Ecological Informatics 55 (2020) 101019

Contents lists available at ScienceDirect

Ecological Informatics

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

Predicting air quality with deep learning LSTM: Towards comprehensive models

Ricardo Navares, José L. Aznarte*

Department of Artificial Intelligence, UNED, Juan del Rosal, 16, 28040 Madrid, Spain

ARTICLE INFO

Keywords:
Air quality
Forecast
Neural networks
Deep learning

ABSTRACT

In this paper we approach the problem of predicting air quality in the region of Madrid using long short term memory recurrent artificial neural networks. Air quality, in this study, is represented by the concentrations of a series of air pollutants which are proved as risky for human health such as CO, NO_2, O_3, PM10, SO_2 and airborne pollen concentrations of two genus (Plantago and Poaceae). These concentrations are sampled in a set of locations in the city of Madrid. Instead of training an array of models, one per location and pollutant, several comprehensive deep network configurations are compared to identify those which are able to better extract relevant information out of the set of time series in order to predict one day-ahead air quality. The results, supported by statistical evidence, indicate that a single comprehensive model might be a better option than multiple individual models. Such comprehensive models represent a successful tool which can provide useful forecasts that can be thus applied, for example, in managerial environments by clinical institutions to optimize resources in expectation of an increment of the number of patients due to the exposure to low air quality levels.

1. Introduction

In the last decades, air quality has been gaining attention due to the health threats produced by high levels of environmental pollution (Ozkaynak et al., 2009). Within the context of this study, air quality is related to both chemical pollutants and biotic factors present in the environment. Concretely, chemical pollutants are considered as the agents released in the environment which disrupt ecosystems such as CO, O_3, NO_2, SO_2 and PM10 which are also considered as the main air chemical pollutants in the studied region (Querol et al., 2012). On the other hand, biotic factors refer to airborne pollen concentrations of the Plantago and Poaceae genus which are the most common and aggressive in terms of allergic and respiratory disorders (Subiza et al., 1995).

Air quality information systems are increasingly used to predict future air pollution levels, which allows for alerting about peaks in admissions in clinical institutions, traffic and environmental management in urban areas or minimizing the exposure for patients in order to prevent adverse effects (Abraham et al., 2009; González et al., 2001; Linares and Díaz, 2008; Ozkaynak et al., 2009).

Field experts have been employing observation-based models to relate records of pollutants to one or more variables which can be measured of predicted, usually meteorological data (Navares and Aznarte, 2016; Sabariego et al., 2012; Schaber and Badeck, 2003; Silva-Palacios et al., 2016; Smith and Emberlin, 2006). Despite the extensive literature, few consider the problem of taking into account both types of pollutants altogether as they are inherently different problems: atmospheric pollen concentrations depend on plant development during previous seasons which, at the same time, depends on the climatological conditions during plant evolution (Cannell and Smith, 1983; Smith and Emberlin, 2006). This implies long and mid-term relations between past atmospheric conditions and current plant status. Contrarily, chemical air pollutant levels are related to recent past atmospheric conditions (Navares et al., 2018). Both pollutants show influence on the development and attack of, for instance, allergic respiratory diseases (D'Amato et al., 2011).

Neural network models have been successfully applied to environmental modeling (Gardner and Dorling, 1998) and air quality problems (Castellano-Méndez et al., 2005; Chaloulakou et al., 1998; Chelani et al., 2002; Grivas and Chaloulakou, 2006; Iglesias-Otero et al., 2015). However, given the nature of the problem (short term influential variables for chemical pollutants and mid-long term influential variables for pollen) this approach requires a thorough research and selection of relevant variables based on expert knowledge (Andersen, 1991; Catalano et al., 2016; Navares and Aznarte, 2016). In addition to the temporal dimension, it is important to take into account the spatial

* Corresponding author.
E-mail address: jlaznarte@dia.uned.es (J.L. Aznarte).

https://doi.org/10.1016/j.ecoinf.2019.101019
Received 22 August 2019; Received in revised form 23 September 2019; Accepted 4 October 2019
Available online 24 October 2019

interactions between observation stations as they might be implicitly related. These approaches imply a new research process which might add a new set of influential variables every time a new kind of pollutant or a new genus of pollen is taken into consideration by the system.

In this paper we propose several long short term memory (LSTM) network setups (Hochreiter and Schmidhuber, 1997) to gain insights on how influential is network design when dealing with interrelated time series of different nature. The study compares network topologies in order to find the most suitable configuration to solve the problem, exposing the advantages and disadvantages of each one. The objective is twofold: on the one hand, we show how to avoid thorough preprocessing steps to find influential features (both long and short term, and with differences at each location as a result of particular environmental conditions of the areas where the observation stations are located) by letting the network extract them regardless of the pollutant type. Such a unified approach avoids manually fitting one specific model per pair of location and pollutant, saving human resources and increasing the scalability of the system. On the other hand, we provide a convenient network topology for accurate forecasts at each location which is able to obtain relevant information from data both temporal and spatial. The problem chosen to prove the validity of the proposal is the prediction of air quality over a dataset which consist of a dozen of time series with different characteristics and regimes, sampled over 13 adjacent locations.

2. Materials and methods

2.1. Data description

Chemical air pollutants were measured using the gravimetric method or an equivalent method (β-attenuation) and were provided by the Madrid Municipal Air Quality Monitoring Grid (http://www.mambiente.munimadrid.es/). The grid consists of a network of 24 urban background stations spread across the city, which capture chemical air pollutants in real time. Hourly data was aggregated to obtain daily mean levels of chemical air pollutants for the study period from 01 to 01-2001 to 31-12-2013. Daily mean concentrations ($\mu g/m^3$) of particulate matter 10 μm in diameter (PM10), carbon monoxide, surface ozone (O_3) and nitrogen dioxide (NO_2) in $\mu g/m^3$.

Pollen observations correspond to daily grains per cubic meter of Poaceae and Plantago pollen registered at Complutense University of Madrid (Pharmacy Faculty). Pollen counts followed the standard methodology of the Spanish Aerobiological Network (Galán Soldevilla et al., 2007) and were provided by Red Palinológica de la Comunidad de Madrid.

Weather observations consist of average daily temperature in Celsius degrees, wind speed measured in *m/s*, daily rainfall in *mm/h*, pressure in *hPa* and degree of humidity in percentage. Data sets for locations (Table 1) consist of observations from 01 to 01-2001 to 31-12-2013. In the presence of missing hourly data, when these observations are less 20% within a day, averages and aggregations were calculated ignoring missing values, otherwise, the daily data is considered as missing. This approach led to complete time series during the study period.

Different pollutants have different nature of the time series having significant seasonal patterns. Some of them show higher variability around the seasonal component as in the case of O_3 and PM10, while others present highers peaks and less variability during the rest of the year as in the pollen counts. It is noticeable that SO_2 and CO concentrations dramatically decreased when compared to early observations due to the progressive transition from coal-powered heating systems to natural gas and the progressive upgrade of the urban car fleet (Díaz et al., 2007).

2.2. Methodology

Compared to traditional neural networks, recurrent neural networks (RNN) are implemented with loops or connections between units allowing information persistence from one step of the network to the next. The ability to map input sequences to output sequences by incorporating past context into their internal state makes them especially promising for tasks that require to learn how to use past information such as time series analysis. RNNs can be thought of as multiple copies of the same neural network, each transferring information to its successor and forming a chain-like architecture which is naturally related to sequences.

RNNs might be able to look at recent information to perform a present task which makes them suitable for time series predictions. However, relevant information might appear further in the past and, as the time gap grows, RNNs are unable to connect the information.

Long short-term memory networks (LSTM) were first introduced in 1997 by (Hochreiter and Schmidhuber, 1997) and improved in 2000 by (Gers et al., 2000). They are a variation of RNNs capable of learning long-term dependencies by including in the architecture special units called memory blocks. In addition, multiplicative units called gates control the flow of information from a LSTM unit to another.

An LSTM unit performs self-loops which enable the flow of the gradient for long durations, enabling it to deal with the vanishing gradient problem. Together with an input gate, an output gate and a forget gate, this architecture models the short-term memory that allows the network to learn over many time steps. For this reason, LSTM had been shown to outperform more traditional recurrent networks on several temporal processing tasks (Gers et al., 2000).

The common scoring rule root mean squared error (RMSE) will be used to measure the average magnitude of the error of several network configurations:

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

where y_i is the observed i^{th} data point, $\hat{y}_i$ the predicted and n the total number of data points in the test set. As benchmark models the LSTMs are compared to the traditional linear regression (LinReg) and computational intelligence technique Random Forest (RF) proposed by (Breiman, 2001). In order to evaluate the results obtained by the algorithms, we use a nonparametric Friedman test (Friedman, 1937) test with a post-hoc procedure as described in (Navares and Aznarte, 2016) to determine, in significance terms, which algorithms are considered the best performers based on their RMSE.

The Friedman test (Friedman, 1937) is a non-parametric analogue of the parametric two-way ANOVA. The objective of the application of the test is to determine if there is a difference among model performances and consequently, whether one (or more) is consistently better than the others. Given that non-parametric hypothesis tests are applied to nominal or ordinal data, the original computed RMSE for each location is converted to its correspondent rank within the set and combined by averaging: $R_j = \frac{1}{n}\sum_i r_i^j$ (where j denotes the model, i refers to each location and n is the total number of pairs $\{\text{model}, \text{location}\}$). Since the error is being used to compare the models, the highest rank 1 will be assigned to the highest error, thus the worst performer. The null hypothesis of equality of medians is tested by the F-statistic

$$F = \frac{12n}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right], \tag{2}$$

where k is the number of algorithms and $F \sim {\chi_{k-1}}^2$. Still, this test is not sufficient as it only indicates the presence of significant differences in the whole variable space. A ranking conversion is computed to obtain the p-value of each pair (Conover, 1999). The former is a valid procedure to compare two models but is not suitable for multiple comparison

Table 1
Availability of variables and locations.

	Long.	Lat.	CO	NO_2	O_3	Plantago	Poaceae	PM10	SO_2	Pr	R	Hum	T	W
ArturoSoria	3° 38′ W	40° 26′ N	*	*	*						*			
BarrioPilar	3° 42′ W	40° 28′ N	*	*	*						*			
CasaCampo	3° 44′ W	40° 25′ N	*	*	*			*	*	*	*	*	*	*
CuatroCaminos	3° 42′ W	40° 26′ N		*					*		*		*	
Farmacia	3° 45′ W	40° 27′ N				*	*							
Farolillo	3° 43′ W	40° 23′ N	*	*	*				*		*		*	
Moratalaz	3° 38′ W	40° 24′ N	*	*				*	*					
PlazaEspana	3° 42′ W	40° 25′ N	*	*					*	*	*	*	*	*
PzadelCarmen	3° 42′ W	40° 25′ N	*	*	*				*					
PzaLadreda	3° 43′ W	40° 23′ N								*	*	*	*	*
RamonyCajal	3° 40′ W	40° 27′ N		*							*			
StaEugenia	3° 36′ W	40° 22′ N									*		*	*
Vallecas	3° 39′ W	40° 23′ N		*				*	*		*			

Pr: Pressure.
R: Rain.
Hum: Hummidity.
W: Wind speed.
T: Average Temperature.
* represent the presence of data in the corresponding location and for the corresponding pollutant

as there is no control of error propagation (Type I errors) when making more than one comparison.

Thus, once the existence of significant differences in the group of models is evidenced, a post-hoc test adjusts the value of the significance level α at each pairwise comparison to allow multiple comparisons. (Holm, 1979) proposed the adjustment by selecting the p-values of each test, starting with the most significant p_i, and test the hypothesis of $H_i : p_i > \alpha/(k - i)$, being k the total number of models in our proposal. If H_i is rejected then allows to test H_{i+1}, being p_{i+1} the next most significant p-value and so on. An extension of this step-down method was proposed by (Shaffer, 1986), which uses a logical relation between the combination of the hypotheses of all pairwise comparisons. For instance, if a model a_1 is better/worse than a_2, it is not possible that a_1 is as good/bad as a_3 and a_2 has the same performance as a_3. Based on this argument and following Holm's method, instead of rejecting $H_i : p_i \leq \alpha/(k - i)$, rejects $H_i \leq \alpha/t_i$, being t_i the maximum number of hypotheses which can be true given the number of false hypotheses in $j \in \{1, \ldots, i\}$.

2.3. Experimental design

The aim is to provide the best one day-ahead forecast in terms of accuracy, and consequently to see if LSTMs are able to efficiently store relevant information over time to position themselves as a strong candidate when deciding which technique to use in such problems. In order to do so, the full historic data set was split into a training set consisting of the period between 01 and 01-2001 and 31-12-2012, leaving the last period (01-01-2013 to 31-12-2013) as a test set.

Selecting the topology of LSTM networks depends on the application domain and there is no general rule of thumb for the amount of hidden nodes that should be used. It has to be figured out case-specifically by trial and error. Thus, starting with the simplest network of one memory cell, the architecture is extended by including units in the layer. The stop condition is given by the validation error from the random selected 10-fold cross-validation on the training set. Even though time series show inherent serial correlation and potential non-stationarity of the data, (Bergmeir et al., 2018) proves that cross-validation empirically compares to out-of-sample or other time-series-specific valuation techniques.

Different architectures are proposed to check the convenience of each one to solve the problem. The first one consists on a fully connected LSTM (FC-LSTM), which is the first and most common approach, where all the input variables are parsed through a 500 LSTM hidden units layer to test the capability of the LSTM to obtain and discriminate relevant information. This layer is connected to a layer of 100 sigmoid units to obtain higher order relations among the locations and pollutants (Fig. 1(a)). The idea is to see if the first layer captures the temporal dependencies while the subsequent dense layer deals with location dimensions which will be then transferred to the output layer.

The second configuration is tailored to ease the discrimination of information by forcing the LSTM units to target different groups of pollutants (GP-LSTM). The core idea is to force each LSTM group to focus on its correspondent pollutant outputs assuming that the same pollutants behave similarly across locations and that the LSTM units discriminate input information to subsequently obtain the relations of the locations only per pollutant group. Thus, the LSTM layer consists of 100 units per group which receive the full set of 1-day lagged input variables and it is connected only to a number of outputs equal to the number of observation stations of that group (Fig. 1(b)). The intention is to facilitate that LSTM units extract the information of each individual group of pollutants and then obtain the relations among groups in the output layer. For instance, in the case of carbon monoxide, the 57 input variables (Table 1) feed 100 LSTM units which are fully connected to 7 outputs, one per station where CO is available. This configuration totals 700 LSTM units, 100 per pollutant.

An alternative setup with a similar configuration and the same aim of easing the discrimination process is also proposed. In this case, the network is assisted by only using as input for each group of LSTM units individual groups of pollutants (IGP-LSTM) and the meteorological variables as seen in Fig. 1(c). Each group consists of 100 LSTM units with a total of 700 LSTM units in the layer, one per group of pollutants, which are fully connected to the output layer. Since the inputs are split by pollutant, the reason behind including a fully connected output layer is to include the interactions among pollutants.

Finally, a more simple version was taken into consideration by training the network using the full set of inputs to feed 7 groups of 66 LSTM units (462 in total), one group per single combination pollutant-station (SP-LSTM, Fig. 1(d)) which outputs one single target variable. This is equivalent to training 31 individual networks (one per target pollutant and station) with a set up 57–66-1 where 57 is the number of 1-day lagged input variables (Table 1), 66 LSTM units and 1 is the output. Fig. 1 summarizes the four different network topologies which were tested against the benchmark algorithms.

In order to obtain the relationship between the inputs and outputs, the well-known *backpropagation* algorithm (Rumelhart et al., 1986) was proposed, employing as loss function the mean squared error since it is a regression problem. As an alternative of the classic Stochastic Gradient Descent (SGD) optimization model to fit network weights, the *Adam* algorithm proposed by (Kingma and Ba, 2019) was used with a

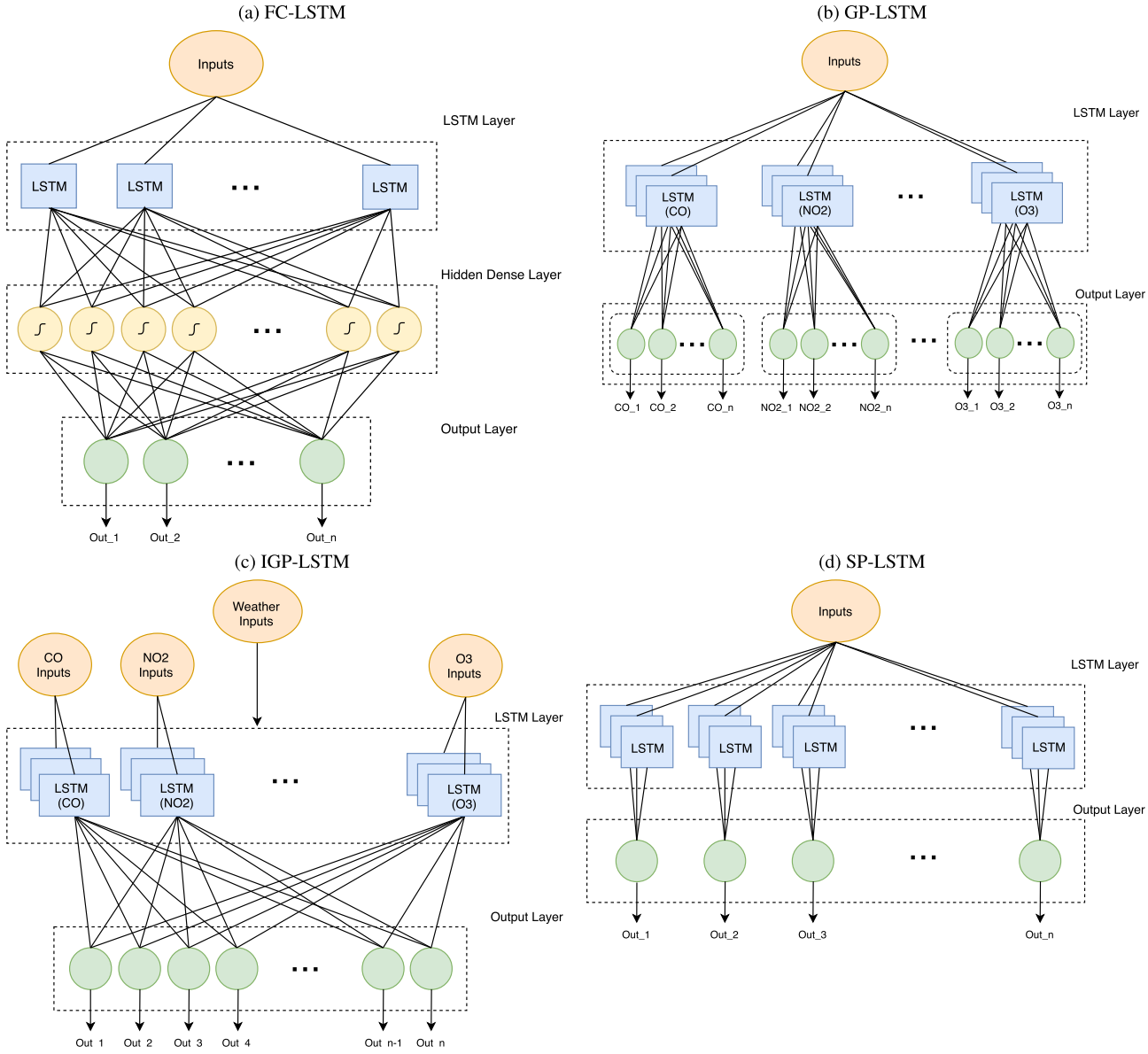

Fig. 1. LSTM architectures: (a) Fully connected LSTM (FC-LSTM), (b) LSTM grouped by pollutant class (GP-LSTM), (c) LSTM fed by pollutant variable (IGP-LSTM), (d) One variable LSTM (SP-LSTM).

learning rate $\alpha = 0.001$, an exponential decays of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ were used as suggested by (Ruder, 2016). As opposed to SGD which maintains a single learning rate α for all weight updates, the method computes individual adaptive learning rates from the estimates of first and second moments of the gradients (Kingma and Ba, 2019). Specifically, the algorithm uses an exponential moving average of the gradient and its square using the parameters β_1 and β_2 to control the decay rates of these moving averages. Performances are compared to the traditional and commonly used linear regression enabling the identification and the characterization of relationships among each pollutant and meteorological variables and its 1-day lagged observation. Therefore, one linear model is fitted per pollutant. As an extension, an identical setup is used to train random forests to compare with another family of computational intelligence models.

3. Results

Table 2 shows prediction errors for each pollutant at each location, while Table 2 shows the average prediction errors for each pollutant. Linear regression obtains an average RMSE of 0.107 across all location for carbon monoxide (CO) while Random Forest manages to diminish this error to 0.086. All LSTM configurations outperform Random Forest results except SP-LSTM which results in an average RMSE of 0.093 mainly due to the error at Farolillo where it underperforms with an RMSE of 0.127. GP-LSTM is the most accurate in forecasting CO levels with an average RMSE of 0.083. In Fig. 3 can be seen the percentage improvement of each algorithm with respect to LinReg. There is an average improvement around 20% with all computational intelligence based algorithms for this pollutant, except for the aforementioned SP-LSTM, which performs poorly (-20.27%) at the observation station of Farolillo.

With regard to Nitrogen dioxide, an average RMSE of 12.07 is shown for LinReg followed by RF and SP-LSTM with an error of 10.60 and 11.00 respectively. The remaining LSTM configurations reduce this error to RMSE levels lower than 9.70 implying an increase in accuracy of around 20% when compared to LinReg as shown in Fig. 3.

Results for ozone are paired and oscillate around an error of 11 $\mu g/m^3$ being RF and FC-LSTM the worst and best performer in average

Table 2
RMSE per location and variable. Highlighted in bold best average performer for each pollutant.

Pollutant	Station	LinReg	RF	SP-LSTM	IGP-LSTM	GP-LSTM	FC-LSTM
CO ($\mu g/m^3$)	ArturoSoria	0.087	0.072	0.086	0.080	0.080	0.087
	BarrioPilar	0.137	0.106	0.120	0.119	0.111	0.114
	CasaCampo	0.072	0.055	0.055	0.054	0.059	0.057
	Farolillo	0.106	0.100	0.127	0.083	0.082	0.077
	Moratalaz	0.110	0.087	0.079	0.081	0.075	0.078
	PzaCarmen	0.124	0.087	0.089	0.080	0.082	0.083
	PzaEspana	0.116	0.093	0.096	0.097	0.091	0.091
Average		0.107	0.086	0.093	0.085	**0.083**	0.084
NO_2 ($\mu g/m^3$)	ArturoSoria	12.310	11.817	13.052	10.103	10.312	10.514
	BarrioPilar	13.685	11.623	11.933	10.560	10.427	10.479
	CuatroCaminos	11.885	10.290	10.422	10.442	9.471	9.474
	CasaCampo	10.047	8.232	7.826	7.515	7.439	7.632
	Farolillo	10.667	8.551	8.966	8.114	8.020	7.862
	Moratalaz	13.163	12.325	12.312	10.607	9.799	9.932
	PzaCarmen	11.002	9.710	9.448	8.300	8.004	8.288
	PzaEspana	12.041	11.454	11.561	10.339	10.005	10.875
	RamonyCajal	13.629	12.515	13.248	11.588	11.197	10.731
	Vallecas	12.327	9.550	11.273	8.983	9.739	8.592
Average		12.076	10.607	11.004	9.655	9.442	**9.438**
O_3 ($\mu g/m^3$)	ArturoSoria	10.501	10.945	10.957	10.836	11.795	10.139
	BarrioPilar	11.442	11.471	11.484	10.796	11.343	10.506
	CasaCampo	12.392	13.809	13.866	12.042	14.108	12.526
	Farolillo	11.471	12.268	12.039	11.550	11.981	10.872
	PzaCarmen	10.063	11.507	9.730	9.742	9.536	9.840
Average		11.174	12.000	11.615	10.993	11.753	**10.777**
PM10 ($\mu g/m^3$)	CasaCampo	8.094	6.314	6.784	5.323	5.350	5.700
	Moratalaz	8.244	6.458	5.969	6.147	5.965	6.382
	Vallecas.	9.005	6.672	6.102	5.856	5.847	5.548
Average		8.447	6.481	6.285	5.776	**5.721**	5.877
SO_2 ($\mu g/m^3$)	CuatroCaminos	2.213	1.815	1.787	1.571	1.695	1.727
	CasaCampo	0.831	1.033	1.188	0.613	0.665	0.742
	Farolillo	1.253	1.149	1.366	0.832	0.809	0.855
	Moratalaz	4.186	4.965	4.396	4.192	4.165	4.065
	PzaCarmen	1.869	1.709	1.779	1.730	1.605	1.748
	PzaEspana	1.336	1.218	1.631	1.122	1.159	1.202
	Vallecas	1.455	1.342	1.206	1.272	1.268	1.422
Average		1.877	1.890	1.908	**1.619**	1.624	1.680
Plantago (grains/m^3)	Farmacia	6.948	9.927	7.927	6.938	6.875	7.296
Poaceae (grains/m^3)	Farmacia	24.344	24.220	25.738	23.686	25.268	24.933
Average		15.646	17.074	16.833	**15.312**	16.072	16.114

respectively. It can be seen in Fig. 3 that RF and SP-LSTM underperform LinReg while the other networks show mixed results depending on the location.

On the other hand, computational intelligence models clearly overcome, in terms of accuracy, linear regression when predicting particulate matter (PM10), having an average improvement between 25% and 30% except for RF and SP-LSTM which perform 23.16% and 6.18% better respectively. Similar situation occurs when using GP-LSTM and FC-LSTM to predict SO_2 averaging a RMSE of 1.624 and 1.68 respectively. Although the best average performer for this pollutant is IGP-LSTM, it fails to improve LinReg results at Moratalaz where the error is 0.16% higher. For the pollen series, IGP-LSTM is the only model which improves LinReg results in both series by 0.15% and 2.70% in Plantago and Poaceae respectively.

In general, both LSTM and RF perform better than linear regression (as seen in the left part of Fig. 2). Table 2 shows that LSTM architectures achieved lower errors. Figs. 2 and 3 clearly show these improvements are higher for IGP-LSTM, GP-LSTM and FC-LSTMS than the other models. Regarding bias, shown in the right part of Fig. 2, we can see how GP.LSTM and IGP.LSTM show zero bias in median while the former produces predictions with very low bias also in mean.

In light of the combined error/bias results, we would be tempted to choose GP.LSTM as the best overall model. In order to investigate this and provide statistical evidence, a Friedman rank test was performed over the errors shown in Table 2. A Friedman statistic of $F = 72.96$, distributed according a χ^2 with 5 degrees of freedom obtains a p-value of 4.34e-11 with $\alpha = 0.05$, which provides strong evidence of the existence of significant difference between the algorithms ranked GP-LSTM as the best performer (as expected), followed by FC-LSTM and IGP-LSTM and RF, SP-LSTM and linear regression with ranks 4, 5 and 6 respectively.

Since Friedman's null hypothesis was rejected, a post-hoc pairwise comparison was carried out to check the differences between the proposed algorithms. Table 3 shows there is strong evidence of differences between GP-LSTM, IGP-LSTM, FC-LSTM and linear regression (hypotheses 1, 2 and 3) and SP-LSTM (hypotheses 4, 5 and 6) which implies, given Friedman ranks, that GP-LSTM, IGP-LSTM and FC-LSTM perform better than the benchmark LinReg and the SP-LSTM configuration. Hypotheses 7, 8 and 9 provide statistical evidence of GP-LSTM, IGP-LSTM and FC-LSTM performing better than RF. Lastly, there is no difference in terms of error between SP-LSTM and the benchmark algorithms.

4. Discussion

As we have seen in Section 3 there is statistical evidence of the outperformance of GP-LSTM, IGP-LSTM and FC-LSTM with respect to the other proposed methods. This situation is clear for CO, NO_2 and PM10 where there is an error reduction higher than 10% at all locations except for Arturo Soria when forecasting CO. With a close look at this location, we have seen a yearly average concentrations of 0.40 $\mu g/m^3$ with a standard deviation of 0.24 $\mu g/m^3$ while this average goes over 0.47 $\mu g/m^3$ with a standard deviation of at least 0.32 $\mu g/m^3$ for the remaining locations, suggesting a lower improvement when the

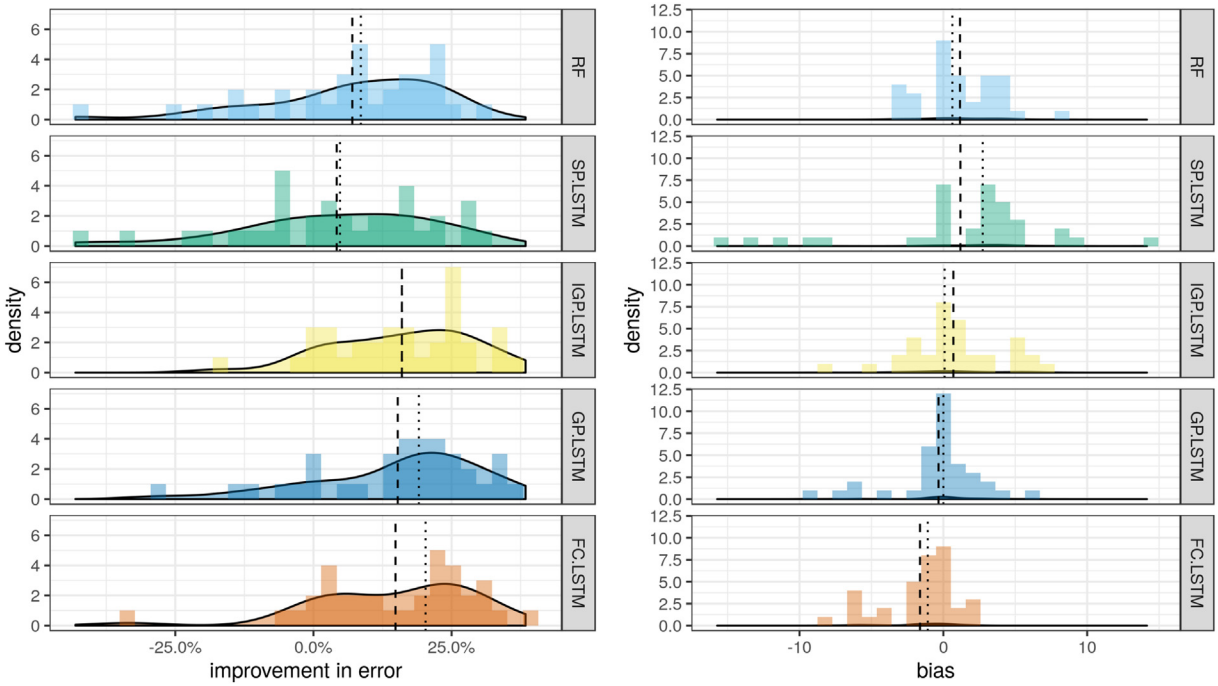

Fig. 2. Relative improvement with respect to LinReg of each of the methods applied (left) and bias (right). Dashed vertical line represents the mean, dotted vertical line represents median.

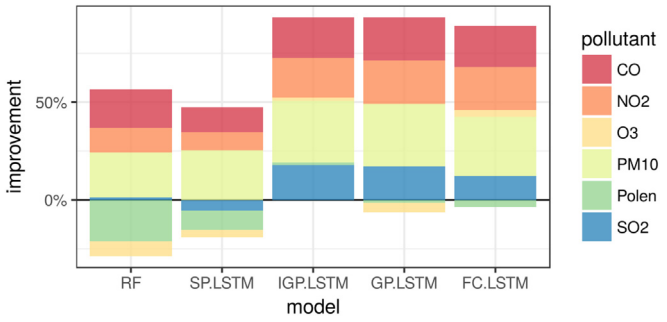

Fig. 3. Average improvement over LinReg per pollutant.

Table 3
Adjusted Holm and Schaffer p-values with pairwise rejected hypothesis at $\alpha = 0.05$.

i	Hypothesis	$p_{unadj.}$	p_{Holm}	p_{Shaf}
1	Linear vs GP-LSTM	5.171E-10	7.756E-9	7.756E-9
2	Linear vs IGP-LSTM	9.131E-9	1.369E-7	9.131E-8
3	Linear vs FC-LSTM	9.131E-9	1.369E-7	9.131E-8
4	SP-LSTM vs GP-LSTM	3.815E-7	5.722E-6	3.815E-6
5	SP-LSTM vs IGP-LSTM	4.021E-6	4.021E-5	4.0215E-5
6	SP-LSTM vs FC-LSTM	4.021E-6	4.423E-5	4.021E-5
7	RF vs GP-LSTM	1.398E-4	0.001	9.787E-4
8	RF vs IGP-LSTM	8.354E-4	0.006	0.005
9	RF vs FC-LSTM	8.354E-4	0.006	0.005
10	Linear vs RF	0.016	0.096	0.096
11	RF vs SP-LSTM	0.204	1.021	0.817
12	Linear vs SP-LSTM	0.256	1.024	1.024
13	IGP-LSTM vs GP-LSTM	0.639	1.919	1.919
14	GP-LSTM vs FC-LSTM	0.639	1.919	1.919
15	IGP-LSTM vs FC-LSTM	1.0	1.919	1.919

variability reduces. Nevertheless, most of the improvements are over 20%.

It is to note the low performance of SP-LSTM when forecasting CO at Farolillo (−20.27%). A simple analysis of the series shows that the standard deviation taken by years goes from 0.73 $\mu g/m^3$ in 2001 to 0.12 $\mu g/m^3$ in 2012 while this change in the behavior of the series is more constant and smoother at other locations. Not including a fully connected layer, either a dense sigmoid or output, weights more past information from the only target series (CO at Farolillo) incurring in overfitting.

The best performing configurations, GP-LSTM, IGP-LSTM and FC-LSTM, manage to obtain good improvements in the case of SO_2 even though, there are dramatic changes on series patterns with respect to past years due to the progressive transition from coal-powered heating to natural gas. An exception is the results at Moratalaz where the error improvement decreases for all models with respect to their performances at other locations. The particular location of this station makes the observed levels of SO_2 behave differently when compared to other areas. It is located between the A3 and R3 highways, which are the two main access to Madrid from the East, and M30 and M40 which are city main beltways. Consequently, SO_2 levels remain high due to the dense traffic. While SO_2 mean values for the most recent years of this study stay around 3.15 $\mu g/m^3$ with average maximums of 18.90 $\mu g/m^3$, at Moratalaz these means double up to 6.43 $\mu g/m^3$ with maximums of 34.66 $\mu g/m^3$.

Random Forest does not improve LinReg when predicting O_3 at all studied locations. A similar situation occurs with SP-LSTM with the exception of Plaza del Carmen, suggesting these two proposals are not able to capture the strong seasonal pattern of this pollutant. These patters are captured by FC-LSTM improving LinReg average performance by 3.5% although this does not apply for the observations at Casa de Campo. The reason is because tropospheric ozone behaves opposite to other pollutants as its concentration levels are higher outside urban centers where the air quality is assumed as clean in general. Not only is it formed due to directly traffic or industrial emissions but also when combined these with airborne pollutants and solar radiation (Sharma et al., 2016), having Casa de Campo all the conditions to concentrate high levels compared to other locations as it is the largest public park in Madrid.

With respect to airborne pollen concentrations, IGP-LSTM is the only model which managed to improve LinReg results for both pollen

genus. Pollen series are particularly characteristic since they show very low concentrations (or almost none) during the calendar year until the main pollination season where the high peaks appear. Nevertheless, IGP-LSTM is able to identify those peaks, specially in the case of Plantago where the test set includes the highest peak (around 120 grains/m^3) among all observed train data.

Fig. 2 (left hand side) shows that IGP-LSTM presents a higher consistency of results as the median improvement with respect to LinReg does not differ much from the average. On top of that, IGP-LSTM presents the shorter negative tail among all the models, while GP-LSTM and FC-LSTM show a heavy-tailed distribution of improvements suggesting these configurations are not as good as IGP-LSTM when obtaining the characteristics of some time series. In fact, Fig. 3 shows that, on average, they do not manage to improve LinReg results for O_3 and Pollen respectively. However, when considering the bias of the predictions (right hand side of Fig. 2) we see how the predictions of GP-LSTM are, in general, slightly better than the rest.

5. Conclusions

This paper presents a comparison study of different LSTM configurations in order to obtain the most suitable to forecast air quality in the region of Madrid. Several pollutants showing very different behaviors were taken into consideration. In addition to the intrinsic differences in the behavior among pollutant types, each pollutant behaves differently at each location given the conditions of the zones studied due to several factors such as traffic congestion or green areas. This adds an extra dimension to the complexity problem which let test the capability of the proposed models to obtain relevant information for forecasting.

We have seen that there is statistical evidence that the LSTM grouped by pollutant class (GP-LSTM) and the LSTM with individual groups of pollutants as inputs (IGP-LSTM) outperform benchmark algorithms and the other two proposals. Furthermore, the GP-LSTM shows smaller bias, but evidence also shows that performing a discriminative input of the groups of pollutants as in IGP-LSTM eases the network to focus on the relevant information and provides more stable results across locations.

By including in the configuration fully connected layers, either as an output or hidden layer, the networks are able to better identify the relations among pollutants with no data preprocessing. However, we have seen that there is still room for improvement as the LSTMs struggle to identify the presence of sudden high peaks as past information weights on the predictions. This situation can be mitigated by capping pollutant observation levels to thresholds over which it implies risk for human health.

References

Abraham, G., Byrnes, G., Bain, C., 2009. Short-term forecasting of emergency inpatient flow. Inf. Technol. Biomed. 13, 380–388.

Andersen, T.B., 1991. A model to predict the beginning of the pollen season. Grana 30, 269–275.

Bergmeir, C., Hyndman, R., Koo, B., 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Comput. Stat. Data Anal. 120, 70–83. https://doi.org/10.1016/j.csda.2017.11.003.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Cannell, M., Smith, R., 1983. Thermal time, chill days and prediction of budburst in *Picea sitchensis*. J. Appl. Ecol. 20, 269–275.

Castellano-Méndez, M., Aira, M.J., Iglesias, I., Jato, V., González-Manteiga, W., 2005. Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air. Int. J. Biometeorol. 49, 310–316.

Catalano, M., Galatioto, F., Bell, M., Namdeo, A., Bergantino, A.S., 2016. Improving the prediction of air pollution peak episodes generated by urban transport networks. Environ. Sci. Pol. 60, 69–83.

Chaloulakou, A., Saisana, M., Spyrellis, N., 1998. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. Sci. Total Environ. 313, 1–13.

Chelani, A., Rao, C., Phadke, K., Hasan, M., 2002. Prediction of sulphur dioxide concentration using artificial neural networks. Environ. Model. Softw. 17, 161–168.

Conover, W.J., 1999. Nonparametric methods. In: Wiley, B., O'Sullivan, M. (Eds.), Practical Nonparametric Statistics. John Wiley and Sons, pp. 233–305. https://doi.org/10.1002/bimj.19730150311.

D'Amato, G., Rottem, M., Dahl, R., Blaiss, M., Ridolo, E., Cecchi, L., Rosario, N., Motala, C., Ansotegui, I., Annesi-Maesano, I., 2011. For the WAO Special Committee on Climate Change, Allergy, climate change, migration, and allergic respiratory diseases: an update for the allergist. World Allergy Organ. J. 4, 120–125.

Díaz, J., Linares, C., Tobías, A., 2007. Short term effects of pollen species on hospital admissions in the city of Madrid in terms of specific causes and age. Aerobiologia 23, 231–238.

Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. 32, 674–701.

Galán Soldevilla, C., Cariñanos González, P., Alcázar Teno, P., Domínguez Vílches, E., 2007. Manual de Calidad y Gestión de la Red Española de Aerobiología. Universidad de CÃ³rdoba.

Gardner, M., Dorling, S., 1998. Artificial neural networks (the multilayer perceptron) a review of applications in the atmospheric sciences. Atmos. Environ. 32, 2627–2636.

Gers, F., Schmidhuber, J., Cummins, F., 2000. Learning to forget: continual prediction with LSTM. Neural Comput. 12, 2451–2471.

González, S., Díaz, J., Pajares, M., Alberdi, J., López, C., Otero, A., 2001. Relationship between atmospheric pressure and mortality in the Madrid autonumus region : a time series study. Int. J. Biometeorol. 45, 34–40.

Grivas, G., Chaloulakou, A., 2006. Artificial neural network models for prediction of pm10 hourly concentrations, in the greater area of Athens, Greece. Atmos. Environ. 40, 1216–1229.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

Iglesias-Otero, M.A., Fernández-González, M., Rodríguez-Caride, D., Astray, G., Mejuto, J.C., Rodríguez-Rajo, F.J., 2015. A model to forecast the risk periods of Plantago pollen allergy by using ANN methodology. Aerobiologia 31, 201–211.

Kingma, D.P., Ba, J., 2019. Adam: a method for stochastic optimization. In: CoRR abs/1412.6980, URL. http://arxiv.org/abs/1412.6980.

Linares, C., Díaz, J., 2008. Impact of high temperatures on hospital admissions: comparative analysis with previous studies about mortality (Madrid). Eur. J. Pub. Health 18, 318–322.

Navares, R., Aznarte, J., 2016. What are the most important variables for poaceae airborne pollen forecasting? Sci. Total Environ. 579, 1161–1169.

Navares, R., Aznarte, J., 2016. Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features. Int. J. Biometeorol. https://doi.org/10.1007/s00484-016-1242-8.

Navares, R., Díaz, J., Linares, C., Aznarte, J., 2018. Comparing Arima and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid. Stoch. Env. Res. Risk A. 1–11. https://doi.org/10.1007/s00477-018-1519-z.

Ozkaynak, H., Qualters, B.G. nd J.R., Strosnider, H., McGeehin, M., Zenick, H., 2009. Summary and findings of the epa and cdc symposium on air pollution exposure and health. J. Expo. Sci. Environ. Epidemiol. 19, 19–29.

Querol, X., Viana, M., Moreno, T., Alastuey, A., 2012. Bases científico-técnicas para un plan nacional de mejora de la calidad del aire. Informes CSIC 1, 19–25.

Ruder, S., 2016. An overview of gradient descent optimization algorithms. In: arXiv (Ed.), CoRR abs/1609.04747. arXiv URL. http://arxiv.org/abs/1609.04747.

Rumelhart, D.E., Hinton, G.E., Ronald, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.

Sabariego, S., Cuesta, P., Fernández-González, F., Pérez-Badia, R., 2012. Models for forecasting airborne cupressaceae pollen levels in Central Spain. Int. J. Biometeorol. 56, 253–258.

Schaber, J., Badeck, F.-W., 2003. Physiology-based phenology models for forest tree species in Germany. Int. J. Biometeorol. 47, 193–201.

Shaffer, J., 1986. Modified sequentially rejective multiple test procedures. J. Am. Stat. Assoc. 81, 826–831.

Sharma, S., Sharma, P., Khare, M., Kwatra, S., 2016. Statistical behavior of ozone in urban environment. Sustain. Eviron. Res. 142–148. https://doi.org/10.1016/j.serj.2016.04.006.

Silva-Palacios, I., Fernández-Rodríguez, S., Durán-Barroso, P., Tormo-Molina, R., Maya-Manzano, J., Gonzalo-Garijo, A., 2016. Temporal modelling and forecasting of the airborne pollen of cupressaceae on the southwestern Iberian peninsula. Int. J. Biometeorol. 60, 1509–1517.

Smith, M., Emberlin, J., 2006. A 30-day-ahead forecast model for grass pollen in North London, UK. Int. J. Biometeorol. 50, 233–242.

Subiza, J., Jerez, M., Jiménez, J., Narganes, M., Cabrera, M., Varela, S., Subiza, E., 1995. Allergenic pollen pollinosis in madrid. J. Allergy Clin. Immunol. 96, 15–23.