https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

Deep learning tool: Reconstruction of long missing climate data based on multilayer perceptron(MLP)

Zhang Yan[2]; Xu Tianxin[1]; Zhang Chenjia[1*]; Ma Daokun[1]

[1]College of Information and Electrical Engineering, China Agricultural University Haidian District, Beijing 100089, China

[2]School of Yi Language and Culture, Xichang University, Xichang, Sichaun 615000, China

* Correspondence to: Zhang Chenjia, Email: zcj136578646@outlook.com

Abstract: Long-term monitoring of climate data is significant for grasping the law and development trend of climate change and guaranteeing food security. However, some weather stations lack monitoring data for even decades. In this study, 62 years of historical monitoring data from 105 weather stations in Xinjiang were used for missing sequence prediction, validating proposed data reconstruction tool. First of all, study area was divided into three parts according to the climatic characteristics and geographical locations. A deep learning tool based on multilayer perceptron (MLP) was established to reconstruct meteorological data with three time scales（Short term, cycle and long term）and one spatio dimension as inputing, filling in long sequence blank data. By designing an end-to-end model to autonomously detect

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

the locations of missing data and make rolling predictions,we obtained complete meteorological monitoring data of Xinjiang from 1961 to 2022. Seven kinds of parameter reconstructed include maximum temperature (Max_T), minimum temperature (Min_T), mean temperature (Ave _ T), average water vapor pressure (Ave _ WVP), relative humidity (Ave _ RH), average wind speed (10 m Ave _ WS), and sunshine duration (Sun_H). The quality of reconstructed data was evaluated by calculating correlation coefficient with the monitored sequences of nearest station. Results show that,proposed model reached satisfied average correlation coefficient for Max_T, Min_T, Ave _ T and Ave _ WVP parameters are 0.969, 0.961, 0.971 and 0.942 respectively. The average correlation coefficient of Sun_H and Ave _ RH are 0.720 and 0.789. Although it is difficult to predict extreme values, it can still capture the period and trend; the reconstruction effect of 10 m Ave _ WS is poor, with the average similarity of 0.488. Finally, we published the trained parameter files and prediction codes as a micro service on the Agricultural Smart Brain platform, which provides firstly a deep learning tool for rapid and reliable reconstruction of meteorological monitoring data.

Keywords: Deep learning; Meteorological monitoring; Data reconstruction; MLP

1 Introduction

Agriculture, as the most fundamental industry for human, is facing a serious threat of climate change.Meteorological disasters account for over 70% of the natural disasters globally,have caused serious economic losses (Qin et al. 2002). More severely, global climate is changing dramatically,due to amount of greenhouse gases human activities producted. Because of climate warming, the frequency and intensity of drought and flood are increasing, and the harm to agricultural production is increasingly intensified (IPCC. 2012). From 2010 to

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

2017, global average annual economic losses due to drought reached US $23.125 billion, with annual grain production cuts ranging from millions of tons to more than 30 million tons (Buda et al. 2018).In past 40 years, flooding events caused over a trillion dollars damage as well (UNDRR. 2020).Monitoring and analysis of meteorological data are able to reduce food and economic losses (Ziolkowska & Zubillaga. 2018).

In order to effectively guide agricultural production, meteorological monitoring usually includes meteorological parameters such as temperature, humidity, water vapor pressure, wind speed and sunshine.What is more, conducting research on climate forecasting requires long-term, large-scale and comprehensive climate data (Bonnet et al. 2020). Governments and scientific communities have been committed to the construction of meteorological databases (Anderson et al. 2008), a large number of professional meteorological monitoring stations have been established around the world, including China. However, there is much missing historical data due to temporal differences of monitoring station establishment, sensor failure, and other reasons.Thus, it is crucial to reconstruct the complete meteorological monitoring data.

Usually, researchers use interpolation method combined with manual correction to reconstruct missing meteorological data.Which not only consumes a lot of manpower, but also, due to the spatial variability of geographical conditions, the data results reconstructed by the traditional method are too smooth and inaccurate (Yao et al. 2023).Machine learning is a better interpolation tool (Li et al. 2020),but which performed poorly when deal with long-sequence missing data scenarios. A simple and efficient method for data reconstruction, deep neural network, has a great potential in meteorological data reconstruction tasks. The

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

neurons in the hidden layers of the neural network can constantly update the weights under supervision of true value, learning high-dimensional association among different data, and more accurately complement missing data (Rajaee et al. 2019). In the task of reconstructing monitoring data of turbomechanical particle flow. Deep learning method was more accurate compared with six commonly used interpolation methods (Ghasem&Nader. 2022). In fact, in the field of weather forecasting, some available deep learning models have been published. Training baseed on large amounts of data, FourCastNet2 can calculate the next 24 hours of climate for 100 sites in just 7s (Jaideep et al. 2022),orders of magnitude faster than the numerical weather prediction (NWP). The Pangu model proposed by Huawei team can accurately and quickly predict the global climate by learning global meteorological monitoring data of past four decades (Bi et al. 2023).

Selection and design of neural networks is a key step in reconstructing climate data. Ghose selected recursive neural network (RNN) for groundwater level prediction (Ghose et al. 2018).Vu reconstructed 50 years groundwater level data in Normandy (France) based on the long and short-term memory (LSTM) (Vu et al. 2020). Differently, meteorological parameters are greatly spatially correlated, as a typical spatiotemporal sequence.Nature Subissue Geoscience published related research, which using image restoration technology combined with HadCRUT4 global historical temperature grid dataset, reconstructed complete global monthly grid temperature, and the reconstructed data sequence has extremely high correlation with the non-reconstructed data (Christopher et al. 2020). Continuity of time and spatial correlation must be considered simultaneously in the data reconstruction.Most of the frontier studies of spatio-temporal prediction are modeling based on graph neural network (GNN) and

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

Transformer (Pan&Li. 2021). But they have high computational complexity and memory overhead. Although MLP is a relatively simple deep learning model, the ability of spatiotemporal prediction is not inferior to complex models in recent studies.The MLPST model shows that, compared with RNN, GNN and Transformer, it can be very accurate even completely based on MLP (Zhang et al. 2023).Usually, different type of time series data have different characteristics, and screening some obvious characteristics can significantly improve the model performance (Tang et al. 2024).Therefore, in specific tasks, feature engineering and special model design need to be carried out to improve the prediction performance of the model.

To meet the demand of meteorological data reconstruction in agricultural productions, we designed a neural network models as a reconstruction tool based on MLP. A total of 143 missing data (43 weather stations) were reconstructed obtained from three divided study areas in Xinjiang.The parameters reconstructed include Max_T, Min_T, Ave _ T, Ave _ WVP, Ave _ RH, 10m Ave _ WS, and Sun_H. Inputs make up with short-term, cyclical, long-term trends and the same time data of weather stations with the highest sequence similarity, length of filled sequences ranged from one month to 38 years.The confidence of the results is measured by the correlation with the most adjacent station. Finally, datasets automatic construction module, automatic training module, missing positions automatic query module, and automatic rolling prediction module are integrated, realizing end-to-end data reconstruction and published as a micro service.

2 Study area and data

The study area is located in Xinjiang, northwest of China. Which is one of the most

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

important cotton production bases in China and most developed drought agricultural technology region(Liu. 2022). Located in the hinterland of Eurasia, due to complex terrain and frequent weather system activity, drought is the main climatic feature of this region (Mao et al. 2008). The Tianshan Mountains crosses the central region, divides Xinjiang into northern Xinjiang and southern Xinjiang. The water vapor could enter northern Xinjiang but hardly reach southern Xinjiang, so the drought degree difference of drought between the north and the south is obvious (Wang.2023). Yili River Valley located in west of the Tianshan Mountains in Xinjiang, surrounded by mountains on three sides, with abundant precipitation, forming special climate (Yan et al. 2017).

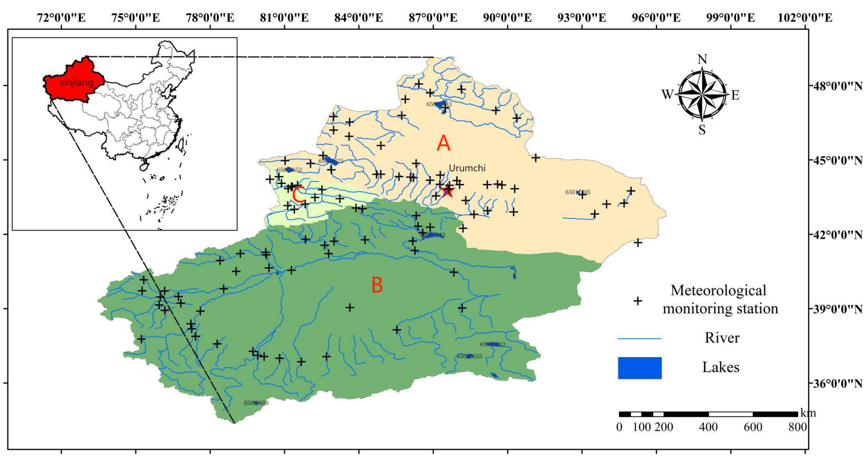

Figure 1. Study area division and location of weather stations

105 weather stations in this study are distributed in three areas: northern Xinjiang (A), southern Xinjiang (B) and Yili (C), having 48, 44, 13 sites respectively, recorded meteorological data for nearly 62 years (Figure 1). Among those, 44 stations exist data missing in varying degrees, with missing parameter types and missing duration varied. Table 1 listed codes and parameters of weather stations with missing data. Of these, 16 stations exist

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

Max_T and Min_T data missing. The number of Ave_T is 13, Ave_WVP and Ave_RH is 23, 10m Ave_WS is 28, Sun_H is 24. In totally, we need to reconstruct a total of 143 sequences,with time spans from 1961 to 2022. Figure 2 corresponding to Table 1, shows that the specific missing period, these missing lengths are long, the missing location is different, increased the difficulty of the reconstruction.

Table 1. Meteorological parameter types of missing data and weather station number of subtasks

Meteorological parameters	**Weather station number of subtasks**
Max_T(°C)	A29,A33,A34,A40,A46,B4,B10,B12,B16,B23,B31,B36,B38,B43,C1,C12
Min_T(°C)	A29,A33,A34,A40,A46,B4,B10,B12,B16,B23,B31,B36,B38,B43,C1,C12
Ave_T(°C)	A29,A33,A34,A40,A46,B10,B12,B16,B23,B31,B43,C1,C12
Ave_WVP(hPa)	A29,A32,A33,A34,A40,A46,B4,B6,B7,B9,B10,B12,B16, B23,B31,B34,B36,B38,B39,B41,B43,C1,C12
Ave_RH(%RH)	A29,A32,A33,A34,A40,A46,B4,B6,B7,B9,B10,B12,B16, B23,B31,B34,B36,B38,B39,B41,B43,C1,C12
10mAve_WS(m/s)	A9,A16,A29,A33,A34,A35,A40,A46,A48,B2,B4,B7,B10,B12,B16， B23,B26,B31,B34,B36,B38,B39,B40,B43,C1,C6,C11,C12
Sun_H（h）	A3,A4,A5,A11,A17,A21,A26,A29,A30,A31,A33,A34,A35,A38, A39,A40,A47,B4,B16,B31,B34,B36,B39,C12

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

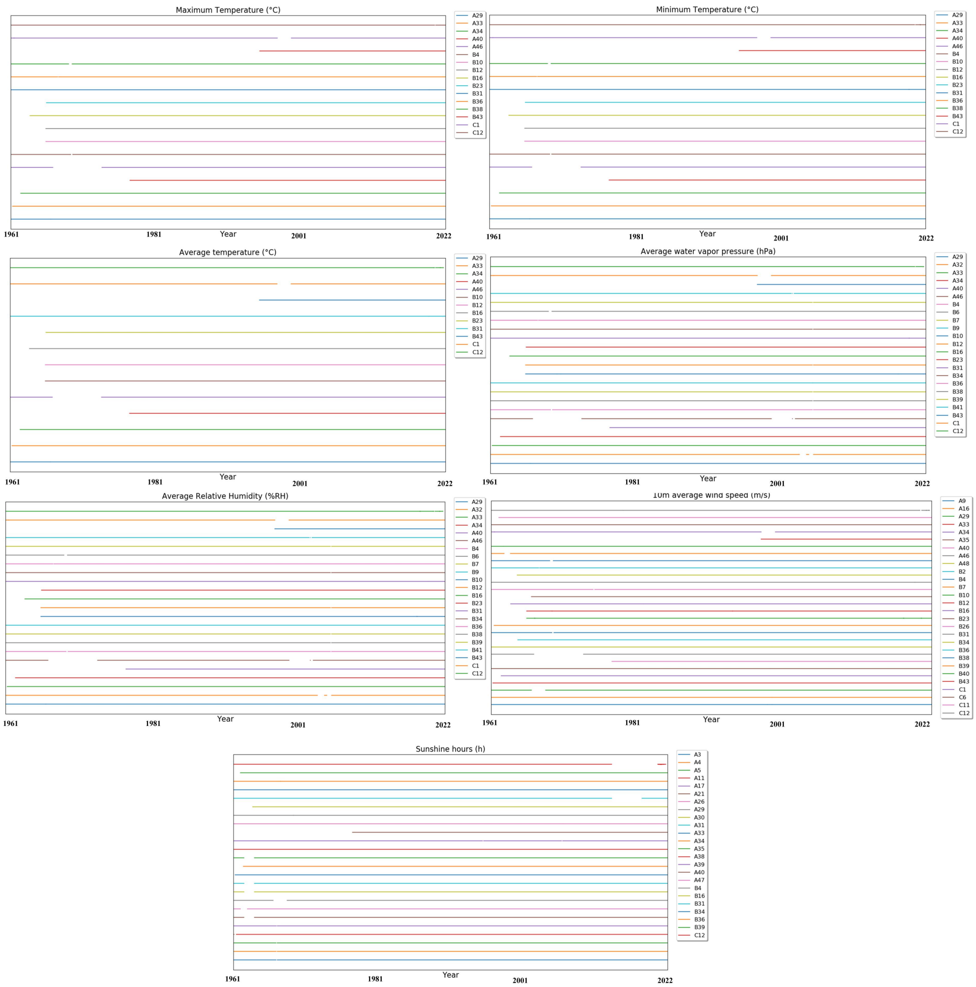

Figure 2. Measurement time-window at 105 weather stations over 62 years from 1961 to 2022

3 Methodology and model design

3.1 MLP

MLP is the most classic deep neural network, widely used to solve the classification and regression problems of nonlinearity. Whose structure (Figure 3 lower-right) includes input layer, hidden layer and output layer, each layer contains several neurons, and neurons in the upper and lower layers are connected to each other for information exchange (Benedict. 1988).

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

When training it, the weight parameters of the neurons are constantly updated until a good fit is achieved.Forward propagation and backpropagation are required to complete each time the weights are updated (Rumelhart et al. 1986). Forward propagation takes the outputs of the previous layer as the inputs of the next layer, calculates the outputs of the next layer according to the weight.Consider the layer1 and layer2 as examples, outputs of the layer2 is:

$$a_i^{layer2} = \sigma(b_i + \sum w_j a_j^{layer1})$$

Where σ is the activation function, which is the key for MLP to achieve nonlinear fitting. The most commonly used activation function, ReLU, is selected in this study (Glorot et al. 2011):

$$f(x) = \max(0, x)$$

Prediction errors is measured by cost function LOSS. The back propagation process is based on the chain conduction law, calculating gradients each layer parameters in the network to represent the influence of the parameters on the prediction errors, and updating the weight through multiply by learning rate α, until the loss value no longer drops. Which can be considered that the MLP model fitting has reached the optimal solution. The initial learning rate selected for this study was 0.001. The backpropagation process is:

$$w_{jnew} = w_j - \alpha \bullet \frac{\partial LOSS(y, \hat{y})}{\partial w_j}$$

3.2 Spatiotemporal MLP

Past studies have proved that,climate shows a short-term dependence, and which is cyclical and shows a trend in the long term (Liu et al. 2020), and closely associated with the adjacent site data. According to these experiences,we designed four modules based on the MLP (Figure 3): Spatial MLP, Short-term MLP, Periodic MLP, Trend MLP. Time series, with

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

different time scales resampled, were entered separately Short-term MLP, Periodic MLP, Trend MLP models, extracting short-term trends, cyclical and long-term trend characteristics of historical data respectively. Monitoring values from nearby stations were fed into the Spatial MLP module to obtain spatial associations between them. Results of the four modules are combined as inputs of predictive header, two fully connected layers. Which enable spatio-temporal association in the sequence is captured.

Dataset size and input sequence length are negatively correlated, need to be balanced when designing the inputs. In our model, all of the inputs length were set to 8, ensuring input format is unified. Inputs of short-term MLP are values last 8 days. Inputs of Periodic MLP and Trend MLP are values resampled according to 90 days intervals and 365 days respectively. Inputs of Spatial MLP were the monitoring values of eight stations with highest similarity to the target sequence within the study region. Using pyramid structure, the number of neurons in each layer is half the number of neurons in the previous layer, and which is usually able to extract features at different scales more effectively (Yang et al. 2020).

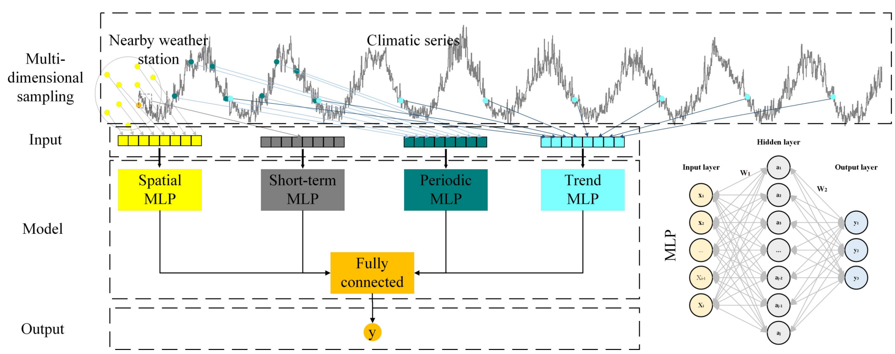

Figure 3. MLP and Model Framework

All meteorological data were standardized according to the following formula:

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

$$x_i^{normal} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

Where x_i^{normal} is normalized value, x_i is actual value, $x_{\max}$ and $x_{\min}$ are maximum and minimum value of sequence, respectively. This normalization method standardizes the values to between 0-1, be able to eliminate the effect of dimension and negative values for model fitting.

When predicting, we restore the results,and output the dimensional results:

$$y_i^{pred} = \hat{y}_i \bullet (x_{\max} - x_{\min}) + x_{\min}$$

Where y_i^{pred} is predicted value, $\hat{y}_i$ is output value of neural network.

3.3 Assessment methods

Meteorological similarity is measured by Euclidean distance of the sequence commonly,but there is a big difference between the different parameters. In order to standardize this index to 0-1, we define similarity based on Euclidean distance of two sequences:

$$SM_{\mathrm{mn}} = \frac{1}{e^{\sum_{\mathrm{i=1}}^{\mathrm{n}} |y_i - y'_i| / 100 * n}}$$

SM_{mn} represents the similarity of m sequence and n sequence, y_i, y'_i are the value of two sequences at the same time, respectively. n is the number of non-missing value. SM is closer to 1, the more similar, the closer to 0, the lower similarity. SM was used to select the inputs for the Spatial MLP module.

We used two common indicators, mean squared error (MSE) and mean absolute error (MAE), to evaluate the quality of the prediction:

$$MAE = \frac{1}{\mathrm{n}} \sum_{\mathrm{i=1}}^{\mathrm{n}} (y_i - \hat{y}_i)$$

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the real measure of GWL; $\hat{y}_i$ is the estimated value of GWL; and $\bar{y}_i$ is the mean of y_i. MAE and MSE were used to select the best number of hidden layers.

When evaluating the credibility of the reconstructed data, we used the correlation coefficient as the evaluation index：

$$corr_{m-n} = \frac{\sum_{i=1}^{n}(h_i^m - \bar{h}^m)(h_i^n - \bar{h}^n)}{\sqrt{\sum_{i=1}^{n}(h_i^m - \bar{h}^m)^2}\sqrt{\sum_{i=1}^{n}(h_i^n - \bar{h}^n)^2}}$$

h_i^m and h_i^n are value of m sequence and n sequence,respectively, $\bar{h}_i^m$ and $\bar{h}_i^n$ is the average value of m sequence and n sequence, respectively. Correlation coefficient is closer to 1, the reconstructed data is more credible, and correlation coefficient closer to 0, the more unreliable it is.

4 Reconstruction of missing climate data

4.1 Sub-task division

The reconstruction task was divided into 21 scenarios, 143 sub-tasks depending on the region and the parameters. Climate region A consists of 53 sub-tasks, among these, Max_T, Min_T and Ave_T take up 5 sub-tasks respectively; Ave_WVP and Ave_RH take up 6 sub-tasks respectively; Ave_WS and Sun_H take up 9 and 17 sub-tasks respectively.Climate region B has 75 sub-tasks, Max_T, Min_T, Ave_T, Ave_WVP, Ave_RH, Ave_WS, Sun_H take up 9, 9, 6, 15, 15, 15, 6 sub-tasks respectively. The numbers in climate region C are 2, 2, 2, 2, 2, 4, 1 respectively, total sub-tasks numbers were 15.

The SM of each sequence was calculated and derived as SM table. We show the SM of the target station and stations in the same region in Figure 4 due to the large amount of table

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

data. The more pronounced the yellow color, the higher the similarity, and the more pronounced blue the lower similarity, medium similarity shows green. Under all of the reconstruction scenarios, overall, the SM of these sequences ranged between 70.6% and 99.48%.Based on SM, 8 stations with the highest similarity to each target task, 143 groups in total. Following the spatiotemporal sampling method shown in Figure 3, build model inputs.

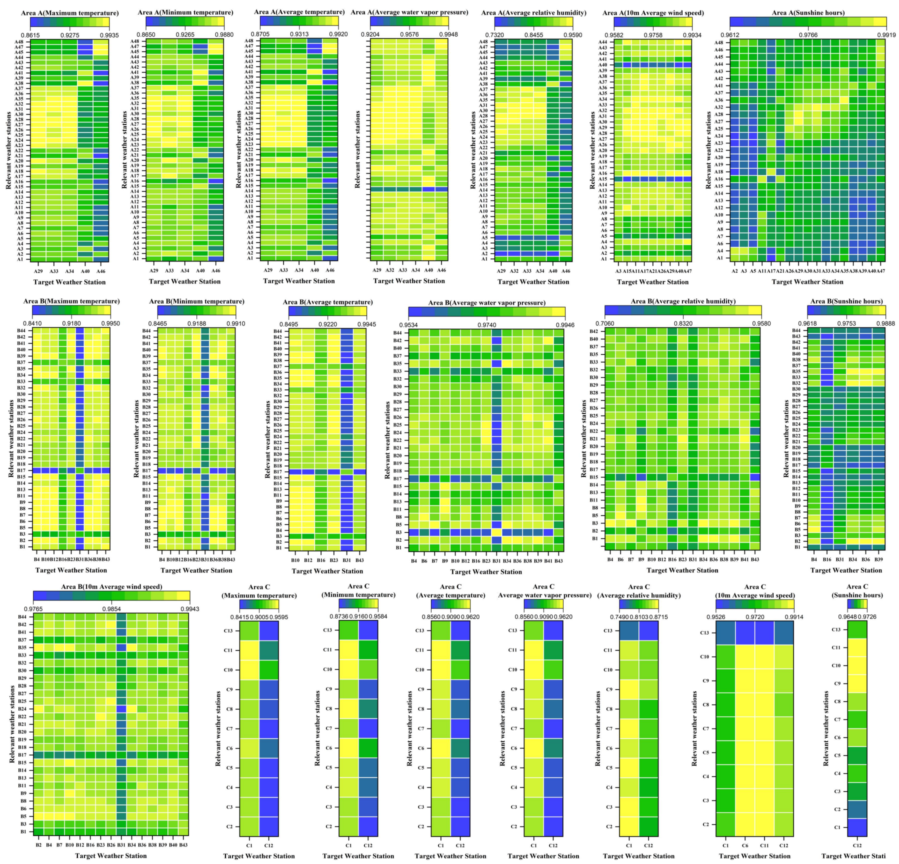

Figure 4. The similarity between the target weather station and the related weather station (under the sequence reconstruction scenario of different regions and different parameters)

4.2 Determine the number of MLP hidden layers

The number of layers of the MLP hidden layer is one of the most important parameters of model.Generally, increasing hidden layers can enhance the fitting ability of model, but

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

enhancement is limited, while largely increasing the number of model parameters, and causing slower run speed of model. To determine the number of hidden layers, we randomly picked five datasets, testing MSE, MAE and training time of the model prediction. Numbers of hidden layers is seted from 1 to 7(training 1000 epochs). Figure 5 displays, when setting 1-4 hidden layers, MAE and MSE of the model did have a significant downward trend as the hidden layer increasing. But when number of hidden layers greater than 4, MAE and MSE showed little improvement, even rose on some datasets. This may be related to the appearance of gradient explosion when model is too deep.

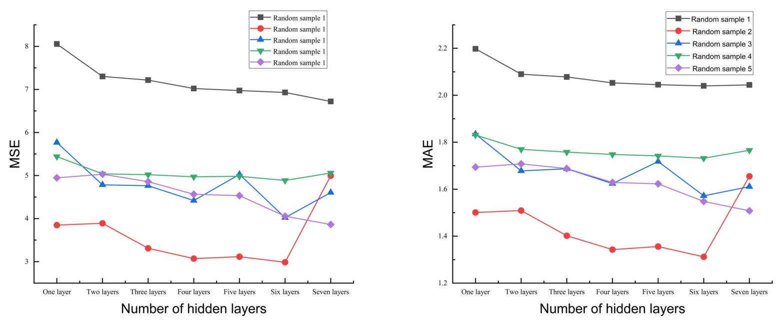

Figure 5. MAE and MSE trend with the number of hidden layers increases

Figure 6shows, time consumption to complete training increased significantly with the increase of the number of hidden layers, on the five randomly selected datasets, which displays almost linear. Considering the results of the above trials, the number of hidden layers chosed for MLP is 4, and the longest time consumption for a single task is 519.35s.

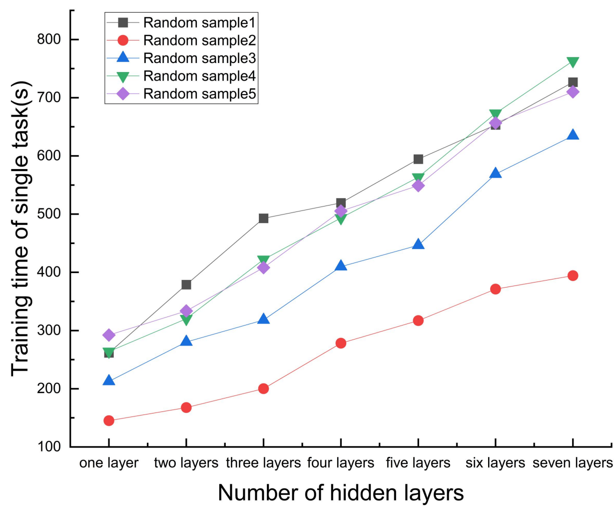

Figure 6. Time-consuming trend of training model with the number of hidden layers increases

4.3 Train model and reconstruct missing data

Figure 7shows our prediction process.Using pycharm2022.1 for our programming, we integrated multiple modules to implement end-to-end programs, with pre-processing data automatically, training model automatically,detecting missing data automatically,and reconstructing data automatically. The situation of missing data is complex, and the task of manually constructing datasets is large. Automatic data pre-processing model could generates datasets by reading data sheets according to task list, and completes normalization.Data were disrupted the order before entering.Automatic training model could Completes multiple tasks and saves as different parameter files.When predicting, missing data detection model could detects location of missing data.Later, according to detecting results, rolling prediction model automatically forward or backward predicting.

Tensorflow (Martín et al. 2016) was selected to be development framework in this study.Using Adaptive Moment Estimation(Adam) optimizer to improve learning efficiency

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

(Kingma&Ba. 2014), it can adjust automatically the learning rate according to historical gradient information. At the beginning of training, the larger learning rate helps the model to converge quickly. While later, learning rate adjusts smaller to improve accuracy of model.Meanwhile, Adam normalized the weight parameters, which also alleviates overfitting. MSE was choosed to be loss function.As a skill of training, Dropout layer can effectively prevent model overfitting (Nitish et al. 2014). In this study, the super-parameter of Dropout layers was set to 0.5.

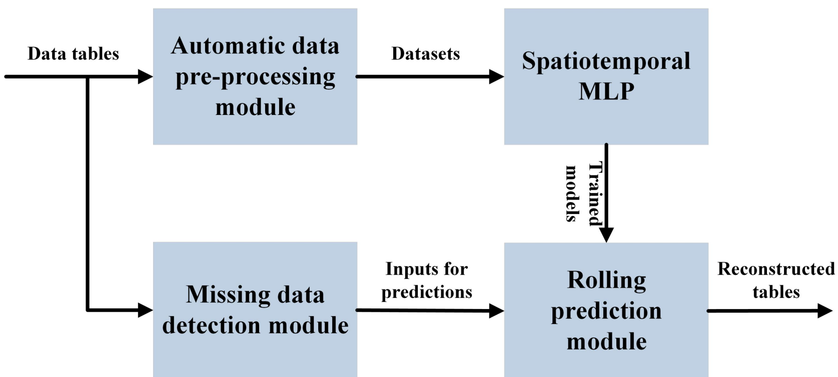

Figure 7. Structure of ensemble forecasting

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

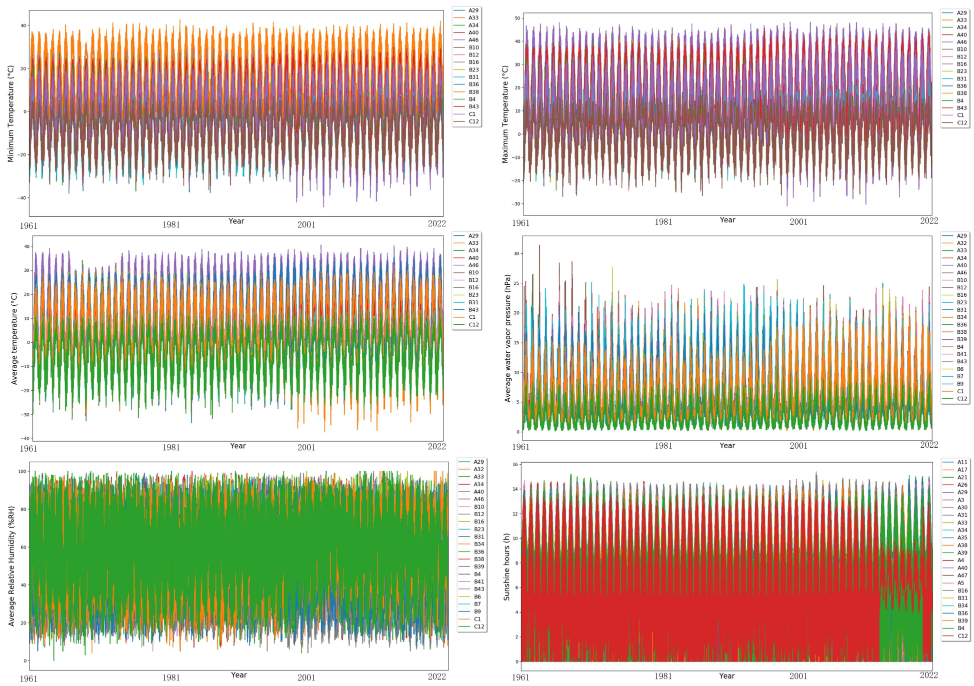

Figure 8. Results of reconstructed sequence(except Ave_WS)

Figure 8 shows reconstruction results of Max_T, Min_T, Ave_T, Ave_WVP, Ave_RH and Sun_H.Except Sun_H,from the figure, reconstructed sequences of other five parameters are indistinguishable from the real sequences. Sun_H usually represents time length,that the solar radiation above certain intensity. Which influenced by all kinds of meteorological factors, especially the change of the clouds. Our prediction values almost no 0 while measured exists some 0 values, can not mining to the occurrence of 0 values. But we can clearly see that, even without filtering, proposed model could dig out the cycle and trend laws. These reconstructed Sun_H data still very useful. For Ave_WS (Figure 9), model is difficult to predict the suddenly extreme wind speed. Which also shows shortage when predicting extreme values.

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

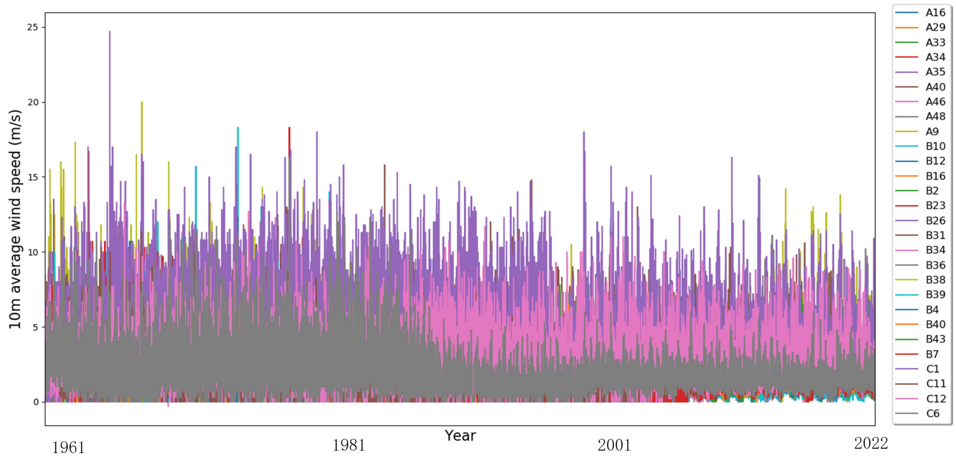

Figure 9. Result of reconstructed sequence(Ave_WS)

4.4 Evaluate quality of reconstructed data

Compared with the visual evaluation, the evaluation index can give more information about the reconstruction results. Assessing the quality of reconstructed data is very difficult due to the difficulty in tracing the real data of the past. By calculating correlation coefficient of sequences between reconstructed data and the nearest weather station data, credibility of reconstructed meteorological data was scientifically evaluated. A higher correlation coefficient indicates a higher confidence.

Table 2. Correlation of reconstructed sequences and nearest neighbor sequences

Weather station	Meteorological parameter Types Max_T (℃)	Min_T (℃)	Ave_T (℃)	Ave_WVP (hPa)	Ave_RH (%RH)	10m Ave_WS (m/s)	Sun_H (h)
A11-A10							0.838
A16-A10						0.521	
A17-A23							0.878
A21-A43							0.755
A26-A27							0.858
A29-A27	0.998	0.992	0.998	0.984	0.948	0.568	0.943
A3-A2							0.818
A30-A37							0.850
A31-A27							0.903
A32-A37				0.967	0.926		
A33-A41	0.937	0.952	0.955	0.961	0.450	0.314	0.774
A34-A37	0.987	0.982	0.989	0.959	0.927	0.371	0.879
A35-A36						0.681	0.944
A38-A37							0.780

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

A39-A27							0.830
A4-A1							0.741
A40-A1	0.937	0.911	0.948	0.917	0.225	0.220	0.466
A46-A36	0.962	0.952	0.966	0.902	0.562	0.196	
A47-A45							0.872
A48-A43						0.261	
A5-A2							0.646
A9-A8						0.517	
B10-B11	0.993	0.975	0.992	0.957	0.764	0.566	
B12-B11	0.994	0.980	0.994	0.967	0.817	0.579	
B16-B17	0.955	0.948	0.962	0.900	0.559	0.373	0.612
B2-B35						0.705	
B23-B22	0.997	0.993	0.997	0.979	0.906	0.608	
B26-B25						0.711	
B31-B33	0.942	0.949	0.964	0.895	0.450	0.589	0.687
B34-B32				0.920	0.819	0.621	0.884
B36-B32	0.996	0.982		0.929	0.886	0.644	0.904
B38-B37	0.973	0.969		0.932	0.535	0.439	
B39-B32				0.935	0.853	0.685	0.851
B4-B5	0.996	0.990		0.958	0.851	0.655	0.865
B40-B21						0.469	
B41-B22				0.978	0.890		
B43-B14	0.984	0.948	0.984	0.928	0.749	0.405	
B6-B5				0.966	0.853		
B7-B5				0.977	0.920	0.681	
B9-B15				0.964	0.871		
C1-C10	0.934	0.923	0.944	0.876	0.435	0.301	
C11-C8						0.298	
C12-C3	0.922	0.925	0.931	0.909	0.356	0.114	0.366
C6-C8						0.585	

Correlation coefficients of all reconstructed sub-tasks are shown in Table 2. From the reconstruction effect of temperature,we very approach the results of Christopher (0.9941) (Christopher et al.2020), even exceed theirs in 4 tasks of temperature reconstruction (45 in total). More importantly, the data we reconstructed are of daily scale, smaller than their time granularity (monthly). Our work demonstrates that MLP with special spatiotemporal design can better reconstructe climate data.

The consistency of evaluation indicators in different tasks is also one of the goals we pursue. Figure 10 shows the distribution of its correlation coefficient index.Which can be easily see, Max_T, Min_T, Ave_T, Ave_WVP shows excellent performance, with average correlation coefficient is over 0.9 and distribution is very concentrated; Ave_RH and Sun_H performance unstable, although average correlation coefficient is over 0.7 but dispersed;

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

Ave_WS shows poor results, average correlation coefficient is around 0.5 and distributing is dispersed.

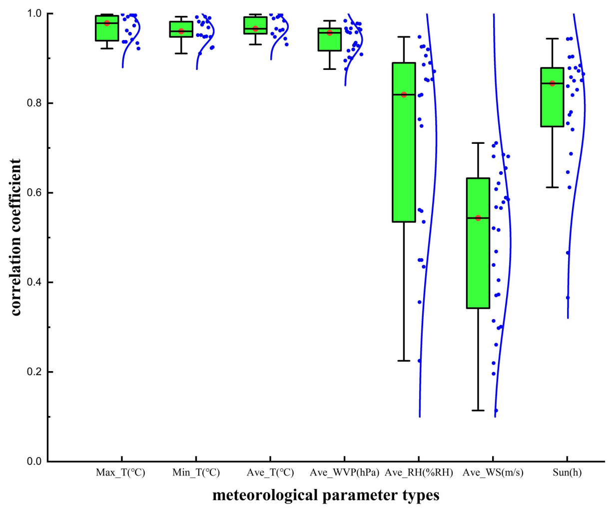

Figure 10. The evaluation indicators distribution of different type parameters reconstruction

4.5 Release model

In order to provide convenient services for people in the agriculture field, the model accomplished in this study will published on the Agricultural Smart Brain platform as a tool.Which developed by Beijing Lianchuang Siyuan Measurement and control Technology Co., LTD, providing scientific research data, computing power and publishing AI micro-services for scientific researchers. Users can obtain it by purchasing access authority to the platform. The link of our microservices as follow: http: // 192. 168. 50. 201: 15000/ app/ services/ visionarytech/ test-1/ alg- 26bc86f42a137f8f

5 Conclusion

In order to reconstruct the long-term missing meteorological monitoring data in agricultural field, proposed an end-to-end rapid reconstruction method based on MLP.

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

Spatio-temporal datasets were built according to the similarity indicator SM, standardized data to ensure good performance of the model. Backbone was designed four MLP modules with 4-hidden layers to jointly learn short-term trends, periodicity, long-term trends, and spatial associations. Predictive head consisted with two fully connected layers. After that, the automatic preprocessing, automatic detection of missing data location, automatic model training and rolling prediction modules are coded and integrated to realize end-to-end long sequence reconstruction. Our model is able to complete a single reconstruction task within 10 minute.

Daily meteorological monitoring data of 44 meteorological stations (143 tasks) in Xinjiang from 1961 to 2022 were reconstructed using our method. The evaluation indexs show that, average correlation coefficient of Max_T, Min_T, Ave_T and Ave_WVP are 0.969,0.961,0.971, and 0.942 respectively, showing high consistency and high credibility; average correlation coefficient of Ave_RH and Sun_H are 0.720 and 0.789 respectively, showing low consistency and general credibility; average correlation coefficient of Ave_WS is 0.488, showing low consistency and low credibility. The results demonstrate that MLP was useful and reliable in the task of rapidly reconstructing missing meteorological data, which will provide an important solution to solve the problem of missing data in agrometeorological field.

Finally, we released our model on Agricultural Smart Brain platform, provided users a tool of data reconstruction, in the form of micro service.

Conflict of interest: None.

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

Code/data availability: Both the code and data are freely available by contacting the corresponding authors.

Author contributions:

ZhangYan: methodology;essay writing;proofreading of dissertations

XuTianxin: data processing;method validation;essay writing;proofreading of dissertations

Zhangchenjia: methodology;method validation;coding;proofreading of dissertations

MaDaokun: financial support;project management

Thanks

Thanks to financial support of Silk Road Economic Belt Innovation-Driven Development Pilot Zone, WuChangShi National Independent Innovation Demonstration Zone project(2022LQ04001).

Thanks to Beijing Lianchuang Siyuan Measurement and Control Technology Co., Ltd. and Beijing Zhiyu Chuangyi Co., Ltd. for their help in the release of micro service.

Reference

Anderson, S.P., Bales, R.C., Duffy, C.J., 2008. Critical Zone Observatories: building a network to advance interdisciplinary study of Earth surface processes. Mineral. Mag. 72 (1), 7 – 10.

Benedict A K. 1988. Learning in the multilayer perceptron. Journal of Physics A: Mathematical and General.21(11).

Bonnet, R., Bo´e, J., Habets, F., 2020. Influence of multidecadal variability on high and low flows: the case of the Seine basin. Hydrol. Earth Syst. Sci. 24, 1611 – 1631.

Bi Kaifeng, Xie Lingxi, Zhang Hengheng, Chen Xin, Gu Xiaotao, Tian Qi. Accurate

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

medium-range global weather forecasting with 3D neural networks. Nature 619, 533–538 (2023). https://doi.org/10.1038/s41586-023-06185-3

Buda Su, Jinlong Huang, T. Fischer, Yanjun Wang, Z. Kundzewicz, J. Zhai, Hemin Sun, Anqian Wang, X. Zeng, Guojie Wang, H. Tao, M. Gemmer, Xiucang Li, T. Jiang. 2018. Drought losses in China might double between the 1.5 ° C and 2.0 ° C warming. Proc Natl Acad Sci USA， 115（42）：10600-10605

Christopher Kadow, David Matthew Hall & Uwe Ulbrich. 2020. Artificial intelligence reconstructs missing climate information. Nat. Geosci. 13, 408 – 413 (2020). https://doi.org/10.1038/s41561-020-0582-5

Ghose, D., Das, U., Roy, P., 2018. Modeling response of run off and evapotranspiration for predicting water table depth in arid region using dynamic recurrent neural network. Groundwater Sustainable Dev. 6, 263 – 269.

Ghasem A ,Nader M .2022. Reconstruction of particle image velocimetry data using flow-based features and validation index: a machine learning approach. Measurement Science and Technology, 2022, 33(1)

Glorot, Xavier, Antoine Bordes, Yoshua Bengio.2011. “Deep Sparse Rectifier Neural Networks.” International Conference on Artificial Intelligence and Statistics (2011).

IPCC. 2012. Summary for policymakers // Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. A Special Report of Working Groups Ⅰ and Ⅱ of the Intergovernmental Panel on Climate Change. Cambridge：Cambridge University Press，1-19

Jaideep Pathak, Shashank Subramanian, Peter Harrington, Sanjeev Raja, Ashesh

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

Chattopadhyay, Morteza Mardani, Thorsten Kurth, David Hall, Zongyi Li, Kamyar Azizzadenesheli, Pedram Hassanzadeh, Karthik Kashinath, Animashree Anandkumar. FourCastNet: a global data-driven high-resolution weather model using adaptive Fourier neural operators. Preprint at https://arxiv.org/abs/2202.11214 (2022).

Kingma P D ,Ba J . 2014. Adam: A Method for Stochastic Optimization. CoRR 2014, abs/1412.6980

Li C ,Ren X ,Zhao G .2023. Machine-Learning-Based Imputation Method for Filling Missing Values in Ground Meteorological Observation Data. Algorithms, 2023, 16 (9):

Liu Kai, Nie Gege, Zhang Sen. 2020. Study on the Spatiotemporal Evolution of Temperature and Precipitation in China from 1951 to 2018. Advances in Earth Science, 2020, 35(11): 1113-1126 DOI:10.11867/j.issn.1001-8166.2020.102

Liu Yi. 2022. Build a national high-quality cotton production base. Xinjiang Daily, 2022-09-05 (001). DOI:10.28887/n.cnki.nxjrb.2022.003680.(in chinese)

Mao Weiyi,Nan Qinghong,Shi Hongzheng. 2008. Research on climate change characteristics and climate zoning methods in Xinjiang. Meteorological Calendar, 2008, (10): 67-73.(in chinese)

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek Gordon Murray, Benoit Steiner, Paul A. Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, Xiaoqiang Zhang. 2016. TensorFlow: A system for large-scale machine learning. CoRR, 2016, abs/1605.08695

https://doi.org/10.5194/egusphere-2024-439
Preprint. Discussion started: 7 March 2024

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 2014, 15 (1): 1929-1958.

Pan Zhisong, Li Wei.2021. Survey of Spatio-temporal Series Prediction Methods Based on Deep Learning. Data Acquisition and Processing, 2021, 36 (03): 436-448. DOI:10.16337/j.1004-9037.2021.03.003.（in Chinese）

Qin D H，Ding Y H，Wang S W，Wang S M，Dong G R，Lin E D，Liu C Q，She Z X，Sun H N ， Wang S R ， Wu G H. 2002. Ecological and environmental change in West China and its response strategy. Adv Earth Sci，17（3）：314-319（in Chinese）

Rajaee, T., Ebrahimi, H., Nourani, V.2019. A review of the artificial intelligence methods in groundwater level modelling. J. Hydrol. 572, 336–351.

Rumelhart, D., Hinton, G., Williams, R. 1986. Learning representations by back-propagating errors. Nature 323, 533 – 536 (1986). https://doi.org/10.1038/323533a0

Tang D ,Zhan Y ,Yang F. 2024. A review of machine learning for modeling air quality: Overlooked but important issues. Atmospheric Research, 2024, 300, 107261-.

UNDRR. Human Cost of Disasters: An Overview of the Last 20 Years 2000–2019; United Nations for Disaster Risk Reduction (UNISDR): Geneva, Switzerland, 2020.

Vu M.T., Jardani A., Massei N., Fournier M. 2020. Reconstruction of missing groundwater level data by using Long Short-Term Memory (LSTM) deep neural network. Journal of Hydrology, 2020, (prepublish): 125776-.

Wang Jiaoyan.2023. Distribution and evolution characteristics of drought under the background of climate warming and humidification in Xinjiang. Arid Environment Monitoring, 2023, 37 (01): 15-21.(in chinese)

Yan Junjie, Yan Min, Cui Dong, Liu Haijun, Chen Chen, Xia Qianqian. 2017. Analysis of temperature and precipitation trends in the Ili River Valley of Xinjiang in the past 55 years. Hydropower Energy Science, 2017, 35 (10): 13-16+12.

Yang C., Xu Y., Shi J., Dai B., Zhou B. 2020. Temporal pyramid network for action recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2020, 588-597.

Yao Ziqiang，Zhang Tao，Wu Li，Wang Xiaoying，Huang Jianqiang. Physics-Informed Deep Learning for Reconstruction of Spatial Missing Climate Information in the Antarctic.Atmosphere, 2023, 14 (4)

Zhang Zijian, Huang Ze, Hu Zhiwei, Zhao Xiangyu, Wang Wanyu, Liu Zitao, Zhang Junbo, Qin S.Joe,Zhao Hongwei. 2023. MLPST: MLP is All You Need for Spatio-Temporal Prediction. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23). Association for Computing Machinery, New York, NY, USA, 3381－3390. https://doi.org/10.1145/3583780.3614969

Ziolkowska Jadwiga R, Zubillaga Jesus. 2018. Importance of weather monitoring for agricultural decision-making - an exploratory behavioral study for Oklahoma Mesonet.. Journal of the science of food and agriculture (13), 4945-4954.