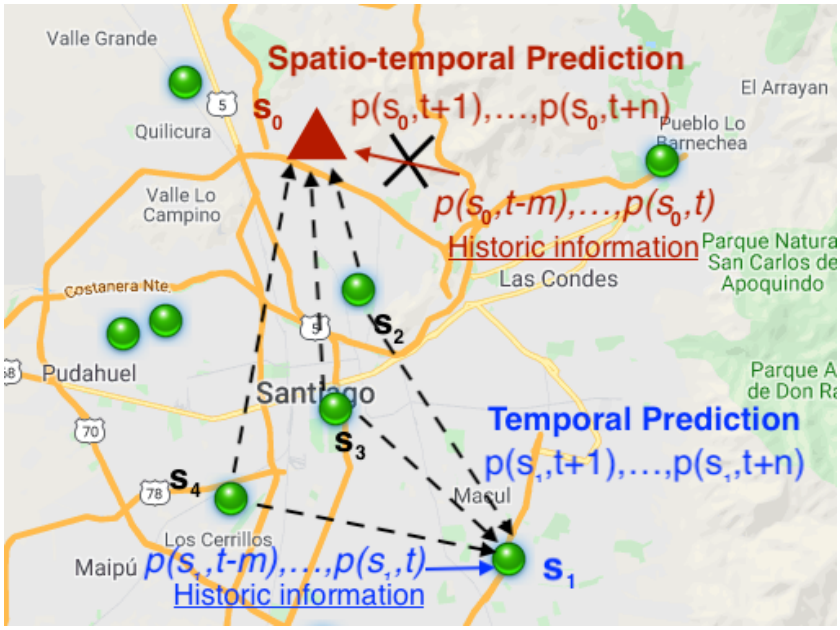

Figure 1. Temporal prediction vs. spatio-temporal prediction. Green points indicate the sites of monitoring stations where historical information (pollution and meteorological variables) is available, $p(\mathbf{s}_1, t-m), \ldots, p(\mathbf{s}_1, t)$ for a target point $\mathbf{s}_1$, and temporal predictions at time $t+1, \ldots, t+n$ can be made at the same points. The red triangle indicates a point where the is not a monitoring station. In this case, the proposed model uses the past (temporal and spatial) information of the other stations for providing the spatio-temporal prediction $p(\mathbf{s}_0, t+1), \ldots, p(\mathbf{s}_0, t+n)$.

3.2. Base Neural Network

The use of recurrent neural networks is proposed to obtain temporal and spatio-temporal predictions, that is, to determine the possible values that a variable can obtain at a certain number of hours ahead, or/and at a certain spatial point. To achieve these predictions, the recurrent neural networks learn through historical information, generating relationships between the delivered variables. For this particular case, it is very important to study how the prediction of meteorological and pollutant variables are directly related to the values of these variables in a time $t-1$. The pre-processing of data will allow one to choose which variables will be used in the models.

We propose the use of long short term memory (LSTM) [29], which has been established as an efficient and scalable model for various types of problems. For this investigation, the LSTM network had additional time distributed in which it not only related the values recorded in a single time $t-1$ but also generated a relationship between all the inputs $t-1$. For example, to obtain the value at time t, it depends on the observations recorded at time $t-1$, $t-1'$, $t-1''$, where each registered value corresponds to a different spatial site but which also influences the prediction of the time variable t.

In Figure 2, we can see an operating scheme of the recurrent neural network LSTM. In a simplified way, an LSTM neural network assumes a multidimensional input x_t that interacts with the previous network state r_{t-1}, where it first passes through the input gate having i_t as output, which enters the memory cell that is affected by the forget gate output given by f_t in order to generate the memory cell output c_t. This, in turn, is fed to the hidden state variable h, which passes by the output gate originating the output variable o_t. Finally, the output o_t in conjunction with the hidden state of the network h determines the current state of the network r_t, which eventually can feed back the previous gates a predefined number of times. This recurrent network is based on sequential processing of information, where particularly long-term information can be controlled through the forgetting gate.

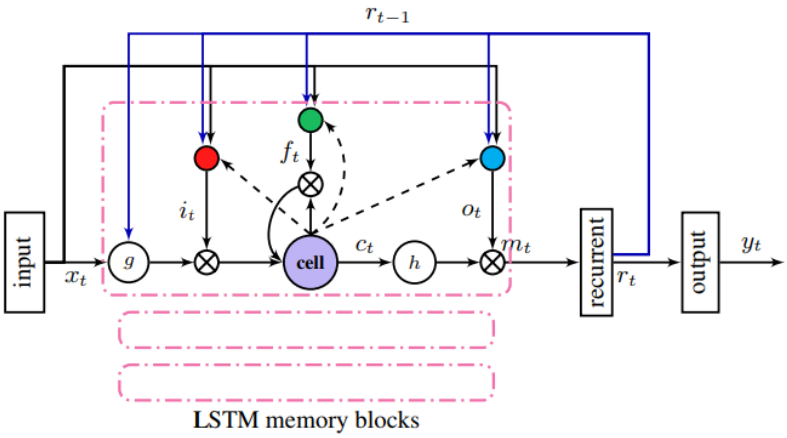

Figure 2. LSTM recurrent network scheme: the current state of the network depends on the previous state and the input, which are weighted by gate functions.

4. Pre-Processing of Data and Selection of Neural Model

4.1. The Dataset

The database used in this work was obtained from monitoring sites of the Ministry of the Environment of Chile (National Air Quality Information System (SINCA), http://sinca.mma.gob.cl, accessed on 1 January 2020).

For this study, the following observations were selected: temperature (temp), relative humidity (hrel), wind direction (dirv), wind speed (velv), particulate matter with diameter less than 10 microns (PM_{10}), particulate matter with diameter less than 2.5 microns ($PM_{2.5}$) and nitrogen dioxide (NO_2).

As the objective of our study focuses on the Metropolitan region, the information recorded from 2010 to 2018 by the meteorological stations located in the capital was used, which are: Cerrillos, Cerro Navia, El Bosque, Independencia, La Florida, Las Condes, Pudahuel, Puente Alto, Quilicura, Parque O'Higgins and Talagante. Figure 3 shows their geographical locations.

Figure 3. Google map of the the Metropolitan region (Chile) with the locations of the 11 meteorological stations.

Table 1 shows the base structure of the data obtained by SINCA from the 11 meteorological stations, in which the name of the file indicates the polluting or meteorological variable that is recorded (a). Within these files, 3 columns are stored, which identify the date (b), the time (c) and the observed value (d).

Table 1. Basic structure of meteorological and pollutant data.

$PM_{2.5}$ (a)			...	Wind Speed (a)		
Date (D-M-Y) (b)	**Hour (H:M) (c)**	**(d)**	**...**	**Date (D-M-Y) (b)**	**Hour (H:M) (c)**	**(d)**
1 January 2010	00:00	65	...	1 January 2010	00:00	2.20
1 January 2010	01:00	70	...	1 January 2010	01:00	1.93
1 January 2010	02:00	41	...	1 January 2010	02:00	1.27
...	...	...	...	...	...	...
31 December 2018	23:00	75	...	31 December 2018	23:00	2.42

4.2. Analysis and Pre-Processing

Once the information of each station was stored, the first data pre-processing step was carried out, where the normalization of the meteorological and pollutant variables corresponding to the date and time of capture was implemented. Table 2 describes the meaning of each of the variables. It is relevant to indicate that the time series that make up the database consider a time step of one hour, which was used in all the experiments carried out.

Table 2. Description of the columns once the first pre-processing step was carried out.

Variable	Description
Station	Register the name of the station
E	East coordinate (UTM)
N	North coordinate (UTM)
Year	Year of registration of the observation
Month	Observation registration month
Day	Observation record day
Hour	Observation record time
N° of week	Records the week number(1–52)
N° of day	Register day of the week (1–7) from Monday
type of day	Register if weekend (0–1)
Season year	Record the season of the year (1–4) from Summer
hrel, NO_2, temp PM_{10}, dirv, velv and $PM_{2.5}$	Record the value of the observation

Once the variables were selected, the appropriate consecutive years were considered as training and testing sets. Using the data provided and considering the percentage of missing values per year (less than 5%), the years 2012 and 2013 were considered in our experiments. Unfortunately, in the 2010–2011 period, not all of the 11 stations indicated yet existed. On the other hand, in 2014 and 2015, some stations were not operational, or measurements of some pollution variables such as NO_2 were not recorded. This generated discontinuity for using later years. Given that the time series require that the study period be contiguous and that we propose using the maximum number of operating stations, 11, the two indicated years were chosen. We hope in a later study to be able to incorporate a greater number of years considering at least approximate data from the stations with missing data using more complex data imputation models.

Given that we only have two years available according to the stated requirements, in our experiments we considered the out-of-sample approach, where an initial dataset formed the training set and the final data corresponded to the test set [47]. As previously indicated, in future works we plan to use more years with which even more exhaustive evaluation schemes could be considered, such as blocked cross-validation [47].

In relation to the experiments and following the out-of-sample approach, the year 2012 was selected as the training set, covering a total of 96,624 observations, and the year 2013 was used for the test set, covering a total of 96,360 observations. The difference in observations between these two datasets is due to the fact that 2012 is considered a leap year, giving an additional 24 h of observations for each of the stations. Note that the complete year is considered in each set because the information of the past months or seasons (winter or summer) can be relevant in the prediction.

The missing values of the selected sets were imputed by using the algorithm MICE [48], considering the same variables at different sites. The time series were generated according to the number of hours necessary for the prediction of future hours. In this particular case, and after some preliminary analysis, we decided to take the observations of the last 24 h in order to predict the values of the following hours.

The time series used for the input and output of the neural network model were generated as follows: first, the entry and output periods were selected, corresponding to the first 24 h available and 25th hour, respectively. The variables to be considered for each hour were extracted and concatenated in an array. Then, the generation process started from the following hour, taking the interval 2–25 for the input period and the 26th hour for the output period. Again, the variables were extracted and a list was generated. This process was repeated, concatenating the arrays in a data matrix, which was used for both training and test data. This procedure is called sliding window [49] and can be seen in Figure 4. It should be noted that the data obtained in each hour correspond to those given by each station. To denote each station, its coordinates (latitude and longitude) were added.

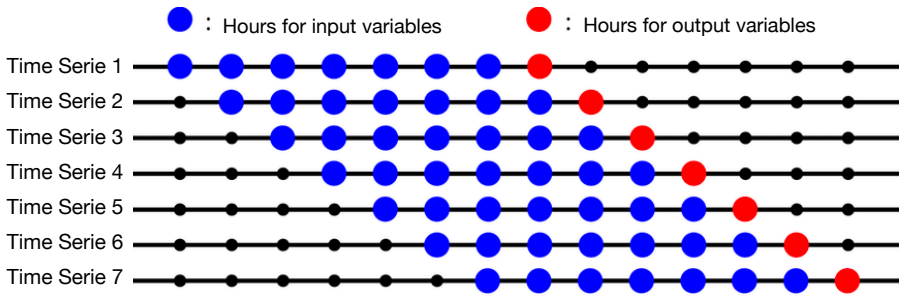

Figure 4. Visualization of sliding window procedure for the generation of time series considering 7 h of input and 1 h of output.

Given that we have 24 h, 11 stations and 7 selected variables (defined in Section 4.3.1), in addition to the two location variables of each station, we have that the number of input variables in each time series corresponds to $11 \times (2 + 7 \times 24) = 1870$. In the output only one variable is required, $PM_{2.5}$. In the experiments where the number of stations varies, the number of data inputs will also be changed. The generated layout can be viewed in the header of the Table 3.

Table 3. Data structure expressed in time series.

E (Es-01)	N (Es-01)	$PM_{2.5}$ (Es-01) (t-23)	...	$PM_{2.5}$ (Es-01) (t-0)	E (Es-02)	...	$PM_{2.5}$ (Es-01) (t + 24)	...	$PM_{2.5}$ (Es-N) (t + 24)
A	B	38.0	...	35.0	C	...	36.0	...	36.5
A	B	34.0	...	33.0	C	...	30.0	...	33.4
A	B	27.0	...	28.6	C	...	30.0	...	25.0

The variables presented in Table 3 refer to those mentioned in the Table 2 during a certain period of time; this means that, for example, columns E(Es-01) and E(Es-02) refer to the easting (longitude) coordinates for stations 01 and 02, respectively, and N(Es-01) is the northing (latitude) coordinate for station 01. The letters A, B and C indicate the UTM coordinates. In general, we differentiate the coordinates by unique letters of the alphabet. Additionally, column $PM_{2.5}$ (Es-01)(t − 23) refers to the $PM_{2.5}$ concentration 23 h before the reference time t for station 01, and column $PM_{2.5}$ (Es-02)(t + 24) refers to the value of $PM_{2.5}$ at 24 h after the reference time *t* for station 02. The pre-processed dataset with the data from the 11 stations is available to the community (https://1drv.ms/u/s!AnkU8l4kGM9wlhWy22fexCGzHByp?e=ZQ9Hqd, accessed on 1 January 2020).

4.3. Training Net Selection

For the selection of the architecture of the network, first we considered models for temporal prediction of $PM_{2.5}$, which is an easier task. The resulting model was used as a basis for the final space-time neural model. For the training network selection, a new dataset was generated based on Table 3.

In this task, we considered 24 h for input and 1 h for output. Input variables are indicated by columns with expression $(t - n)$, where n takes values from 0 to 23, and output variables are indicated by expression $(t + n)$ in an analogous way. Note that each input series was concatenated with the 11 stations which are identified by the coordinates E (longitude) and N (latitude). To achieve a temporal forecast, we first (i) selected a station to forecast. In this case we eliminated all the columns of the output variables that do not correspond to the station we sought to predict. Then (ii) we modified the input variables of the station to be predicted by replacing them with -1, with the exception of the coordinate of the station itself. Finally, (iii) the two previous steps (i and ii) were repeated for all the stations in all the data of Table 3. In this way it was expected that the neural network learned that the values to be predicted correspond to the locations that contain values of -1. Relevantly, it should be noted that this step only considered the original training data, that is, from the year 2012, and that it was aimed at finding an initial network structure for our models; that is, the weights learned in the experiment were not used anymore in space-time. The reason for this step was the difficulties in finding a good initial structure in the space-time problem, which is why it was decided to first obtain a solution in a more relaxed problem, in this case only temporarily. An alternative is to eliminate a station in each test, which required a high computational cost, since the experiment had to be repeated for each station. However, we expected that the bias effect would be reduced, since it was only considered in the selection of the network structure. Subsequently, a new pre-processing stop will be carried out to adapt it to a spatio-temporal modeling. Table 4 presents a summary of the results of this procedure.

Table 4. Structure of the data expressed in time series. The columns to predict are indicated using padding with value −1.

E (Es-01)	N (Es-01)	$PM_{2.5}$ (Es-01) (t-23)	...	$PM_{2.5}$ (Es-01) (t-0)	E (Es-02)	N (Es-02)	$PM_{2.5}$ (Es-02) (t-23)	...	$PM_{2.5}$ (Es-02) (t-0)	...	$PM_{2.5}$ (t+1)
A	B	−1	...	−1	C	D	36.5	...	40.2	...	36.0
A	B	34.0	...	33.0	C	D	**−1**	...	**−1**	...	22.0
A	B	27.0	...	28.6	C	D	30.2	...	36.0	...	25.0

In Table 4 the new dataset is presented. The first row has the predicted values of $PM_{2.5}$ of station 01; therefore, its columns (Es-01) have values of −1. For the second row, the values of predicted $PM_{2.5}$ for station 02, and again, its corresponding columns have values −1.

4.3.1. Network Selection

For the selection of the architecture of the LSTM network, a group of networks with different architectures were generated, where the dropout values varied between 0 and 0.3 with steps of 0.05 and the number of neurons per layer was between 100 and 250 (both included) in multiples of 50, except for the last LSTM layer, for which there were 25, 50, 75, 100, 150, 200 or 250 neurons. The activation functions of the intermediate-layer neurons correspond to hyperbolic tangent functions. Finally, a final time distribution layer was added using a linear activation function. This layer is very important because the input variables are not only directly related to their previous values, but they are also related to the previous observations of the other meteorological stations. The mean square error was used as the objective function, optimizing the cost function with the adaptive moment estimation (Adam) algorithm [50].

We used a greedy strategy for the architecture selection of the neural network to reduce the search space of the structure due to the exponential number of possible configurations. We started by optimizing the first layer and after finding the best configuration according to the 7 neuron values and 6 dropout values. We optimized the network by adding a second layer and fixing the configuration of the first, and we continued until a third layer. In this way, we tested a total of 39 neural network configurations according to the number of neurons and the dropout operator. In this case, we used as the network input data configuration the schema given in Table 4. Eleven meteorological stations were used. The variables temp and $PM_{2.5}$ of the last 24 h were used as input, and $PM_{2.5}$ forecast 1 h ahead was output.

Next, Table 5 provides a summary of the 6 networks with the highest values of R^2 of the 39 tested in total. The networks that stand out in terms of R^2 in their training set are identified. Finally, although the values of R^2 do not suffer variations of more than 2 points, we selected R36 (Table 5) as our candidate because, in addition to having the highest values of R^2 in training and validation, we grant the possibility that being a deep network, it can learn a greater number of relationships that exist between the data.

Table 5. Selection of six neural networks that present the highest values of R^2 in the dropout selection training group, number of neurons and LSTM layers.

Code LSTM Network	Neurons for Layer	Dropout	R^2 in Train	R^2 in Validation
R02	200		0.7661	0.6681
R07	200	0.1	0.7611	0.6544
R15	200	0.15	0.7599	0.666
R23	200–100	0.1–0.15	0.7786	0.6554
R30	200–100	0.1–0.3	0.7725	0.6567
R36	200–100–50	0.1–0.15–0.3	0.7794	0.6682

The selection of input variables was based on the R36 network considering the value of R^2 obtained in the validation set for the prediction of $PM_{2.5}$. As a result of this experiment, the input variables were $PM_{2.5}$, PM_{10}, NO_2, temp, velv, hrel and hour, which are detailed in Table 2. To optimize results, only $PM_{2.5}$ was left as the output variable.

4.4. Spatio-Temporal Prediction

4.4.1. Test and Training Sets

In order to carry out the space-time prediction of $PM_{2.5}$, the following steps were considered: (i) First, the training and testing set were modified (see, Table 4) by inserting 2 new columns, (E′ and N′), representing the spatial coordinates of the location where we need to predict the $PM_{2.5}$. The inclusion of these two new input variables is necessary for associating the geographic coordinate to the station to be predicted, which we call the target station. Then, (ii) the variables of the target station were filled with -1. The previous step (ii) was replicated for all stations (iii) which generated the database that was the input of the neural network. For example, if we have 11 stations, during the training there will be 10 stations that generate the input variables with values different from -1, and the target station to predict, that is, the 11th, will be the only station that will provide the output. In addition, in the input variables, the coordinates of the target station will appear at the beginning, and in contrast, the historical variables of contamination or others of this station will not appear. By not including the historical information of the target station to be predicted as input variables, we prevented this prediction using the information which corresponds to a typical temporal model. An example of the resulting input can be seen in Table 6. Since the coordinates vary in the input variables, and this can change arbitrarily even when the other input variables remain the same, this process allows generating a space-time prediction. Note that this process is repeated for both the training and testing processes; the only difference is that in testing the station to consider in the output variables corresponds to the target station. To strengthen the prediction, so that most of the stations were considered, high dropout values (greater than 0.4) were used in the neural networks tested. At the same time, this complete procedure was repeated for all the stations in order to evaluate the model in different target stations. Finally, it should be noted that in this process the weights of the neural networks were initialized randomly—that is, only the structure of the previous step given in Section 4.3 was considered.

Table 6. Structure of dataset used in spatio-temporal experiments.

E′	N′	E (Es-02)	N (Es-02)	$PM_{2.5}$ (Es-02) (t-23)	…	$PM_{2.5}$ (Es-02) (t-0)	E (Es-03)	N (Es-03)	$PM_{2.5}$ (Es-03) (t-23)	…	$PM_{2.5}$ (Es-03) (t-0)	…	$PM_{2.5}$ (t+24)
C	D	C	D	−1	…	−1	E	F	25.1	…	18.2	…	32.1
E	F	C	D	35.1	…	32.8	E	F	−1	…	−1	…	35
G	H	C	D	25.3	…	30.4	E	F	17.4	…	22.9	…	24.1

Initially, we eliminated the input data of the station to be predicted during the training to prevent the network from looking only at the historical data of the station; however, the results improved when considering it. We believe this happened because valuable information was not lost during training, and on the other hand, the dropout operator allowed the network to consider all available stations.

4.4.2. LSTM Neural Network Tuning

Finally, we fit the multilayer LSTM neural network according to the two experiments specified in Section 5 which consider 7 and 11 input stations with 1 and 24 h as prediction outputs, respectively. The network needs to be specified because the numbers of input and output variables varies. In the case of the experiments with seven stations and 1 h as output, the network R36 was taken as the base considering $\{200, 100, 50\}$ neurons and dropout rates

of $\{0.1, 0.3, 0.5\}$ in a greedy way, to avoid trying an exponential number of combinations. In addition, the entry of a fourth layer was allowed. The selection was based on the value of R^2 on a validation set given by the 30% of the original test set (details in Section 5). The experiments indicated that the network with the best results turned out to have four LSTM layers with $\{200, 50, 100, 30\}$ neurons with dropout rates of $\{0, 0.5, 0.5, 0.5\}$, respectively. We call this network LSTM-4.The training of the proposed model typically manages to control overfitting by considering the indicated dropout rates. As an example, Figure 5 shows the evolution of the mean squared error (MSE) loss function using the LSTM-4 network for the 1 h forecast at Cerro Navia station using six closest stations as input. In the graph it can be seen that the training and validation loss curves tend to converge.

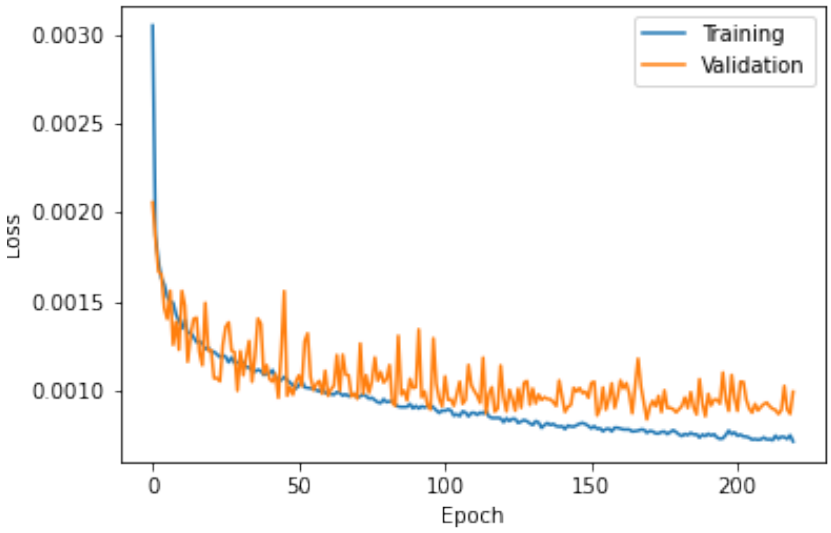

Figure 5. Evolution of loss function during training of LSTM-4 neural network model for one-hour-ahead $PM_{2.5}$ prediction of station Cerro Navia using the 6 closest stations as input.

5. Experiments

In this section, we present two experiments used to validate the proposed models and analyze the results of the space-time predictions for Santiago. First, a short-term prediction is proposed, one hour in the future (Section 5.1). This was expected to be the most reliable prediction due to the short time in the future to be predicted. In addition, a medium-term prediction was performed, 24 h into the future (Section 5.2). In this case, the predictions from (t+1) to (t+24) were done simultaneously considering the data recorded in the previous 24 h (from (t-23) to (t)). The reason for this experiment was to test how the network behaves over a longer horizon, in addition to analyzing its behavior when it has multiple outputs. Indeed, the prediction of a 24 h period did not produce the best prediction for any particular hour. On the other hand, we also considered different areas within Santiago as input for the neural network. We first studied the set of seven nearby stations: Cerrillos, Cerro Navia, El Bosque, Independencia, La Florida, Parque O'Higgins and Pudahuel. We then studied all eleven stations available in the database. The reason for this differentiation is that we observed that there are stations very far from others, such as Talagante (see Figure 3). Normally, very distant stations contain contamination values that are poorly correlated with the others due to their particular conditions (sources of emissions or meteorological factors). Since the predictions in a given location are based on input variables which are collected in different spatial points, the space-time prediction quality could be affected by farthest stations. Hence, we removed the four furthest stations: Las Condes, Quilicura, Puente Alto and Talagante.

Regarding the design of experiments, in addition to the indicated training and testing sets, we proceeded to divide the testing set: 30% for validation and the rest for testing. In relation to the compared models, we tested five models: a multivariable linear regression

(LR), a feed forward neural network (FFN), an LSTM neural network (LSTM), a multilayer gated recurrent units (GRU) neural network and a multilayer LSTM neural network. This last network is identified by LSTM-X, where X is the number of stacked LSTM layers. The validation of the results was carried out considering the following standard metrics: the coefficient of determination (R^2), the square root of the squared error (RMSE) and the median absolute error (MAE). A demo source code with the tested models is made available to the community (https://github.com/sagagk/ExperimentosPM25, accessed on 1 January 2020).

Finally, regarding the configurations of the stacked LSTM neural network, in the prediction experiments at 1 and 24 h, the LSTM-4, detailed in Section 4.4.2, was used. Regarding the GRU model, the same number of layers and dropout operators given in the stacked LSTM configurations where the LSTM layer was replaced by a GRU were used. The LSTM model had the configuration of the first layer of the LST3 and LSTM-4 model, having 200 neurons without dropout. The FFN model was multilayered and consisted of 512 and 64 neurons. This setup was based on a greedy search using a validation set with $\{64, 128, 256, 512\}$ neurons in each layer. On the other hand, the LR model did not require configuration, since it corresponds to a linear combination of the input variables. This section is divided into two subsections which detail the short-term forecast (1 h in the future) and medium-term forecast (24 h ahead), each considering both seven and eleven stations.

5.1. One-Hour-Ahead Forecasting

We applied the different models to forecast the $PM_{2.5}$ one hour ahead. This task is the prediction which had the highest expected certainty because there was less discontinuity with the historical data. We used two experiments: (i) first we considered the seven nearby stations, and then (ii) the eleven available stations. We complement these experiments by visually displaying the prediction results at a particular station. In this way we hope to see the effect of the more distant stations on the quality of the predictions.

(i) Prediction considering the seven closest stations: In Table 7, each of the seven trained networks is displayed, where the column Station refers to the station to be predicted. In relation to metric R^2, the results show that LSTM-4 obtained the best results in comparison to other alternative methods. In particular, the station Cerro Navia delivered the highest prediction quality of the $PM_{2.5}$ pollutant with R^2 of 0.844, followed by Pudahuel, Cerrillos and El Bosque. On the other hand, the lowest quality of prediction was obtained on La Florida with an R^2 of 0.679. On average, the prediction quality remained relatively high with a R^2 of 0.741. We think that these good results are explained by these stations being close. Regarding RMSE and MAE, the results are similar to the R^2 results: LSTM-4 was the best method again. LSTM obtained the best results for MAE. We also tested a simple baseline prediction based on the mean of the last value of $PM_{2.5}$; however, the results are poor: the average R^2, RMSE and MAE were 0.002, 25.7 and 20.4. In general, we found that the methods based on recurrent neural networks perform better than simpler methods such as FFN or LR.

Table 7. Metrics of space-time predictions considering 7 stations, of 1 h in the future. Bold font indicates best results.

Station	Metric	LR	FFN	LSTM	GRU	LSTM-4
Cerrillos	R2	0.533	0.699	0.725	0.732	**0.733**
	RMSE	13.35	10.73	10.25	10.12	**10.10**
	MAE	7.44	4.83	4.60	4.43	**4.38**
Cerro Navia	R2	0.678	0.707	0.828	0.836	**0.844**
	RMSE	14.27	13.62	10.43	10.19	**9.93**
	MAE	6.80	4.96	3.82	**3.82**	3.83
El Bosque	R2	0.523	0.646	0.684	0.706	**0.707**
	RMSE	15.52	13.37	12.63	12.18	**12.17**
	MAE	8.57	5.96	5.49	**5.43**	5.58
Independen.	R2	0.041	0.529	0.688	0.698	**0.699**
	RMSE	14.27	10.00	8.13	8.01	**8.00**
	MAE	7.27	4.66	**4.12**	4.15	4.20
La Florida	R2	0.169	0.464	0.667	**0.679**	0.676
	RMSE	15.41	12.38	9.75	**9.57**	9.63
	MAE	8.57	5.96	5.49	**5.43**	5.58
P.O'Higgins	R2	0.568	0.696	0.717	0.691	**0.729**
	RMSE	11.69	9.81	9.46	9.88	**9.26**
	MAE	6.56	4.53	4.33	4.76	**4.24**
Pudahuel	R2	0.511	0.691	0.785	0.800	**0.802**
	RMSE	15.49	12.33	10.29	9.92	**9.87**
	MAE	9.11	4.78	**3.75**	4.09	3.79
Average	R2	0.432	0.633	0.728	0.735	**0.741**
	RMSE	14.29	11.75	10.13	9.98	**9.85**
	MAE	7.74	5.08	4.41	4.48	**4.40**

(ii) Prediction considering all 11 stations: Table 8 displays the results for each of the 11 trained networks, where the Station column indicates the station to predict. When considering the metric R^2, it appears again that the LSTM-4 algorithm obtained the highest value, it being very close to the one given by GRU. As in the previous experiment, the Cerro Navia station delivered the highest $PM_{2.5}$ forecast quality with a maximum R^2 of 0.816 (LSTM-4), followed by Pudahuel (LSTM), Quilicura (LSTM-4) and Cerrillos (GRU). On the other hand, the lowest prediction quality was obtained in the stations of Puente Alto and Las Condes with R^2 of 0.461 and 0.489, when considering the LSTM-4 method. On average, the quality of the predictions was still relatively high, given an R^2 of 0.657, although clearly lower than the result obtained using the seven nearby stations. We believe that this decrease in accuracy can be explained by the incorporation of very distant stations, which could have confused the neural network. Regarding RMSE, the results are similar to those of R^2: the LSTM-4 algorithm was again the best method, closely followed by GRU. In MAE, the LSTM network obtained the best average result. Additionally, we also tested a simple benchmark prediction based on the mean of the last values of $PM_{2.5}$. In this case, the results were not good, given average R^2, RMSE and MAE of -0.501, 22.3 and 13.77, respectively. We found that methods based on recurrent neural networks with multiple layers obtained the best performance on average.

Table 8. Metrics of space-time predictions considering 11 stations, of 1 h in the future. Bold font indicates best results.

Station	Metric	LR	FFN	LSTM	GRU	LSTM-4
Cerrillos	R2	0.566	0.656	0.708	**0.712**	0.710
	RMSE	12.94	11.51	10.62	**10.53**	10.57
	MAE	7.08	5.20	**4.42**	4.63	4.67
Cerro Navia	R2	0.619	0.662	0.811	0.802	**0.816**
	RMSE	15.59	14.67	10.96	11.23	**10.82**
	MAE	7.36	5.36	**3.96**	3.82	4.07
ElBosque	R2	0.522	0.641	0.668	**0.688**	0.675
	RMSE	15.58	13.50	12.98	**12.59**	12.85
	MAE	8.33	6.07	**5.57**	5.89	5.69
Independencia	R2	0.329	0.624	0.619	0.656	**0.667**
	RMSE	12.06	9.03	9.09	8.64	**8.50**
	MAE	7.22	4.63	4.48	4.33	**4.31**
LaFlorida	R2	0.407	0.574	0.663	0.671	**0.673**
	RMSE	13.12	11.12	9.89	9.77	**9.75**
	MAE	7.80	5.37	**4.61**	5.31	5.18
Las Condes	R2	−0.595	−0.404	0.451	0.485	**0.489**
	RMSE	15.91	14.93	9.33	9.04	9.00
	MAE	8.92	6.70	4.82	4.73	4.79
P.O'Higgins	R2	0.498	**0.678**	0.677	0.665	0.674
	RMSE	12.69	**10.16**	10.17	10.37	10.22
	MAE	6.97	**4.60**	4.71	**4.60**	4.91
Pudahuel	R2	0.538	0.658	**0.779**	0.777	0.759
	RMSE	15.13	13.03	**10.46**	10.50	10.93
	MAE	8.47	4.50	**3.98**	4.51	4.42
Puente Alto	R2	0.277	0.357	0.398	**0.461**	**0.461**
	RMSE	19.73	18.60	18.01	**17.04**	**17.04**
	MAE	10.64	10.55	8.79	8.70	**8.69**
Quilicura	R2	0.590	0.689	0.710	0.718	**0.720**
	RMSE	11.01	9.59	9.26	9.14	**9.11**
	MAE	6.10	4.44	**4.07**	4.08	4.12
Talagante	R2	0.370	0.532	**0.587**	0.573	0.581
	RMSE	14.06	12.11	**11.38**	11.58	11.47
	MAE	7.21	4.97	4.68	**4.45**	4.66
Average	R2	0.375	0.515	0.643	0.655	**0.657**
	RMSE	14.35	12.57	11.10	10.95	**10.93**
	MAE	7.83	5.67	**4.92**	5.00	5.05

Next, we visually analyze the results of the 1 h spatio-temporal prediction of the LSTM-4 neural network considering the Cerro Navia station. Figure 6 shows the forecast for 24 h ahead—9 August 2013. We chose this day because it corresponds to a winter one, which is a season where greater pollution changes occur. In particular, we note that the

network was able to predict the peak of $PM_{2.5}$ that usually occurs around 9 am ([8]) due to the high traffic. In general, we note that the network managed to reasonably predict pollution, especially during the hours when the people are most exposed (i.e., from 7 am to 9 pm).

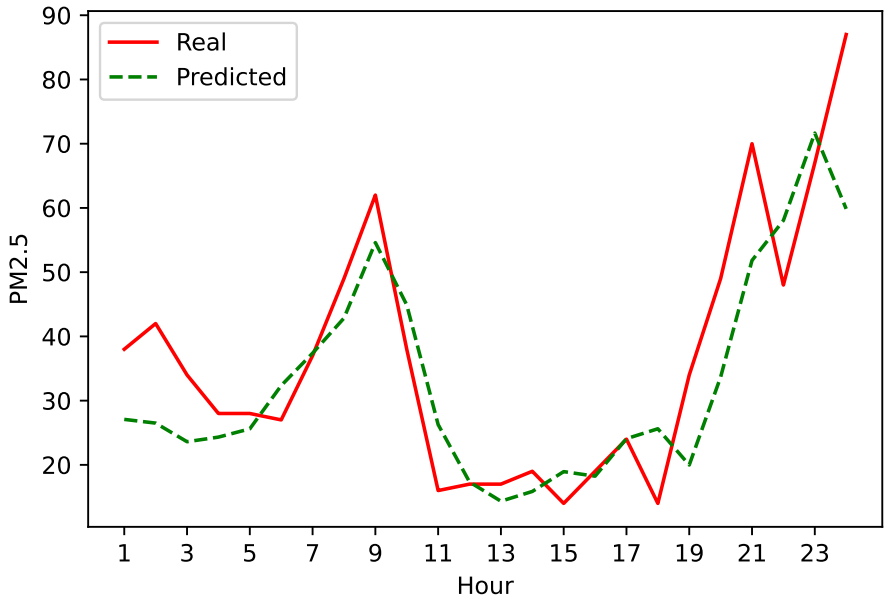

Figure 6. Comparison of real and one-hour-ahead values of $PM_{2.5}$ at the Cerro Navia station during 9 August 2013.

5.2. 24-h-Ahead Forecasting

In this experiment we performed space-time prediction of 24 h in the future. This problem is more difficult than the 1 h prediction because there were more variables to adjust simultaneously, which was verified in the results. In this case, we again performed two experiments, as in the previous Section: (i) we first evaluated the seven nearby stations, and then (ii) the eleven available stations. We also complement these experiments by visually displaying the results of the evolution of R^2 and the 24 h predictions at a particular station.

(i) Prediction considering the seven closest stations: Table 9 presents the results considering all the techniques proposed for the 24 h prediction. The results of the R^2 metric show that the LSTM-4 neural network obtained the best results, closely followed by the GRU network. As when predicting 1 h ahead, the Cerro Navia station provided the highest $PM_{2.5}$ prediction quality, with an R^2 of 0.411 using the LSTM method, followed by El Bosque (LSTM) and Cerrillos (LSTM-4), which provided values higher than 0.4. On the other hand, the lowest prediction quality was obtained in Florida, with an R^2 of 0.330. On average, the quality of the prediction was relatively low, with a R^2 of 0.38. Note that this R^2 corresponds to the mean of 24 values of R^2, one for each hour. The values for the first hour were much better than for hours 12 or 24. For example, in Cerro Navia, the R^2 was 0.62 for hour 1 and 0.37 for hour 24. Regarding RMSE and MAE, the results are similar for R^2: the LSTM-4 algorithm appears to been the method with the best results, being closely followed by GRU. Again, we tested a simple prediction based on the mean of the last 24 $PM_{2.5}$ values for all training stations. The results were very poor, being on average -0.588, 24.4 and 18.8 for the R^2, RMSE and MAE, respectively. In summary, the methods based on recurrent networks appear to be the ones with the best performances.

Table 9. Metrics of space-time prediction considering 7 stations, of 24 h in the future. Bold font indicates best results.

Station	Metric	LR	FFN	LSTM	GRU	LSTM-4
Cerrillos	R2	−0.153	0.347	0.389	0.392	**0.401**
	RMSE	20.99	15.79	15.28	15.24	**15.12**
	MAE	12.85	7.20	6.57	6.85	**6.39**
Cerro Navia	R2	0.193	0.393	**0.411**	0.383	0.408
	RMSE	22.43	19.46	**19.17**	19.63	19.21
	MAE	11.80	7.40	7.26	8.38	**6.67**
ElBosque	R2	−0.145	0.387	**0.403**	0.399	0.402
	RMSE	24.02	17.58	**17.36**	17.41	17.37
	MAE	14.99	8.25	7.51	7.71	**7.23**
Independencia	R2	−0.767	0.133	0.282	**0.377**	0.352
	RMSE	19.44	13.62	**12.39**	11.54	11.77
	MAE	12.37	7.18	6.16	6.50	**6.10**
LaFlorida	R2	0.085	0.137	0.284	**0.330**	0.327
	RMSE	16.20	15.73	14.32	**13.86**	13.89
	MAE	8.73	8.32	6.78	7.25	**6.72**
P.O'Higgins	R2	−0.122	0.305	0.370	**0.391**	0.377
	RMSE	18.86	14.84	14.13	**13.90**	14.05
	MAE	11.70	7.62	6.65	6.74	**6.37**
Pudahuel	R2	0.112	0.360	0.383	0.373	**0.393**
	RMSE	20.69	17.56	17.24	17.38	17.11
	MAE	11.31	7.15	7.03	7.36	6.40
Average	R2	−0.114	0.295	0.360	0.378	**0.380**
	RMSE	20.38	16.37	15.70	15.57	**15.51**
	MAE	11.96	7.59	6.85	7.26	**6.56**

(ii) Prediction considering 11 stations: Table 10 shows the results all the methods for 24 h predictions. In relation to metric R^2, in this case the neural network GRU obtained the best results. In particular, Cerro Navia delivered the highest prediction quality of the $PM_{2.5}$ pollutant with an R^2 of 0.45, followed by Quilicura, Pudahuel, and El Bosque. On the other hand, the lowest-quality predictions were obtained for Las Condes, with an R^2 of 0.292. On average, the prediction quality remained relatively low with an average R^2 of 0.38. We note that this R^2 corresponds to the average of 24 R^2, one for each hour. The values for the first hour are much better than for the 12th or 24th hours. For example, for Cerro Navia, the R^2 was 0.74 for hour 1 and 0.35 for hour 24. This behavior was repeated for all stations. In subsequent analysis, we will focus on this point for one station. Regarding RMSE and MAE, the results are similar to those of R^2: the GRU algorithm was the best method again. We also tested a simple baseline prediction based on the mean of the last 24 values of $PM_{2.5}$ for all training stations. In this case, the results were very poor; the average R^2, RMSE and MAE were −1.12, 40.3 and 18.2. Similarly to the previous experiment, we found that the methods based on recurrent neural networks preformed better than simpler methods, such as FFN and LR.

Table 10. Metrics of space-time prediction considering 11 stations, of 24 h in the future. Bold font indicates best results.

Station	Metric	LR	FFN	LSTM	GRU	LSTM-4
Cerrillos	R2	0.343	0.375	0.383	0.392	**0.393**
	RMSE	15.89	15.50	15.40	15.29	**15.27**
	MAE	7.66	7.20	6.51	**6.46**	**6.46**
Cerro Navia	R2	0.364	0.384	**0.450**	0.449	0.445
	RMSE	20.04	19.73	**18.64**	18.67	18.73
	MAE	8.77	8.17	6.71	**5.98**	6.13
El Bosque	R2	−0.268	0.394	0.417	**0.422**	0.416
	RMSE	25.36	17.54	17.19	**17.12**	17.22
	MAE	15.45	7.97	7.63	**7.29**	7.47
Independen.	R2	−1.220	0.371	0.386	**0.402**	0.401
	RMSE	21.85	11.62	11.49	**11.34**	**11.34**
	MAE	13.64	6.23	**5.99**	6.16	6.08
La Florida	R2	−0.862	0.352	0.379	**0.425**	0.386
	RMSE	23.20	13.69	13.40	**12.90**	13.33
	MAE	14.93	6.98	**6.52**	6.60	6.53
Las Condes	R2	−2.686	−0.351	−0.137	**0.292**	0.248
	RMSE	24.21	14.66	13.45	**10.62**	10.94
	MAE	15.65	7.81	6.72	**5.87**	5.91
P.O'Higgins	R2	−0.634	0.374	0.387	0.383	**0.392**
	RMSE	22.88	14.15	14.01	14.05	**13.95**
	MAE	14.61	7.19	6.89	**6.57**	**6.57**
Pudahuel	R2	0.348	0.375	0.407	**0.410**	0.404
	RMSE	17.85	17.48	17.01	**16.98**	17.06
	MAE	7.08	6.95	6.67	**6.28**	6.54
Puente Alto	R2	−0.434	0.108	0.112	**0.185**	0.160
	RMSE	27.75	21.89	21.83	**20.92**	21.24
	MAE	16.92	10.74	10.61	10.53	**10.47**
Quilicura	R2	−0.747	0.404	0.424	0.428	**0.439**
	RMSE	22.62	13.21	12.99	12.94	**12.81**
	MAE	14.29	6.60	5.94	**5.93**	6.03
Talagante	R2	−0.672	0.294	0.333	**0.367**	0.339
	RMSE	22.98	14.93	14.51	**14.13**	14.44
	MAE	14.27	7.39	**6.08**	6.11	6.38
Average	R2	−0.588	0.280	0.322	**0.378**	0.366
	RMSE	22.24	15.85	15.45	**15.00**	15.12
	MAE	13.03	7.57	6.93	**6.71**	6.78

We note that the most distant stations, in particular, Las Condes, Puente Alto and Talagante, but not Quilicura, have the poorest metrics. We believe that this was due to a variety of phenomena, such as the great distances between stations, the different emission sources and the altitude of each (see, for example, the works of [6,51]).

In Figure 7 we visually analyze the results of R^2 for each of the 24 h at Cerro Navia station. It can be seen that the recurrent networks have high values of R^2 for the first hour (0.71), and then they steadily decay. Around hour 12, the value of the metric slowly decays. This is explained by the greater uncertainty due to the accumulated time. The same behavior is evident for the simpler methods, such as FFN and LR, those their results are generally much lower in quality compared to those of the recurrent networks. This behavior was repeated for the other stations and in the other metrics; the global metrics must be carefully analyzed.

In Figure 8 we visually analyze the 24 h forecasts for 1 July 2013 at the Cerro Navia station. It can be seen that the LSTM-4 model managed to predict $PM_{2.5}$ values quite well; however, for the afternoon it tended to be imprecise. This behavior was expected due to the greater uncertainty in hours far from those of the training set.

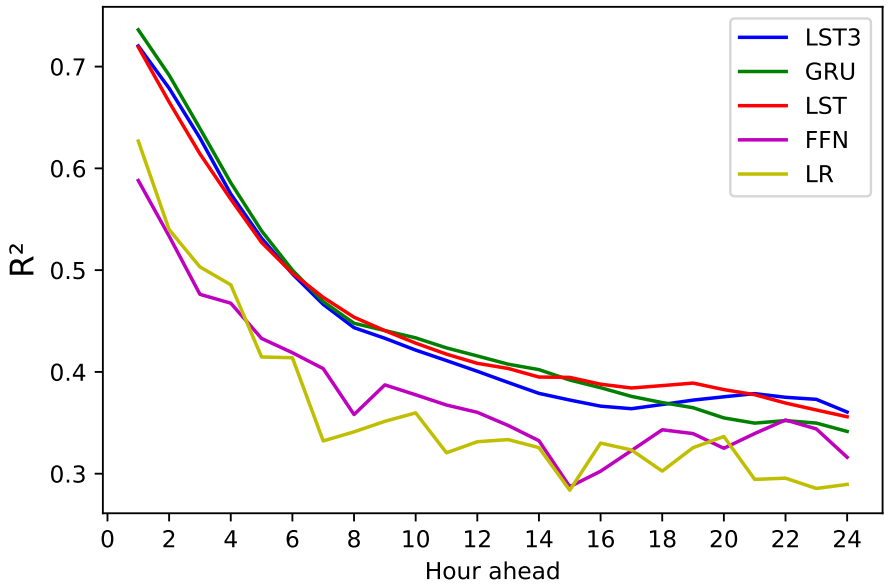

Figure 7. Behavior of R^2 for the $PM_{2.5}$ 24 h forecasts at Cerro Navia station during 2013 for each hour of the day using different methods.

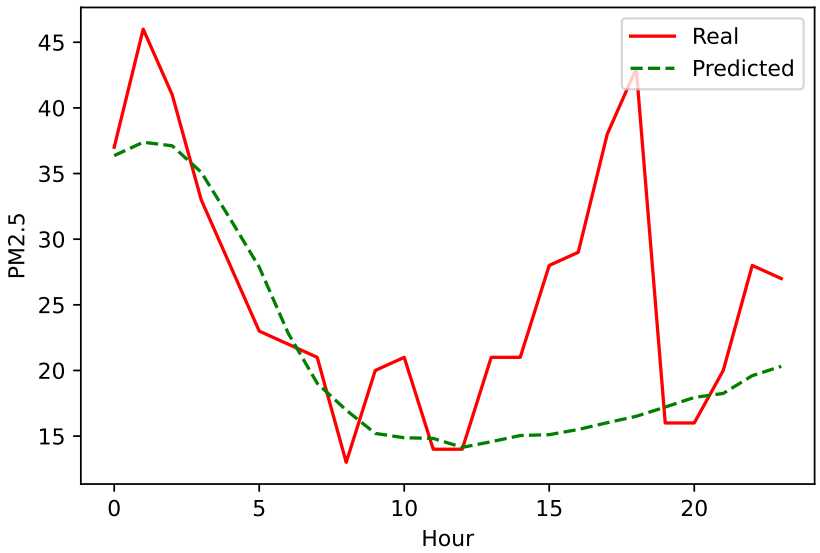

Figure 8. 24 h-ahead forecast for the day 1 July 2013 at Cerro Navia's meteorological station, using the information of the other 10 stations as input data.

Note that the 24 h prediction lowers the 1 h-prediction accuracy, which is because the neural network seeks to predict multiple outputs simultaneously. This problem has the difficulty that the pollution data from the same station were not used to make the prediction, which made it difficult to obtain accurate relationships between the available stations with respect to the location of the target station. For example, there may be local events that can alter measurement levels that are not captured by the model input variables. We believe that the use of additional information such as altitude can improve the prediction.

6. Discussion

When analyzing the results of the experiments considering the prediction for one hour in the future, it can be observed that the use of seven stations allowed us to obtain better predictions than considering eleven stations, which included the four furthest stations: Las Condes, Puente Alto, Quilicura and Talagante. We believe that this was due to the fact that the four outer stations observe values that have lower correlations with those of the seven closest stations. Therefore, this theoretically affects the quality of the short-term forecasts enough to reduce their quality. For example, regarding the R^2 for Cerro Navia and Cerrillos, they decreased from 0.84 and 0.73 to 0.82 and 0.71, respectively, when considering the data from the 11 stations. We think that this may have been due to the presence of factors that could alter pollution, such as some particular sources of emissions (traffic, heating, factories, etc.) or particular meteorological conditions (temperature, wind, humidity, etc.).

Remarkably, we observe that the 24 h forecasts did not differ majorly when considering seven and eleven stations. In particular, we observe that the best average values of R^2 and MAE occurred when considering seven stations, though slightly, but the best RMSE occurred for 11 stations. We can suggest that there is no best configuration in this case, unlike the case of predicting one hour in the future. We think that this result is explained because when considering 24 h, it is possible that the use of more stations can influence the prediction. That is, the patterns of the remote stations may be more correlated when considering a longer horizon than just 1 h in the future. However, in practical terms, although there is no greater difference in performance, naturally a model that considers more stations appears to be more applicable.

When comparing the results of predictions both 1 h ahead and 24 h ahead, it is evident that the results were much better for 1 hour ahead for all tested models. We believe there are two main reasons for this difference in performance. Firstly, when predicting 24 h simultaneously, the neural network generates a weight configuration that tries to respond to all the hours of the output variables simultaneously. In other words, learning seeks to adjust many outputs simultaneously, which may mean that performance must be reduced in one output variable to improve another. The second reason is that the prediction at a longer temporal distance is naturally more uncertain. This can be seen in Figure 7, where as the hour to be predicted moves further away from the input hours, the performance drops noticeably. The R^2 is greater than 0.7 in the first hour, and ends up less than 0.4 at hour 24. A possible solution is to make predictions for every hour in order to maximize the use of the neural network; however, this can be expensive. Another way would be to explore new neural architectures that better handle adaptation to different output variables.

In relation to the pre-processing, the training process considered all the stations, though the information of the target station was canceled while maintaining its coordinates. In this way we hoped that the model would learn to use the other stations to predict a particular station, and thus we avoided typical temporal modeling. The use of all stations can be seen as similar to other $PM_{2.5}$ spatio-temporal prediction work [41], although they did not perform applied information override. An interesting alternative is to remove a training station while it is being used for testing. The problem that appears is that when training, it is necessary to eliminate the station associated with the output variable (in order to imitate the procedure in the testing process), for which the number of stations is reduced by one. In further experiments, we found that this reduction causes the performance of the tested network to drop. We believe that reducing the already reduced number of stations

in the city under study will affect the quality of the predictions; however, we hope to carry out a comparative experiment in a later work.

In relation to the models, the results suggest that in the case of prediction 1 h into the future, the multilayer LSTM model is better on average, although the GRU and LSTM models produced similar results. On the other hand, in the case of the 24 h forecast, the results varied, LSTM-4 and GRU being better on average depending on whether 7 or 11 stations were processed, although their performances were very similar. Therefore, we believe that the implementation of these models requires an optimization of the neural network structure for both architectures, LSTM and GRU, according to the available data.

Finally, in relation to the space-time models that are the state of the art, although they produce excellent results, most of them are not comparable, since they have been applied in areas where a large number of monitoring stations are available or simulated meteorological variables are provided for each point of the space. When considering the predictions for Santiago de Chile, temporal models are often used which do not include the space component. Although these models are particularly good for modeling $PM_{2.5}$ concentrations when the historical values are known, they are not adequate for forecasting pollution at new sites where data are not available. See, for example, the works of [8,14,26]. During the preliminary analysis of data, we considered a temporal model, where we obtained on average an R^2 greater than 0.95 when predicting one hour ahead with seven stations' data. With the proposed space-time model, we obtained on average an R^2 of 0.74 (Table 7) in the same hour-ahead predictions with the same stations' data. This implies that the space-time problem is a more difficult task, since the historical information of the spatial points to be predicted is not included in the training process. If we compare the proposed model with the results of a statistical space-time model proposed by Nicolis et al. [6], in some cases, our RMSE values are lower. However, many differences characterize the two models: (i) while the RMSE in our model is evaluated for one year of prediction using one year for training the model, in the the statistical model of [6], the RMSE is evaluated for one day after using two and a half months for training the model; (ii) the statistical model uses WRF simulations for predicting the meteorological variables, whereas in our case we only use past values collected by monitoring stations located at points of interest for the forecasting of $PM_{2.5}$. We think that our method could improve their predictions if simulations of WRF models or/and other variables are considered as additional input.

7. Conclusions

In this work we have used the LSTM recurrent network for a prediction of $PM_{2.5}$ concentration levels for the city of Santiago de Chile with 11 meteorological stations. By using a recurrent model composed of space-time pollutant and meteorological data, we were able to predict the concentrations of $PM_{2.5}$ in Santiago during the next hour and the next 24 h, reaching values of 0.74 and 0.38 for average R^2 when considering 7 and 11 stations, respectively. This research shows that the task of spatio-temporal prediction of $PM_{2.5}$ pollutant concentration is a more difficult task than typical temporal forecasting because the historical data of the target station are not used in the input data of the training process of the model.

As future work, we propose to improve the prediction quality of these networks by increasing the number of training variables and looking for the relationship that exists in each of the monitoring stations through the distance and positioning of each of these. We also plan to consider different input variables, such as the outputs of meteorological models. Finally, the use of new attention-based architectures or the use of spatial information through convolutional networks stacked with recurrent networks could be proposed for further improving the results. In conclusion, we think that the proposed model constitutes a new approach for space-time pollution prediction that could be used by governmental decision makers for implementing restrictive measures which could prevent negative effects on the human health.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available from: https://1drv.ms/u/s!AnkU8l4kGM9wlhWy22fexCGzHByp?e=ZQ9Hqd, accessed on 1 January 2020.

Acknowledgments: Billy Peralta appreciates the support of the National Center for Artificial Intelligence CENIA FB210017, Basal ANID. Orietta Nicolis appreciates the support of the Research Center for Integrated Disaster Risk Management (CIGIDEN), ANID/FONDAP/15110017 and the ANID-FONDECYT 1201478.

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

References

1. The International Energy Agency. World Energy Outlook Special Report. Available online: https://www.iea.org/weospecialreports/ (accessed on 23 September 2019).
2. Kampa, M.; Castanas, E. Human health effects of air pollution. *Environ. Pollut.* **2008**, *151*, 362–367. [CrossRef] [PubMed]
3. Landrigan, P. Air pollution and health. *Lancet Public Health* **2016**, *2*, E4–E5. [CrossRef]
4. Manisalidis, I.; Stavropoulou, E.; Stavropoulos, A.; Bezirtzoglou, E. Environmental and health impacts of air pollution: A review. *Front. Public Health* **2020**, *8*, 14. [CrossRef]
5. Fuller, R.; Landrigan, P.; Balakrishnan, K.; Bathan, G.; Bose-O'Reilly, S.; Brauer, M.; Caravanos, J.; Chiles, T.; Cohen, A.; Corra, L.; et al. Pollution and health: A progress update. *Lancet Planet. Health* **2022**, *6*, E535–E547. [CrossRef]
6. Nicolis, O.; Díaz Peña, M.; Sahu, S.; Marín, J. Bayesian spatiotemporal modeling for estimating short-term exposure to air pollution in Santiago de Chile. *Environmetrics* **2019**, *30*, e2574. [CrossRef]
7. Rutllant, J.; Garreaud, R. Meteorological air pollution potential for Santiago, Chile: Towards an objective episode forecasting. *Environ. Monit. Assess.* **1995**, *34*, 223–244. [CrossRef] [PubMed]
8. Perez, P.; Gramsch, E. Forecasting hourly PM2.5 in Santiago de Chile with emphasis on night episodes. *Atmos. Environ.* **2015**, *124*, 22–27. [CrossRef]
9. Ostro, B.; Sánchez, J.; Aranda, C.; Eskeland, G. Air pollution and mortality: Results from Santiago, Chile. *Expos. Anal. Environ. Epidemiol.* **1995**, *6*.
10. Valdés, A.; Zanobetti, A.; Halonen, J.; Cifuentes, L.; Morata, D.; Schwartz, J. Elemental concentrations of ambient particles and cause specific mortality in Santiago, Chile: A time series study. *Environ. Health Glob. Access Sci. Source* **2012**, *11*, 82. [CrossRef]
11. Soza, L.; Jordanova, P.; Nicolis, O.; Strelec, L.; Stehlík, M. Small sample robust approach to outliers and correlation of Atmospheric Pollution and Health Effects in Santiago de Chile. *Chemom. Intell. Lab. Syst.* **2018**, *185*, 73–84. [CrossRef]
12. Sahu, S.; Nicolis, O. An evaluation of European air pollution regulations for particulate matter monitored from a heterogeneous network. *Environmetrics* **2008**, *20*, 943–961. [CrossRef]
13. World Health Organization. WHO Air Quality Database 2022. 2022. Available online: https://www.who.int/data/gho/data/themes/air-pollution/who-air-quality-database. (accessed on 1 January 2020).
14. Menares, C.; Perez, P.; Parraguez, S.; Fleming, Z.L. Forecasting PM2.5 levels in Santiago de Chile using deep learning neural networks. *Urban Clim.* **2021**, *38*, 100906. [CrossRef]
15. Ministerio del Medio Ambiente de Chile. Plan de Prevención y Descontaminación Atmosférica Para la Región Metropolitana de Santiago (Decreto 31). 2022. Available online: https://ppda.mma.gob.cl/region-metropolitana/ppda-region-metropolitana/. (accessed on 1 January 2020).
16. Mullins, J.; Bharadwaj, P. Effects of Short-Term Measures to Curb Air Pollution: Evidence from Santiago, Chile. *Am. J. Agric. Econ.* **2015**, *97*, 1107–1134. [CrossRef]
17. Zivin, J.; Neidell, M. The Impact of Pollution on Worker Productivity. *Am. Econ. Rev.* **2011**, *102*, 3652–3673. [CrossRef] [PubMed]
18. Hao, R.; Wan, Y.; Zhao, L.; Liu, Y.; Sun, M.; Dong, J.; Xu, Y.; Wu, F.; Wei, J.; Xin, X.; et al. The effects of short-term and long-term air pollution exposure on meibomian gland dysfunction. *Sci. Rep.* **2022**, *12*, 6710. [CrossRef] [PubMed]
19. Catalano, M.; Galatioto, F.; Bell, M.; Namdeo, A.; Bergantino, A.S. Improving the prediction of air pollution peak episodes generated by urban transport networks. *Environ. Sci. Policy* **2016**, *60*, 69–83. [CrossRef]
20. Liu, Q.; Gao, J. Public Health co-benefits of reducing greenhouse gas emissions. In *Health of People, Health of Planet and Our Responsibility*; Springer: Cham, Switzerland, 2020; pp. 295–307.

21. Sun, W.; Zhang, H.; Palazoglu, A.; Singh, A.; Zhang, W.; Liu, S. Prediction of 24-Hour-Average PM2.5 Concentrations Using a Hidden Markov Model with Different Emission Distributions in Northern California. *Sci. Total. Environ.* **2012**, *443*, 93–103. [CrossRef] [PubMed]
22. Zhu, H.; Lu, X. The Prediction of PM2.5 Value Based on ARMA and Improved BP Neural Network Model. In Proceedings of the 2016 International Conference on Intelligent Networking and Collaborative Systems (INCoS), Ostrava, Czech Republic, 7–9 September 2016; pp. 515–517. [CrossRef]
23. Sudumbrekar, A.; Kale, R.; Kaurwa, T.; Mule, V.; Devkar, A., Feasibility Study of ARIMA Model for PM2.5 Prediction using Real-world Data Gathered from Pune Region. In *New Frontiers in Communication and Intelligent Systems*; SCRS: New Delhi, India, 2021; pp. 105–111. [CrossRef]
24. Mahajan, S.; Chen, L.J.; Tsai, T.C. An Empirical Study of PM2.5 Forecasting Using Neural Network. In Proceedings of the 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), San Francisco, CA, USA, 4–8 August 2017. [CrossRef]
25. Subramaniam, S.; Raju, N.; Ganesan, A.; Rajavel, N.; Chenniappan, M.; Prakash, C.; Pramanik, A.; Basak, A.K.; Dixit, S. Artificial Intelligence Technologies for Forecasting Air Pollution and Human Health: A Narrative Review. *Sustainability* **2022**, *14*, 9951. [CrossRef]
26. Perez, P.; Trier, A.; Reyes, J. Prediction of PM2.5 Concentrations Several Hours in Advance Using Neural Networks in Santiago, Chile. *Atmos. Environ.* **2000**, *34*, 1189–1196. [CrossRef]
27. Diaz-Robles, L.; Ortega Bravo, J.C.; Fu, J.; Reed, G.; Chow, J.; Watson, J.; Moncada, J. A Hybrid ARIMA and Artificial Neural Networks Model to Forecast Particulate Matter in Urban Areas: The Case of Temuco, Chile. *Atmos. Environ.* **2008**, *42*, 8331–8340. [CrossRef]
28. Sahu, S.K. 16—Hierarchical Bayesian Models for Space—Time Air Pollution Data. In *Time Series Analysis: Methods and Applications*; Handbook of Statistics; Subba Rao, T.; Subba Rao, S.; Rao, C., Eds.; Elsevier: Amsterdam, The Netherlands, 2012; Volume 30, pp. 477–495.
29. Hochreiter, S.; Schmidhuber, J. Long Short-term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
30. Hamdi, A.; Shaban, K.B.; Erradi, A.; Mohamed, A.; Rumi, S.K.; Salim, F.D. Spatiotemporal data mining: A survey on challenges and open problems. *Artif. Intell. Rev.* **2022**, *55*, 1441–1488. [CrossRef] [PubMed]
31. Kyriakidis, P.; Journel, A. Geostatistical Space—Time Models: A Review. *Math. Geol.* **1999**, *31*, 651–684. [CrossRef]
32. Cressie, N.; Wikle, C. *Statistics for Spatio-Temporal Data*; Wiley: Hoboken, NJ, USA, 2015.
33. Amato, F.; Guignard, F.; Robert, S.; Kanevski, M.F. A novel framework for spatio-temporal prediction of environmental data using deep learning. *Sci. Rep.* **2020**, *10*, 22243. [CrossRef] [PubMed]
34. Ghaderi, A.; Sanandaji, B.M.; Ghaderi, F. Deep Forecast: Deep Learning-based Spatio-Temporal Forecasting. *arXiv* **2017**, arXiv:1707.08110.
35. Bai, L.; Yao, L.; Li, C.; Wang, X.; Wang, C. Adaptive Graph Convolutional Recurrent Network for Traffic Forecasting. *arXiv* **2020**, arXiv:2007.02842.
36. Shang, Z.; Deng, T.; He, J.; Duan, X. A novel model for hourly PM2. 5 concentration prediction based on CART and EELM. *Sci. Total. Environ.* **2019**, *651*, 3043–3052. [CrossRef] [PubMed]
37. Wang, D.; Wei, S.; Luo, H.; Yue, C.; Grunder, O. A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. *Sci. Total. Environ.* **2017**, *580*, 719–733. [CrossRef] [PubMed]
38. Fan, J.; Li, Q.; Hou, J.; Feng, X.; Karimian, H.; Lin, S. A Spatiotemporal Prediction Framework for Air Pollution Based on Deep RNN. *ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci.* **2017**, *IV-4/W2*, 15–22. [CrossRef]
39. Huang, C.J.; Kuo, P.H. A Deep CNN-LSTM Model for Particulate Matter (PM2.5) Forecasting in Smart Cities. *Sensors* **2018**, *18*, 2220. [CrossRef]
40. Hong, K.; Pinheiro, P.; Weichenthal, S. Predicting Global Variations in Outdoor PM2.5 Concentrations using Satellite Images and Deep Convolutional Neural Networks. *arXiv* **2019**, arXiv:1906.03975.
41. 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]
42. Franceschi, F.; Cobo, M.; Figueredo, M. Discovering relationships and forecasting PM 10 and PM 2.5 concentrations in Bogotá, Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering. *Atmos. Pollut. Res.* **2018**, *9*, 912–922. [CrossRef]
43. Belavadi, S.V.; Rajagopal, S.; R, R.; Mohan, R. Air Quality Forecasting using LSTM RNN and Wireless Sensor Networks. *Procedia Comput. Sci.* **2020**, *170*, 241–248. [CrossRef]
44. 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]
45. Zou, X.; Zhao, J.; Zhao, D.; Sun, B.; He, Y.; Fuentes, S. Air quality prediction based on a spatiotemporal attention mechanism. *Mob. Inf. Syst.* **2021**, *2021*, 6630944. [CrossRef]
46. Zhang, Q.; Lam, J.C.; Li, V.O.; Han, Y. Deep-AIR: A hybrid CNN-LSTM framework forFine-grained air pollution forecast. *arXiv* **2020**, arXiv:2001.11957.

47. Cerqueira, V.; Torgo, L.; Mozetič, I. Evaluating time series forecasting models: An empirical study on performance estimation methods. *Mach. Learn.* **2020**, *109*, 1997–2028. [CrossRef]
48. van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **2011**, *45*, 1–67. [CrossRef]
49. Hota, H.; Handa, R.; Shrivas, A. Time series data prediction using sliding window based RBF neural network. *Int. J. Comput. Intell. Res.* **2017**, *13*, 1145–1156.
50. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
51. Díaz, M.; Nicolis, O.; Marín, J.C.; Baran, S. Statistical post-processing of ensemble forecasts of temperature in Santiago de Chile. *Meteorol. Appl.* **2019**, *27*, e1818. [CrossRef]