Karaelmas Science and Engineering Journal
Journal home page: https://dergipark.org.tr/tr/pub/karaelmasfen
DOI: 10.7212/karaelmasfen.1637034

Research Article
Received / Geliş tarihi : 10.02.2025
Accepted / Kabul tarihi : 18.03.2025

Time Series and Machine Learning Methods for Estimating Missing Meteorological Data: The Western Black Sea Basin Case

Eksik Meteorolojik Verilerin Zaman Serisi ve Makine Öğrenmesi Yöntemleri Kullanılarak Tahmin Edilmesi: Batı Karadeniz Havzası Örneği

Yusuf Kaya[1*], Berna Aksoy[2], İsmail Hakkı Özölçer[2]

[1]Istanbul Beykent University, Faculty of Engineering Architecture, Department of Civil Engineering, Istanbul, Türkiye
[2]Zonguldak Bulent Ecevit University, Faculty of Engineering, Department of Civil Engineering, Zonguldak, Türkiye

Abstract

The development of artificial intelligence applications is rapidly advancing today. It enables making existing data meaningful through a wide range of applications. This increases the importance of data for all artificial intelligence subfields. The accuracy, continuity and meaningfulness of data are very important for training and testing the models. Discontinuities or errors are likely to occur in data obtained from long-term physical measurements. It is inevitable for meteorological data, which are highly influenced by external factors to have gaps. Mentioned situation causes negative effects on the analysis reliability of meteorological data, which plays a significant role in climate change and hydrological modelling. Within the scope of this study, monthly precipitation measurements of meteorological observation stations in the city centers of seven different provinces in the Western Black Sea basin were examined. The gaps in the monthly rainfall data measured between 2000 and 2023 were estimated using time series, statistical and machine learning approaches. In the modelling process, ARIMA, SARIMA, ARIMAX, XGBOOST, and mean imputation methods were employed. The analyses revealed that SARIMA models, which consider seasonal effects, provided more consistent results, as demonstrated by performance metrics. The completed data form the basis for advanced drought analysis. Thus, impact of deviations due to data loss in future drought analyses is minimized.

Keywords: Missing data, precipitation, sarima, time series.

Öz

Yapay zekâ uygulamalarının gelişimi günümüzde hızla devam etmektedir. Geniş bir uygulama yelpazesiyle elde var olan verinin anlamlı hale getirilmesini sağlamaktadır. Bu durum tüm yapay zekâ alt grupları için verinin önemini arttırmaktadır. Verinin doğruluğu, sürekliliği ve anlamlılığı modellerin eğitilmesi ve test edilmesi için oldukça önemlidir. Uzun süreli fiziksel ölçümlere dayalı elde edilen verilerde süreksizlikler veya hatalar olması olasıdır. Dışsal faktörlerin oldukça etkili olduğu meteorolojik verilerde eksikler olması kaçınılmazdır. Bu durum iklim değişikliği ve hidrolojik modellemelerde önemli bir yer tutan meteorolojik verilerin analiz güvenliğini negatif etkilemesine neden olmaktadır. Bu çalışma kapsamında Batı Karadeniz havzasında yer alan yedi farklı ilin şehir merkezlerindeki meteorolojik gözlem istasyon aylık yağış ölçümleri incelenmiştir. Ölçümleri yapılan 2000-2023 yıllarındaki aylık yağış verilerindeki eksiklikler zaman serisi, istatiksel ve makine öğrenmesi yaklaşımları ile tahmin edilmeye çalışılmıştır. Modellemelerde ARIMA, SARIMA, ARIMAX, XGBOOST ve ortalama ile tamamlama yöntemleri kullanılmıştır. Yapılan analizlerde mevsimsel etkileri dikkate alan SARIMA modellerinin daha uyumlu sonuçlar verdiği performans metrikleri ile ortaya konulmuştur. Tamamlanan veriler ileri düzey kuraklık analizlerine altlık oluşturmaktadır. Böylece ileride yapılacak kuraklık analizlerinin veri kayıplarından kaynaklı sapmalardan etkilenmesi minimize edilmiştir.

Anahtar Kelimeler: Eksik veri, sarima, yağış, zaman serisi

*Corresponding author: yusufkaya@beykent.edu.tr

Yusuf Kaya orcid.org/0000-0002-0923-2217
Berna Aksoy orcid.org/0000-0001-6925-1594
İsmail Hakkı Özölçer orcid.org/0000-0002-8404-0522

1. Introduction

Due to growing demands for the Artificial Intelligence (AI) applications and their wide use, data is becoming more and more significant in the modern era. Data is the primary component of AI applications since these systems need vast volumes of data for testing, training and ongoing self-improvement procedures in order to function properly and efficiently. AI models are unable to learn and perform as intended in the absence of comprehensive and continuous data. The continuity and significance of the data collection are just as crucial as the gathering of information that belongs to any entity. The data gathered by long-term physical measurements and observations very certainly contains flaws, inaccuracies and discontinuities. The elimination of these shortcomings is crucial.

1.1. Literature Review

There are various methods in the literature for completing the measured data. These methods have differentiated with the methodological approaches and have taken their current final form. The studies for completing missing data, which started with statistical methods based on the mean and median values of the data, were followed by correlation-based regression methods, time series models, machine learning-based methods, deep learning and hybrid methods. The studies for completing the missing data are given in Table 1 according to the time structure.

About %30 of the 30-year rainfall data has missing values. Both statistical methods (averaging, multiple linear regression, normal ratio, and linear interpolation) and artificial intelligence (Support Vector Regression, Neural Network) are employed to estimate missing data (Wangwongchai et al, 2023). Statistical methods have yielded good results at some points because mentioned methods are simple and understandable. 19 years precipitation data used for making hybrid Artificial Neural Networks models (Waqas et al, 2024). According to performance metrics (MAE, RMSE and R^2) Long-Short Term Memory Recurrent Neural Network model have high accuracy.

Autoregressive integrated moving average (ARIMA) and seasonal autoregressive integrated moving average (SARIMA) models were used for estimating surface runoff over an extended period of time (Valipour, 2015). It was demonstrated that SARIMA models produced superior results with a relative error of less than 5%. Using time series, the observation data is identified required for precipitation estimation (Valipour, 2012). According to the findings of the error analysis (R^2), time series models performed better in semiarid climates. SARIMA models are used to estimate

Table 1. Studies in literature.

Authors	Year	Article	Method	Aim
Schafer and Graham	2002	Missing Data: Our View of the State of the Art	Mean and Median	For the completion of missing data
Dempster et al.	1977	Maximum Likelihood from Incomplete Data via the EM Algorithm	Expectation-Maximization (EM)	Introduction to the EM Algorithm
García-Laencina et al.	2009	K nearest neighbours with mutual information for simultaneous classification and missing data imputation	K-Nearest Neighbors (KNN)	Effectiveness of the KNN Method in Missing Data Imputation
Stekloven and Bühlmann	2012	MissForest—Non-parametric Missing Value Imputation for Mixed-type Data	Random Forest	Introduces MissForest based on Random Forest
Vincent et al.	2008	Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion	Autoencoders	Missing data imputation with the relevant method
Wangwongchai et al.	2023	Imputation of missing daily rainfall data; A comparison between artifical intelligence and statistical techniques	Support Vector Regression and Statistical Tecniques (Arithmetic Averaging)	Comparing Method Performances

monsoon time series over a 14-year span (Dabral and Murry, 2017). The ARIMA model reflected the climate change model, according to a study done in the Las Vegas area (Huntra and Keener, 2017).

There are seven city center meteorological stations spread throughout seven provinces in the Western Black Sea region that provide monthly average precipitation data from January 2020 to December 2023. These precipitation data have a number of shortcomings. These shortcomings could have been brought on by things like malfunctioning measurement equipment, shifting the station's position, etc.

To ensure continuity in monthly rainfall data, the four methods listed above are used to fill in the missing data within parameters of this study. Average completion, time series models (ARIMA, SARIMA, and ARIMAX), and machine learning techniques (XGBoost) were used to obtain missing data. The models that performed the best with the test data were then chosen.

Examining the research in the literature reveals that they use a single group as the primary focus. Some data groups were subjected to machine learning, while others were subjected to time series, and yet others were subjected statistical evaluation tecniques. There are limitations to the study where all method groups were assessed collectively. Furthermore, models were created based on a single model in the research that were reviewed. Monthly models were created for this investigation, and seasonality was assessed within each model.

2. Material and Methods

2.1. Dataset

Data from five sub-basins of the Western Black Sea basin's seven provincial meteorological observation stations were used in this investigation. In figure 1, the locations of the weather stations are displayed.

The data set identification of these observation stations is as stated in Table 2.

The analyses used data from seven different precipitation stations from January 2000 to December 2023 as targets and attributes. These included monthly average total precipitation, monthly relative humidity, monthly average temperature, monthly minimum temperature, monthly max-

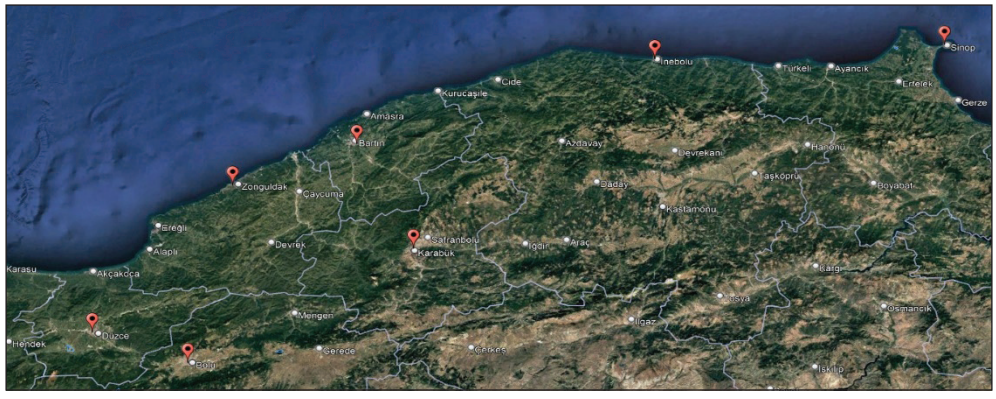

Figure 1. Locations of the meteorological stations (Google Earth).

Table 2. Western Black Sea basin.

Sub-basin	Station No	Station ID	Latitude	Longitude	Altitude
Eregli Sub-basin	17022	Zonguldak	41.4492	31.7779	123
Bartın Sub-basin	17020	Bartin	41.6248	32.3569	36
Filyos Sub-basin	17078	Karabuk	41.2044	32.6328	270
Devrakani-Sinop Sub-basin	17024	Inebolu	41.9789	33.7636	48
Filyos Alt Sub-basin	17070	Bolu	40.7328	31.6021	743
Melen Sub-basin	17072	Duzce	40.8436	31.1487	145
Devrakani-Sinop Sub-basin	17026	Sinop	42.0299	35.1544	28

imum temperature, monthly average wind speed, monthly maximum wind speed, and monthly maximum precipitation amount.

2.2. Method

To fill in the gaps, a variety of techniques based on distinct approaches in the literature have been created. The scenario shown in Table 3 arises when these techniques categorized under the five major areas.

Mean imputation, time series models (ARIMA, SARIMA, and ARIMAX), and machine learning-based techniques (XGBoost) were used in this study to fill in the missing data on the meteorological data set. The methods that produced the best outcomes based on model performance were identified. Figure 2 illustrates the models' flow diagram.

Table 3. Missing data imputation methods.

Methods	Sub-methods
Statistical Methods	Mean Imputatıon, Linear Regression
Correlation-Based Methods	Correlation-Based Regression
Advanced Statistical Methods	Time Series Models
Machine Learning-Based Methods	K-Nearest Neighbors (KNN), Support Vector Regression (SVR)
Deep Learning	Artificial Neural Networks (ANN), Autoencoders

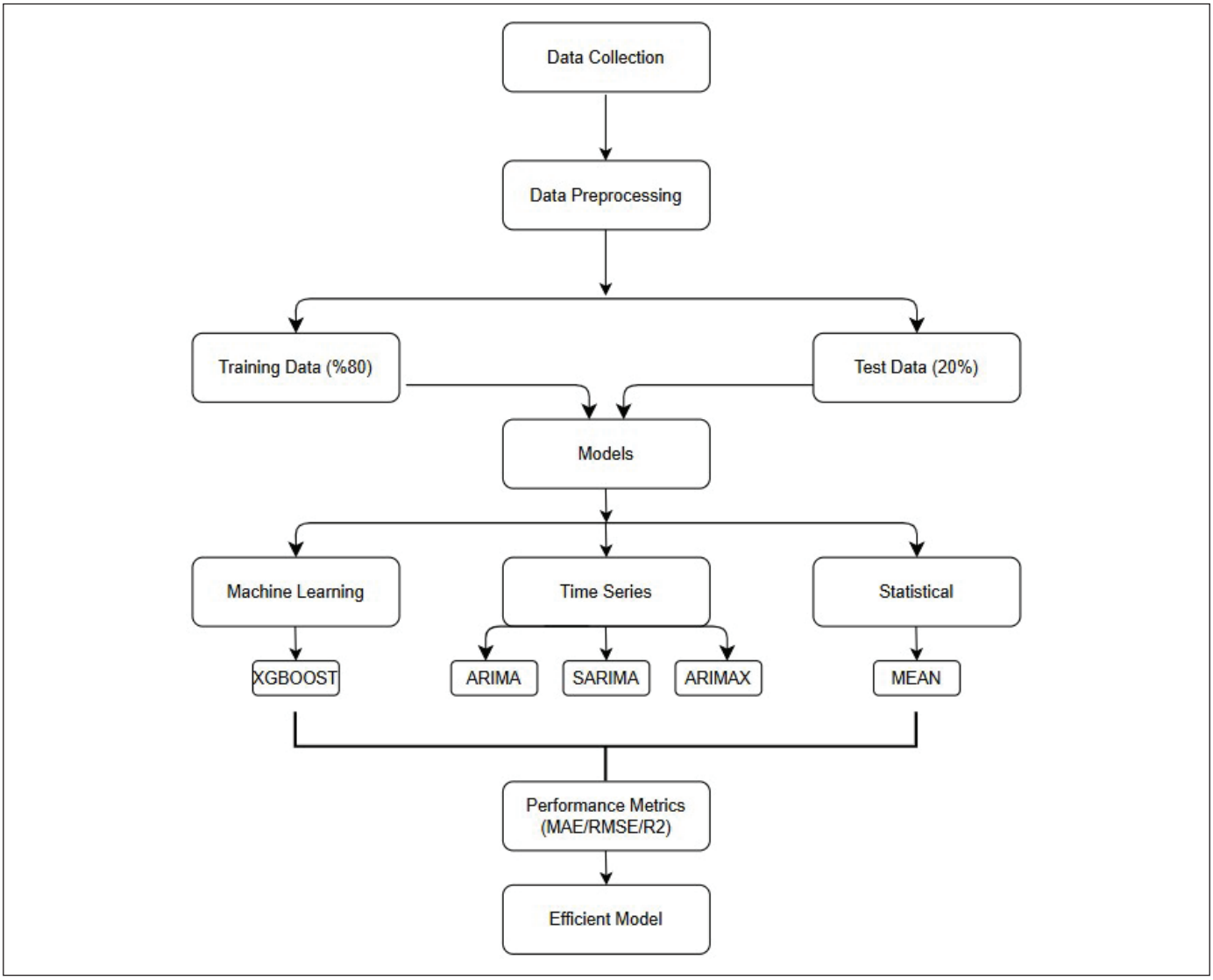

Figure 2. Model flow chart.

The purpose of employing the chosen techniques is to determine which approaches are better suited to the data's structure, how well simple and sophisticated models work, and how models alter when subjected to seasonal impacts.

2.2.1. ARIMA

This approach is a time series model that was put forth (Box and Jerkins, 1970). Without requiring outside influences on the data, it displays performance based on historical values. Stationarity of the data is required by the ARIMA model. Equation 1 provides the ARIMA model's mathematical expression (Zhang, 2003).

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q} \quad (1)$$

In equation 1, y_t is the value of the time series at time t, ε_t is the value of the error at time t (random error), ϕ is a coefficient (AR, AutoRegressive) that allows the series to be associated with past values, θ is the moving average coefficient (MA, Moving Average), and p and q are the degrees of the coefficients. The ARIMA model consist of AR (AutoRegressive), MA (Moving Average) and I (Integrated) components. AR is the relationship of the series with past values, MA is the effect of past values of error terms on the series, and I is the difference process to make the data stationary.

2.2.2. ARIMAX

Modeling the ARIMA model on only one variable has a negative effect on the results. With the ARIMAX model, multiple independent variables of the data be used (Fan et al., 2009). A generalization of the ARIMAX approach was made (Bierens, 1987). The mathematical expression of the ARIMAX model is given in equation 2.

$$\left(1 - \sum_s^p \alpha_s L_s\right)\Delta y_t = \mu + \sum_{s=1}^q \beta_s L^s x_t + \left(\sum_{s=1}^r \gamma_s L^s\right) e^t \quad (2)$$

Here, $\gamma_s L^s = y_{t-s}$ is expressed as $\Delta y_t = y_t - y_{t-1}$, e^t is the error, L is the lag operator; β, α, γ, μ are the model parameters. While the ARIMAX model uses the components of the ARIMA model (AR, MA, I), it also includes the exogenous variable (X) component (Sutthichaimethee and Ariyasaijakorn, 2017).

The ARIMAX model's representation of the time series is determined by the parameters that comprise its components (AR, MA, and I). Time series analysis and statistical analysis are frequently used for parameter selection. The Partial Autocorrelation Function (PAFC) plot is analyzed when selecting the AR parameter. The ADF (Augmented Dickey-Fuller) test is used to verify stationarity while choosing the I parameter. Lastly, the Autocorrelation Function (ACF) graph is analyzed in order to determine the MA parameter. The "pmdarima" module in Python is utilized to optimize the parameters.

2.2.3. SARIMA

The SARIMA model is a product of seasonal and non-seasonal polynomials and is expressed as SARIMA (p, d, q) x (P, D, Q)s. Here (p, d, q) and (P, D, Q) are the non-seasonal components, respectively, and s represents the seasonality period. The SARIMA model is defined by equation 3 in academic sources (Box et al., 2008; Cryer and Chan, 2008; Wang et al., 2008).

$$\phi(B^s)\phi(B)(1 - B^s)^D(1 - B^s)^d yt = \theta(B^s)\theta(B)\varepsilon_t \quad (3)$$

In this equation, Φ and φ are seasonal and non-seasonal autoregression (AR) parameters, Θ and Θ are seasonal and non-seasonal moving average (MA) parameters, B is the backshift operator, $B(y_t)=y_{t-1}$, $(1-B^s)^D$:S is the D-th difference of seasonality, $(1-B^s)^d$ is the non-seasonal d-th difference. SARIMA can be used in series that contain seasonality and are non-stationary.

2.2.4. XGBOOST

XGboost (Extreme Gradient Boosting) method was introduced by optimizing the gradient boosting method (Chen and Guestrin, 2016). The XGBoost method is widely used for complex problems and large data sets. XGboost uses the modified loss function (Equation 4 and Equation 5) (Bentéjac et al., 2021).

$$L_{xgb} = \sum_{i=1}^N L(y_i, F(x_i)) + \sum_{m=1}^M \Omega(h(x; a_m)) \quad (4)$$

$$\Omega(h(x; a_m)) = \gamma T + \frac{1}{2}\lambda \| w \|^2 = \gamma T + \frac{1}{2}\sum_{j=1}^T w_j^2 \quad (5)$$

Here, γ denotes the minimum loss reduction required for a partition to be performed; T is the size of the tree (number of leaves); w is a vector of leaf scores; and λ controls the L_2 regularization strength. The hyperparameter γ is specific to XGBoost.

The usage differences for the time series models ARIMA, SARIMA and ARIMAX are given in Table 4.

3. Results and Discussion

ARIMA, SARIMA, ARIMAX, XGBOOST, and average usage methods were used in this study to estimate missing data for each of the seven distinct meteorological observation sites. The open source Python programming language was used for the modelling. Python programming language libraries for machine learning and embedded statistics were utilized. In the Python, Sklearn library used for machine learning model and Statmodels library used for time series models. Various measures were employed to assess the developed models' performance. Appropriate assessment measures were chosen for each model for the sake of a common evaluation, and the model with the best performance was then given further performance tests.

3.1. Basin Stations Model Comparisons

Every month, models for weather stations were developed to assess each time frame separately. There are 23 data values per month on a monthly basis since the dataset as part of examination spans the years 2000-2023. Five test data were produced since the training data was 80% and the test data was 20%.

Table 5 shows the outcomes of five distinct method modeling conducted at Zonguldak station number 17022 for each approach. The Sarima approach produced the best estimation results for the Zonguldak station; Figure 3 shows the test estimation graph for this method.

Table 6 presents the outcomes of five distinct technique modeling conducted at Bartin station number 17020 for each approach. Figure 4 shows the model graph that produced the most consistent results.

Table 7 shows the outputs of the five distinct technique modeling conducted at Karabuk station number 17078 for each approach. Figure 5 shows the model graph that produces the most consistent results.

Table 4. Variations in usage of ARIMA, ARIMAX, and SARIMA.

Feature	ARIMA	ARIMAX	SARIMA
Model Structure	AR, I, MA	AR, I, MA, External	AR, I, MA, Seasonal
Stationarity	Necessary	Necessary	Necessary
Seasonal Data	Not Supported	Not Supported	Supported
External Variable	None	Exist	None
Advantages	Simple	Includes external Factors	Seasonality
Disadvantages	Non-seasonal	External factor identification	Complexity

Table 5. Model output of Zonguldak station.

Test	Original	Arimax	Sarima	Arima	Mean	XGBoost
1	43.6	44.11	30.02	104.95	60.61	64.39
2	121.4	135.64	119.12	84.05	120.92	105.24
3	110.8	46.54	108.34	106.81	128.19	128.38
4	66.6	38.92	63	94.33	52.85	30.86
5	108.4	123.09	92.29	113.55	95.09	106.34

Table 6. Model output of Bartin station.

Test	Original	Arimax	Sarima	Arimax	Mean	XGBoost
1	67.8	83.08	43.55	86.42	56.11	41.96
2	140.4	92.83	87.16	83.36	88.87	26.91
3	91.4	165.22	86.44	85	106.67	134.74
4	47.6	25.07	61.08	87.48	62.23	19.73
5	76.8	15.21	79.81	92.76	86.35	72.19

Table 7. Model outputs of Karabuk station.

Test	Original	Arimax	Sarima	Arima	Mean	XGBoost
1	52.4	62.55	32.52	39.69	22.83	42.59
2	46.4	27.75	50.76	37.47	44.29	33.31
3	20.5	46.65	33.74	35.14	28.84	39.85
4	11.9	15.45	29.2	40.21	26.98	14.49
5	36.1	26.93	48.82	52.95	34.1	25.15

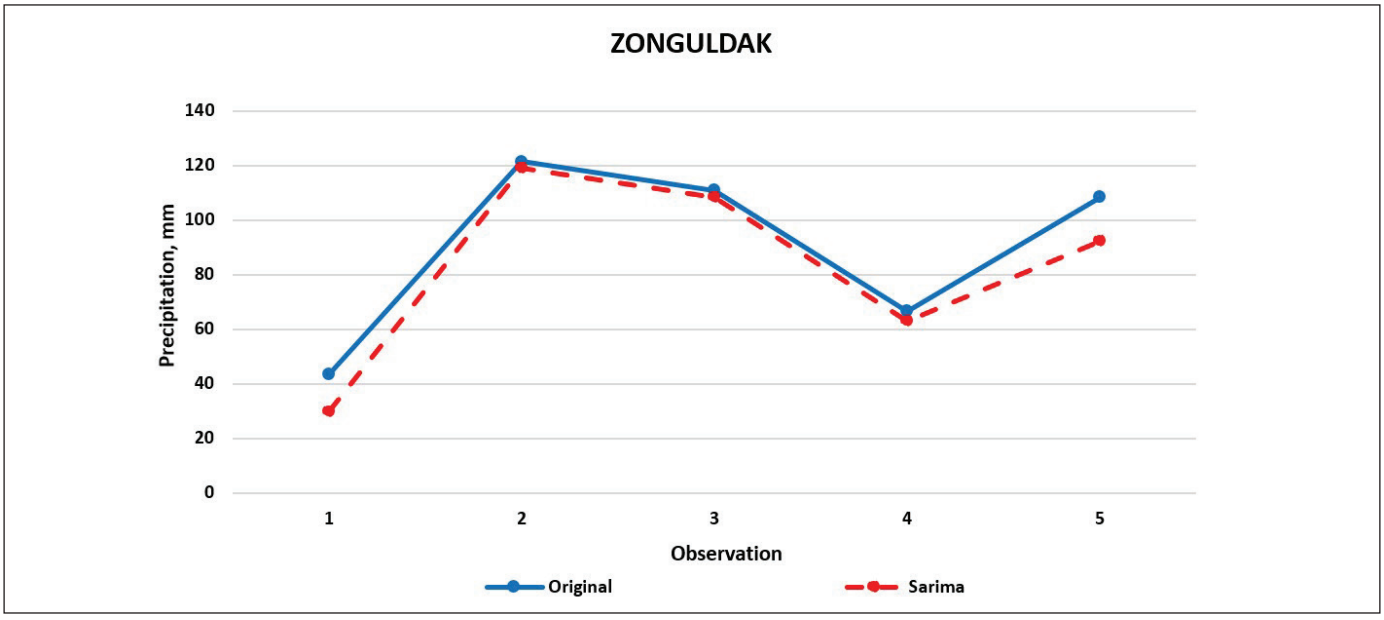

Figure 3. Model comparison (Zonguldak station).

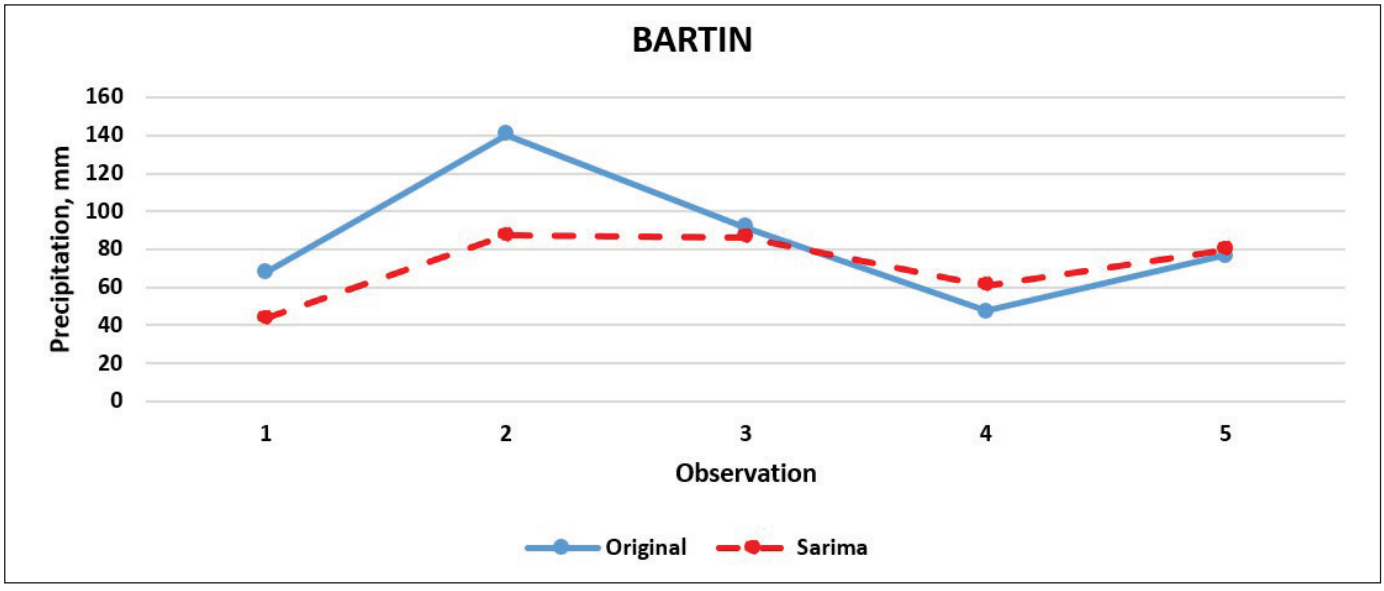

Figure 4. Model comparison (Bartin station).

The results obtained for each method as a result of five different method modeling performed at Inebolu station number 17024 are given in Table 8. The model graph giving the most consistent results in Figure 6.

Table 8. Model outputs of Inebolu station.

Test	Original	Arimax	Sarima	Arima	Mean	XGBoost
1	37.4	19.53	34.62	87.25	53.37	35.03
2	104.2	94.49	110.63	83.03	100.35	92.79
3	71.4	45.34	106.08	75.14	103.6	162.27
4	12.8	3.06	35.84	78.26	52.18	11.49
5	88	142.52	80.86	85.87	86.43	134.25

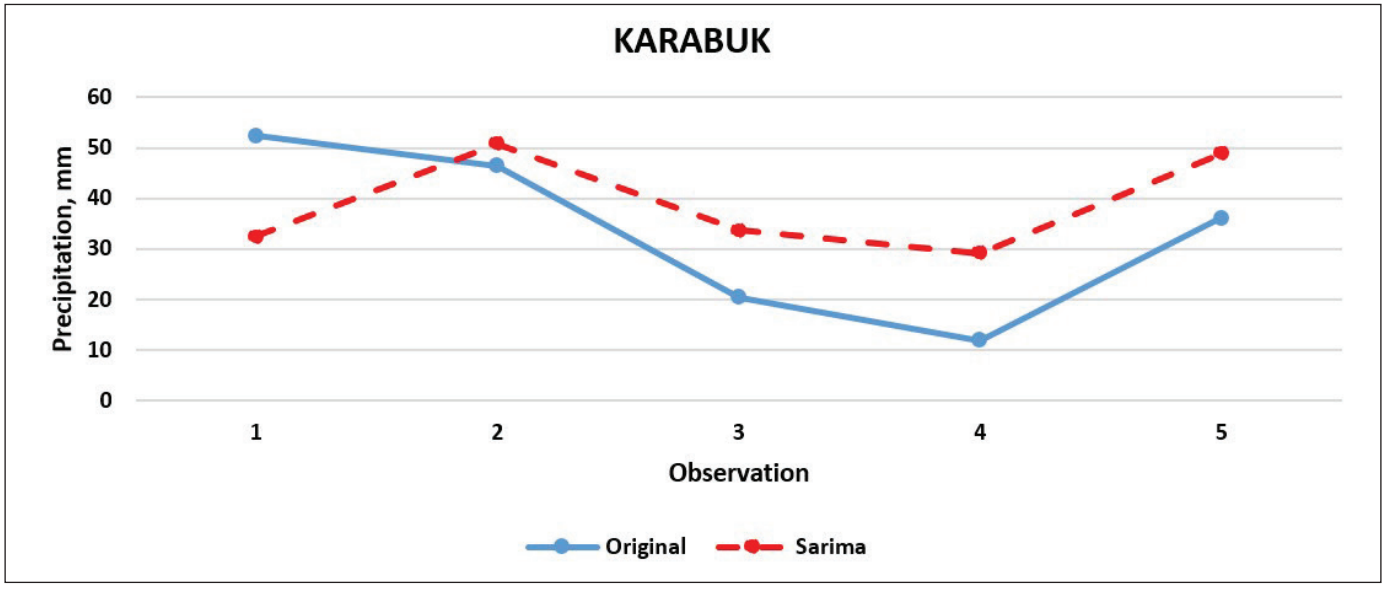

Figure 5. Model comparison (Karabuk station).

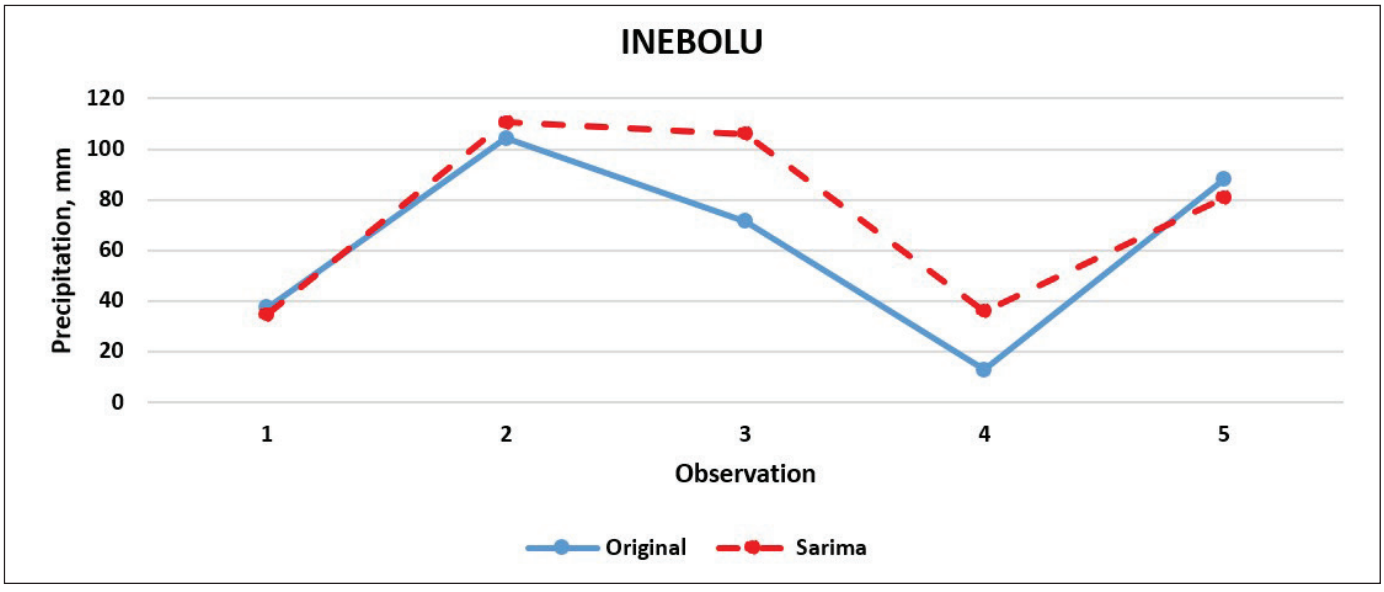

Figure 6. Model comparison (Inebolu station).

Table 9 shows the outcomes of five distinct technique modeling conducted at Bolu station number 17070 for each approach. Figure 7 shows that produces the most consistent results.

Table 10 shows the findings of five distinct methods modeling performed at Duzce station number 17072 for each approach. Figure 8 presents the model graph that produced the most consistent results.

Table 11 provides the findings of five distinct methods modeling conducted at Sinop station number 17026 for each approach. Figure 9 shows the model graph that provides the most consistent results.

3.2. Error Distributions of Prediction Models

The Root Mean Squared (RMSE), Mean Absolute Error (MAE), and R^2 error approaches were used to determine performance outputs of the models created to fill in the missing precipitation data. In table 12, performance metrics are provided for test estimation result.

Table 9. Model outputs of Bolu station.

Test	Original	Arimax	Sarima	Arima	Mean	XGBoost
1	63	32.43	48.34	46.11	51.57	62.05
2	22.5	33.38	27.79	40.36	25.45	14.89
3	3.6	89.45	25.29	41.1	35.66	24.39
4	58.1	4.89	25.16	45.38	21.48	39.49
5	29	27.3	44.62	44.45	46.19	70.68

Table 10. Model outputs of Duzce station.

Test	Original	Arimax	Sarima	Arima	Mean	XGBoost
1	32.8	31.98	30.34	69.06	61.29	81.01
2	78.4	48.84	59.36	61.21	54.63	44.06
3	22.6	109.58	53.08	64.54	66.96	44.46
4	7.8	16.57	27.6	58.4	40.92	70.41
5	34.6	11.49	63.83	67.65	72.32	26.62

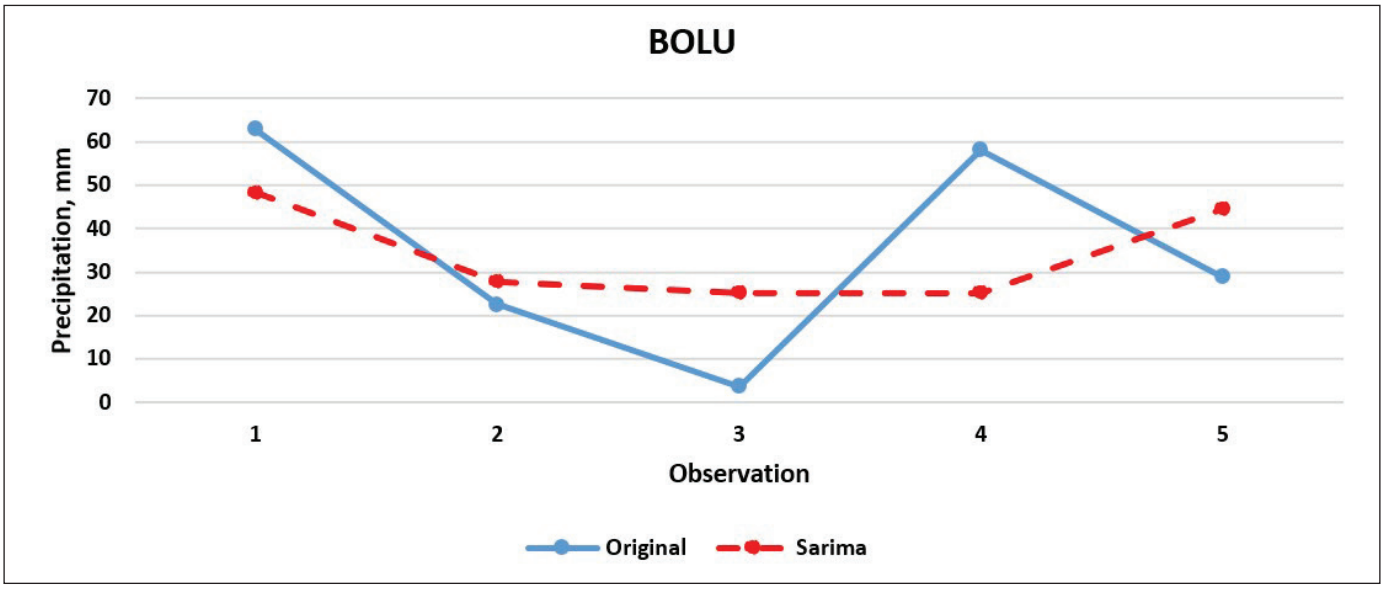

Figure 7. Model comparison (Bolu station).

Table 11. Model outputs of Sinop station.

Test	Original	Arimax	Sarima	Arima	Mean	XGBoost
1	54.8	46.36	35.25	66.59	36.86	58.82
2	106.9	27.17	79.44	57.11	71.67	28.42
3	74.6	118.04	85.78	56.94	83.85	62
4	10.6	3.71	29.36	55.38	38.58	4.66
5	40.3	63.56	49.44	71.81	48.72	52.2

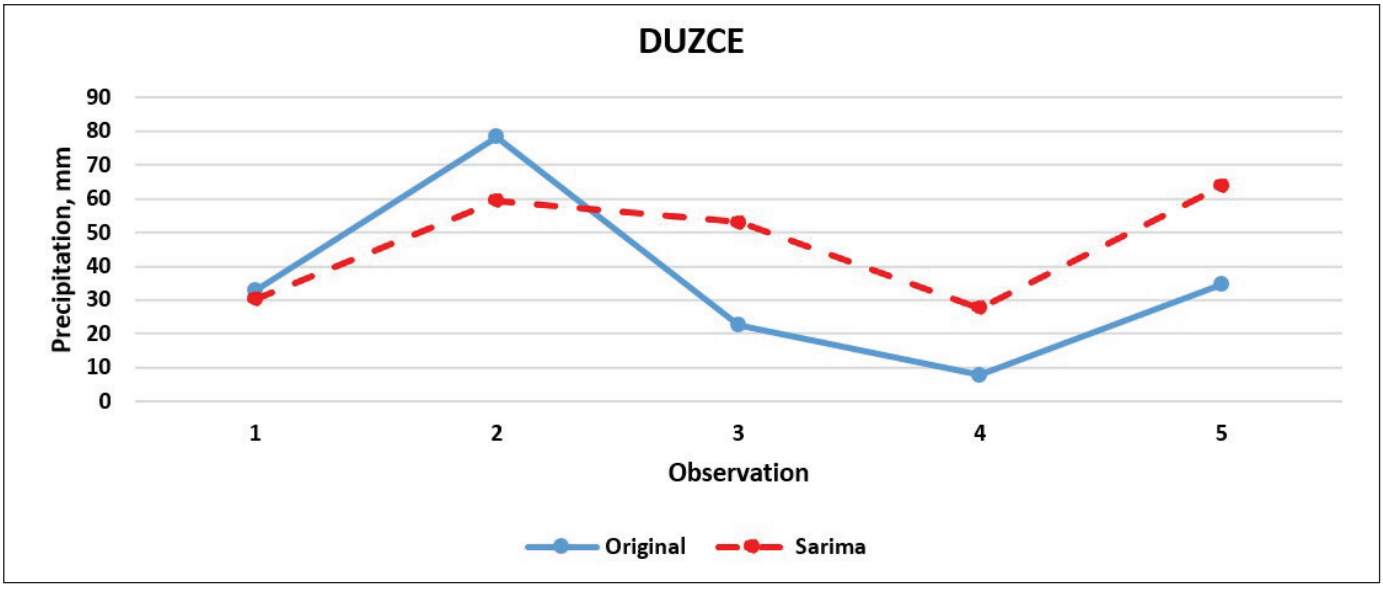

Figure 8. Model comparison (Duzce station).

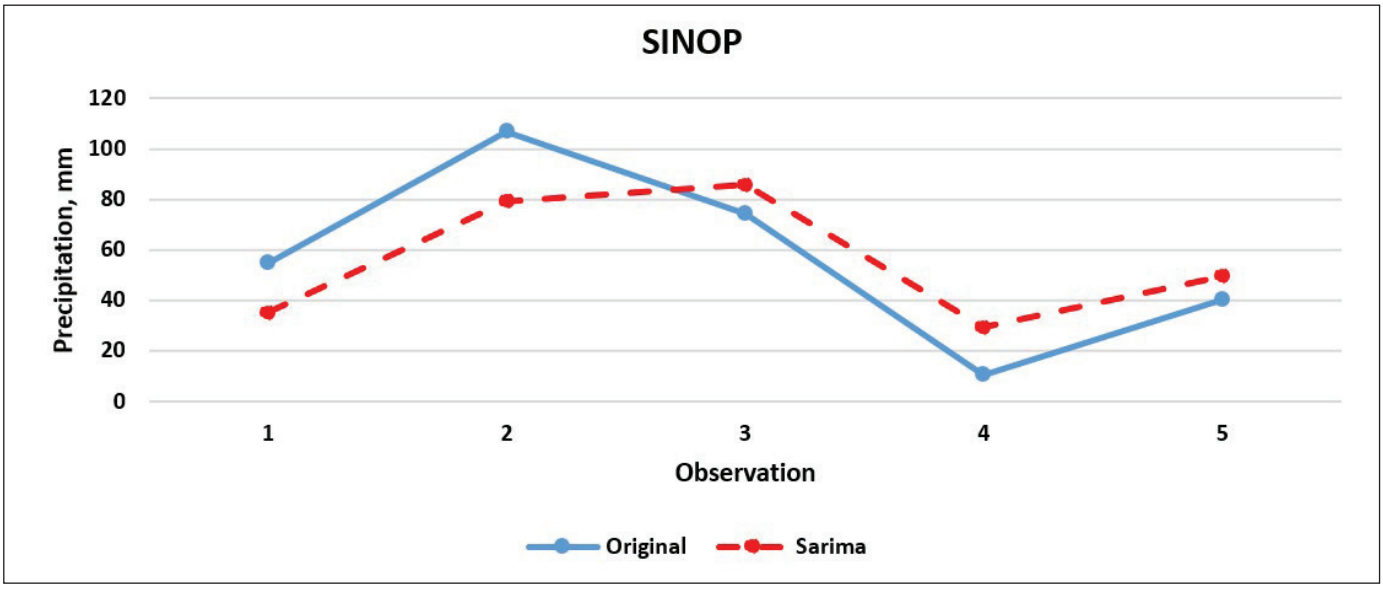

Figure 9. Model comparison (Sinop station).

Table 12. Performance metrics of models.

Station	Error/Model	Arimax	Arima	Sarima	Mean	XGBoost
Zonguldak	**MAE**	24.27	27.11	7.6	12.38	18.46
	RMSE	32.6	34.55	9.67	13.84	21.37
	R2	-0.19	-0.33	0.89	0.78	0.48
Bartin	**MAE**	44.15	27.58	19.78	20.53	43.03
	RMSE	49.49	33.12	26.97	25.8	56.96
	R2	-1.51	-0.12	0.25	0.31	-2.33
Karabuk	**MAE**	13.53	16.28	13.5	11.41	11.15
	RMSE	15.69	17.55	14.49	15.36	12.39
	R2	-0.05	-0.32	0.09	-0.01	0.34
Inebolu	**MAE**	23.58	28.47	14.81	18.59	30.44
	RMSE	28.84	38.04	19.14	23.91	45.89
	R2	0.25	-0.29	0.67	0.48	-0.89
Bolu	**MAE**	36.44	20.08	18.04	20.05	17.92
	RMSE	47.44	21.95	20.21	23.67	22.69
	R2	-3.51	0.03	0.18	-0.12	-0.03
Duzce	**MAE**	29.84	35.8	20.2	33.49	35
	RMSE	42.54	37.47	22.55	34.24	39.91
	R2	-2.25	-1.52	0.08	-1.1	-1.86
Sinop	**MAE**	32.35	31.1	17.21	19.76	22.58
	RMSE	42.19	34.43	18.42	22.37	36.08
	R2	-0.7	-0.13	0.67	0.52	-0.24

According to models' error measures the Sarima model performs better overall. Sarima is the most effective technique, while averaging comes in second. The Arima approach is the least effective. Since Sarima takes seasonality into account, it is at the forefront. It performs similarly to Sarima since averaging eliminates variation. Due to the absence of data, the machine learning approach performed rather poorly because it was developed as monthly models. Training and, thus, test success were low because of limited amount of data. While time series models can use lagged values directly, for the XGBoost model, lag features may need to be defined manually to train the time dependence. The model can be supplemented by adding different lag features (e.g. by adding data from the past 3, 6, 9, 12 months as features). XGBoost models may face the problem of overfitting in small data sets. In order for XGBoost models to understand seasonal variables, the correct assignment of additional variables may be required. In addition, while differencing can be performed to ensure stationarity, seasonality can be customized by using the Fourier transform. Within the scope of the study, feature engineering was performed while creating the XGBoost model and different attributes were used as input. In addition, lag features can be used in future studies. Box plots are used to presents the station-based evaluations of error distributions. For all stations, the average Sarima model MAE error was 15.87, the average RMSE error was 18.77, and the average R^2 error was 0.40. Figures 10 and 11 exhibit the stations' error distributions.

Examining the models' error distributions reveals that the Sarima model performs similarly at every station. The error distribution is more constrained in the Sarima model. This demonstrates how more balanced the model is. The average distribution of the errors is balanced, as indicated by the median value being near zero. It is evident that the outliers produced by Arimax and XGBoost might result in significant deviations.

When the performance metrics are analyzed on a station basis, the SARIMA time series model performed better than the other models, since it reflects seasonality best. One

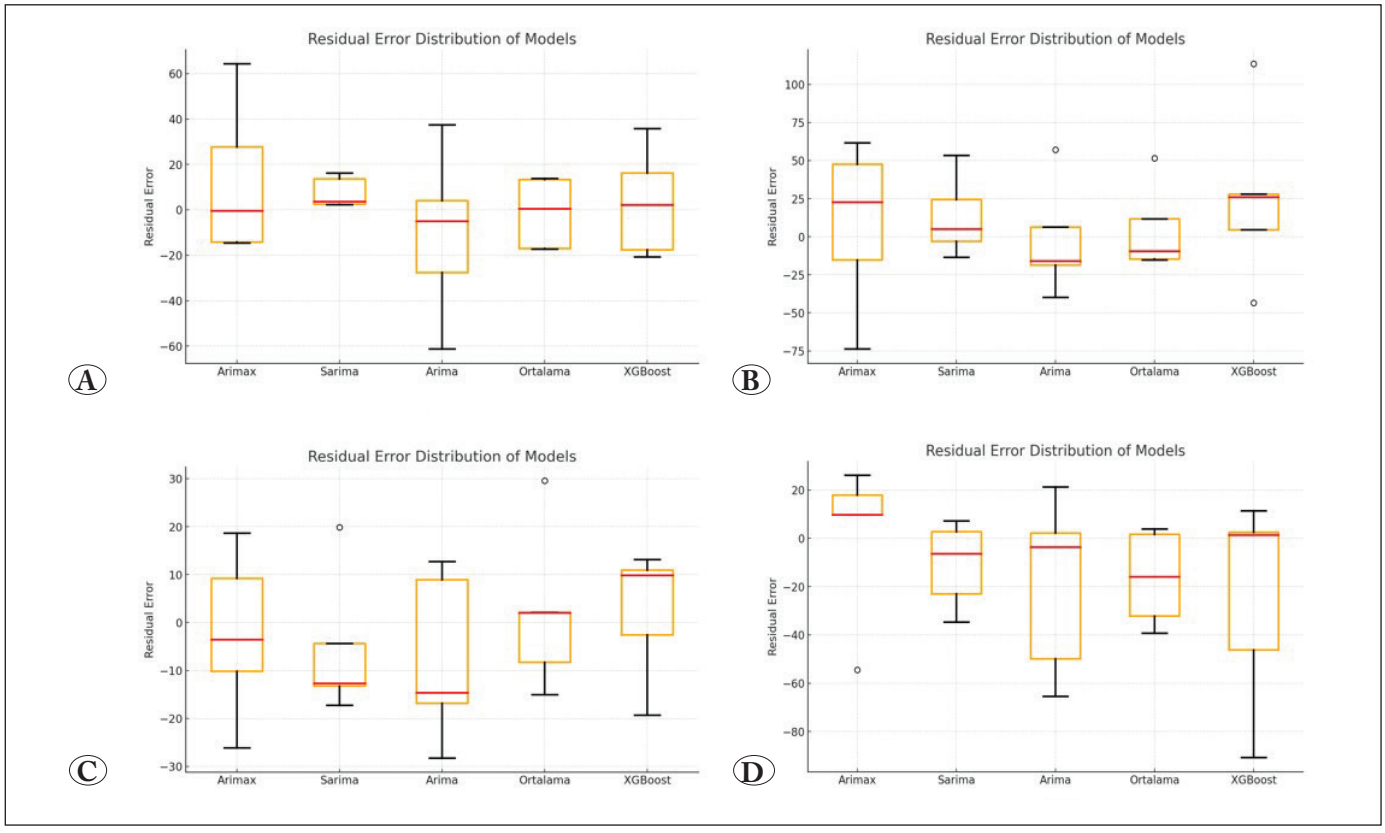

Figure 10. Error distributions; **(A)** Zonguldak station, **(B)** Bartin station, **(C)** Karabuk station, **(D)** Inebolu station.

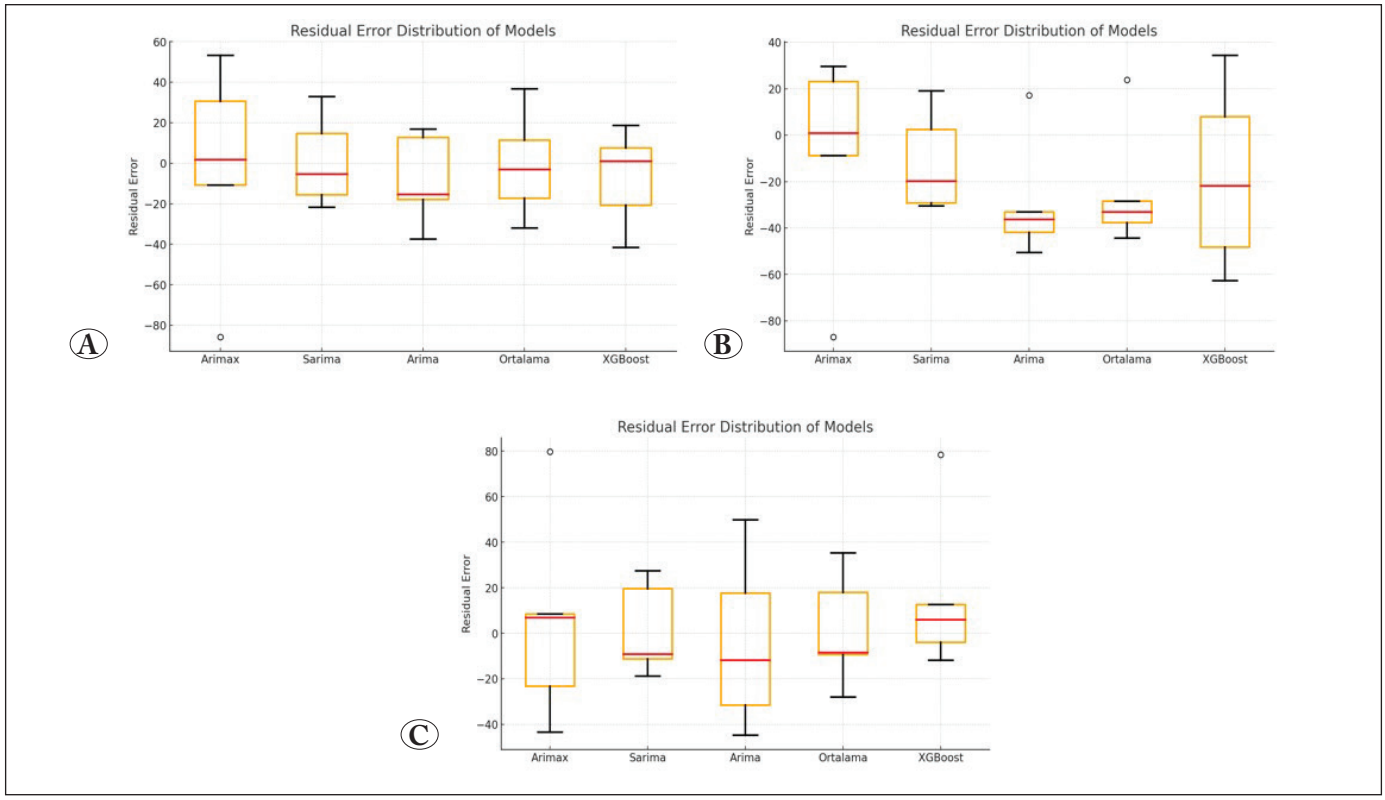

Figure 11. Error distributions; **(A)** Bolu station, **(B)** Duzce station, **(C)** Sinop station.

of the important reason why XGBoost models give different results on a station basis is that the distribution of missing data varies on a station basis, which means that the number of training data varies on a station basis. In addition, the regularity of precipitation patterns facilitates forecasting. This may have a reducing effect on error rates at stations with rainfall regularity. In future studies, station locations (Latitude-Longitude) can be added to the models. In machine learning models, cross-validation approaches are preferred instead of classical statistical tests. In order to evaluate models with similar approaches, common performance measures (MAE, RMSE, and R^2) are used in this study.

4. Conclusions

In this study, time series, statistical, and machine learning techniques were used to fill in the missing values in the precipitation data from 2000 to 2023 that were recorded at principal meteorological stations in the provinces that make up the western Black Sea basin. The Sarima time series approach the most consistent model in the analyses, which investigated test data and error metrics for every precipitation station. The trend in the first precipitation data was captured using Sarima models.

Sarima is highlighted in the results when the performance metrics are looked at. Since Sarima models account for seasonality, their error performance is lower than that of other time series methods. Since the SARIMA model forecasts time series based on past trends and seasonal patterns, its performance can be significantly degraded in situations involving unexpected, sudden and unusual changes, such as extreme weather events. When the data contains outliers, it can be overly sensitive to past data and may have difficulty predicting similar events in the future. XGBoost modeling requires large data sets. In limited data sets, XGBoost's performance is insufficient. This situation demonstrates the need to validate machine learning techniques using several approaches in models that assess months independently. XGBoost does not forecast directly from past observations like time series models. In order for XGBoost to learn time dependence, lag features need to be created manually. In small data sets, overfitting for XGBoost models can significantly reduce model performance. For XGBoost models, feature engineering were made by feature engineering within the scope of the study.

The effects of seasonality have been made clear by setting up the models independently for each month. This characteristic has been reflected in the results of the Sarima model, one of the well-known models that takes seasonality into account.

Author contribution: Author Yusuf Kaya: wrote the article by analyzing the study, Author Berna Aksoy: Collected and analyzed data about the study, Author İsmail Hakkı Özölçer: planned and designed the study.

5. References

Bentéjac, C., Csörgo, A., Martínez-Muñoz, G. 2021. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54. https://doi.org/10.1007/s10462-020-09896-5

Bierens, HJ. 1987. ARMAX model specification testing, with an application to unemployment in the Netherlands. Journal of Econometrics, *35*(1):161–190. https://doi.org/10.1016/0304-4076(87)90086-8

Box, GE., Jenkins, GM., Reinsel, G. 1970. Forecasting and control. Time Series Analysis, 3, 75.

Box, GEP., Jenkins, GM., Reinsel, GC. 2008. Time series analysis: Forecasting and control (4th ed.). New Jersey: John Wiley & Sons, Inc. ISBN: 978-0470272848

Chen, T., Guestrin, C. 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, s. 785–794. ACM. https://doi.org/10.1145/2939672.2939785

Cryer, JD., Chan, KS. 2008. Time series analysis: With applications in R. New York: Springer-Verlag. ISBN: 978-0-387-75958-6.

Dabral, PP., Murry, MZ. 2017. Modelling and forecasting of rainfall time series using SARIMA. Environmental Processes, 4:399–419. https://doi.org/10.1007/s40710-017-0226-y

Dempster, AP., Laird, NM., Rubin, DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), *39*(1):1–38. https://doi.org/10.2307/2984875

Fan, J., Shan, R., Cao, X., Li, P. 2009. The analysis of the tertiary industry with the ARIMAX model. Journal of Mathematics Research, 1(2):156. https://doi.org/10.5539/jmr.v1n2p156

GarcíaLaencina, PJ., SanchoGómez, JL., FigueirasVidal, AR., Verleysen, M. 2009. K nearest neighbours with mutual information for simultaneous classification and missing data imputation. Neurocomputing, **72**(7–9):1483–1493. https://doi.org/10.1016/j.neucom.2008.11.026

Huntra, P., Keener, TC. 2017. Evaluating the impact of meteorological factors on water demand in the Las Vegas Valley using time-series analysis: 1990–2014. International Journal of Geo-Information, 6(8), 249. https://doi.org/10.3390/ijgi6080249

Schafer, JL., Graham, JW. 2002. Missing data: Our view of the state of art. Psychological Methods, 7(2):147-177. https://doi.org/10.1037/1082-989X.7.2.147

Stekhoven, DJ., Bühlmann, P. 2012. MissForest: Nonparametric missing value imputation for mixedtype data. Bioinformatics, *28*(1):112–118. https://doi.org/10.1093/bioinformatics/btr597

Sutthichaimethee, P., Ariyasajjakorn, D. 2017. Forecasting energy consumption in short-term and long-term periods by using the ARIMAX model in the construction and materials sector in Thailand. Journal of Ecological Engineering, *18*(4):52-59. https://doi.org/10.12911/22998993/74396

Valipour, M. 2012. Number of required observation data for rainfall forecasting according to the climate conditions. American Journal of Scientific Research, *74*:79–86.

Valipour, M. 2015. Long-term runoff study using SARIMA and ARIMA models in the United States. Meteorological Applications, *22*:592–598. https://doi.org/10.1002/met.1491

Wangwongchai, A., Waqas, M., Dechpichai, P., Hlaing, PT., Ahmad, S., Humphries, U. W. 2023. Imputation of missing daily rainfall data: A comparison between artificial intelligence and statistical techniques. MethodsX, 11, 102459. https://doi.org/10.1016/j.mex.2023.102459

Vincent, P., Larochelle, H., Bengio, Y., Manzagol, PA. 2008. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, s. 1096–1103. ACM. DOI:10.1145/1390156.1390294

Wang, J., Du, YH., Zhang, XT. 2008. Theory and application with seasonal time series (1st ed.). Nankai: Nankai University Press.

Waqas, M., Humphries, UW., Hlaing, PT., Wangwongchai, A., Dechpichai, P. 2024. Advancements in daily precipitation forecasting: A deep dive into daily precipitation forecasting hybrid methods in the tropical climate of Thailand. MethodsX, 12, 102757. https://doi.org/10.1016/j.mex.2024.102757

Zhang, G. P. 2003. Time series forecasting using a hybrid ARIMA and neural network model. Neurocomputing, 50:159–175. https://doi.org/10.1016/S0925-2312(01)00702-0