Manuscript number: gmd-2019-180

Xueling Wu *, Ying Wang, Siyuan He, Zhongfang Wu: "$PM_{2.5}$ / PM_{10} Ratio Prediction Based on a Long Short-term Memory Neural Network in Wuhan, China".

Dear editor and reviewer,

We would like to thank you for the positive and constructive comments concerning our revised manuscript (ID: gmd-2019-180). These comments are all valuable and very helpful for revising and improving our paper. We have studied the comments carefully. Based on the comments, we have made corrections which we hope will meet with your approval. In the revised manuscript, all the corrections are marked in red. The responses to the reviewers' comments are as follows (in blue font):

Responses to the Referee comment 2:

Point-by-point responses to the comments:

1. Comment: *In particular, the authors should clarify the way they split their data. They state on line 271, "80% of the data were used as the training sample for modelling, and 20% of the data were used as the verification sample." It is important to specify how the hyper-parameters of their model were chosen. If they were chosen by optimising the performance against the verification dataset then it is possible that the algorithm has over-fit the hyper-parameters.*

Lines 272-274 seem to imply there was at least some hyper-parameter tuning performed. Ideally, the data should be split three ways, into a training, verification, and test, so that the hyper-parameters are tuned against the verification data, and the algorithm

scored against the test data. The authors should reassure the reader that they have taken measures to ensure they have not overfit the hyper-parameters - for instance, perhaps they further split their training set.

In addition, they state on line 279 that for Section 5.1 that there were 570 samples in the training data, and what I infer is 20 samples in the verification data (two days multiplied by ten sites, as per the training data). This appears not to be an 80%/20% split. In any case, the verification data are from one period in the season (end of December)-the algorithm may simply be good at predicting air quality in December but not the rest of the year. A more convincing approach would be to test against multiple cases from throughout the year. This is similar for the spatial prediction, which appears to only be tested at one site.

As such, the authors must reassure us that their approach to validation guards against overfitting in order for this to be suitable for publication.

Response: Thank you for pointing this out. After considering your suggestions, we readjusted our model and code. Except for the data used for prediction, we divided the data set involved in the model construction into three parts: 40% of the data were used as the training samples for modeling, 30% of the data were used as the test samples, and the remaining 30% of the data was used as verification data. We tried to add a regularization term, but the effect did not improve. After adjusting the number of neurons, the number of epochs, and the batch size, the loss function we obtained has converged without overfitting.

Moreover, the revised model obtained higher prediction accuracy than the original one. We have generated the following learning curves for three prediction patterns, but in order to avoid the manuscript being too verbose, we do not intend to add the learning curve to the manuscript. The three curves in this figure are the losses of training samples, test samples and verification samples (train, test, and verifi) that increase with the number of epochs. We understand that our spatial and time prediction patterns do not completely cover the whole year and all regions, so we have added a random prediction pattern which randomly selects data from the whole year and the entire region for prediction to reduce fortuity of the other two prediction patterns. The correction is as follows (line 304-313):

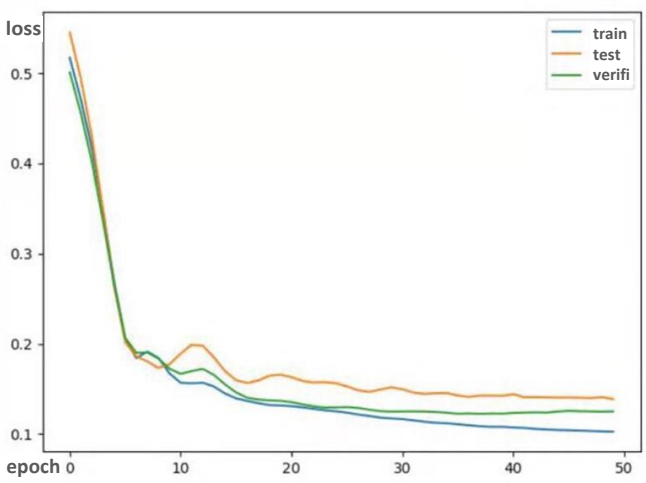

Figure 1 Learning curves

To determine the appropriate number of layers for the LSTM method, except for the data used for prediction, we divided the data set involved in the model construction into three parts: 40% of the data were used as the training samples for modeling, 30% of the data were used as the test samples, and the remaining

30% of the data was used as verification data. We tried to use various LSTM architecture layers for the comparison. After obtaining the results of various LSTM architecture layers, we found that the results obtained using the LSTM architecture with four layers were the best, with the first three layers and the dense layer as the last layer. The role of the dense layer is to complete the final output of unique values. Because the LSTM uses the activation function as the gate, the outputs of the gates must be between 0 and 1, and the output ranges of both types of activation functions must be satisfied. We determined that the activation function for setting the forget gate and the input gate was defined as a sigmoid function. After adjusting the number of neurons, the number of epochs, and the batch size, the loss function we obtained has converged without overfitting.

2. Comment*: Could you discuss why you have focused on the PM2.5/PM10 ratio as opposed to considering them separately?*

Response: Thank you for pointing this out. The explanation of using the $PM_{2.5}$-PM_{10} scale is as follows (line 43-47):

Since fine and coarse particles come from different sources, the $PM_{2.5}$-PM_{10} scale model has different physicochemical properties, which can not only distinguish the type of aerosol in the PM but also provide the mixing ratio of dust and artificial aerosols (Sugimoto et al., 2015). The $PM_{2.5}$-PM_{10} scale is the main indicator for macro analysis of the source of particulate pollution in

a region, which is more practical than considering $PM_{2.5}$ and PM_{10} separately.

***3. Comment**: line 75: "random precision", and Section 5.3 "random pattern prediction". Please could you clarify what this is - it was unclear to me. Are you randomly selecting a subset of points in space and then predicting them with the remaining, contemporaneous points? If so, how is this significantly different from the spatial prediction in Section 5.2? Please better explain this task near the beginning of the manuscript.*

Response: Thank you for pointing this out. We have added the explanation of these three prediction patterns to the Section 1-Introduction (line 80-84):

The time precision mentioned in this article refers to the accuracy of inputting time-series data to predict the subsequent period results; the spatial precision refers to the accuracy of inputting all-time data of spatial points to predict the result of another spatial point; the random accuracy refers to the accuracy of inputting data of any time and space to predict the random selection data.

***4. Comment**: line 112,113: I found this sentence confusing. Are "monitoring station" and "monitoring site" different things? I'm unclear what the definition is on an "inspection standard"? Does this mean the "truth" data you are using for the verification. I'm unclear what "correlation factors" means.*

Response: Thank you for your thoughtful insights. "Monitoring station" and "monitoring site" have the same meaning, and we replaced "monitoring site" with "monitoring station". "Inspection standard" means truly data and

"Correlation factors" refer to $PM_{2.5}$, PM_{10}, and gaseous pollutant data detected by the stations. We have redefined these meanings in the article (line 126-127):

Therefore, we used the truly air quality data from the ground monitoring stations as the inspection standard for verification and extracted the values of $PM_{2.5}$, PM_{10}, and gaseous pollutant with the data from the monitoring stations.

5. Comment*: line 132: You verify your data processing against NASA data. Is NASA has a product, why not just use that?*

Response: Thank you for your thoughtful insights. The commonly used aerosol automatic observation network AERONET jointly established by NASA and CNRS is of good quality and easy to obtain, but the number of stations is limited, and there is no station coverage in the study area. The remote sensing data we collected are better in time and space continuity, and the AOD retrieval algorithm is also applicable to the study area.

6. Comment*: line 175: "higher trend" and "lower trend". The word "trend" changes the meaning. I presume this should read "average temperature is higher in summer and lower in winter", as expected. Otherwise, I don't understand what it means.*

Response: Thank you for your thoughtful insights. We have modified the expression of this sentence (line 197-198):

The average surface temperature and average temperature were higher in summer and lower in winter.

***7. Comment**: line 185: Is this standard published. If so, please cite. In my opinion, this can be a technical paper as opposed to a peer review paper (but the editor may feel differently).*

Response: Thank you for your thoughtful insights. This standard has been published, so we added a citation as follows (line 206-208):

The national standard method is performed according to the method specified in the Chinese national standard GB/T13201-91 (http://www.mee.gov.cn/gzfw_13107/kjbz/qthjbhbz/qt/201605/t20160522_342349.shtml).

***8. Comment**: line 230: Could you clarify the optical subset approach? Was the R^2 score performed on the output of the LSTM as compared with the observations. If so, it may be better to put this section after the description of the LSTM and make that clear.*

Response: Thank you for your thoughtful insights. We clarified the optical subset approach. In addition, we did not perform R^2 scoring on the output of the LSTM, because the relative error rate can also reflect the accuracy intuitively. Since the maximum relative error and minimum relative error needs to be analyzed at the same time, it is more neatly to display the three relative error rates in a table. The explanation of optimal subset method is as follows (line 249-252):

The process of the optimal subset method is that in a set containing multiple independent variables, freely selecting and combining from each independent

variable, combining all independent variables and dependent variables to establish all possible equations, and then the best independent variable combination model is selected from all the fitted regression equations.

9. Comment*: Table 4: I presume this table shows only the top 10 scoring selections? Presumably you scored all combinations of predictors. Please explain.*

Response: Thank you for your thoughtful insights. I added an explanation for table 4 as follows (line 257-258):

This table shows the top 10 scores for R^2 scores and the corresponding factor combinations.

10. Comment*: Line 258: Normally the first gate is expressed at deciding what to forget, rather than what to remember (I appreciate they are equivalent). Figure 7 shows a "Forget gate" so it would be helpful to standardise the terminology.*

Response: Thank you for your thoughtful insights. The correction is as follows (line 286-288):

The input gate determines how much of the input x_t of the network is saved to the cell state c_t at the current moment, the forget gate determines how much the cell state c_{t-1} at the previous moment is retained to the current moment c_t, and the output gate controls how much the cell state c_t is output to the current output value h_t of the LSTM.

11. Comment*: "and the third switch controls whether c is the current output of the*

LSTM model" I'm not sure about this - correct me if I'm wrong, but isn't c_t combined with h_t-1, and x_t to create h_t, which is the output?

Response: Thank you for your thoughtful insights. It is true that h_t is created by x_t with the combination of c_t and h_{t-1}. The correction is as follows (line 282-288):

Fig. 6 emphasizes the calculation process of the cell state c, and the overall process of the LSTM model is shown in Fig. 7. The key point of the LSTM model is how to control the state c. The idea of the LSTM model is to use three control switches to control it. The switches implemented in the algorithm are known as "gates", which are fully connected layers whose input is a vector, and the output is a real vector between 0 and 1 (Srivastava and Lessmann, 2018). The input gate determines how much of the input x_t of the network is saved to the cell state c_t at the current moment, the forget gate determines how much the cell state c_{t-1} at the previous moment is retained to the current moment c_t, and the output gate controls how much the cell state c_t is output to the current output value h_t of the LSTM.

***12. Comment**: Section 4: I see the link to your code, that you used Keras and their LSTM implementation, which is great. If possible, could you cite Keras directly in the manuscript, and state that you used their implementation of LSTMs.*

Response: Thank you for your thoughtful insights. I cited Keras in the manuscript as follows (line 301-302):

The implementation of the LSTM models is based on Keras which is a high-level neural network Application Programming Interface written in Python.

13. Comment*: line 273: "with the first three layers being the LSTM layer and the last layer being the dense layer". I presume this means that you have three LSTM units follows by a dense layer. However, each LSTM unit can be thought of comprising multiple layers, so this terminology is confusing. In addition, could you explain to the reader the purpose of the final dense layer.*

Response: Thank you for your thoughtful insights. This sentence does means that we have three LSTM units follows by a dense layer. To avoid misunderstandings, we adjusted the description and explained the purpose of the dense layer as follows (line 307-310):

We tried to use various LSTM architecture layers for the comparison. After obtaining the results of various LSTM architecture layers, we found that the results obtained using the LSTM architecture with four layers were the best, with the first three layers and the dense layer as the last layer. The role of the dense layer is to complete the final output of unique values.

14. Comment*: line 282: I don't understand the difference between a "multilayer perceptron" and an "artificial neural network". In addition, I would have thought the LSTM, multilayer perceptron and artificial neural network, all rely on back propagation, so I don't understand the terminology "back propagation artificial neural network". Please clarify.*

Response: Thank you for your thoughtful insights. The reason why we discriminate between MLP and BP is that MLP is a back propagation neural network model with a three-layer architecture after adjusted in Python by us, while the BP neural network acquired by the clementine software with non-adjustable parameters. After careful consideration, we decided to delete the comparison with MLP to avoid ambiguity.

15. Comment*: Section 5.3: as mentioned above, I don't understand what is being done in this section.*

Response: Thank you for your thoughtful insights. The output of the random prediction pattern in Section 5.3 is randomly selected data in any time and space. This pattern is different from the time and space pattern. The applicability of LSTM reflected by the random prediction pattern is more obvious than the other two patterns. The explanation for section 5.3 is in the manuscript as follows (line 358-362):

The random pattern prediction was based on the completely random selection of time and space aspects and can reflect the effect of air quality prediction under various climatic conditions well. The superiority of the LSTM model prediction in the random prediction pattern was more obvious than in the other patterns, which indicates that under irregular conditions, the LSTM model is more suitable for making predictions.

16. Comment*: line 24: "Aerosols are a general term" -> "Aerosol is a general term".*

Response: Thank you for your thoughtful insights. We apologize for our mistakes. The correction is as follows (line 25):

Aerosol is a general term for solid and gas particles suspended in air.

17. Comment*: Please define all your acronyms on first use, for instance RH, DVV etc.*

Response: Thank you for your thoughtful insights. The correction is as follows (line 12-15):

First, the aerosol optical depth (AOD) in 2017 in Wuhan was obtained based on Moderate Resolution Imaging Spectroradiometer (MODIS) images, with a 1 km spatial resolution, by using the Dense Dark Vegetation (DDV) method. Second, the AOD was corrected by calculating the planetary boundary layer height (PBLH) and relative humidity (RH).

18. Comment*: line 74: "intelligent models", please clarify what this means.*

Response: Thank you for your thoughtful insights. We replaced "intelligent models" with "machine learning models". The correction is as follows (line 79-80):

However, with the introduction of new machine learning models, the traditional regression model reflects the inability to balance time, space and random precision.

19.Comment*: line 89: "classic" -> "classical".*

Response: Thank you for your thoughtful insights. The correction is as

follows (line 99-101):

Finally, the space and time scales and random $PM_{2.5}$/PM_{10} predictions were determined and performed, respectively, via the LSTM model, and the prediction results of the LSTM model and other classical models were compared and analyzed.

20. Comment*: line 108: -> "Our environmental monitoring station only monitors data in real-time". I'm also not sure what this sentence is meant to mean. Are you saying that the instrument doesn't provide information about the future? Please clarify.*

Response: Thank you for your thoughtful insights. The supplement is as follows (line 122-123):

The data that our environmental monitoring station can monitor is only real-time data with no predicting subsequent data in advance.

21. Comment*: line 110: "which have", is ambiguous - is this referring to the AOD or the atmospheric aerosols? It seems redundant to say that PM is linked to atmospheric aerosols. Please clarify this sentence.*

Response: Thank you for your thoughtful insights. "Which have" refers to AOD. The correction is as follows (line 123-124):

The AOD which has a great relationship with PM is an important parameter in the study of atmospheric aerosols.

22. Comment*: line 140: "The shortest distance between points exceeds 3 km, and the*

average distance exceeds 10 km." I don't know what the word "exceeds" means in this context. Can we replace it with "is" instead?

Response: Thank you for your thoughtful insights. The correction is as follows (line 155):

The shortest distance between points is more than 3 km, and the average distance is about 10 km.

A special thanks to you for your insightful and valuable comments.

We greatly appreciate the editors and reviewers for their helpful work and hope that the corrections will meet your approval. Once again, thank you very much for your valuable and helpful comments and suggestions.

With best wishes

Yours sincerely,

Xueling Wu

Institute of Geophysics and Geomatics, China University of Geosciences

No. 388 Lumo Road, Wuhan 430074, P. R. China

Email: snowforesting@163.com

The corrected tables are as follows:

Table 5 The results and relative error rates of the time pattern predictions.

Measured value	Predicted value				Relative error rate (%)			
	LSTM	BP	SVM	CHAID	LSTM	BP	SVM	CHAID
0.8212	0.6335	0.7786	0.6698	0.4853	22.8604	5.1875	18.4364	40.9036
0.7436	0.5610	0.6961	0.7841	0.4853	24.5491	6.3878	5.4465	34.7364
0.6629	0.7346	0.7074	0.8353	0.6753	10.8069	6.7129	26.0069	1.8706
0.6950	0.7949	0.6850	0.5628	0.6753	14.3746	1.4388	19.0216	2.8345
0.7816	0.7347	0.6871	0.8092	0.5145	5.9982	12.0906	3.5312	34.1735
0.6311	0.7605	0.5864	0.7032	0.6487	20.5089	7.0829	11.4245	2.7888
0.7959	0.7347	0.6870	0.8568	0.6973	7.6931	13.6826	7.6517	12.3885
0.8743	0.8067	0.6474	0.7451	0.6973	7.7307	25.9522	14.7775	20.2448
0.7204	0.6553	0.8106	0.7446	0.8206	9.0291	12.5208	3.3592	13.9089
0.9854	0.7128	0.7154	0.6760	0.8206	27.6610	27.4000	31.3984	16.7242
0.7079	0.7249	0.8321	0.6089	0.7959	2.4048	17.5449	13.9850	12.4311
0.9455	0.7790	0.7064	0.7285	0.7959	17.6108	25.2882	22.9508	15.8223
0.7200	0.4924	0.6692	0.8172	0.6931	31.6131	7.0556	13.5000	3.7361
0.8600	0.6521	0.6192	0.6907	0.6931	24.1694	28.0000	19.6860	19.4070
0.6571	0.6432	0.7011	0.8522	0.5812	2.1242	6.6961	29.6911	11.5508
0.9189	0.7175	0.6195	0.7146	0.5812	21.9150	32.5824	22.2331	36.7505
0.7640	0.7673	0.6549	0.5406	0.7870	0.4291	14.2801	29.2408	3.0105
0.9273	0.7896	0.6354	0.7155	0.7870	14.8513	31.4785	22.8405	15.1299
0.6277	0.4614	0.7308	0.5392	0.6951	26.4993	16.4250	14.0991	10.7376
0.8896	0.6904	0.6685	0.6694	0.7534	22.3909	24.8539	24.7527	15.3103
Mean:					15.7613	16.1330	17.7017	16.2230
Maximum:					31.6111	32.5824	31.3984	40.9036
Minimum:					0.4319	1.4388	3.3592	1.8706

Table 6 The results and relative error rates of the spatial pattern prediction.

Models	LSTM	BP	SVM	CHAID
Mean:	27.9231	34.1333	34.0207	33.7718
Maximum:	178.0639	222.9295	204.7317	230.1367
Minimum:	0.0764	0.1124	0.9026	0.2396

Table 7 The results and relative error rates of the random pattern prediction.

Measured value	Predicted value				Relative error rate (%)			
	LSTM	BP	SVM	CHAID	LSTM	BP	SVM	CHAID
0.5870	0.6031	0.5762	0.6091	0.4928	2.7428	1.8399	3.7649	16.0477
0.6213	0.6581	0.6561	0.6826	0.6795	5.9231	5.6012	9.8664	9.3675
0.9843	0.4662	0.6247	0.6185	0.7422	52.6364	36.5336	37.1635	24.5962
0.8000	0.4198	0.4772	0.5231	0.4928	47.5250	40.3500	34.6125	38.4000
0.4638	0.4654	0.4773	0.5136	0.4928	0.3450	2.9107	10.7374	6.2527
0.7010	0.5762	0.6811	0.6675	0.6795	17.8031	2.8388	4.7789	3.0670
0.2222	0.2470	0.4292	0.3971	0.3737	11.1611	93.1593	78.7129	68.1818
0.5929	0.6418	0.6543	0.6598	0.6795	8.2476	10.3559	11.2835	14.6062
0.9571	0.5875	0.6246	0.6698	0.6164	38.6167	34.7404	30.0178	35.5971
0.7576	0.7095	0.5959	0.6398	0.4928	6.3490	21.3437	15.5491	34.9525
0.6277	0.6368	0.6935	0.6802	0.6795	1.4497	10.4827	8.3639	8.2523
0.8896	0.6508	0.7551	0.7353	0.7422	26.8435	15.1192	17.3449	16.5692
Mean:					18.3036	22.9396	21.8496	22.9909
Maximum:					52.6364	93.1593	78.7129	68.1818
Minimum:					0.3450	1.8399	3.7649	3.0670