International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 4996-5007

https://doi.org/10.70517/ijhsa463428

Using time series analysis to study the cumulative distribution function of meteorological data and selecting typical meteorological cycles for normalized modeling

Zesen Wang[1,2], Shuaihao Kong[1,2], Qi Li[1,2], Jingrong Guo[1,2], Hao Liang[1,2], Tianqi Zhao[1,2], Jiayu Ding[3] and Runfeng Zhang[3,*]

[1] North China Electric Power Research Institute Co., Ltd., Beijing, 100045, China
[2] State Grid Jibei Electric Power Research Institute, Beijing, 100052, China
[3] Nanjing Tode Technology Co., Ltd., Nanjing, Jiangsu, 210094, China

Corresponding authors: (e-mail: ZrunFfeng@163.com).

Abstract Short-term prediction of meteorological data needs to extract effective information from complex time-series features, which centers on improving data quality through preprocessing, constructing prediction models adapted to seasonal variations, and optimizing fuzzy delineation of data distributions in order to improve prediction accuracy. Based on the hour-by-hour meteorological data of an international airport in China, this study proposes a time series analysis framework that integrates data cleaning, normalization, seasonal ARIMA modeling, and cumulative distribution domain delineation, aiming to improve the accuracy of short-term meteorological forecasts. For the missing values and outliers in the raw data, the segmented linear interpolation and truncation strategies are used to reconstruct the features, and the seasonal segmentation strategies of cold, hot, and transitional seasons are combined to enhance the data integrity. The min-max normalization is used to eliminate the differences in the magnitudes of multiple sources of meteorological elements, and a seasonal ARIMA product model is constructed to capture the cyclical fluctuation pattern of the temperature data. The cumulative probability distribution method is further introduced to divide the thesis domain, and the temperature data are mapped into interpretable fuzzy intervals to optimize the model's ability to express uncertainty. The experimental results show that the method in this paper significantly outperforms the traditional RNN and LSTM models in the wind speed and temperature prediction task, in which the MAE, MSE, and MAPE of the daily maximum temperature prediction are reduced to 0.0554, 0.00604, and 20.13%, respectively, which verifies the model's utility in complex meteorological time-series features.

Index Terms time series analysis, seasonal variation, ARIMA model, meteorological data, cumulative probability distribution

I. Introduction

Time series refers to the data series obtained by sampling at a certain time interval, such as economic indicators, meteorological data, etc. [1], [2]. In time series, there often exists a certain periodicity characteristic, i.e., the data in a certain time interval will show a recurring regularity [3], [4]. In recent years, in the field of meteorology, the cumulative distribution function of meteorological data and the analysis of typical meteorological cycles based on time series analysis methods have gradually become one of the important studies [5]-[7].

Meteorological data play a vital role in atmospheric science research [8]. These data contain rich information, such as temperature, humidity, barometric pressure, wind speed, wind direction, etc. By analyzing these indicators, we can better understand the behavior of the atmosphere, predict weather changes, and study climate characteristics [9]-[11]. Typical meteorological cycle is a “virtual meteorological cycle” composed of typical weather in a certain period, and the analyzed meteorological factors are dry bulb temperature, dew point temperature, wind speed and total radiation on the horizontal surface, and the accurate typical meteorological cycle is of great significance for architectural design, solar energy technology utilization and so on [12]-[15]. The use of time series analysis to study the cumulative distribution function of meteorological data and select the typical meteorological cycle for normalized modeling can comprehensively analyze the statistical characteristics and cycle patterns of meteorological data, thus providing support for climate change prediction and disaster warning [16]-[18].

In this study, the hour-by-hour meteorological data of an international airport is taken as the research object, and a methodological framework integrating data cleaning, normalization processing, seasonal ARIMA modeling and cumulative distribution domain delineation is systematically proposed. Aiming at the problems of missing values,

outliers and feature redundancy in the raw data, the segmented linear interpolation and truncation strategies are proposed to reconstruct the wind direction and cumulative wind speed features, and the data quality is verified by statistics. The data are also divided into cold, hot, and transitional seasons to ensure data integrity by combining the bimodal distribution characteristics of air temperature. It also eliminates the magnitude differences of multi-source meteorological elements through normalization. Further, for the bimodal characteristics and cyclical pattern of temperature data, a seasonal ARIMA product model is constructed by decomposing the long-term trend, seasonal variation and residual components of the time series. The model structure is optimized by combining the difference operation with the seasonal cycle parameters to capture the cyclical fluctuation pattern of temperature data. Finally, based on the cumulative probability distribution method, the boundaries of the domain are determined by the normality test, and the temperature data are mapped into interpretable fuzzy intervals by combining the linguistic value interval division and fuzzy distance parameter adjustment, so as to enhance the ability of the prediction model to express the uncertainty.

II. Meteorological data time series preprocessing with seasonal ARIMA modeling and domain delimitation

II. A. Raw data preprocessing

II. A. 1) Data cleansing

The research topic of this paper is short-term, quantitative forecasting based on historical meteorological data, and the main task is to give hour-by-hour temperature forecasts for the next 24 hours by analyzing historical data of limited length. The meteorological data studied in this paper are observed at an international airport in China and are recorded automatically by sensors with a time resolution of 1 hour. The dataset elements and their meaningful contents are shown in Table 1.

Table 1: Dataset elements and their meanings

Name of data element	Meaning
No.	Record the time point number
year, month, day, hour	Data recording time information, 24-hour system
season	Season number
PM_(Station)	Pollution value information
DEWP	Dew point temperature (Celsius)
HUMI	Relative humidity (percentage)
PRES	Atmospheric pressure (hPa)
TEMP	Temperature (Celsius)
cbwd	Combined wind directions (NE, NW, SE, etc.)
Iws	Cumulative wind speed (meters per second)
precipitation	Precipitation per hour (millimeters per hour)
Iprec	Single cumulative precipitation (millimeters per hour)

As the data collection equipment may be down, sensor failure, data storage errors, etc., there may be missing values, abnormal values, etc., in the raw data, which do not meet the requirements of subsequent analysis. Therefore, it is necessary to check and clean the data. Since the pollution value information is not within the scope of this paper, it will be ignored in the subsequent analysis. In the dataset used in this paper, the missing values of each meteorological element data are occupied by “NA”.

The length of consecutive missing data is less than or equal to three hours more frequently, and the length of missing data is greater than three hours less often. By looking at the raw data, data with missing length less than or equal to three hours appear sporadically, and data with missing length greater than three hours are mostly concentrated on the same date. For the missing length less than or equal to three hours of data segments used linear interpolation patch method, patch calculation process as shown in equation (1).

$$x_{t+l} = x_t + \frac{x_{t+L+1} - x_t}{L+1} \times l, l \in (0, L], L \geq 1, L \in N^+ \tag{1}$$

where x_t is the data value at moment t, x_{t+l} is the missing data value at l hours from moment t, and L is the missing data length.

For temperature, dew point temperature, barometric pressure, and relative humidity data segments with missing lengths greater than three hours, their missing lengths are basically concentrated in greater than six hours and are

not numerous, and the use of patching may introduce features that are not uniform with the data, so the data will be truncated directly.

For the combined wind direction feature, since its value is symbolic direction information, its missing value will be directly made up with the wind direction feature of the previous moment. The recording process of the single cumulative wind speed information is as follows: when the wind direction changes or the wind speed changes from zero to a non-zero value, the cumulative wind speed is cleared to zero and the recording of the new hour starts; if the current wind direction remains unchanged, the cumulative wind speed of the current hour is recorded. In the raw data, due to the definition of the missing wind direction value “NA” as a change in wind direction, the data will appear “wind direction has not changed, but the cumulative wind speed suddenly changed to a smaller value”. Therefore, a differential method is used to transform the single cumulative wind speed to hourly cumulative wind speed during data processing. Hourly cumulative wind speed will be used as the element name “ws”.

Cumulative precipitation is the cumulative value of hourly precipitation, which contains repetitive information, so only the hourly precipitation is retained.

After cleaning, the statistics of the distribution of values of each meteorological data are shown in Table 2.

Table 2: Statistical information of meteorological elements

Statistical elements	DEWP	HUMI	PRES	TEMP	ws	precipitation
Mean value	2.23	54.3	1016.2	12.9	2.93	0.0662
Standard deviation	14.3	26.0	10.220	12.1	2.33	0.866
Minimum value	-40.0	2.00	991.00	-19.0	0.45	0
25% percentile	-10.0	31.0	1008.0	2.00	0.89	0
50% quantile	3.00	55.0	1016.0	14.0	1.79	0
75% quantile	15.0	78.0	1024.0	23.0	4.02	0
Maximum value	28.0	100	1046.0	42.0	20.1	69.2

The 25%, 50%, and 75% quartiles of the hourly cumulative precipitation data in the table are all zero because of the low annual precipitation in the area, with no precipitation on most of the days. It shows that there are still deficiencies in analyzing the data only from the statistical information, and it is difficult to reflect some special situations of the data.

It can be seen from the statistics that there are no outliers in the data. It shows that the data are of high quality after processing and can be further analyzed.

The forecast target of this paper is the temperature, and it can be observed that the distribution of the temperature in the data set has an obvious bimodal characteristic. This bimodal characteristic is caused by the annual average daily temperature change rule is roughly similar to the sinusoidal curve. The temperature data are decomposed into cold and hot seasons. The cold season is taken as January, February, March and December, the hot season is taken as June, July, August and September, and the transition season is taken as April, May, October and November.

It can be observed that the temperature obeys roughly normal distribution in the cool season, hot season and transition season. In the subsequent analysis, this paper defaults to temperatures obeying normal distribution in the short-term prediction to simplify the analysis process.

II. A. 2) Data normalization

The meteorological element dataset contains features such as temperature (°C), relative humidity (%), and barometric pressure (hPa), and the units of each element are different, and the range of values is also different. Depending on the subsequent analysis method chosen and the model constructed, the range of values will have an impact on the prediction performance of certain models.

For artificial neural networks, there are two advantages of data normalization:

(1) Easy information transfer within the neural network. Artificial neural networks use a large number of nonlinear activation functions, and there are dead zones in the input range of nonlinear activation functions, such as Fig. 1 shows the inputs and outputs of the sigmoid activation function and the tanh activation function. The data normalization can avoid the activation function output losing sensitivity to changes due to excessive changes in the input data value range.

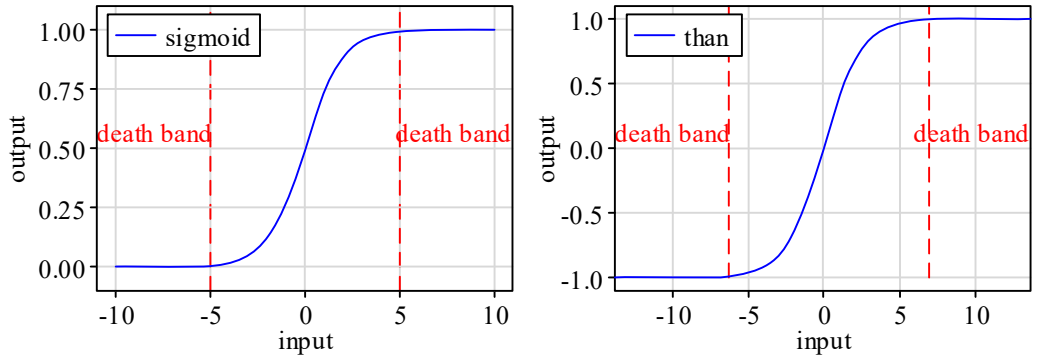

Figure 1: The input and output of the sigmoid and tanh activation functions

(2) Facilitate the network to find the optimal parameters in multi-dimensional features. At present, most of the neural networks are trained using the backward error propagation algorithm. Features in the same value domain, its related parameters in each training process to get the opportunity to update the approximation, is conducive to speeding up the training speed of the network.

As shown in the schematic fitting of the original features and normalized features in Fig. 2, for the two features x, y in the gradient descent process, compared with the original features, the normalized features have a greater speed of convergence in the training.

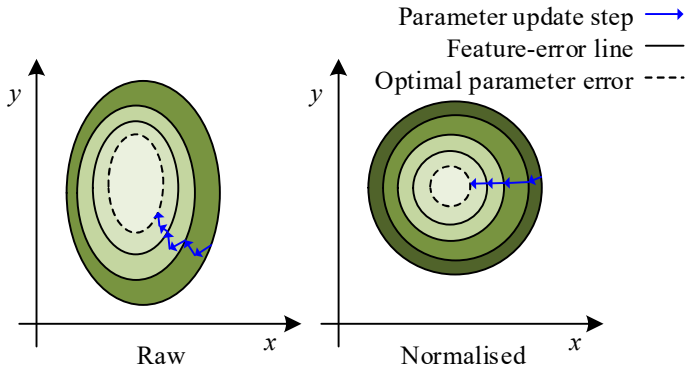

Figure 2: Illustration of fitting between original features and the normalized features

The data normalization method used in this paper is min-max normalization, and its calculation process can be described as equation (2).

$$\{x_{i,t}\}_{new} = 2 \cdot \frac{\{x_{i,t}\}_{origin} - \overline{x}_i}{x_{i,\max} - x_{i,\min}} \tag{2}$$

where $\overline{x}_i$ is the time series mean of feature i in the data set, $x_{i,\max}$ is the maximum value in the time series, and $x_{i,\min}$ is the minimum value in the time series. The distribution of the normalized meteorological data only undergoes a change in the value domain and does not change the type of distribution. The new value domain of the data is $[-1,1]$.

II. B.Construction of the time series ARIMA seasonality model

After data cleaning and normalization, the completeness and consistency of the meteorological elements are guaranteed, which provide reliable inputs for time series modeling. On this basis, it is necessary to further explore the cyclical pattern in the data, and in this section, the seasonal ARIMA model will be constructed to realize the trend-period separation of the temperature series through the difference operation and seasonal term decomposition.

II. B. 1) Seasonal overview

The data in a time series is usually caused by a combination of many different influencing factors, in other words that is, in a time series, the data usually contains many different influencing factors. Here we assume that X represents the original data in the time series, then the combined model determined by multiple factors can be expressed as (3).

$$\begin{aligned}&\text{Additive modelling: } X = T + S + C + I\\&\text{Multiplicative models: } X = TSCI\end{aligned} \tag{3}$$

T long-term trend, S seasonal variation component (cyclical), C cyclical variation component, and I irregularity factor (residual).

(1) In the additive model, the effects of the four factors are assumed to be independent of each other, and each component is expressed in absolute terms;

(2) In the multiplicative model, the relationship between the influencing factors is more complex compared to that of the additive model. It is assumed that the four factors are influential on the development of the phenomenon and they are not independent of each other, the absolute amount of T is regarded as the base amount, and the other three factors are expressed as ratios.

A seasonal stochastic time series interval is a relatively strong correlation between random variables at two points in time with a period length of m, or a seasonal time series exhibiting periodic correlation. For example, for quarterly data, $m = 4$, X_t is correlated with X_{t-4}, and so it is possible to utilize this inter-periodic correlation for the fit. Similarly, monthly data can be represented here as a 12-month period, and weekly data can be set to either 7 or 5 days. This subsection focuses on the characteristics of seasonal data and the points to be noted when forecasting data containing seasonality.

For the modeling process of time series with seasonality (periodicity), the seasonal effect can be realized by seasonal difference and product seasonal term, which is the difference operation between the current period's series value X_t and the last year's series value of the same period X_{t-m} (m stands for the period value of each season), also known as the m-step difference, and the seasonal effect of the time series can only be realized by seasonal The seasonal effect of a time series can only be directly portrayed by the seasonal difference or the autoregressive or moving average term of the m-period, which is called the seasonal model.

The general expression of the seasonal model is:

$$\nabla^d \nabla_m^D X_t = \frac{\varphi(B)}{\phi(B)} \varepsilon_t \tag{4}$$

$$\varphi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_q B^q = 1 - \sum_{i=1}^{q} \varphi_i B^i \tag{5}$$

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p = 1 - \sum_{i=1}^{p} \phi_i B^i \tag{6}$$

where $\{\varepsilon_t\}$ is Gaussian white noise and $\nabla^d \nabla_m^D X_t$ denotes that the time series $\{X_t\}$ is smooth after d th order differencing and D th m-step differencing. $\varphi(B)$ autoregressive coefficient polynomial, $\phi(B)$ sliding average coefficient polynomial.

If the time series $\{X_r\}$ contains a stochastic trend component, in order to remove the stochastic trendiness, Box and Jenkins came up with the ARIMA autoregressive summation sliding average model. In simple terms, this model is to obtain a stationary time series after the d order simple integer sequence undergoes d order differences, and then construct the $ARMA(p,q)$ model, denoted as the $ARIMA(p,d,q)$ model. Here, d represents the order of the difference, p represents the autoregressive order, and q represents the moving average order. It is a kind of time series forecasting analysis model established by considering the relationship between the dependent variable and its own lagged value on one hand, and the relationship between the present value and the lagged value in the random error on the other hand, in the process of stabilizing the non-stationary time series.

II. B. 2) Fundamentals of ARIMA modeling

The ARIMA model is a model that uses autocorrelation between time and describes this through a mathematical model that characterizes the continuity of the development of the object of prediction by forecasting the future values of the series based on the past and present values of the time series.

$ARIMA(p,d,q)$ The expression of the model is (7).

$$\left(1-\sum_{i=1}^{p}\varphi_i B^i\right)(1-B)^d X_t=\left(1-\sum_{i=1}^{q}\phi_i B^i\right)\varepsilon_t \tag{7}$$

The $ARIMA(p,d,q)$ model is a typical time series forecasting model, which is an extension of the ARMA model, and can be viewed as consisting of three parts: the AR model, the I-differential, and the MA model.

$ARIMA(p,d,q)$ in which p is denoted as the autoregressive number of the AR "autoregression"; q is denoted as the number of sliding average terms for the MA for, d is the number of differences, also called the order, that need to be done to make the unsteady time series a smooth one, B is the lag operator, and $d\in Z, d>0$.

Let $\{x_t, t=0,1,\cdots\}$ denote a random time series if there exists a nonnegative integer d such that (8) holds.

$$\nabla^d x_t = X_t \tag{8}$$

Substituting Eqs. (5) and (6) into (7), and the difference operator $\nabla=1-B$, so (7) is converted to the form (9), and $\{\varepsilon_t\}$ is a white noise sequence.

$$\varphi(B)X_t=\phi(B)\varepsilon_t \tag{9}$$

If $\{x_t, t=0,1,\cdots\}$ satisfies the stochastic equation (7), then $\{x_t, t=0,1,\cdots\}$ is considered to be an autoregressive summed sliding average series denoted as an $ARIMA(p,d,q)$ model.

II. B. 3) ARIMA's seasonal modeling

The principle of ARIMA model lies in the fact that non-stationary time series are smoothed, and in the process of smoothing, the lagged values of the dependent variable as well as the random error term are done regression analysis and modeled.

Seasonality model, a series of changes in the seasonality of the time series with seasonality, specifically, the estimation and elimination of seasonality in the event series data. The purpose is mainly to better reflect the cyclical characteristics of the original series (quarterly, monthly, weekly). Comparing the data with the seasonal factor removed with the original, it is found that the data after removing the influence of seasonality can show the following two advantages: first, it can reflect the trend of the original time series itself more clearly. By using scientific and effective methods to measure, separate and adjust the seasonal factors in the original data, the trend of the data itself can be revealed. Secondly, the data is comparable because the data after removing seasonal factors eliminates seasonal disturbances, so the data in different periods can be directly used for comparison.

From the time series plot of the raw data, it is not difficult to find that both precipitation and wind speed have very obvious seasonal components. For time series with seasonal components, there are two models used to separate the seasonal factors in most cases: the additive decomposition model and the multiplicative decomposition model. The multiplicative decomposition model is used for time series that show a trend with the seasons, while the additive decomposition model is used for time series with a relatively consistent trend. The multiplicative model is used in this thesis.

The ARIMA multiplicative seasonal model is built in three steps as shown in Figure 3:

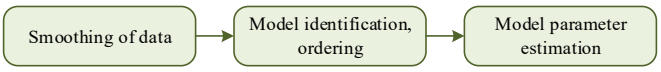

Figure 3: Model establishment steps

The ARIMA multiplicative seasonal model has good generalization in time series forecasting and is a combination of stochastic seasonal model and ARIMA. Assuming a certain time series X_t, the ARIMA model is denoted as:

$$\varphi(B)\nabla^d X_t=\phi(B)\varepsilon_t \tag{10}$$

X_t denotes the time series, B is the backward shift operator, $\nabla = 1 - B$, p, d, and q denote the autoregressive order, difference order, and moving average order, respectively; $\varphi(B)$ denotes the autoregressive operator; and $\phi(B)$ denotes the sliding average operator.

A product seasonal model of order $(p,d,q)(P,D,Q)_m$ can be expressed as (11).

$$\varphi(B)\Phi(B^m)\nabla^d\nabla_m^D X_t = \Theta(B^m)\phi(B)\varepsilon_t \tag{11}$$

where P, Q are the orders of the autoregressive model and the order of the moving average model for time series with seasonality, D is the order of the seasonal difference, m is the seasonal period, and ε_t is the random error.

$$\begin{aligned}
\varphi(B) &= 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_p B^p \\
\phi(B) &= 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_q B^q \\
\Phi(B^m) &= 1 - \Phi_1 B^m - \Phi_2 B^{2m} - \cdots - \Phi_P B^{Pm} \\
\Theta(B^m) &= 1 - \Theta_1 B^m - \Theta_2 B^{2m} - \cdots - \Theta_Q B^{Qm}
\end{aligned} \tag{12}$$

$\Phi(B^m)$ is the product seasonal autoregressive coefficient polynomial and $\Theta(B^m)$ is the product seasonal sliding average coefficient polynomial.

II. C.Define and delimit the thesis domain

Although the ARIMA model can effectively capture the linear trend and seasonal fluctuation of the series, there are still limitations in the fuzzy representation of the distributional properties of the data. For this reason, this section introduces the cumulative probability distribution method, which transforms the statistical properties of temperature into a fuzzy rule base by defining the thesis domain and dividing the linguistic value intervals to provide interpretable semantic support for the subsequent prediction models.

In the research on the problem of dividing the domain, the cumulative probability distribution method is a popular fuzzy division method, which has been successfully used in many aspects. If it is applied to the fuzzy delineation of domain intervals in the intuitionistic fuzzy time series prediction model, it can not only obtain better intuitionistic fuzzy numbers, i.e., linguistic value intervals, but also improve the accuracy of prediction. The method is specifically realized through the following four steps.

II. C. 1) Testing the Normal Distribution

The experimental dataset must be estimated using the critical value method, Equation (13), to estimate whether the sample data obey a normal distribution. The cumulative probability distribution method requires that all observations in the dataset are approximately normally distributed, e.g. Divore's simulation experiment provides convincing evidence that a sample size of 30 is sufficient to overcome the overall skewness and give an approximately normal distribution.

$$z = \frac{\bar{x} - x_i}{\sigma / \sqrt{n}} \tag{13}$$

where $\bar{x}$ is the mean of the sequence history data, x_i is the sequence history data, σ is the standard deviation of the sequence history data, and n is the number of sequence history data.

II. C. 2) Defining the Theory X

Assuming that X is the domain of the argument and $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the sequence history data, respectively, then

$$X = \left[x_{\min} - \sigma_1, x_{\min} + \sigma_2\right] \tag{14}$$

where σ_1 and σ_2 are suitable positive integers.

II. C. 3) Determining the number of subintervals

Based on the actual meaning of the sequence history data and the uncertainty that exists in the way people deal with real problems, the domain X is partitioned in a way that can be expressed in natural language. That is, it is assumed that the number of subintervals divided is

$$k=\left[\frac{(n-1)\left|x_{\max}-x_{\min}\right|}{\left(\sum_{i=1}^{n}\left|X_{i+1}-X_i\right|\right)^p}\right] \tag{15}$$

where n is the dimension of the sample data set; "[ ]" is the upward and downward rounding operation; p is the distance parameter adjustment factor; and when the parameter $p=1$, k is a constant, and the fuzzy distance between the sequence data points is kept unchanged; when $p<1$, k decreases gradually, and the fuzzy distance between the sequence data points is decreases; when $p>1$, k gradually increases, and the fuzzy distance between sequence data points is enlarged. It can be seen that k can be obtained by adjusting the distance parameter p.

II. C. 4) Determining the length of the language value interval

In this step, the fuzzy partitioning method of cumulative probability distribution is used to determine the length of each argument interval. The method has the following 2 sub-steps.

(1) Calculate the cumulative probability of the upper and lower bounds of each language value interval. For each language value interval, the cumulative probability of the lower bound $\left(P_{LB}^{i}\right)$ and the cumulative probability of the upper bound $\left(P_{UB}^{i}\right)$ are computed with the help of equation (16):

$$\begin{cases} P_{LB}^{1}=1-\sum_{i=1}^{n}P_{LB}^{i}, i=1 \\ P_{LB}^{i}=\dfrac{2i-3}{2n}, 2\le i\le n \\ P_{UB}^{j}=\dfrac{j}{n}, 1\le j\le n \end{cases} \tag{16}$$

where $P\in[0,1]$, n is the number of language-valued intervals, and $i=j$ is the order of language-valued intervals.

(2) Calculate the upper and lower bounds of each language-valued interval. The upper and lower bounds of the language value intervals are obtained in this sub-step by inverting the normal cumulative distribution function (8):

$$p=F(x\mid\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}-\frac{(t-\mu)^2}{2\sigma^2}dt \tag{17}$$

where μ denotes the mean of the serial data and σ denotes the standard deviation of the serial data.

III. Empirical analysis of ARIMA-based seasonal modeling and forecasting of meteorological data

In Chapter 2, a complete time series analysis framework is formed through the preprocessing of meteorological data, the construction of seasonal ARIMA model, and the introduction of the cumulative distribution thesis division method. In order to further verify the effectiveness of the framework, this chapter will combine the simulation experiments and actual meteorological data to carry out an empirical study from the three levels of model parameter optimization, prediction performance comparison, and application scenario expansion, and systematically evaluate the robustness and practicality of the method.

III. A. Simulation experiment

III. A. 1) Experimental setup

According to the data normalization process in section 2.1, then set the learning rate of the model, the number of iterations and other parameters, because the output of the wind speed data, so the output layer dimension is set to 1. The more the number of neurons, the better the prediction effect of the model, but the number of neurons, the model of the training time is long, easy to overfitting phenomenon, in order to balance the model prediction effect and the length of the training time, to avoid the problem of overfitting In order to balance the model prediction effect with the training time and avoid the overfitting problem, the number of neurons in the time series ARIMA is set to 50, the proportion of abandoned neurons in the Dropout layer is set to 0.2, and the output dimension of the Dense layer is set to 1.

In order to verify the effectiveness of the model, 2000 wind speed data of this field in February 2024 were firstly screened from the automatic weather station to carry out the experimental analysis. This data records the wind speed information of this field every 10 minutes, and the data is divided into 1 part, the first 1600 data are used for training data, and the last 400 data are used for testing data, the strategy used for data prediction is to predict xi+31 using [xi…xi+30], where xi represents the wind speed data at the ith moment.

III. A. 2) Model wind speed forecast analysis

The decomposed wind speed series are predicted using the ARIMA seasonal model, and the prediction results at the ya4, yb4 and yb2 layers are shown in Figs. 4, 5 and 6, respectively. In the ya4 and yb4 layers, the model predicted data basically fit the real data, and in the yb2 layer, although the model predicted results can reflect the trend of the real value changes, the prediction results at each inflection point are smaller than the real situation. As the frequency increases, the gap between the prediction results at the inflection points and the true value further increases. Therefore, for the data in the low-frequency region, the model is able to better predict the transformation trend of the data, and for the data in the high-frequency region, as the frequency rises, the prediction results of the ARIMA model are not good, and they cannot reflect the intrinsic relationship of the original data, which directly affects the accuracy of the prediction results.

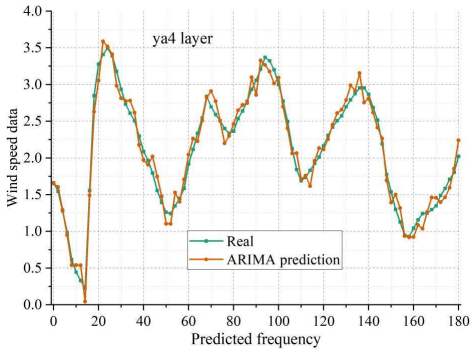

Figure 4: The prediction results of ARIMA at the ya4 layer

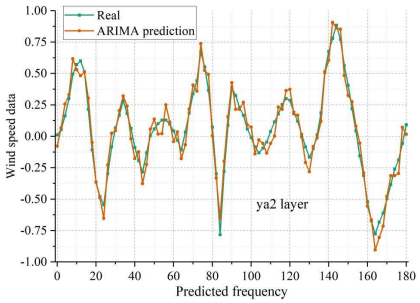

Figure 5: The prediction results of ARIMA at the yb4 layer

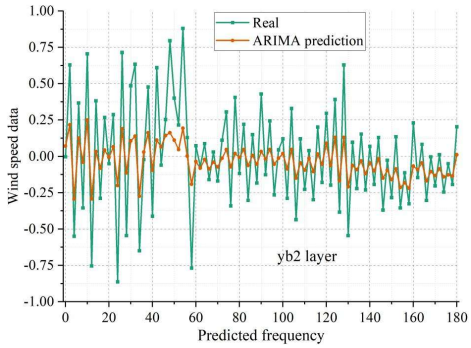

Figure 6: The prediction results of ARIMA at the yb2 layer

III. B. Research on the application of time series analysis in meteorology

After completing the simulation experiments on wind speed data, in order to further validate the universality of the time series analysis method in meteorological forecasting, this section extends the research object to the monthly average air temperature data, and comprehensively evaluates the practical application effect of the method through data preprocessing, periodicity elimination and model evaluation.

III. B. 1) Data pre-processing

In this paper, the monthly average temperature data of a region in China from January 2010 to December 2024 are collected, after doing data cleaning and processing. The monthly average temperature data from 2010 to 2019 are also selected as training samples, and the data from 2021-2024 are used as test samples; according to the above conventions, the time series of monthly average temperature from 2010 to 2020 are firstly plotted as shown in Fig. 7.

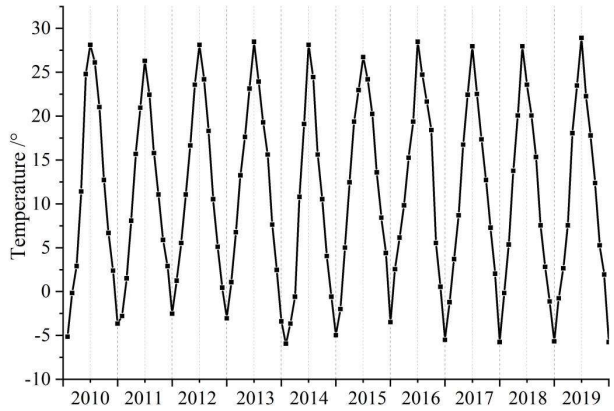

Figure 7: The monthly average temperature time series chart of from 2010 to 2019

From the above figure, it can be seen that the temperature data have no obvious upward or downward trend, and show obvious periodicity, the period of 12 months, consistent with the actual situation. It can be seen that the sequence is a non-stationary sequence, the need to first carry out the series for the smoothing process: first of all, the original sequence to do 12-step difference, to eliminate the impact of the seasonal effect. The trend after eliminating the periodic effect is shown in Figure 8.

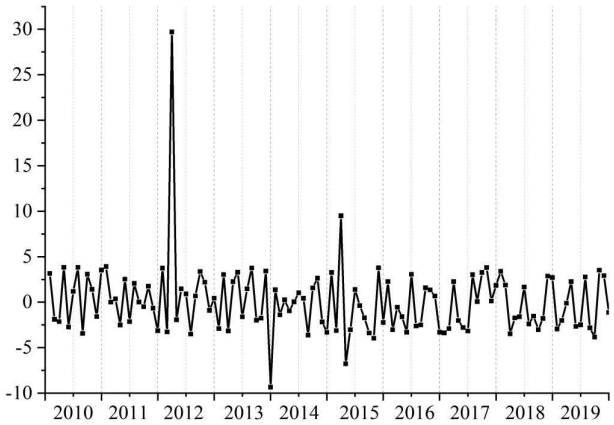

Figure 8: The trend after eliminating the periodic effect

Observing the image, we perform a 12-step differencing of the original sequence, and we find that the differenced timing diagram is essentially smooth except for a few individual points.

III. B. 2) Parameter setting

The computer hardware configuration and machine learning network parameters were set as described previously. The software programming language used for the experimental data simulation is Python, which has the Keras

package specialized for deep neural network learning as well as a variety of data preprocessing and drawing image packages. Simple to use, easy to learn and other characteristics, beginners and research scholars are widely used.

Specific network settings for the number of iterations epoch for 50/100 times, activation function activation for relu, the number of samples selected for a training batch_size for 32, optimizer for Adam = 0.0001, the size of the hidden layer inside the unit unint for 32, fully connected layer sense for 1;

III. B. 3) Comparative experimental analysis

The experiment is mainly based on the long and short-term memory neural network and self-attention mechanism and long and short-term memory neural network fusion of two kinds of machine learning to compare and analyze the figure; in addition to using three evaluation indexes Mean Absolute Error (MAE), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE) for the comparative analysis of meteorological data, multidisciplinary application scenarios to verify the time series analysis method proposed in this paper The effectiveness of ARIMA model is verified.

Next, the algorithms of the three neural networks are visually compared to verify the accuracy effect of the prediction for time series data analysis.

The daily maximum temperature of the region RNN, LSTM, this paper's method corresponding to the evaluation of the index value shown in Table 3.

Table 3: The corresponding evaluation index values of the daily maximum temperature

	MAE	MSE	MAPE
RNN	0.06785	0.00691	23.9441
LSTM	0.06032	0.00708	26.0143
OURS	0.05540	0.00604	20.1264

Table 3 demonstrates the performance comparison of the three models in the daily maximum temperature prediction task. From the data results, the method proposed in this paper outperforms the traditional RNN and LSTM models in all three metrics. Specifically, the MAE of this paper's method is 0.05540, which is about 18.3% and 8.2% lower than 0.06785 of RNN and 0.06032 of LSTM, respectively; the MSE is 0.00604, which is 12.6% and 14.7% lower than 0.00691 of RNN and 0.00708 of LSTM, respectively; and the MAPE is 20.1264%, which is also lower than 23.9441% for RNN and 26.0143% for LSTM. This indicates that the method in this paper significantly improves the prediction accuracy by integrating the seasonal modeling of time series with the fuzzy theory domain delineation strategy, especially in reducing the error magnitude while enhancing the model's adaptability to complex meteorological fluctuations. In addition, the optimization of MAPE further validates the ability of the method to control the relative error in practical applications, providing more reliable quantitative support for meteorological forecasting.

Figure 9 shows the histogram of the corresponding evaluation index values of RNN, LSTM, and this paper's method for the daily maximum temperature in the region, which can more intuitively see the prediction accuracy and superiority of this paper's method.

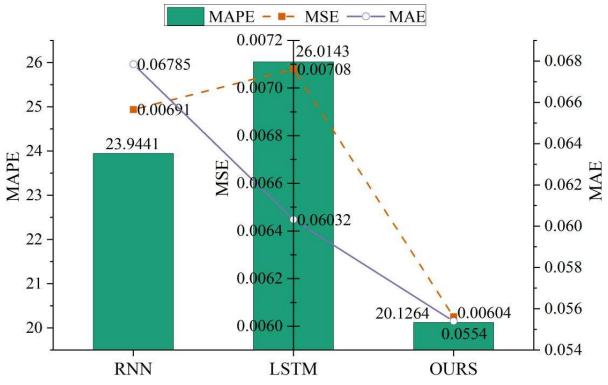

Figure 9: The corresponding evaluation index value of the daily maximum temperature

IV. Conclusion

This study experimentally verifies the effectiveness of fusing seasonal ARIMA modeling with cumulative distributional domain delineation method in meteorological data prediction. The experiments for wind speed data

show that for the data in the low-frequency region, the model is able to predict the transformation trend of the data better, with an accuracy of more than 95%, but the model has errors in predicting the inflection point in the high-frequency region, such as the yb2 layer, and the predicted value is 15% lower than the true value on average, which reveals the model's insufficiency of modeling the high-frequency components.

In the daily maximum temperature prediction task, the MAE=0.0554, MSE=0.00604, and MAPE=20.13% of this paper's method compared to the MAE=0.06785, MSE=0.00691, and MAPE=23.94% of the traditional RNN and the MAE=0.06032, MSE=0.00708, and MAPE=26.01% of the LSTM, with significant prediction accuracy. 26.01%, the prediction accuracy is significantly improved, and the MAE, MSE, and MAPE are reduced by 18.3%, 12.6%, and 16.1%, respectively. This indicates that the seasonal ARIMA model successfully captures the cyclical fluctuation pattern of temperature (with a period of 12 months), while the cumulative probability distribution method enhances the model's ability to express uncertainty through fuzzy interval partitioning.

Funding

Technology for Continuous Simulation and Quantitative Evaluation of the Adequacy of Supply-Regulation Capability in New Power Systems across Multiple Scenarios (KJZ2023117).

References

[1] Chatfield, C., & Xing, H. (2019). The analysis of time series: an introduction with R. Chapman and hall/CRC.
[2] Fulcher, B. D., Little, M. A., & Jones, N. S. (2013). Highly comparative time-series analysis: the empirical structure of time series and their methods. Journal of the Royal Society Interface, 10(83), 20130048.
[3] Osmanoğlu, B., Sunar, F., Wdowinski, S., & Cabral-Cano, E. (2016). Time series analysis of InSAR data: Methods and trends. Isprs journal of photogrammetry and remote sensing, 115, 90-102.
[4] Fulcher, B. D. (2018). Feature-based time-series analysis. In Feature engineering for machine learning and data analytics (pp. 87-116). CRC press.
[5] Mudelsee, M. (2019). Trend analysis of climate time series: A review of methods. Earth-science reviews, 190, 310-322.
[6] Baranowski, P., Krzyszczak, J., Slawinski, C., Hoffmann, H., Kozyra, J., Nieróbca, A., ... & Gluza, A. (2015). Multifractal analysis of meteorological time series to assess climate impacts. Climate Research, 65, 39-52.
[7] Murat, M., Malinowska, I., Gos, M., & Krzyszczak, J. (2018). Forecasting daily meteorological time series using ARIMA and regression models. International agrophysics, 32(2)
[8] Colston, J. M., Ahmed, T., Mahopo, C., Kang, G., Kosek, M., de Sousa Junior, F., ... & Network, T. M. E. (2018). Evaluating meteorological data from weather stations, and from satellites and global models for a multi-site epidemiological study. Environmental research, 165, 91-109.
[9] Ukhurebor, K. E., Azi, S. O., Aigbe, U. O., Onyancha, R. B., & Emegha, J. O. (2020). Analyzing the uncertainties between reanalysis meteorological data and ground measured meteorological data. Measurement, 165, 108110.
[10] Dawson, A. (2016). eofs: A library for EOF analysis of meteorological, oceanographic, and climate data. Journal of Open Research Software, 4(1).
[11] Greene, C. A., Thirumalai, K., Kearney, K. A., Delgado, J. M., Schwanghart, W., Wolfenbarger, N. S., ... & Blankenship, D. D. (2019). The climate data toolbox for MATLAB. Geochemistry, Geophysics, Geosystems, 20(7), 3774-3781.
[12] Fink, A. H., Engel, T., Ermert, V., Van Der Linden, R., Schneidewind, M., Redl, R., ... & Janicot, S. (2017). Mean climate and seasonal cycle. Meteorology of tropical West Africa: The forecasters' handbook, 1-39.
[13] Daniel, J. S., Portmann, R. W., Solomon, S., & Murphy, D. M. (2012). Identifying weekly cycles in meteorological variables: The importance of an appropriate statistical analysis. Journal of Geophysical Research: Atmospheres, 117(D13).
[14] Zang, H., Wang, M., Huang, J., Wei, Z., & Sun, G. (2016). A hybrid method for generation of typical meteorological years for different climates of China. Energies, 9(12), 1094.
[15] Cano, A., Pardo, J. J., Montero, J., & Domínguez, A. (2022). Determining Irrigation Requirements of Extensive Crops Using the Typical Meteorological Year Adjusted to the Growing Cycle Period. Agronomy, 12(9), 2208.
[16] Şen, Z. (2024). Moving trend analysis methodology for hydro-meteorology time series dynamic assessment. Water Resources Management, 38(11), 4415-4429.
[17] Faisal, A. F., Rahman, A., Habib, M. T. M., Siddique, A. H., Hasan, M., & Khan, M. M. (2022). Neural networks based multivariate time series forecasting of solar radiation using meteorological data of different cities of Bangladesh. Results in Engineering, 13, 100365.
[18] Krzyszczak, J., Baranowski, P., Zubik, M., & Hoffmann, H. (2017). Temporal scale influence on multifractal properties of agro-meteorological time series. Agricultural and Forest Meteorology, 239, 223-235.