Proceedings of the 5th International Conference on Materials Chemistry and Environmental Engineering
DOI: 10.54254/2755-2721/129/2025.20299

Drought Prediction of Henan Province in China

Jiahao Liu[1,a,*]

[1]*Beijing National Day School. Beijing, 100039, China*
a. liujiahao12789@163.com
**corresponding author*

Abstract: Henan province is a key agricultural region in China, influenced strongly by sudden drought. There are relatively few methods to help farmers to make drought projections for the future. The paper tried to use previous monthly SPEI data of time series from 12 stations in Henan from 1978 to 2018 to achieve this purpose. The paper applied Seasonal ARIMA (Autoregressive Integrated Moving Average) to capture patterns of time series and forecast future 5-years value from 2019 to 2024. The terms of models including autoregressive (p), integrated (d), moving average (q) were found based on different tests in order to construct models by removing trends and seasonality of datasets. The final results are ARIMA models of 12 stations in Henan as well as future 5-years predictions of each station. Models effectively capture the main trends and patterns of time-series data; however, some complex patterns and hidden laws still exist in datasets, influencing the accuracy of forecasting. The paper proves Seasonal ARIMA and studying information from past data of SPEI can be useful for preparing drought in Henan. However, the progress in improving the accuracy of forecasting needs to be continued in the future.

Keywords: ARIMA, SPEI, Drought, Henan

1. Introduction

1.1. Basic introduction

Henan province is a key agricultural region in China, contributing approximately 10% of the nation's grain output. The National Bureau of Statistics estimates that summer grain production in Henan will reach 37.857 million tons in 2024. Its agricultural performance is crucial for feeding hundreds of millions. However, drought poses a significant threat to grain production, with history showing frequent drought and flood events that are hard for local farmers to predict. Such disasters lead to substantial agricultural losses, negatively impacting grain supply and farmers' livelihoods. This research utilizes Standardized Precipitation Evapotranspiration Index (SPEI) data from previous years to predict drought occurrences in Henan. Given the data's seasonality and time series nature, a Seasonal Autoregressive Integrated Moving Average Model (SARIMA) is employed to model and forecast future SPEI.

The Standardized Precipitation Evapotranspiration Index (SPEI), introduced by Vicente-Serrano, Beguería, and López-Moreno in 2010, is a multiscalar drought index derived from the difference between precipitation and potential evapotranspiration (PET) across various timescales. The resulting standardized value indicates drought (negative values) or wet conditions (positive values).

DOI: 10.54254/2755-2721/129/2025.20299

SPEI incorporates temperature variability and extremes, making it advantageous for analyzing and monitoring drought across different climates, relative to the Standardized Precipitation Index (SPI). The SPEI values categorize conditions as follows: ⩾2 extreme wet, 1.5-1.99 very wet, 1-1.49 moderate wet, -0.99 to 0.99 normal, -1 to -1.49 moderate dry, -1.5 to -1.99 very dry, and ⩽-2 extreme dry[1].

The Autoregressive Integrated Moving Average (ARIMA) model, introduced by Box and Jenkins in 1970, is utilized for the analysis and modeling of time series data.[2] The ARIMA model posits that historical data and errors within a time series can be leveraged to capture data trends and forecast future values. For a more comprehensive understanding, this model can be subdivided into three components: AR (Autoregressive), I (Integrated), and MA (Moving Average). Essentially, it seeks to unify the AR and MA models to enhance projection accuracy.[3,4] The Seasonal ARIMA model incorporates the periodic nature of the dataset, thereby mitigating the impact of seasonality on the fitting outcomes.[5]

2. Constructing SARIMA model

2.1. Data preprocessing and finding seasonality

The research utilizes the SPEI dataset from the National Cryosphere Desert Data Center, encompassing multi-scale daily SPEI data for drought assessment across mainland China from 1961 to 2018. It focuses on 6-month scale data from 1978 to 2018 at twelve stations in Henan province, including Xinxiang, Anyang, Sanmenxia, Zhengzhou, Xuchang, Kaifeng, Xixia, Nanyang, Baofeng, Xihua, Shangqiu, and Gushi.[6] The selected timeframe is justified by: 1. missing SPEI data from 1968 to 1977; 2. earlier data may obscure recent trends and skew model training outcomes; 3. the 6-month scale SPEI effectively captures drought trends and seasonal fluctuations for enhanced forecasting.

The data set is collection of the daily SPEI of 12 stations from 1978 to 2018 in Henan province. However, there are some problems if the research uses these data to do predictions: (1)Daily data is highly susceptible to many insignificant factors that produce short-term fluctuations or noise, which can affect the stability of the data and the overall accuracy of the forecast model;(2)It's more easy for monthly SPEI data to reflect seasonality and climate long-term trend of the data set to make the model become more exact; (3)Data recorded in days with a large number of data points increases the computational load of the overall model, which is suitable for short-term prediction. For these reasons, the research changes daily data set to monthly data set.

In order to find P,D, and Q for seasonal ARIMA, the research has to find the seasonality that is removed from the data set. The research decides to use m = 12 to remove the seasonality of the data set. (The period of meteorological data or index typically is a year)

2.2. Trend test and stationary test to find integrated(d)

After getting monthly SPEI time-series data of 12 stations, the next step to construct the model is to find Integrated or differencing(d) in ARIMA(p, d, q). The order of d is related to whether values from previous years are stable. An appropriate differencing(d) is able to stabilize the series of values and minimize the standard deviation in order to construct an accurate ARIMA model and get more precise predictions. For the purpose of testing stationary of data set, the research decides to use two methods.

In order to test whether these data have a total long-term trend from 1978-2018, the research decides to use P value test of Augmented Dickey Fuller test and P value test of . Mann-Kendall test is a statistical test that always be used to access if time series data is stationary and whether it has

DOI: 10.54254/2755-2721/129/2025.20299

long-term trend. Augmented Dickey Fuller test (ADF) is used to examine stability of time series data further with P value test. The ADF test is essential method for time series data analysis when preparing data for model. The null hypothesis is that the time series data has a unit root that means data is non-stationary.[7] If P value is lower 0.05, then H0 is refused and these data are stationary; otherwise, time-series values are not stationary when P value is larger than 0.05. Mann-Kendall test(MK) is designed to identify whether there is a monotonic increasing or decreasing trend in the data collections. The null hypothesis(H0) is that there is no trend in the data.[8] If P value is less than 0.05, then H0 is refused and the data set has a monotonic trend, and vice versa. Following table shows P value of MK test and ADF test of data sets of 12 stations in Henan Province:

Table 1: ADF Test and Mann-Kendall Test of SPEI Data from 12 Stations in Henan Province

Name of Station	ADF Test (P value)	Stability	Mann-Kendall Test (P value)	Trend
Anyang	6.68×10-5	Stationary	0.00	Increasing
Xinxiang	3.23×10-7	Stationary	8.75×10-9	Increasing
Sanmenxia	6.68×10-5	Stationary	0.68	No Trend
Zhengzhou	8.13×10-5	Stationary	8.17×10-9	Increasing
Xuchang	6.00×10-4	Stationary	0.0018	Increasing
Kaifeng	2.58×10-5	Stationary	4.57×10-8	Increasing
Xixia	2.06×10-5	Stationary	7.99×10-6	Decreasing
Baofeng	1.41×10-6	Stationary	0.012	Increasing
Nanyang	2.16×10-9	Stationary	0.0087	Increasing
Shangqiu	9.46×10-7	Stationary	6.26×10-11	Increasing
Gushi	1.33×10-9	Stationary	0.25	No Trend
Xihua	4.15×10-7	Stationary	6.03×10-7	Increasing

From Table 1 it can know that these SPEI data from 12 station are stationary. Because their stability, in the SARMIA model, no further differential processing of the data is required to minimize the standard deviation. The Integrated(d) is selected to be 0 for models of SPEI data in 12 stations including Xinxiang, Anyang, Sanmenxia, Zhengzhou, Xuchang, Kaifeng, Xixia, Nanyang, Baofeng, Xihua, Shangqiu, and Gushi. According to MK test, most of stations have a increasing trend of SPEI except Sanmenxia, Xixia and Gushi. Since seasonal differencing(D) is used to remove trend or non-stationary from seasonal period, the research uses D = 1 to get a stable data set of SPEI value and achieve this purpose.

2.3. Selecting appropriate AR(p), and MA(q) into SARIMA

The forecasting is based on the SPEI values from 1978 to 2018 in 12 stations and is done for future five years, 2019-2024. The research has to select SARIMA with most appropriate autogressive(p), integrated(d), and moving average(q) to fit the previous data well and give a accurate forecasting results. After determination of d, AR(p) and MA(q) are selected depended on the AIC and BIC. The Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are metrics developed to assess the relative quality of statistical models in the context of a specific time-series dataset and are frequently employed for model selection purposes.[9,10] AIC and BIC are all based on the likelihood function that help them to measure how well model fits the data set. A lower AIC and BIC indicates there are smaller differences between fitting results and true values of data. Hence, the best seasonal ARIMA can also be find depended on AIC and BIC. A model with minimum AIC and BIC among models with different AR(p) and MA(q) is the most appropriate for

DOI: 10.54254/2755-2721/129/2025.20299

time-series data collections. The research uses auto_arima method from pmdarima library in Python to achieve this. This method is able to be used to try different p and q automatically. The output results of SPEI data in two station is that: Table 2 shows that the final seasonal ARIMA models of 12 station:

Table 2: The SARIMA Model of SPEI of 12 Stations in Henan Province

Name of Station	SARIMA Model	AIC	BIC
Anyang	ARIMA(5, 0, 0)(2, 1, 0)12	721.71	755.10
Xinxiang	ARIMA(1, 0, 1)(2, 1, 0)12	793.80	814.67
Sanmenxia	ARIMA(5, 0, 5)(2, 1, 0)12	737.45	791.71
Zhengzhou	ARIMA(1, 0, 1)(2, 1, 0)12	753.22	774.08
Xuchang	ARIMA(5, 0, 0)(2, 1, 0)12	782.13	815.52
Kaifeng	ARIMA(5, 0, 0)(2, 1, 0)12	767.07	800.46
Xixia	ARIMA(1, 0, 5)(2, 1, 0)12	724.12	761.68
Baofeng	ARIMA(4, 0, 2)(2, 1, 0)12	707.61	749.35
Nanyang	ARIMA(2, 0, 0)(2, 1, 0)12	780.04	800.91
Shangqiu	ARIMA(1, 0, 1)(2, 1, 0)12	668.38	689.25
Gushi	ARIMA(5, 0, 0)(2, 1, 0)12	760.02	793.41
Xihua	ARIMA(4, 0, 2)(2, 1, 0)12	696.45	734.02

3. Results and reflection

3.1. Prediction results

The research uses SPEI data in 12 stations in Henan province from 1978 to 2018 to train the seasonal ARIMA models m and dynamically forecast future 5 years values from 2019 to 2024 by models. From SPEI index, it's able to find when the drought would occur in this duration. Graphs from Figure1 to Figure12 show results of future five-year SPEI values of 12 stations in Henan province:

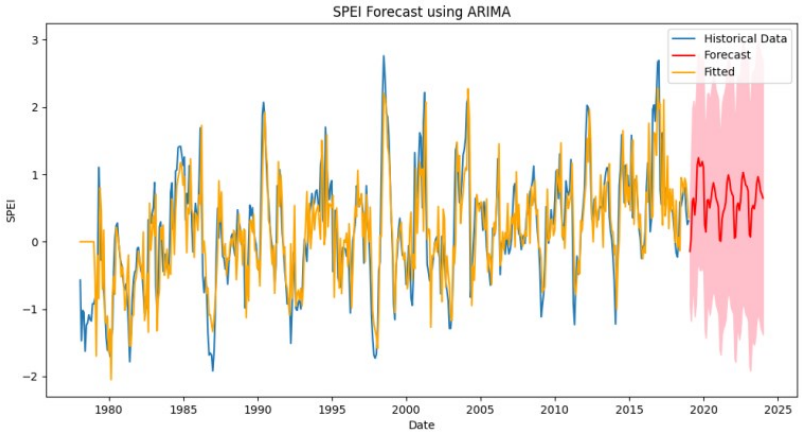

Figure 1: SPEI Forecast Results for the Next Five Years in Anyang Station

DOI: 10.54254/2755-2721/129/2025.20299

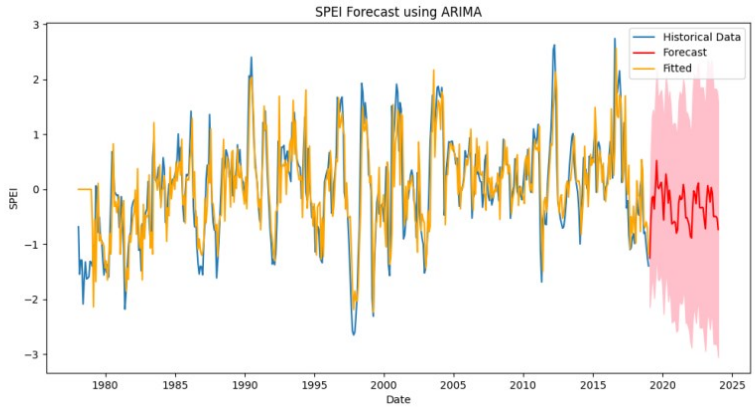

Figure 2: SPEI Forecast Results for the Next Five Years in Xinxiang Station

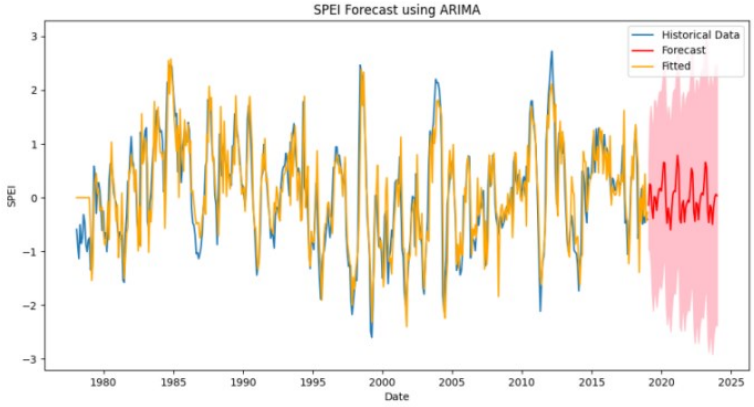

Figure 3: SPEI Forecast Results for the Next Five Years in Sanmenxia Station

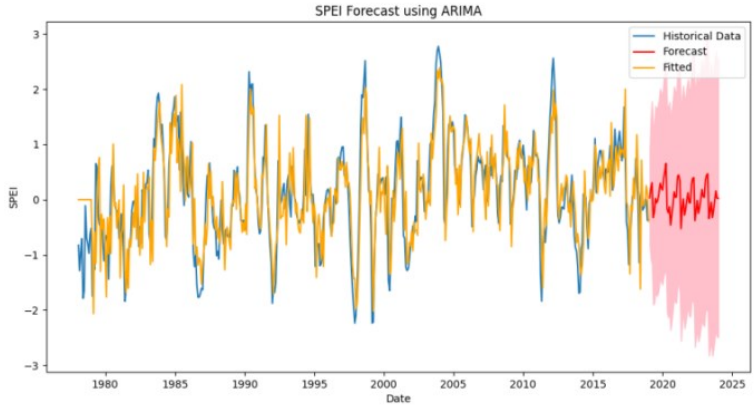

Figure 4: SPEI Forecast Results for the Next Five Years in Zhengzhou Station

DOI: 10.54254/2755-2721/129/2025.20299

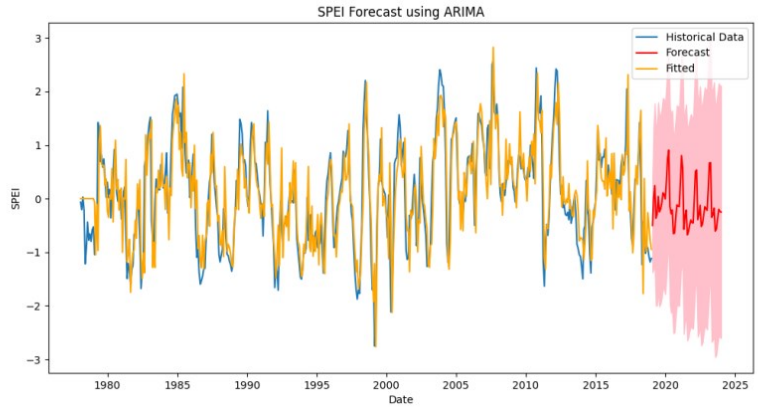

Figure 5: SPEI Forecast Results for the Next Five Years in Xuchang Station

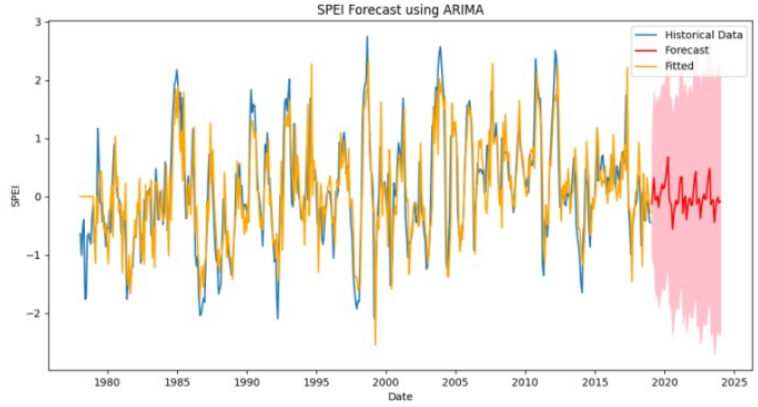

Figure 6: SPEI Forecast Results for the Next Five Years in Kaifeng Station

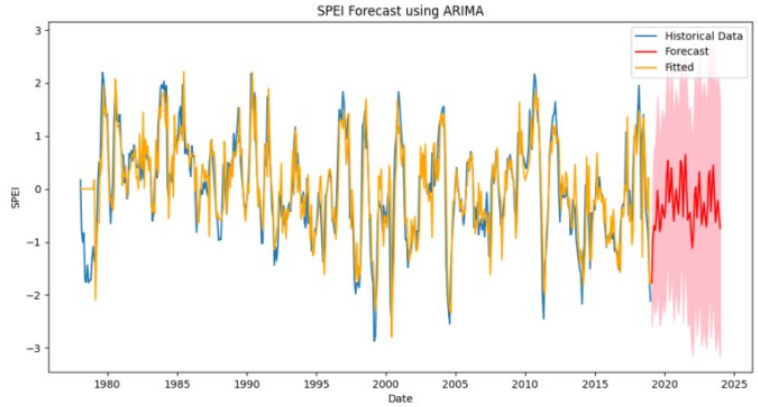

Figure 7: SPEI Forecast Results for the Next Five Years in Xixia Station

DOI: 10.54254/2755-2721/129/2025.20299

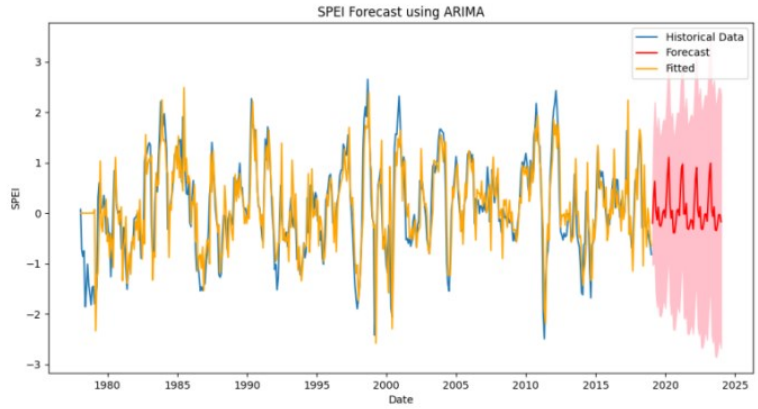

Figure 8: SPEI Forecast Results for the Next Five Years From 1978 to 2018 in Baofeng Station

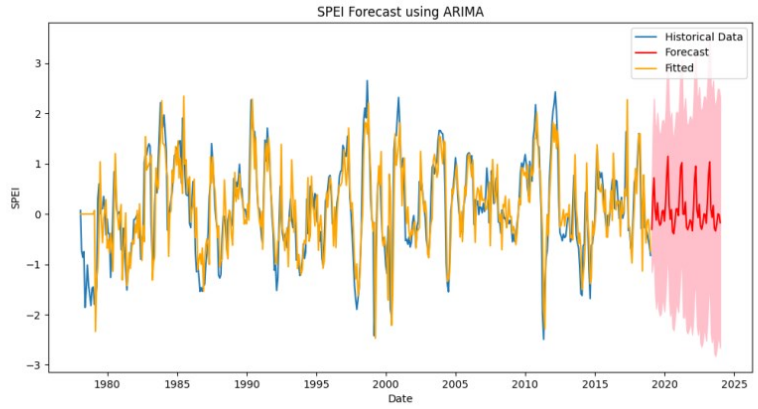

Figure 9: SPEI Forecast Results for the Next Five Years From 1978 to 2018 in Nayang Station

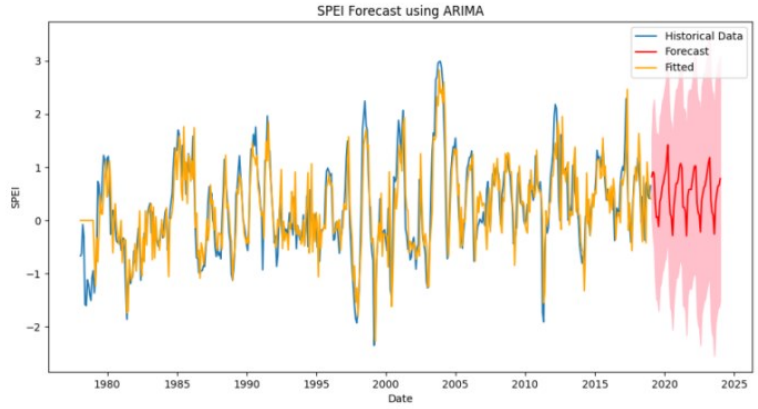

Figure 10: SPEI Forecast Results for the Next Five Years From 1978 to 2018 in Shangqiu Station

DOI: 10.54254/2755-2721/129/2025.20299

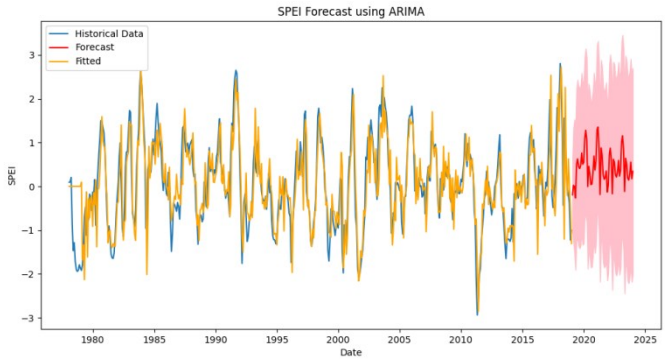

Figure 11: SPEI Forecast Results for the Next Five Years From 1978 to 2018 in Gushi Station

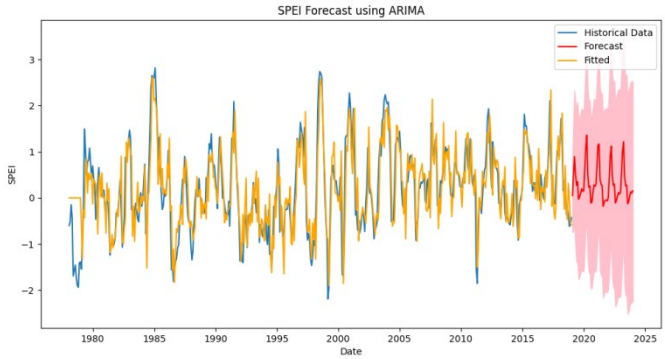

Figure 12: SPEI Forecast Results for the Next Five Years From 1978 to 2018 in Xihua Station

Blue line represents the true previous values of SPEI from 1978 to 2018; yellow line represents the fitted data of previous years; red line represents the mean of forecasting SPEI values from 2019 to 2024; and red region represents the predicted value of a 90% confidence interval.

3.2. Tests of seasonal ARIMA models

The research decides to use 3 dimensions to evaluate the seasonal ARIMA models of 12 stations in Henan province: test for the stationary of residuals of fitted results, the residual series analysis, and the fitted rate. ADF test with P value is used for testing of the stationary of residuals of fitted results. Ljung-Box with P value is able to examine whether there is a autocorrelation in residuals in order to test if the noise type in data is white noise. For fitted rate, the research just uses the last 80% of data to calculate, because, at the beginning of ARIMA fitting, model doesn't capture the pattern and trend of data.

Table 3 shows that the results of these tests:

Table 3: Tests by SARIMA Model of SPEI of 12 Stations in Henan Province

Name of Station	Stability of Residuals	Residual Series Analysis	Degree of Fitting//%
Anyang	Stationary	Not White Noise	62.84
Xinxiang	Stationary	Not White Noise	62.99
Sanmenxia	Stationary	White Noise	73.57
Zhengzhou	Stationary	Not White Noise	69.04
Xuchang	Stationary	White Noise	71.23

DOI: 10.54254/2755-2721/129/2025.20299

Table 3: (continued).

Kaifeng	Stationary	Not White Noise	67.98
Xixia	Stationary	White Noise	70.62
Baofeng	Stationary	Not White Noise	69.24
Nanyang	Stationary	Not White Noise	67.90
Shangqiu	Stationary	Not White Noise	67.78
Gushi	Stationary	Not White Noise	70.18
Xihua	Stationary	Not White Noise	69.41

3.3. Discussion and reflection

From results of tests, the research is able to get that all models have a stable residuals. However, except Sanmenxia station, Xuchang station and Xixia station, the noises of data from other stations are not white noise. These things indicates that residuals series doesn't have seasonality and trend but there are still some patterns of residuals that are not be captured by seasonal model. Most degrees of fitting are around 68%, which illustrates that existed SARIMA models already captured main trend and patterns of SPEI time series data, and there is not overfitting phenomenon; however, these models have not obtained complex patterns or noises, indicating that they can be better and more accurate by modifying. Besides, by observing graphs, if selecting a 90% confidence level, the range of SPEI is relatively large that means that it's hard to use these models to give accurate SPEI values. The accuracy of using mean of forecasting results to do predictions of drought is not high enough.

The above things demonstrates that although ARIMA models already are able to do easy predictions, there are a lot of areas needed to improve further. The next step to make models become better may be to introduce nonlinear models or exogenous variables. Moreover, this method for forecasting is lack the support of meteorological knowledge. Prediction results can be improved to gather or apply more meteorological data and models.

4. Conclusion

The Standardized Precipitation-Evapotranspiration Index (SPEI) serves as a crucial metric for assessing the dryness or humidity levels within a specific region. In the current study, it analyze historical SPEI time series data from 12 monitoring stations located in Henan Province, China. Through a data preprocessing approach, it eliminate seasonal influences and identify the most suitable Seasonal Autoregressive Integrated Moving Average (ARIMA) models by carefully selecting the appropriate moving average (MA) and autoregressive (AR) components.

The research uses these ARIMA models to obtain predictions of SPEI values of duration from 2019 to 2024. After doing some tests for residuals and results, the research finds that seasonal ARIMA models can capture main patterns and trends of SPEI data to forecast. Nevertheless, they still have many flaws, such as ignoring complex patterns of data sets and exogenous factors. In summary, seasonal ARIMA models in the present research can be one of the methods to make drought predictions, and in the short term and a certain range, they have effective precision; but, they need to be improved and are considered combined with other methods to forecasting drought. Future research should explore the integration of machine learning techniques alongside traditional ARIMA approaches to enhance predictive performance.

DOI: 10.54254/2755-2721/129/2025.20299

References

[1] *Vicente-Serrano, S. M., et al. (2010). A multiscalar drought index sensitive to global warming: The standardized precipitation evapotranspiration index. Journal of Climate, 23(7), 1696–1718.*

[2] *Box, G. E. P., & Jenkins, G. M. (2015). Time series analysis: Forecasting and control. Hoboken, NJ: John Wiley & Sons.*

[3] *Li, Y. (2016). The application of ARIMA model in forecasting of PDSI in Henan Province. Agricultural Science & Technology, 17(3), 760–764.*

[4] *Shmueli, G. (2018). Practical time series forecasting: A hands-on guide. Axelrod Schnall Publishers.*

[5] *Dimri, T., Ahmad, S., & Sharif, M. (2020). Time series analysis of climate variables using seasonal ARIMA approach. Journal of Earth System Science, 129(1), Article 149.*

[6] *Wang, Q., et al. (2021). A multi-scale daily SPEI dataset for drought characterization at observation stations over mainland China from 1961 to 2018. Earth System Science Data, 13(2), 331–341.*

[7] *Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366a), 427–431.*

[8] *Mann, H. B. (1945). Nonparametric tests against trend. Econometrica, 13(3), 245.*

[9] *Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723.*

[10] *Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461–464.*