Predicting the Solar Activity Cycle Based on A LSTM-ARIMA Hybrid Model

Yusen Zhou *

College of Astronautics NUAA Nanjing University of Aeronautics and Astronautics Nanjing, China

* Corresponding Author Email: nuaazys@nuaa.edu.cn

Abstract. Solar activity has a profound impact on the Earth, and the development of effective models to accurately predict solar activity will help us to further our understanding of space weather changes caused by solar activity and their possible consequences.In this paper, data on the area of the solar magnetic field, the number of sunspots and the area of sunspot regions were collected and processed, and histograms and box plots of the distributions of the three types of data were obtained, and it was found that the former showed a normal distribution and the latter two showed a skewed distribution. Subsequently, the data were further processed using the sliding average method to obtain the smoothed trend time series and fluctuation time series of the above three groups of data.The use of a single common time series forecasting model to predict solar activity proves to be unreliable due to the uncertainty and non-linearity of the solar activity cycle and intensity. Therefore, in this paper, separate ARIMA and LSTM forecasting models are first constructed to predict the general and fluctuating trends of solar activity indicators, respectively. Then, a hybrid prediction model is built to integrate the prediction results of the two models through denormalization, weight adjustment and noise removal to obtain the final prediction results.This paper focuses on predicting the changes in the solar magnetic field area, the number of sunspots, and the sunspot area to further speculate on the beginning and end of solar activity and its intensity. Finally, this paper provides a comparative analysis, reliability validation, and parameterization of the established models. From the perspectives of sliding and rolling validation, this paper compares the stability and accuracy of the independent ARIMA prediction model and LSTM prediction model as well as the hybrid model in processing raw data. In terms of model parameter determination, this paper analyzes and determines the number of cycles, hidden layer grid size in the LSTM model and the autoregressive term, moving average term and difference order in the ARIMA model, respectively.

Keywords: solar activity, LSTM, ARIMA, Fusion Model.

1. Introduction

Solar activity, as an indispensable natural driving force, has important implications for the surface environment and human space exploration missions. It has been shown that increased solar activity can lead to an increase in the flux of high-energy protons, resulting in the formation of super geomagnetic storms [1]. Meanwhile, the strength of solar activity also directly affects the on-orbit lifetime and operational performance of spacecraft [2]. As an important external feature of solar activity, the relative number of sunspots is an indicator of the strength of solar activity [3]. Sunspots usually appear as a pair of magnetic poles opposite to each other in the active region, and their period coincides with the solar cycle of about 11 years. At the same time, solar activity is also inextricably linked to the magnetic field of the Sun and the size of sunspot regions. The inherent instability of the solar activity cycle and the variability of the maximum intensity of activity lead to a lack of accuracy in the predictions of many existing prediction models. Therefore, it is of great significance to establish a reliable mathematical prediction model on the basis of existing data to make a reasonable prediction of the pattern of change of solar activity, which is important for the prediction of space weather and the state of the ionosphere.

Sanjay B. Waykar [4]et al. designed a deep LSTM model for predicting solar activity using sunspot number (SSN) and solar radio flux (SRF), which is mainly involved in the extraction of technological indicators and the prediction of solar activity; Krasheninnikov and S. Chumakov [5] applied the Elman artificial neural network platform to the historical data series of observational data

and analyzed the possibility of predicting the dependence of sunspot number (SSN) in the solar activity cycle. A method for normalizing the initial data is proposed, i.e., constructing a virtual idealized cycle using the time scaling factor and the maximum value in the solar activity cycle.

2. Data Preprocession

2.1. Basic Preprocessing Operations

For the raw data, this paper first averaged the default values in the data to fill them in and converted the strings present in the raw data. Then a general trend analysis was performed on all the data, and histograms and box plots of the raw data for the solar magnetic field, sunspot number, and sunspot area were obtained as follows: Figures 1 and 2 show the distribution of the solar magnetic field data, Figures 3 and 4 show the distribution of the sunspot number data, and Figures 5 and 6 show the distribution of the sunspot area data, as shown in the following figures.

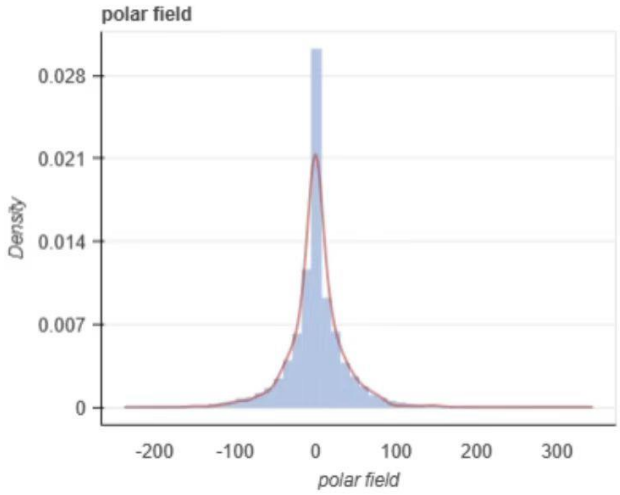

Fig. 1 Polar field distribution

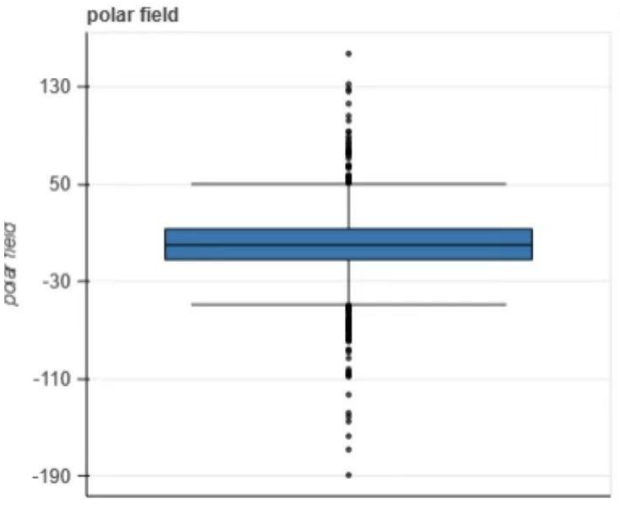

Fig. 2 Polar field distribution

Observing Figures 1 and 2, it can be found that the magnetic pole strength of the solar magnetic field is concentrated around 0. The changes in the magnetic poles of the solar magnetic field may correspond to the beginning and end of the solar cycle. Observing Fig. 1, it can be found that the magnetic pole intensity of the solar magnetic field is more concentrated, but there are still some large anomalies, which may be related to the intensity of solar activity.

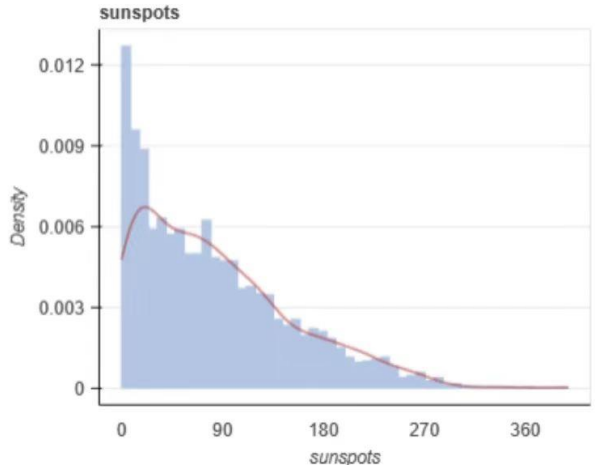

Fig. 3 Sunspot number distribution

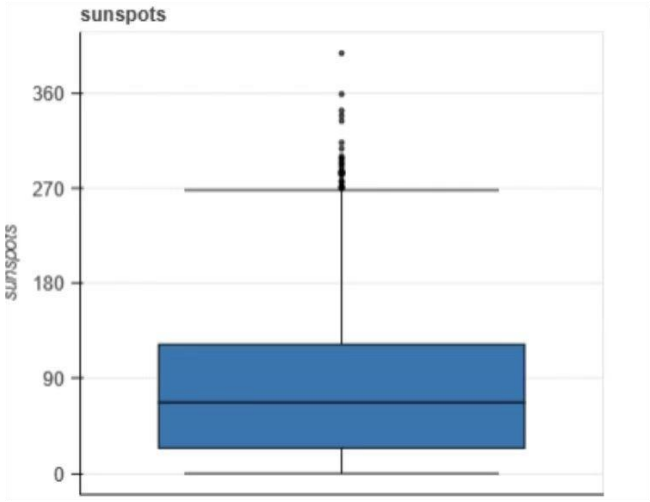

Fig. 4 Sunspot number distribution

From the analysis of Figs. 3 and 4, it can be seen that the sunspot numbers are relatively concentrated in the range of 20 to 140. In addition, there are some outlier points in the image, which may be the sunspot number corresponding to the solar maximum.

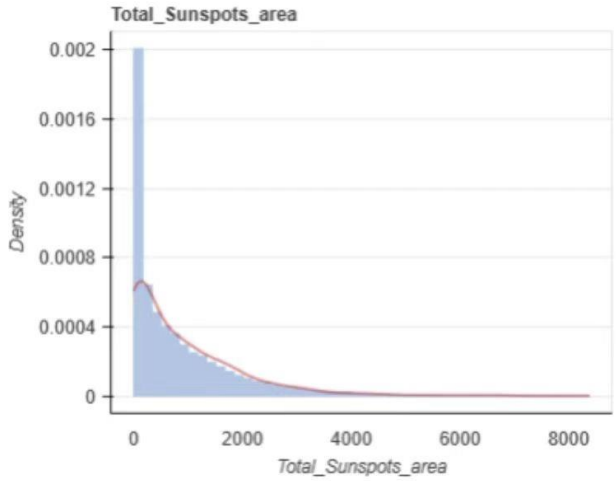

Fig. 5 Sunspot area distribution

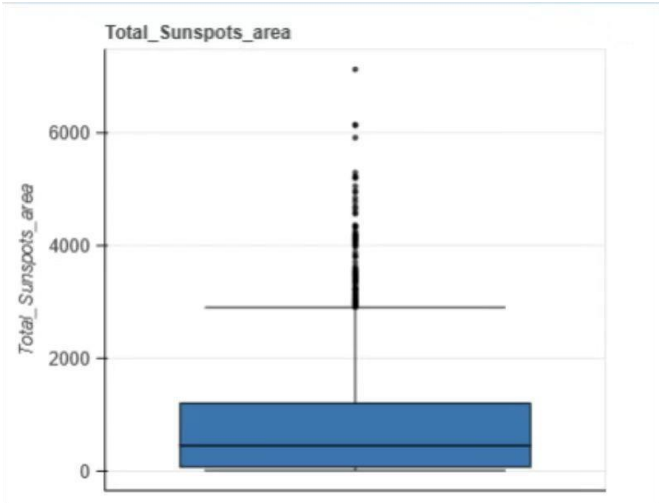

Fig. 6 Sunspot area distribution

From the analysis of Figs. 5 and 6, it can be seen that most of the sunspot areas are in the range of 0 to 1000, and since the intensity of solar activity is closely related to the sunspot area [6], it means that the intensity of solar activity is at a low level most of the time. In addition, there are outliers in Figure 8, which may be the sunspot areas corresponding to the most intense moments of solar activity.

2.2. Data Decomposition Based on Moving Average

In order to make the model prediction more accurate, this paper uses the moving average operation to further decompose the preprocessed basic data into smoothed time series and fluctuation time series. Among them, the moving average window is taken as 13 months, and the fluctuation time series is obtained by subtracting the smoothed time series from the preprocessed time series.

The formula for calculating the moving average of a time series:

$$(X)_t = \frac{1}{k} \sum_{i=t-k+1}^{t} X_i \tag{1}$$

The moving average result $(X)_t$ is the moving average at the t point in time, k is the size of the sliding window (the number of data points contained in the window), and X_i is the value of the time series at the i point in time.

In this paper, Python is used to further process the data obtained after the first step of processing to obtain the smoothed time series, the fluctuation time series of the magnetic pole distribution of the sun, the number of sunspots, and the area of sunspots, respectively. Due to the limited space of the article, some of the results (figure7,figure 8 and figure 9) are shown as follows:

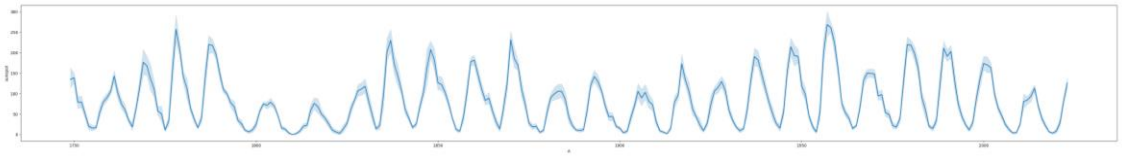

Fig. 7 Sunspot number time series

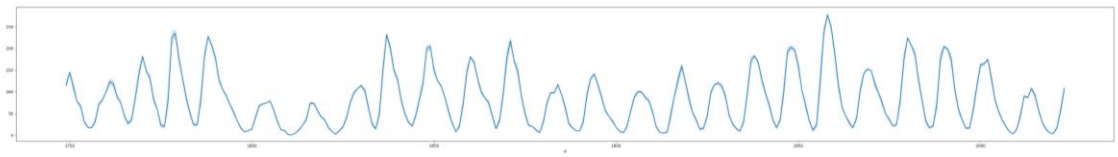

Fig. 8 Sunspot number smoothed time series

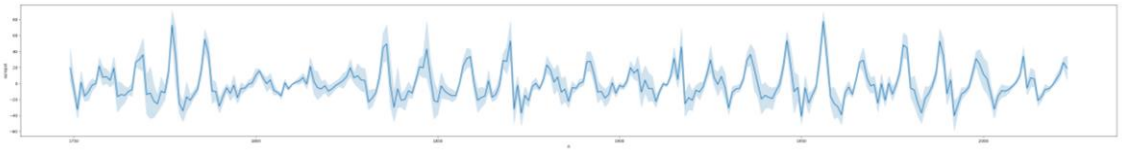

Fig. 9 Sunspot number fluctuation time series

3. Introduction to the Modeling Algorithm

3.1. The Establishment of the ARIMA Model

Auto-Regressive Integrated Moving Average is a statistical model widely used for time series analysis and forecasting [7].The ARIMA model combines both auto-regressive (AR) and moving average (MA) components with an integrated component.The model is a common used for analyzing and forecasting model for analyzing and forecasting time series data. It can capture the trend, seasonality, and random fluctuation components of the data [8].

In this paper, the cleaned monthly average data are firstly processed by moving average with a window period of 13. By comparing the two sets of data before and after the processing, it can be found that the smoothing of the data after the moving average processing is enhanced, the trend is more obvious, the cyclical characteristics are more prominent, the outliers are suppressed with obvious effect, and the data volatility is weakened.

Auto-regressive:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + \xi_t \tag{2}$$

Moving Average:

$$Y_t = \mu = \mu_t + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_q Y_{t-q} \tag{3}$$

Differential:

$$Y_t = Y_t - Y_{t-d} \tag{4}$$

ARIMA model expression:

$$Y_t = c + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + \ldots + \varphi_p Y_{t-p} + Y_t - Y_{t-d} + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \tag{5}$$

Where, $\varphi 1, \varphi 2, \ldots, \varphi_p$ represents the auto-regressive coefficient, ε_t is the white noise error at the current moment, $\theta_1, \theta_2, \ldots, \theta_q$ represents the moving average coefficient, Y_t is the current value of the time series, and c is a constant term.

3.2. The Establishment of the LSTM Model

Long Short-Term Memory is a variant of Recurrent Neural Network (RNN) specially designed to solve the problem of long sequence dependence and gradient vanishing [9].LSTM performs well in dealing with time series and sequential data, and is able to deal with the noise and uncertainty in the data to a certain extent, and is widely used in time series data prediction tasks in the financial field, meteorology [10], and stock market [11].

In this paper, a set of volatility time series data is obtained by subtracting the cleaned data from the smoothed data after moving average processing. Compared with the moving average processed data, this data removes the trend of the data and makes the volatility of the original data more obvious, and the data can better reflect the cyclical components in the original data, highlight the outliers in the original data, and retain the noise information, reflecting the instantaneous changes in the original data.

Since the above data features match the features of LSTM applicable data types, the LSTM model is chosen to analyze and predict these data in this paper. The mathematical expression [12] and flowchart 10 of this model are as follows:
Input Gate:

$$\begin{aligned} i_t &= \sigma\sigma(\ _{ii}x_t + b_{ii} + W_{hi}\ h_{t\ -1} + b_{hi}\ ) \\ c_t &= \tanh\left(W_{ic}x_t + b_{ic} + W_{hc}h_{t-1} + b_{hc}\right) \end{aligned} \tag{6}$$

Forget Gate:

$$f_t = \sigma\sigma(\ _{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hc}\ ) \tag{7}$$

Cell State Update:

$$c_t = f_t c_{t\ -1} + i_t \tilde{c}_t \tag{8}$$

Output Gate:

$$\begin{aligned} o_t &= \sigma\sigma(\ _{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}\ ) \\ h_t &= o_t\ \tanh\left(c_t\ \right) \end{aligned} \tag{9}$$

Where, i_t denotes the output of the input gate. $\tilde{c}_t$ denotes the candidate cell state of the input gate. f_t denotes the output of the forgetting gate. x_t is the current cell state. h_t is the hidden state and the output of the LSTM model. x_t is the input of the t step of the input sequence. h_{t-1} is the hidden state of the previous time step. W_{ii}, W_{if}, W_{io}, W_{ic} represents the weight matrix of the input gate. $W_{hi}, W_{hf}, W_{ho}, W_{hc}$ represents the weight matrix of the hidden state. $b_{ii}, b_{if}, b_{io}, b_{ic}$ represents the bias of the input gate. $b_{hi}, b_{hf}, b_{ho}, b_{hc}$ represents the bias of the hidden state. σ represents the sigmoid activation function. $\tanh$ is the hyperbolic tangent activation function.

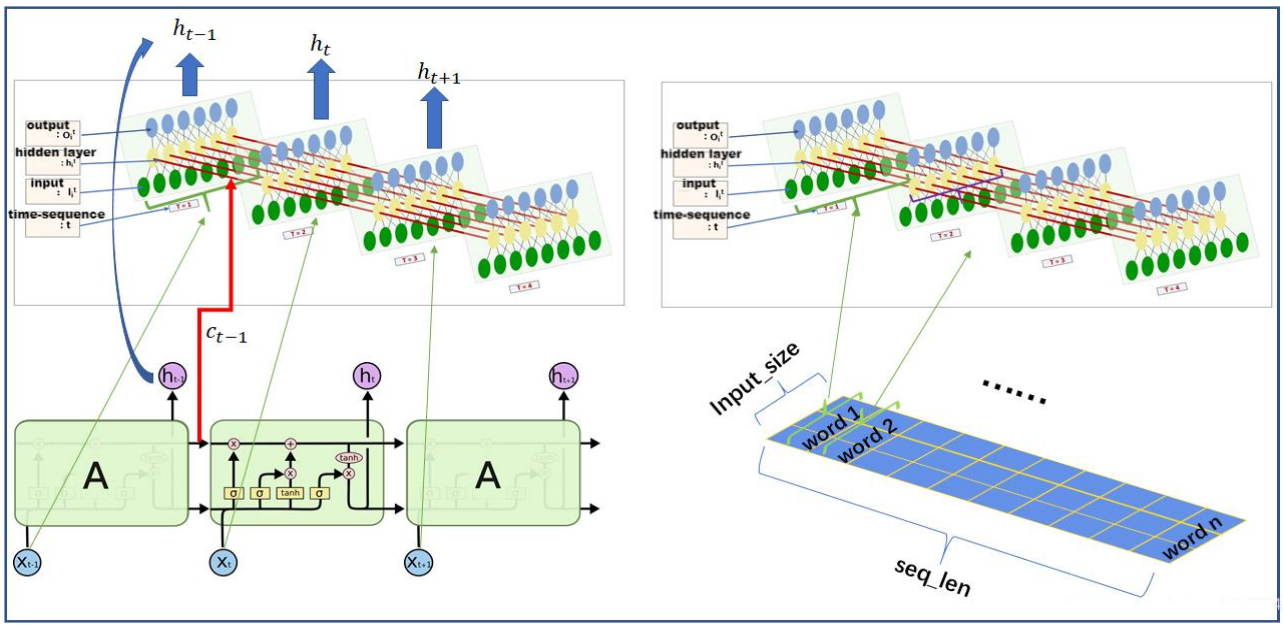

Fig. 10 Schematic Diagram of LSTM

4. Model Fusion

4.1. The Establishment of the Fusion Model

In fusing the trend time series obtained from the ARIMA model and the fluctuation time series obtained from the LSTM model, this paper adopts the following method: firstly, inverse normalization is performed to return the two series to the original data measurement space, and then weight adjustment is performed. Considering that sequences with a lot of noise in the forecast may lead to larger errors, this paper reduces their weights, while enhancing the weights related to the main trend, realizing a combination of main and fine-tuning effects. Finally, this paper employs a simple denoising process to emphasize the obvious trends in the ultimate cycle. This approach enhances the robustness and performance of the hybrid model.

This process primarily comprises three steps: denormalization, weight adjustment, and noise removal.

4.2. Predicted Results

After the above modeling steps and analysis, the final prediction results are shown as follows:

In this paper, the solar magnetic field, sunspot number and sunspot area trend time series data are imported into the ARIMA model to obtain the solar magnetic field, sunspot number and sunspot area potential time series prediction data, and then the solar magnetic field, sunspot number and sunspot area fluctuation time series data are imported into the LSTM model to obtain the solar magnetic field, sunspot number and sunspot area fluctuation time series prediction data. Finally, the hybrid model strategy is used to integrate the above data and plot the results, as shown in the following figures. Figures 11, 12 and 13 show the prediction results of the solar magnetic field, sunspot number and sunspot area, respectively:

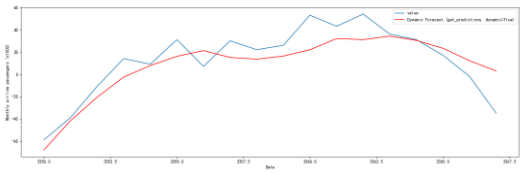

Fig. 11 Results of magnetic field prediction of fusion models

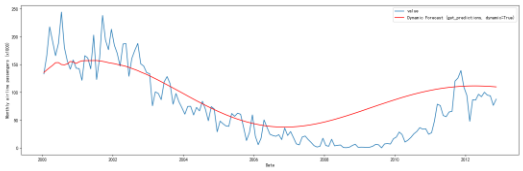

Fig. 12 Results of sunspot number prediction from fusion models

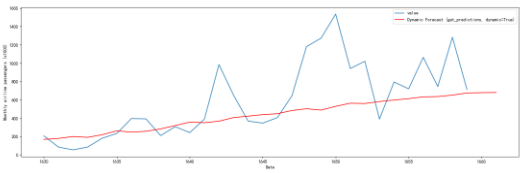

Fig. 13 Results of sunspot area prediction of fusion models

5. Model Comparison and Reliability Verification

5.1. Sliding Validation

It can be seen from the observation of fig.14 that the same LSTM model has a good fitting relationship in the magnetic field of the data set, but there are obvious errors in the sunspot number

data set and sunspot area data set. It can be seen that the difficulty of the problem situations in the three data sets is different, which also provides interpretability for the effect improvement in the figure15 below. In the less difficult magnetic field data set scenario, the improvement of the fusion model is not obvious, and there is only a small optimization at the corners; but in the more difficult problem scenario, the sunspot area and number data sets both show large translations and the prediction of data is more accurate in terms of the overall trend.

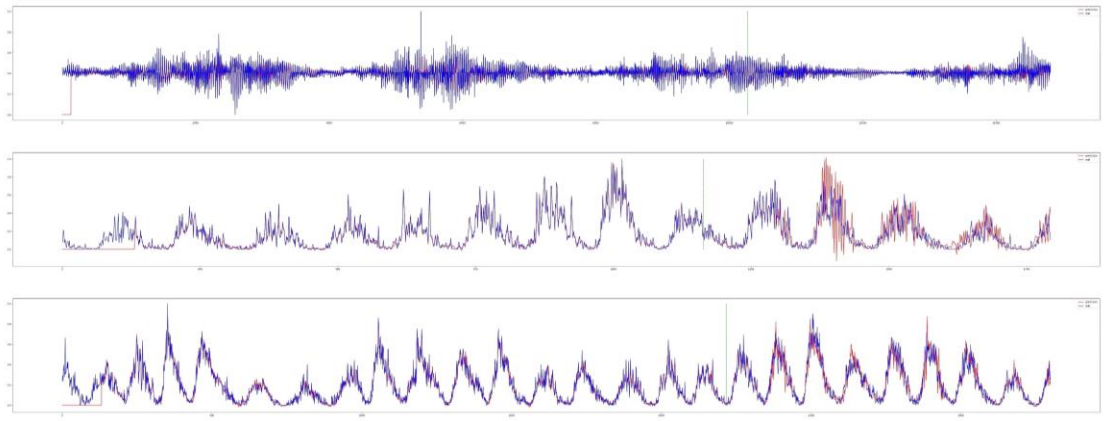

Fig. 14 The fitting effect of LSTM on the data sets of terrestrial magnetic field (row 1), sunspot area (row 2), and sunspot number (row 3) (the test set is after the green dotted line)

5.2. Rolling validation

The example data set selected in figure 15 has similar segments, LSTM is a separate segment, and ARIMA and the fusion model are in the same segment. After the ARIMA prediction results are generated, this paper fuses the model results immediately. Horizontal comparison of various methods: The fusion model method (column 3), based on ARIMA (column 2), has been optimized to varying degrees under different data set scenarios. The first two rows of data set scenarios are more complex and the optimization is relatively obvious, while the third row has little impact. Vertical comparison of each data set: As the abscissa (time series) increases, the prediction curve gradually deviates from the true trend. This is because in actual problem scenarios, researchers need to continuously update new prediction values into the sequence of time steps for extracting information. Thus, using the "predicted value" to predict the "new predicted value" is unavoidable, which will cause the error to gradually accumulate and increase.

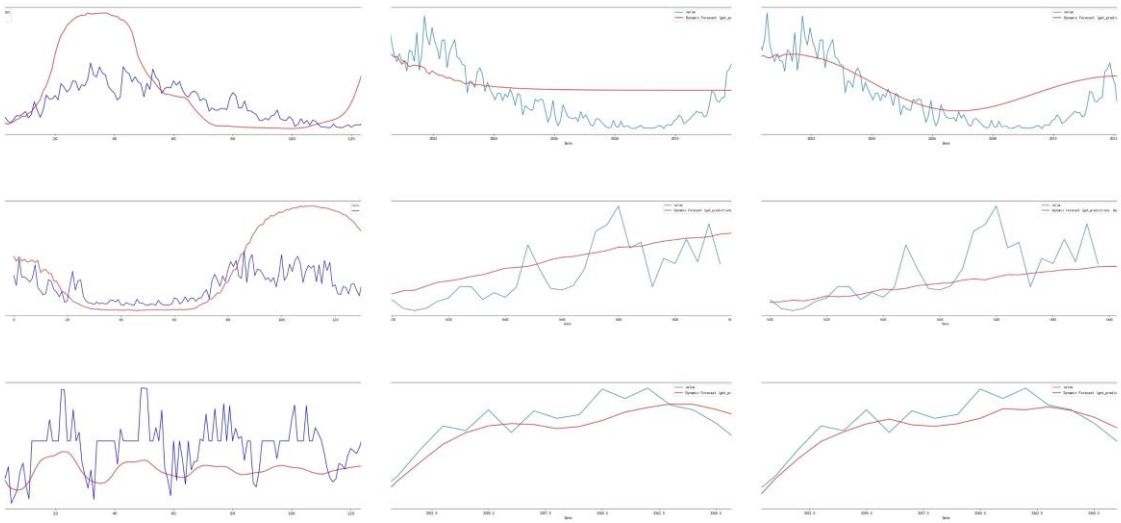

Fig. 15 Effect diagram of LSTM (column 1), ARIMA (column 2), and fusion model (column 3) under the sunspot number (row 1), sunspot area (row 2), and magnetic field (row 3) data sets

6. Parameter determination

6.1. LSTM Parameters

By changing the number of model iterations, the trend of the mean square error of the LSTM fluctuation prediction model with the number of iterations were obtained, and the results are shown in fig.16:

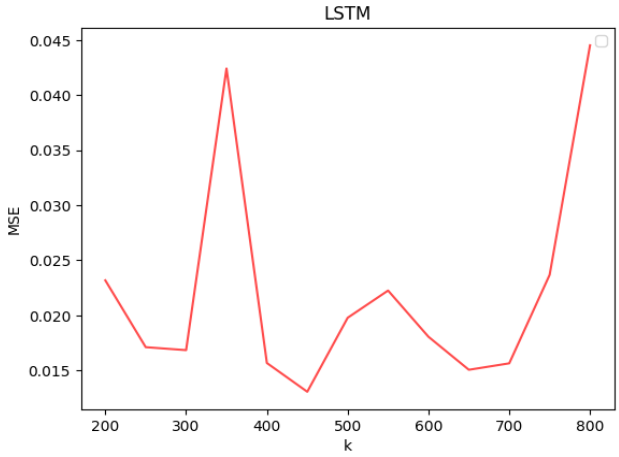

Fig. 16 The relationship between the number of iterations and MSE

It is observed that the number of iterations k corresponding to the two smaller mean square error values in the image are 450 and 650, respectively, and since the smaller error does not mean that the fluctuation trend is more pronounced, the predictions corresponding to the number of iterations at these two points were further calculated in this paper, as shown in the figure 17:

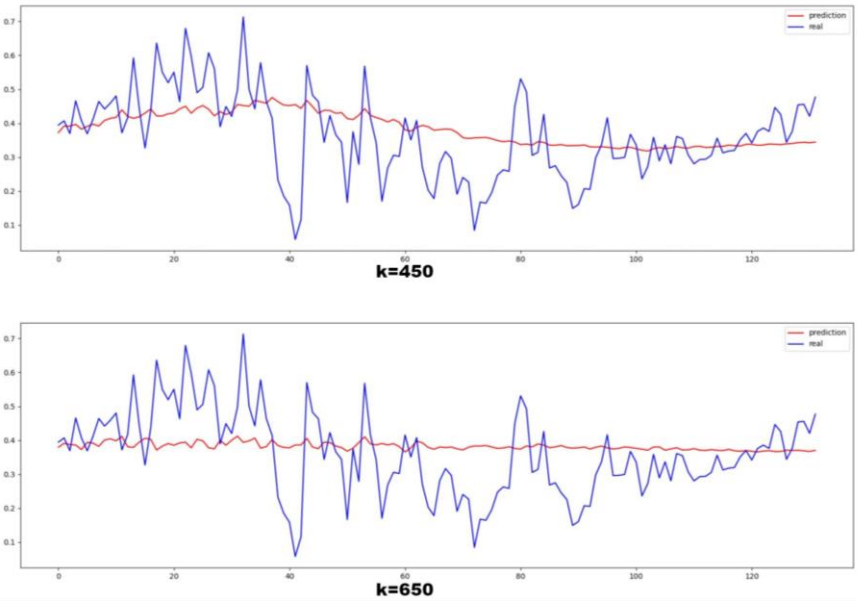

Fig. 17 Prediction effect of each iteration number

Through the comparison of the above figure, it is found that when k=450, the fluctuation trend of the prediction result is more obvious, closer to the real value, and more accurate, so this paper adopts the number of iterations k=450 as the final parameter of the model.

By changing the hidden layer size of the model， the trend of the mean square error of the LSTM fluctuation prediction model with the number of iterations were obtained, and the results are shown in the figure 18:

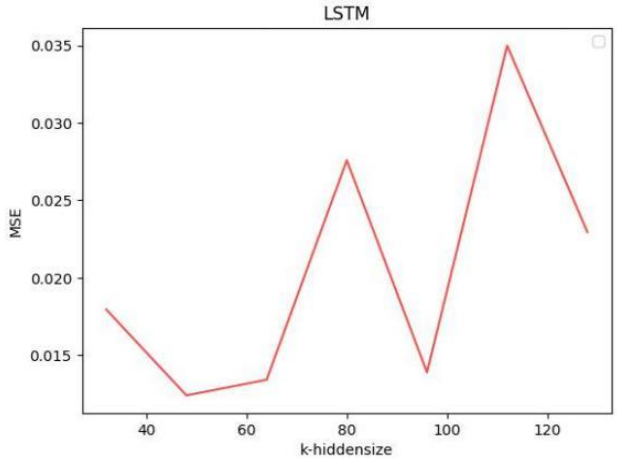

Fig. 18 Hidden layer size and MSE relationship diagram

Through observation, it can be found that the k-hidden size of the hidden layer corresponding to the three smaller mean square error values in the image are 48, 64, and 96, respectively, and since the smaller error does not mean that the fluctuation trend is more pronounced, the prediction results corresponding to the hidden layer sizes at these three points were further calculated, as shown in the following figure19:

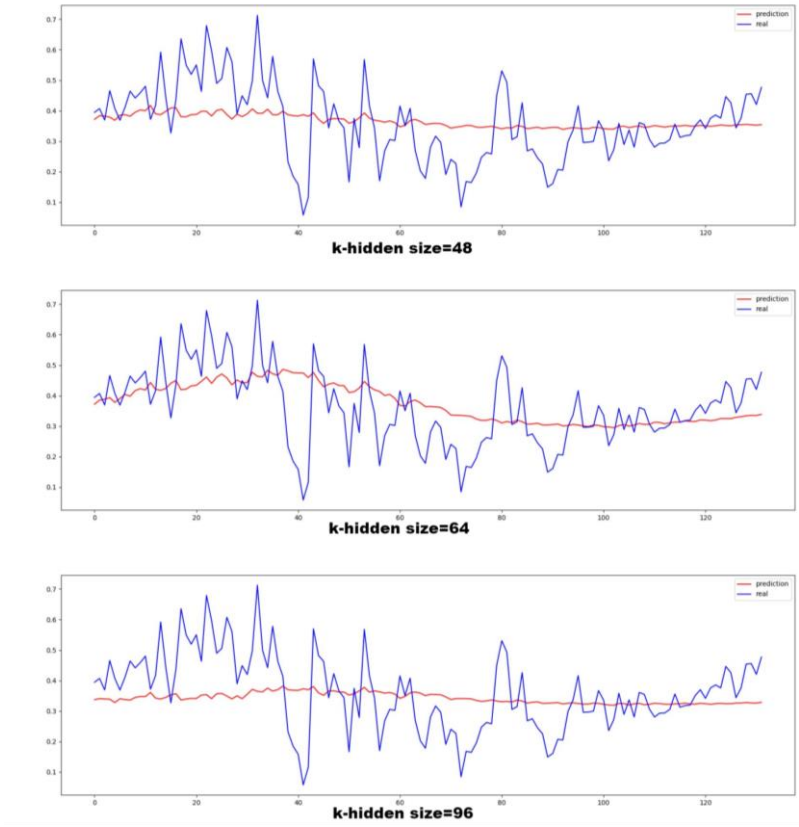

Fig. 19 Prediction results of each hidden layer size

Through the comparison of the above figure, it can be found that when k-hidden size=64, the fluctuation trend of the prediction result is more obvious, closer to the real value, and more accurate, so this paper adopts the number of iterations k-hidden size=64 as the final parameter of the model.

6.2. ARIMA Parameters

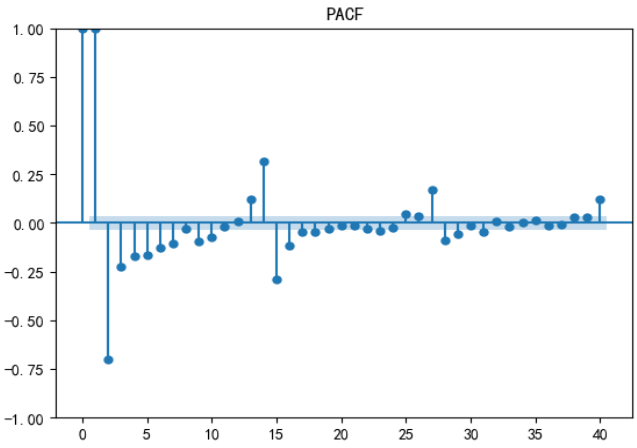

Fig. 20 PACF

By observing the figure 20, it can be found that the horizontal axis is the order of the autoregressive term p and the vertical axis is the value of the PACF. The dashed line represents the 95% confidence interval. Here lag = 40 and the maximum value is the 40th order. Different orders represent points with different lags, i.e., the correlation of the same series at different orders. Here the third order is about -0.25, i.e., it is negatively correlated with itself, and the correlation coefficient is about 0.25.

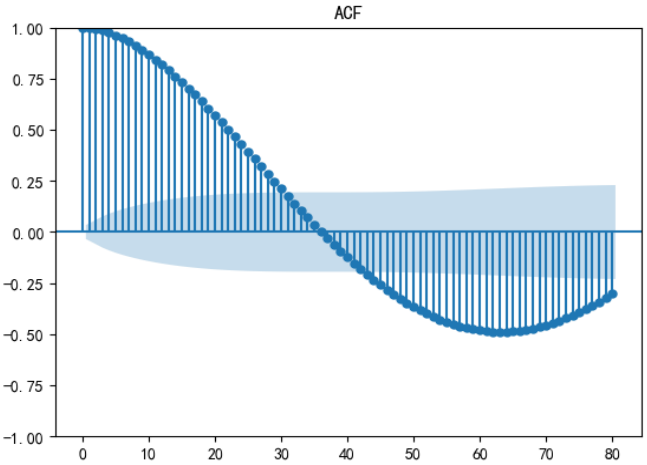

Fig. 21 ACF

By observing the figure 21, it can be found that the horizontal axis is the order of the moving average term q, and the vertical axis is the value of ACF. The dashed line represents the 95% confidence interval. Here Lag=40, the maximum is 40th order. The ACF chart shows obvious oscillation characteristics without censoring which embodies significant cycle-related information. It can be seen from the observation that when q=35, the relevant information of a quarter cycle can be roughly summarized. Since the training cost is too high and the training cycle is too long when q=65, q=35 is selected as the optimal parameter.

Tab. 1 Differential effect evaluation index

d value	ADT-result	t-statistic	delay	tests number
d=0	-9.4132625	5.7589499	28	3270
d=1	-10.1172401	9.6519675	27	3270
d=2	-12.7647541	8.0019908	28	3268

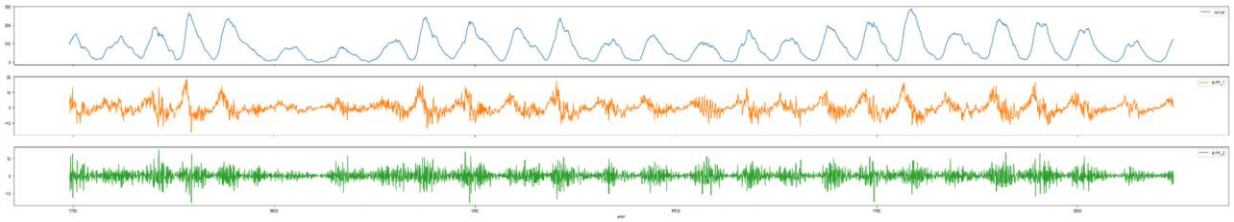

Fig. 22 Data difference renderings (from top to bottom are the original data and d=1, d=2)

The difference degree d is also an important indicator of the ARIMA model. If the original data is not stationary enough, you can find its stationary components through first difference or second difference. The greater the number of differences, the greater the inverse difference error that may result. Generally, researchers will choose the minimum difference value d that makes the data have good stability. As shown in the table above, the evaluation index ADT-statistic should be less than "-3.4324" at 1% confidence level, less than "-2.8624" at 5% level, and less than "-2.5672" at 10% level. The t-statistic needs to be less than 0.05, and the closer it is to zero, the more stable it is. After verification, after noise decomposition, the three data sets all showed good stability when d=0, so the parameter d was selected as 0.

7. Conclusions

In this paper, the original data were first analyzed and the distribution characteristics of the three types of data were analyzed. Based on the commonality of the above data distribution, ARIMA model and LSTM model are established to realize the prediction of trend time series and fluctuation time series respectively, and a reasonable fusion strategy is adopted to combine the results of the models to realize a more accurate prediction of solar activity. At the same time, two different prediction methods, namely sliding and rolling validation, are used to verify the reliability of the models, and it is concluded that the fusion strategy is more accurate than the single prediction model. Finally, this paper focuses on the method and process of determining the important parameters of each model, which further improves the credibility of the model.

Due to the complexity of the LSTM prediction model, the interpretability of the prediction results is poor. Random forests, decision trees and other more interpretable regression algorithms can be considered to further optimize the prediction of solar activity. At the same time, the data fusion method selected by the fusion model strategy proposed in this paper only considers the prediction characteristics of ARIMA and LSTM. The subsequent data fusion processing should be further optimized by considering more objective physical conditions about solar activities.

References

[1] E. W. Cliver, “Solar activity and geomagnetic storms: The first 40 years,” Eos, Transactions American Geophysical Union, vol. 75, no. 49, pp. 569–575, 1994.

[2] S. W. Samwel, E. A. El-Aziz, H. B. Garrett, A. A. Hady, M. Ibrahim, and M. Y. Amin, “Space radiation impact on smallsats during maximum and minimum solar activity,” Advances in Space Research, vol. 64, no. 1, pp. 239–251, Jul. 2019.

[3] Wei Sun, G. Li, J. Wang, Shuting Xu, Shiaoxiao Wei, and F. Liu, "Construction of magnetic polarity index for sunspot magnetic field and its periodicity analysis," Geophysical Progress, vol. 37, no. 4, pp. 1475-1483, Aug. 2022.

[4] S. B. Waykar, A. W. Kale, and N. P. Karlekar, “Deep Learning-Based Solar Activity Prediction Using Sunspot Number and Solar Radio Flux,” in Advances in Parallel Computing, D. J. Hemanth, T. N. Nguyen, J. Indumathi, and S. Lakshmanan, Eds., IOS Press, 2022.

[5] I. Krasheninnikov and S. Chumakov, “Forecasting the Sunspots Number Function in the Cycle of Solar Activity Based on the Method of Applying Artificial Neural Network,” in Artificial Intelligence Application in Networks and Systems, vol. 724, R. Silhavy and P. Silhavy, Eds., in Lecture Notes in Networks and Systems, vol. 724. Cham: Springer International Publishing, 2023, pp. 824–835.

[6] C. J. Schrijver and C. Zwaan, Solar and Stellar Magnetic Activity. in Cambridge Astrophysics. Cambridge: Cambridge University Press, 2000.

[7] G. Jensen, "Applied forecasting with an autoregressive integrated moving average (ARIMA) model," 1977.

[8] P. R. Low and E. Sakk, "Comparison between autoregressive integrated moving average and long short-term memory models for stock price prediction," IAES International Journal of Artificial Intelligence (IJ-AI), vol. 12, no. 4, Art. no. 4, Dec. 2023.

[9] A. Sherstinsky, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network," Physica D: Nonlinear Phenomena, vol. 404, p. 132306, Mar. 2020.

[10] D. S. Mohanty, "An International Study of Application of Long Short-Term Memory (LSTM) Neural Networks for the prediction of stock and forex markets," IJFMR - International Journal for Multidisciplinary Research, vol. 5, no. 3.

[11] X.-H. Le, H. V. Ho, G. Lee, and S. Jung, "Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting," Water, vol. 11, no. 7, Art. no. 7, Jul. 2019.

[12] J. Gonzalez and W. Yu, "Non-linear system modeling using LSTM neural networks," IFAC-PapersOnLine, vol. 51, no. 13, pp. 485–489, Jan. 2018.