COMPARISON OF MISSING DATA IMPUTATION METHODS APPLIED TO DAILY TEMPERATURE AND PRECIPITATION DATA IN TURKEY

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

DİDEM GEZGEN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
STATISTICS

AUGUST 2023

Approval of the thesis:

COMPARISON OF MISSING DATA IMPUTATION METHODS APPLIED TO DAILY TEMPERATURE AND PRECIPITATION DATA IN TURKEY

submitted by **DİDEM GEZGEN** in partial fulfillment of the requirements for the degree of **Master of Science** i**n Statistics, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Özlem İlk Dağ
Head of the Department, **Statistics** _____

Prof. Dr. Ceylan Yozgatlıgil
Supervisor, **Statistics, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Burçak Başbuğ Erkan
Statistics, METU _____

Prof. Dr. Ceylan Yozgatlıgil
Statistics, METU _____

Assoc. Prof. Dr. K. Demirberk Ünlü
Industrial Engineering, Atılım University _____

Date: 07.08.2023

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name Last name : Didem Gezgen

Signature :

ABSTRACT

COMPARISON OF MISSING DATA IMPUTATION METHODS APPLIED TO DAILY TEMPERATURE AND PRECIPITATION DATA IN TURKEY

Gezgen, Didem
Master of Science, Statistics
Supervisor : Prof. Dr. Ceylan Yozgatlıgil

August 2023, 120 pages

A significant portion of the data under analysis contains missing values, which hinders the generation of meaningful results, particularly when dealing with time-dependent data where the order of observations is crucial. This issue leads to unreliable outcomes in statistical analyses applied in fields such as meteorology and economy. To address this challenge, handling missing values meticulously in time-dependent data is imperative. In this thesis, daily average temperature and total precipitation data, obtained from the General Directorate of Meteorology of Turkey, were utilized. The primary objective was to impute the missing values in these datasets using various methods and subsequently compare their performance. Missing values were intentionally introduced into the temperature and precipitation data. The methods employed for imputation included Simple Arithmetic Average Method (SAA), K-Nearest Neighbor Method (KNN), Random Forest Method (RF), Multiple Imputation by Chained Equation Method (MICE), and Generalized Adversarial Imputation Network (GAIN). The outcomes were assessed based on the Root Mean Square Error (RMSE), Coefficient of Variation of Root Mean Square Error (CVRMSE), and Nash-Sutcliffe Efficiency (NSE). The results indicated that

Random Forests exhibited superior performance in most cases, followed by KNN and GAIN.

Keywords: General Adversarial Imputation Network (GAIN), Multiple Imputation by Chained Equation (MICE), Nash-Sutcliffe Efficiency (NSE), Meteorological Data, Random Forest (RF)

ÖZ

TÜRKİYE'DE GÜNLÜK SICAKLIK VE YAĞIŞ VERİLERİNE UYGULANAN KAYIP VERİ ATAMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Gezgen, Didem
Yüksek Lisans, İstatistik
Tez Yöneticisi: Prof. Dr. Ceylan Yozgatlıgil

Ağustos 2023, 120 sayfa

Bir çok veri analizinde kullanılan verilerin önemli bir kısmı eksik değerler içermekte ve özellikle zaman bağımlı verilerde gözlem sırasının önemli olduğu durumlarda anlamlı sonuçların elde edilmesini engellemektedir. Bu durum, meteoroloji ve ekonomi gibi alanlarda uygulanan istatistiksel analizlerde güvenilmez sonuçlara yol açmaktadır. Bu zorluğun üstesinden gelmek için zaman serisi verilerinde eksik değerlerin titizlikle ele alınması gerekmektedir. Bu tez çalışmasında, Türkiye Meteoroloji Genel Müdürlüğünden elde edilen günlük ortalama sıcaklık ve toplam yağış verileri kullanılmıştır. Temel amaç, bu veri setlerindeki eksik değerleri çeşitli yöntemlerle tamamlamak ve performanslarını karşılaştırmaktır. Sıcaklık ve yağış verilerine kasıtlı olarak eksik değerler eklenmiştir. Kayıp verileri doldurmak için kullanılan yöntemler arasında Basit Aritmetik Ortalama Yöntemi (SAA), K-En Yakın Komşu Yöntemi (KNN), Rastgele Orman Yöntemi (RF), Zincir Denklemle Çoklu Tamamlama Yöntemi (MICE) ve Genelleştirilmiş Rakip Tamamlama Ağı (GAIN) yer almaktadır. Sonuçlar, Kök Ortalama Kare Hatası (RMSE), Kök Ortalama Kare Hatasının Değişim Katsayısı (CVRMSE) ve Nash-Sutcliffe

Verimliliği (NSE) temel alınarak değerlendirilmiştir. Sonuçlar, çoğu durumda Rastgele Ormanların üstün performans sergilediğini, onu KNN ve GAIN yöntemlerinin takip ettiğini göstermiştir.

Anahtar Kelimeler: Üretken Düşman Ağları (GAIN), Zincirli Denklemlerle Çok Değişkenli Atama (MICE), Nash-Sutcliffe Verimliliği (NSE), Meteoroloji Veri, Rastgele Orman (RF)

To my family…

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to the following individuals who have provided invaluable support and guidance throughout the completion of this thesis:

I am deeply grateful to my supervisor, Prof. Dr. Ceylan Yozgatlıgil, for their continuous support, valuable insights, and patient guidance. Their expertise and encouragement have been instrumental in shaping the direction of this research. I am truly grateful to her for entrusting me with the opportunity to work under her guidance. I sincerely appreciate her for believing in me and giving me the chance to learn from her.

I would also like to thank my colleagues Erdinç Erdoğan, İlknur Ceyda Pulatsü, Zahide Merve Karabacak and Cem Kağan Yaşar for encouraging me all the time.

I am grateful for the unwavering support and encouragement of my dear, Berkay Filiz, during this challenging time. He has been a constant source of strength and guidance, and I feel truly blessed to have him by my side.

Finally, I want to express my heartfelt appreciation to my parents Sibel and Soner GEZGEN, and my brother Doğukan GEZGEN for their unwavering support and understanding. Their presence in my life always gives me strength and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

ABBREVIATIONS

MAR	:	Mediterranean Region
AR	:	Aegean Region
BSR	:	Black Sea Region
MR	:	Marmara Region
EAR	:	Eastern Anatolia Region
CAR	:	Central Anatolia Region
SAR	:	Southeastern Anatolia Region
MAR	:	Missing Data at Random
MCAR	:	Missing Data Completely at Random
MNAR	:	Missing Data Not at Random
GAN	:	General Adversarial Network
MI	:	Multiple Imputation
KNN	:	K Nearest Neighbor
RF	:	Random Forest
MICE	:	Multiple Imputation by Chain Equation
SAA	:	Simple Arithmetic Average
GAIN	:	General Adversarial Imputation Network
MCMC	:	Monte Carlo Markov Chain
MLP	:	Multilayer Perceptron

PPCA	:	Probabilistic Principal Component Analysis
PMM	:	Predictive Mean Matching
CART	:	Classification and Regression Trees
ANN	:	Artificial Neural Network
NSE	:	Nash-Sutcliffe Efficiency
RMSE	:	Root Mean Square Error
CVRMSE	:	Coefficient of Variation of the Root Mean Square Error
MAE	:	Mean Absolute Error
MSE	:	Mean Square Error
KNTT	:	K-Nearest Temperature Trends
NRMSE	:	Normalized Root Mean Square Error
GAMIN	:	General Adversarial Multiple Imputation
SD	:	Standard Deviation
CV	:	Coefficient of Variation

LIST OF SYMBOLS

SYMBOLS

N,n	:	Number of Observations
K	:	Constant
$\mathcal{X}$	:	Random Variable
M	:	Random Variable
$\widetilde{\mathbf{X}}$	:	Random Variable
Z	:	Noise Variable
$\mathcal{L}$	:	Loss Function
d	:	Predicted Value
f	:	Actual Value
$\bar{f}$	:	Mean of Actual Values
$\bar{X}$	:	Vector of Imputations
$\hat{X}$	:	Completed Vector
H	:	Hint Vector
D	:	Discriminator
G	:	Generator

CHAPTER 1

INTRODUCTION

Analyzing data to discover valuable insights has become progressively crucial in the contemporary world. This data can be used to create models for scientific studies in a variety of fields, including economy, meteorology, and industry. By analyzing data, researchers can gain a better understanding of the world and make informed decisions based on this knowledge. With the help of technology, people are able to collect and interpret data on a scale that was once impossible, paving the way for new discoveries and advancements in a wide range of fields. Dealing with missing data has become a significant challenge in scientific studies. It is a common issue that hinders the accuracy of models and results in unreliable outcomes. The problem has been recognized for some time, and since 1989, researchers have been working to develop techniques to address it [1]. By finding ways to handle missing data, scientists can improve the quality of their research and gain a better understanding of the world.

Missing values can be a challenge when analyzing datasets, and this is especially true for meteorological data. These missing values can arise from various sources such as environmental or machine problems that occurred in the past. It's crucial to address this issue to ensure the accuracy of models and reliable outcomes. Since handling missing data is crucial in time series analysis, as all data points are analyzed sequentially. The literature suggests two main approaches to address this issue: deleting the missing values or replacing them with meaningful values. Both methods have their advantages and disadvantages, and it is important to handle missing data with meticulousness. However, deleting these values may not be the best solution due to the correlation between the time series values. It's better to replace or impute

these missing values with approximate values based on the data distribution. This approach usually leads to better results.

1.1 General Information

Temperature and precipitation values are crucial factors in various fields, including climate change, agricultural studies, and groundwater studies. Accurate data analysis plays a vital role in obtaining reliable results in these studies. Unfortunately, many meteorological stations in certain regions have lost their temperature and precipitation data due to various issues. To address this problem, missing data imputation studies are conducted to recover usable data that is as close to the original data as possible. These studies have been carried out using various methods in the past and are still ongoing today, utilizing new techniques such as deep learning, machine learning, and traditional methods. When filling in incomplete meteorological data, it is essential to exercise caution since these data are time and location-dependent.

1.2 Objectives of the Thesis

Various methods have been utilized to address the lack of meteorological data in different parts of the world. One such approach involves utilizing data from neighboring stations to obtain more accurate results. Simply relying on a single station's data may not yield precise outcomes due to the absence of spatial information. Hence, filling in missing data in meteorological stations requires the use of different techniques [2].

In this study, the aim is to test different techniques on daily meteorological data with the help of neighboring stations. The main objective is to determine the accuracy of these methods and compare the results to the original data. The data to be used in this study is obtained from the General Directorate of Meteorology of Turkey, and

these data include daily average temperature and daily total precipitation amount of the stations belonging to seven different geographical regions of Turkey. Therefore, the goal is to complete the missing data separately for these seven regions that have different climatological traits. Daily series are prefered here because working with daily series are crucial to prevent flood type events and there are already studies on imputation of monthly series (Yozgatligil et al, 2013). The techniques that work well for monthly series may not yield satisfactory results when applied to daily series due to their frequency and volatility.

The methods are used in this study are K-Nearest-Neighbors (KNN), Random Forest (RF), Multiple Imputation by Chain Equation (MICE), Simple Arithmetic Average method (SAA) and also Generative Adversarial Network (GAN) methods.

1.3 Thesis Outline

The problems of missing values in meteorological data namely daily total precipitation and average temperature were mentioned, and information was given about how to fill these missing values with different methods discussed in In Chapter 1.

In Chapter 2, by making a literature review, studies on the methods to be used in this thesis were mentioned. The results of these studies were also shown.

In Chapter 3, ways to deal with missing values and the basic information of missing data structure are explained. The methodologies of the models to be used were given and explained in detail. The methodologies of the selected methods to compare the results of these models were given and detailed explanations were made for these methods as well.

In Chapter 4, general information about the data was given. Stations were separated according to regions; data sets were created to be used by making data adjustments. Correlation values between stations were examined for both precipitation and temperature data. Target and reference stations were selected for each region and prepared for the models.

In Chapter 5, data sets were used in the models and each model was tuned to give the best results. After finding the values that gave the best results, the model performances were compared for both precipitation and temperature for each region. The outline of the thesis is given in Figure 1-1.

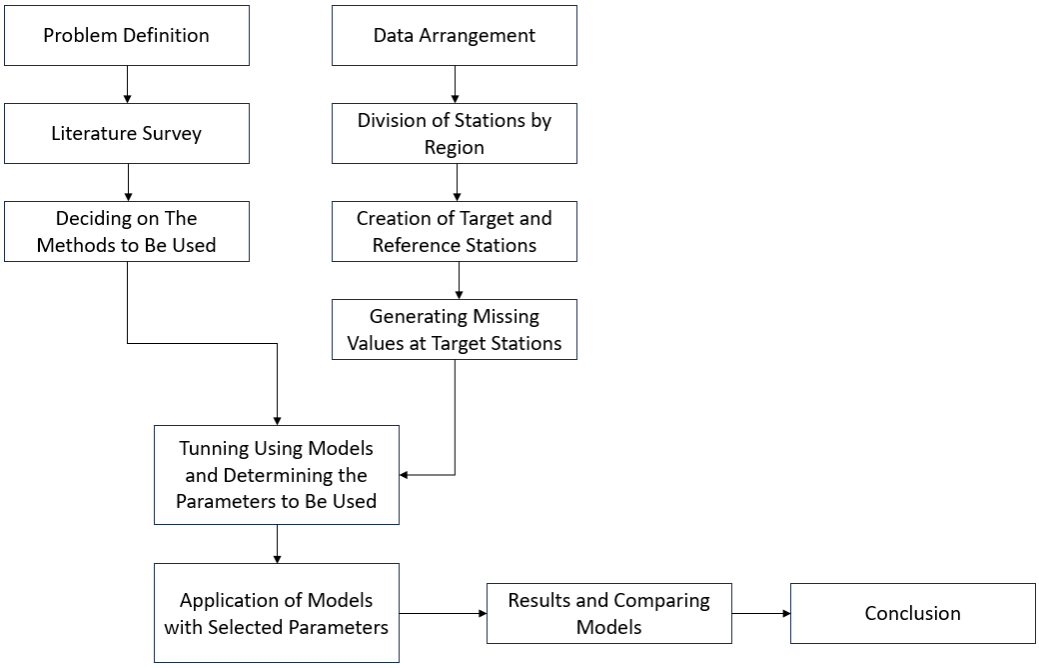

Figure 1-1. The Outline of the thesis

CHAPTER 2

LITERATURE REVIEW

In this chapter, literature review of the related topic is given in detail with the help of different studies.

2.1 K-Nearest Neighbors Method

Aieb et al. (2019) studied different types of missing value imputation methods like hot-deck, k-nearest-neighbors (KNN), simple average method (SAM), multiple imputation (MI) and linear regression (LR) to impute daily rainfall data in Algeria. In this study, these methods applied to data showing daily precipitation from January 1982 to until the end of 2014 were examined, and Root Mean-Square-Error (RMSE) was applied to examine which method was better. According to the study, these methods were adapted to datasets with different loss percentages (4%, 8%, 12%, 16%) and it was examined which method gave better results. When some stations were examined, it was observed KNN gave the best results for all loss percentages [3].

In 2017, Kiani and Saleem studied data imputation methods for daily temperature meteorology data. Their proposed method was K-nearest-temperature-trends (KNTT) and compare it with KNN for their 30 years daily temperature data for Pakistan. They used RMSE to evaluate their models. According to the study, KNTT gave better results than KNN for all stations [4].

Jadhav et al. in 2019 studied the performances of different imputation methods. Mean imputation, predictive mean matching, linear regression, median imputation,

Bayesian Linear Regression methods were used on five different data sets. To evaluate models, RMSE was used. The results of these models at different missing percentages were investigated and a comparison was made by looking at the Normalized Root Mean Square Error (NRMSE) value. According to the study, among these models, it can be said that the KNN imputation method gave the best results [5].

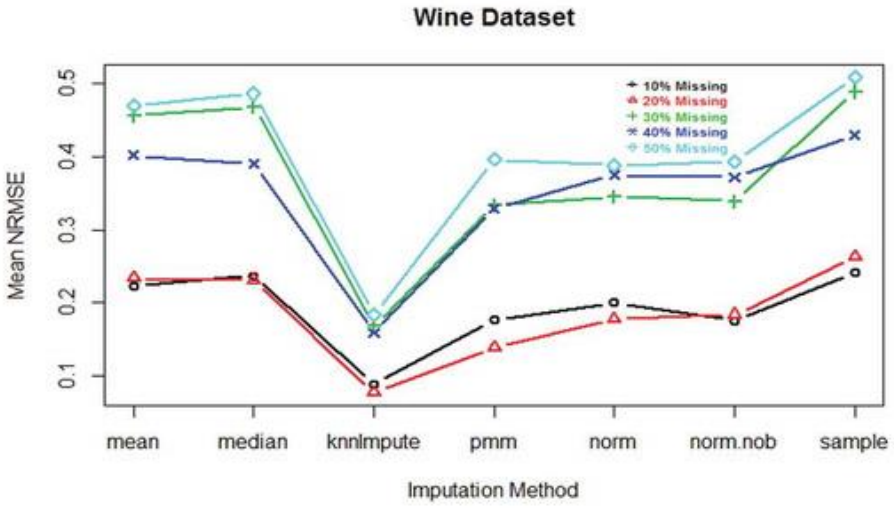

Figure 2-1. Mean NRMSE values for imputation methods [5]

In 2014, Thirumahal and Patil studied KNN and ARL (Autoregressive Model) to impute missing values and to evaluate the results NRMSE is used. Below results are obtained [6].

Table 2-1. NRMSE values for KNN and ARL, adapted from [6]

% missing	ARL	KNN
10	0.08004	0.12
15	0.0912	0.28
20	0.17	0.33

According to the study, KNN imputation gave better results when *k* is between 10 and 20. For overall, ARL imputation method gave better results than KNN imputation method [6].

Another study on KNN imputation is the study done by Jerez et al (2010). They studied missing data imputation techniques for breast cancer problem. The study was divided into as traditional and modern methods and the data of 3679 women from 32 different hospitals from Spanish Breast Cancer Research Group were examined. Multilayer perceptron (MLP), self-organization maps (SOM) and KNN were used to impute values. As a result of the study, it has been revealed that machine learning methods give better results than traditional methods [7].

2.2 Random Forest

Another missing value imputation study is missing precipitation imputation using Random Forest which is done in 2020 by Mital et al (2020). This study was carried out using 10 years of daily data and 97 stations. To impute missing values the Random Forest method was used with reference stations and to examine the performance of the model Nash-Sutcliffe Efficiency (NSE) was used. According to this study, it is said that the correlation between the target and reference stations is important, and results demonstrated that a few highly correlated reference stations are more significant than many of weakly correlated reference stations [8].

In a missing data study for long daily precipitation data, KNN, Probabilistic-Principal-Component-Analysis (PPCA), Random Forest, Mean, and Multiple Imputation by Chain Equation methods were used for different rates of missing data (5%, 10%, 20%, 30%) (Addi et al., 2022). In their study they used data which is from 1976 to 2012 and contains 40 stations in Ghana. Using methods such as Kolmogorov-Smirnov, RMSE, and MAE they compared the performance of the

methods that they used. According to the paper, Random Forest and PPCA performed better than other methods for all missing percentages can be said [9].

To impute environmental missing data, Dixneuf et al. (2021) studied RF, MICE and KNN. In this study, it was seen that RF had a better performance [10].

In 2009, Pantanowitz and Marwala used methods such as RF, Multilayer Perceptron (MLP), Fuzzy Inference Systems (FISs), and Genetic Algorithms (GAs) to a dataset which is a study that was made in 2001 and it is about HIV. The data types contain both integer and binary values. As a result of this study, one can be said that the RF model outperforms other models that are investigated in [11].

The study done by Shah et al. in 2014, aimed to compare missing value imputation methods like RF and MICE to a data set based on electronic health record. It was found that when RF is used within the framework of MICE, it gave better results, especially in imputing data with binary and continuous variables. In addition, they concluded that when missing data is non-ignorable, using the RF method to impute missing values is a better way [12].

Another study about RF imputation was made by Tang and Ishwaran in 2017. They compared different RF-based models with the models used in single imputation and multiple imputation. They used RMSE when making this comparison of performances. To summarize the results of this study, one can be said that using RF-based models is more competitive and gives better results when imputing lost data than other models [13].

In 2019, Kokla et al. conducted a study on imputing the missing data due to device problems and technical reasons by applying the RF method. They also used KNN, mean imputation and least squares imputation to compare with RF. To measure the performance of methods and compare them, they used the correlation between

imputed values and the correct values and NRMSE. It is observed that the RF model gives better results than other methods in the studied data set [14].

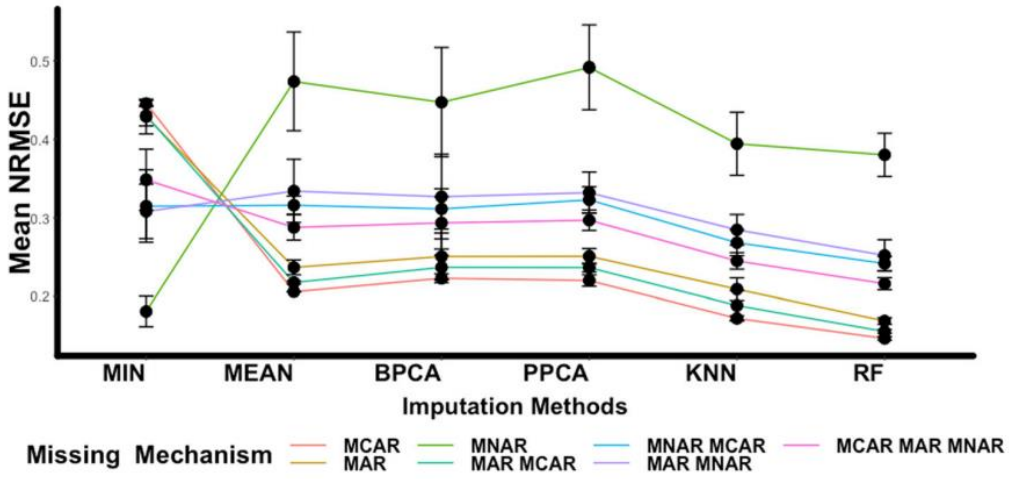

Figure 2-2. Mean NRMSE values for imputation methods [14]

2.3 Simple Arithmetic Average

The simple Arithmetic Average Method was used to compare with other missing value imputation methods for the daily rainfall dataset in Bangladesh (Jahan et al., 2019). For this method, it was used for the target station by taking the average values for the same day using reference stations [15].

Another study was done by Rahman et al in 2017 to impute missing daily precipitation data in the Kelantan region and the period was between 1975 to 2014. In this study, five different imputation methods (Simple Arithmetic Average (SAA), Inverse Distance Weighting method, Normal Ratio method, Geographical Coordinate and Correlation Coefficient Weighting method) were used for data that have different percentages of missing values. For performance criteria RMSE, CVRMSE and MAE were used. To conclude this study, all five imputation methods gave similar results [16].

Sattari et al. also conducted a study to estimate missing values. Different methods were used to impute missing monthly rainfall data from six stations. SAA, the multiple linear method gave good results [17].

Another study related to the imputation of monthly precipitation data is done by Sanusi et al. (2017). SAA, Normal Ratio (NR), Inverse Distance and Coefficient of the Correlation Weighting methods were used to impute the missing precipitation data of stations located in the city of Makassar. When looking at the RMSE and MAE values used to compare the methods, the NR methods is more suitable for imputing the data. [18]

2.4 Multiple Imputation by Chained Equation

Aguilera et al. (2020) worked on estimating the missing values in precipitation data. When the percentage of this missing data was large, it became more and more difficult to impute. Aguilera et al. used three different methods in this study while imputing precipitation data which is based on 1975-2017 with high loss percentage. The methods used are Spatio Temporal Kriging (STK), RF and Multiple Imputation by Chained Equations with PMM (MICE-PMM) [19].

Another study for imputation with MICE was done by Norazizi and Deni in 2019. A target station was selected from the precipitation information obtained from 8 different stations in the state of Pahang, Malaysia, and MICE, Artificial Neural Network (ANN) and Expectation Maximization algorithm were used in the study for data with different loss percentages. RMSE and MAE were used to compare the results of the methods [20].

In 2021, Abdullah et al. studied missing value imputation for both daily minimum and maximum temperature and daily precipitation. The methods that they used for their studies were MICE and PMM [21].

Carvalho et al. studied MICE to impute daily precipitation data in Brazil. They compared multiple imputation techniques with geo-statistical methods. To compare results MSE was used. As a result of this study in 2017, it was found that the MICE method gave better and more consistent results [22].

The study which was done by Turrado et al. was based on imputing missing values in solar radiation. According to the study, solar radiation had high variability, so imputing missing values was a bit more difficult. Looking at RMSE and MAE values Inverse Distance Weighting (IDW) and Multiple Linear Regression (MLR) had similar results, on the other hand MICE represented better results than others [23].

Another study which was about MICE was done by Wesonga (2015). In their study, there were daily wind speed data recorded from 1995 to 2008 with missing values. After imputation, they reached reliable daily wind speed data close to actual values [24].

Diouf et al. (2022) made research to compare different missing value imputation methods like KNN, MICE, RF, Probabilistic Principal Component Analysis (PPCA), and Time Series Missing Value Imputation to their daily rainfall and temperature data recorded between 1973 and 2020 in Senegal. They used these methods for different percentages of missing values (5%, 10%, 20%, 30%, 40%). To compare the results of these methods, they used the values from Taylor's Diagram, Kling-Gupta Efficiency (KGE). According to the study, when the percentage of missing values is low, it was seen that all the methods used gave similar outputs [25].

Table 2-2. Table of Kling Gupta Efficiency results for 5% missing for temperature and rainfall, adapted from [25]

Stations	Methods	Tmax	Tmin	Tmoy	Rainfall
Diourbel	imputeTS	0.9835714	0.9831029	0.9824979	0.9787246
	Knn	0.9872866	0.9922031	0.9958493	0.9533001
	Mice	0.9820596	0.9857274	0.9929311	0.90575048
	missForest	0.9882202	0.9922049	0.9957484	0.9780226
	ppca	0.9614748	0.9631652	0.9651033	0.9802646

2.5 General Adversarial Network Imputation

Another method to impute meteorological time series data is Generative Adversarial Imputation Method (GAIN). For this method, Popolizio et al. used temperature data which has 98 stations in Italy in 2019. At the end of the study, by looking at RMSE, when GAIN was used to impute missing values, they found that the results were very close to the real data [26].

Low et al. (2020) conducted a study to impute the gaps in the data to be used in parking and observation study conducted in Singapore, which includes information on drivers and trucks. General Adversarial Multiple Imputation (GAMIN), GAN and KNN were used to impute and MAE was used to evaluate the results. At the end of the research, it was noted that the GAMIN model gave better results [27].

Another study that can be shown as an example of GAN imputation was done by Dong et al. (2021). The aim of this study was to use GAN to impute missing data and compare the outputs of the GAN model with the outputs of other models used in this field like random forest and multiple imputation by chained equation. In the study performed on data with different percentages of loss (20% and 50%), random forest and GAN produced better results than MICE when there was 20% missing

data. When the percentage of missing data was 50, GAN gave better results than the others. NRMSE was used to evaluate these results [28].

Wang et al. (2018) used the GAN imputation method to impute missing values in two different data sets. In this study, different loss percentages were studied (10%, 30%, 50%, 70%, 90%). They also compared all outputs using mean imputation, linear regression and KNN. According to the study, if there are powerful correlations between variables in datasets, linear regression, and KNN are suitable for these datasets. Also, when there are high loss rates, the GAN method is more competitive than other methods. The GAN method can generate infinite data to impute missing data after training is completed. To evaluate the performances of methods cosine similarity was used [29].

Table 2-3. Table of Cosine similarity for different missing percentages of methods, adapted from [29]

Missing Rate	KNN	Mean	LR	GAN
10.00%	0.99720	0.99878	0.99708	0.98182
30.00%	0.99156	0.99555	0.99155	0.98226
50.00%	0.98975	0.99236	0.98596	0.98194
70.00%	0.98426	0.98927	0.97876	0.98121
90.00%	0.98295	0.98645	0.97346	0.98117

According to Table 2-3, one can say that for a small loss percentage all methods gave similar results but as the lost data rate increased, the values of the methods other than GAN deviated a little from the real data, and the GAN method remained stable [29].

A study on the filling of lost data in the sensors on a bridge in China was made by Jiang et al. (2021) Although the errors increased with the increase in the missing data rate, it was seen that the filled data gave results close to the real data [30].

2.6 Studies in Turkey

Sahin and Cigizoglu conducted a study in 2010 to impute the missing values in monthly precipitation and temperature data in Turkey. The study utilized temperature and precipitation data of 232 stations between 1974-2002 were used. Linear Regression (LR) and EM are used to impute missing values. The results showed that the EM method was more reliable in imputing missing values compared to the LR method [31].

Yozgatlıgil et al. (2013) studied about the comparison of missing value imputation methods for monthly mean temperature and total precipitation data in Turkey. Turkey was divided into 7 regions and target stations were determined for each region. Reference stations with a high correlation with these target stations were also selected. It was aimed to impute the missing data in these target stations from the reference stations with complete data by using SAA, MLP, Monte Carlo Markov Chain based on Expectation Maximization (EM-MCMC, namely MICE), Normal Ratio (NR). To compare the results Coefficient of Variation of the Root Mean Square Error (CVRMSE), RMSE were used. In this study, MICE gives the best results for all regions under all missing data percentages [32].

Another study conducted by Dikbas in 2017 utilized precipitation data from 70 stations in 21 basins in Turkey and used a frequency-based imputation model (FBI), which was compared to EM and MLR models. The study found that the FBI model produced more accurate results compared to the other models. It's important to have reliable methods for imputing missing data to ensure accurate analysis and decision-making [33].

One study which was done in Turkey by Kalkan et al. in 2018, where they focused on imputation methods in handling missing data. The study utilized IRT-based imputation (MBI), EM, MI, and regression imputation methods to compare their

accuracy. The result showed that the MBI method produced better results compared to the other methods [34].

Katipoğlu and Acar (2021) conducted a study to impute missing values in monthly temperature data in Horasan station. By using neighboring stations with the same characteristics, the temperature values at the Horasan station were imputed using ANN. The study found that the ANN model produced good results when imputing missing values in monthly temperature data [35].

Başakın et al. (2023) conducted a study on the imputation of missing values in solar radiation data, which is one of the important meteorological variables. In this study for Konya province, data sets with different missing percentages were created (5%, 10%, 20%, 30%) and these data were imputed using machine learning methods such as Extreme Gradient Boosting (XGBoost), RF and Multivariate Adaptive Regression Spline (MARS) and interpolation techniques. MAE, RMSE, NSE and Kling-Gupta Efficiency (KGE) were used to compare the performances of these methods. It has been seen that machine learning methods give better results [36].

In the context of missing data imputation, GANs can be useful tool. The task of missing data imputation involves filling in the missing values in a dataset. GANs offer a unique approach by leveraging their generative capabilities to generate plausible missing data based on the available observed data.

One advantage of using GANs for missing data imputation is that they can capture complex dependencies and distributions present in the data, allowing for more accurate imputations compared to simpler methods like mean imputation or regression-based approaches. GANs also have the potential to generate diverse imputations, providing more flexibility in handling uncertainty in the missing data. However, it is important to note that GANs can be computationally expensive to train, and their performance heavily relies on the availability and quality of the

training data. Careful validation and selection of appropriate GAN architectures and training strategies are crucial to ensure reliable imputations in real-world scenarios.

In this study, we want to see the effectiveness of GAN method over other methods chosen in the literature to impute daily meteorological variables under different climatologic regions of Turkey.

CHAPTER 3

METHODOLOGY

3.1 Ways to Deal with Missing Values

When dealing with missing data, it's crucial to consider various approaches. Two common methods are deletion and imputation. There are two data deletion methods. The first is the listwise deletion approach, where observations with missing values in a variable are removed from the data. However, this method may result in losing important information. The second method is pairwise deletion, which excludes missing data on a variable basis. This method aims to use the data to the fullest extent. Alternatively, data can be imputed to prevent information loss. There are different advantages and disadvantages to deleting or imputing lost data. In this study, it is aimed not to lose information by using different imputation methods.

3.2 Missing Data Imputation

Imputation is a technique to replace missing values with meaningful values in order to preserve most of the values in the dataset and get efficient results in further processing. A large amount of missing data may cause distortions in the distribution of the variables, decrease or increase their values and erroneous analyzes can be made by using these data. Some imputation methods are used to prevent such situations. To sum up, the purpose of imputation methods is to create data sets with no missing data that can be used for other analyzes without giving any compromising the accuracy of the results.

There are many types of missing value imputation methods. Some of them are single imputation methods and some of them are multiple imputation methods. The single

imputation is the replacement of missing values with a single value based on observed data. Mean imputation can be an example of this type of imputation. Multiple imputation, on the other hand, creates multiple data sets using statistical models and combines them using appropriate models.

In this study, K-Nearest Neighbor imputation (KNN), Random Forest imputation (RF), Simple Arithmetic Average imputation (SAA), Multiple Imputation by Chain Equation imputation (MICE) and Genereal Adversarial Networks imputation (GAIN) methods are used and compared their results.

3.3 Types of Missing Data

It is very important to deal with missing data in statistical analyses, for this it is necessary to determine the factors that may cause the missing data. It is divided into two; ignorable and non-ignorable. We can say that if the data is randomly missing or if the parameters of interest and the parameters containing these missing data are not related to each other, they can be ignored. When the missing values are not random, it becomes impossible to ignore them when there is a relationship between the missing values and other values.

Rubin (1976) classified the problems causing incomplete data into three categories as Missing at Random (MAR), Missing Not at Random (MNAR), and Missing Completely Random (MCAR). If the probability of data loss is the same in all possible situations, it can be said to be MCAR. So this means that the lack of data has nothing to do with the data. If the probability of missing data within groups defined by the observed data is the same, it is not called MCAR, it is called MAR. When comparing MAR and MCAR, MAR is more general than MCAR. Generally, the MAR assumption is made [37].

3.3.1 Missing Completely at Random (MCAR)

It is the case that the loss of data is not dependent on any variable. In other words, losses in data are purely coincidental and are not related to the underlying characteristics of the data. Therefore, it can be called the simplest type of missing data pattern. It can be expressed as follow:

$$P(M) = P(M|X)$$

where *M* represents missing data, *X* stands for complete data, and *P* is the probability indicator. This means if the probability of being missing equals to the probability of missing complete data, then the pattern of being missing is MCAR.

3.3.2 Missing at Random (MAR)

Missing at Random is a type of missing data model in which the loss is due not to the unobserved variables in the data, is due to the observed variables in the data. That is, the probability of the data being missing is not related to any missing variables, but it is related to the values of other variables in the data. It can be expressed as follows:

$$P(M) = P(M|X)$$

where *M* represents missing data, *X* stands for complete data, and *P* is the probability indicator. This means if the probability of being missing equals to the probability of missing observed data, then the pattern of being missing is MAR. In short, it can be said that the missing data in MCAR is lost by chance, while the missing data for MAR is related to other observed data.

3.3.3 Missing Not at Random (MNAR)

It is a missing data model in which the loss of data depends on the missing data itself or other unobserved variables. It can be said that the missing data is not completely

random or dependent on the observed variables, on the contrary, there is a relationship between the unobserved variables and the values of the missing data. Imputation of missing data in MNAR is more complicated than in MCAR and MAR, as it has a relationship with unobserved variables. In this missing data type, it is possible to obtain incorrect and biased results if the missing data is not properly examined.

3.4 Imputation techniques

The methodology of missing data imputation techniques to be used in this study is explained below.

3.4.1 Simple Arithmetic Average Method (SAA)

In this method, stations that have high correlations and show the same characteristics at the same time as data that has missing values are investigated. The arithmetic averages are calculated by taking the data from these stations, and the missing values are filled with these calculated values.

$$x_m = \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right). \tag{3.1}$$

For the above formula, x_m represents the stations which have missing and x_i represents the other stations. *N* is the number of stations that have high correlation with the target station.

3.4.2 K Nearest Neighbor (KNN)

KNN as an imputation method is used to impute the empty values using the nearby points which means the method based on the concept of closeness between observations of the variables [38]. Missing data is filled with values obtained from

the related stations in the general data by looking observations proximity and the proximity is the definition of Euclidean distance. The algoritm behind KNN is ;

Firstly, missing value is chosen and the other values which is on the same row with the selected missing value.

Then, the k-nearest neighbors of the observations are found.

$$k: \quad 1 \leq k \leq n. \tag{3.2}$$

Secondly, the distance between the target and the neighbor values is calculated.

$$d(x_{i^*}, x_i) = \sqrt{\sum_{i=1}^{n} (x_{i^*} - x_i)^2} \ . \tag{3.3}$$

The distances are ordered and the *k*-nearest neighbors are found based on minimum distances.

$$x_{(i1)}, \dots, x_{(ik)}. \tag{3.4}$$

For these selected missing values, the mean of the *k* closest values found is assigned (Diouf et al., 2022) [25].

$$\frac{1}{k}\left(x_{(i1)} + x_{(ik)}\right). \tag{3.5}$$

In this study, there are five reference stations and the algorithm of KNN is used according to these stations. If there is no data at the target stations, data imputation is provided from the reference stations according to this algorithm.

In R, "caret" package is used to impute missing values. preProcess is the function which is used to impute data in KNN imputation. The method is set to knnImpute to specify it is KNN imputation, *k* is set to optimal *k* value which gives the best value and knnSummary is set by selecting the summary statistic used to load these missing values.

3.4.3 Random Forest (RF)

The machine learning method that fills in missing data in datasets using decision trees is called Random Forest imputation. It can basically be explained as follows.

The random forest model consists of decision trees and is constructed in such a way that the mean square error between the output of each tree and the output of observed can be minimized. The random forest result is obtained by combining the results from all decision trees. It can be used against mixed data types, that is, in data sets containing both continuous and categorical variables. RF also maintains correlations between variables. It reduces bias by using the information provided.

To give more information about decision trees, it can be said that it is a supervised machine learning method used for both classification and regression. There is a tree structure where branches show decision rules, internal nodes properties, and leaves show nodes result. It creates a tree like structure by recursively segmenting data based on features. Each node applies a rule about which branch to follow. This process continues until the result reaches a leaf node. The number of trees that will be generated and the number is the hyperparameters to tune to receive the best performance.

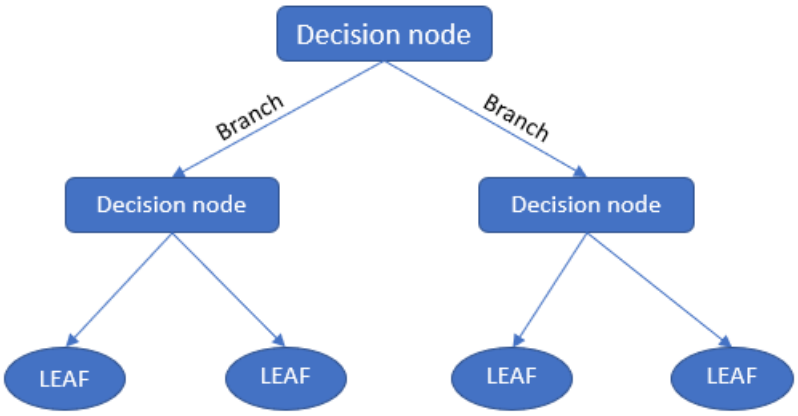

Figure 3-1. Decision Tree structure, adapted from [39]

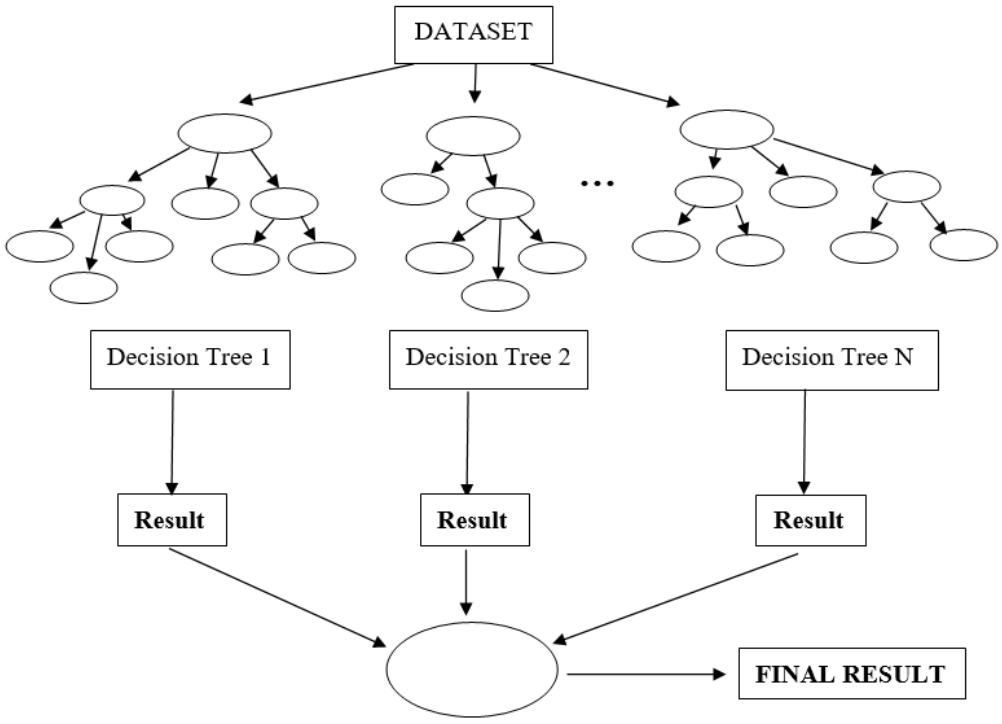

Figure 3-2. Random Forest structure, adapted from [40]

In this study, by using informations from reference stations, the imputation of target stations by random forest model is aimed.

In R, "missRanger" package is used to impute missing values. The methodology behind is the same. When using this method, the formula for imputation is specified, number of trees and other parameters can be used. In addition, the number of nearest neighbors to be used for predictive mean matching assignment and the number of iterations of the model can also be specified. Thus, missing values can be imputed flexibly and conveniently, which is one of the advantages of this package.

Predictive Mean Matching (PMM) is a method used to impute missing data. The idea behind this is to impute by finding the k closest observations with non-missing data and then randomly selecting observed values to use to impute values. These values are known as matched values. The *k* value used to select the matching value is known as the donor pool. The matched values are adjusted to be consistent with the observed values in the donor pool. When used in conjunction with Random Forest imputation, PMM can improve the accuracy of predictions, as the Random Forest will help identify the appropriate donor pool for missing values.

The default value for *pmm.k* is 5 and this means that when choosing the imputed value for each missing value, the closest 5 observed values will be looked at.

3.4.4 Multiple Imputation by Chain Equations (MICE)

Another method used in the literature is to impute missing values with Multiple Imputation by Chain Equations (MICE). It creates datasets by replacing missing data in each variable in the dataset with predicted values in those variables and other variables in the data. In other words, it produces more than one value by replacing the missing data with the values obtained because of a model that captures the relationship between variables in the data.

The algorithm is as follows:

- First, missing values in the data are imputed using methods such as regression or mean imputation. This is to generate complete data for the algorithm to be used here.
- A regression model is built using all variables and missing values are calculated separately. This model is established with the observed data and thus missing data is estimated.
- Imputed values are added to the data and a missing value in data is imputed.
- For the remaining missing data, the two items mentioned above continue until there is no missing data.
- As a result of these items, multiple data sets are formed. The results from these datasets are combined.

There are several advantages of using the MICE algorithm. It can be used in various data with binary, categorical or continuous variables and it is also used for nonlinear data variables. This algorithm assumes data is missing at random, so missing data can be imputed with observed data.

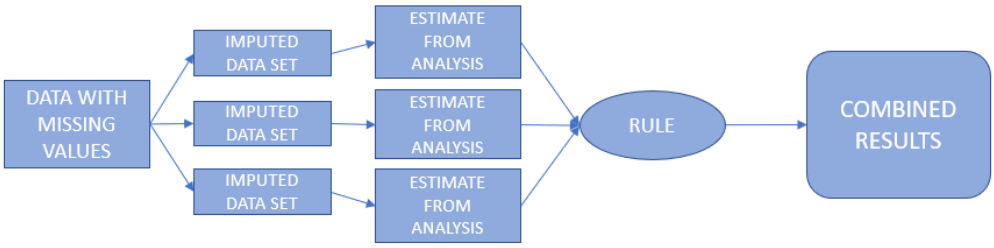

Figure 3-3. The MICE process [41]

In R, “mice” package is used to impute missing values. In this package there is a mice function. This function contains maxit, meth and number of m which is the number of multiple imputations. “maxit” is used to specify the number of iterations required to run the code. The number of m is the number of assignments to be

created. The default value is 5, but this may not always give accurate results. Thus, tunning is made to find better *m* values that gives results that are more reliable. “meth” is default “pmm” but “meth” is used in methods like “cart”, “rf”, “mean”, “norm”. In this study “cart” is selected because of giving better results. It is a decision tree application that can also be used to impute missing values. The difference between pmm method and cart method is that pmm tries to find values to missing data based on distance, while cart method creates a decision tree and uses these decision trees to predict missing data.

When the CART method is used in MICE, this algorithm is used to predict the conditional distribution of missing data in the observed data and creates a decision tree that subdivides the data according to the observed variables. The steps are as above. Data imputation was made with decision trees by using CART model instead of PMM model only.

3.4.5 General Adversarial Networks Imputation (GAIN)

Yoon et al. adopted a new method to impute missing values in 2018 by adapting Generalized Adversarial Nets and named it GAIN (Generalized Adversarial Imputation Nets) [34]. According to the results, they observed in different data sets, they reached the output that this method performs well. In this method, there are two parts as generator and discriminator. To summarize the method;

Components of real data are observed by the generator, missing data is determined, and complete data is obtained. The discriminator takes this completed data and tries to distinguish which is imputed and which is real data. The hint (clue) vector is given to the discriminator so that the generator learns the distribution of the data. The hint ensures that the generator generates data in the correct distribution. The purpose of the discriminator is to distinguish between the observed and imputed values.

The GAIN formulation of Yoon et al. is as follows [34].

For *d*-dimensional space $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_d$, a random variable $(\mathrm{X} = (X_1, \dots, X_d))$ takes values in this space with P(X), this means distribution. There is also another random variable $(M = (M_1, \dots, M_d))$ with taking values $\{0,1\}$. Here, *M* represents the mask vector and X represents the data vector.

For a new space $\tilde{\mathcal{X}}_i = \mathcal{X}_i \cup \{*\}$, $i \in \{1, \cdots d\}$, represents unobserved values and * is not in $\mathcal{X}_i$. $\tilde{\mathcal{X}} = \tilde{\mathcal{X}}_1 \times ... \times \tilde{\mathcal{X}}_d$ and $\tilde{\mathbf{X}} = \left(\tilde{X}_1 \cdots \tilde{X}_d\right) \in \tilde{\mathcal{X}}$ is a random variable with

$$\tilde{X}_i = \begin{cases} X_i & if \quad M_i = 1 \\ * & otherwise \end{cases}.$$

M indicates which components of X are observed. The *M* can be recovered by using $\tilde{\mathbf{X}}$. *n* i.i.d. copies of $\tilde{\mathbf{X}}$ are realized and denoted $\tilde{\mathrm{X}}^1, \dots, \tilde{\mathrm{X}}^{\mathrm{n}}$ in the imputation. The dataset is defined as $D = \left\{\left(\tilde{\mathrm{x}}^i, m^i\right)\right\}^n_{i=1}$, m^i is simply recovered realization of *M* corresponding to $\tilde{\mathrm{x}}^i$. The goal is to load unobserved values in each $\tilde{\mathrm{x}}^i$. To fill in the missing data values in D, samples are generated according to $P\left(X|\tilde{X} = \tilde{\mathrm{x}}^i\right)$ which is the conditional distribution of X given $\tilde{X} = \tilde{\mathrm{x}}^i$ for each *i*.

The generator G takes inputs and outputs. $\tilde{\mathbf{X}}$, *M* and *Z*, which is a noise variable are taken as inputs, and $\bar{X}$, which is a vector of imputations, is taken as outputs. Let G : $\tilde{\mathcal{X}} \times \{0,1\}^d \times [0,1] \to \mathcal{X}$ be a function and $Z = (Z_1, \dots, Z_d)$ be a d-dimensional noise which is independent of all other variables. $\bar{X}$ and $\hat{X}\ \epsilon\ X$ random variables can be defined using the following formulas.

$$\bar{X} = G\left(\tilde{X}, M, (1 - M) \circ Z\right). \tag{3.6}$$

$$\hat{X} = M \circ \tilde{X} + (1 - M) \circ \bar{X}. \tag{3.7}$$

where ∘ element-wise multiplication, $\bar{X}$ corresponds to the vector of imputed values, $\hat{X}$ corresponds to the completed vector.

The discriminator D is used to train G. The discriminator tries to distinguish which components are fake (imputed) or real rather than determining whether an entire vector is fake or real, which means estimating ***m*** mask vectors. The mask vector is predetermined by the dataset ***M***.

The discriminator $D: X \rightarrow [0,1]^d$ is a function with the i-th component of $D(\hat{x})$ corresponding to the probability that the i-th component of $\hat{x}$ was observed.

It is necessary to create hint vector that is a random variable **H** and it takes values in a space ℍ which is defined. **H** is allowed to depend on M and h is drawn according to the distribution $H|M = m$ for each imputed sample. h is passed as an additional input to the D and it becomes a function $D: X \times H \rightarrow [0,1]^d$, where the i-th component of $D(\hat{\mathrm{x}}, \mathrm{h})$ corresponds to the probability of $\hat{\mathrm{x}}$ was observed conditional on $\hat{X} = \hat{x}$ and $H = h$.

To maximize the probability of predicting M correctly, D is trained and to minimize the probability of D predicting M, G is trained. The quantity $V(D, G)$ is defined as below.

$$V(D, G) = E_{\hat{\mathrm{x}}, M, H}\left[M^T \, log\, D\left(\hat{\mathrm{X}}, \mathrm{H}\right) + (1 - M)^T \, log\left(1 - D\left(\hat{\mathrm{X}}, \mathrm{H}\right)\right)\right] \quad (3.8)$$

where log is an elementwise logarithm and dependence on G is through $\hat{\mathrm{X}}$. At the end of these formulas, the purpose of GAIN can be defined as follows.

$$\min_G \max_D \; V(D, G). \quad (3.9)$$

The binary cross-entropy loss function is defined $\mathcal{L}: \{0,1\}^d \times [0,1] \rightarrow \mathbb{R}$ by

$$\mathcal{L}(a,b) = \sum_{i=1}^{d} [a_i \, log(b_i) + (1 - a_i) \, log(1 - b_i)]. \qquad (3.10)$$

The architecture of GAIN schema is as follows by Yoon et al [42].

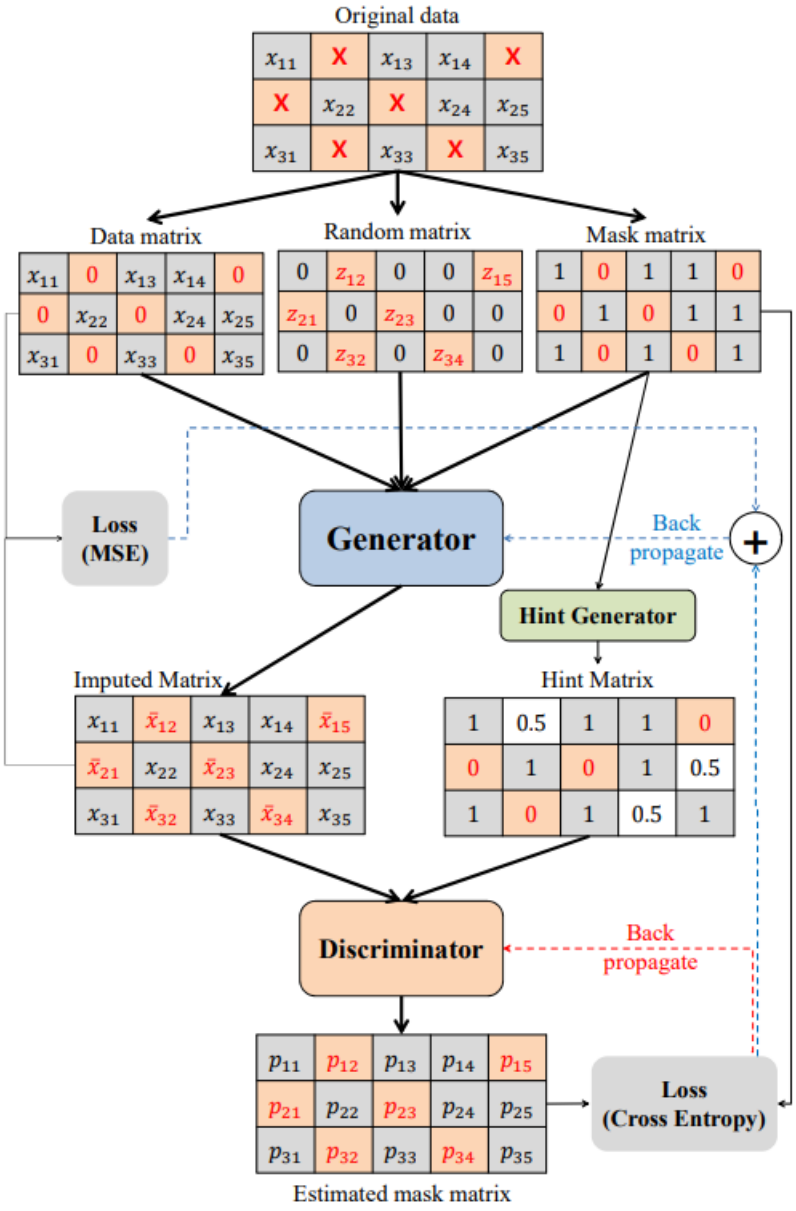

Figure 3-4. The architecture of GAIN [34]

Briefly, the algorithm can be summarized as follows.

- When there is data to be imputed, X, M is the matrix to identify the missing data and it has the same shape as X and Z is a noise vector which has the same shape as X.

- Variables are standardized so that different scales do not affect the weight of the variables.
- A multi-layer generator is created, and *X*, *M* and *Z* are given as inputs. Its output is shown by the formula (3.6). Training is shown by the formula (3.7).
- A multi-layer discriminator is created. The inputs of *D* are the data set which is imputed by the generator and the hint matrix. The hint matrix provides more information to ensure that the hint matrix distinction discriminates well. Then, *D* gives the probability that a value is true or false.
- *N* samples from the dataset are randomly taken as a mini-batch included in the training process.
- After training *D*, *D* and *G* are trained alternately to optimize the respective loss functions.

In this study, GAIN that is created by Yoon et al. in Python is used. The parameters given for tuning the input data are made with Weights and Biases [43]. This is a library and a platform that helps with visualization, tuning in machine learning experiments. This library is abbreviated as “wandb”.

The mini-batch size, hint rate, epoch and learning rate are defined to be tuned. Initialization is the process of determining initial values of weights for models. Xavier initialization is also known as Glorot is used as an initialization. This is a widely used technique for initializing weights in neural network aimed at improving network convergence and performance during training. The Xavier initialization technique helps to avoid exploding and vanishing gradient problems by keeping gradients and activations constant across the network. Xavier Normal Distribution is used in this study.

Adam Optimizer (Adaptive Moment Estimation) is a commonly used algorithm for updating the parameters of a neural network during training.

3.5 Evaluation Metrics

3.5.1 RMSE

Root Mean Square Error is the standard deviation of the estimating results. RMSE is used to calculate the distance between the predicted and the actual values of the estimator, to measure the accuracy of the predictions, and to evaluate performance of the models. RMSE provides a measure of the mean deviation of predicted values from actual values. The formula of RMSE is as follows;

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(d_i - f_i)^2} \tag{3.11}$$

where *n* is the number of observations, d is predicted value, f is actual value. It gives a value in the same units as actual values and predicted values in the RMSE result and allows the results to be interpreted easily. A low RMSE occurs when the deviations between the actual and the predicted values are small. In such a case, it can be said that the model performs well, and the estimated values are close to the real values.

3.5.2 MAE

Mean Absolute Error is used to calculate the average size of the errors in the models and also to measure the accuracy of the predicted values that result from the models. The formula of MAE is as follows;

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|d_1 - f_i| \tag{3.12}$$

where *n* is the number of observations, d is predicted value, f is actual value. When there are outliers, MAE is less sensitive than other metrics. Allow MAE occurs when the average absolute deviations between the actual and predicted values are small. In such a case, it can be said that the model performs well, and the estimated values are close to the real values.

3.5.3 CVRMSE

Coefficient of Variation of Root Mean Square is used to evaluate the relative performance of models, taking into account the variability of the RMSE relative to the mean of the observed values. The formula of CVRMSE is as follows;

$$CVRMSE = \frac{RMSE}{\bar{f}_{(n)}} = \frac{\sqrt{\sum_{i=1}^{n}(d_i - f_i)^2 \,/\, n}}{\bar{f}_{(n)}} \tag{3.13}$$

where *n* is the number of observations, d is predicted value, f is actual value and $\bar{f}_{(n)}$ shows the average of the true value for the generated in missing period. By dividing RMSE by the mean of true values, CVRMSE is obtained.

3.5.4 NSE

Nash-Sutcliffe Efficiency was found by Nash and Sutcliffe (1970). NSE is used to measure the performance of environmental models, including imputation models. NSE serves to measure the accuracy of models by comparing the actual values and predicted values. The formula of NSE is as follows [44];

$$NSE = 1 - \frac{\sum_{i=1}^{n}(d_i - f_i)^2}{\sum_{i=1}^{n}\left(d_i - \bar{f}\right)^2} \tag{3.14}$$

where *n* is the number of observations, d is predicted value, f is actual value and $\bar{f}$ is the mean of the actual values. The range of this metric is from -Inf to 1. If the NSE result is equal to 1, it means that there is a very good match between true values and predicted values. An NSE of 0 indicates that the predicted values are as accurate as the average of the true values. An NSE value less than 0 indicates that the model is worse than using the average of the observed values for prediction.

CHAPTER 4

INTRODUCING THE DATA AND PREPROCESSING

In this section, the data used in the study is introduced and data preprocessing stages that are applied for each method are given.

4.1 Selecting Stations

As mentioned before, in the daily average temperature and total precipitation meteorology data, there are missing values due to the equipment or weather conditions, as there may be problems in the devices that measure the air temperature and precipitation. The aim of this study is to impute those values with five different methods, to find out which model's outputs give close values to the real values by looking at the imputed data, and to impute all the missing data with this method that gives the best result. The missing values of any station in the Turkish Meteorological Database can be estimated from the associated station database by simultaneous observations (Yozgatlıgil et al., 2013). The precipitation and temperature data in this study are the data with missing values recorded from January 1, 2005, to September 7, 2022.

In the data obtained, there are only days with data. The creation and correction of the data was done using SAS software [45]. For each station, calendar data were created from January 1, 2005, to September 7, 2022, and the values of the days with average temperature and precipitation values were obtained. According to the station information, the cities in which the stations are located were discussed. Latitude, longitude, and altitude information has been added.

Based on the 7 geographical regions of Turkey, these stations are distributed according to their regions, and they represent different climatological characteristics of Turkey. These regions are the Marmara Region (MAR), Black Sea Region (BSR), Aegean Region (AR), Mediterranean Region (MR), Central Anatolia Region (CAR), Eastern Anatolia Region (EAR), Southeastern Anatolia Region (SAR). These physical classifications are important because each region have its own characteristics. Therefore, it is necessary to consider these regions separately and to estimate the missing values. The target stations were selected for each region. While selecting these target stations, it was checked that they do not have missing values and that they can be in the center of the other reference stations to be selected. After the target stations were selected, the other 5 stations with high correlation with the target stations and no or little missing data were selected as reference stations. Pearson correlation was used when looking at the correlations between target stations and other stations. After selecting the target and reference stations, separate data sets were created for each of the 7 regions with both temperature and precipitation data. To examine the performance of the models on data with different missing percentages, 5%, 10%, 20% and 30% missing data periods were created for each data set for both precipitation and temperature. The selected target and reference stations and the longitude, latitude, and altitude values of these stations are listed below.

Table 4-1. List of meteorological stations

Region	Station Number	Station Name	Latitude-Longitude- Altitude
	17220 (Target)	İzmir Bölge	38°39' - 27°08' - 29
	17180	Dikili	39°07' - 26°88' - 3
	17186	Manisa	38°61' - 27°40' - 71
AR	17232	Kuşadası	37°85' - 27°26' - 25
	17221	Çeşme	38°30' - 26°37' - 5
	17792	Salihli	38°48' - 28°12' - 111

Table 4-1 (continued).

	17310 (Target)	Alanya	36°55' - 31°98' - 6
	17954	Manavgat	36°78' - 31°44' - 38
	17330	Silifke	36°38' - 33°93' - 10
MR	17974	Gazipaşa	36°27' - 32°30' - 21
	17320	Anamur	36°06' - 32°86' - 2
	17956	Mut	36°65' - 33°43' - 340
	17699 (Target)	Manyas	40°04' - 27°97' - 50
	17674	Balıkesir/Gönen	40°11' - 27°64' - 37
	17673	Karacabey	40°13' - 28°33' - 15
MAR	17114	Bandırma	40°33' - 27°99' - 63
	17705	Susurluk	39°91' - 28°16' - 47
	17158	Balıkesir Akçaldede Radar Sahası	39°74' - 27°61' - 631
	17020 (Target)	Bartın	41°62' - 32°35' - 33
	17022	Zonguldak	41°44' - 31°77' - 135
	17613	Devrek	41°23' - 31°96' - 100
BSR	17602	Amasra	41°75' - 32°38' - 73
	17604	Kastamonu/Cide	41°88' - 32°94' - 36
	17615	Ulus	41°58' - 32°63' - 162
	17099 (Target)	Ağrı	39°72' - 43°05' - 1646
	17720	Doğubeyazıt	39°53' - 44°01' - 1640
	17740	Hinis	39°36' - 41°69' - 1715
EAR	17780	Malazgirt	39°14' - 42°53' - 1540
	17784	Erciş	39°01' - 43°33' - 1678
	17100	Iğdır	39°92' - 44°05' - 856

Table 4-1 (continued).

	17966 (Target)	Birecik	37°01' - 37°97' - 347
	17262	Kilis	36°70' - 37°11' - 640
	17261	Gaziantep	37°05' - 37°35' - 854
SAR	17871	Gölbaşı	37°78' - 37°65' - 900
	17270	Şanlıurfa	37°16' - 38°78' - 550
	17944	Bozova	37°36' - 38°51' - 622
	17245 (Target)	Konya Bölge	37°86' - 32°47' - 1029
	17191	Cihanbeyli	38°65' - 32°92' - 973
	17242	Beyşehir	37°67' - 31°74' - 1141
CAR	17902	Karapınar	37°71' - 33°52' - 996
	17832	Ilgın	38°27' - 31°89' - 1036
	17900	Cumra	37°56' - 32°79' - 1014

Five stations were chosen as reference stations. These stations have higher correlations with target stations than other unselected stations. One can be said that these highly correlated reference stations were chosen to be surround the target stations. However, this does not apply to stations located in coastal areas such as the Black Sea and Mediterranean. It is aimed at selecting the reference stations in these regions as close to the target as possible. Stations in the same climatic conditions were chosen as reference stations. Selected target and reference stations were shown both for Turkey in general and for each region separately. The summary statistics of the target and reference stations in given Table 4-2.

Table 4-2. Summary statistics of the target and reference stations for daily temperature

Region	Station Number	Mean	SD	CV
AR	*17220 (Target)*	18.38	7.65	41.59
	17180	17.20	7.17	41.66
	17186	17.10	8.52	49.84
	17232	18.21	6.79	37.28
	17221	17.88	6.57	36.73
	17792	17.18	8.42	49.02
MR	*17310 (Target)*	20.50	6.43	31.37
	17954	19.19	6.88	35.84
	17330	19.64	7.20	36.67
	17974	18.87	6.79	35.97
	17320	20.19	6.64	32.90
	17956	18.62	9.33	50.10
MAR	*17699 (Target)*	15.26	8.27	54.21
	17674	15.83	8.27	52.26
	17673	16.04	8.37	52.20
	17114	15.84	8.13	51.30
	17705	15.92	8.26	51.88
	17158	13.24	8.16	61.67
BSR	*17020 (Target)*	13.34	7.74	58.00
	17022	14.32	7.30	50.96
	17613	14.31	8.02	56.02
	17602	14.32	7.28	50.80
	17604	14.57	7.20	49.39
	17615	13.10	8.03	61.32

Table 4-3 (continued).

	17099 (Target)	6.38	12.21	191.50
	17720	8.98	10.12	112.79
	17740	6.81	10.94	160.45
EAR	17780	7.48	12.04	161.00
	17784	7.39	9.84	133.07
	17100	12.36	11.06	89.48
	17966 (Target)	18.05	9.53	52.79
	17262	17.54	8.69	49.52
	17261	15.84	9.32	58.84
SAR	17871	14.55	9.75	67.04
	17270	18.86	10.01	53.10
	17944	16.71	10.16	60.79
	17245 (Target)	14.80	9.24	62.43
	17191	13.10	9.39	71.67
	17242	12.09	9.00	74.41
CAR	17902	12.69	9.09	71.64
	17832	12.69	8.84	69.65
	17900	13.36	9.02	67.54

Based on the data presented in Table 4-2, it appears that the average temperature values and standard deviations for the selected target and reference stations are similar. The region with the highest coefficient of variation is the EAR region, which suggests that the temperature data has a wide distribution. On the other hand, the MR region has the lowest coefficient of variation.

Table 4-4. Summary statistics of the target and reference stations for the precipitation

Region	Station Number	Mean	SD	CV
AR	*17220 (Target)*	2.17	7.76	357.98
	17180	1.62	5.49	338.83
	17186	1.79	6.68	373.66
	17232	1.95	7.49	384.06
	17221	1.83	6.32	344.46
	17792	1.34	4.23	314.97
MR	*17310 (Target)*	1.79	6.55	365.59
	17954	1.45	10.10	696.87
	17330	0.75	3.45	459.82
	17974	1.62	6.53	404.24
	17320	1.78	6.70	376.81
	17956	0.51	2.43	476.37
MAR	*17699 (Target)*	1.85	7.49	405.26
	17674	1.69	6.75	400.54
	17673	1.54	5.34	346.81
	17114	1.63	5.67	346.87
	17705	1.98	6.95	351.42
	17158	1.79	6.49	362.14
BSR	*17020 (Target)*	3.00	8.63	287.29
	17022	2.89	7.62	263.66
	17613	1.98	5.59	282.36
	17602	2.71	7.25	267.35
	17604	3.89	13.53	347.91
	17615	2.63	6.70	255.04

Table 4-5 (continued).

	17099 (Target)	1.44	3.50	242.25
	17720	1.07	2.98	271.50
	17740	1.19	2.99	250.42
EAR	17780	1.70	3.77	321.97
	17784	0.98	3.09	315.76
	17100	0.60	2.31	386.48
	17966 (Target)	0.63	3.38	534.77
	17262	0.90	3.27	364.28
	17261	1.02	3.15	308.02
SAR	17871	1.47	5.08	346.04
	17270	0.71	3.13	444.36
	17944	0.64	2.75	426.32
	17245 (Target)	0.87	2.91	333.91
	17191	0.78	2.65	338.30
	17242	1.24	4.61	372.61
CAR	17902	0.78	3.20	409.77
	17832	1.05	3.76	356.67
	17900	0.68	2.53	373.64

According to Table 4-3, the region with the highest coefficient of variation is the MR region, while the lowest region is the BSR region.

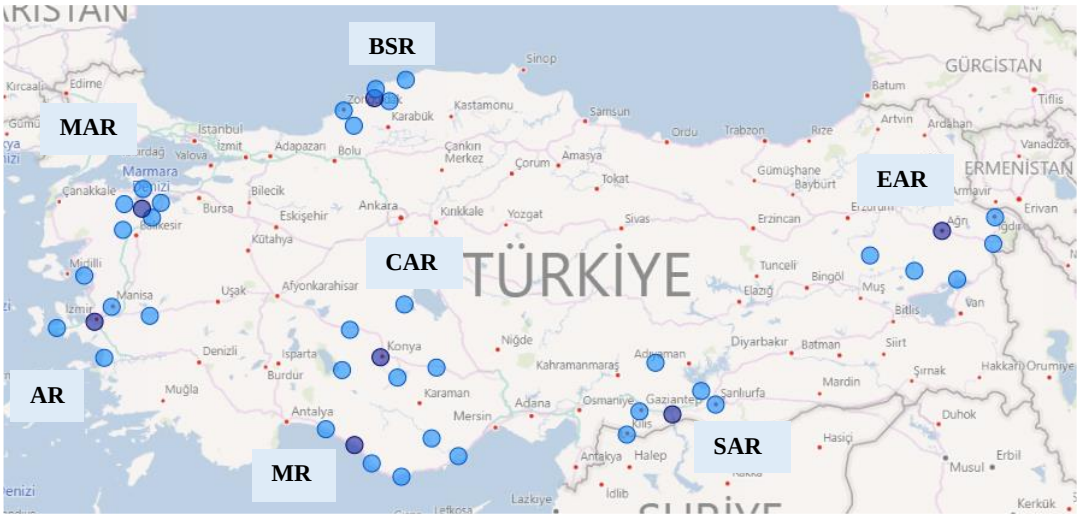

Figure 4-1. Locations of stations on Turkey

The dark blue ones represent the positions of the target stations, while the light blues represent the positions of the reference stations. Below, target stations and reference stations were shown in more detail on the basis of regions. In this study, the target and reference stations were chosen from the western region of the Black Sea instead of the eastern region. This decision was made because the western region has fewer missing values and is better suited for the purposes of this study.

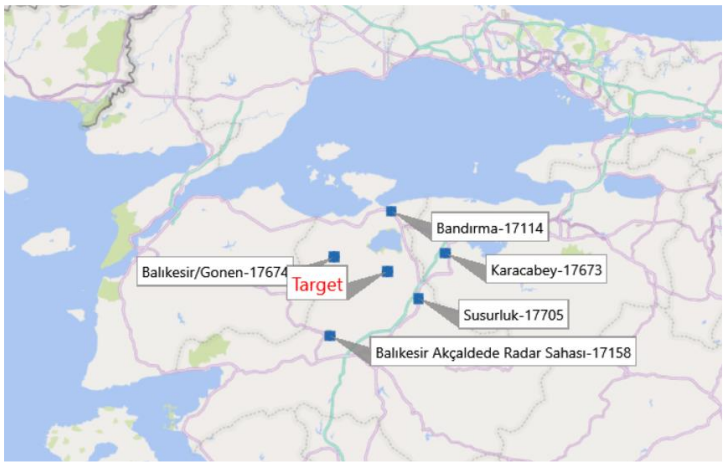

Figure 4-2. MAR region target and reference stations

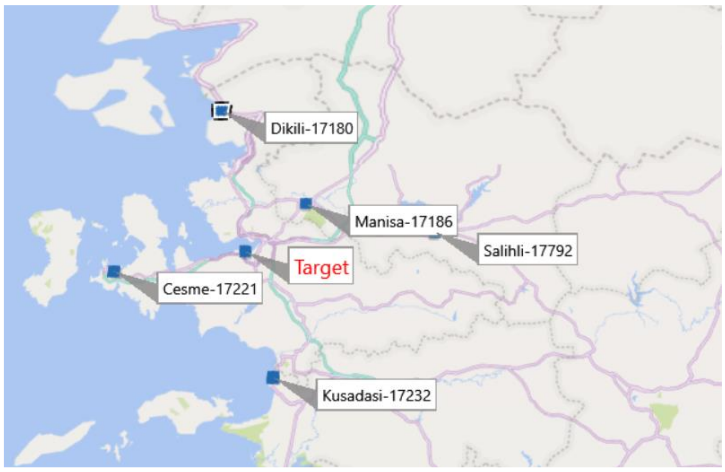

Figure 4-3. AR region target and reference stations

Figure 4-4. BSR region target and reference stations

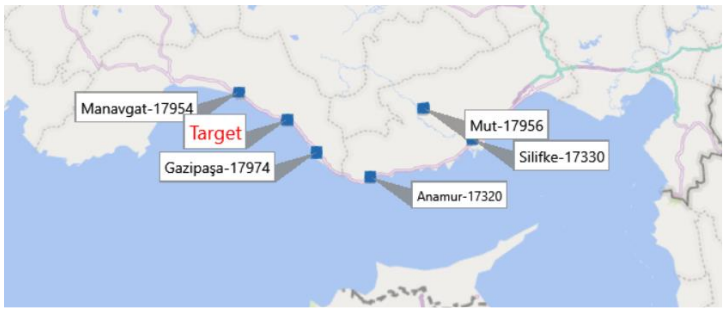

Figure 4-5. MR region target and reference stations

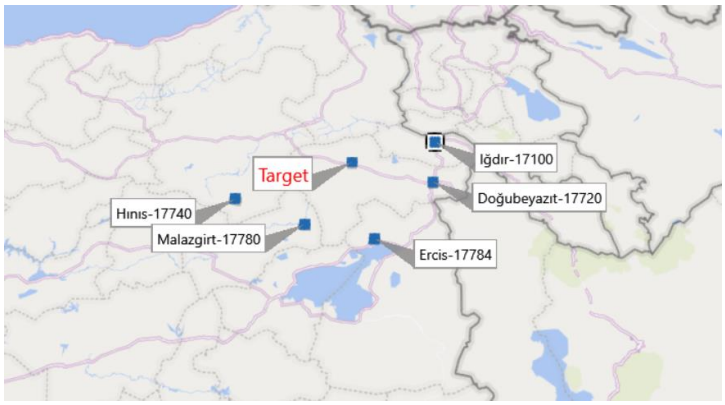

Figure 4-6. EAR region target and reference stations

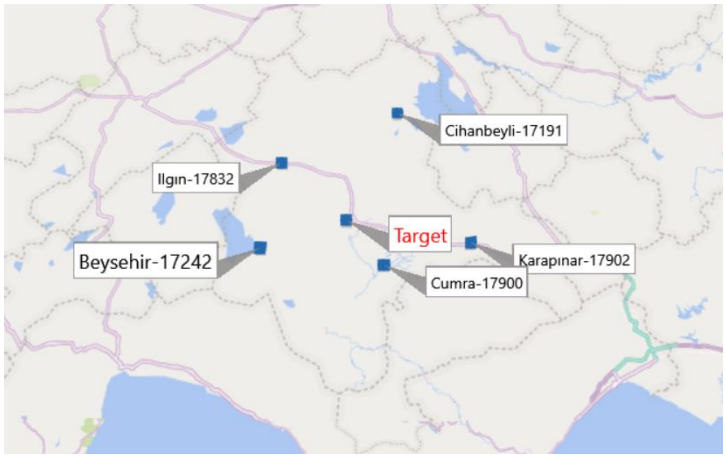

Figure 4-7. CAR region target and reference stations

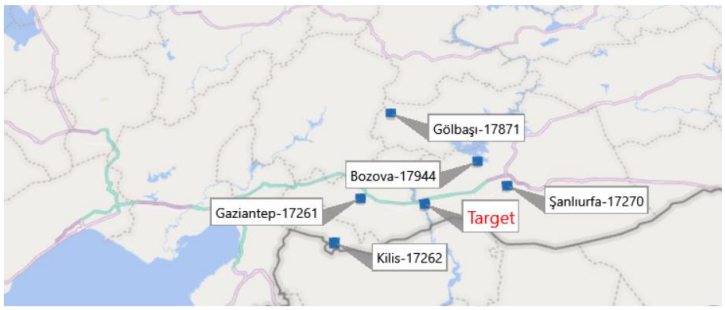

Figure 4-8. SAR region target and reference stations

Reference and target stations were shown by region. As mentioned before, stations with high correlation were selected. The values found using Pearson correlation were shown in two tables for both precipitation and temperature as follows.

Table 4-6. Correlations between target and reference stations for temperature

Region	Reference Stations	Correlation	Region	Reference Stations	Correlation
		İzmir Bölge			***Alanya***
	Dikili	0.991215		Manavgat	0.989538
	Manisa	0.987757		Silifke	0.983054
AR	Kuşadası	0.990606	**MR**	Gazipaşa	0.990399
	Çeşme	0.989214		Anamur	0.990222
	Salihli	0.987968		Mut	0.969166
		Manyas			***Bartın***
	Balıkesir/Gönen	0.991653		Zonguldak	0.956047
	Karacabey	0.994048		Devrek	0.984923
MAR	Bandırma	0.986736	**BSR**	Amasra	0.956508
	Susurluk	0.991739		Kastamonu/Cide	0.9555
	Balıkesir	0.965199		Ulus	0.986569
	Akçaldede				
	Radar Sahası				

Table 4-7 (continued).

		Ağrı			***Birecik***
	Doğubeyazıt	0.972212		Kilis	0.97467
	Hinis	0.986325		Gaziantep	0.986053
EAR	Malazgirt	0.992696	**SAR**	Gölbaşı	0.985835
	Erciş	0.976126		Şanlıurfa	0.986122
	Iğdır	0.961572		Bozova	0.9883
		Konya Bölge			
	Cihanbeyli	0.99262			
	Beyşehir	0.989347			
CAR	Karapınar	0.985684			
	Ilgın	0.987325			
	Cumra	0.994288			

As can be seen in Table 4-2 above, the correlations between the daily temperature values of the stations are high. The lowest correlation with 0.9555 is between Bartın and Kastamonu/Cide stations located in BSR region. The correlations between the target and reference stations in the AR region and the CAR region are high. Therefore, it can be said that especially in these regions, these reference stations can be helpful in imputing missing temperature data at the target stations.

Table 4-8. Correlations between target and reference stations for precipitation

Region	**Reference Stations**	**Correlation**	**Region**	**Reference Stations**	**Correlation**
		İzmir Bölge			***Alanya***
	Dikili	0.737112		Manavgat	0.570066
	Manisa	0.799482		Silifke	0.552973
AR	Kuşadası	0.635431	**MR**	Gazipaşa	0.605719
	Çeşme	0.61649		Anamur	0.65533
	Salihli	0.50756		Mut	0.571924

Table 4-9 (continued).

		Manyas			***Bartın***
	Balıkesir/Gönen	0.779867		Zonguldak	0.715711
	Karacabey	0.766686		Devrek	0.63945
MAR	Bandırma	0.664616	**BSR**	Amasra	0.767806
	Susurluk	0.775929		Kastamonu/Cide	0.691046
	Balıkesir Akçaldede Radar Sahası	0.678156		Ulus	0.702063
		Ağrı			***Birecik***
	Doğubeyazıt	0.458572		Kilis	0.624553
	Hinis	0.518358		Gaziantep	0.624899
EAR	Malazgirt	0.634241	**SAR**	Gölbaşı	0.58697
	Erciş	0.466684		Şanlıurfa	0.641802
	Iğdır	0.493703		Bozova	0.706307
		Konya Bölge			
	Cihanbeyli	0.523926			
	Beyşehir	0.621896			
CAR	Karapınar	0.553926			
	Ilgın	0.533594			
	Cumra	0.675831			

The stations with the highest correlation with the target stations were tried to be taken as reference for the daily precipitation series, and the results given in Table 4.3 were obtained. It can be said that the correlation between the reference stations and target stations in some regions is low. When the correlation values were examined, one can be said that the regions with high correlation are MAR and BSR, respectively. The lowest correlation is 0.45 with Doğubeyazıt and this station is in the EAR region. In addition, the correlation values of other stations in this region are low. It can be seen from the tables above that both temperature correlations and precipitation correlations are high for the AR region.

4.2 Data Preprocessing

To carry out this study, there should be no missing data in both target stations and reference stations. Therefore, the longest non-lost date ranges of temperature and precipitation data at each station were determined and these data were used in the study. In other words, the dates and the number of observations taken from each region are different from each other. In the table below, it can be seen in which date ranges the data for the regions were taken to be used in the study.

Table 4-10. Selected date intervals of regions for temperature

Region	Date Intervals	Number of Observations
AR	May 15, 2011 – May 17, 2015	1444
MR	January 1, 2005 – January 6, 2009	1467
MAR	March 3, 2011 – September 7, 2012	535
BSR	May 17, 2011 – April 1, 2013	686
EAR	January 1, 2005 – May 15, 2010	1961
SAR	January 1, 2005 – April 9, 2011	2290
CAR	January 28, 2020, September 7, 2022	954

According to this table, the region with the longest date range for temperature without missing data is the SAR region. The region with the shortest date range is the MAR region. For temperature data, 5%, 10%, 20% and 30% missing data were created for each target station in each region. In this way, while there is lost data created at the target stations, there is no lost data at the reference stations.

Table 4-11. Selected date intervals of regions for precipitation

Region	Date Intervals	Number of Observations
AR	September 9, 2018 – August 13, 2021	1057
MR	July 21, 2015 – December 28, 2016	527
MAR	July 20, 2015 – November 24, 2017	859
BSR	June 6, 2015 – November 19, 2016	533
EAR	October 28, 2018 – November 15, 2018	353
SAR	March 4, 2015 – July 7, 2016	492
CAR	January 28, 2020 – September 7, 2022	954

The date ranges in the data created for the precipitation above have been selected to not contain any lost data. It was a bit difficult to select the date ranges without missing data, as the precipitation dataset consists of a lot of missing data. Therefore, the number of observations is less than the number of observations for temperature. According to this table, the region with the longest date range for temperature without missing data is the AR region. The region with the shortest date range is the EAR region. For precipitation data, 5%, 10%, 20% and 30% missing data were created for each target station in each region. In this way, while there is lost data created at the target stations, there is no lost data at the reference stations. These created missing data are randomly created within the data set.

Data with different missing percentages for temperature and precipitation are kept in different data frames. This procedure was done for each region. The data sets to be used in the GAIN method are normalized in R, saved in Excel files, and prepared for use in Python.

In addition to creating this random missing data, a certain date range for both precipitation and temperature in a selected region (AR) was selected and deleted from the complete data. In this way, it is desired to examine the performances of the

selected models when imputing the missing data in the daily meteorological data, when the data in block form is lost. These data sets are also saved in Excel files.

CHAPTER 5

APPLICATION AND RESULTS

This chapter discusses the application of the methods to be used to assign data to the generated data with missing values. As mentioned before, the map of Turkey was divided into 7, and target and reference stations the same characteristics and having high correlation with each other were selected. These selected stations did not contain missing data. Random missing values at target stations were generated according to different percentages of missing (5%, 10%, 20%, 30%). These missing imputation methods were used to estimate the missing data at the target station using reference stations. Tuning was done when using these methods. How tuning is done in applications and which parameters give good results according to the tuning result are examined. In all regions, these processes were performed separately for the data with all missing percentages and the results were compared.

5.1 Application of Missing Data Imputation Methods

5.1.1 SAA Application

Reference stations were given while applying this method. On the days when the missing values at the target station were found, the averages of the values of the other stations on that day were taken and the missing data found in the target station were printed instead. This process was performed on temperature and precipitation data for all regions and with all missing values.

5.1.2 KNN Application

While applying this method, it is aimed to find the *k* value that gives the best result. For this, a loop was created, and the value of k was adjusted to increase by 10 from 1 to 500. Data imputation was performed with each *k* value and a MAE value obtained by comparing the outputs of the imputed with the raw data. As a result of the *k* value adjusted from 1 to 500, in which *k* range this model gives the best MAE value, the loop is run again for that range. For example, for temperature data, looking at the BSR region, the MAE value for the 20th *k* value is low. Therefore, the loop was run again to include the 20th *k* value and the k value that gave the lowest MAE value was selected by looking at the MAE values as a result of which *k* value within this range. As a result of the selected *k* values, the missing data were imputed. This process was performed on temperature and precipitation data for all regions and with all missing values. The *k* values used for both temperature and precipitation data in each region are given in the table below.

Table 5-1. Best *k* values for temperature

Region	**5%**	**10%**	**20%**	**30%**
AR	15	8	6	8
MR	15	14	11	17
MAR	12	3	5	4
BSR	4	2	6	7
EAR	20	28	16	24
SAR	18	22	28	16
CAR	9	6	7	5

Table 5-2. Best *k* values for precipitation

Region	5%	10%	20%	30%
AR	14	3	4	17
MR	3	31	8	7
MAR	1	5	6	4
BSR	20	2	2	12
EAR	4	5	9	13
SAR	7	4	13	11
CAR	36	18	4	2

5.1.3 RF Application

It is aimed to find the parameters that give the best MAE value in the RF model, as was done in the KNN model. It was aimed to find the number of trees that would give the best result, and for this, a cycle was established in which these tree numbers were increased by 100 from 1 to 1000. The default value of pmm.k was taken. It has been observed that the output values give better results when maxit is equal to 5. For this reason, a study was carried out to determine the number of trees by keeping these values constant. As mentioned above in the KNN method, the range of the number of trees that give the smallest MAE value in this method was found according to the number of trees that were run from 1 to 1000 and increased by 100. Then the loop was run again for this interval. The number of trees with the lowest MA result was taken. As a result of the selected number of trees, the missing data were imputed. This process was performed on temperature and precipitation data for all regions and with all missing values. The number of trees used for both temperature and precipitation data in each region are given in the table below.

Table 5-3. Number of trees for temperature

Region	5%	10%	20%	30%
AR	210	159	373	188
MR	779	272	180	171
MAR	446	118	25	289
BSR	214	131	475	121
EAR	384	435	139	65
SAR	48	477	347	835
CAR	332	350	486	575

Table 5-4. Number of trees for precipitation

Region	5%	10%	20%	30%
AR	633	104	294	635
MR	65	817	165	913
MAR	389	202	11	67
BSR	357	150	101	53
EAR	453	285	138	256
SAR	209	74	352	148
CAR	130	265	246	27

5.1.4 MICE Application

In the MICE method, it is aimed to find the *m* value that will give the lowest MAE value by looping like other models. In this model, the CART model was used as the method, and it was observed that it gave better results when the maxit value was equal to 10. Therefore, this value has been kept constant. As in the above methods, a loop is established in this method. In this loop, the number of *m* is set to increase by 10 from 1 to 100. The range of the *m* value that gives the lowest MAE value is found and another cycle is established for this m value range. At the end of this, the

number of *m*, which gives the smallest MAE value, was found. As a result of the selected number of *m*, the missing data were imputed. This process was performed on temperature and precipitation data for all regions and with all missing values. The number of *m* used for both temperature and precipitation data in each region are given in the table below.

Table 5-5. Number of *m* for temperature

Region	**5%**	**10%**	**20%**	**30%**
AR	46	33	10	45
MR	10	51	68	14
MAR	41	15	15	16
BSR	39	35	41	5
EAR	44	21	61	45
SAR	41	64	69	5
CAR	19	72	17	21

Table 5-6. Number of *m* for precipitation

Region	**5%**	**10%**	**20%**	**30%**
AR	44	44	11	44
MR	41	5	71	31
MAR	11	36	10	10
BSR	40	5	43	5
EAR	37	30	5	39
SAR	17	81	31	23
CAR	28	20	21	14

5.1.5 GAIN Application

Initially, the GAN is trained on a dataset with complete observations. The generator learns to generate synthetic samples that resemble the original data distribution, while the discriminator learns to distinguish between real and generated samples. To be able to reach the best performance, architecture of the network is crucial. First of all, 2 hidden layer architectures were used in the model. To explain the reason, it was desired to make an architectural tuning in this model first. Other parameters were kept the same by taking the number of hidden layers as 2, 3, 4 and 5. As a result, it has been observed that an architectural structure with two hidden layers gives better results than the others. The results were worse when 3 hidden layers were used, however, 4 and 5 hidden layers gave better results than three. As a result of this study, the number of hidden layers was taken as 2, and generator and discriminator architectures were created in this way. The next step is to tune other parameters and find the parameter values that give the best results. These parameters are mini-batch size, probability of hint, the number of epoch and the learning rate required for optimization. Tuning of these parameters was done using weights and biases tool. Its graphics are given in the appendix. Weights and biases help to try each combination of these parameters to find the combination of parameters that gives good results, i.e., low MSE. To give brief information about these parameters, the proper selection of the mini-batch size increases the performance and the effectiveness of the training process. The probability of hint helps to determine the extent to which missing data is hidden. In the prediction process, this parameter controls how many hints the algorithm will be used. Using hints makes it easier to predict these missing data when imputing data. The epoch number is the number of iterations that the data training algorithms perform using data. This parameter is important because increasing the number of epochs can improve the performance of the model. For this, the most optimal epoch number should be found. Another parameter is the learning rate. This parameter serves to show values that can affect the training speed, performance and process of the model. For this, it is important to find the optimal learning rate value.

In this way, training loss can be reduced, and better performance can be achieved. By using weights and biases, ranges were defined for these parameters, and it was found out at which values the parameters gave good results. In other words, it was aimed to find the parameter values that will minimize the losses. As a result of the selected parameter values, the missing data were imputed. This process was performed on temperature and precipitation data for all regions and with all missing values. The selected parameter values used for both temperature and precipitation data in each region are given in the table below.

Table 5-7. Tuned Hyperparameter values for temperature

Region	Parameter	5%	10%	20%	30%
AR	***Minibatch size***	59	316	346	72
	epoch	1406	4129	6976	4734
	learning rate	0.08216	0.08164	0.01536	0.01566
	probability of hint	0.653	0.7454	0.773	0.65112
MR	***Minibatch size***	215	33	75	464
	epoch	3519	5066	3923	1575
	learning rate	0.008758	0.04178	0.009437	0.04017
	probability of hint	0.594	0.8973	0.5071	0.5978
MAR	***Minibatch size***	90	416	418	57
	epoch	6569	6235	9789	9073
	learning rate	0.0323	0.05967	0.0183	0.00120
	probability of hint	0.8416	0.8067	0.6996	0.5921
BSR	***Minibatch size***	313	299	45	241
	epoch	9083	9850	6569	2988
	learning rate	0.01053	0.002713	0.01832	0.03523
	probability of hint	0.5704	0.6764	0.698	0.6322

Table 5-8 (continued).

EAR	***Minibatch size***	138	400	483	109
	epoch	5405	8000	8762	4215
	learning rate	0.007941	0.01	0.01233	0.00715
	probability of hint	0.8658	0.8	0.5607	0.7241
SAR	***Minibatch size***	30	32	167	456
	epoch	12000	9000	8500	5000
	learning rate	0.01	0.0013	0.007919	0.0276
	probability of hint	0.51	0.71	0.7213	0.7239
CAR	***Minibatch size***	32	30	91	80
	epoch	9765	9500	9261	6678
	learning rate	0.00535	0.043	0.004445	0.00116
	probability of hint	0.8867	0.99	0.5401	0.5355

Looking at the above table, which is about temperature, it is concluded that a low probability of hint, mini-batch size and learning rate will give good results for all regions in general. Nothing definite can be said about epoch numbers. In some regions, the high epoch number gives better performance, in some regions lower epoch number gives better results.

Table 5-9. Tuned Hyperparameter values for precipitation

Region	**Parameter**	**5%**	**10%**	**20%**	**30%**
AR	***Minibatch size***	36	32	129	256
	epoch	2102	3500	5150	5607
	learning rate	0.0342	0.01	0.003406	0.04656
	probability of hint	0.8477	0.98	0.6503	0.728

Table 5-10 (continued).

MR	***Minibatch size***	120	32	43	56
	epoch	1500	1719	2500	8456
	learning rate	0.05	0.001199	0.001	0.0784
	probability of hint	0.97	0.98	0.999	0.7834
MAR	***Minibatch size***	276	200	43	53
	epoch	2466	8987	8017	2329
	learning rate	0.02486	0.003669	0.03943	0.03325
	probability of hint	0.9538	0.9534	0.923	0.589
BSR	***Minibatch size***	23	72	58	219
	epoch	9332	6592	4824	9984
	learning rate	0.003303	0.002807	0.03265	0.01544
	probability of hint	0.9254	0.6249	0.8727	0.5836
EAR	***Minibatch size***	120	250	176	23
	epoch	1500	2500	2005	1724
	learning rate	0.01	0.01	0.03115	0.00103
	probability of hint	0.99	0.51	0.5936	0.778
SAR	***Minibatch size***	67	41	45	31
	epoch	7348	7570	4618	1825
	learning rate	0.00764	0.001	0.07554	0.01754
	probability of hint	0.7639	0.99	0.7315	0.6744
CAR	***Minibatch size***	173	38	145	25
	epoch	6149	7992	2403	9980
	learning rate	0.02911	0.005065	0.002521	0.001
	probability of hint	0.6626	0.5292	0.6678	0.51

Looking at the parameter values to be used for the precipitation table, the probability of hints to be given is higher than the probability of hints to be used for the temperature. Learning rates are low, as are the values to be used in temperature. The number of epochs varies. Low mini-batch sizes generally give better results.

5.1.6 Block Missing Application

In this part of study, target and reference stations in the AR region were used. For the temperature and precipitation data sets, a random 8-month time interval as a block was deleted from the data and missing data was created. It was aimed to use 5 methods and compare these methods in the imputation of this 8-month period. It was also aimed to create missing values in the data to cover summer, autumn and winter seasons. The loss data range created for temperature is between 1 July 2011 and 1 March 2012. The parameter values of the five models to be used for temperature are as follows.

Table 5-11. Parameter values for temperature

Method	Parameters
KNN	*k* : 8
RF	*Number of tree* : 491
MICE	*m :* 61
GAIN	*Minibatch size :* 74 *Epoch :* 9813 *learning rate :* 0.01349 *probability of hint :* 0.5255

It is aimed to create missing values in precipitation to cover the winter, spring and summer seasons. The loss data range created for temperature is between 1 December 2018 and 1 August 2019. The parameter values of the 5 models to be used for precipitation are as follows.

Table 5-12. Parameter values for precipitation

Method	Parameters
KNN	*k* : 5
RF	*Number of tree* : 814
MICE	*m :* 21
GAIN	*Minibatch size :* 18 *Epoch :* 2188 *learning rate :* 0.006438 *probability of hint :* 0.5674

5.2 Results

In Section 5.1 the parameters showing the best performances for the models are given. Using these parameters, the models were run for each region. Model comparisons for each region are as follows.

5.2.1 Results for Izmir (AR region)

Izmir is located in the Aegean region. For its climate, it can be said that it is hot, dry in the summer, warm, and rainy in the winter. As can be seen in Figure 4-3, the reference stations for this region are to cover this target station. Correlation values between stations are shown in Table 4-2 and Table 4-3.

Table 5-13. NSE values between the imputed and original values for AR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17220 İzmir	*5*	0.9988865	0.9997143	**0.9997727**	0.9989932	0.999618
	10	0.9980612	0.9993841	**0.9993969**	0.9974217	0.999255
	20	0.9958025	0.998692	**0.9987269**	0.9941363	0.998558
	30	0.9938329	0.9977303	**0.9978983**	0.9911904	0.997569

According to the above table, the RF model gives the best results for all missing percentages. RF is followed by KNN and GAIN methods. NSE values in these three models are very close to each other.

Table 5-14. RMSE values between the imputed and original values for AR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17220 İzmir	*5*	0.2550466	0.1291888	**0.1152199**	0.2425197	0.149366
	10	0.3365352	0.1896837	**0.1876927**	0.3893136	0.208577
	20	0.4951795	0.2764192	**0.2727083**	0.5852646	0.290264
	30	0.6002128	0.361268	**0.3503856**	0.7173696	0.376822

Table 5-15. CVRMSE values between the imputed and original values for AR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17220 İzmir	*5*	0.01391047	0.007029518	**0.006268887**	0.0131988	0.008125474
	10	0.01838978	0.01031433	**0.01020653**	0.02118059	0.01137719
	20	0.02719726	0.0150407	**0.01493193**	0.03185305	0.01575629
	30	0.03311797	0.01980138	**0.01905898**	0.03908368	0.02047639

By looking Tables 5-9 to 5-11, it can be said that the missing values in the temperature data in the AR region give better results when imputed by the RF method. It is seen that when the missing data percentage is increasing, RMSE values are also increasing. On the other hand, it is seen that KNN and GAIN methods give results close to RF method. Figure 5-1 shows the graph in which the real data and the imputed data were created only for the days with missing values (30% missingness). In this plot randomly created missing values are combined so that one can see the overall performance of the models.

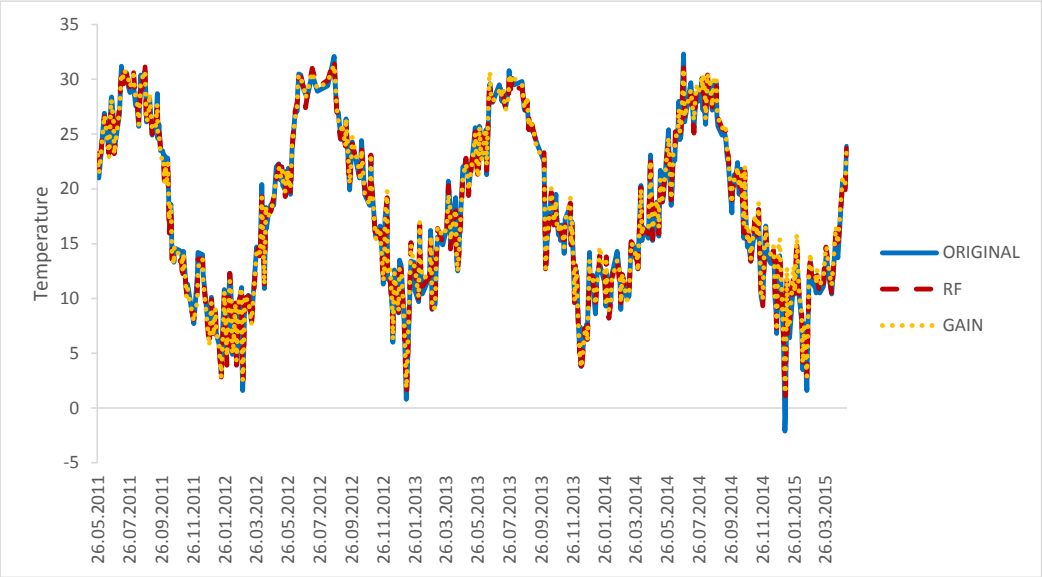

Figure 5-1. Temperature imputations for AR region for 30% missing percentages

Table 5-16. NSE values between the imputed and original values for AR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17220 İzmir	*5*	0.9929391	0.9901587	**0.9983468**	0.9892946	0.99578
	10	0.9634383	0.9448999	**0.9805925**	0.9105104	0.979972
	20	0.9688185	0.9585148	0.9657784	0.9489545	**0.978049**
	30	0.9217509	0.9077059	**0.9333552**	0.8333712	0.930042

According to Table 5-12, RF model gives the best results for all missing percentages like temperature data. RF is followed by GAIN method. While RF is better in precipitation data with 5%, 10% and 30% missing values, GAIN method gives better results in precipitation data with 20% missing values.

Table 5-17. RMSE values between the imputed and original values for AR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17220 İzmir	*5*	0.6517683	0.7694682	**0.315374**	0.8025385	0.5038747
	10	1.483124	1.820708	**1.080558**	2.320333	1.097706
	20	1.369658	1.579832	1.434874	1.752438	**1.149177**
	30	2.16972	2.35641	**2.002383**	3.166205	2.51553

Table 5-18. CVRMSE values between the imputed and original values for AR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17220 İzmir	*5*	0.3016547	0.3617041	**0.1458863**	0.3730521	0.2284227
	10	0.6984708	0.8743698	0.4963928	1.059893	**0.4746703**
	20	0.6638399	0.769265	0.6853732	0.8428866	**0.5308117**
	30	1.045093	1.206513	**0.9021433**	1.498334	0.9027198

By looking Tables 5-12 to 5-14, it can be said that the missing values in the precipitation data in the AR region give better results when imputed by the RF method for 5%, 10% and 30% missing values. On the other hand, it is seen that GAIN method gives results close to RF method and for 20% missing values, GAIN method gives better results than RF method. However, to summarize in general, RF and GAIN methods give close results. Figure 5-2 shows the graph in which the real data and the imputed data were created only for the days with missing values (30%

missingness). It is seen that some of the high peak values cannot be captured by any models.

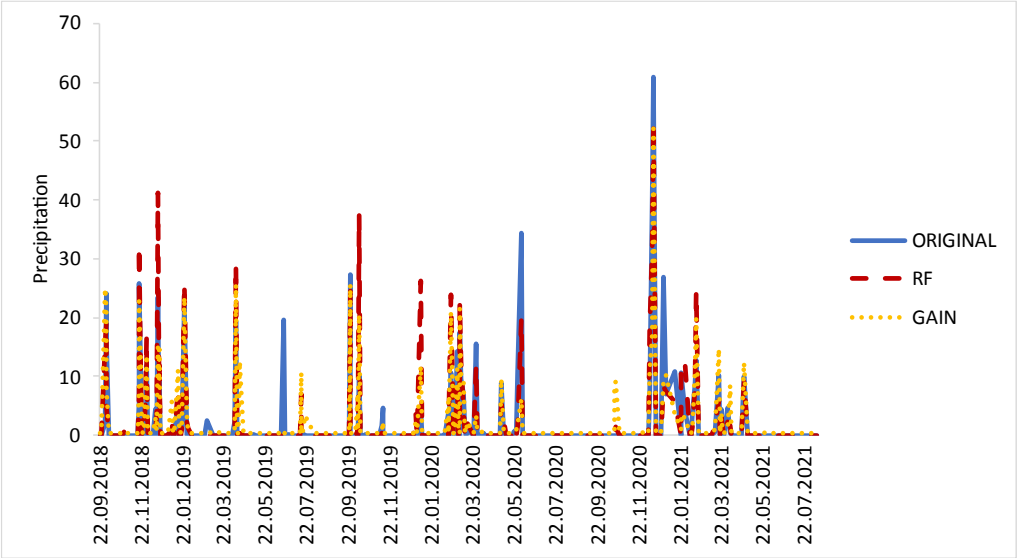

Figure 5-2. Precipitation imputations for AR region for 30% missing percentages

5.2.2 Results for Alanya (MR region)

Alanya is located in the Mediterranean region. For its climate, it can be said that like the Aegean region, it is hot and dry in the summer and warm and rainy in winter. As can be seen in Figure 4-5, the reference stations for this region do not cover the target station. Correlation values between stations are shown in Table 4-2 and Table 4-3 and according to tables, the correlations of the stations according to the temperature values are generally above 0.98 and are high, but it can be said that the correlation values for precipitation are a little low.

Table 5-19. NSE values between the imputed and original values for MR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17310 Alanya	*5*	0.9960956	0.999546	**0.9996237**	0.9988066	0.999598
	10	0.9919838	0.9990043	**0.9991471**	0.9980549	0.998926
	20	0.9858797	0.9980394	0.9978839	0.9957286	**0.998017**
	30	0.975396	0.9962363	0.996357	0.993003	**0.996383**

According to Table 5-17, while RF method gives good results in 5% and 10% missing data, GAIN method gives better results in 20% and 30%. In addition, KNN results are close to these two models.

Table 5-20. RMSE values between the imputed and original values for MR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17310 Alanya	*5*	0.4014989	0.1369026	**0.1246399**	0.2219699	0.128809
	10	0.5752948	0.2027545	**0.1876559**	0.2833845	0.210624
	20	0.7635318	0.2845128	0.2955821	0.4199442	**0.286172**
	30	1.007879	0.3941979	0.387826	0.5374775	**0.386451**

Table 5-21. CVRMSE values between the imputed and original values for MR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17310 Alanya	*5*	0.01965786	0.0066814	**0.006083353**	0.01083598	0.006263
	10	0.02826845	0.0098410	**0.009160189**	0.01382869	0.010228
	20	0.03771293	0.0138869	0.01443352	0.02049449	**0.013967**
	30	0.05016096	0.0192404	0.01894013	0.02624512	**0.018846**

By looking Tables 5-17 to 5.19, it can be said that the missing values in the temperature data in the MR region give better results when imputed by the RF method and GAIN. On the other hand, it is seen that KNN method gives results close to RF and GAIN. Figure 5-3 shows the graph in which the real data and the imputed data were created only for the days with missing values (30% missingness). According to Figure 5-3, it can be said that the models generally captured the peak values.

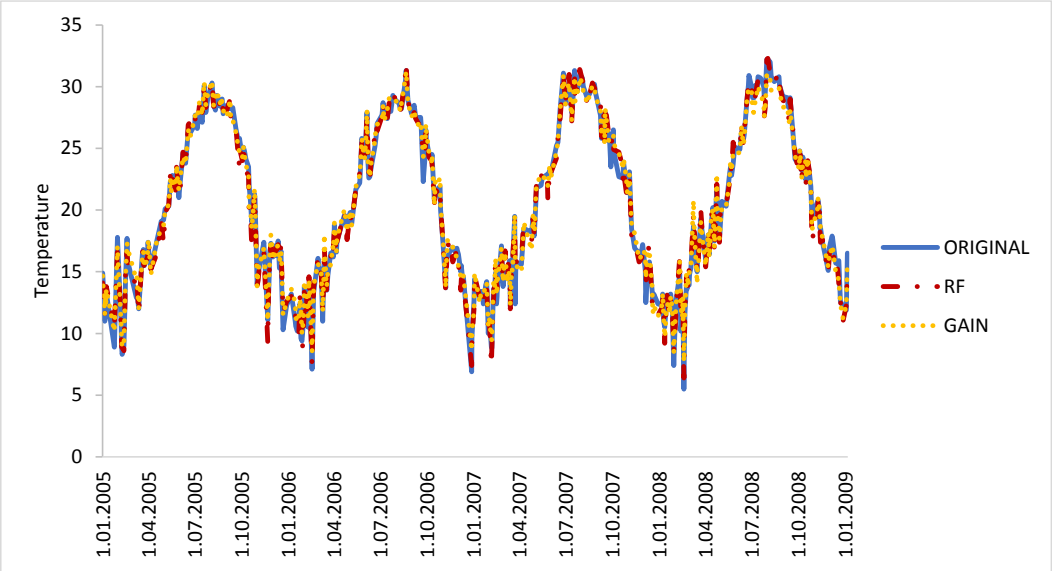

Figure 5-3. Temperature imputations for MR region for 30% missing percentages

Table 5-22. NSE values between the imputed and original values for MR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17310 Alanya	*5*	0.9970356	0.9972119	0.9999481	0.9740076	**0.999965**
	10	0.9943777	0.9986015	0.9826995	0.9515113	**0.999128**
	20	0.9911332	0.9917423	0.988257	0.9594863	**0.991156**
	30	0.954005	0.949344	0.9507305	0.920567	**0.960635**

When the coefficients of variation of the stations are examined, it is seen that the stations with the highest values are the stations in the MR region. According to Table 5-20, the GAIN model gives the best results for all missing percentages. RF is followed by KNN and GAIN methods. NSE values in these 3 models are very close to each other.

Table 5-23. RMSE values between the imputed and original values for MR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17310 Alanya	*5*	0.3561751	0.3454228	0.04711808	1.054681	**0.0389958**
	10	0.4905181	0.2446389	0.8604531	1.440516	**0.1931594**
	20	0.6160017	**0.5944677**	0.7089025	1.316737	0.6086673
	30	1.402985	1.472358	1.452068	1.843734	**1.297928**

Table 5-24. CVRMSE values between the imputed and original values for MR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17310 Alanya	*5*	0.1984441	0.1935954	0.02635452	0.6012734	**0.02027805**
	10	0.2665309	0.1371102	0.465994	0.750822	**0.1060669**
	20	0.3397661	**0.332133**	0.3925518	0.7059208	0.3331585
	30	0.8151856	0.8441389	0.7772877	1.010975	**0.6924665**

According to RMSE and CVRMSE tables for precipitation in MR region, GAIN gives the best results for almost all percentages of missing data, but also KNN results give results close to GAIN at certain percentages of missing. It has been observed that certain models are unable to capture some of the highest peak values.

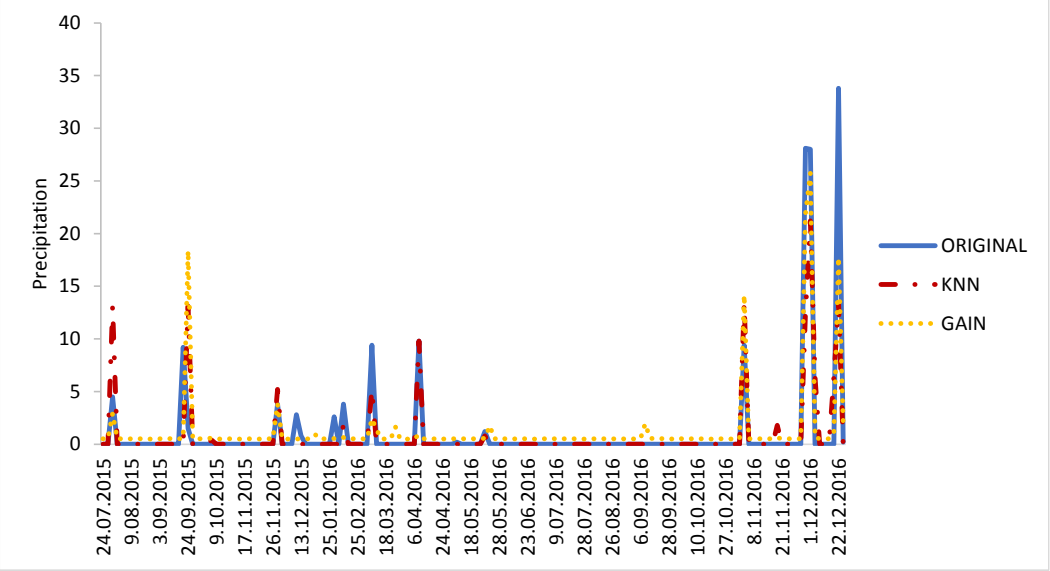

Figure 5-4. Precipitation imputations for MR region for 30% missing percentages

5.2.3 Results for Manyas (MAR region)

The climate of the Marmara region is transitional between the climates of the Black Sea region and Mediterranean region. Although it has a climate that changes frequently, it has a semi-arid climate. The air temperature is low and rainy in winter, while in summer is hot. As can be seen in Figure 4-2, the reference stations for this region are to cover this target station (Manyas). Correlation values between stations are shown in Table 4-2 and Table 4-3. The correlation between the reference stations and target station seems very high when looking for temperature. On the other hand, correlation values for precipitation have values higher than 0.70 on average.

Table 5-25. NSE values between the imputed and original values for MAR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17699 Manyas	*5*	0.9997218	0.9998695	**0.9999124**	0.9955369	0.999842
	10	0.9993493	**0.9995192**	0.999247	0.9942157	0.999181
	20	0.9984788	**0.9988926**	0.9985887	0.9845625	0.998038
	30	0.9981395	**0.9985879**	0.9984983	0.9827226	0.998118

Looking at the NSE table for the MAR region, it is seen that the KNN method gives better results than other methods. On the other hand, it is also seen that RF method gives slightly better results than KNN in temperature data with only 5% missing values. In general, GAIN and RF methods give very close results to KNN method.

Table 5-26. RMSE values between the imputed and original values for MAR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17699 Manyas	*5*	0.1378147	0.0943744	**0.07733892**	0.5519549	0.103749
	10	0.2107521	**0.1811696**	0.2267198	0.6283638	0.236443
	20	0.3222398	**0.2749341**	0.3103811	1.026536	0.365946
	30	0.3563659	**0.3104715**	0.3201635	1.085985	0.358459

Table 5-27. CVRMSE values between the imputed and original values for MAR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17699 Manyas	*5*	0.009034207	0.0061885	**0.005070379**	0.03619487	0.006803
	10	0.01381718	**0.0118845**	0.01486678	0.04117365	0.015489
	20	0.02109477	**0.0180182**	0.02032882	0.06737204	0.024044
	30	0.0233403	**0.0203469**	0.02100965	0.07112107	0.023557

According to the tables above, KNN gives good results for almost all missing percentages. While GAIN and RF give results close to KNN, RF method gives better results than GAIN. According to Figure 5-5, the results of the models are close to the original values.

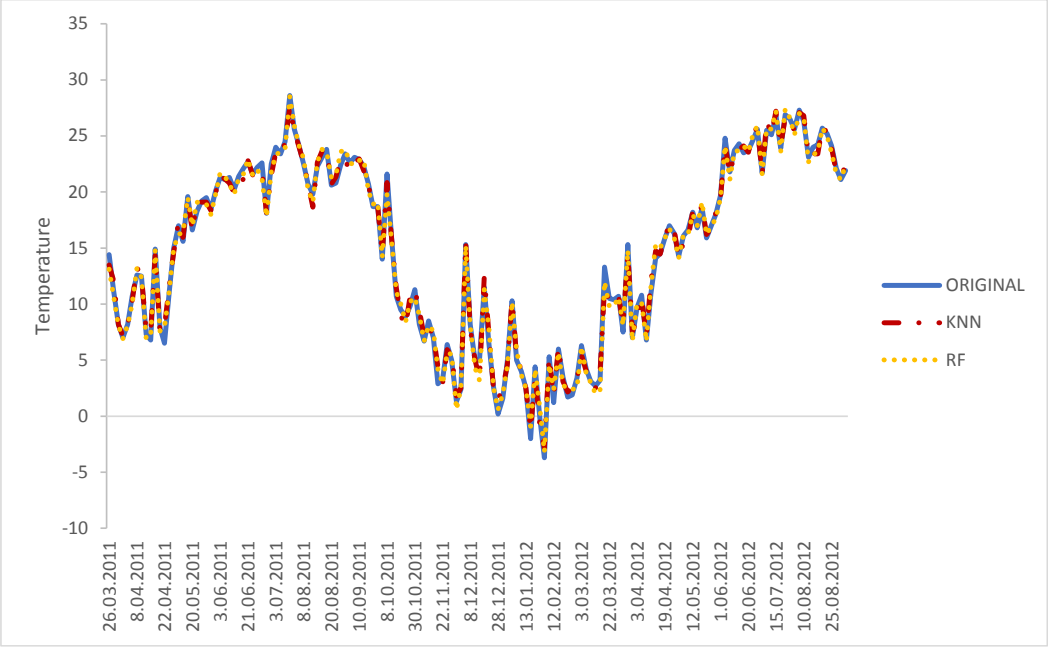

Figure 5-5. Temperature imputations for MAR region for 30% missing percentages

Table 5-28. NSE values between the imputed and original values for MAR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17699 Manyas	*5*	0.9747495	0.9788372	0.9896543	0.9753259	**0.991465**
	10	0.9122646	0.8971388	**0.9792053**	0.8507028	0.962138
	20	0.9455106	0.9461734	**0.9686388**	0.7729988	0.949814
	30	0.8345188	0.8151665	0.8434346	0.7712549	**0.873669**

According to the NSE table, GAIN gives better output in data with 5% and 30% missing values, while RF gives better results in data with 10% and 20% missing values. However, when using the GAIN method while imputing the precipitation values in the MAR region, it gives outputs close to the RF method.

Table 5-29. RMSE values between the imputed and original values for MAR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17699 Manyas	*5*	1.189429	1.088906	0.761348	1.175774	**0.69154**
	10	2.217128	2.400652	**1.079392**	2.892206	1.456492
	20	1.747268	1.736608	**1.325561**	3.566297	1.676847
	30	3.044933	3.218057	2.96177	3.579969	**2.660492**

Table 5-30. CVRMSE values between the imputed and original values for MAR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17699 Manyas	*5*	0.6616328	0.589024	0.415844	0.6456498	**0.3817686**
	10	1.235702	1.393352	**0.591778**	1.690071	0.8006597
	20	0.9338968	0.9715055	**0.7284609**	1.727249	0.891026
	30	1.75626	1.959115	1.702805	1.939084	**1.573118**

Considering the RMSE and CVRMSE values obtained for precipitation, RF and GAIN gave good results for all missing percentages. In data with some missing percentages, GAIN gave better outputs than RF, and in some, RF gave better outputs than GAIN. It was noticed that some models may have difficulty capturing the highest peak values.

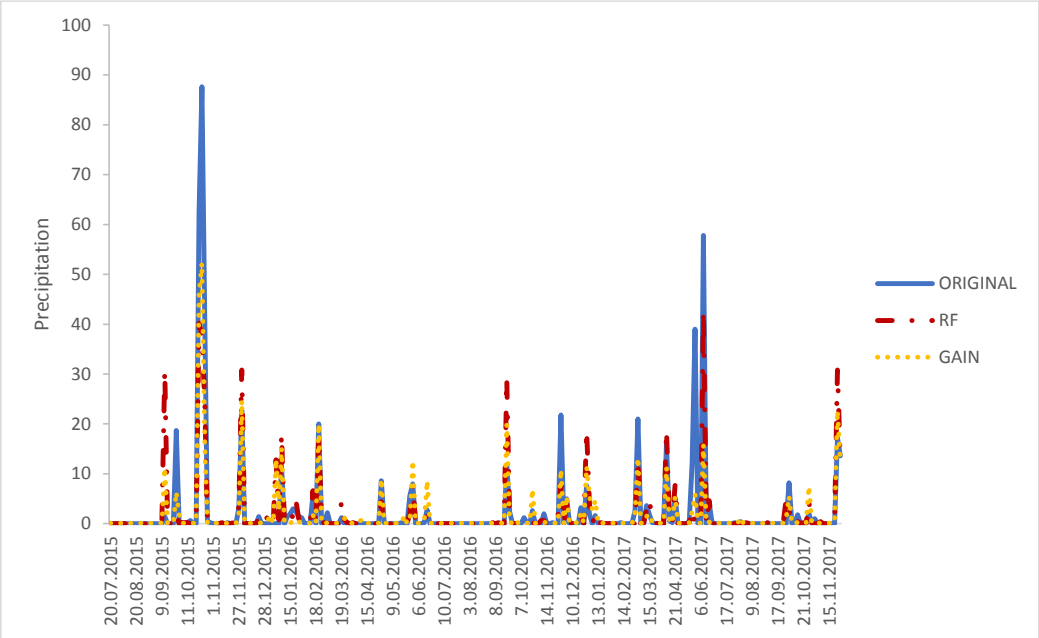

Figure 5-6. Precipitation imputations for MAR region for 30% missing percentages

5.2.4 Results for Bartın (BSR region)

Every season of the Black Sea region is rainy and there is no period when it is dry. While the most precipitation is in autumn and winter, the least precipitation is in summer. The annual temperature difference is less when compared to other regions. As can be seen in Figure 4-4, the reference values for this region do not cover the target station (Bartın). Correlation values between stations are shown in Table 4-2 and Table 4-3. Although the correlation values between the reference stations and the target station for temperature remain low compared to other regions, they are still high. For precipitation, however, correlation values are slightly higher than for other regions.

Table 5-31. NSE values between the imputed and original values for BSR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17020 Bartın	*5*	0.9972082	**0.9997049**	0.9996966	0.9917406	0.999443
	10	0.9964047	0.9984163	**0.9984583**	0.9882326	0.998381
	20	0.9939825	0.9979692	**0.9982398**	0.9830134	0.997916
	30	0.9896612	0.9963506	0.9966888	0.9830134	**0.996813**

According to the NSE results above, the model that gave the best output was the RF model. After the RF model, GAIN and KNN models gave good outputs.

Table 5-32. RMSE values between the imputed and original values for BSR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17020 Bartın	*5*	0.4085136	**0.13281**	0.134663	0.7026501	0.182521
	10	0.4635904	0.3076763	**0.3035744**	0.8386961	0.311116
	20	0.5997536	0.3484137	**0.3243743**	1.007667	0.352916
	30	0.7861401	0.4670639	0.4448933	1.00766	**0.436506**

Table 5-33. CVRMSE values between the imputed and original values for BSR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17020 Bartın	*5*	0.03050687	**0.009952767**	0.01009295	0.05248397	0.01368619
	10	0.03454646	0.02305245	**0.02274686**	0.06273257	0.02331393
	20	0.04450953	0.02610012	**0.02431602**	0.07593592	0.02650542
	30	0.05788135	0.03486424	0.03324439	0.075935	**0.03254033**

According to the results of Table 5-29 to 5-31, the three models (KNN, RF and GAIN) gave close results to each other. RF model can be used for data with 5%, 10% and 20% missing values, while GAIN can be used for data with 30% missing values. According to the Figure 5-7, It appears that the peak value could not be captured by any model on a single date.

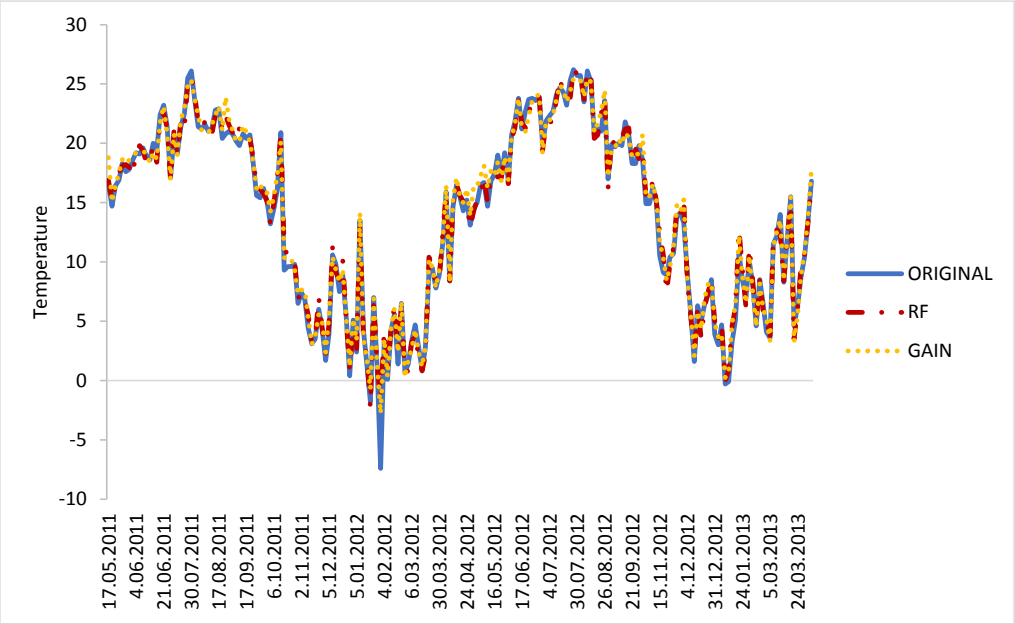

Figure 5-7. Temperature imputations for BSR region for 30% missing percentages

Table 5-34. NSE values between the imputed and original values for BSR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17020 Bartın	*5*	0.9902319	0.9894793	0.9880879	**0.9945504**	0.991627
	10	0.9546664	0.93766616	0.9831147	0.941424	**0.988772**
	20	0.9218899	0.8893289	**0.9414467**	0.8276338	0.920602
	30	0.8935196	0.8176955	**0.9211632**	0.8397731	0.893942

According to Table 5-32, GAIN gave better results in precipitation data with low missing percentages, while RF gave better results when the missing percentage was higher. It can be said that the MICE method gave good results for data with 5% missing value.

Table 5-35. RMSE values between the imputed and original values for BSR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17020 Bartın	*5*	0.8524433	0.8846725	0.9413574	**0.6367129**	0.7892188
	10	1.836413	2.153466	1.120766	2.087468	**0.913939**
	20	2.410537	2.869309	**2.087064**	3.580851	2.430333
	30	2.814459	3.682636	**2.421724**	3.452454	2.808866

Table 5-36. CVRMSE values between the imputed and original values for BSR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17020 Bartın	*5*	0.2861197	0.3000321	0.3186685	**0.2119858**	0.2663636
	10	0.6196478	0.7190586	0.3715439	0.67938	**0.2978943**
	20	0.8494316	1.03899	**0.7391886**	1.350547	0.8417217
	30	1.00967	1.461538	**0.8438669**	1.292518	0.989121

RMSE and CVRMSE results show that using MICE method for imputing 5% missing data will give better outputs, GAIN method will give better outputs for imputing 10% missing data and using RF method for 20% and 30% missing data will give better outputs. It is seen that some of the high peak values cannot be captured by any models.

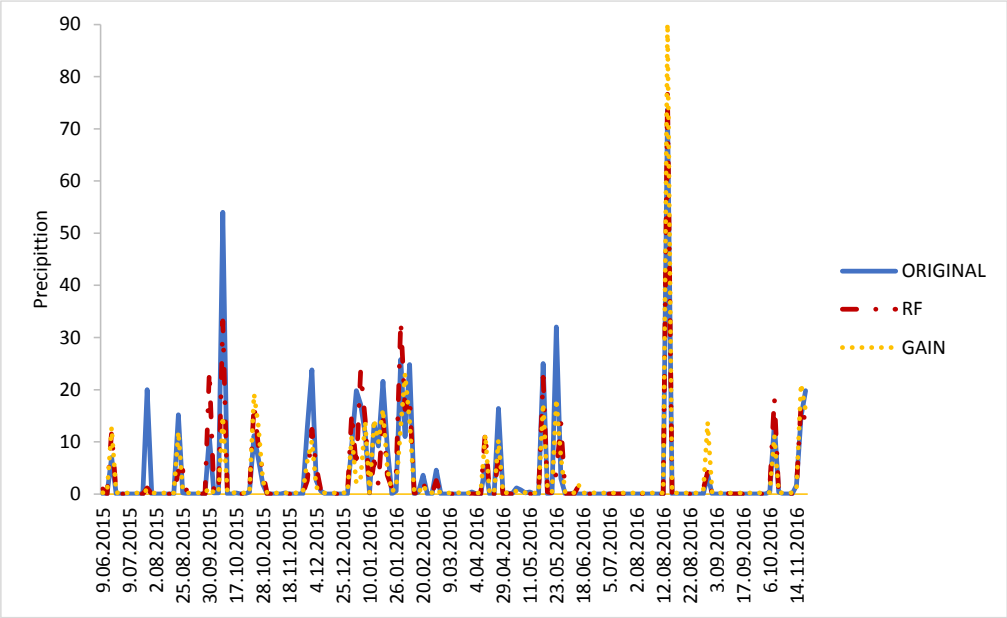

Figure 5-8. Precipitation imputations for BSR region for 30% missing percentages

5.2.5 Results for Ağrı (EAR region)

The target station is located in a region where Turkey's longest winter months are experiences, and the weather is very cold. There is little precipitation here, but many snow falls. Summer months are hot. As can be seen in Figure 4-6, the reference values for this region do not cover the target station (Agri). Correlation values between stations are shown in Table 4-2 and Table 4-3. Reference stations do not cover the target station but are very close to the target station. Temperature correlations are at least 0.96 and very high. On the other hand, it can be said that they have lower correlations for precipitation than other regions. In general, the lowest values of all correlation values for precipitation are in this region.

Table 5-37. NSE values between the imputed and original values for EAR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17099 Ağrı	*5*	0.9971968	**0.9994032**	0.9994048	0.9988381	0.999102
	10	0.9925921	**0.9989632**	0.9989475	0.9965789	0.998474
	20	0.9883985	**0.9977051**	0.9976587	0.9947074	0.997658
	30	0.9764984	0.9960396	**0.9962541**	0.9861596	0.995684

Among the methods used in temperature imputation in this region, it can be said that KNN gives better results than other methods. RF method values are also very close to KNN values, even in the data with some missing percentages, they gave almost the same results. The outputs of the GAIN method are also very close to these two methods.

Table 5-38. RMSE values between the imputed and original values for EAR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17099 Ağrı	*5*	0.6465284	**0.2983103**	0.2979019	0.4162372	0.365933
	10	1.051013	**0.3931901**	0.396157	0.7142321	0.47701
	20	1.315275	**0.584979**	0.5908619	0.8883718	0.590929
	30	1.872012	0.7684747	**0.7473734**	1.436593	0.802197

Table 5-39. CVRMSE values between the imputed and original values for EAR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17099 Ağrı	*5*	0.09984865	**0.04680547**	0.04674046	0.06536257	0.0574477
	10	0.1592107	**0.06157209**	0.06206073	0.1117591	0.4627544
	20	0.1934919	**0.0917928**	0.09263069	0.1388291	0.09248127
	30	0.2644477	0.119555	**0.1167334**	0.2083715	0.1278061

According to the RMSE and CVRMSE tables, KNN and RF give close results.

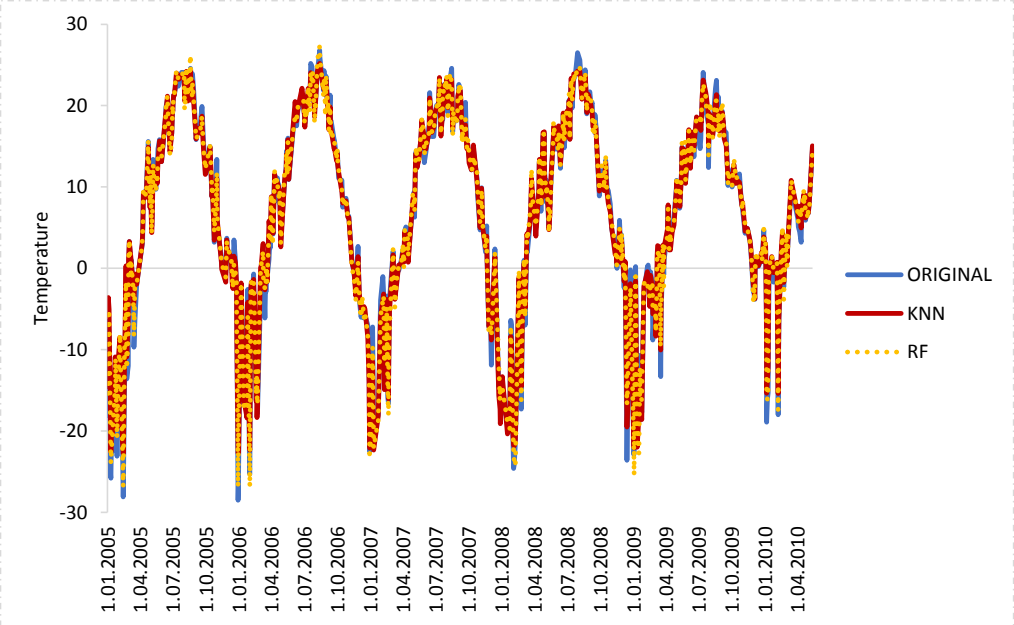

Figure 5-9. Temperature imputations for EAR region for 30% missing percentages

Table 5-40. NSE values between the imputed and original values for EAR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17099 Ağrı	*5*	0.9936697	0.9964418	0.9962139	**0.9967674**	0.994574
	10	0.9310589	0.900995	0.9455347	0.8938903	**0.972272**
	20	0.8998842	0.9092183	**0.9351905**	0.8539343	0.913004
	30	0.9051223	0.8866205	0.8522855	0.6978361	**0.891097**

When the NSE values obtained as a result of the imputation of the data of this target station with all the missing percentages are examined, it is seen that different methods give good results for each missing percentage. In general, it is possible to say that the GAIN method gives good results.

Table 5-41. RMSE values between the imputed and original values for EAR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17099 Ağrı	*5*	0.2776898	0.2081893	0.2147533	**0.1984358**	0.2570993
	10	0.9164002	1.098183	0.8145281	1.136904	**0.581172**
	20	1.104326	1.051587	**0.8885166**	1.33389	1.029425
	30	1.095254	1.175204	1.341398	1.918525	**1.151769**

Table 5-42. CVRMSE values between the imputed and original values for EAR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17099 Ağrı	*5*	0.1932964	0.144667	0.2236301	**0.1371068**	0.17765
	10	0.6846913	0.8363723	0.1506217	0.8163689	**0.3898477**
	20	0.8182083	0.7671219	**0.6425863**	0.8641279	0.6951252
	30	0.8089738	1.442789	0.9558205	1.374547	**0.7554962**

In general, it can be said that the GAIN method gives good outputs. After GAIN method, RF method also gives outputs. It has been observed that certain models are unable to capture certain high peak values.

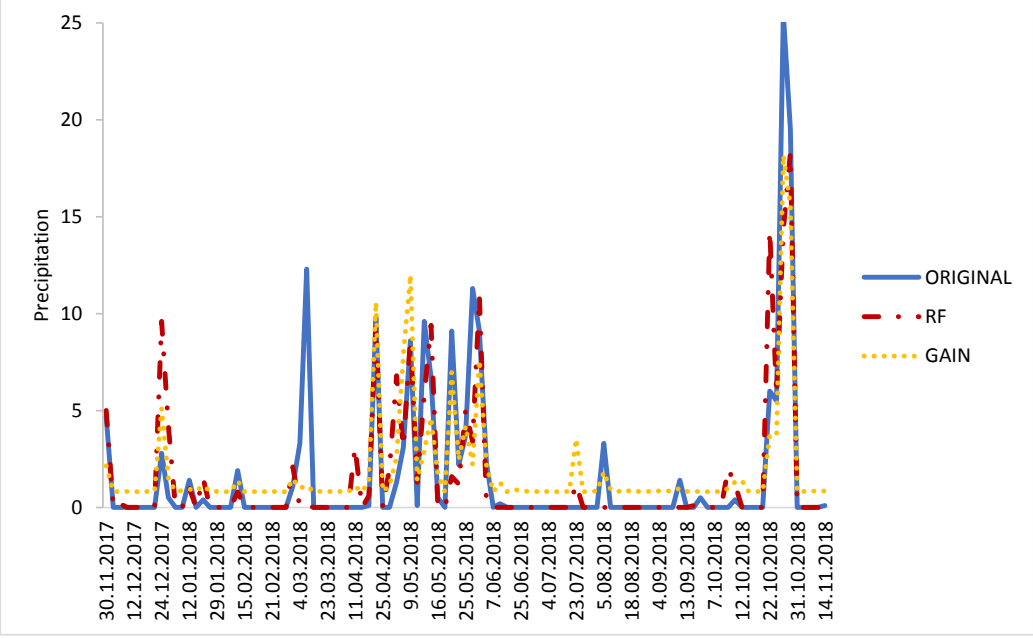

Figure 5-10. Precipitation imputations for EAR region for 30% missing percentages

5.2.6 Results for Konya (CAR region)

In the Central Anatolia region, the summer months are dry and hot, and the winter months are cold and snowy. The target station is in one of the provinces with the least rainfall. As can be seen in Figure 4-7, the reference stations for this region do to cover this target station (Konya). Correlation values between stations are shown in Table 4-2 and Table 4-3. While the correlation of temperature values is high, the correlation values of precipitation values is lower.

Table 5-43. NSE values between the imputed and original values for CAR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17245 Konya	*5*	0.9973421	**0.9997799**	0.9997763	0.998339	0.99967
	10	0.9948012	**0.9995311**	0.9994071	0.9972905	0.999259
	20	0.9880597	0.9980722	**0.998302**	0.9938895	0.998126
	30	0.9835443	0.9974027	0.997264	0.9904236	**0.997677**

The NSE outputs of the imputation methods applied to the temperature data in the CAR region are as above. Accordingly, RF, GAIN and KNN models provide similar results in filling in missing data. While KNN and RF are better for low missing percentages, GAIN and RF give better outputs at high missing percentages.

Table 5-44. RMSE values between the imputed and original values for CAR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17245 Konya	*5*	0.4759635	**0.1369785**	0.1380836	0.376261	0.1677
	10	0.6656621	**0.1999148**	0.224799	0.4805548	0.257973
	20	1.008808	0.4053494	**0.3804306**	0.7216697	0.399649
	30	1.184291	0.4705031	0.4829048	0.9034467	**0.444948**

Table 5-45. CVRMSE values between the imputed and original values for CAR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17245 Konya	*5*	0.0323879	**0.009256024**	0.009333412	0.02540343	0.01133674
	10	0.04559546	**0.01350618**	0.01519536	0.03249616	0.01716245
	20	0.07014609	0.02740245	**0.02574197**	0.0487587	0.02709696
	30	0.08343418	0.03176061	0.03262362	0.06124758	**0.0300739**

RMSE and CVRMSE values confirm the results mentioned above. As a result, for temperature data in the CAR region, KNN and RF can be used for imputation in data with low missing percentages, while RF and GAIN models can be used for imputation in data with high missing percentages.

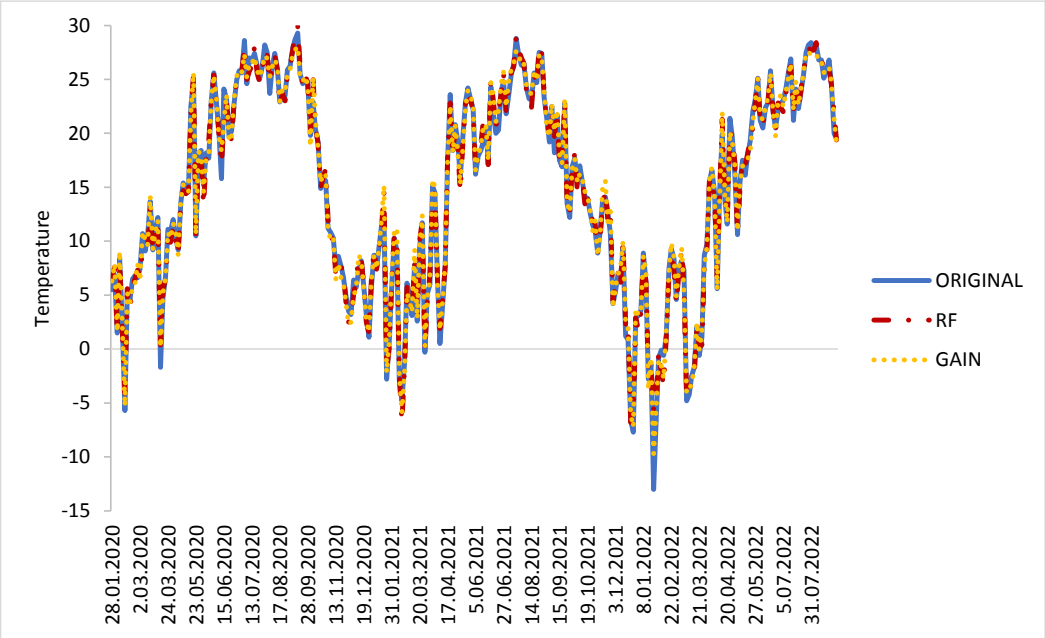

Figure 5-11. Temperature imputations for CAR region for 30% missing percentages

Table 5-46. NSE values between the imputed and original values for CAR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17245 Konya	*5*	0.9738951	0.9470093	**0.9882785**	0.9365672	0.973226
	10	0.9775423	**0.9877558**	0.9580781	0.9686056	0.986822
	20	**0.8939047**	0.847981	0.8632558	0.7965731	0.889615
	30	**0.879225**	0.8567647	0.8255449	0.5654972	0.858484

For the imputation of precipitation data, SAA gives better outputs than other regions. While KNN, RF and GAIN give good NSE values in data with low missing percentages, SAA and GAIN give good NSE values in data with high missing

percentages. GAIN's imputation of precipitation data in this region is more stable than other models.

Table 5-47. RMSE values between the imputed and original values for CAR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17245 Konya	*5*	0.4700381	0.6696866	**0.3149656**	0.7327041	0.4760271
	10	0.4359678	**0.3219112**	0.5956516	0.5154632	0.3339658
	20	**0.9475884**	1.134281	1.075786	1.312127	0.9665535
	30	**1.011021**	1.101024	1.215104	1.917644	1.094394

Table 5-48. CVRMSE values between the imputed and original values for CAR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17245 Konya	*5*	0.5470894	0.8025135	**0.36272**	0.8663854	0.5341576
	10	0.4861072	**0.3777408**	0.4113609	0.5934732	0.3888861
	20	**1.125105**	1.540362	1.418325	1.589346	1.132608
	30	**1.10422**	1.333812	1.487692	1.99938	1.313206

RMSE and CVRMSE values are similar to NSE values. It confirms the above-mentioned results. It is seen that some of the high peak values cannot be captured by any models.

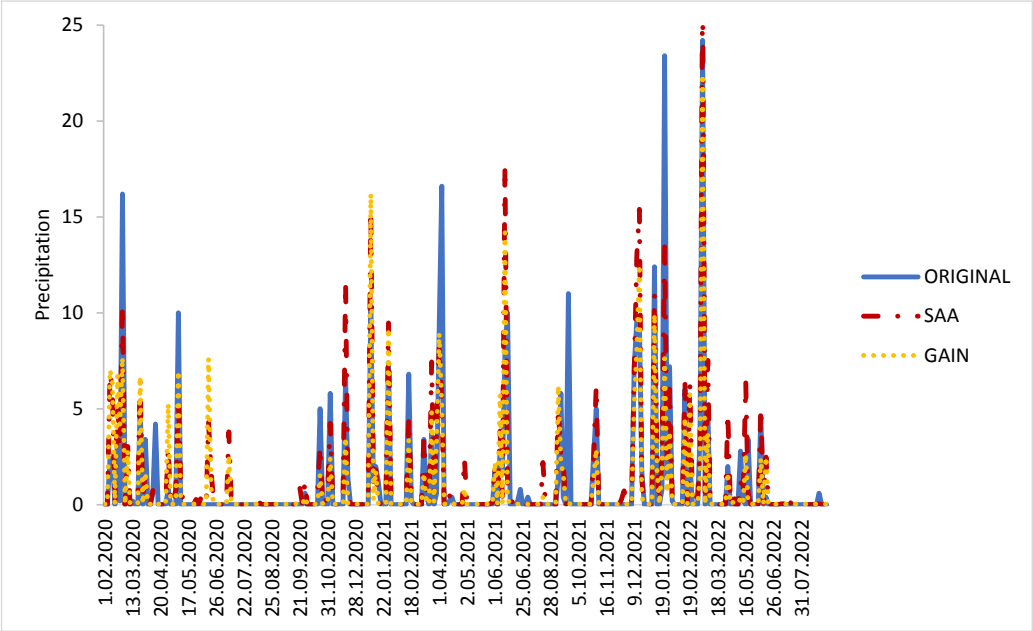

Figure 5-12. Precipitation imputations for CAR region for 30% missing percentages

5.2.7 Results for Birecik (SAR region)

For the climate of this region, it can be said that the summer months are hot and dry, and the winter months are abundantly rainy. The target station is in one of the provinces with the least rainfall. As can be seen in Figure 4-8, the reference stations for this region do cover this target station (Birecik). Correlation values between stations are shown in Table 4-2 and Table 4-3. When the correlation values for temperature are considered, it can be said that the values are generally 0.98 and higher. The average correlation values for the precipitation are approximately 0.64.

Table 5-49. NSE values between the imputed and original values for SAR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17966 Birecik	*5*	0.9980832	**0.9992577**	0.9992054	0.998602	0.998943
	10	0.9964839	**0.9986758**	0.9985328	0.9968696	0.997891
	20	0.9927221	**0.9967356**	0.9965658	0.9939375	0.996449
	30	0.9892091	**0.9956666**	0.9941658	0.9918009	0.99501

When the NSE outputs of the models used in the imputation of the temperature data of SAR region are examined, it is seen that KNN model makes the best imputation. Apart from KNN model, RF and GAIN models also give results close to KNN.

Table 5-50. RMSE values between the imputed and original values for SAR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17966 Birecik	*5*	0.4171852	**0.259608**	0.2686128	0.3562787	0.309802
	10	0.5650266	**0.3467472**	0.3649909	0.5331377	0.437594
	20	0.8129095	**0.5444299**	0.5584093	0.7419319	0.567813
	30	0.9898504	**0.6272679**	0.727831	0.8628252	0.673105

Table 5-51. CVRMSE values between the imputed and original values for SAR region temperature data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17966 Birecik	*5*	0.0231917	**0.01437278**	0.01487141	0.0197203	0.01716203
	10	0.0315219	**0.01919932**	0.02021704	0.02950773	0.02447223
	20	0. 0456826	**0. 03011504**	0. 03090458	0. 0410653	0. 03134387
	30	0.05604659	**0.0351692**	0.04031057	0.04772886	0.03738701

When CVRMSE and RMSE values are examined, it is seen that KNN model gives the best results and RF model gives better results than GAIN. GAIN model gives better results than RF for data with only 30% missing value.

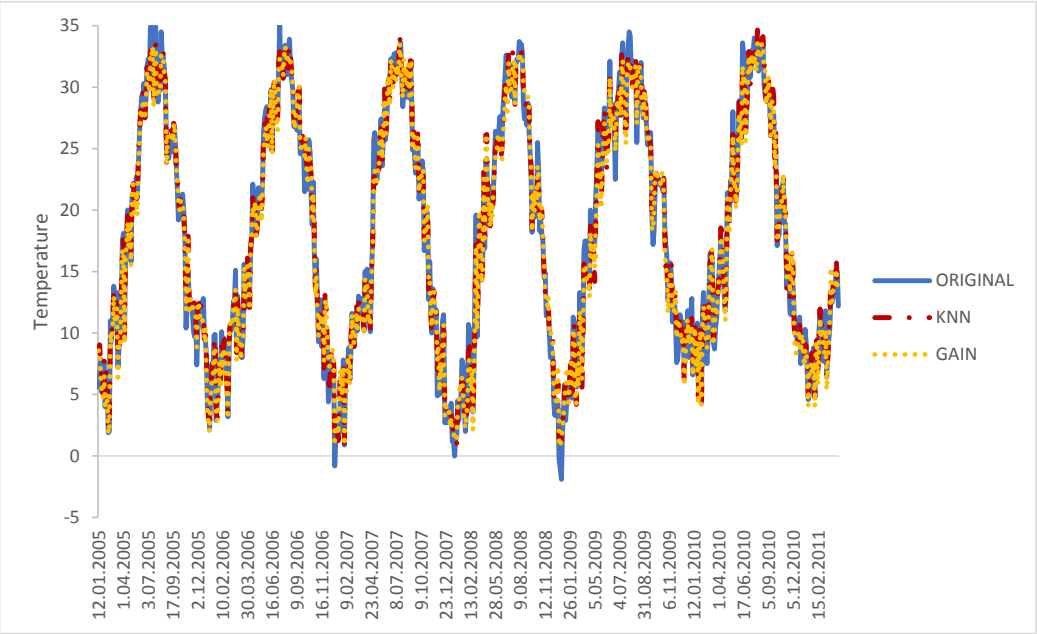

Figure 5-13. Temperature imputations for SAR region for 30% missing percentages

Table 5-52. NSE values between the imputed and original values for SAR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17966 Birecik	*5*	0.9816342	0.9953006	**0.9992666**	0.9944817	0.996896
	10	0.9901267	0.9942272	**0.9966036**	0.9840149	0.993486
	20	0.9500764	**0.9969596**	0.9907436	0.9787458	0.991082
	30	0.9771766	**0.9912206**	0.9894263	0.9820924	0.985681

RF model gives better results than other models for data with 5% and 10% missing values according to NSE values. For data with 20% and 30% missing values, the KNN model gives better results than other models.

Table 5-53. RMSE values between the imputed and original values for SAR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17966 Birecik	*5*	0.4579337	0.2316431	**0.09150947**	0.2510142	0.182496
	10	0.3357591	0.2567385	**0.1969276**	0.427224	0.2727239
	20	0.7550061	**0.1863208**	0.3251016	0.4926286	0.3191112
	30	0.5104908	**0.3166132**	0.3474647	0.4521853	0.4043523

Table 5-54. CVRMSE values between the imputed and original values for SAR region precipitation data

Target Station	Missing Percentage (%)	SAA	KNN	RF	MICE	GAIN
17966 Birecik	*5*	0.6765461	0.3714746	**0.1460826**	0.3910671	0.290942
	10	0.5114665	0.4197917	**0.3160091**	0.658916	0.4138756
	20	1.007767	**0.3033416**	0.5126602	0.8020293	0.5182508
	30	0.7045202	**0.5367804**	0.5427067	0.7071684	0.6892407

The 2 tables above show similar features to the outputs of the NSE table, showing that using RF model for data with low missing percentages and using KNN model for data with high missing percentages will yield better results. It was noticed that some models may have difficulty capturing the highest peak values.

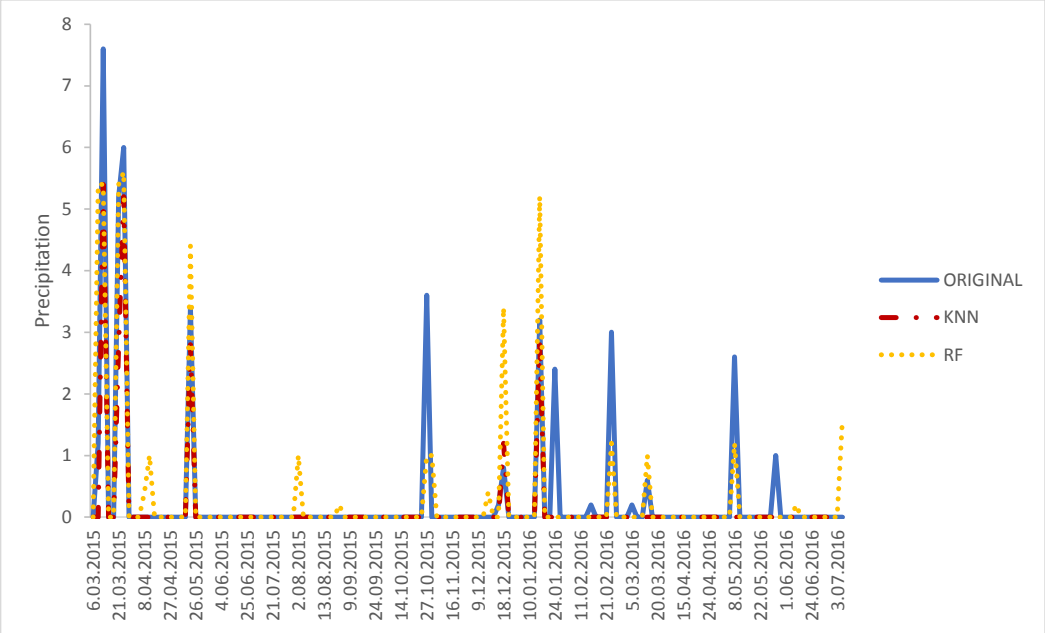

Figure 5-14. Precipitation imputations for SAR region for 30% missing percentages

5.2.8 Results of Block of Missing Data Imputation

In this part of the study, we want to see the performance of the methods when the block data is missing. Usually, meteorological data has this type of block-missing values. At the target station in the AR region, an 8-month missing period was created and imputed in both precipitation and temperature data sets with the parameter values specified in Section 5.1.6. The reason for choosing this region is that the variations in temperature and precipitation values in this region are more stable than other regions.

The results of the models established for the temperature data are as follows.

Table 5-55. NSE values between the imputed and original values for AR region temperature data

Target Station	SAA	KNN	RF	MICE	GAIN
17220 İzmir	0.9947967	**0.9987793**	0.9987271	0.9948582	0.9971718

Looking at the NSE values, it was seen that the KNN and RF models gave the best results. On the other hand, GAIN gave values close to these 2 models. There is no obvious difference between these models for temperature.

Table 5-56. RMSE values between the imputed and original values for AR region temperature data

Target Station	SAA	KNN	RF	MICE	GAIN
17220 İzmir	0.5544926	**0.2670307**	0.2726829	0.5480512	0.4064621

Table 5-57. CVRMSE values between the imputed and original values for AR region temperature data

Target Station	SAA	KNN	RF	MICE	GAIN
17220 İzmir	0.03049729	**0.01456505**	0.0148562	0.0299109	0.0222655

The results of the RMSE and CVRMSE tables confirm the results of the NSE table. Accordingly, it can be said that the model that gives the best result in the data with an 8-month missing values period for temperature is the KNN model.

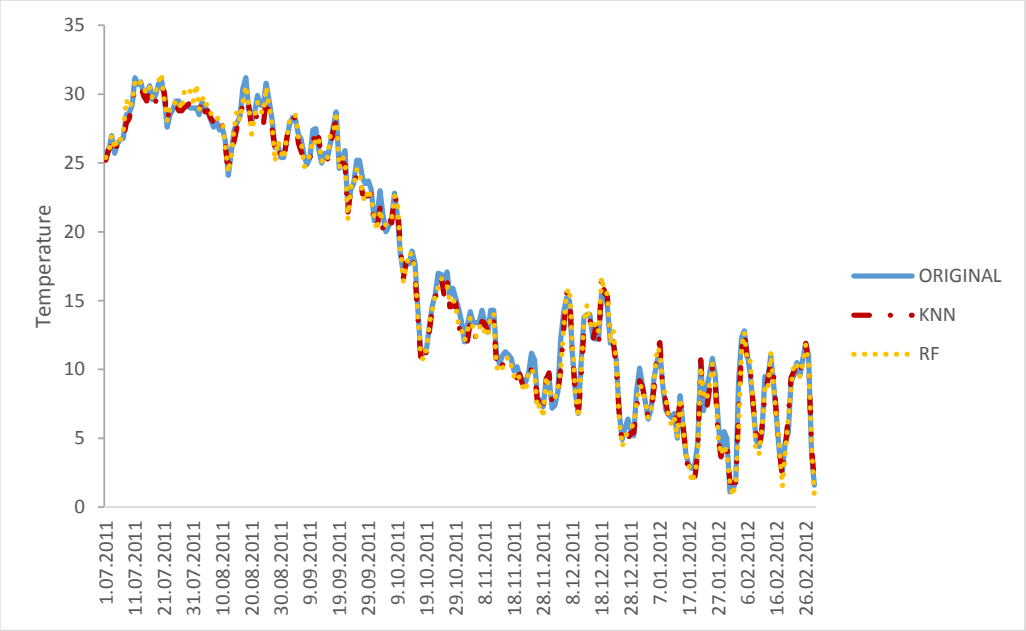

Figure 5-15. Temperature imputations for AR region

The results of the models established for the precipitation data are as follows.

Table 5-58. NSE values between the imputed and original values for AR region precipitation data

Target Station	SAA	KNN	RF	MICE	GAIN
17220 İzmir	0.9257152	0.8889561	0.9168596	0.8226968	**0.9407895**

Looking at the NSE values, it has been observed that the GAIN method gives much better results than other results.

Table 5-59. RMSE values between the imputed and original values for AR region precipitation data

Target Station	SAA	KNN	RF	MICE	GAIN
17220 İzmir	2.114044	2.584603	2.236506	3.266046	**1.887397**

Table 5-60. CVRMSE values between the imputed and original values for AR region precipitation data

Target Station	SAA	KNN	RF	MICE	GAIN
17220 İzmir	1.039952	1.376077	1.094945	1.670559	**0.8738868**

By looking at the 2 tables above, it can be said that GAIN gives the best results, then RF model gives good results. It is seen that any models cannot capture some of the high peak values.

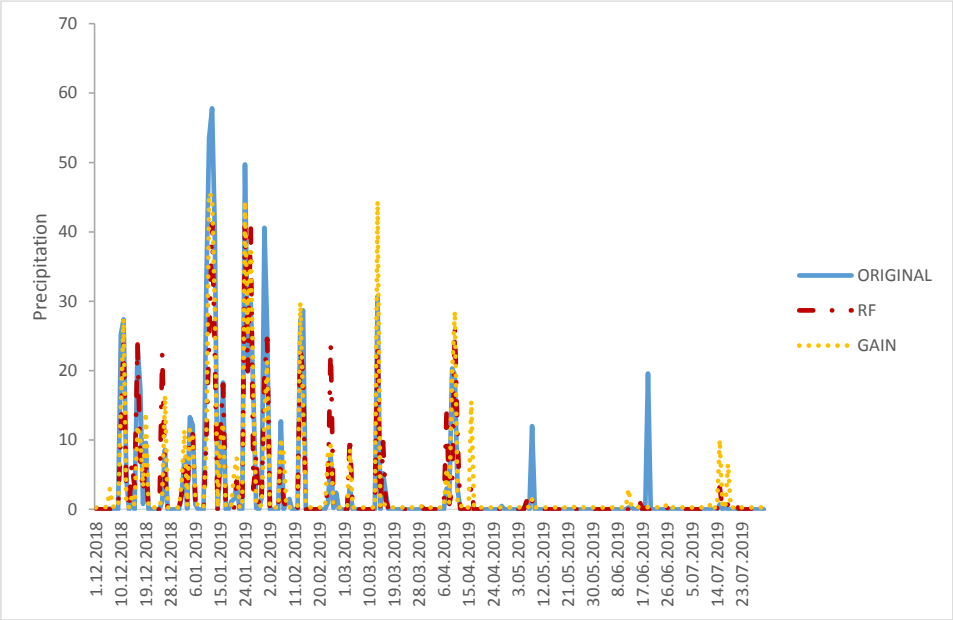

Figure 5-16. Precipitation imputations for AR region

5.3 Discussion

In this section, the performance of five different models at the specified stations was compared. To make these comparisons, NSE, RMSE, and CVRMSE values were used based on varying percentages of missing data.

In the AR region, the temperature values remain relatively stable. Through our imputation study, it has been found that the RF method outperforms other methods in all cases of missing data. For precipitation data in the AR region, it has been analyzed NSE and RMSE values and determined that the RF model provides more accurate results for data with 5% and 30% missing values, while the GAIN model is more accurate for data with 10% and 20% missing values. In terms of CVRMSE values, the RF model excels in data with 5% and 30% missing values, while the GAIN model performs better in data with 10% and 20% missing values.

In the MR region, which generally has a hot climate, the performance of the RF method in data with 5% and 10% missing data is better than other methods, while

the GAIN method is better in data with 20% and 30% missing values. When analyzing precipitation values with a high coefficient of variation, the GAIN method was found to be more accurate at predicting missing values for all percentages. While the GAIN method captured some peak points better than others, it also created non-existent peak points.

After analyzing the NSE, RMSE, and CVRMSE values for the MAR region, it was discovered that the RF model performed well in the dataset with 5% missing values, while the KNN model worked well in the remaining data with other missing percentages, producing results that were similar to the original values. The study revealed that the GAIN method was more effective in datasets with 5% and 30% missing values, while the RF method worked better in datasets with 10% and 20% missing percentages, based on their NSE, RMSE, and CVRMSE values.

Based on the analysis of various models in different datasets with varying percentages of missing values, it has been observed that the KNN model works well in the BSR region with 5% missing values, while the RF model is more effective in datasets with 10% and 20% missing percentages. Moreover, the GAIN model performs better in datasets with 30% missing values. In the case of the least precipitation variation in this region, the MICE model is more effective in the dataset with 5% missing values, while the GAIN model works well in datasets with 10% missing values. Additionally, the RF models are more effective in datasets with 20% and 30% missing percentages.

In the EAR region where temperature values are very variable, the KNN model is effective in datasets with 5%, 10% and 20% missing values. However, in datasets with 30% missing values, the GAIN model performs better than other models. For precipitation data in the EAR region, where there is the least correlation between target and reference stations, the MICE model with 5% missing value, GAIN with 10% and 30% missing value, and RF models with 20% missing value have better

outputs. It should be noted that no model has been able to accurately predict the peaks in the data.

Based on the NSE, RMSE, and CVRMSE values obtained from imputing temperature data in the CAR region, it was found that KNN methods are effective for datasets with 5% and 10% missing values, while RF and GAIN methods perform better for datasets with 20% and 30% missing values, respectively. As for precipitation data in the same region, RF method works best for datasets with 5% missing values, while KNN method is suitable for datasets with 10% missing values. For the remaining missing values, the SAA method provides the closest outputs to the actual values. Figures 5-12 demonstrate that both models are capable of accurately estimating peak points, as evidenced by their results being close to the original values.

According to the temperature data of the SAR region, the KNN method provides the best results for all missing values. Additionally, the GAIN and RF methods also perform well, with results close to those of the KNN method. For precipitation data, the study found that the RF method is more effective for datasets with 5% and 10% missing values, while the KNN method is better suited for datasets with 20% and 30% missing values, based on the NSE, RMSE, and CVRMSE values. Notably, neither the KNN nor RF models for precipitation data appear to converge to any peak values.

Another study for the Aegean region was the complete extraction of an 8-month part from both temperature and precipitation data. According to this study, it is seen that the KNN model gives more consistent results compared to other models at the stage of imputing the temperature values, and this model is followed by the RF model. On the other hand, in the study conducted for precipitation data, it was revealed that the GAIN model gave better results compared to other models by far. It was found that the GAIN model mostly predicted the peaks better.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 General

This thesis focuses on the imputation of missing values for the daily average temperature and total precipitation data obtained from the Turkish Meteorology General Directorate. The inspiration of the thesis is the investigation and imputation of daily meteorological data with different missing percentages. According to the research, no study on this subject has been found in Turkey. Within the scope of the thesis, it was employed five different model to impute missing values. To accurately evaluate the efficacy of these models, three distinct evaluation methods were applied.

The study started with the data taken from the stations in seven regions of Turkey since the temperature and precipitation distribution of these regions have different characteristics. The meteorological data were arranged using SAS software. In the ongoing studies, a specific station was chosen as the target for each region. Subsequently, five reference stations that were in close proximity and had high correlations with the target station were selected. Following that, some values were subtracted from the target stations to create missing values. The percentage of missing values was selected as 5, 10, 20, and 30 percent of the overall data. Five distinct imputation methods were then used to fill in the missing data using data from the reference stations. The five methods employed were the Simple Arithmetic Average Method (SAA), K-Nearest Neighbor Method (KNN), Random Forest Model (RF), Multiple Imputation by Chained Equation Method (MICE), and General Adversarial Imputation Network (GAIN). The first four methods were utilized in R, while the last method was utilized in Python. The methods' performances were compared using NSE, RMSE, and CVRMSE methods.

As a result of the study, the following outputs were obtained.

- RF gave the best results for the AR region.
- In the study conducted for temperature values in the MR region, RF gave results close to the real values in data with low missing percentages, while GAIN gave better results at high values. For precipitation data with high variation, GAIN gave results close to the true values.
- It was found that KNN performed well in predicting temperature values in the MAR region. However, when it comes to precipitation values with varying missing percentages, GAIN and RF were found to be more effective in providing accurate results.
- In the study conducted for the temperature data of the BSR region, GAIN gave good outputs in data with high missing percentages, while RF gave low ones. Additionally, the study showed that RF performed well in precipitation data with less variability, generating values that were close to the original results.
- The EAR region's temperature data showed high variability, and the study found that GAIN yielded good results for data with high missing percentages, while KNN gave good outputs for data with low missing percentages.
- Based on the study conducted on temperature data in the EAR region, it was discovered that GAIN produced accurate results for data with high missing percentages, while KNN performed well for data with low missing percentages. Similarly, in precipitation data, the SAA method was found to yield good outputs.
- It has been observed that the KNN method for the study of temperature values in the SAR region, and the RF and KNN methods for precipitation give good results.
- In the AR region, where the 8-month division was extracted, KNN in temperature data and GAIN in precipitation data gave good results.

Separate tables were created for temperature and precipitation to display which model performed better at each loss percentage in all regions.

Table 6-1. Results of temperature

Regions	**5%**	**10%**	**20%**	**30%**
AR	RF	RF	RF	RF
MR	RF	RF	GAIN	GAIN
MAR	RF	KNN	KNN	KNN
BSR	KNN	RF	RF	GAIN
EAR	KNN	KNN	KNN	RF
CAR	KNN	KNN	RF	GAIN
SAR	KNN	KNN	KNN	KNN

Table 6-2. Results for precipitation

Regions	**5%**	**10%**	**20%**	**30%**
AR	RF	RF	GAIN	RF
MR	GAIN	GAIN	GAIN	GAIN
MAR	GAIN	RF	RF	GAIN
BSR	MICE	GAIN	RF	RF
EAR	MICE	GAIN	RF	GAIN
CAR	RF	KNN	SAA	SAA
SAR	RF	RF	KNN	KNN

To sum up, it has been observed that the performance of the GAIN model is superior to other models in regions with high variation. In regions with low variation, it was observed that KNN and RF models gave better outputs.

The results indicated that Random Forests exhibited superior performance in most cases, followed by KNN and GAIN.

6.2 Future Work

The following studies can be done in the future for the imputation of daily average temperature and total precipitation data.

- Data with higher missing percentages can be created, and selected models can be applied to these data.
- More reference stations can be selected for the study.
- By creating a block of missing values for the target stations in each region, the selected models can impute these data, and the performances of these models can be compared.
- While generating the missing values in the form of blocks, the first parts, middle parts, or the last parts of the data can be removed from the data. Imputation models can be applied to these three different data and compared.
- The study can be extended to other meteorological variables.

REFERENCES

[1] Lahiff, M., Little, R. J. A., & Rubin, D. B. (1989b). Statistical Analysis With Missing Data. Journal of the American Statistical Association, 84(405), 332. https://doi.org/10.2307/2289883

[2] Ramos-Calzado, P., Gómez-Camacho, J., Pérez-Bernal, F., & Pita-López, M. F. (2008). A novel approach to precipitation series completion in climatological datasets: application to Andalusia. International Journal of Climatology, 28(11), 1525–1534. https://doi.org/10.1002/joc.1657

[3] Aieb, A., Madani, K., Scarpa, M., Bonaccorso, B. & Lefsih, K., 2019, 'A new approach for processing climate missing databases applied to daily rainfall data in Soummam watershed,Algeria', Heliyon 5(2019), e01247. https://doi.org/10.1016/j.heliyon.2019.e01247

[4] Kiani, K., & Saleem, K. (2017b). K-Nearest Temperature Trends. https://doi.org/10.1145/3077584.3077592

[5] Jadhav, A. G., Pramod, D., & Ramanathan, K. (2019b). Comparison of Performance of Data Imputation Methods for Numeric Dataset. Applied Artificial Intelligence, 33(10), 913–933. https://doi.org/10.1080/08839514.2019.1637138

[6] Thirumahal, R., & Patil, D. (2014b). KNN and ARL Based Imputation to Estimate Missing Values. Indonesian Journal of Electrical Engineering and Informatics, 2(3). https://doi.org/10.11591/ijeei.v2i3.117

[7] Jerez, J. M., Molina, I. J., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., & Franco, L. (2010b). Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artificial Intelligence in Medicine, 50(2), 105–115. https://doi.org/10.1016/j.artmed.2010.05.002

[8] Mital, U., Dwivedi, D., Brown, J. B., Faybishenko, B., Painter, S. L., & Steefel, C. I., 2020, 'Sequential Imputation of Missing Spatio-Temporal Precipitation Data Using Random Forests' Front. Water, 2, 2624-9375, https://doi.org/10.3389/frwa.2020.00020

[9] Addi, M., Gyasi-Agyei, Y., Obuobie, E., & Amekudzi, L. K. (2022). Evaluation of imputation techniques for infilling missing daily rainfall records on river basins in Ghana. Hydrological Sciences Journal-journal Des Sciences Hydrologiques, 67(4), 613–627. https://doi.org/10.1080/02626667.2022.2030868

[10] Dixneuf, P. (2021, August 21). A computational study on imputation methods for missing environmental data. arXiv.org. https://arxiv.org/abs/2108.09500

[11] Pantanowitz, A., & Marwala, T. (2009b). Evaluating the Impact of Missing Data Imputation. In Lecture Notes in Computer Science (pp. 577–586). Springer Science+Business Media. https://doi.org/10.1007/978-3-642-03348-3_59

[12] Shah, A. M., Bartlett, J. W., Carpenter, J. R., Nicholas, O., & Hemingway, H. (2014b). Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. American Journal of Epidemiology, 179(6), 764–774. https://doi.org/10.1093/aje/kwt312

[13] Tang, F., & Ishwaran, H. (2017). Random forest missing data algorithms. Statistical Analysis and Data Mining, 10(6), 363–377. https://doi.org/10.1002/sam.11348

[14] Kokla, M., Virtanen, J. K., Kolehmainen, M., Paananen, J., & Hanhineva, K. (2019). Random forest-based imputation outperforms other methods for imputing LC-MS metabolomics data: a comparative study. BMC Bioinformatics, 20(1). https://doi.org/10.1186/s12859-019-3110-0

[15] Jahan, F., Sinha, N. C., Rahman, M., Rahman, M., Mondal, M. S. H., & Islam, M. S. (2019). Comparison of missing value estimation techniques in rainfall data of Bangladesh. Theoretical and Applied Climatology, 136(3–4), 1115–1131. https://doi.org/10.1007/s00704-018-2537-y

[16] Rahman, N. H. A., Deni, S. M., & Ramli, N. M. (2017b). Generalized linear model for estimation of missing daily rainfall data. https://doi.org/10.1063/1.4981003

[17] Sattari, M. T., Rezazadeh-Joudi, A., & Kusiak, A. (2017b). Assessment of different methods for estimation of missing data in precipitation studies. Hydrology Research, 48(4), 1032–1044. https://doi.org/10.2166/nh.2016.364

[18] Sanusi, W., Zin, W. Z. W., Mulbar, U., Danial, M., & Side, S. (2017). Comparison of the Methods to Estimate Missing Values in Monthly Precipitation Data. International Journal on Advanced Science, Engineering and Information Technology. https://doi.org/10.18517/ijaseit.7.6.2637

[19] Aguilera, H., Guardiola-Albert, C., & Serrano-Hidalgo, C. (2020b). Estimating extremely large amounts of missing precipitation data. Journal of Hydroinformatics, 22(3), 578–592. https://doi.org/10.2166/hydro.2020.127

[20] Norazizi, N., & Deni, S. M. (2019b). Comparison of Artificial Neural Network (ANN) and Other Imputation Methods in Estimating Missing Rainfall Data at Kuantan Station. In Communications in computer and information science (pp. 298–306). Springer Science+Business Media. https://doi.org/10.1007/978-981-15-0399-3_24

[21] Abdullah, A. S., Bhuian, M. H., Kiselev, G., Dewan, A., Hasan, Q. M., & Rafiuddin, M. (2021b). Extreme temperature and rainfall events in Bangladesh: A comparison between coastal and inland areas. International Journal of Climatology, 42(6), 3253–3273. https://doi.org/10.1002/joc.6911

[22] De Carvalho, J. R. M., Monteiro, J., Nakai, A. M., & Assad, E. D. (2017b). Model for Multiple Imputation to Estimate Daily Rainfall Data and Filling of Faults. Revista Brasileira De Meteorologia, 32(4), 575–583. https://doi.org/10.1590/0102-7786324006

[23] Turrado, C. C., Del Carmen Meizoso Lopez, M., Las-Heras, F., Gómez, B., Rolle, J. L. C., & De Cos Juez, F. J. (2014b). Missing Data Imputation of Solar Radiation Data under Different Atmospheric Conditions. Sensors, 14(11), 20382–20399. https://doi.org/10.3390/s141120382

[24] Wesonga, R. (2015b). On multivariate imputation and forecasting of decadal wind speed missing data. SpringerPlus, 4(1). https://doi.org/10.1186/s40064-014-0774-9

[25] Diouf, S., Deme, E. H. B., & Deme, A. (2022b). Imputation methods for missing values: the case of Senegalese meteorological data. African Journal of Applied Statistics, 9(1), 1245–1278. https://doi.org/10.16929/ajas/2022.1245.267

[26] Popolizio, M., Amato, A., Politi, T., Calienno, R., & Di Lecce, V. (2021). Missing data imputation in meteorological datasets with the GAIN method. In 2021 IEEE International Workshop on Metrology for Industry 4.0 & IoT (MetroInd4.0&IoT). https://doi.org/10.1109/metroind4.0iot51437.2021.9488451

[27] Low, R., Tekler, Z. D., & Cheah, L. (2020b). Predicting Commercial Vehicle Parking Duration using Generative Adversarial Multiple Imputation Networks. Transportation Research Record, 2674(9), 820–831. https://doi.org/10.1177/0361198120932166

[28] Dong, W., Fong, D. Y. T., Yoon, J., Wan, E. Y. F., Bedford, L., Tang, E., & Lam, C. L. K. (2021b). Generative adversarial networks for imputing missing data for big data clinical research. BMC Medical Research Methodology, 21(1). https://doi.org/10.1186/s12874-021-01272-3

[29] Wang, H., Chen, Y., Shen, B., Wu, D., & Ban, X. (2018b). Generative Adversarial Networks Imputation for High Rate Missing Values. https://doi.org/10.1109/cybermatics_2018.2018.00121

[30] Jiang, H., Wan, C., Yang, K., Ding, Y., & Xue, S. (2021b). Continuous missing data imputation with incomplete dataset by generative adversarial networks–based unsupervised learning for long-term bridge health monitoring. Structural Health Monitoring-an International Journal, 21(3), 1093–1109. https://doi.org/10.1177/14759217211021942

[31] Şahin, S., & Cigizoglu, H. K. (2010). Homogeneity analysis of Turkish meteorological data set. Hydrological Processes, 24(8), 981–992. https://doi.org/10.1002/hyp.7534

[32] Yozgatligil, C., Aslan, S., Iyigun, C., & Batmaz, İ. (2013b). Comparison of missing value imputation methods in time series: the case of Turkish meteorological data. Theoretical and Applied Climatology, 112(1–2), 143–167. https://doi.org/10.1007/s00704-012-0723-x

[33] Dikbaş, F. (2016). Frequency based imputation of precipitation. Stochastic Environmental Research and Risk Assessment, 31(9), 2415–2434. https://doi.org/10.1007/s00477-016-1356-x

[34] Kalkan, Ö. K., Kara, Y., & Kelecioğlu, H. (2018). Evaluating Performance of Missing Data Imputation Methods in IRT Analyses. International Journal of Assessment Tools in Education, 403–416. https://doi.org/10.21449/ijate.430720

[35] Katipoğlu, O. M., & Acar, R. (2021). Eksik sıcaklık verilerinin Yapay Sinir Ağları (YSA) ile tahmin edilmesi. Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi. https://doi.org/10.24012/dumf.852821

[36] Başakın, E. E., Ekmekcioğlu, Ö., & Özger, M. (2023). Developing a novel approach for missing data imputation of solar radiation: A hybrid differential evolution algorithm based eXtreme gradient boosting model. Energy Conversion and Management, 280, 116780. https://doi.org/10.1016/j.enconman.2023.116780

[37] Rubin, D. B. (1976). Inference and missing data. Biometrika, 63(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

[38] Troyanskaya, O. G., Cantor, M. N., Sherlock, G., Brown, P. O., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics, 17(6), 520–525. https://doi.org/10.1093/bioinformatics/17.6.520

[39] Decision Tree Algorithm in Machine Learning - Javatpoint. (n.d.). www.javatpoint.com. https://www.javatpoint.com/machine-learning-decision-tree-classification-algorithm

[40] What is a Random Forest? (n.d.). TIBCO Software. https://www.tibco.com/reference-center/what-is-a-random-forest

[41] Campion, W. J., & Rubin, D. B. (1989). Multiple Imputation for Nonresponse in Surveys. Journal of Marketing Research, 26(4), 485. https://doi.org/10.2307/3172772

[42] Yoon, J. (2018, June 7). GAIN: Missing Data Imputation using Generative Adversarial Nets. arXiv.org. https://arxiv.org/abs/1806.02920

[43] L. Biewald, “Experiment Tracking with Weights and Biases,” Weights & Biases. [Online]. Available: http://wandb.com/.

[44] Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology, 10(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

APPENDICES

A. Weights and Biases

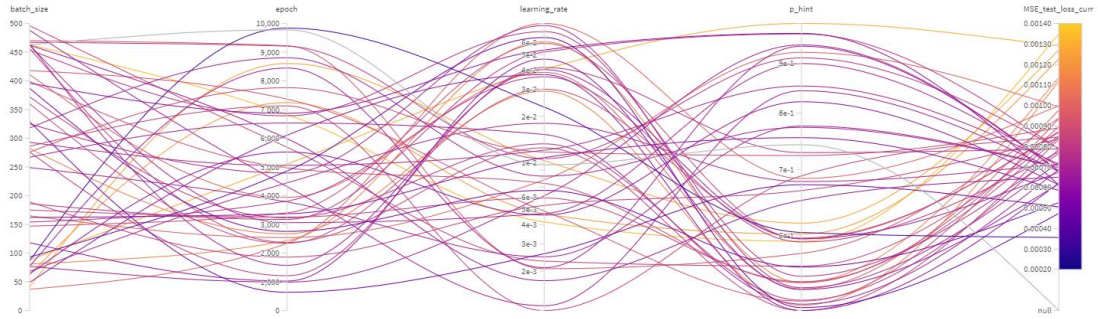

Figure A-1. Parameter tuning graph of MR region temperature data with 30% missing value

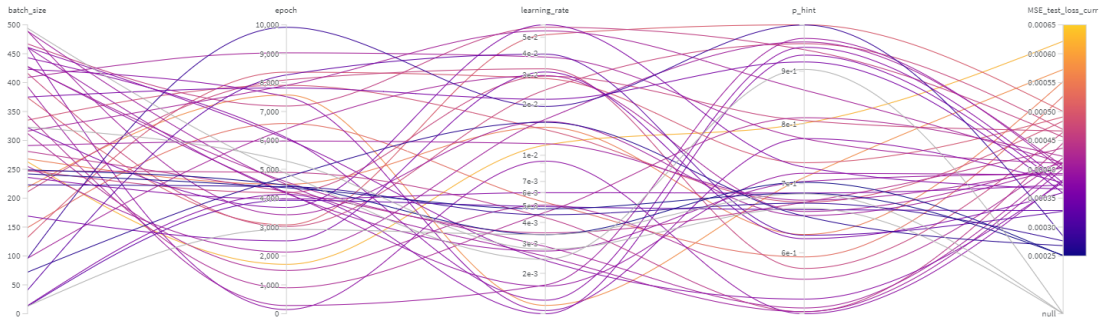

Figure A-2. Parameter tuning graph of AR region temperature data with 30% missing value

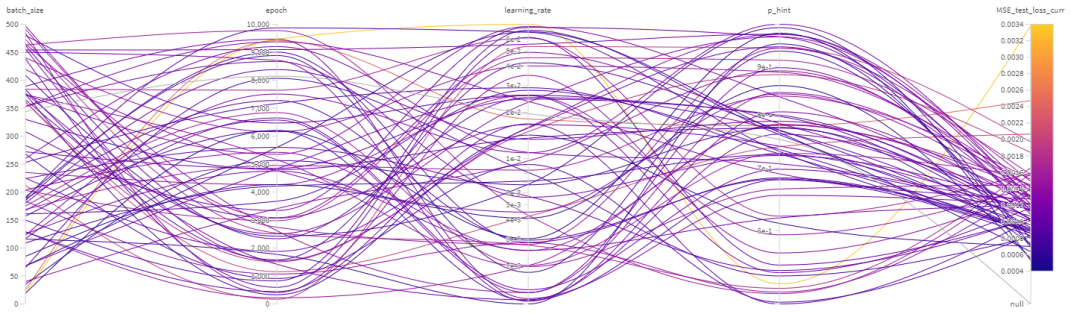

Figure A-3. Parameter tuning graph of EAR region temperature data with 30% missing value

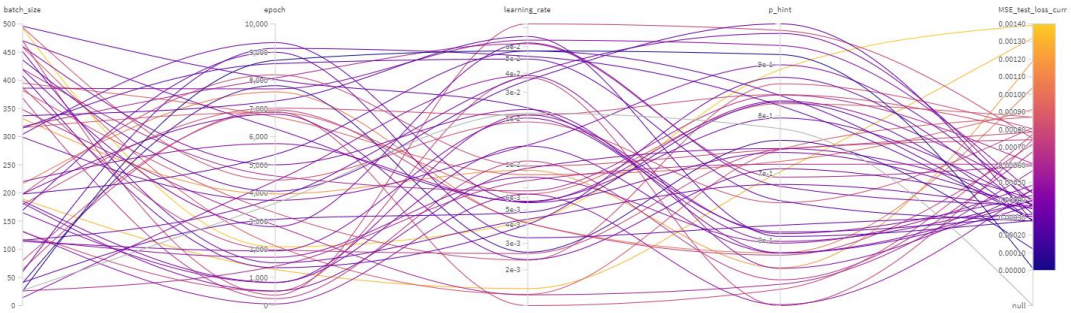

Figure A-4. Parameter tuning graph of CAR region temperature data with 30% missing value

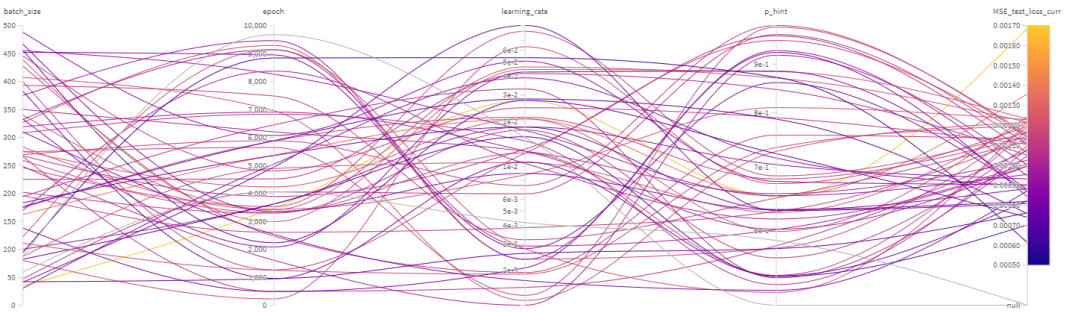

Figure A-5. Parameter tuning graph of SAR region temperature data with 30% missing value

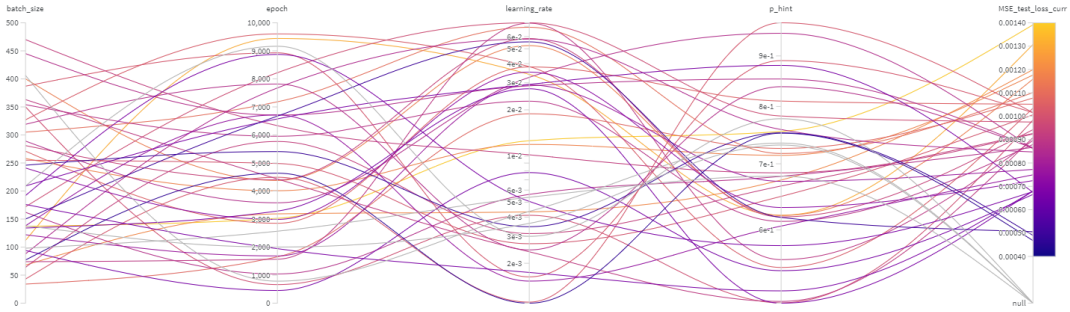

Figure A-6. Parameter tuning graph of BSR region temperature data with 30% missing value

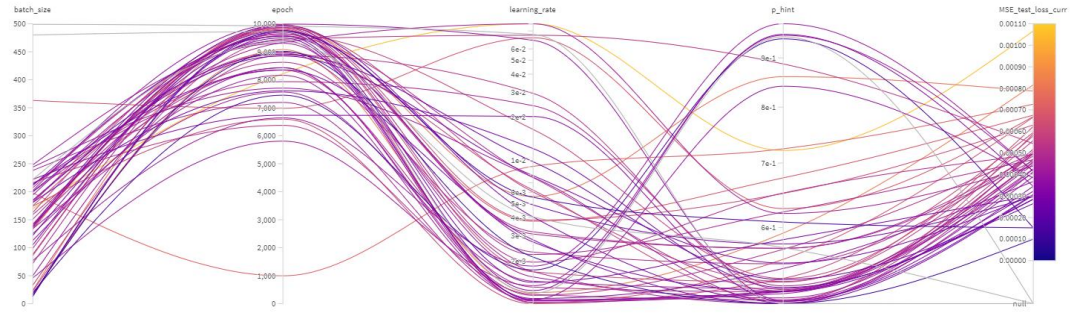

Figure A-7. Parameter tuning graph of MAR region temperature data with 30% missing value

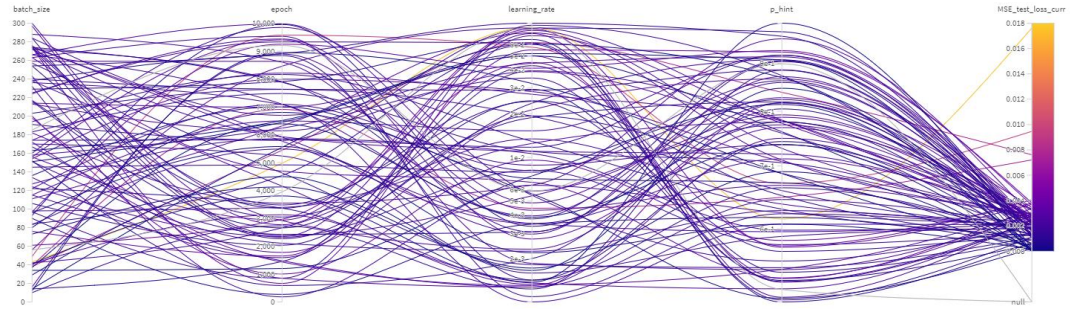

Figure A-8. Parameter tuning graph of MR region precipitation data with 30% missing value

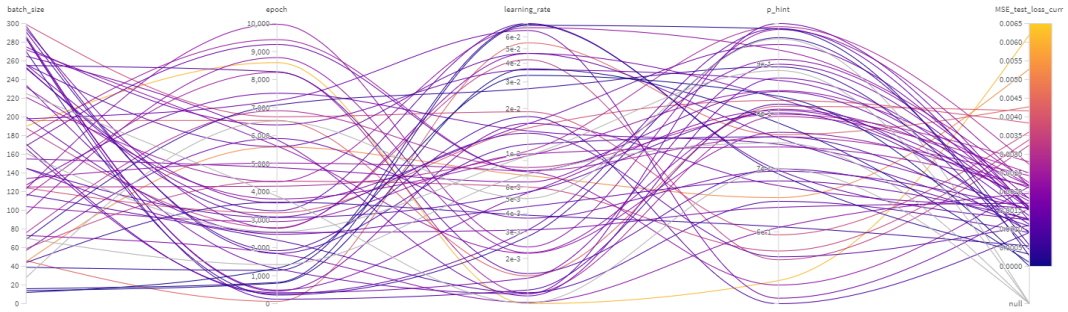

Figure A-9. Parameter tuning graph of AR region precipitation data with 30% missing value

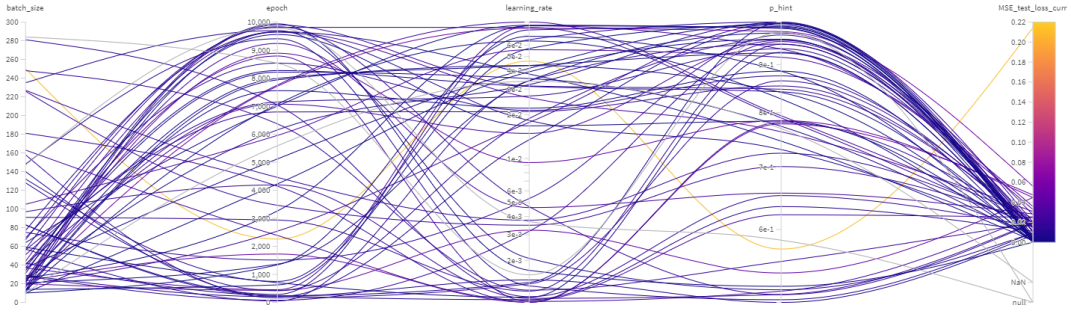

Figure A-10. Parameter tuning graph of EAR region precipitation data with 30% missing value

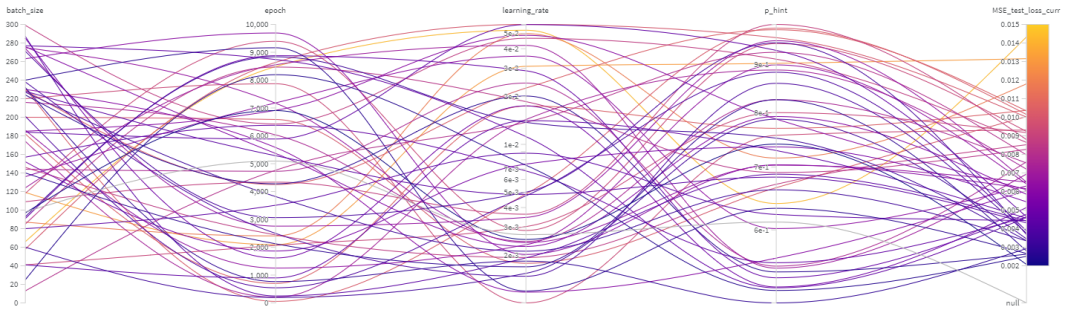

Figure A-11. Parameter tuning graph of CAR region precipitation data with 30% missing value

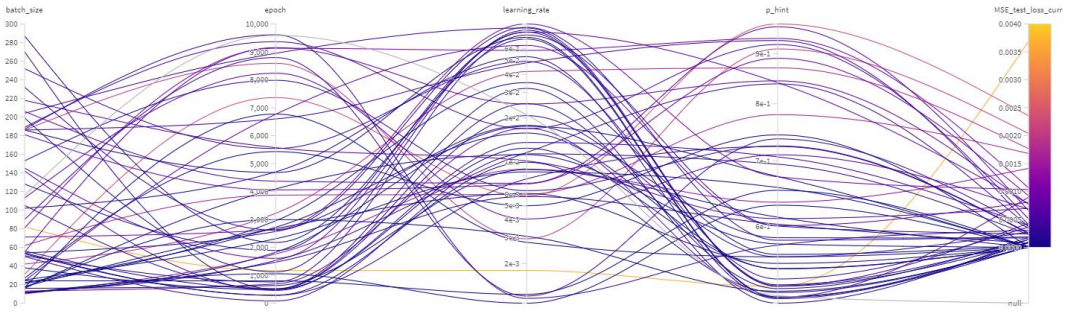

Figure A-12. Parameter tuning graph of SAR region precipitation data with 30% missing value

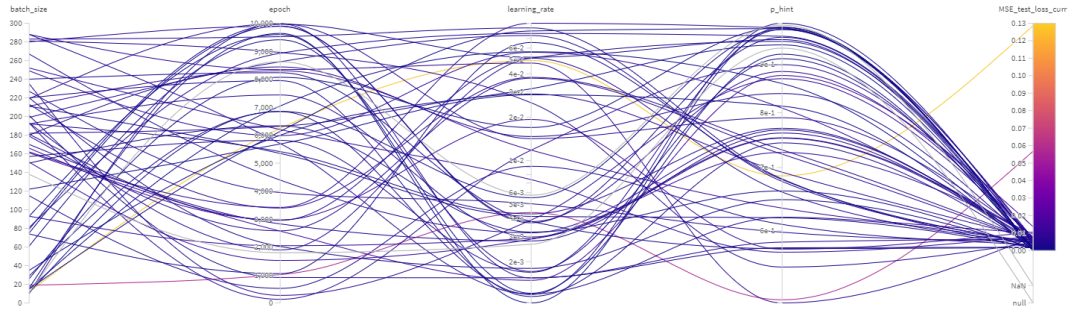

Figure A-13. Parameter tuning graph of BSR region precipitation data with 30% missing value

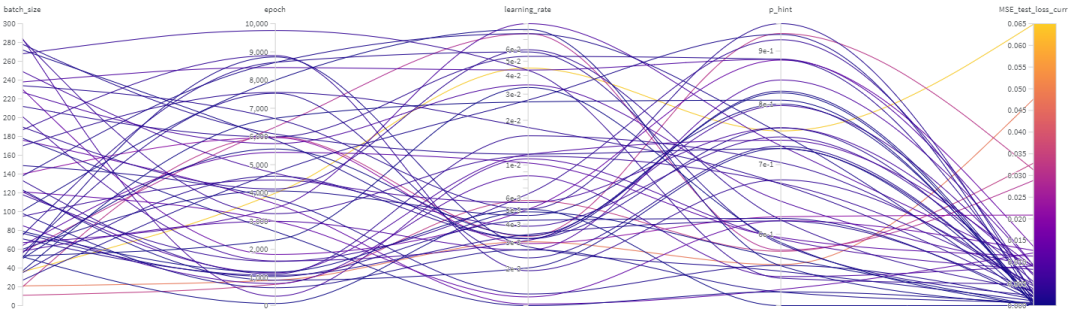

Figure A-14. Parameter tuning graph of MAR region precipitation data with 30% missing value

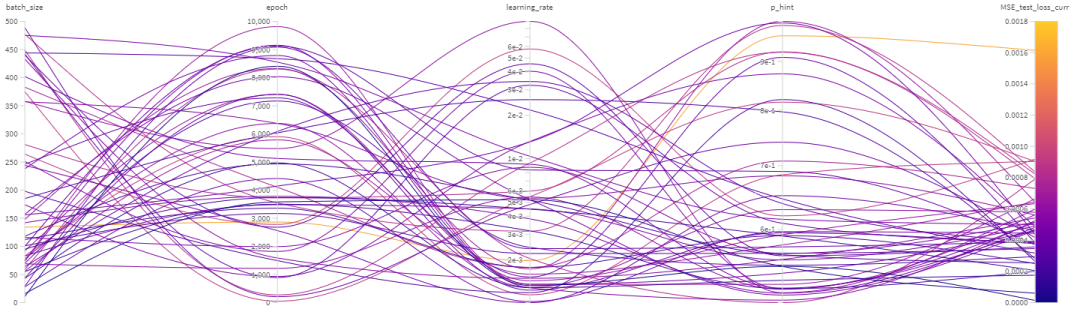

Figure A-15. Parameter tuning graph of AR region temperature data with block missing value

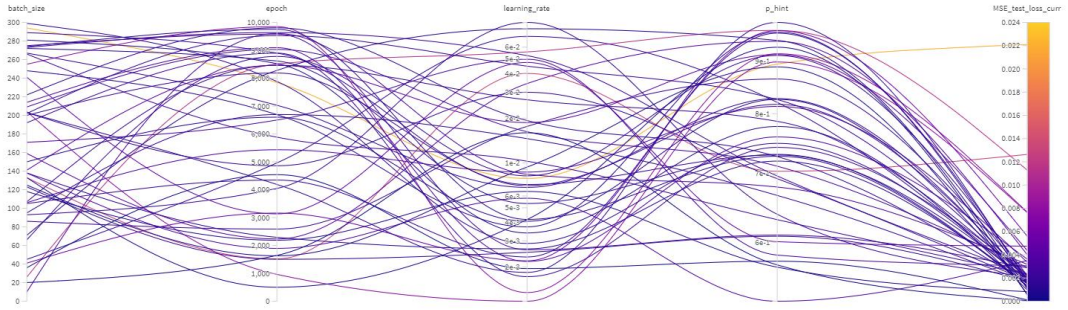

Figure A-16. Parameter tuning graph of AR region precipitation data with block missing value