Research on a Prediction Method for Carbon Dioxide Concentration Based on an Optimized LSTM Network of Spatio-Temporal Data Fusion

Jun MENG†, Member, Gangyi DING†, and Laiyang LIU†, Nonmembers

SUMMARY In view of the different spatial and temporal resolutions of observed multi-source heterogeneous carbon dioxide data and the uncertain quality of observations, a data fusion prediction model for observed multi-scale carbon dioxide concentration data is studied. First, a wireless carbon sensor network is created, the gross error data in the original dataset are eliminated, and remaining valid data are combined with kriging method to generate a series of continuous surfaces for expressing specific features and providing unified spatio-temporally normalized data for subsequent prediction models. Then, the long short-term memory network is used to process these continuous time- and space-normalized data to obtain the carbon dioxide concentration prediction model at any scales. Finally, the experimental results illustrate that the proposed method with spatio-temporal features is more accurate than the single sensor monitoring method without spatio-temporal features.

key words: carbon emissions, wireless carbon sensor network, optimized LSTM network, multi-source data fusion

1. Introduction

In recent years, people have paid increasing attention to the issue of global warming, and the relationship between global warming and the increase in greenhouse gas emissions is no longer in doubt. Carbon dioxide (CO$_2$) is one of the most important greenhouse gases in the atmosphere, so monitoring the CO$_2$ content in the environment and scientifically analysing the data are of great significance to the prediction of future climate change trends. In January 2009, Japan launched the GOSAT (Greenhouse gases Observing SATellite), the world’s first satellite dedicated to monitoring the distribution of greenhouse gas concentrations from space [1]. In July 2014, NASA launched the Orbiting Carbon Observatory-2 (OCO-2) satellite [2]. In addition, there are ground-based TCCON (Total Carbon Column Observing Network) under the NDACC-IRWG and GAW plans [3]. These monitoring methods provide strong data support for the data-driven environmental carbon data prediction method.

At present, for different environments, the majority of CO$_2$ data prediction methods introduced by studies have been statistical theory methods, but in the face of multi-source and heterogeneous, multi-modal data of space and time, traditional data fusion methods possess certain shortcomings [4]. At the same time, we have also noticed that some scholars have applied mobile sensing system [5], big data [6], deep distributed fusion network [7] and other technologies to air quality prediction and achieved good results. However, the data sources of these studies are mainly based on ground-based sensor monitoring network, without considering the relevant satellite monitoring data. Among the data provided by multi-source sensors, the data collected by ground sensors are continuous in time at discrete locations in the spatial distribution, but due to the limited number of installation locations, the spatial continuity of the data on a global scale cannot be guaranteed. In the coordinate system of Earth, the position of a carbon satellite is constantly changing with time, so the data provided by it cannot guarantee the continuity of both space and time at a specific location. Therefore, the scientific data fusion of these multi-source data and the design of more accurate prediction methods based on data-driven ideas is of great significance to the prediction of future climate change trends and now-casting.

2. Proposed Method

2.1 Model Framework

Aimed at solving the problems of different spatial and temporal resolutions and the uncertain observation quality of multi-source heterogeneous CO$_2$ data, a prediction model of CO$_2$ concentration is established. The framework of the model is shown in Fig. 1, which mainly includes the following contents:

- Data sources: Heterogeneous data sets based on time and space distributions from multi-source sensors are obtained. These mainly include the observation data provided by the OCO-2 and GOSAT satellite sensors data products, the observation data provided by the TCCON observation network data products, and the related data collected by the self-built wireless carbon sensor network.
- Data fusion: Firstly, according to the Grubbs criterion, the gross error data in the original dataset are eliminated [8]. Secondly, the multi-source data based on temporal and spatial distributions are fused, and a se-
ries of continuous surfaces are created through the kriging method to fill in the missing values in the data of discontinuous space-time series [9]. Lastly, the surface data are normalized to improve the convergent rate of the prediction model.

- Model training and evaluation: The continuous surface data based on time series are divided into two groups according to the selected spatial-temporal scale. One group is used as train dataset to train the optimized LSTM network [10], and the other group is used as test dataset to verify the prediction model; the relevant evaluation indicators are used to evaluate the prediction effect.

2.2 Data Processing Algorithm

The multi-source carbon data (CO2 concentration, temperature, humidity and so on) are processed according to the Grubbs criterion: set the sample as $s_i \ (i = 1, 2, 3, \cdots, n)$, where $n$ is the total number of samples. Then, for the $i$-th sample value:

- Arrange the $s_i$ in ascending order.
- Calculate the mean $\bar{s}$ and variance $\sigma$.
  \[
  \bar{s} = \frac{\sum_{i=1}^{n} s_i}{n} \tag{1}
  \]
  \[
  \sigma = \sqrt{\frac{\sum_{i=1}^{n} (s_i - \bar{s})^2}{n-1}} \tag{2}
  \]
- Calculate the lower Grubbs number $g_{(1)}$ and the upper Grubbs number $g_{(n)}$.
  \[
  g_{(1)} = \frac{\bar{s} - s_{(1)}}{\sigma} \tag{3}
  \]
  \[
  g_{(n)} = \frac{s_{(n)} - \bar{s}}{\sigma} \tag{4}
  \]

- Select the larger residual error of the two and select the significance $\alpha$ to obtain the critical value $g_0(n, \alpha)$ through the critical value test table. If $g(i) \geq g_0(n, \alpha)$, there is gross error in the measured value, which should be eliminated.

The deep neural network used in the experiment is mainly composed of an LSTM layer and a fully connected layer, and the test data are obtained from the continuous data surface generated [11]. The location data with a latitude and longitude of (36.604 N, 97.486 W) from September 2014 to August 2015 are selected as the experimental data by the model, which is the location of Lamont, the TCCON observation site. Among them, the data from September 2014 to April 2015 are used as train set and validation set to train the optimized LSTM network, and the data from May 2015 to August 2015 are used as test set. The structure diagram of the calculation unit is shown in Fig. 2.

$\hat{z}(s)$ is obtained by interpolation of relevant data using ordinary Kriging method, after screening, it is used as the input variable $x_t$ of each gate. The sub-structure diagram of each gate is shown in Fig. 3.

Its core processing flow can be expressed as follows:

- Calculate the CO2 concentration of the missing points:
  \[
  \hat{z}(s) = \sum_{i=1}^{n} \lambda_i \hat{z}(s_i) \tag{5}
  \]
- Calculate the weight coefficient:
  \[
  \begin{cases}
  \sum_{i=1}^{n} \lambda_i C(s_i - s_j) + \mu = C(s_j - s) \\
  \sum_{i=1}^{n} \lambda_i = 1
  \end{cases}
  \tag{6}
  \]
In this process, $\hat{z}(s)$ is the concentration of $\text{CO}_2$ at the point to be measured, $s_i$ is the weight corresponding to each $\text{CO}_2$ observation value, and $z(s_i)$ is the $\text{CO}_2$ observation value at each sample point. To calculate the weight coefficient, two conditions of unbiasedness and optimality should be met. $W_f$, $U_f$, $U_i$, $W_o$, $U_o$, $W_c$, $U_c$ and $V_o$ are coefficients of the linear relationships, $b_f$, $b_i$, $b_c$ and $b_o$ are bias vectors, $\sigma$ is the sigmoid activation function, and $\otimes$ is the Hadamard product. The details of model training procedure are shown in Table 1.

### Table 1: Model training procedure.

| Input: | A set of features $X = \{X_{or}, X_c, X_h, X_t\}$, input_length = 360, input_dim = 4, learning rate = 0.001, the number of hidden layers is 2, the number of hidden units is 30 and the temporal granularity is day. |
| Output: | Learned prediction model. |
| 1. | The array constructor is used to convert the dataset $X$ and get the matrix $X_c$. |
| 2. | Data normalization. |
| 3. | Divide and create data sets, 67% for the train dataset, and 33% for the test dataset. |
| 4. | look_back = 1 |
| 5. | Rescale input to be [samples, time_steps = 5, features] |
| 6. | Create and fit the LSTM network. |
| 7. | loss = 'mean_squared_error' |
| 8. | The Adam optimizer is used, |
| 9. | epochs = 25 |
| 10. | batch_size=30 |
| 11. | Make predictions |
| 12. | Reverse normalization the prediction results. |
| 13. | Calculate the mean squared error of the train dataset and test dataset. |
| Return | Epoch: 25 (of 25), loss: 0.0001457373466342688 |

### 3. Experimental Results

Based on the needs of this subject, the research group designed a wireless carbon sensor device. It mainly integrates ARM control module, carbon dioxide data acquisition module, temperature data acquisition module, humidity data acquisition module, GPS positioning module and GPRS data transmission modules. The prototype equipment is shown in Fig. 4.

Since August 2010, the research group has successively set up 14 environmental monitoring sites in Genhe City of Hulunbeier City, Inner Mongolia Autonomous Region, and 5 monitoring sites have been set up in the Xincheng area of Huhehaote City, Inner Mongolia Autonomous Region, since April 2014. To obtain more complete data sets, this number will be further increased in the future. The distribution of the sites is shown in Fig. 5, and the experimental data information of the Wireless Carbon Sensor Network (WCSN) is shown in Table 2.

After calculating the weight coefficients, the variogram is fitted with empirical data to minimize the weighted square deviation between each point and the function curve. An interpolation model that quantifies the spatial correlation curve. An interpolation model that quantifies the spatial correlation of the Wireless Carbon Sensor Network (WCSN) is shown in Table 2.
Table 2  Details of the datasets.

| Sources       | OCO-2 | GOSAT | WCSN  |
|---------------|-------|-------|-------|
| Number of Points | 21258 | 10867 | 52560 |
| Length of Data Sequence | 85032 | 43468 | 210240 |
| Ratio of Missing Values (%) | 62.5  | 80.83 | 90.21 |

| Lamont       | Processed Data | TCCON |
|--------------|----------------|-------|
| Number of Points | 360  | 54938 |
| Length of Data Sequence | 1440 | 219752 |
| Ratio of Missing Values (%) | 0    | 0.06  |

Fig. 6  Semi-variant function generation model.

Fig. 7  Continuous surface generation map of CO2 concentration.

Fig. 8  Comparison of CO2 data between TCCON and OCO-2.

Fig. 9  Comparison of CO2 data between TCCON and GOSAT.

Fig. 10  Comparison of CO2 data between TCCON and prediction model.

Fig. 8 through Fig. 10.

By comparing the line graphs, it can be seen intuitively that the OCO-2 data and TCCON data exhibit a certain deviation, and the change trend of the data is quite different; there is a large deviation between the GOSAT data and the TCCON data, but the data change trend is basically the same. It can be seen that the effects of the two measurement methods are not very satisfactory. The prediction trend of the multi-source data fusion prediction model is basically consistent with that of the TCCON data, and the prediction results are closer. To quantitatively evaluate the prediction results of the model, two evaluation indicators, Root Mean
Table 3  Statistical error analysis results of TCCON and the other three data sources.

| Data Sources   | RMSE  | MAPE (%) |
|----------------|-------|----------|
| OCO-2          | 1.390 | 23.5     |
| GOSAT          | 1.971 | 40.6     |
| Prediction-Model | 0.776 | 16.5     |

Square Error (RMSE) and Mean Absolute Percentage Error (MAPE), are used for statistical comparison and analysis of the data provided by each data source. The statistical results are shown in Table 3.

It can be seen from the statistical results that the deep learning prediction model based on multi-source data fusion achieves higher data accuracy than the two satellite products and can fit the CO2 concentration trend well. This also shows that the model fully incorporates the unique advantages of the optimized LSTM network. It can extract abstract and deep-seated features from high-dimensional data through a variety of nonlinear operations, which causes the prediction model to show better prediction and fitting effects.

4. Conclusion

In this letter, a data fusion prediction method for CO2 concentration is proposed, and the experimental results confirmed that the variation trend of CO2 concentration is related to environmental factors such as temperature and air humidity. These data exhibit a natural continuity in time and achieve strong correlation and causality before and after the time series. The multi-source data fusion deep learning prediction method with spatio-temporal features is more accurate than the single sensor monitoring method without spatio-temporal features. Using this method to achieve CO2 concentration prediction can not only use the correlation of the data in the time dimension but also automatically mine the potential correlations between the data and improve the accuracy of carbon emission data prediction.

References

[1] L. Lei, X. Guan, Z. Zeng, B. Zhang, F. Ru, and R. Bu, “A comparison of atmospheric CO2 concentration GOSAT-based observations and model simulations,” Science China Earth Sciences, vol.57, pp.1393–1402, 2014.
[2] L.L. Zhang, M.W. Zhao, N. Zhao, et al., “Modeling the spatial distribution of XCO2 with high accuracy based on OCO-2’s observations,” Journal of Geo-information Science, vol.20, no.9, pp.1316–1326, 2018.
[3] P.O. Wennberg, D. Wunch, C.M. Roehl, J.-F. Blavier, G.C. Toon, and N.T. Allen, “TCCON data from Lamont (US), Release GGG2014.R1 (Version GGG2014.R1) [Data set],” CaltechDATA, 2016. doi.org/10.14291/TCCON.GGG2014.LAMONT01.R1/1255070
[4] F. Han, L. Zhu, and X.J. Zhi, “Measurement of multi-sensor data fusion method Based on fuzzy theory,” Journal of Applied Optics, vol.30, no.6, pp.988–991, 2009.
[5] Y. Jiang, L. Shang, K. Li, L. Tian, R. Piedrahita, X. Yun, O. Mansata, Q. Lv, R.P. Dick, and M. Hannigan, “MAQS: A Personalized Mobile Sensing System for Indoor Air Quality Monitoring,” UbiComp2011, pp.271–280 2011.
[6] Y. Zheng, F. Liu, and H.-P. Hsieh, “U-Air: When Urban Air Quality Inference Meets Big Data,” KDD2013, pp.1436–1444, 2013.
[7] X. Yi, J. Zhang, Z. Wang, T. Li, and Y. Zheng, “Deep Distributed Fusion Network for Air Quality Prediction,” KDD2018, pp.965–973, 2018.
[8] B. Li, ”Application of Grubbs criterion in project evaluation system,” Computer & Telecommunication, vol.6, pp.61–63, 2019.
[9] L. Che, H.Q. Wang, T. Fei, et al., “Kriging interpolation method optimized by multi-scale least squares support vector machine,” Journal of Geo-information Science, vol.19, no.8, pp.1001–1010, 2017.
[10] S.Q. Wang, G. Hua, G.S. Hao, et al., “Long short term memory networks based on grey wolf optimizer in time series forecasting,” China Sciencepaper, vol.20, no.12, pp.2309–2314, 2017.
[11] G.T. Yang, T. Zhang, Y.N. Wang, et al., “Prediction model for NOx emissions from thermal power plants based on long-short term memory neural network,” Thermal Power Generation, vol.10, pp.12–17, 2018.