UDC 517.55

Statistical analysis and forecasting of meteorological data

Sharipov O. Sh.
National University of Uzbekistan named after Mirzo Ulugbek, Tashkent;
V.I. Romanovskiy Institute of Mathematics of the Academy of Sciences of the Republic of Uzbekistan, Tashkent,
osharipov@yahoo.com
Khudoykulova H. B.
National University of Uzbekistan named after Mirzo Ulugbek, Tashkent,
Tashkent State University of Economics, Tashkent
xudoyqulova_h@nuu.uz

RESUME

This article synthesizes data on the daily average air temperature for December from the Tashkent-Observatory meteorological station over the years 1904–2023 (120 years of observations), and analyzes these data using statistical methods to explore forecasting possibilities. The paper details preliminary examinations of the time series (exploratory analysis; decomposition of trend and seasonality using STL), tests for stationarity (ADF and KPSS), identification of the correlation structure (via ACF and PACF), and the selection, parameter estimation, diagnostics, and assessment (both as model diagnostics and forecasts) of classical models including AR, MA, ARMA, ARIMA, and SARIMA.

Key words: Time series; weather forecasting; ARIMA; SARIMA; STL; Augmented Dickey–Fuller (ADF); ACF/PACF.

Forecasting time series is a significant branch of data science, applied in various domains from economics to meteorology. Forecasting is highly useful because, based on historic data, it allows predictions about the future. However, selecting a suitable model for forecasting is the most critical step. In our analyses, the process of selecting appropriate models for the data and verifying their reliability is carried out in accordance with rigorous statistical testing principles. This approach helps to more effectively predict future meteorological trends and aids decision-making across multiple sectors.

Analyses are based on daily average air temperature (for December) observed at the Tashkent-Observatory meteorological station over the years 1904-2023 (120 years of data).

Methodology

To select appropriate models for the data and to confirm their reliability, we carried out the following key statistical tests in a step-by-step sequence:

- **Stationarity testing**: This is an essential step for forecasting time series. It involves checking whether statistical properties such as mean and variance remain constant over time. We used the widely-applied Augmented Dickey-Fuller (ADF) test to determine whether the series is stationary.
- **Correlation analysis**: To understand the relationships between observations at various time lags, two main functions are used: the autocorrelation function (ACF) and the partial autocorrelation function (PACF).
- **Seasonality testing**: Because repeated patterns at regular intervals (daily, monthly, quarterly, or yearly) are an integral part of many time series, detecting seasonality is important for forecasting. We used STL (Seasonal-Trend decomposition using Loess) to decompose the series into trend, seasonal, and residual (irregular) components.

After conducting initial analyses such as stationarity tests, correlation structure (ACF and PACF), and seasonality identification, the next step is selecting a suitable forecasting model. Based on the outcomes of the previous stages, we decided whether to apply AR (autoregressive), MA (moving average), ARMA, ARIMA (autoregressive integrated moving average), or SARIMA (seasonal ARIMA) models.

Literature Review

Below is a brief analysis of key literature that underpins methodology. One of the most important sources is the model construction methodology for ARIMA and SARIMA as proposed by Box, Jenkins, and Reinsel (2015), which remains among the most widely used methods. Chatfield (2004) and Brockwell & Davis (2016) treat extensively the issues of seasonality and periodicity. Wei (2006) expands the practical capabilities of time series analysis through seasonal indices and both univariate and multivariate approaches. The econometric frameworks described by Gujarati & Porter (2009) and Hamilton (1994) are used in this work to assess the statistical significance of model parameters via t-tests, p-values, etc. With respect to checking model residuals, the Ljung–Box test (Ljung & Box, 1978) plays an important role. In forecasting, methods developed by Makridakis, Wheelwright & Hyndman (1998), as well as contemporary forecasting approaches in Hyndman & Athanasopoulos (2018), are widely used. Notably, Hyndman's *Forecasting: Principles and Practice* is heavily applied in time series forecasting using open-source software (e.g. R, Python). Overall, the above sources cover the various aspects of time series analysis—from theoretical foundations through to practical forecasting methods.

Analysis and Results

Initially, the analyses are performed on the daily average temperature data for December for the years 1904-2023 (see Figure 1).

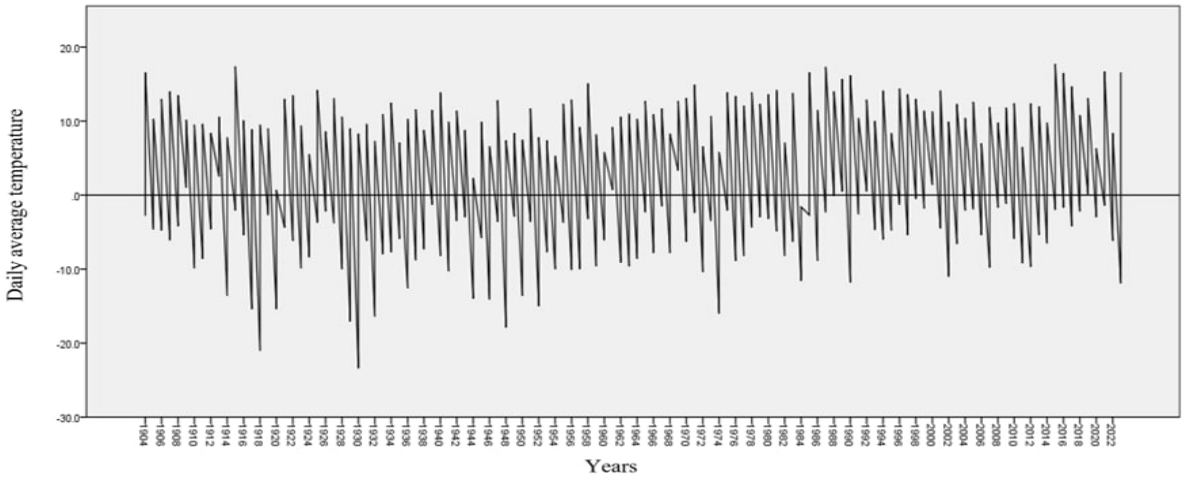

Figure 1. Daily average air temperature in December (1904-2023) at the Tashkent-Observatory, Uzbekistan

According to the results shown in Figure 1, although there are some seasonal effects, there is no obvious long-term trend in the time series. To confirm whether the series is stationary, we apply the Augmented Dickey-Fuller test.

Table 1.The Augmented Dickey-Fuller test

Dickey-Fuller test.	P-Value	Maximum lag order for terms in the regression model
-17.1452	0,000	300

The results show a very low p-value (0.000), indicating statistical significance. The p-value is below the 0.01 significance level, which confirms that the time series is stationary, implying stable and consistent behavior of the data.

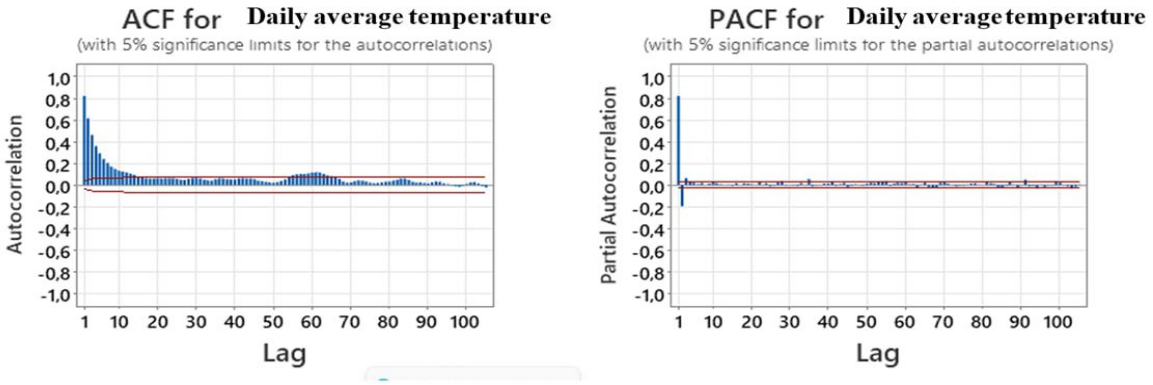

Figure 2. Autocorrelation function (ACF) and partial autocorrelation function (PACF)

The plot shows that the ACF exhibits statistically significant values for several lags, meaning there is persistent correlation among past values in the temperature series. For this dataset, this suggests cyclic components with approximately monthly or bi-monthly periodicity (around 30- and 60-day lags). The fact that the ACF remains significant up to long lags, and that recurrent seasonal peaks are present, motivates consideration of a SARIMA (Seasonal ARIMA) model. To visually express and better understand seasonality, we decompose the time series into trend, seasonal, and residual (irregular) components using an STL decomposition (or SDTS decomposition) (see Figure 3).

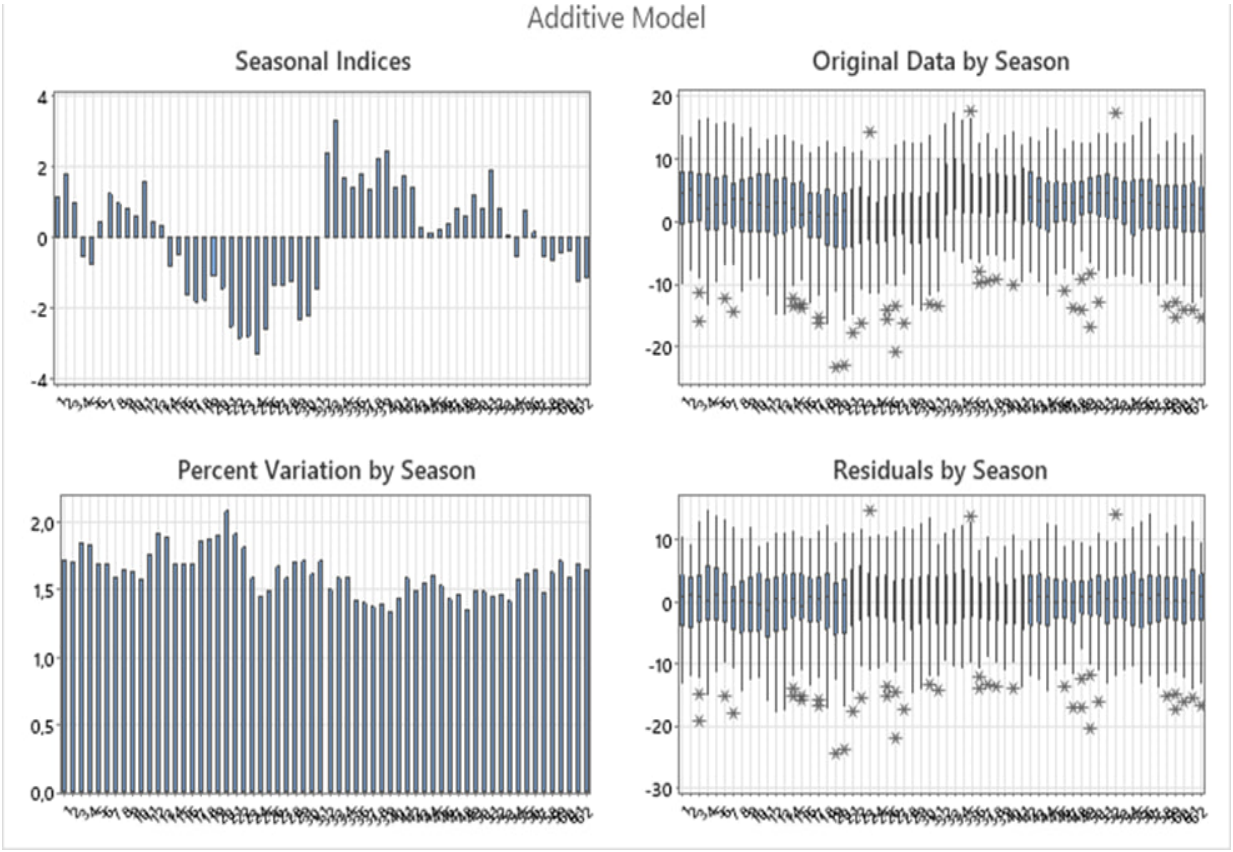

Figure 4. Seasonal index

The seasonal index graph indicates periodic oscillations, with alternating positive values (periods where temperature is above seasonal norm) and negative values (periods below the norm). Sharp changes in the index are observed in the lag intervals of 30–32 days and 60-62 days, which confirms the presence of certain periodicity in the time series.

From the results of stationarity testing, correlation analysis (ACF and PACF), and seasonality identification, we conclude that for the daily average December temperatures, the SARIMA(3,0,0)(1,0,0)[seasonal period] model is the most appropriate. We further validate this choice through reliability tests.

Table 2. Final parameter estimates

Type	Coef	SE Coef	T-Value	P-Value
AR 1	0,9943	0,0164	60,70	0,000
AR 2	-0,2562	0,0227	-11,27	0,000
AR 3	0,0607	0,0164	3,70	0,000
SAR 62	0,0428	0,0165	2,59	0,010
Constant	0,4720	0,0518	9,10	0,000
Mean	2,450	0,269		

Table 2 shows that parameters AR(1), AR(2), AR(3), and seasonal AR(62) are statistically significant, indicating that the time series exhibits both short-term and seasonal dependence.

Table 3. Residual summary

DF	SS	MS
3715	37145,3	9,99874

According to Table 3, the sum of squares of residuals (SS) is 37,145.3, with degrees of freedom (DF) equal to 3,715. The mean square (MS) is 9.99874, showing that average residuals are quite small. This suggests that the chosen model fits the data well.

Table 4. Ljung–Box test for residual autocorrelation

Lag	12	24	36	48
Chi-Square	6,64	13,58	35,69	44,08
DF	7	19	31	43
P-Value	0,468	0,808	0,257	0,426

Table 4 presents the modified Box–Pierce (Ljung–Box) χ^2 statistics for various lags (12, 24, 36, 48). Their corresponding p-values are 0.468, 0.808, 0.257, and 0.426. Since all these p-values exceed 0.05, we conclude that there is no significant autocorrelation in the residuals. This means that after fitting, the residuals behave like random noise, confirming that the chosen SARIMA model adequately explains the time series.

Based on all reliability tests, the selected model **SARIMA(3,0,0)(1,0,0)** is confirmed reliable. Therefore, it can be used for forecasting future data. Forecast values for December daily average temperature for the years 2024-2027 are generated (see Figure 5 and Table 6).

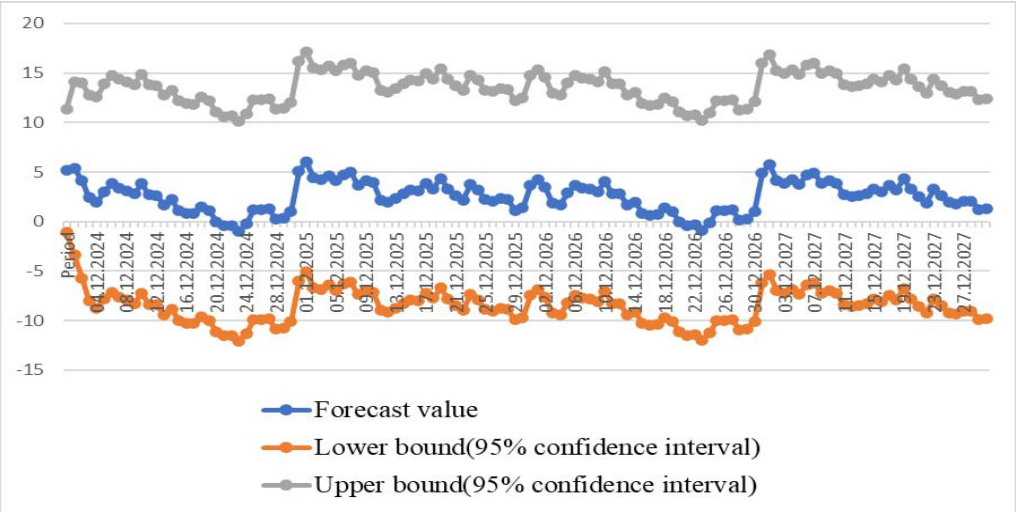

Figure 5. Forecast for December daily average temperature, 2024-2027

These results are calculated using the SARIMA model, one of the seasonal time series models. From the graph it is evident that the confidence intervals remain stable over time, which suggests that the forecasting ability of the model is reliable. Also, recurrent cold periods in December are clearly visible every year. This demonstrates that the seasonal component of the time series is highly stable. The advantage of using a SARIMA model is also shown — the model accounts not only for variability, but for the regular recurring dynamics of temperature.

Conclusion and Recommendations

All the above analyses show that the SARIMA model used for forecasting (daily average December temperature) has a high level of accuracy. The ACF and PACF functions reveal that there are significant correlations at important lags. Seasonal index graphs show that temperature changes repeat in particular periods of the year. The indices display stable oscillations, indicating a strong seasonal component in the series.

These forecasting results have practical implications:

1. **Energy Sector**: Based on forecasts, energy consumption during winter can be planned more precisely; stability of heating supply systems should be ensured.
2. **Agriculture**: Considering extreme cold days observed in December, protective measures for crops should be strengthened.
3. **Transport and Infrastructure**: Forecast results should be used to develop safety strategies for possible issues during cold days (e.g. road icing, vehicle failures).
4. **Climate Monitoring**: Forecasts can be compared to long-term climate observations to detect long-term trends in climate change.

REFERENCES

1. Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. Time Series Analysis: Forecasting and Control. John Wiley & Sons, 2015.
2. Hamilton, J. D. Time Series Analysis. Princeton University Press, 1994.
3. Chatfield, C. The Analysis of Time Series: An Introduction. Chapman & Hall/CRC, 2004.
4. Wei, W. W. S. Time Series Analysis: Univariate and Multivariate Methods. Pearson Addison Wesley, 2006.
5. Brockwell, P. J., & Davis, R. A. Introduction to Time Series and Forecasting. Springer, 2016.
6. Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. Forecasting: Methods and Applications. John Wiley & Sons, 1998.

7. Gujarati, D. N., & Porter, D. C. Basic Econometrics. McGraw-Hill/Irwin, 2009.

8. Brockwell, P. J., & Davis, R. A. Introduction to Time Series and Forecasting. Springer, 2016.

9. Ljung, G. M., & Box, G. E. P. On a Measure of Lack of Fit in Time Series Models. Biometrika, 1978, 65(2), P.297–303.

10. Hyndman, R. J., & Athanasopoulos, G. Forecasting: Principles and Practice. OTexts, 2018.

REZYUME

Ushbu maqola 1904—2023 yillarga oid Toshkent-Observatoriya meteostansiyasining dekabr oyidagi kunlik o‘rtacha havo harorati ma’lumotlarini (120 yillik kuzatuvlar) statistika usullari yordamida tahlil qilib, prognozlash imkoniyatlarini o‘rganadi. Maqolada vaqt qatorlarining dastlabki tekshiruvlari (eksplorator tahlil, STL yordamida trend va mavsumiylikni ajratish), statsionarlik sinovlari (ADF va KPSS), korelyatsion strukturani aniqlash (ACF va PACF), shuningdek AR, MA, ARMA, ARIMA va SARIMA kabi klassik modellarni tanlash, parametr baholash, hamda modelni diagnostika va prognoz sifatida baholash bosqichlari batafsil bayon etilgan.

Kalit so‘zlar: Vaqt qatori; ob-havo prognozi; ARIMA; SARIMA; STL; Augmented Dickey–Fuller (ADF); ACF/PACF.

РЕЗЮМЕ

В данной статье данные о месячной (за декабрь) суточной средней температуре воздуха на метеорологической станции Ташкент-Обсерватория за период 1904–2023 годов (120 лет наблюдений) анализируются с помощью статистических методов для изучения возможностей прогнозирования. В работе подробно описаны предварительные исследования временного ряда (эксплоративный анализ; декомпозиция тренда и сезонности с использованием STL), тесты на стационарность (ADF и KPSS), идентификация корреляционной структуры (через ACF и PACF), а также выбор классических моделей, оценка параметров, диагностика моделей и оценка их прогностических возможностей, включая модели AR, MA, ARMA, ARIMA и SARIMA.

Ключевые слова: Временной ряд; прогнозирование погоды; ARIMA; SARIMA; STL; тест Дики-Фуллера в дополнении (ADF); ACF/PACF.