# COMPARISON OF MISSING VALUE IMPUTATION METHODS FOR METEOROLOGICAL TIME SERIES DATA

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

BY

SİPAN ASLAN

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

SEPTEMBER 2010

Approval of the thesis:

**COMPARISON OF MISSING VALUE IMPUTATION METHODS FOR METEOROLOGICAL TIME SERIES DATA**

submitted by **SİPAN ASLAN** in partial fulfillment of the requirements for the degree of **Master of Science in Statistics Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences** ____________________

Prof. Dr. Öztaş H. Ayhan
Head of Department, **Department of Statistics** ____________________

Assist.Prof.Dr. Ceylan Yozgatlıgil
Supervisor, **Department of Statistics, METU** ____________________

**Examining Committee Members:**

Assoc. Prof.Dr. İnci Batmaz
Department of Statistics, METU ____________________

Assist.Prof.Dr. Ceylan Yozgatlıgil
Department of Statistics, METU ____________________

Assist.Prof.Dr. Zeynep Işıl Kalaylıoğlu
Department of Statistics, METU ____________________

Assist.Prof.Dr. Berna Burçak Başbuğ Erkan
Department of Statistics, METU ____________________

Assist.Prof.Dr. Kasırga Yıldırak
Department of Economics, Trakya University ____________________

**Date: 17.09.2010**

**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: Sipan ASLAN

Signature :

# ABSTRACT

## COMPARISON OF MISSING VALUE IMPUTATION METHODS FOR METEOROLOGICAL TIME SERIES DATA

Aslan, Sipan
M.Sc., Department of Statistics
Supervisor: Assist.Prof.Dr. Ceylan Yozgatlıgil

September 2010, 86 pages

Dealing with missing data in spatio-temporal time series constitutes important branch of general missing data problem. Since the statistical properties of time-dependent data characterized by sequentiality of observations then any interruption of consecutiveness in time series will cause severe problems. In order to make reliable analyses in this case missing data must be handled cautiously without disturbing the series statistical properties, mainly as temporal and spatial dependencies.

In this study we aimed to compare several imputation methods for the appropriate completion of missing values of the spatio-temporal meteorological time series. For this purpose, several missing imputation methods are assessed on their imputation performances for artificially created missing data in monthly total precipitation and monthly mean temperature series which are obtained from the climate stations of Turkish State Meteorological Service. Artificially created missing data are estimated by using six methods. Single Arithmetic Average (SAA), Normal Ratio (NR) and NR

Weighted with Correlations (NRWC) are the three simple methods used in the study. On the other hand, we used two computational intensive methods for missing data imputation which are called Multi Layer Perceptron type Neural Network (MLPNN) and Monte Carlo Markov Chain based on Expectation-Maximization Algorithm (EM-MCMC). In addition to these, we propose a modification in the EM-MCMC method in which results of simple imputation methods are used as auxiliary variables. Beside the using accuracy measure based on squared errors we proposed Correlation Dimension (CD) technique for appropriate evaluation of imputation performances which is also important subject of Nonlinear Dynamic Time Series Analysis.

**Keywords:** Correlation Dimension, Expectation Maximization (EM) Algorithm, Markov Chain Monte Carlo (MCMC), Meteorological Time Series, Multiple Imputation, Nonlinear Dynamic Time Series.

# ÖZ

## METEOROLOJİK ZAMAN SERİSİ VERİLERİNDE KAYIP VERİ ATAMA YÖNTEMLERİNİN KARŞILAŞTIRILMASI

Aslan, Sipan
Yüksek Lisans, İstatistik Bölümü
Tez Yöneticisi: Yrd.Doç.Dr. Ceylan Yozgatlıgil

Eylül 2010, 86 sayfa

Alansal ve zamansal özellikler taşıyan zaman serilerinde karşılaşılan kayıp verileri incelemek kayıp veri probleminin önemli bir branşını oluşturmaktadır. Zamana bağlı verilerin istatistiksel özellikleri gözlemlerin ardışıklığıyla karakterize edilebildiğinden zaman dizisi gözlemlerinde ki ardışıklığın herhangi bir nedenden ötürü kesintiye uğraması çözümü zor problemlere neden olmaktadır. Bu tür durumlarda güvenilir analizler yapabilmek için kayıp verileri, serilerin içermiş olduğu alansal ve zamansal özellikleri dikkate alarak değerlendirmek gerekmektedir.

Biz bu çalışmada, alansal ve zamansal özellikte ki meteorolojik zaman serilerinde karşılaşılan kayıpların uygun şekilde tamamlanması için bazı yöntemleri karşılaştırmayı amaçladık. Bu amaçla, Devlet Meteoroloji Genel Müdürlüğüne ait klimatoloji istasyonlarından elde edilen aylık toplam yağış ve aylık ortalama sıcaklık serilerinde yapay olarak oluşturulan kayıp verileri kullanarak yöntemlerin kayıp tahmin başarımlarını değerlendirdik. Yapay olarak oluşturulan kayıp veriler altı yöntemle tahmin edildi. Çalışmada basit yöntemler olarak adlandırabileceğimiz Basit Artimetik

Ortalama (BAO), Normal Oran (NO), ve Korelasyon Ağırlıklı Normal Oran (KANO) yöntemleri kullanılmıştır. Ayrıca hesap yoğun yöntemlerden Çok Katmanlı Yapay Sinir Ağları (ÇKYSA) Modeli ve Beklenti Maksimizasyonu tabanlı Monte Karlo Markov Zinciri (BM-MCM) algortiması kullanılmıştır. Yukarıdaki yöntemlere ek olarak basit yöntemlerden elde edilen sonuçların yardımcı değişken olarak kullanıldığı modifiye edilmiş BM-MCMC algoritmasını önerdik. Yöntemlerin uygun biçimde karşılaştırılması için hata kareler hesabına dayalı doğruluk ölçümü yanında Doğrusal Olmayan Zaman Serileri Analizinin önemli bir konusu olan Korelasyon Boyutu tekniğini kullandık.

**Anahtar kelimeler:** Korelasyon Boyutu, Beklenti Maksimizasyonu (BM) Algoritması, Markov Zinciri Monte Karlo (MZMK), Meteorolojik Zaman Serileri, Çoklu Atama, Doğrusal Olmayan Dinamik Zaman Serileri.

*To the Monument of the Unknown Student and for all Gangways!*

# ACKNOWLEDGEMENTS

I am heartily thankful to my supervisor, Assist.Prof. Ceylan Yozgatlıgil, whose encouragement, guidance, patience and enthusiasm from the initial to the final level of my master study enabled me to write this thesis and without her perpetual positive energy and persistent help, this thesis would not have been completed. I sincerely appreciate her for giving me the opportunity to work with her.

I wish to express my sincere gratitude to Assoc.Prof. İnci Batmaz, Assist.Prof. Cem İyigün and NINLIL research group for their insightful criticisms, and patient encouragement during my study. Every meeting with them made a deep impression on me. I consider myself fortunate to have worked with NINLIL research group in which I have involved.

I would like to express gratitude to my committee members Assist.Prof. Zeynep Işıl Kalaylıoğlu, Assist.Prof. Berna Burçak Başbuğ Erkan and Assist.Prof. Kasırga Yıldırak for their advice, support and for the time they spent in reviewing this thesis. Their academic persona, sage advice and words of encouragement kept me motivated.

I would like to thank all people who have helped and inspired me during my master study. I would like to thank staff and students in the department of statistics at METU; it has been most valuable and positive learning experience for me.

Last, but certainly not least, I would like to thank my family for their unconditional support.

## TABLE OF CONTENTS

CHAPTERS

# LIST OF TABLES

## TABLES

## LIST OF FIGURES

FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 Overview

There is one indisputable scientific law can be mentioned only, which is that the collections of observations from any process, regardless that the observations are obtained from social, business, atmospheric, psychology, education, health or archeological sciences, always contain more or less missing information. Missing data are widespread problem in any discipline. Since the scientific activity aims to explore the uncertainty, all scientific or especially statistical problems can be formulated as a missingness problem (Little and Rubin, 2002), if we called the uncertainty as a missing data problem (Tan et al., 2010).

Amount of the missing information that we have has restrict our inferences in an unfavorable way. Covering the missing information and decreasing the effect of missingness in such an appropriate way constitutes well known problem of scientific researches and huge number of research has been done for over three decades.

Another inevitable circumstance is encountering with the missing data on time domain observations. Actually, we cannot prevent the missingness in time series observations because we often deal with continuous processes and observing the continuity itself is impossible. However, what we do in time series observation is eliminating the problem by equally spaced observations. Key point in time series analysis is the sequentiality. Statistical information of time series is included in serial properties of observations. If the sequentiality is interrupted by any reason we faced with the problem of missing data that we must overcome somehow. Understanding the temporal and/or spatial behavior of any observable system, which is governed by deterministic or mostly stochastic principles, is highly related with the available observations of the system that is to be analyzed. Therefore, complete data and large sample size are also indecisive position for time series data based analysis and statistical inferences such as forecasting and time series clustering.

Within this study, the only reconstruction of the missing parts of the any environmental incomplete time series over a certain period is under investigation. Since the environmental time series carry out the temporal and spatial characteristics together, reconstruction and assessment of missing period should be done carefully without disturbing the series statistical properties, mainly as temporal and spatial dependencies. In this manner, appropriate missing data estimation is crucial for making the reliable inferences and improving the forecasting performance of the statistical analysis.

The investigation of adequate and efficient missing data estimation method is related with missingness mechanism which leads to missing data (Little and Rubin, 1987). Assumption that the information of missing part carried out by available observations is very useful in order to estimate missing data. Once we have decided to assume that the missing information can be covered from available cases of interested phenomena, the missing data imputation will be simplified. In this context, determining the underlying missingness mechanism of incomplete data and nature of missing data are important problems which have to be handled cautiously.

In general, handling missing data methods can be divided in two classes. The first one is deleting approaches and the other one is imputation strategies. Deleting procedures sometimes can be appropriate in large sample size and if the observations are independent of time domain. Deleting the missing data will cause a little bias in case of large samples and for a time series data autocorrelation. Imputation strategies including estimation of the missing data itself and the estimation of the parameters. Parameter estimation can only be sufficient in some missing data problems. However in some cases, the estimation of missing data is objective. The basic principle for imputation methods is to replace missing value with plausible one and carry out the analysis as if there were no missing data. If we deal with the estimation of missing data, single stochastic imputation or multiple stochastic imputation procedures can be concerned (McKnight et al., 2007).

## 1.2 Literature Review

Despite the long standing of missing data problem, dealing with the problem in statistical approach is based on recent works after the 1970s. Before the statistical approach to the missing data, problems were solved by simple methods mostly by deleting schemes or ad-hoc methods and using these approaches poses bias results in estimations under missing data (Marwala, 2009).

Missing data handling methods can be considered in two categories through published works. The first category contains single approaches and generally called traditional methods such as case deletion algorithms, hot-deck methods, mean imputation and etc (Little and Rubin, 2002). The second category contains more complex approaches and mostly based on parameter estimation and inferential methods under incomplete data. Model based estimations and prediction of missing data methods still under investigation and too many approaches have been used in the literature. Increase in the number of studies mainly caused by developments in computer science. Since the missing data problems vary with different disciplines, the corresponding handling approaches are also vary in this manner. So far there is no single best method or methods emerged in the literature to deal with missing data and for the estimation of missing values. The structure of missing data itself is more important to determine the best method for handling missingness. Thereby, we can say that the most of the studies in missing data literature relies on comparative features and most of them aimed to determine the best approach to specific missingness problem.

A huge number of methodological studies on missing data problem focused on maximum likelihood estimation and multiple imputations. Preliminary conceptual study on incomplete data is published by Dempster, Laird and Rubin in 1977. The methodology is based on likelihood estimation.  For computing the maximum likelihood estimates from incomplete data, firstly they used and explained the Expectation-Maximization (EM) algorithm. Following this pioneer study, a plenty of applications of likelihood based studies have been published in the literature.

In likelihood based missing data analysis literature, the first studies on missing data handling methods primarily concerned with incomplete survey data problems and focuses on identifying missing data and prevent potential biases. (Allison, 2001) Therefore, main aim in the most of these researches is making reliable parameter estimation and reliable inferences under incomplete data rather than missing value imputation. In this context, probability model based procedures can be considered under the general class of data augmentation (DA). DA is a Bayesian methodology. DA as Bayesian estimation aims to find posterior density of parameter(s) given the prior information about the complete data and then emphasizes on sampling from the posterior distribution (Tan et al., 2010). Generally, in these studies imputation of any missing score/point itself carry out to secondary purpose as well as in data augmentation. More detailed studies on extensive evaluation of methods for statistically handling missing data studies of Little and Rubin, (2002); Schafer, (1997); MacLachlan and Krishnan, (1997); McKnight et al., (2007); Tanner and Wong, (1987); Tan et al., (2010) and  Allison, (2001) provide substantial sources of information.

Moreover, another important modern missing data handling approach is the Multiple Imputation (MI) class algorithms. Recent developments on Single and Multiple Imputation methods that are alternative to likelihood based approaches recommended by researchers and attract many more researchers in different areas into missing data problem (Schafer and Graham, 2002). Generally, imputation approaches deal with the construction of complete data sets by replacing the missing data with plausible value or values throughout the simple methods or more complex likelihood estimation based imputation methods. Another key point that we have to say is likelihood based imputation methods relies on Bayesian estimation principles (Enders, 2010).  This valuable progress on ever increasing researches is provided by computational opportunities that are supplied by the softwares and the computer technology.

As previously stated the most of the studies in the literature have comparative features and in fact, it has to be comparative because of the different missing data structures. For instance, a method that can give the best results for a specific study while in the other study, the same method can give bad results and when we look at the literature we often face with this situation. Aikl and Zanuiddin (2008) had been conducted a comparative study on time series data and non-time series data. They used Local Least Square Estimation, Bayesian Principle Analysis and Radial Basis Neural Network imputation approaches. They compared the methods in terms of accuracy measure and within different missing value proportions which are artificially created. Bayesian Principle Analysis is performed best in this study. Musil et al. (2002) modeled cross sectional data set on stress and health of older adults. Simple and complex approaches are compared. They state that regression and EM algorithm performed best. Lin (2008) in his study compared the performance of multiple imputations with EM algorithm and Monte Carlo MCMC method. He reported that according to results methods are indistinctive. Nelwamondo et al. (2007) compared the EM algorithm and Genetic algorithm (GA) Neural Networks on industrial power plant data, industrial winding process and HIV prevalence survey data. Results show that the EM algorithm is convenient and performs better especially when there is no interdependency between the input variables. Nelwamondo et al. (2009) modeled same HIV survey data again and they proposed a new approach which is the combination of neural networks and GA algorithm. They report that the proposed method significantly outperforms a similar method. Li and Parker (2008) used fuzzy Adaptive Resonance Theory (ART) neural network approach for estimating missing observations in spatio-temporal wireless sensor network data. They compared results with EM algorithm and simple methods.

There can be found numerous studies in the literature for dealing with the different subjects of areas. For example, Sehgal et al. (2005), Choong et al. (2009), Tuikkala et al. (2008) and Hu et al. (2006) had worked on micro array data and they used and compared various methods such as Bayesian Principle Component Analysis (BPCA), K-nearest neighbor (KNN) and Least Square Imputation for handling missing data problem on micro array data. Results of the studies does not indicate a single method that outperforms to other approaches.

When we look at the situation on time series imputation studies, we have seen that there is no one method adopted by everyone, yet. Approaches are still under investigation and studies preserve their comparative characteristics. Plaia and Bondi (2006) proposed a new approach that Side Dependent Effects Method (SDEM) for longitudinal air pollution data and compared its performance to other single and multiple imputation methods. It was shown that the results are comparable. Weerasinghe (2009) analyzed pollution data set with SARIMA, Kalman gain and EM algorithm. Hopke et al. (2001) developed three statistical methods which are the Cross-Sectional Model (Multivariate Normal Model based Multiple Imputation), the Multivariate Integrated Moving Average (IMA) and the Multivariate Seasonal Integrated Moving Average (SIMA) for multivariate air pollution time series data. They report that SIMA model gives more realistic imputations that supported by cross-validation checks. Chiewchanwattana et al. (2007) proposed a method on similarity measure. The proposed algorithm is based on the univariate time-series data. The imputation of missing observations are obtained by using sub samples which are similar to the missing parts of data and imputing the missing samples obtained by these sub samples. Researchers note that the proposed algorithm has its drawbacks because it assumes that the data are sufficiently oversampled and they have periodical characteristics. Another comprehensive comparison of imputation methods were done by Moffat et al. (2007). Study compared 15 different methodology including MI and ANN methods on net ecosystem CO2 exchange (NEE) in eddy covariance time series. They reported that the performance of reviewed methods varies due to different artificial gap scenarios but in addition they noted that the artificial neural network (ANN) model based algorithms performs well.

Dastorani et al. (2010) also reported that the ANN performs well for estimation of missing flow observations compared with simple methods. Chiewchanwattana and Lursinsap (2002) and Figueroa Garcia et al. (2008) propose neural network based algorithms as modern time series imputation methods in their studies.

Data loss from observations obtained by meteorological stations is almost inevitable. Studies on missing observation estimation of meteorological time series data far back as 1950s. In the first studies missing values were estimated by simple approaches such as mean imputation or simple linear regression estimates. For instance, Paulhus and Kohler (1952) used average of concurrent observations at nearby stations for imputation of missing precipitation records. The idea of using the information of nearby stations for missing estimation  is still a valid approach because of the spatial properties that meteorological variables have.  Applications using nearby stations’ information  can be considered as a kind of interpolation algorithms. For the purpose of interpolation/estimation, observations at the adjacent observatories serve as predictors and the target station that has missing observations represent the predictands.

Young (1992) compared three different interpolation methods for missing precipitation records. Used methods in the study were Multiple Discriminant Analysis (MDA), performed well in the study, Multiple  Linear Regression and Normal Ratio method. Makhuvha et al. (1997)  compared six different methods for estimating missing rainfall data and reported that the EM algorithm gives most accurate results. Xia et al. (1999a,b) reviewed six different approaches to estimating missing climatological data (daily maximum temperature, minimum temperature, air temperature, water vapor pressure, wind speed and precipitation). The Multiple Linear Regression approach (using the five closest station as predictor) gave best estimates as well as reportedly. Eischeid et al. (2000) used simple approaches to create complete national daily time series of temperature and precipitation for the Western United States and resultant estimates show no systematic bias statistically. Another comparative study on simple methods for completion of precipitation series has been done by Teegavarapu and Chandramouli (2005). Commonly used approach Inverse Distance Weighting Method (IDWM) is reviewed. This method use distance information between target and reference stations to estimate missing data. They made some Improvements on IDWM.

Schneider (2001) in his valuable and most cited study recommended that the Regularized EM algorithm for imputation of incomplete climate data especially when the number of variables in data set typically exceeds sample size. This study can be considered as the first study that extensively reviewed of the EM algorithm for climate data. Another invaluable study that is proposed Multiple Imputation (MI) conducted by Cano and Andreu (2010). They reported that using the appropriate data structure MCMC based MI can be successfully applied on time series data.

The other works that needs to be kept in mind have been done by Junninen et al. (2004); Ramos-Calzado et al. (2008); Smith and Aretxabaleta (2007); Coulibaly and Evora (2007); Lucio et al. (2007); Lo Presti et al. (2008); Aly et al. (2009); Kalteh and Berndtsson (2006) and Kalteh and Hjorth (2009). Besides the EM based MI algorithm and simple missing estimation methods, they have proposed and compared many modern approaches such as Self Organizing Map (SOM) and Neural Network based Models in order to estimate missing observations in climatological data set. In the light of these studies, the use of EM algorithm, MCMC based MI and neural network models seems suitable for the purpose of imputation under incomplete meteorological data.

### 1.3 Missingness Mechanism

As we mentioned before inferential applications under incomplete data should be handled carefully. Before starting any analysis with incomplete data, we have to clarify the nature of missingness mechanism which causes some values being missing. Previously, there was common belief that the mechanism was random but it was really as it was thought? Generally, there are two notions accepted for missingness mechanism by all researchers: **ignorable** and **non-ignorable** missingness mechanism. If the mechanism is ignorable we don’t have to care about it and we can ignore it confidently before missing data analysis but if it is not we have to model the mechanism also as part of the parameter estimation. Identifying the missingness mechanism with a statistical approach is still being a tough problem and so try to develop some diagnostic procedure on missingness mechanism is an important research topic.

Establishing probabilistic relationships between measured variables and missing data emerged with likelihood based analysis of incomplete data (McKnight et al., 2007). Frequently used classification system for missing data proposed first by Rubin (1976) and widely discussed in Schafer (1997); Little and Rubin (1987); Little and Rubin (2002). Rubin (1976) specified three types of assumptions on missingness mechanism: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). MCAR and MAR are in class of ignorable missingness mechanism but MNAR is in class of non-ignorable mechanism. MCAR assumption is generally difficult to meet in reality and it assumes that there is no statistically significant difference between incomplete and complete cases. In other words, the observed data points can only be considered as a simple random sample of the variables you would have to analyze. It assumes that missingness is completely unrelated to the data (Enders, 2010). In this case, there is no impact of missingness affecting on the inferences. Little (1988) proposed a chi-square test for diagnosing MCAR mechanism so called Little's MCAR test. Failure to confirm the assumption of MCAR using statistical tests means that the missing data mechanism is either MAR or MNAR. Unfortunately, it is impossible to determine whether a mechanism is MAR or MNAR. This is an important practical problem of missing data analysis and classified untestable assumption because we do not know the values of the missing scores, we cannot compare the values of those with and without missing data to see if they differ systematically on that variable (Allison, 2001). The most of the missing data handling approaches especially EM algorithm and MI relies on MAR assumption (Schafer, 1997). If we can decide that the mechanism that causes missingness is ignorable in such a way, then assuming the mechanism is MAR seems suitable for further analysis. Conducting the EM algorithm and MCMC based MI under MCAR assumption will be also appropriate, since the mechanism of missingness is ignorable (Schafer, 1997).

For a better explanation of the relationship between measured variables and the probability of missing data we give detailed descriptions of missing data mechanisms in the next three sub-sections.

Let $Y_{com} = \left(Y_{ij}\right), \quad i = 1,2,\ldots,N \quad j = 1,2,\ldots,p$ with size of $N \times p$ be the complete data matrix. By using the word complete data, we want to note that the matrix can be consist of two components, $Y_{obs}$ and $Y_{mis}$ respectively, $Y_{com} = (Y_{obs}, Y_{mis})$ and $f(Y|\theta) \equiv f(Y_{obs}, Y_{mis}|\theta)$ denote the density of $Y_{com}$ (Little and Rubin , 1987). The marginal density of $Y_{obs}$ is obtained by;

$$f(Y_{obs}|\theta) = \int f(Y_{obs}, Y_{mis}|\theta)\, dY_{mis},$$

and likelihood of parameters with observed values discarding that the missingness mechanism is proportional to $f(Y_{obs}|\theta)$, then in this case

$$L(\theta|Y_{obs}) \propto f(Y_{obs}|\theta). \tag{1.1}$$

In order to determine the missingness mechanism statistically and show the relationship between measured variables and the probability of missing data, Rubin (1976) used a dummy indicator matrix.

Let $M_{com} = \left(M_{ij}\right), \quad i = 1,2,\ldots,N \quad j = 1,2,\ldots,p$ with size of $N \times p$ be the indicator matrix whose elements are zero or one by;

$$M_{\mathrm{ij}} = \begin{cases} 1, & \text{if } Y_{ij} \text{ missing} \\ 0, & \text{other wise} \end{cases}$$

and can be shown as $M_{com} = (M_{obs}, M_{mis})$. The joint distribution of $Y_{com}$ and $M_{com}$ is

$$f(Y, M|\theta, \psi) = f(Y|\theta)f(M|Y, \psi) \tag{1.2}$$

and the joint distribution of the observed data and indicator matrix can be obtained by integrating out $Y_{mis}$ ,

$$f(Y_{obs}, M|\theta, \psi) = \int f(Y_{obs}, Y_{mis}|\theta)\, f(M|Y_{obs}, Y_{mis}, \psi)dY_{mis} \tag{1.3}$$

and $f(M|Y_{obs}, Y_{mis}, \psi)$ here shows the missingness mechanism as probability model. The likelihood of $\theta$ and $\psi$ is proportional to (1.3) (Little and Rubin , 1987)

$$L(\theta, \psi|Y_{obs}, M) \propto f(Y_{obs}, M|\theta, \psi). \tag{1.4}$$

### 1.3.1 Missing Completely at Random (MCAR)

In fact, probability model for $M_{com}$ would not be unrelated to $Y_{com}$ but the missing data mechanism is said to be missing at completely random, if the probability of missing data on a variable $Y_i$ is unrelated to other measured variables and is unrelated to the values of $Y_i$ itself (Enders, 2010). In probabilistic point of view this means that

$$f(M|Y_{obs}, Y_{mis}, \psi) = f(M|\psi). \tag{1.5}$$

Therefore the distribution of observed data is

$$
\begin{aligned}
f(Y_{obs}, M|\theta,\psi) &= \int f(Y_{obs}, Y_{mis}|\theta)\, f(M|Y_{obs}, Y_{mis}, \psi) dY_{mis} \\
&= f(M|\psi) \times \int f(Y_{obs}, Y_{mis}|\theta)\, dY_{mis} \\
&= f(M|\psi) f(Y_{obs}|\theta).
\end{aligned}
$$

We also assume that the parameters of the modeled data, $\theta$, and the parameters of the missingness mechanism, $\psi$, are distinct (Schafer, 1997). If $\theta$ and $\psi$ are distinct, then the likelihood based inferences for $\theta$ from $L(\theta,\psi|Y_{obs}, M)$ will be the same as likelihood based inferences for $\theta$ from $L(\theta|Y_{obs})$ (Little and Rubin, 1987). That is, if Equation $(1.5)$ and distinctness of parameters are satisfied then the missingness mechanism is ignorable, $L(\theta,\psi|Y_{obs}, M) \propto f(Y_{obs}|\theta)$, and likelihood based inferences can be based on $f(Y_{obs}|\theta)$ in case of MCAR (Schafer, 1997;Little and Rubin, 1987). This means that the missingness is completely unrelated to the data and under this assumption inferential analysis for only observed data will give valid results (Enders, 2010).

### 1.3.2 Missing at Random

If the probability of missing data on a variable $Y_i$ is related to other measured variables and if it is not related to the missing values of $Y_i$ itself, then the missingness mechanism called MAR (Enders, 2010).

That in case of MAR assumption probability model for $M_{com}$ does not depend on $Y_{mis}$ but depends on $Y_{obs}$,

$$f(M|Y_{obs}, Y_{mis}, \psi) = f(M|Y_{obs}, \psi). \tag{1.6}$$

Under the MAR assumption, the distribution of the observed data is

$$\begin{aligned} f(Y_{obs}, M|\theta,\psi) &= \int f(Y_{obs}, Y_{mis}|\theta)\, f(M|Y_{obs}, Y_{mis}, \psi) dY_{mis} \\ &= f(M, Y_{obs}|\psi) \times \int f(Y_{obs}, Y_{mis}|\theta)\, dY_{mis} \\ &= f(M, Y_{obs}|\psi) f(Y_{obs}|\theta). \end{aligned}$$

Thus by satisfying Equation (1.6) and distinctness of parameters $(\theta,\psi)$, the missingness mechanism is ignorable and inferences based on $L(\theta,\psi|Y_{obs}, M)$ are the same as based on $L(\theta|Y_{obs})$ (Little and Rubin, 1987). Knowing the distribution of the missingness mechanism is not needed for likelihood based estimation for $\theta$ (Little and Rubin, 1987). However in the case of making reliable inferences based on the sampling distribution of $Y_{com} = (Y_{obs}, Y_{mis})$, we have to take into account missing data mechanism because $\theta$ refers to the parameters of the model for the complete data (Schafer, 1997; Laird, 1988). Under the MAR assumption $Y_{obs}$ and $M_{com}$ are dependent as we said before. For example, drawing inferences about parameters of complete data, not for the parameters governing the marginal distribution of $Y_{obs}$ only, the missingness mechanism must be taken into account (Schafer, 1997).

### 1.3.3 Missing Not at Random

MNAR assumption is valid where the relation between the missingness of the variable and that variable itself can be specified. In this case of MNAR the missingness mechanism depends on $Y_{mis}$ and $Y_{obs}$ .

$$f(M|Y_{obs}, Y_{mis}, \psi) = f(M|Y_{obs}, Y_{mis}, \psi) \qquad (1.7)$$

$f(Y_{obs}, M|\theta,\psi)$ can not be factored as in MAR case because $f(M|Y_{obs}, Y_{mis}, \psi)$ must be specified for integral calculation. Specifying the model in this case is very tough problem and beyond the scope of this thesis. (See Daniels and Hogan (2007) and Enders (2010) for more detailed discussion of the topic).

## 1.4 Missing Data Patterns

Many different missingness patterns can be found in the literature. A missing data pattern refers to the scheme of observed and missing values in a data set and gives some prior information about the structure of data set to see if an orderly pattern emerges and for choosing an appropriate data handling methods but it does not provide any information about why the data are missing (Little and Rubin, 2002). Missingness may occur in any pattern.

Some of missingness pattern are shown in Figure 1, (a) monotone missingness often shown in repeated measures and longitudinal studies, sometimes a non-monotone dataset can be made monotone or nearly so by reordering. Reordering missingness pattern to the monotone pattern reduces the mathematical complexity of maximum likelihood and multiple imputation (Schafer, 1997) which we will review in chapter 3. Univariate missingness pattern (b) is not common but the arbitrary missingness pattern (c) are frequently seen in researches. The data analyzed in this study exhibited mostly univariate and arbitrary missingness pattern.

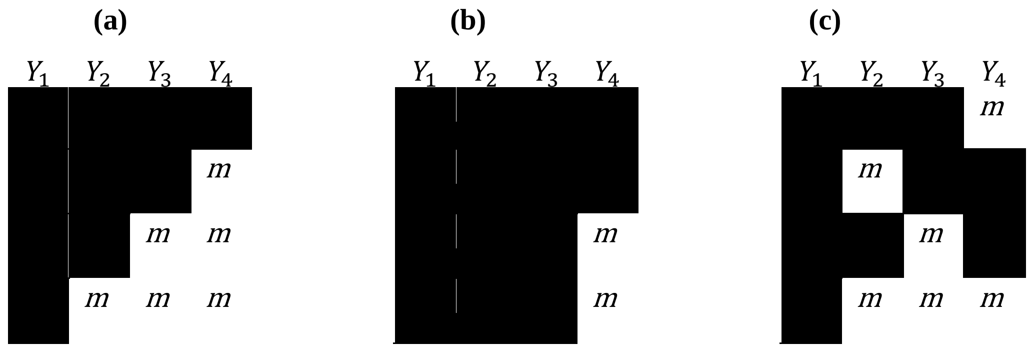

**Figure 1: Three prototypical missing data pattern: (a) monotone missingness, (b) univariate missingness, (c) arbitrary missingness**

## 1.5 Objectives of the Study

Climate related researches, especially clustering and modeling based ones, are required complete time series data to be used. On the other hand, considerable number of observations is missing in meteorological data. Missing data are frequently encountered in climate variables due to many reasons including failure in the observatory instruments, meteorological extremes and observation recording errors. This conflicting problem, however, can be overcome by estimating missing values using observations of correlated adjacent climate stations.

Climatological studies such as determining the effects of climate change on climate variables, homogeneity analysis of series, cluster and multivariate statistical analysis need complete data which are recorded in many stations spread out the whole region of interest in the long run. For that reason, it is crucial for the success of these studies that the missing data exist in series should be handled carefully (Schneider, 2001). An extensively used method in the literature for this purpose is the mean value imputation within a series. There are also several other methods which consider the temporal behavior of the series works as well for gap filling of missing values. However, considering temporal correlations only to impute missing values may cause loss of spatial information (Ramos-Calzado et al., 2008). So that missing values in a series can be imputed by also using correlated neighbor stations' complete observations. The main objective of this study is to compare the performances of conventional and modern methods for imputing monthly meteorological series to determine the most suitable approach or approaches on meteorological data set obtained from stations located in seven different regions of Türkiye.

Single Arithmetic Average (SAA), Normal Ratio (NR) and Normal Ratio Weighted with Correlations (NRWC) are the three simple methods used in the study. On the other hand, Neural Network based ones and Monte Carlo Markov Chain based on Expectation-Maximization Algorithm (EM-MCMC) methods are examined. These approaches can be evaluated in the class of sophisticated methods and mostly used in missing data literature. The modern approaches used in the study are preferred for their capability to handling of the spatial properties of the series. In addition to these, we used a modified EM-MCMC method in which results of above simple imputation methods are used as auxiliary series for missing data estimation.

Most of the comparative studies in the missing data literature relies on accuracy measurement based performances. Accuracy measures are useful for determining the central tendencies but in case of time series imputation, comparison of methods should take into account the temporal behavior or namely autocorrelation of the series. According to our opinion clarifying the disturbance impact of imputation methods on the autocorrelations is also needed. Therefore, another important objective of the study is evaluation of the imputation performances of the used methods in terms of temporal dependencies.

In order to evaluate imputation performances of methods, first, artificially missing periods based on three different missingness scenarios are created which are correspond to 10%, 20% and 50% missingness on the spanning period of the data and then artificially created missing values are re-estimated by above mentioned methods. For completed series, one of the accuracy measure based on squared errors Coefficient of Variation Root Mean Suare (CVRMSE) values are obtained, and compared. For examining the temporal changes of imputed series, we proposed the Correlation Cimension (CD) technique beside the CVRMSE values which is provided by nonlinear dynamic time series analysis.

First chapter of the study aimed to introduce missing data problem in time series and reviews of the noteworthy studies in the literature. In Chapter 2, data and climatological variables used in the study are described. Imputations methods used in the study are explained in Chapter 3. In Chapter 4, findings of the study and in Chapter 5, conclusion and discussions are presented.

# CHAPTER 2

# DATA

## 2.1 Meteorological data set of Türkiye: Overview

As we stated before, the climate researches based on land observatories database need long term and homogeneous records of meteorological variables and require the stations which are well spread out in the area of interest for the purpose of representativeness. In climatology literature, term of homogeneous refers to stations that is not exposed the unnatural change such as change in station location, observation equipment and observation procedure other than climatic changes (Peterson et al., 1998).

Term of climate refers to average weather, and climate can be specified by the central tendencies and variability of meteorological variables such as temperature, precipitation or through combinations of relevant elements (WMO, 2007). Meteorologists define a climate normal as the average of a climate variable over a universally accepted interval. For detecting the any changes on climate regimes or detecting any tendency on meteorological variables such as precipitation and temperature regimes and especially for clustering researches, at least 30 year averages of climatological variables are in need (WMO, 1983).

Total number of active climate station in Türkiye is nearly 270 and well distributed to all area of Türkiye but considerable number of the stations suffers from scarcity of lengthy records and incomplete data. Another challenge for researchers who deal with the Turkish meteorological database is the homogeneity problem of the observations. In case of the existence of the metadata, information provided by metadata is considerably helpful for determining the non-homogeneities easily. However, available stations' metadata have often insufficient information about stations chronology. Therefore, before starting any climatologic research in Türkiye, researcher must find solutions for missing data and homogeneity problem. Due to above mentioned problems, most of the researchers have had to restrict their studies to use very few numbers of climate stations and relatively short period of data. For example Tayanç et al. (2009) and Şen and Habib (2000, 2001a,b) used only 52 stations observations for detecting climate change in Türkiye and searching for a spatial analysis, respectively. As far as we know Türkeş is most involved with the quality control of Turkish meteorological data. Especially, Türkeş and his collaborators in 1995, 1996 and 1998 year publications specifically interested in expanding the meteorological data base of Türkiye. Türkeş and his colleagues in their later researches used expanded database as possible as by updating the data. They published plenty of study on their updated database which is consisted of 99 stations for precipitation and 70 stations for temperature and containing relatively long data period that covered to 1930 – 2002. Climatology of Türkiye and the long-term variability, trends and changes in the precipitation and temperature series were investigated in these studies by Türkeş (1999,2003), Türkeş et al. (2002,2009), Türkeş and Sümer (2004) and Türkeş and Erlat (2003, 2005, 2008, 2009). For clustering climate regions of Türkiye Ünal et al. (2003) report that the 113 stations were used after quality control assessment due to non-homogeneity and missing values. Evrendilek and Berberoğlu, (2008) used 124 climate stations for quantifying spatial patterns in Türkiye. They have had excluded 148 climate stations from the research after quality control study. In the light of these studies, we have seen that almost all of the researchers have excluded 50 to 60 percent stations in their studies due to considerable number of missing values and small amounts of missing values on remained stations were estimated by simple methods.

Another noteworthy study was published recently by Sahin and Cigizoglu (2010). They primarily aimed to detect non-homogeneous stations by several tests that are examined in the literature previously. Sahin and Cigizoglu used 250 station observations with six different climate elements on the period of 1974 – 2002. Secondary purpose of this study is to complete missing values by Linear Regression and EM algorithm. We want to note that the idea of using EM algorithm for handling missing data problem of Turkish meteorological database occurred simultaneously with our previous work (Aslan et al., 2010). In this point of view, we can say that this is the first study that using the latest approaches to handle with missing data. One of the drawbacks about the implementation of the study is that they are attempting to estimate missing values before homogeneity assessment of the stations observations but we know that the imputations or the estimations of missing values are valid under homogeneous series. (Schneider, 2001; Kalteh and Berndtsson, 2006; Ramos-Calzado et al., 2008; Aly et al., 2009).

## 2.2 Selected Stations and Meteorological variables used in the study

As we stated before, the main aim of the study is to find appropriate method for filling gaps in the Turkish Meteorological Database and additionally we claim that the missing values of any station can be estimated by concurrent observations from correlated stations database. In this study, we used two variables which are having the most priority in climatological studies. These variables are monthly total precipitation and monthly mean temperature variables spanning the period of 1965 - 2006.

For an experimental framework we chose 7 physical geography regions called South Eastern Anatolia (SEA), Eastern Anatolia (EA), Mediterranean Region (MR), Central Anatolia (CA), Aegean Region (AR), Western Thrace (WT) and North West Black Sea (NWB). These physical classifications of Türkiye also reflect the different climatic regimes, this is important because we want to compare performances to see if there is any adverse impact on estimations caused by the spatial variability of the variables. Selected stations on each region are homogeneous with respect to studies of Karabörk et al. (2007), Göktürk et al. (2008) and Şahin and Cigizoğlu (2010). Stations contain no missing data for the studied period and we chose a target and reference stations for each of different climate zone. Target stations are selected to be in the center of the reference stations while the reference ones are selected among the ones which are highly correlated with the target station's series. Encountered missingness pattern in Turkish

Meteorological database is generally long term missingness pattern such as three four or more than six years missingness. Therefore, three different missingness periods with 10%, 20% and 50% are created artificially on the target stations owing to comparing performances of used methods on the short and the long term missingness. Data used in the study are obtained from the Turkish State Meteorological Service (TSMS) and station information's given in Table 1. Target stations for each region showed in *italic* and underlined.

**Table 1: List of Stations**

| name | number | Latitude - Longitude | name | number | Latitude - Longitude |
|---|---|---|---|---|---|
| | SEA | | | E. Anatolia | |
| ***Akcakale*** | **17980** | 36°43' - 38°56' | ***Ağrı*** | **17099** | 39°43' - 43°03' |
| **Birecik** | **17966** | 37°01' - 37°57' | **Kars** | **17097** | 40°34' - 43°06' |
| **Ceylanpınar** | **17968** | 36°50' - 40°01' | **Doğubeyazıt** | **17720** | 39°33' - 44°05' |
| **Urfa** | **17270** | 37°09' - 38°47' | **Horasan** | **17690** | 40°03' - 42°10' |
| **Kilis** | **17262** | 36°42' - 37°06' | **Igdır** | **17100** | 39°55' - 44°03' |
| **Siverek** | **17912** | 37°45' - 39°19' | **Mus** | **17204** | 38°41' - 41°29' |
| | Mediterranean | | | C.Antolia | |
| ***Alanya*** | **17310** | 36°33' - 32°00' | ***Konya*** | **17244** | 37°59' - 32°33' |
| **Antalya** | **17300** | 36°52' - 30°42' | **Karaman** | **17246** | 37°12' - 33°13' |
| **Finike** | **17375** | 36°18' - 30°09' | **Akşehir** | **17239** | 38°21' - 31°25' |
| **Manavgat** | **17954** | 36°47' - 31°26' | **Beyşehir** | **17896** | 37°41' - 31°44' |
| **Anamur** | **17320** | 36°05' - 32°50' | **Cihanbeyli** | **17191** | 38°39' - 32°57' |
| **Silifke** | **17330** | 36°23' - 33°56' | **Karapınar** | **17902** | 37°43' - 33°32' |
| | Aegean Region | | | WT | |
| ***İzmir*** | **17220** | 38°23' - 27°04' | ***Uzunkopru*** | **17608** | 41°15' - 26°41' |
| **Kuşadası** | **17232** | 37°52' - 27°15' | **İpsala** | **17632** | 40°55' - 26°22' |
| **Çeşme** | **17221** | 38°18' - 26°18' | **Kırklareli** | **17052** | 41°44' - 27°13' |
| **Dikili** | **17180** | 39°04' - 26°53' | **Lüleburgaz** | **17631** | 41°24' - 27°21' |
| **Bornova** | **17790** | 38°28' - 27°13' | **Tekirdağ** | **17056** | 40°59' - 27°30' |
| **Salihli** | **17792** | 38°29' - 28°08' | **Edirne** | **17050** | 41°41' - 26°33' |
| | | | | NWB | |
| | | | ***Bartın*** | **17020** | 41°38' - 32°22' |
| | | | **Akçakoca** | **17612** | 41°05' - 31°10' |
| | | | **Düzce** | **17072** | 40°50' - 31°10' |
| | | | **İnebolu** | **17024** | 41°59' - 33°47' |

For the purpose of missing data estimation, finding reference stations for which the station has missing data cause a problem that needs to be examined. Although we able to find more reference stations for some of target stations, we have restricted the research for maximum five reference stations to use for estimation with the aim of reflecting the natural situation. For instance, one can find more than 10 reference stations in case of İzmir and Konya regions. Actually, for completion of missing data, it is beneficial to use more than five reference station, if it is possible, but for example as in the case of Bartın in this study and for Blacksea and South Eastern regions for future work finding suitable reference stations will be too problematic. In addition to this, sometimes reference stations may not surround the target station as it’s desired and they can be spread out unevenly. Bartın, Akçakale and Alanya stations are good examples for this situation. Within the scope of this study we also aimed to compare capability of methods in case of the scarcity and unevenly distributed reference stations.

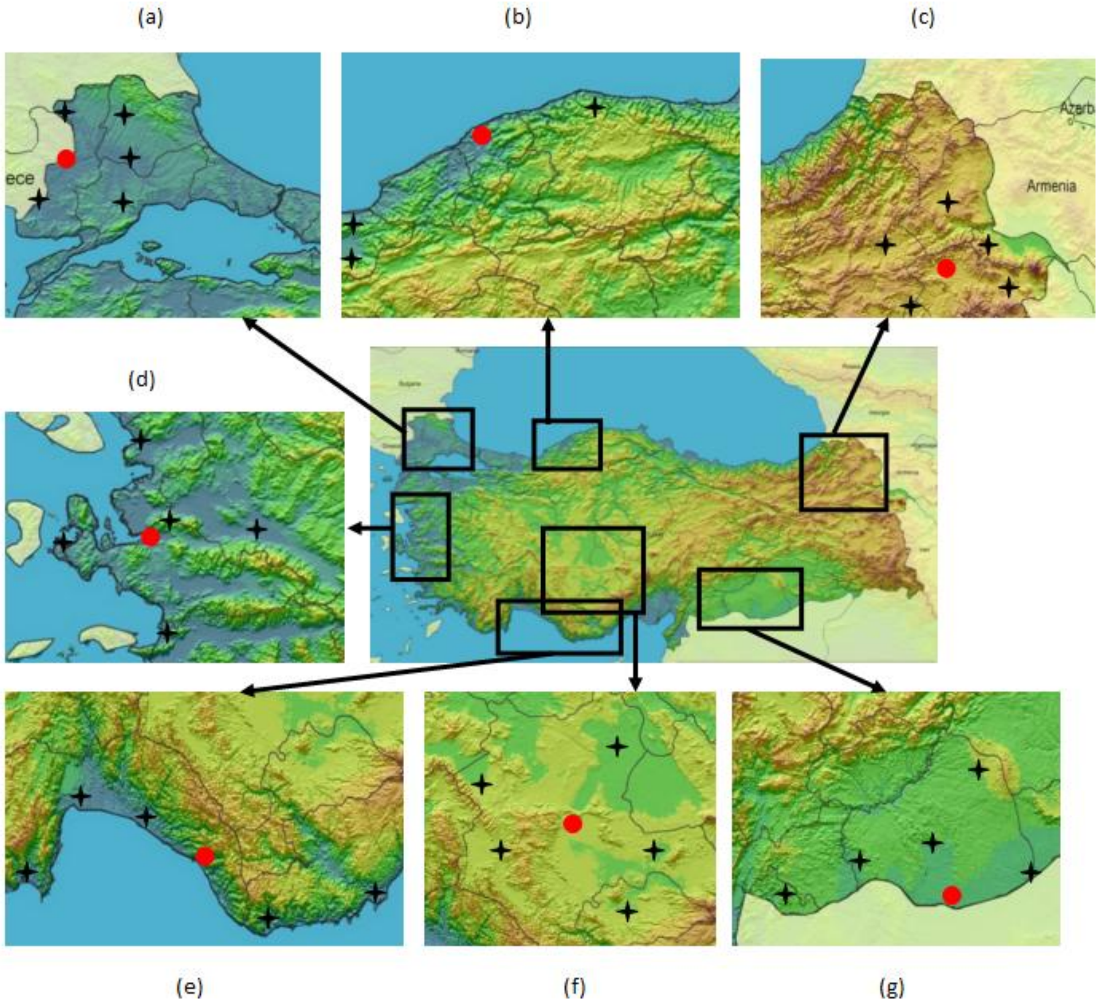

**Figure 2: Location of stations: (a) WT (b) NWB (c) EA (d) ER (e) MR (f) CA (g) SEA**

Locations of stations are shown on topographic map of Turkey in Figure 2. Target stations are marked as red circle and reference stations are marked as black star. We had previously stated that used stations have no missing data for studied period and they are well conditioned principle stations. For selection of reference stations we pay attention to the high correlations between stations; however correlations vary with respect to the meteorological variables. Precipitation highly depends on orography and it is highly changeable even in the same area. For example, correlations for temperature variable do not fall below 0.9 in the Eastern Anatolia Region, while for the precipitation variable correlations seem to be very modest. We give the correlations of stations in the following tables, Table 2 shows the temperature correlations and Table 3 shows the precipitation correlations.

**Table 2: Monthly Temperature Correlations Between Stations (1965 – 2006)**

| Reference | | Reference | | Reference | |
|---|---|---|---|---|---|
| | ***Akcakale*** | | ***Ağrı*** | | ***Alanya*** |
| **Birecik** | 0.9985 | **Kars** | 0.9938 | **Antalya** | 0.9892 |
| **Ceylanpınar** | 0.9984 | **Doğubeayzıt** | 0.9902 | **Finike** | 0.9948 |
| **Urfa** | 0.9971 | **Horasan** | 0.9945 | **Manavgat** | 0.9942 |
| **Kilis** | 0.9936 | **Igdır** | 0.9824 | **Anamur** | 0.9951 |
| **Siverek** | 0.9958 | **Mus** | 0.9958 | **Silifke** | 0.9905 |
| | ***İzmir*** | | ***Konya*** | | ***Uzunkopru*** |
| **Kuşadası** | 0.9965 | **Karaman** | 0.9973 | **İpsala** | 0.9985 |
| **Çeşme** | 0.9981 | **Akşehir** | 0.9967 | **Kırklareli** | 0.9989 |
| **Dikili** | 0.9987 | **Beyşehir** | 0.9975 | **Lüleburgaz** | 0.9987 |
| **Bornova** | 0.9987 | **Cihanbeyli** | 0.9968 | **Tekirdağ** | 0.9911 |
| **Salihli** | 0.9959 | **Karapınar** | 0.9971 | **Edirne** | 0.9937 |
| | ***Bartın*** | | | | |
| **Akçakoca** | 0.9924 | | | | |
| **Düzce** | 0.9967 | | | | |
| **İnebolu** | 0.9892 | | | | |

As we seen by Table 2, temperature correlations between stations are quite high. Temperature is found to be uniform spatial distribution over areas having similar surface features (Von Hann, 2009). It is a continuous variable in space and time, and has a strong correlation with topographical characteristics (Dobesch et al., 2009). Therefore, temperature variable provides convenience for missing temperature imputation of any station so one can get reliable estimations easily even if applying  simple methods through information of nearby stations. This meteorological variable has its importance for detecting long term behavior of the annual cycle of air. In the recent decade, scientific discuss on the issue of ‘‘global climate change,’’ greatly has been done based on the parameter of temperature (Potter and Colman, 2003).  A meteorological station observes more than nine variants of temperature variables such as minimum, maximum and monthly mean minimum and maximums of temperatures.

**Table 3: Monthly Total Precipitation Correlations Between Stations (1965 – 2006)**

| Reference | | Reference | | Reference | |
|---|---|---|---|---|---|
| | ***Akcakale*** | | ***Ağrı*** | | ***Alanya*** |
| **Birecik** | 0.8831 | **Kars** | 0.4081 | **Antalya** | 0.7918 |
| **Ceylanpınar** | 0.9098 | **Doğubeyazıt** | 0.6355 | **Finike** | 0.8462 |
| **Urfa** | 0.9290 | **Horasan** | 0.7378 | **Manavgat** | 0.8894 |
| **Kilis** | 0.8371 | **Igdır** | 0.6664 | **Anamur** | 0.9144 |
| **Siverek** | 0.8819 | **Mus** | 0.7131 | **Silifke** | 0.8558 |
| | ***İzmir*** | | ***Konya*** | | ***Uzunkopru*** |
| **Kuşadası** | 0.9074 | **Karaman** | 0.7340 | **İpsala** | 0.7841 |
| **Çeşme** | 0.8834 | **Akşehir** | 0.7005 | **Kırklareli** | 0.8673 |
| **Dikili** | 0.9059 | **Beyşehir** | 0.7589 | **Lüleburgaz** | 0.7884 |
| **Bornova** | 0.9641 | **Cihanbeyli** | 0.7457 | **Tekirdağ** | 0.8282 |
| **Salihli** | 0.8133 | **Karapınar** | 0.7549 | **Edirne** | 0.8001 |
| | ***Bartın*** | | | | |
| **Akçakoca** | 0.7674 | | | | |
| **Düzce** | 0.7432 | | | | |
| **İnebolu** | 0.7045 | | | | |

As in Table 3, the precipitation correlations for Eastern Anatolia and North West Blacksea are considerably weak. This can be seen as an indication for spatial variability of precipitation variable. In contrast to temperature, precipitation is non-continuous in space and time. Precipitation varies depending on orography and these two regions are highly affected by mountainous district. In this situation Kars station should be excluded from the analysis for precipitation variable. Precipitation phenomena can occurs in two different types, precipitation by convective process or by large-scale troposphere circulation. Convective precipitation is usually randomly distributed in space and could be limited over a narrow area, whereas stratiform precipitation is more evenly distributed in space (Dobesch et al., 2009). Due to high spatial variability, it is hard to find suitable imputation method for precipitation.

### 2.3 Data Preprocessing

Observations for monthly total precipitation and monthly mean temperature variables are obtained from TSMS in the spanning period of 1950-2009 but in this study we have used the period of 1965 – 2006 to work with nonmissing data. Actually, raw data supplied from TSMS have several quality controls and file format problems.

First, supplied raw data were given in one column for a variable and for all climate stations in .txt file format which is not easy to use in statistical data analyses directly. Therefore, we had to convert .txt file format into several file formats in order to use data in different software.

Secondly, the most difficult part of the raw data problem is the identification of no rainy days and missing values in monthly total precipitation series. Since TSMS use zero to describe the amount of precipitation under 10mm, they have used blank to represent no precipitation days. Hence, it was very difficult to distinguish these two cases directly. For dealing with this problem, firstly we considered whole data set as a matrix and each cell of matrix correspond to observation of variable. As we were informed by meteorologists, some of blank cells in whole data set may contain zero value because climate stations of TSMS do not assign any value for the rainless months and leave the corresponding measure cell as blank like in missing value.  Zero values cannot be treated as missing value because it is a measure of rainless day or rainless month. Therefore, one could mistakenly treat the zero value as missing value in this situation. We followed a particular algorithm to distinguish zero values and missing values in monthly precipitation data. We compared the monthly total precipitation data together with the monthly mean temperature and the daily precipitation data set by checking the each blank cells of precipitation data set with latter one by one. Temperature is a continuous variable and it has to be measured at each station continuously and it cannot be missing theoretically unless the missingness caused by another problem. Problem that causes missingness in the temperature observation probably will affect all other observable variables on that station, so we use temperature series as reference series for determining zero values. Therefore, if any of the blank cell is in the precipitation data matched with observed cell in the temperature data set, then it is suspected that in that month there was no precipitation and blank cells of the precipitation as in this situation may filled by value zero. To strengthen our decision for filling the blank cell with zero value, we use other information such as information about arid years, observations of neighbor stations humidity, pressure and daily precipitation series of the same station. If all these informations give the same results then we decided to use zero value for the corresponding cell of precipitation data. We have also requested stations' metadata and annual bulletins for determining missing values. We have used available metadata and bulletins for some stations and in some doubtful situations, chronology of stations checked by meeting of station chief. By preprocessing the data, we have prepared data not containing any missing value for the studied period 1965-2006.

# CHAPTER 3

# METHODS

Number of methods has been developed to handle missingness problem in meteorological time series data but in the context of this study, we focused only on Multiple Imputation based algorithms and Neural Network Based estimations. These two approaches are classified as model based imputation algorithms and widely considered in the missing data literature and such noteworthy studies published by Schneider (2001); Junninen et al., (2004); Coulibaly and Evora (2007) and Kalteh and Hjorth (2009). These approaches give comparable results if applied correctly to corresponding data. Beside the computational intensive methods, we have also tried simple methods for missing data estimation which are easy to implement and often used by researchers such as Paulhus and Kohler (1952) ; Young (1992); Eisched et al., (1999); Xia et al., (1999); Teegavarapu (2005); Aly et al., (2009).

Within the aim of comparing relative performance of missing data techniques, we introduce Nonlinear Dynamic Time Series Analysis and Correlation Dimension (CD) technique in this chapter. Most of the published study compares the missing value estimation methods by accuracy measures that rely on central tendencies. We believe that the better comparison methods, along with the accuracy measures, should be considered in case of times series imputation studies that are taking into account the temporal properties of the data. Therefore, we proposed the CD technique to compare missing data estimation in time series applications.

The imputation methods used in the study are namely Single Arithmetic Average (SAA), Normal Ratio (NR), Normal Ratio Weighted with Correlations (NRWC), and model based computationally intensive methods, Multi Layer Perceptron Neural Network (MLPNN) and Monte Carlo Markov Chain based on Expectation-Maximization Algorithm (EM - MCMC) . In addition to these, we have proposed a modification on the EM - MCMC method which uses the results of different imputation methods as auxiliary stations.

## 3.1 Simple Arithmetic Average (SAA) Method

In this method, missing values are imputed by the arithmetic average of concurrent observations in the surrounding stations having similar features with the target station. (Paulhus and Kohler, 1952).

Let $\alpha_m$ be the missing value for target station and $\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_N$ are concurrent observed values at the adjacent stations and $N$ is the number of used auxiliary station. Then,

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

This method can give reliable estimates if the meteorological variable does not have spatial variability and if correlated surrounding stations are available.

## 3.2 Normal Ratio (NR) Method

When concurrent values for station pairs are compared, it is found that in some cases their ratio (e.g. precipitation, wind speed) or their difference (e.g. temperature) tends to be constant (WMO, 1983). In the light of this information, developed model imputes missing data with the help of weights which are obtained from the ratios of nearby stations (Paulhus and Kohler, 1952). Let $T_1, T_2, \ldots, T_N$ are totals obtained at the adjacent stations and $T_m$ is total value of interested variable at target station then,

$$\alpha_m = \frac{1}{N}\left[\left(\frac{T_m}{T_1}\right)\alpha_1 + \left(\frac{T_m}{T_2}\right)\alpha_2 + \cdots + \left(\frac{T_m}{T_N}\right)\alpha_N\right] \qquad (3.2)$$

The Equation 3.2 is valid if the proportion of total amounts tends to be constant. In case of temperature variable, we may change proportions to average differences thus Equation 3.2 becomes

$$\alpha_m = \frac{1}{N}\left[\left(\alpha_1 + (\mu_m - \mu_1)\right) + \left(\alpha_2 + (\mu_m - \mu_2)\right) + \cdots + \left(\alpha_N + (\mu_m - \mu_N)\right)\right] \qquad (3.3)$$

$\mu_1, \mu_2, \dots, \mu_N$ are averages obtained at the adjacent stations and $\mu_m$ is average of the variable of interest at target station.

### 3.3 Normal Ratio Weighted with Correlations (NRWC) Method

NRWC method is similar to NR method described above. However, in this method, weights are obtained from the correlations between the target and reference stations (Young, 1992). Let $r_i$ denotes the correlation between target and $i$ th neighbour station and $n_i$ denotes the number of observations that correlation calculated based on so corresponding weight is derived according to

$$w_i = \frac{{r_i}^2(n_i - 2)}{1 - {r_i}^2},$$

and missing value at station $m$ is estimated by

$$\alpha_m = \frac{1}{\sum w_i}[w_1\alpha_1 + w_2\alpha_2 + \cdots + w_N\alpha_N]. \qquad (3.4)$$

### 3.4 Multi Layer Perceptron Neural Network (MLPNN)

Artificial Neural Network (ANN) model can be considered as a semi-parametric nonlinear function which corresponds the input to an output data to determine mathematical relationship of the input and output data. ANNs simulate human brain learning algorithm with parallel distributed processors or called neurons. In the ANN literature, learning or training of neural network is a mathematical optimization task that relies on modification of synaptic weights by supplying its input and corresponding output data. They have been widely used to model complex relationships of data (Haykin, 1999).

There are many ANN architecture which has been developed in the literature to deal with different type of problems such as clustering and pattern recognition. See Bishop, (1995); Haykin, (1999) and Patterson, (1996) for more detailed information. For time series modeling and atmospheric science based researches, Multi Layer Perceptron (MLP) architecture is much used type of neural network model than other type of architectures because of its potential to capture unknown relationship of an input and output data (e.g. Gardner and Dorling, 1998; ASCE, 2000; French et al., 1992; Toth et al., 2000). Therefore, it is also used in missing data and time series imputation researches due to reported benefits (e.g Junninen et al., 2004;  Coulibaly and Evora, 2007; Kalteh and Berndtsson, 2006; Kalteh and Hjorth, 2009; Sorjaama, 2009; Nelwamondo and Marwala, 2007).

General MLP architecture consists of three layers which are called input, hidden and output layers. Hidden layer can be designed to contain more than one layer, if it is appropriate. Each layer consists of processors called neurons and fully connected to each other with synaptic weights. Actually, neuron is nothing than a function which is called activation function and it can be linear or non-linear. Appropriate number of hidden layers and required number of neurons determined by trial and error procedure related to the learning\training algorithm. Weights are calculated and adjusted iteratively by selected learning algorithm; learning relies on minimization of deviations between output and actual value that released during learning process (Haykin, 1999).

Preventing the overfitting problem during training of MLP data must be divided into training and validation set. Training of network is stopped when the mean square error achieve its minimum both for training and validation set which is called early stopping rule, depicted in Figure 3. Thus, training algorithm of MLP can be considered as a generalization of the mean square error minimization and called back-propagation algorithm in NN literature. Errors are propagated to back for adjusting the weights by scaled conjugate gradient descent method (Haykin, 1999). Simplified architecture of MLP is illustrated in Figure 4. Arrows in the figure represents synaptic weights.

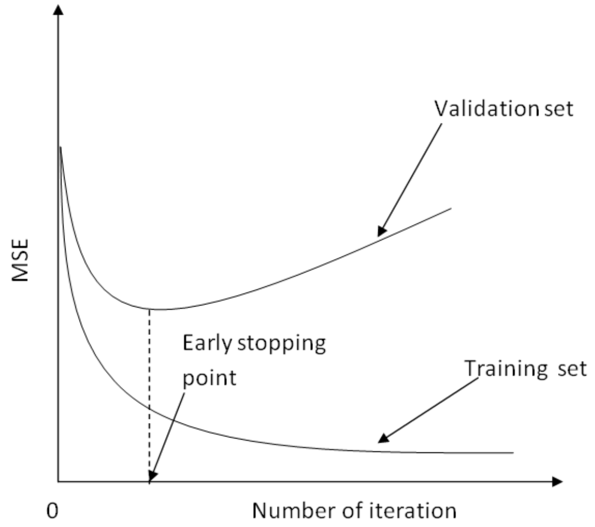

**Figure 3: Early-Stopping rule.**

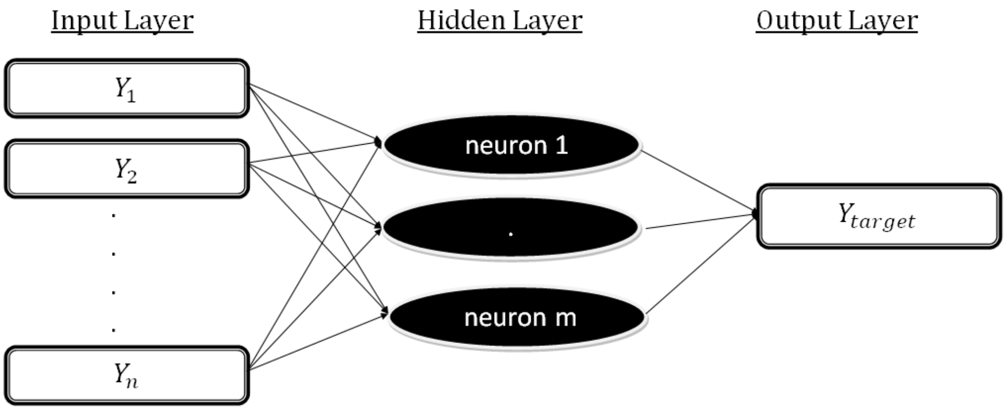

**Figure 4: Simplified architecture of MLP**

Mathematical representation of MLP is given as follows:

$$\hat{Y}_{target} = f_{output}\left(\sum_{j=1}^{M} w_j f_{hidden}\left(\sum_{i=1}^{N} w_{ji} Y_i + \alpha_0\right) + \beta_0\right) \qquad (3.5)$$

Where $w_{ji}$ denotes weights that carry inputs to the hidden layer and $w_j$ denotes weights that carry resultants of the hidden layer to the output layer. One can choose more than one hidden layer after trial and error procedure. $\alpha_0$ and $\beta_0$ denote threshold (bias) values. $f_{hidden}$ and $f_{output}$ are activation functions.

There are many activation functions used in literature and appropriate activation functions are also determined by trial and error process. Most used functions are hyperbolic tangent, sigmoid, sine, identity, and softmax functions. Softmax functions are used in case of categorical variables. We listed the most used activation functions in Table 4.

**Table 4: List of activation functions**

| name | definition | range |
|---|---|---|
| Hyperbolic Tangent | $\tanh(x) = [(e^{x} + e^{-x})/(e^{x} - e^{-x})]$ | $(-1, +1)$ |
| Sigmoid function | $\mathrm{g}(x) = [1/(1 - e^{-x})]$ | $(0, +1)$ |
| Sine function | $\mathrm{sine}(x) = \sin(x)$ | $(0, +1)$ |
| Identity function | $\mathrm{id}(x) = x$ | $(-\infty, +\infty)$ |
| Softmax function | $\gamma(x_k) = \exp(x_k)/\left(\sum_{i} x_i\right)$ | $(0, +1)$ |

## 3.5 EM Algorithm

EM algorithm is first developed by Dempster, Laird and Rubin (1977) and used for Maximum-Likelihood based parameter estimation under incomplete data. EM essentially deals with parameter estimation under ignorable missingness mechanism MCAR and MAR. Main objective in the algorithm is not filling in missing data to get complete data. EM takes into account the missingness as proportional information in order to estimate the parameter of interest via conditional expectations of missing data (Little and Rubin, 2002).

EM algorithm utilizes relationships of unknowns $Y_{mis}$ and $\theta$. Actually $Y_{mis}$ has information on $\theta$ and $\theta$ includes information about $Y_{mis}$ that how $Y_{mis}$ likely should be. Thus, algorithm implements iteratively by augmentation of observed data, $Y_{obs}$, with initial estimate of $\theta$ and then with re-estimation of $\theta$ by the augmented data $Y_{com} = (Y_{obs}, Y_{mis})$ (Schafer, 1997).

The distribution function of the complete data, $Y_{com}$, can be factored as

$$f(Y_{com}|\theta) = f(Y_{obs}|\theta).f(Y_{mis}|Y_{obs}, \theta) \qquad (3.6)$$

The term $f(Y_{obs}|\theta)$ is the distribution of observed data and $f(Y_{mis}|Y_{obs},\theta)$ is the conditional predictive distribution of $Y_{mis}$ given $Y_{obs}$ and $\theta$ that keeps relationships of $Y_{mis}$ and $\theta$. The corresponding log-likelihood of $Y_{com}$ is

$$l(\theta|Y_{com}) = l(\theta|Y_{obs},Y_{mis}) = l(\theta|Y_{obs}) + \log f(Y_{mis}|Y_{obs},\theta). \qquad (3.7)$$

The complete data log-likelihood, $(\theta|Y_{com})$, cannot be calculated because of the right hand side of Equation (3.7) due to unknown $Y_{mis}$. To deal with this complexity first write Equation (3.7) as

$$l(\theta|Y_{obs}) = l(\theta|Y_{com}) - \log f(Y_{mis}|Y_{obs},\theta), \qquad (3.8)$$

and then taking the expectations of both sides of Equation (3.8) over the conditional predictive distribution of $Y_{mis}$ given $Y_{obs}$ and a current estimate of $\theta$, say $\theta^{(t)}$, $f\left(Y_{mis}\middle|Y_{obs},\theta^{(t)}\right)$, yields

$$l(\theta|Y_{obs}) = Q\left(\theta\middle|\theta^{(t)}\right) - H\left(\theta\middle|\theta^{(t)}\right), \qquad (3.9)$$

where

$$Q\left(\theta\middle|\theta^{(t)}\right) = \int [l(\theta|Y_{obs},Y_{mis})] f\left(Y_{mis}\middle|Y_{obs},\theta^{(t)}\right) dY_{mis}$$

and

$$H\left(\theta\middle|\theta^{(t)}\right) = \int \log f(Y_{mis}|Y_{obs},\theta) f\left(Y_{mis}\middle|Y_{obs},\theta^{(t)}\right) dY_{mis}.$$

According to the key result of Dempster, Laird and Rubin (1977), $\theta^{(t+1)}$ is better estimate than $\theta^{(t)}$, because the change from $\theta^{(t)}$ to $\theta^{(t+1)}$ in each iteration increases the log likelihood,

$$l\left(\theta^{(t+1)}\middle|Y_{obs}\right) \geq l\left(\theta^{(t)}\middle|Y_{obs}\right). \qquad (3.10)$$

Therefore, iteration of EM algorithm can be considered in two steps: **Expectation Step** and **Maximization Step**.

a. **E-Step:** In this step, the function $Q\left(\theta\middle|\theta^{(t)}\right)$ is calculated as the conditional expectation of complete data log likelihood over the conditional predictive distribution, $f\left(Y_{mis}\middle|Y_{obs},\theta^{(t)}\right)$, of $Y_{mis}$ given $Y_{obs}$ and a current estimate of $\theta$, say $\theta^{(t)}$.
b. **M-Step:** In this step, estimation of $\theta^{(t+1)}$ is carried out as if there were no missing data which is achieved by maximizing $Q\left(\theta\middle|\theta^{(t)}\right)$ from E-step.

In order to define convergency of iterations, differences of parameter estimations derived in the each iteration are considered. If the difference of consecutive estimates less than selected threshold value, then iterations are stopped. Estimations from the last iteration are used as parameter estimations.

### 3.6 Multiple Imputation (MI) and MCMC

As we mentioned before missing data handling methods can be divided into two general classes. The first one is deleting or ignoring approaches and the other one is imputation approaches. The basic idea in imputation approach is substituting a plausible value for each missing data and carrying out the desired analysis on completed data (McKnight et al., 2007). The term imputation can be considered as estimation or interpolation in some situations. For example, above mentioned conventional methods and MLP approaches are single imputation methods. What we try to do in the reviewed approaches are estimation of the missing values only once and if the estimation or interpolation is decided to appropriate, the bias will be reduced in some extent compared to the bias using the incomplete data. In fact, without considering random component, single imputations generally underestimates the parameters which is called imputation uncertainty (Allison, 2001; Little and Rubin, 2002). One of the major disadvantage of single imputation is that imputing a single value treats that missing value as known, so there is an uncertainty occurred by insufficient evaluation of sampling variability of missingness (Little and Rubin ,2002). MI is developed by Rubin (1987) on Bayesian framework to overcome this uncertainty by replacing each missing value with a set of plausible values and creating $m \geq 2$ complete data set (Figure 5).

Imputation of any missing value with MI procedure is achieved by drawing samples from the posterior predictive distribution of missing data. Then, complete data set with imputed values can be analyzed by standard complete data methods and resultants are combined for final inferences. Major advantage of MI is the reflection uncertainty caused by missing data even with small sample size (Little and Rubin, 2002). The steps of MI procedure are illustrated in Figure 6. The first step is imputation that relies on Bayesian principle and then following with the analysis step of each imputed complete data set by standard procedures. Finally results of these analyses are pooled to get a valid inference that reflects variability.

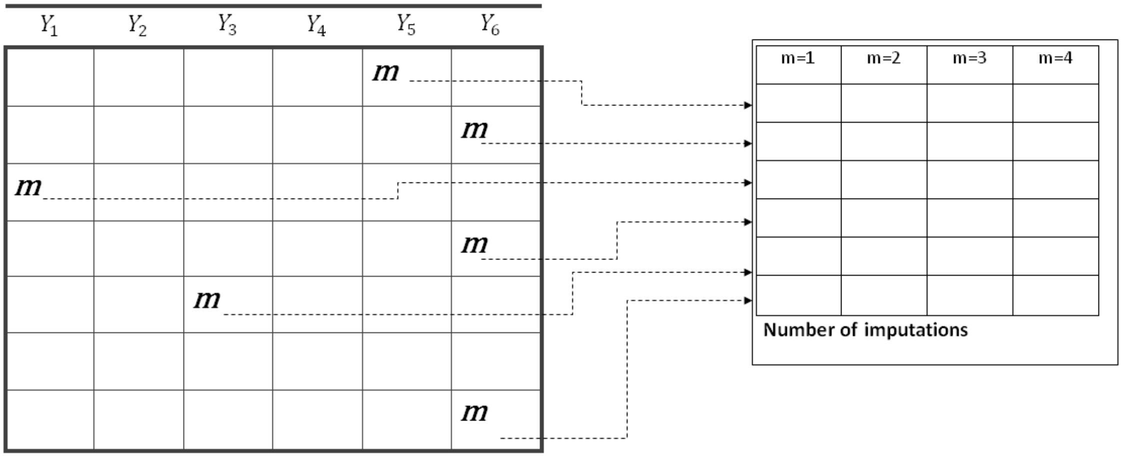

**Figure 5: Multiply imputed data set for *m*=4.**

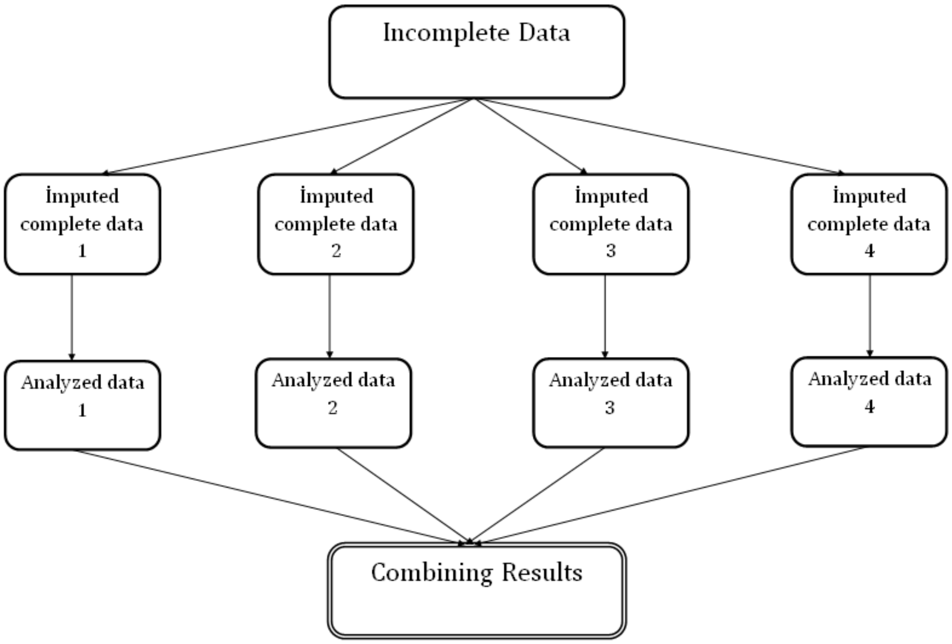

**Figure 6: Three steps of standard MI procedure**

For implementing MI procedure, there are many different statistical methods developed, see Tanner (1996),  Enders (2010), Schafer (1997) and Tan et al., (2010) for more details. In the scope of this study we only focus on MCMC method based on EM algorithm which is called data augmentation (Tanner and Wong, 1987; Schafer, 1997). MCMC method has been used on different statistical studies and in most cases with Bayesian framework, the MCMC is applied to estimate the posterior distribution parameters. Unlike the Frequentists point of view, Bayesians view parameter as a random variable and focused on posterior distribution of the parameter. Posterior distribution of the parameter given complete data, $Y_{com}$ ,   given as

$$f(\theta|Y_{com}) = \frac{f(Y_{com}, \theta)}{f(Y_{com})} = \frac{f(Y_{com}|\theta)f(\theta)}{\int f(Y_{com}|\theta)f(\theta)d\,\theta} \qquad (3.11)$$

MCMC sampling in Bayesian inferences used to simplify posterior expectations of any function of parameter, see Gilks, Richardson and Spiegelhalter, (1996) and Tanner, (1996) for more details.

For incomplete data problems, Schafer (1997) handled the MCMC in a different way. In order to implement MI in missing data problems, he proposed a method that uses data augmentation developed by Tanner and Wong (1987) which can be considered as a Bayesian counterpart of the EM algorithm for incomplete data (Tan et al, 2010). Under complete data inference on parameter, $\theta$, can be done easily by using $f(Y_{com}|\theta)$ and likelihood function $L(\theta|Y_{com})$, but in case of missingness, $Y_{com} = (Y_{obs}, Y_{mis})$, MI procedure based on Bayesian framework treats the unknown $\theta$ and $Y_{mis}$ as random variables to overcome incomplete data problem such a proper way (Schafer, 1997).

$Y_{mis}$ and $\theta$ could be simulated from the joint conditional distribution of $Y_{mis}$ and $\theta$ given $Y_{obs}$, $f(Y_{mis}, \theta|Y_{obs})$, if it is possible to find such a distribution. However, in order to implement MI in this situation, the posterior predictive distribution has to be examined by looking at the marginal distribution using the joint conditional distribution,$f(Y_{mis}, \theta|Y_{obs})$, Schafer (1997).

$$f(Y_{mis}|Y_{obs}) = \int f(Y_{mis}|Y_{obs}, \theta)\, f(\theta|Y_{obs}) d\theta, \qquad (3.12)$$

where $f(Y_{mis}|Y_{obs})$ is called the posterior predictive distribution $f(Y_{mis}|Y_{obs}, \theta)$ is the conditional predictive distribution of $Y_{mis}$ and $f(\theta|Y_{obs})$ is the posterior distribution of $\theta$ with respect to the observed data.

The posterior predictive distribution $f(Y_{mis}|Y_{obs})$ is examined as the average of the conditional predictive distribution of $Y_{mis}$ $f(Y_{mis}|Y_{obs}, \theta)$ over the posterior distribution of $\theta$ with respect to the observed data.

Summarizing the observed data, posterior distribution, $f(\theta|Y_{obs})$, under these expressions conducted by an augmentation of $Y_{obs}$ with an assumed value of $Y_{mis}$ as in two implementation steps: **The imputation I- Step** and **The posterior P- Step**.

a. **The imputation I- Step:** In this step, $Y_{mis}$ is drawn from of the conditional predictive distribution of $Y_{mis}$, $f(Y_{mis}|Y_{obs}, \theta)$, i.e.

$$Y_{mis}^{(t+1)} \sim f\left(Y_{mis}\middle|Y_{obs}, \theta^{(t)}\right). \tag{3.13}$$

b. **The posterior P- Step:** In the posterior step, a new value of $\theta$ drawn from its complete data posterior distribution given completed data $Y_{obs}, Y_{mis}^{(t+1)}$, i.e.

$$\theta^{(t+1)} \sim f\left(\theta\middle|Y_{obs}, Y_{mis}^{(t+1)}\right). \tag{3.14}$$

Repetition of the steps given in Equations (3.13) and (3.14) iteratively starting from an initial value of $\theta^{(0)}$, supplied to the EM algorithm in this study, yields a Markov Chain, i.e.

$$\left(Y_{mis}^{(1)}, \theta^{(1)}\right), \left(Y_{mis}^{(2)}, \theta^{(2)}\right), \left(Y_{mis}^{(3)}, \theta^{(3)}\right), \dots \dots \dots \dots \dots \dots \dots.$$

The distribution of this chain will converge to the joint conditional distribution of $Y_{mis}$ and $\theta$ given$Y_{obs}$, $f(Y_{mis}, \theta|Y_{obs})$. If the convergence of distribution is satisfied, then $\theta^{(t)}$ can be considered as an appropriate draw from the observed- data posterior distribution, $f(\theta|Y_{obs})$, and $Y_{mis}^{(t)}$ can be considered as an appropriate selection from the posterior predictive distribution $f(Y_{mis}|Y_{obs})$. Simplified procedure of the procedure is illustrated in Figure 7.

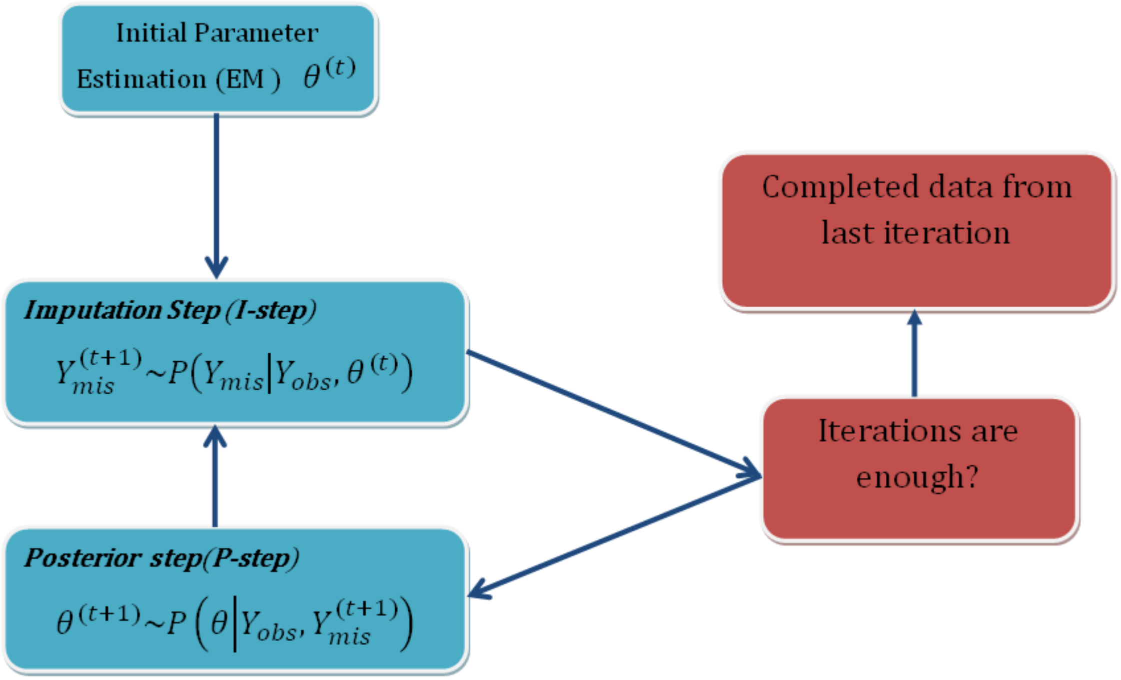
Initial Parameter
Estimation (EM ) $\theta^{(t)}$
*Imputation Step (I-step)*
$Y_{mis}^{(t+1)} \sim P\left(Y_{mis} \middle| Y_{obs}, \theta^{(t)}\right)$
*Posterior step(P-step)*
$\theta^{(t+1)} \sim P\left(\theta \middle| Y_{obs}, Y_{mis}^{(t+1)}\right)$
Iterations are enough?
Completed data from last iteration

**Figure 7: Simplified MI procedure based on EM – MCMC**

### 3.7 Nonlinear Dynamic Time Series (NDTS) Analysis

In order to evaluate imputation performance on time series data, we have proposed the correlation dimension (CD) technique which is constitute an important branch of the Nonlinear Dynamic Analysis (NDA). Significant deviations between calculated CD from observed and imputed time series will give information about imputation approach. Comparisons of methods in published studies have been usually examined with the accuracy measures which are only considering central tendencies. NDTS analyses are especially deal with underlying dynamics of time series, with sensitiveness on nonlinearity, which is considered to govern the observed time series rather than modeling to get future projections of time series (Small, 2005).

Any system that is evolving on time domain is considered as dynamical system. Dynamical systems are characterized by their phase spaces, which give information about long term behavior of system, stability conditions and sensitive dependence on initial conditions. These are primary research areas of dynamical system analyses. Phase spaces are constructed by realizations of the dynamics or by solutions of differential equations of the dynamics if it is possible to evaluate. Since the derivation of differential equations for any observed time series almost impossible NDTS methodology treat the observed times series as univariate or multivariate realization of a dynamic system and try to re-construct the phase space using this scalar time series with phase space reconstruction methods (Kantz and Schreiber, 2003).

Let us consider three tuple set $(\boldsymbol{\mathcal{M}}, \boldsymbol{\Phi}, \boldsymbol{\mathcal{T}})$ that describes any dynamic system. $\boldsymbol{\mathcal{M}}$ shows the $m$ dimensional real phase space on which the dynamics evolve and $\boldsymbol{\Phi}$ is the time dependent evolution operator and $\boldsymbol{\mathcal{T}}$ for the time. Main goal in the dynamical system analysis is to get information about the nature of dynamic structure via analyzing $\boldsymbol{\mathcal{M}}$, if it is not possible to get $\boldsymbol{\mathcal{M}}$ mathematically the nature of the dynamic structure will be understood by approximation to the $\boldsymbol{\mathcal{M}}$. Approximation is nothing but than trying to re-construction of phase space.

### 3.7.1 Phase Space Reconstruction with Time Delay Embedding

Let $z_n \in \boldsymbol{\mathcal{M}}$ and $\mathscr{g}: \boldsymbol{\mathcal{M}} \to \mathbb{R}$ then $y_n$ denote the univariate scalar time series in ℝ that realization of $\boldsymbol{\mathcal{M}}$ via function $\mathscr{g}(\,.\,)$, i.e. $y_n = \mathscr{g}(z_n)$. So re-construction of $\boldsymbol{\mathcal{M}}$ in dimension $d$ can be done by using Time Delay Embedding procedure (Takens, 1981;Small, 2005). Figure 8 simplifies the re-construction (Small, 2005).

In time delay embedding procedure univariate scalar time series are transformed into $d_e$ dimensional re-construction vectors with time delay $\tau$.

Let $y_n$ be univariate scalar time series and $n = 1, 2, 3, \dots \dots \dots, N$. Then, $N - (d_e - 1)\tau$ number of re-construction vectors are

$$\boldsymbol{x_i} = \left[y_i, y_{i-\tau}, y_{i-2\tau}, \dots \dots \dots \dots, y_{i-(d_e-1)\tau}\right] \tag{3.15}$$

for $i = 1{,}2{,}3, \dots \dots \dots \dots, N - (d_e - 1)\tau$.

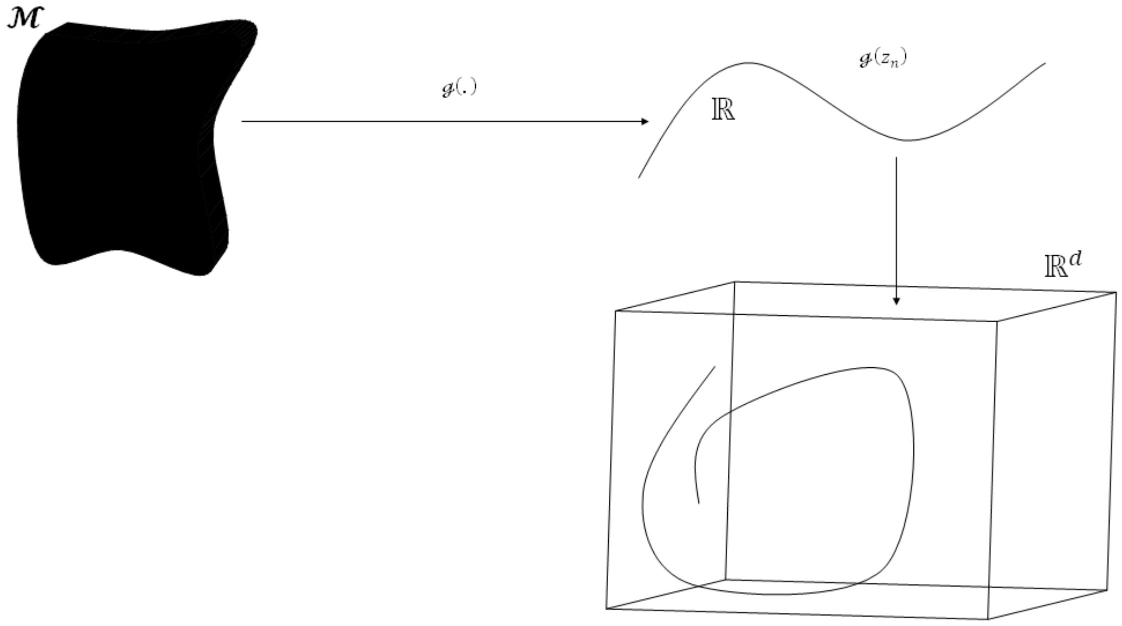

**Figure 8: Re-construction of univariate time series in dimension $d$**

Re-constructed vectors in phase space constitute a geometric figures and are called strange attractors in Chaos Theory (Takens, 1981). Nonlinear dynamic theory states that the systems, which are not random, will constitute well shaped attractors on phase space and trajectories or evolution of system will be attracted on this geometric figure. Therefore, geometric properties of attractors are main investigation topic of nonlinear dynamic system theory.

Embedding dimension, $d_e$, in $(3.15)$ is called minimum embedding dimension and $\tau$ is known as time delay. In order to have a proper re-construction of phase space these two indicators have to be calculated by an appropriate way.

### 3.7.1.1 Mutual Information Criterion

For selecting a proper time delay, Mutual Information Criterion (MIC) can be used (Fraser and Swinney, 1986). Mutual information for scalar time series can be calculated by

$$I(\tau) = \sum_{ij} p_{i,j}(\tau) log \frac{p_{i,j}(\tau)}{p_i p_j} \tag{3.16}$$

$p_i$ indicates the probability of $x_n$ being in $i$th bin of histogram, $p_j$ indicates the probability of $x_{n+\tau}$ being in $j$th bin of histogram and $p_{i,j}(\tau)$ indicates the joint probability of $x_n, x_{n+\tau}$, if $x_n$ in $i$th bin of histogram and $x_{n+\tau}$ in $j$th bin of histogram concurrently. $I(\tau)$ calculates amount of information on knowing $x_{n+\tau}$, if we are already on state $x_n$ in another saying, how much information carried by $x_n$ for knowing $x_{n+\tau}$.

The main aim of using time delay is constructing uncorrelated but statistically dependent vectors. So if time delay is selected too large, then re-construction vectors will be strictly uncorrelated and they will diverse randomly on phase space. If the time delay is selected too small then re-construction vectors will be strictly correlated and they will be too close to each other on phase space which means that the dynamic does not evolve. Therefore, first minimum of $I(\tau)$ is selected for optimum time delay value (Kantz and Schreiber, 2003).

#### 3.7.1.2 Minimum Embedding Dimension

Another component of re-construction procedure is the minimum embedding dimension, $d_e$. Most used methods for choosing appropriate embedding dimension is false nearest neighbor method (Kennel et al., 1992). This method aims to detect false neighbor vectors during the change of dimension $d_e$ to $d_e + 1$. Optimum minimum embedding dimension selected that when it gives the smallest percent of false nearest neighbor. Figure 9 shows false nearest neighbor points.

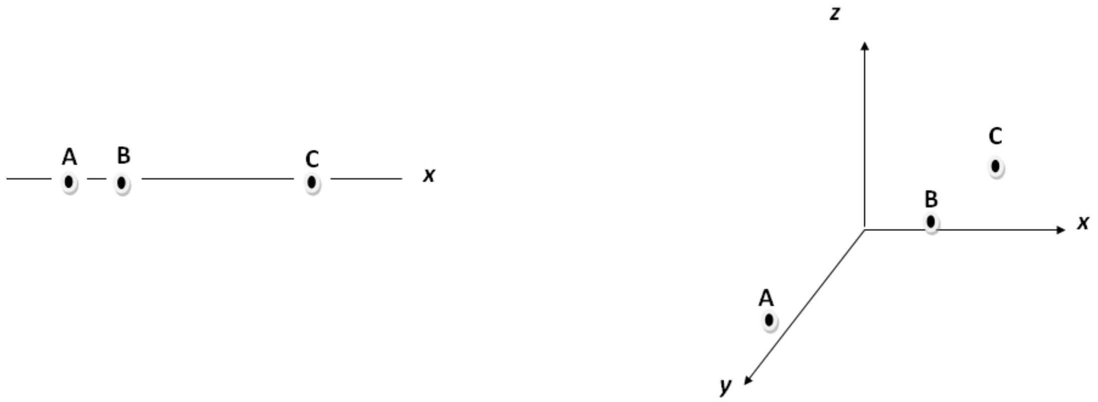

**Figure 9: Points A and B are false nearest neighbors if they are examined in 3 dimension**

Let $\boldsymbol{y_i}$ be the reconstruction vectors in $\mathbb{R}^n$ **,**

$$\boldsymbol{y_i} = [x_i, x_{i-\tau}, x_{i-2\tau}, \dots \dots \dots \dots, x_{i-n\tau}]$$

and $\boldsymbol{y_i}^{\boldsymbol{NN}}$ be the nearest neighbor of $\boldsymbol{y_i}$ in $\mathbb{R}^n$,

$$\boldsymbol{y_i}^{\boldsymbol{NN}} = [x_{i^{\cdot}}, x_{i^{\cdot}-\tau}, x_{i^{\cdot}-2\tau}, \dots \dots \dots \dots, x_{i^{\cdot}-n\tau}]$$

changing the dimension $n$ to $n+1$, then vectors will be

$$\boldsymbol{\hat{y}_i} = \left[x_i, x_{i-\tau}, x_{i-2\tau}, \dots \dots \dots \dots, x_{i-n\tau}, x_{i-(n+1)\tau}\right]$$

and

$$\boldsymbol{\hat{y}_i}^{\boldsymbol{NN}} = \left[x_{i^{\cdot}}, x_{i^{\cdot}-\tau}, x_{i^{\cdot}-2\tau}, \dots \dots \dots \dots, x_{i^{\cdot}-n\tau}, x_{i^{\cdot}-(n+1)\tau}\right].$$

Euclidean distance between vectors can be shown as

$$\left\|\widehat{\boldsymbol{y}}_{\boldsymbol{i}}-\widehat{\boldsymbol{y}}_{\boldsymbol{i}}{}^{\boldsymbol{NN}}\right\|^2-\|\boldsymbol{y}_{\boldsymbol{i}}-\boldsymbol{y}_{\boldsymbol{i}}{}^{\boldsymbol{NN}}\|^2=\left(x_{i-(n+1)\tau}-x_{i^{\cdot}-(n+1)\tau}\right)^2$$

and $R_T$ be threshold value between 10 and 30, i.e. $10 \leq R_T \leq 30.$
These two points are false nearest neighbor, if the proportion is bigger than $R_T$.

$$\frac{\left|x_{i-(n+1)\tau}-x_{i^{\cdot}-(n+1)\tau}\right|}{\|\boldsymbol{y}_{\boldsymbol{i}}-\boldsymbol{y}_{\boldsymbol{i}}{}^{\boldsymbol{NN}}\|}\geq R_T$$

### 3.7.2 Correlation Dimension

Correlation dimension technique is the dimension calculation of attractor on the re-constructed phase space. CD is called topological invariant in NDS analysis literature. CD technique claims that the data are generated by a finite dimensional attractor. It is aimed to summarize how often different parts of attractor is visited by the same trajectories. Finding infinite CD is the evidence of noise and finding integer CD is evidence of the strict periodicity (all vectors are on the same trajectory) but finding fractional CD is generally falsely considered as an evidence of chaos, in this manner CD is misunderstood concept by many researchers. Therefore, CD technique can be used for only detecting of what a system unlikely be (Small, 2005). In scope of this study, we are not dealing with the chaotic behavior of meteorological data. We use the CD for only its invariant property that is stated in NDA literature with respect to uniqueness of attractor.

We use Grassberger and Procaccia (1983) algorithm to find CD. Correlation dimension is calculated by using correlation sum. The technique uses resultants of time delay embedding procedure.

Let $W = N-(d_e-1)\tau$ , $\varepsilon$ denotes the radius. $y_i$ and $y_j$ are re-construction vectors.
Then, the correlation sum is

$$C(\mathcal{W}, d_e, \varepsilon) = \frac{2}{\mathcal{W}(\mathcal{W}-1)} \sum_{i=1}^{\mathcal{W}} \sum_{j=i+1}^{\mathcal{W}} \Theta\left(\varepsilon - \left(\|y_i - y_j\|\right)\right)$$

where $\Theta(x)$ is heaviside (step) function,

$$\Theta(x) = \begin{cases} 0, & x \leq 0 \\ 1, & x > 0 \end{cases}$$

Correlation sum is a function of time delay, $\tau$, and embedding dimension, $d_e$. Correlation sum count the vector pairs $(y_i, y_j)$ which are close to each other up to radius $\varepsilon$.

Grassberger and Procaccia, (1983) state that for $\varepsilon \to 0$ and $N \to \infty$ ,

$$D_c = \lim_{\varepsilon \to 0} \lim_{N \to \infty} \frac{\log C(\varepsilon)}{\log \varepsilon} \tag{3.17}$$

or due to power law

$$C(\mathcal{W}, d_e, \varepsilon) \propto \alpha \varepsilon^{D_c} \Longrightarrow \log C(\mathcal{W}, d_e, \varepsilon) \sim \log \alpha + D_c \log \varepsilon \tag{3.18}$$

where $D_c$ denotes the correlation dimension. Slope of Equation (3.18) will provide the correlation dimension$D_c$.

# CHAPTER 4

# RESULTS

In this chapter, we give results of applied methods on a regional basis. Selected stations cover the different areas of Türkiye reflecting different climate status, Figure 2 shows the location of used stations on Türkiye topological map. Two meteorological elements, total precipitation and mean temperature, were selected with respect to their significant impact on defining the climate of any region. Dataset covered the period of 1965-2006 on a monthly basis with no missing value. In order to evaluate imputation performances on short and long term missingness period we studied on three missingness scenarios which are mostly seen in Türkiye meteorological data set. We artificially create 10, 20 and 50 percent missingness on each target station that correspond to 4, 8 and 21 year missing period. All methods utilize the concurrent observations of reference stations to impute missing value. Target stations were sure to be surrounded by correlated stations. Table 2 and Table 3 list target and reference stations with their correlations and location parameters.

## 4.1 Implemented Procedures

Conventional methods are easy to implement and mostly used to filling gaps in climatology researches. However their efficiencies are questionable. SAA and NRWC methods can be identical in some cases, if the correlations of reference stations are nearly the same. NR method is better than the other two methods since the NR method utilizes total precipitation ratios and average temperature differences as weights. According to climatologists, precipitation ratios and mean temperature differences of stations' observations tend to be constant during the time. We used also proportions of monthly totals and seasonal totals to obtain weights however there was no extra gain on estimation accuracies.

MLPNN is a useful tool for input and output mapping. Percent of training and validation dataset, training algorithm, architecture of network, number of necessary neurons and activation functions must be considered cautiously before imputation. In this study, we investigate the appropriate architecture and percent of training dataset by trial and error. We used scaled conjugate gradient descent algorithm for training algorithm, 65 percent of data used for training and 35 percent of data used for validation with excluding missing data. We have seen that using large validation sample for training of the network is more suitable, if it is available. However, in case of 50 percent missingness we used 70 percent of data for training to get better estimations. During the trial and error procedure, we decided to use only one hidden layer in network architecture because it was sufficiently enough for capturing the complexity of input and output data. Using more than one hidden layer did not provide extra improvement while number of used neurons in hidden layer varied between 2 and 8. Although resultants from MLPNN method are similar, they vary slightly after each training period even if we apply the same procedure. This is natural corollary because of the error optimization algorithm. Therefore, we averaged the estimation of ten different training procedures to get stable estimation for each missing value. We used software PASW 18 for implementation of neural network.

In this study, EM-MCMC analyses implemented in software SAS 9.1.3. EM algorithm used to provide initial estimates for MCMC. In this method imputations and parameters are estimated iteratively. To get stable estimations and eliminating imputation uncertainty in MCMC method, we provide over dispersed initial estimates for each chain by bootstrap sampling. In Bayesian analyses it is recommended that using non informative prior is appropriate when there is no strong prior information about parameters. Therefore, we used non informative prior and under non informative prior distribution corresponding complete data posterior distribution has normal inverted Wishart distribution. Non informative prior used in this study is

$$f(\theta) \propto |\Sigma|^{-(p+1)/2},$$

and corresponding posteriors are

$$(\Sigma|Y) \sim W^{-1}(n-1,(nS)^{-1}),$$

$$(\mu|\Sigma,Y) \sim N(\bar{y},n^{-1}\Sigma).$$

Procedure for MI based EM-MCMC is given in Chapter 3. For detailed information on prior and posterior distributions please see Schafer (1997). We used average of 1000 imputation for each missing value with running multiple chains and for each chain the number of burn-in iterations before the first imputation is settled to 500. In this study, we assumed that data are from a multivariate normal distribution and the missingness mechanism is ignorable especially MAR assumption. In fact considered meteorological variables are not from a multivariate normal distribution. However, according to Demirtas et al. (2008) and Schafer (1997) violation of normality assumption may not cause invalid estimations. We use two procedure to implement EM – MCMC algorithm which we named EM – MCMC 1 and EM – MCMC 2. In first procedure, we used only reference stations as covariates and in the second procedure we used results of simple imputation (e.g. SAA, NR and NRWC) methods as extra auxiliary reference station.

With the aim of comparison of imputation performances we used the Coefficient of Variation of the Root Mean Square Error (CVRMSE). Root Mean Square Error (RMSE) is commonly used accuracy measure for forecasting performances in time series analysis. We used CVRMSE accuracy measure in order to eliminate scale dependency. CVRMSE can be shown as

$$\text{CVRMSE} = \frac{\text{RMSE}}{\overline{a_{(n)}}} = \frac{\sqrt{\sum_{i=1}^{n}(a_i - e_i)^2/n}}{\overline{a_{(n)}}},$$

where $a_i$ denotes the actual value of variable and $e_i$ denotes the corresponding imputed value. The term $\overline{a_{(n)}}$ denotes the mean of actual value for artificially created missing period. CVRMSE is obtained by dividing RMSE with mean of actual values.

Another scale independent accuracy measures are based on percentage errors such as Mean Absolute Percentage Error and symmetrical Mean Absolute Percentage Error. However, we did not use these measures that are based on percentage errors due to their reported disadvantages of being undefined under zero values and due to having skewed distribution when observed and estimated values are near to zero (Hyndman and Koehler, 2006).

For another comparison tool for imputation performance we have proposed the correlation dimension technique which we introduced in chapter 3. CVRMSE is sensitive to central tendencies so for example using coarsened average to fill in missing data could give also smaller CVRMSE values as those of used other imputation techniques. For assessing the impact of imputation methods on temporal behavior of time series we consider the CD technique besides the CVRMSE values. Actually, CD is not taking into account temporal correlations directly as determined by autocorrelation function. It is more than temporal correlation. CD also give information about dimensionality of interested variable and it is assessed as degrees of freedom of a dynamic system. For example, finding a CD 4.23 could be understood as there is at least 4 variables or dimensions to be considered for producing the observed phenomena. Under appropriate phase space re-construction indicators, which are minimum embedding dimension $d_e$ and time delay $\tau$ , the CD is unique.

## 4.2 Findings

Findings are given in regional basis. We used term of basin in this section for each area instead of their regional name. This is because that the selected stations are not fully reflecting the climate conditions of the region where they are located. However they differ from each other. Term of basin means that each studied area is narrower than the whole region that containing the interested area. Therefore, using term of basin is more suitable.

Temperature variable has very low spatial variability. Therefore, imputation performance for temperature series is quite good for all implemented methods. We did not need to evaluate CD deviations for imputed temperature series because results are found to be very close to the original series. We give only CVRMSE table and time series plot for temperature variable.

### 4.2.1 Findings for Ağrı Basin

Ağrı is located at Eastern Anatolia Region and has an extremely hard climate conditions Ağrı basin is classified in continental climate region. Winters are very cold and summers are hot. Spring and autumn terms are relatively short. Major form of precipitation in this basin is snow and nearly 125 days of the year is covered with snow.

Imputation results of monthly total precipitation variable for Ağrı station are given in Table 5. Correlations of stations are not very high and we excluded Kars station in the imputation procedure due to its low correlation for precipitation variable. Conventional methods SAA and NRWC gave bad results however NR is slightly better than these two methods.

MLPNN method seemed more appropriate for relatively short missingness period however its performance decreased with small size of training data set. EM-MCMC 1 and EM-MCMC 2 gave more consistent estimations for different missingness periods. We calculated only correlation dimensions of original data and imputed data by MLPNN, and EM – MCMC 2 due to their consistent results.

**Table 5: CVRMSE values of precipitation imputations for Ağrı station**

| | 17099 AĞRI STATION | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **SAA** | 0.4568 | 0.4375 | 0.4687 |
| **NR** | 0.4937 | 0.4897 | 0.4466 |
| **NRWC** | 0.4551 | 0.4316 | 0.4610 |
| **MLPNN** | **0.4344** | **0.3886** | 0.4314 |
| **EM-MCMC 1** | 0.4449 | 0.3952 | 0.4292 |
| **EM-MCMC 2** | 0.4447 | 0.3929 | **0.4286** |

CDs are listed in Table 6. Original data CD is 3.78 with embedding dimension $d_e = 10$ and time delay $\tau = 2$. Because of the high spatial variability of precipitation in this basin, estimated embedding dimension is found to be high. CDs for imputed data were not disturbed for 10 and 20 percent missingness period.

**Table 6: Correlation Dimensions**

| | 17099 AĞRI STATION | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **MLPNN** | 3.8 | 3.76 | 4.22 |
| **EM-MCMC 2** | 3.8 | 3.75 | 3.93 |

We gave the time series plot of original and imputed data for 20 percent missingness of precipitation variable in Figure 10. Results for temperature imputation at Ağrı station is shown in Table 7. Time series plot of original and imputed data for 20 percent (correspond to period 1970 – 1977) missingness of monthly mean temperature variable shown in Figure 11. Due low spatial variability and high correlations imputation results seem very satisfactory.

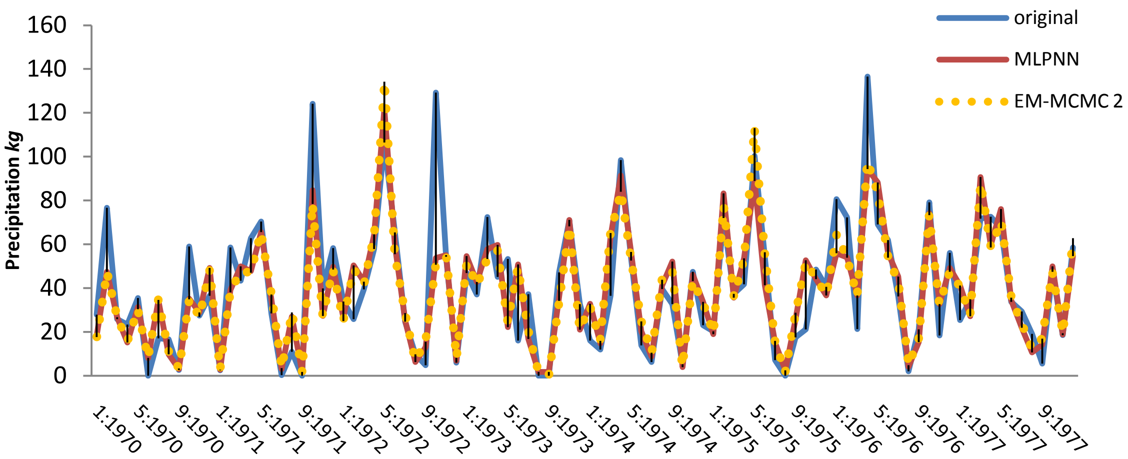

**Figure 10: Time series plot of precipitation imputations for Ağrı with 20% missingness**

**Table 7: CVRMSE values of temperature imputations for Ağrı station**

| | **17099 AĞRI STATION** | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **SAA** | 0.6580 | 0.5707 | 0.4276 |
| **NR** | 0.3611 | 0.3181 | 0.2372 |
| **NRWC** | 0.6559 | 0.5688 | 0.4261 |
| **MLPNN** | 0.2221 | 0.1914 | 0.1534 |
| **EM-MCMC 1** | **0.1775** | **0.1623** | **0.1400** |
| **EM-MCMC 2** | 0.1826 | 0.1631 | 0.1407 |

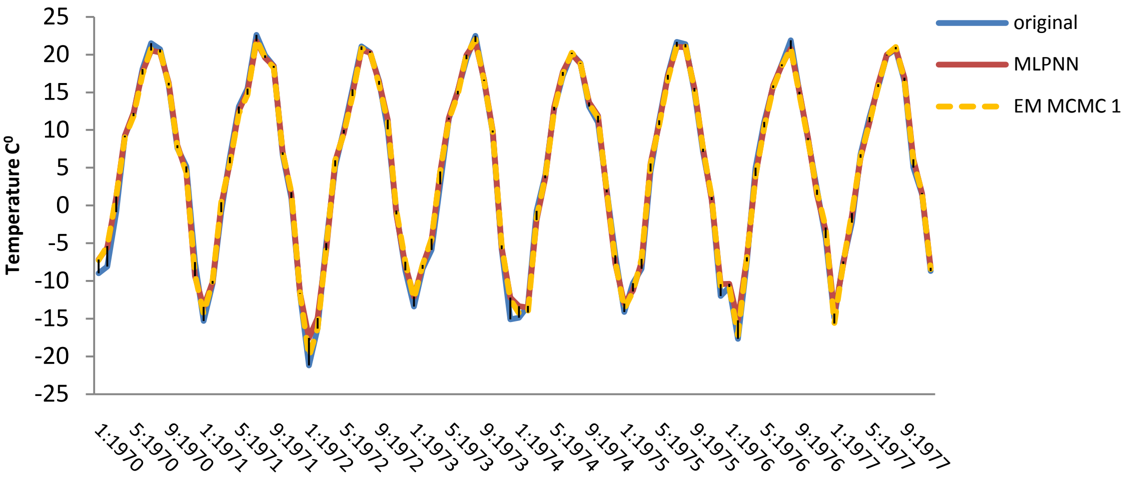

**Figure 11: Time series plot of temperature imputations for Ağrı with 20% missingness**

### 4.2.2 Findings for Akçakale Basin

Akçakale is a district of Şanlıurfa Province. Akçakale station is located at the border of Syria on Southeastern of Anatolia and reflects the southeastern continental climate conditions with cold winters and extremely hot and dry summers. The temperature in this basin exceeds 30 degree Celsius in the average number of 236 days and for nearly 125 days falls below zero Celsius. Major form of precipitation in this basin is rainfall. The average number of rainy days is 70 days in the average. Table 3 shows the precipitation correlations in this basin. Correlations of stations considerably high but unevenly distributed in this basin, they are not surrounding the target station as desired. Imputation results are given in Table 8.

Conventional methods were exhibit very poor results. However, estimation results of modern methods were good and were exhibit near the same performance in this basin and proposed EM – MCMC 2 procedure was slightly better than MLPNN. We can connect these results to low spatial variability of precipitation in this basin. Considering the all missingness periods, we can say that CDs are not disturbed much, see Table 9. CD for original precipitation variable is estimated as 2.58. Time delay is found to be $\tau = 3$ and corresponding embedding dimension is $d_e = 5$. We gave the time series plot of original and imputed precipitation data for 20 percent missingness in Figure 12.

**Table 8: CVRMSE values of precipitation imputations for Akçakale station.**

| | 17980 AKÇAKALE STATION | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **SAA** | 0.9253 | 0.9058 | 1.0035 |
| **NR** | 0.7271 | 0.6568 | 0.7183 |
| **NRWC** | 0.9183 | 0.9015 | 0.9974 |
| **MLPNN** | **0.6677** | 0.5774 | 0.5510 |
| **EM-MCMC 1** | 0.6740 | 0.5732 | 0.5374 |
| **EM-MCMC 2** | 0.6700 | **0.5701** | **0.5353** |

Considering the all missingness periods, we can say that CDs are not disturbed much, see Table 9. CD for original precipitation variable is estimated as 2.58. Time delay is found to be $\tau = 3$ and corresponding embedding dimension is $d_e = 5$. We gave the time series plot of original and imputed precipitation data for 20 percent missingness in Figure 12.

**Table 9: Correlation Dimensions**

| | 17980 AKÇAKALE STATION | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **MLPNN** | 2.58 | 2.64 | 2.59 |
| **EM-MCMC 2** | 2.58 | 2.60 | 2.63 |

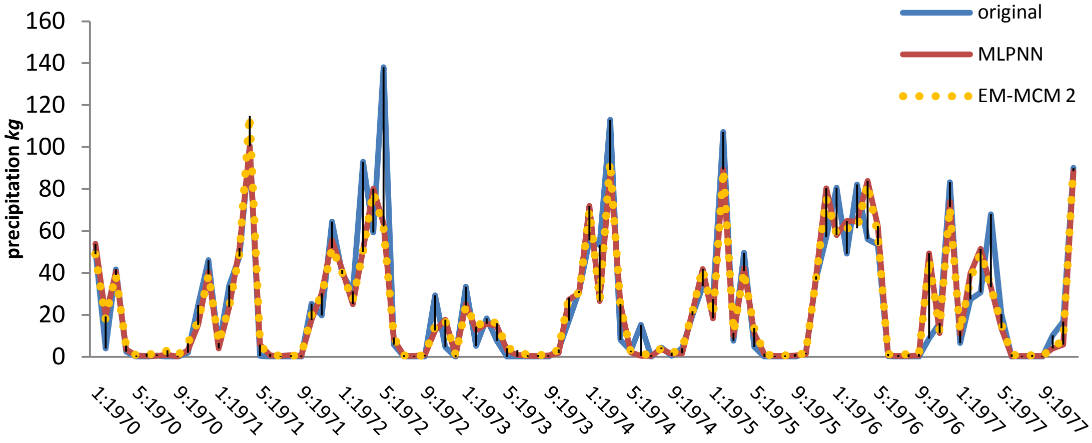

**Figure 12: Time series plot of precipitation imputations for Akçakale with 20% issingness**

Results for temperature imputation at Akçakale station is shown in Table 10. Time series plot of original and imputed data for 20 percent (correspond to period 1970 – 1977) missingness of monthly mean temperature variable shown in Figure 13.

**Table 10: CVRMSE values of temperature imputations for Akçakale station.**

| | **17980 AKÇAKALE STATION** | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **SAA** | 0.0424 | 0.0428 | 0.0449 |
| **NR** | 0.0283 | 0.0277 | 0.0369 |
| **NRWC** | 0.0423 | 0.0427 | 0.0449 |
| **MLPNN** | 0.0266 | 0.0208 | 0.0355 |
| **EM-MCMC 1** | **0.0162** | **0.0163** | **0.0345** |
| **EM-MCMC 2** | 0.0174 | 0.0170 | 0.0346 |

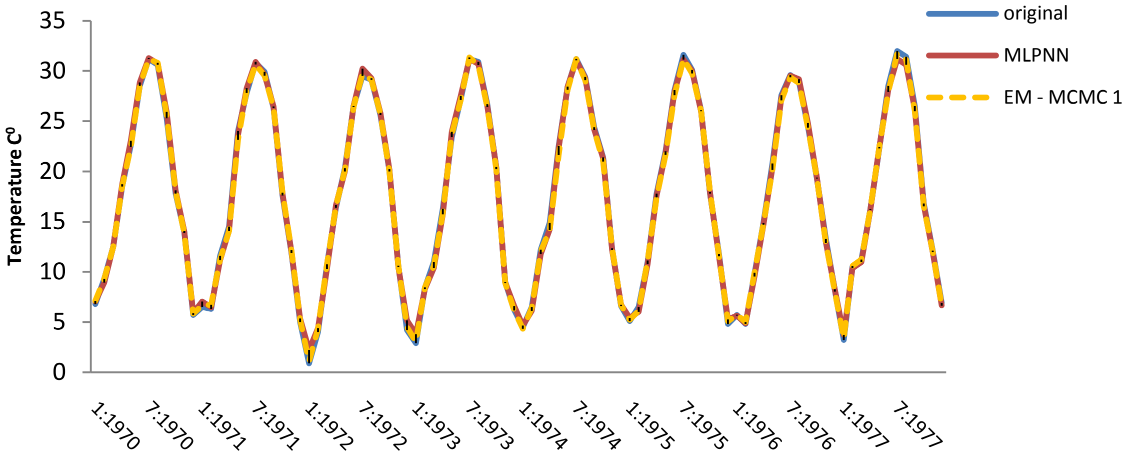

**Figure 13: Time series plot of temperature imputations for Akçakale with 20% missingness**

### 4.2.3 Findings for Alanya Basin

Alanya is located on Mediterranean region of Anatolia and southern of the Alanya is bordered by the Mediterranean Sea and its climate conditions highly affected by the Sea. Therefore, humidity rate is high both in winter and summer. Alanya basin reflects the typical Mediterranean climate with hot and dry summers while mild winters with rainy days. Excessive rainfall mostly seen in spring and autumn terms. The spatial variability of the precipitation is high for Alanya and for selected reference stations which are located across the Mediterranean Sea. Correlations of stations are high but they are not surrounding the target station as we stated before. Results of imputation procedure are given in Table 12.

**Table 11: Correlation Dimensions**

| | 17310 ALANYA STATION | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **MLPNN** | 3.49 | 3.46 | 3.49 |
| **EM-MCMC 2** | 3.53 | 3.49 | 3.95 |

As we have seen in previous results MLPNN performs better in short term missingness and EM-MCMC estimates shows stability. CDs were little affected by imputations.

Estimated CD are given in Table 11. CD for original precipitation variable is estimated as 3.62. Time delay is found to be $\tau = 3$ and corresponding embedding dimension is $d_e = 10$. The spatial variability of the precipitation is supported with high value of minimum embedding dimension.

**Table 12: CVRMSE values of precipitation imputations for Alanya station**

| | 17310 ALANYA STATION | | |
|---|---|---|---|
| | 10% | 20% | 50% |
| **SAA** | 0.5455 | 0.5351 | 0.6099 |
| **NR** | 0.6955 | 0.6612 | 0.7309 |
| **NRWC** | 0.5355 | 0.5284 | 0.6011 |
| **MLPNN** | **0.4309** | **0.4297** | 0.5812 |
| **EM-MCMC 1** | 0.4650 | 0.4662 | **0.5113** |
| **EM-MCMC 2** | 0.4651 | 0.4651 | 0.5120 |

Time series plot of imputations for 20% percent missingness showed in Figure 14.

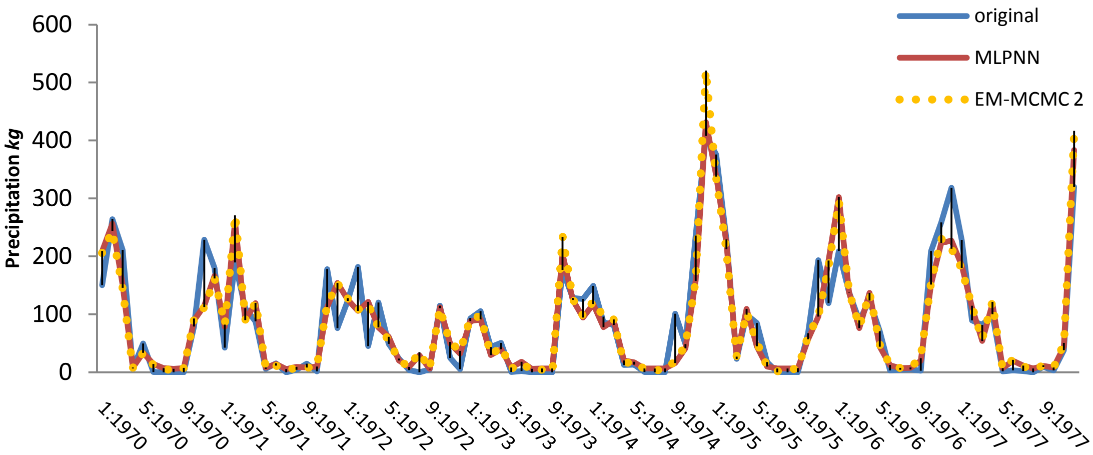

**Figure 14: Time series plot of precipitation imputations for Alanya with 20% missingness**

Table 13: CVRMSE values of temperature imputations for Alanya station

| | 17310 ALANYA STATION | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **SAA** | 0.0337 | 0.0328 | 0.0524 |
| **NR** | 0.0470 | 0.0447 | **0.0381** |
| **NRWC** | 0.0337 | 0.0327 | 0.0524 |
| **MLPNN** | 0.0270 | 0.0322 | 0.0502 |
| **EM-MCMC 1** | **0.0260** | **0.0304** | 0.0463 |
| **EM-MCMC 2** | 0.0291 | 0.0318 | 0.0451 |

Results for temperature imputation at Alanya station is shown in Table 13. Time series plot of original and imputed data for 20 percent (correspond to period 1970 – 1977) missingness of monthly mean temperature variable shown in Figure 15.

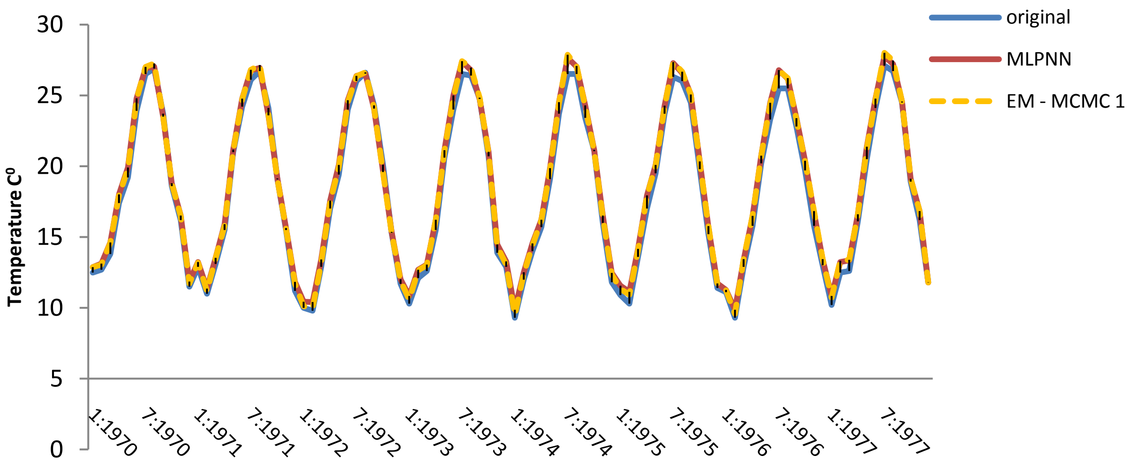

Figure 15: Time series plot of temperature imputations for Alanya with 20% missingness

### 4.2.4 Findings for Bartın Basin

Bartın is located on the North West Black Sea region of Anatolia and Northern of the Bartın bordered by the Black Sea. Bartın's climate reflects the temperate maritime climate. Summers are hot and winters are mild with rainy days. Bartın basin receives high amount of precipitation throughout the year and the spatial variability of precipitation is also high. Annual precipitation rate is nearly 1000 kg per $m^2$. Finding suitable reference station in this basin and for all Black Sea region of Anatolia are quite problematic. Location of stations depicted in Figure 2 and correlations are given in Table 3. Correlations for precipitation variable are not very high and we found only three suitable reference station for Bartın station. Imputation results are given in Table 14 and CD values are given in Table 17. CD values were not disturbed so much for imputed series. CD for observed series found to be 3.41. Time series plot of imputations for 20 percent missingness showed in Figure 16.

In this basin, estimations from NR method were slightly better than modern approaches. However comparing the overall performance of NR method to the modern methods does not support using the NR method due to inconsistent estimations. EM-MCMC approaches are preserve the consistency.

**Table 14: CVRMSE values of precipitation imputations for Bartın station**

| | **17020 BARTIN STATION** | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **SAA** | 0.3674 | 0.4398 | 0.4497 |
| **NR** | **0.3469** | **0.4227** | 0.4533 |
| **NRWC** | 0.3670 | 0.4389 | **0.4495** |
| **MLPNN** | 0.4002 | 0.4625 | 0.4588 |
| **EM-MCMC 1** | 0.3546 | 0.4276 | 0.4558 |
| **EM-MCMC 2** | 0.3562 | 0.4272 | 0.4575 |

**Table 15: Correlation Dimensions**

| | **17020 BARTIN STATION** | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **MLPNN** | 3.52 | 3.42 | 3.35 |
| **EM-MCMC 2** | 3.47 | 3.52 | 3.25 |

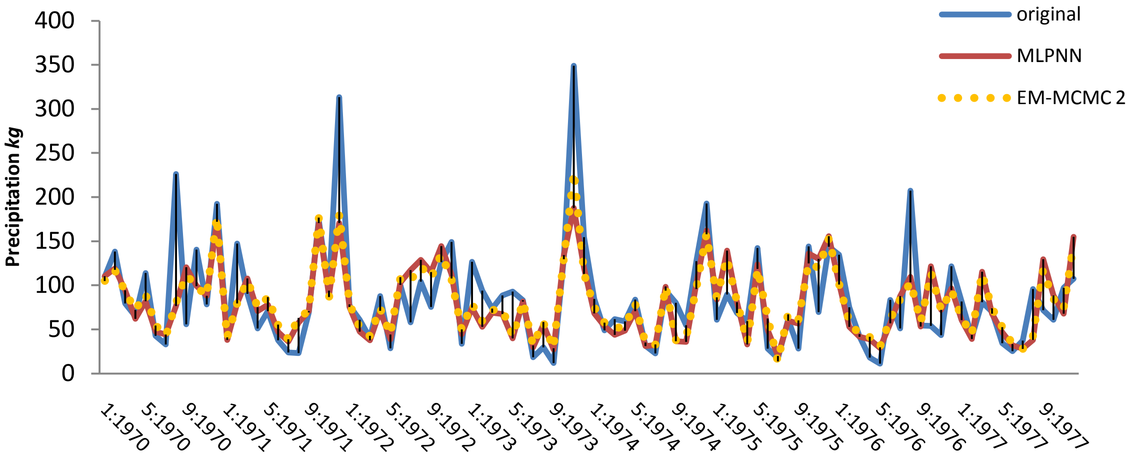

**Figure 16: Time series plot of precipitation imputations for Bartın with 20% missingness**

**Table 16: CVRMSE values of temperature imputations for Bartın station**

| | **17020 BARTIN STATION** | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **SAA** | 0.0648 | 0.0580 | 0.1066 |
| **NR** | 0.0501 | 0.0477 | 0.0865 |
| **NRWC** | 0.0647 | 0.0579 | 0.1064 |
| **MLPNN** | 0.0398 | 0.0342 | 0.0731 |
| **EM-MCMC 1** | **0.0348** | **0.0313** | 0.0680 |
| **EM-MCMC 2** | **0.0348** | **0.0313** | **0.0678** |

Results for temperature imputation at Bartın station is shown in Table 16. Time series plot of original and imputed data for 20 percent (correspond to period 1970 – 1977) missingness of monthly mean temperature variable shown in Figure 17.

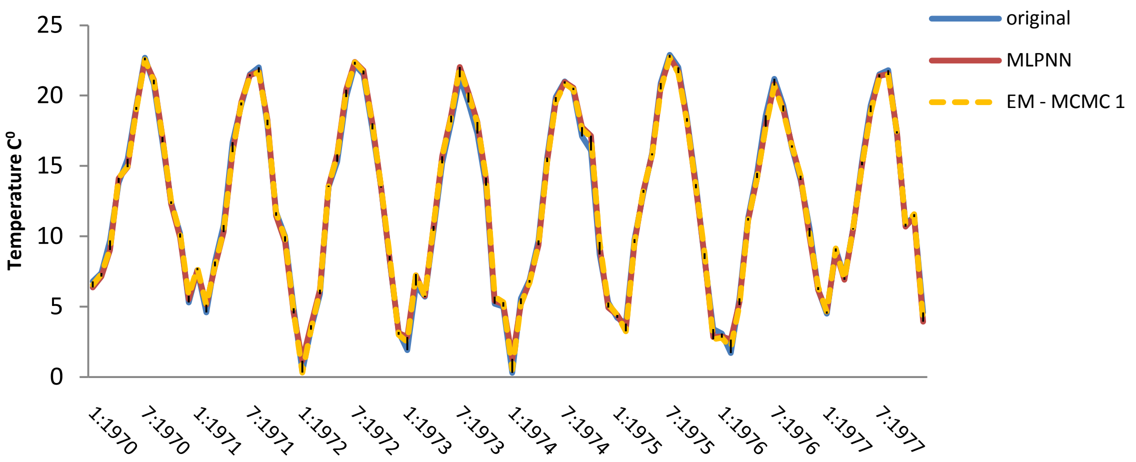

**Figure 17: Time series plot of temperature imputations for Bartın with 20% missingness**

### 4.2.5 Findings for İzmir Basin

İzmir is located on the Aegean coastline. Climate of İzmir basin is classified in the Mediterranean climate zone. Summers are hot and dry with warm winters. Humidity rate is also high due to the Aegean Sea. Major form of precipitation in this basin is rainfall. The spatial variability of precipitation is moderate in this basin. Correlations are given in Table 3. Imputation results are given in Table 17. EM-MCMC method performs better than MLPNN in this basin.

**Table 17: CVRMSE values of precipitation imputations for İzmir station**

| | **17220 İZMİR STATION** | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **SAA** | 0.3628 | 0.3816 | 0.4763 |
| **NR** | 0.3320 | 0.3357 | 0.3939 |
| **NRWC** | 0.3562 | 0.3760 | 0.4699 |
| **MLPNN** | 0.2928 | 0.2761 | 0.3037 |
| **EM-MCMC 1** | **0.2882** | **0.2725** | 0.2963 |
| **EM-MCMC 2** | 0.2909 | 0.2726 | **0.2959** |

Observed data CD is found to be 2.27. Both EM-MCMC and MLPNN do not affect on CD estimation for all missingness cases. This can be explained by high corraleted reference stations and low variability of precipitation in this basin. CDs are given in Table 18.

**Table 18: Correlation Dimensions**

| | 17220 İZMİR STATION | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **MLPNN** | 2.27 | 2.27 | 2.27 |
| **EM-MCMC 2** | 2.27 | 2.27 | 2.19 |

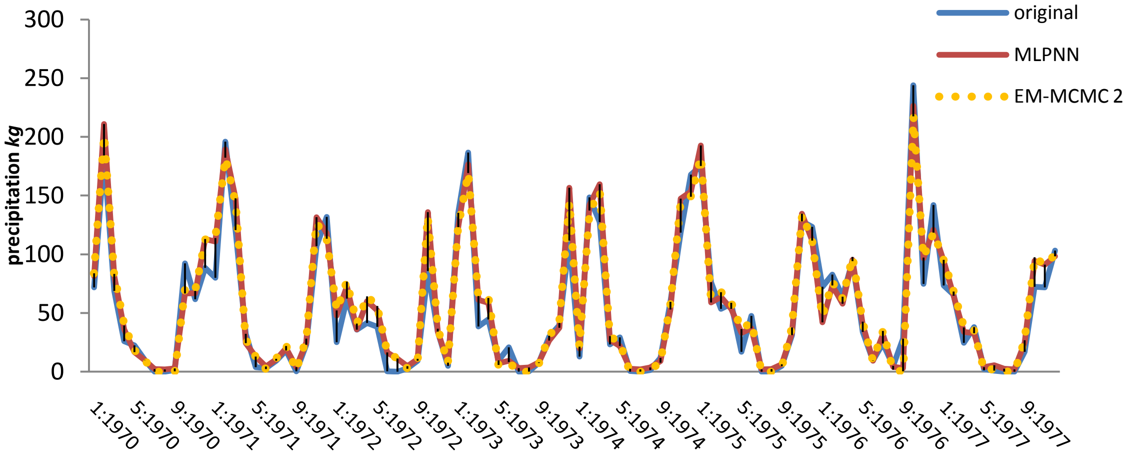

**Figure 18: Time series plot of precipitation imputations for İzmir with 20% missingness**

Time series plot of imputations for 20 percent missingness showed in Figure 18. Time series plot of observed and imputed data shows high compatibility.

**Table 19: CVRMSE values of temperature imputations for İzmir station**

| | 17220 İZMİR STATION | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **SAA** | 0.0611 | 0.0598 | 0.0600 |
| **NR** | 0.0262 | 0.0266 | 0.0210 |
| **NRWC** | 0.0611 | 0.0598 | 0.0600 |
| **MLPNN** | 0.0160 | 0.0156 | 0.0196 |
| **EM-MCMC 1** | 0.0158 | 0.0152 | 0.0177 |
| **EM-MCMC 2** | **0.0151** | **0.0145** | **0.0176** |

Results for temperature imputation at İzmir station is shown in Table 19. Time series plot of original and imputed data for 20 percent (correspond to period 1970 – 1977) missingness of monthly mean temperature variable shown in Figure 19.

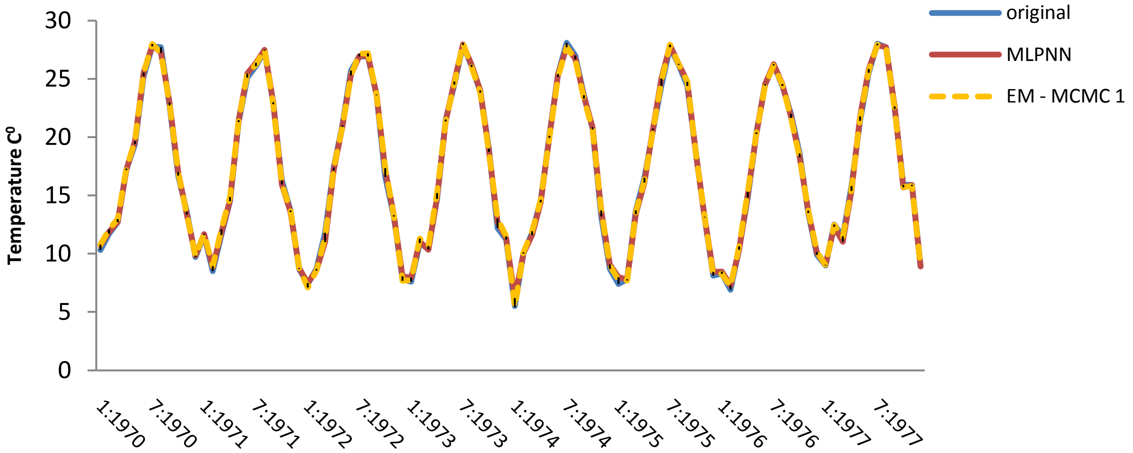

**Figure 19: Time series plot of temperature imputations for İzmir with 20% missingness**

### 4.2.6 Findings for Konya Basin

Konya basin is located on the Central Anatolia. This basin reflects the continental climate regime with hot and dry summers and cold winters. Correlations stations are given in Table 3. Major form of precipitation in this basin is rainfall. Although location of Konya is in plain, we expect that the spatial variability of precipitation will not affect the imputations. However, the number correlated reference stations are not high. In our opinion this is caused by the distance effect. Stations are far from the Konya station and they are located at high altitude. As we mentioned before, the precipitation variable is highly affected by orography and altitude of the area.

Table 20 shows the imputation results. NR method perform slightly better than both MLPNN and EM-MCMC methods. Despite the lack of reliability of estimations in this basin we propose to consider the CD estimations. Observed series CD is found to be 3.96. In short missingness period, CDs are not disturbed so much as we see in Table 21. Imputed period reflects the general behavior of the observed data.

**Table 20: CVRMSE values of precipitation imputations for Konya station**

| | **17244 KONYA STATION** | | |
|---|---|---|---|
| | **10**% | **20**% | **50**% |
| **SAA** | 0.4530 | 0.45944615 | 0.645315 |
| **NR** | **0.4449** | 0.4021071 | **0.537685** |
| **NRWC** | 0.4502 | 0.45358731 | 0.644689 |
| **MLPNN** | 0.4783 | 0.39175761 | 0.623299 |
| **EM-MCMC 1** | 0.4779 | **0.3893582** | 0.560829 |
| **EM-MCMC 2** | 0.4784 | 0.39171364 | 0.564387 |

**Table 21: Correlation Dimensions**

| | 17244 KONYA STATION | | |
|---|---|---|---|
| | 10 | 20 | 50 |
| MLPNN | 3.94 | 3.76 | 3.93 |
| EM-MCMC 2 | 3.81 | 3.70 | 4.23 |

Time series plot of imputations for 20 percent missingness showed in Figure 20. Time series plot of observed and imputed data shows high compatibility.

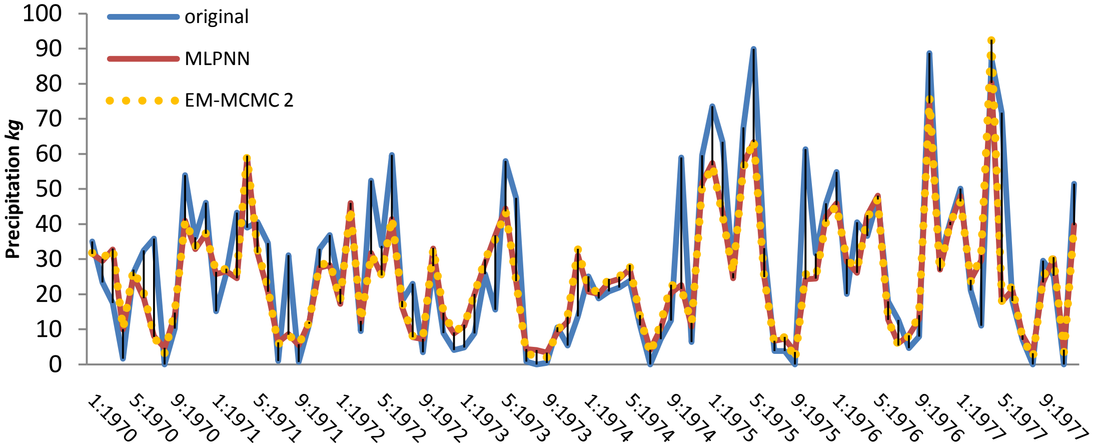

**Figure 20: Time series plot of precipitation imputations for Konya with 20 missingness**

**Table 22: CVRMSE values of temperature imputations for Konya station**

| | 17244 KONYA STATION | | |
|---|---|---|---|
| | 10% | 20% | 50% |
| SAA | 0.0560 | 0.0490 | 0.0488 |
| NR | **0.0484** | **0.0430** | **0.0525** |
| NRWC | 0.0560 | 0.0490 | 0.0488 |
| MLPNN | 0.0494 | 0.0434 | 0.0710 |
| EM-MCMC 1 | 0.0520 | 0.0454 | 0.0696 |
| EM-MCMC 2 | 0.0513 | 0.0445 | 0.0622 |

Results for temperature imputation at Konya station is shown in Table 22. Time series plot of original and imputed data for 20 percent (correspond to period 1970 – 1977) missingness of monthly mean temperature variable shown in Figure 21.

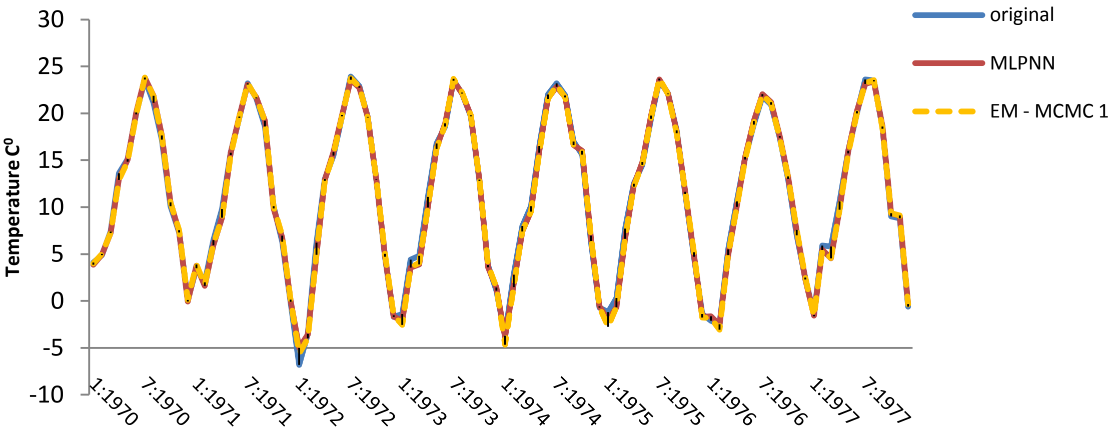

**Figure 21: Time series plot of temperature imputations for Konya with 20% missingness**

### 4.2.7 Findings for Uzunköprü

Uzunköprü is district of Edirne Province and it is located on the Western Thrace Region of Türkiye. Climate of Uzunköprü basin and the Western Thrace is not consistent. Usually, the continental climate characteristics have been reflected in this basin however in some years the Black Sea climate or the Mediterranean climate regime can have an effect on this area. This basin is known as transition zone of climate systems. Correlations are moderate and Uzunköprü station surrounded by reference station with short distances. Imputation results are given in Table 23. and estimated CDs are given in Table 24. Observed date CD is found to be 5.31. Time series plot of observed and imputed data given in Figure 22.

**Table 23: CVRMSE values of precipitation imputations for Uzunköprü station**

| | 17608 UZUNKÖPRÜ STATION | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **SAA** | 0.4441 | 0.4472 | 0.3962 |
| **NR** | 0.4032 | 0.4037 | 0.3802 |
| **NRWC** | 0.4428 | 0.4461 | 0.3927 |
| **MLPNN** | 0.3926 | 0.4012 | 0.3876 |
| **EM-MCMC 1** | **0.3852** | **0.4010** | **0.3753** |
| **EM-MCMC 2** | 0.3880 | 0.4020 | 0.3758 |

**Table 24: Correlation Dimensions**

| | 17608 UZUNKÖPRÜ STATION | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **MLPNN** | 5.26 | 5.07 | 5.29 |
| **EM-MCMC 2** | 5.33 | 5.17 | 5.52 |

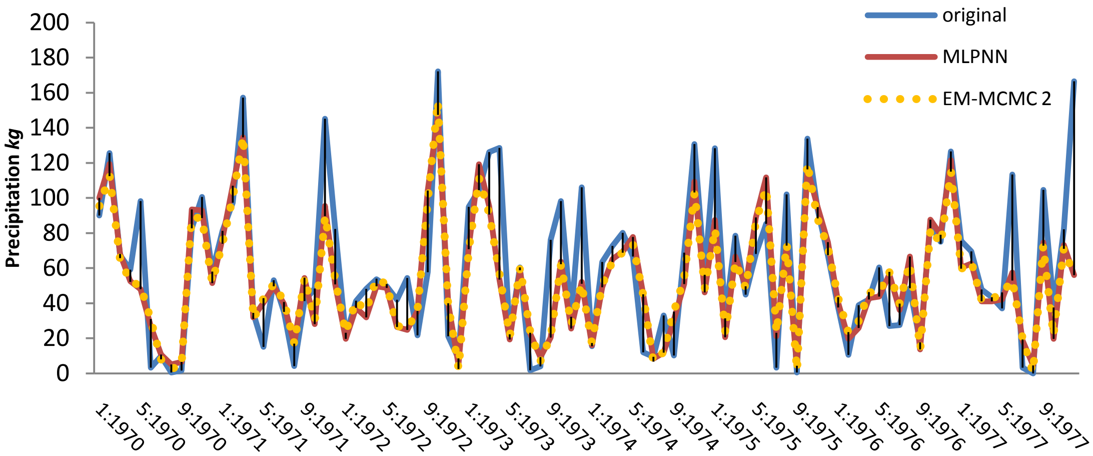

**Figure 22: Time series plot of precipitation imputations for Uzunkopru with 20% missingness**

Results for temperature imputation at Uzunköprü station is shown in Table 25. Time series plot of original and imputed data for 20 percent (correspond to period 1970 – 1977) missingness of monthly mean temperature variable shown in Figure 23.

**Table 25: values of temperature imputations for Uzunköprü station**

| | 17608 UZUNKÖPRÜ STATION | | |
|---|---|---|---|
| | **10%** | **20%** | **50%** |
| **SAA** | 0.0208 | 0.0234 | 0.0353 |
| **NR** | 0.0225 | 0.0258 | 0.0321 |
| **NRWC** | 0.0207 | 0.0234 | 0.0352 |
| **MLPNN** | 0.0190 | 0.0254 | 0.0351 |
| **EM-MCMC 1** | **0.0167** | **0.0174** | 0.0422 |
| **EM-MCMC 2** | 0.0166 | 0.0190 | **0.0396** |

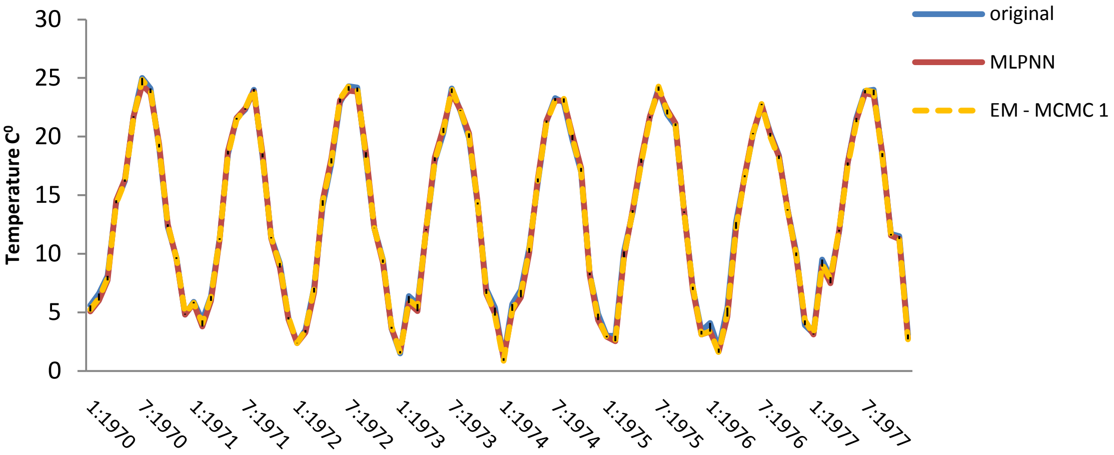

**Figure 23: Time series plot of temperature imputations for Uzunkopru , 20% missingness**

# CHAPTER 5

# CONCLUSION

Existence of missing values in any statistical data analysis causes biased results so appropriate handling of missing data becomes very important to decrease the bias effect of missingness up to a certain degree on data analyses. In this context, numbers of research have been published in literature to deal with missing data subject to many disciplines and it seems that will continue to publish. However, main objective of this study is that the comparing several imputation techniques to fill in missing values in Türkiye meteorological time series data set.

Although the official foundation of the Turkish State Meteorological Service (TSMS) as far back to 1925, unfortunately a very few number of climate stations have complete data and they are suffered with lack of metadata that carry out the historical information of stations. Currently Türkiye has nearly 270 actively working principal climate stations which are spread out the whole area of the country but the meteorological observations before the year 1950 does not supplied to researchers by TSMS due to undone quality control work of the data. According to descriptive study of Asar et al., (2010) on available Türkiye meteorological database, these stations have nearly 50% missingness between 1950-1960 period and percentage of missingness gradually decrease after year 1960. Locations of some actively working stations were changed during the years and several equipments of stations were renewed or changed also. Realization of such changes occurred on climate stations will causes the problem of homogeneity of the stations' observations. Homogeneity of observations in meteorological variables refers to some artificial changes on climate observations rather than the climatic changes.

Thus, due to lack of metadata homogeneity assessments of stations have to be done before starting the any climatology research.

Conducting reliable statistical analyses in order to determine long term behavior of climate conditions for an interested basin, existence of the long period of observations of some important meteorological variables such as precipitation and temperature are crucial. Within this study we compare the performances of six imputation methods which are widely used in missing data literature. Single Arithmetic Average (SAA), Normal Ratio (NR) and NR Weighted with Correlations (NRWC) are the three simple methods used. On the other hand, computationally complex model based methods Multi Layer Perceptron type Neural Network (MLPNN) and Monte Carlo Markov Chain based on Expectation-Maximization Algorithm (EM-MCMC) are used. In addition to these, we propose a modification in the EM - MCMC method in which results of simple imputation methods are used as auxiliary series.

In this study, we emphasize the importance of spatio-temporal assessment of imputed time series. Therefore, beside the coefficient of variation root mean square (CVRMSE) measure we propose Correlation Dimension (CD) technique for performance evaluation of imputation methods. CD technique is an important branch of Nonlinear Dynamic Time Series Analysis (NDTSA) and it is widely used in literature for determining the nature of underlying dynamic of a time series. We consider that CD technique can be good evaluation criteria under well establishment of phase space. Large deviations from the CD of original series will carry out the some quantitative information of imputation performance. Unlike the autocorrelation function and NRMSE measure, CD technique is more sensitive to small changes in the series so we think that the disturbance affects of imputation method on time series imputation can be detected more precisely by using the CD technique.

In order to see the performance of imputation techniques we studied monthly total precipitation and monthly mean temperature series of Türkiye in the spanning period 1965-2006. Data are obtained from the stations of TSMS. Due to spatial variability of meteorological items that are under consideration, we choose 7 different basins from 7 physical geography regions of Türkiye called South Eastern Anatolia (SEA), Eastern Anatolia (EA), Mediterranean Region (MR), Central Anatolia (CA), Aegean Region (AR), Western Thrace (WT) and North West Blacksea (NWB) which reflect also different climate regimes. Selected stations on each region are homogeneous with respect to studies of Karabörk et al., (2007), Göktürk et al., (2008) and Şahin and Cigizoğlu (2010). Stations contain no missing data for the studied period and we chose a target and reference stations for each of different basin. Target stations are selected to be in the center of the reference stations while the reference ones are selected among the ones which are highly correlated with the target station's series. Encountered missingness pattern in Türkiye Meteorological database is generally long term missingness pattern such as three, four or more than six years missingness. Therefore, three different missingness periods with 10%, 20% and 50% are created artificially on the target stations owing to comparing performances of used methods on the short and the long term missingness.

For three distinct missingness scenario, imputations from EM-MCMC method have seemed to be more consistent when compared to the other methods and in case of 10% missingness for Ağrı, Akçakale and Alanya Basins imputations from MLPNN method were more close to the real values. NR method for Bartın basin that is located at NWC of Türkiye was give slightly better results than other methods. Actually, all methods gave very near results. Bartın basin receives high amount of precipitation throughout the all year and the spatial variability of precipitation is also high. Annual precipitation rate is nearly 1000 kg per $m^2$. Finding suitable reference station in this basin and for all of the Black Sea region quite problematic. Naturally, extreme precpitation amounts could not be catched by all methods however deviations from original CD were not so dramatic. When we look at the results from Konya Basin NR method performs better for 10% and 50% missingness than other methods and EM- MCMC is better for 20% percent missingness. We connect these inconsistent results to the inappropriate selection of some reference stations. Although correlations are high for selected reference stations, total amount of annual precipitation is varying a lot for Akşehir and Beyşehir reference stations.  Konya basin has nearly 320 kg precipitation rate annually but reference stations Akşehir and Beyşehir receives 572 kg and 490 kg amount of annual precipitation respectively.  These variations of reference stations with respect to total amount of precipitation cause inconsistent imputations from MLPNN and EM – MCMC.

We have seen that if the correlations of stations were high and the total amount of precipitations were near for each of station than MLPNN and EM – MCMC methods give better results. Except for Konya Basin results from EM – MCMC and MLPNN for other basins stay consistent and CDs were not diverse so much. This situation gives us information about selection criteria of reference stations. Selected references must be correlated with target stations and in addition to this they have to be similar monthly total precipitation amounts.

In case of suitable selection of reference stations, imputations will be more reliable from EM – MCMC and MLPNN. Selection procedure of reference stations is important as imputation procedure used. For short term missingness MLPNN and EM – MCMC can be used and for long term missingness EM – MCMC method is more suitable than other imputation methods with respect to considered performance evaluation criteria. Using imputation results from simple methods as auxiliary reference stations in EM – MCMC algorithm will improve imputation performance slightly if it is not possible to find more than two appropriate reference stations. We name this procedure as EM – MCMC 2 in the study and we use abbreviation of EM –MCMC 1 for implementation of the algorithm with only available reference stations. Results from EM – MCMC 2 were almost similar to EM – MCMC 1 or better than EM – MCMC 1 in some basins. We think that using more appropriate reference stations will improve EM – MCMC 2 results. As a result, we think that the using missing data imputation from EM – MCMC algorithm for statistical analyses of meteorological data to detect long term change in climate regimes and using it for any climatological study will give opportunity to decrease amount of uncertainty.

For further analyses to improve imputation performances we will consider spatio-temporal sensitive methods deeply for missing data imputation in meteorological database. Finding more reliable selection procedure and improving selection procedure of reference stations are also important for further analyses. Empirical studies for handling missing data in meteorology may expanded with containing more than 7 basins with using more than two meteorological variables in analyses. For further analyses, we consider that inferences on imputation results have to be done with an interdisciplinary framework. Considering all these together, we finally aim to construct web site that shares completed meteorological database of Türkiye imputed with different methods and with discussions.

# REFERENCES

Aikl, L.E. and Zainuddin, Z. (2008). A comparative study of missing value estimation methods: Which method performs better. *2008 International Conference on Electronic Design, ICED 2008* , art. no. 4786656

Allison, P.D. (2001). *Missing Data.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-136. Thousand Oaks, CA: Sage.

Aly, A., Pathak, C., Teegavarapu, R. S. V., Ahlquist, J., and Fuelberg, H. (2009). Evaluation of improvised spatial interpolation methods for infilling missing precipitation records. Paper presented at the *, 342* 5914-5923.

Asar, Ö., Kartal, E., Aslan, S., Öztürk, M.Z., Yozgatlıgil, C., Çınar, İ., Batmaz, İ., Köksal, G., Türkeş, M., and Tatlı, H., (2010). Descriptive Analysis of Turkish Precipitation data with Data Mining Methods. *Proceedings of the 7th National Symposium of Statistics Days, Ankara, Middle East Technical University.*

Aslan, S., Yozgatlıgil, C., İyigün, C., Batmaz, İ., Türkeş, M., and Tatlı, H., (2010). Comparison of Missing Value Imputation Methods for Turkish Monthly Total Precipitation Data. *Proceedings of the 9th International Computer Data Analysis and Modeling: Complex Stochastic Data and Systems Conference, Minsk, Belarus State University, Vol(2) p137-141.*

Bishop, C. M. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press, New York, USA

Cano, S., and Andreu, J. (2010). Using multiple imputation to simulate time series: A proposal to solve the distance effect. *WSEAS Transactions on Computers, 9*(7), 768-777.

Chiewchanwattana, S., and Lursinsap, C. (2002). FI-GEM networks for incomplete time-series prediction. Paper presented at the Proceedings of the International Joint Conference on Neural Networks , *2* 1757-1762.

Chiewchanwattana, S., Lursinsap, C., and Henry Chu, C. (2007). Imputing incomplete time-series data based on varied-window similarity measure of data sequences. *Pattern Recognition Letters, 28*(9), 1091-1103.

Choong, M.K., Charbit, M., Yan, H. (2009). Autoregressive-model-based missing value estimation for DNA microarray time series data. *IEEE Transactions on Information Technology in Biomedicine,* 13 (1), pp. 131-137

Coulibaly P., Evora N.D., (2007). Comparison of neural network methods for in filling missing daily weather records. *Journal of Hydrology*, Vol. 341, pp. 27-41.

Daniels, M.J. and Hogan, J.W. (2008). *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis.* ChapmanandHall (CRC Press).

Dastorani, M. T., Moghadamnia, A., Piri, J., and Rico-Ramirez, M. (2010). Application of ANN and ANFIS models for reconstructing missing flow data. *Environmental Monitoring and Assessment, 166*(1-4), 421-434.

Demirtas, H., Freels, S.A., Yucel, R.M., (2008). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation*, 78 (1), pp. 69-84

Dempster A.P., Laird N.M., Rubin D.B., (1977). Maximum likelihood from incomplete data via the EM algorithm**.** *Journal of the Royal Statistical Society .B.*, **39**, pp. 1-38.

Dobesch, H., Dumolard, P., Dyras, I. (2009) *Spatial Interpolation for Climate Data The Use of GIS in Climatology and Meteorology*.

Enders, C. (2010). *Applied Missing Data Analysis.* 1st edition. The Guilford Press.

Eischeid, J. K., Pasteris, P. A., Diaz, H. F., Plantico, M. S., and Lott, N. J. (2000). Creating a serially complete, national daily time series of temperature and precipitation for the western united states. *Journal of Applied Meteorology, 39*(9), 1580-1591.

Evrendilek, F., and Berberoglu, S. (2008). Quantifying spatial patterns of bioclimatic zones and controls in turkey. *Theoretical and Applied Climatology, 91*(1-4), 35-50.

Figueroa García, J. C., Kalenatic, D., and Lopez Bello, C. A. (2008). *Missing data imputation in time series by evolutionary algorithms Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 5227 LNAI, pp. 275-283

Fraser, A. M. and Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information. *Phys. Rev.* A, 33, 1134

French M.N., Krajewski W.F., Cuykendall R.R., (1992). Rainfall forecasting in space and time using a neural network. *Journal of Hydrology*,Vol. 131, pp. 1–31.

Gardner, M. W., and Dorling, S. R. (1998). Artificial neural networks (the multiplayer perceptron)--a review of applications in the atmospheric sciences. Atmospheric Environment, 32, 2627-2636.

Gilks, W.R., Richardson, S., Spiegelhalter, D. (1996). *Markov Chain Monte Carlo in Practice: Interdisciplinary Statistics*.Chapman and Hall/CRC Interdisciplinary Statistics

Göktürk O.M., Bozkurt D., Şen Ö.L., Karaca M., (2008). Quality Control and Homogeneity of Turkish Precipitation Data. *Hydrological Processes*,Vol. 22., pp. 3210-3218

Govindaraju, R. S. (2000a). Artificial neural networks in hydrology. I: Preliminary concepts. *Journal of Hydrologic Engineering, 5*(2), 115-123.

Govindaraju, R. S. (2000b). Artificial neural networks in hydrology. II: Hydrologic applications. *Journal of Hydrologic Engineering, 5*(2), 124-137.

Haykin S., (1999) *Neural Networks: A Comprehensive Foundation*. 2nd Edition, Prentice-Hall

Hopke, P. K., Liu, C., and Rubin, D. B. (2001). Multiple imputation for multivariate data with missing and below-threshold measurements: Time-series concentrations of pollutants in the arctic. *Biometrics, 57*(1), 22-33.

Hu, J., Li, H., Waterman, M. S., and Zhou, X. J. (2006). Integrative missing value estimation for microarray data. *BMC Bioinformatics, 7*

Hyndman, R. J., and Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679-688.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment, 38*(18), 2895-2907.

Kadioğlu, M. (2000). Regional variability of seasonal precipitation over turkey. *International Journal of Climatology, 20*(14), 1743-1760.

Kalteh, A. M., and Berndtsson, R. (2007). Interpolating monthly precipitation by self-organizing map (SOM) and multilayer perceptron (MLP). *Hydrological Sciences Journal, 52*(2), 305-317.

Kalteh, A. M., and Hjorth, P. (2009). Imputation of missing values in a precipitation-runoff process database. *Hydrology Research, 40*(4), 420-432.

Kantz, H., Schriber, T. (2003). *Nonlinear Time Series Analysis*. 2nd Edition. Cambridge University Press.

Karabörk, M. C., Kahya, E., and Kömüşçü, A. U. (2007). Analysis of turkish precipitation data: Homogeneity and the southern oscillation forcings on frequency distributions. *Hydrological Processes, 21*(23), 3203-3210.

Kennel, M. B., Brown, R. and Abarbanel, H. D. I. (1992). Determining embedding dimension for phase-space reconstruction using a geometrical construction. Phys. Rev. A, 45, 3403. Reprinted in [Ott et al. (1994)].

Laird, N. M. (1988). Missing data in longitudinal studies. *Statistics in Medicine, 7*(1-2), 305-315.

Li, Y., Parker, L.E. (2008). A spatial-temporal imputation technique for classification with missing data in a wireless sensor network. *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS* , art. no. 4650774, pp. 3272-3279

Lin, T.H., (2010). A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. *Quality and Quantity*,44 (2), pp. 277-287.

Little R.J.A., Rubin D.B., (1987) *Statistical Analysis with Missing Data*. Chichester: Wiley.

Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data*. 2nd Edition. Chichester: Wiley

Lo Presti, R., Barca, E., Passarella, G. (2010). A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environmental Monitoring and Assessment* ,160 (1-4), pp. 1-22.

Lucio, P. S., Conde, F. C., Cavalcanti, I. F. A., Serrano, A. I., Ramos, A. M., and Cardoso, A. O. (2007). Spatiotemporal monthly rainfall reconstruction via artificial neural network - case study: South of brazil. *Advances in Geosciences, 10*, 67-76.

Makhuvha, T., Pegram, G., Sparks, R., and Zucchini, W. (1997). Patching rainfall data using regression methods. 1 best subset selection, EM and pseudo-EM methods: Theory. *Journal of Hydrology, 198*(1-4), 289-307.

Makhuvha, T., Pegram, G., Sparks, R., and Zucchini, W. (1997). Patching rainfall data using regression methods. 2. comparisons of accuracy, bias and efficiency. *Journal of Hydrology, 198*(1-4), 308-318.

Marwala, T. (2009). *Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques.* Information Science Reference, USA.

McKnight E.P., McKnight M.K., Sidani S., Figueredo J.A., (2007) *Missing Data*. The Guilford Press

McLachlan G., Krishnan T., (1997) *The EM Algorithm and Extension*. New York: Wiley

Moffat, A. M., Papale, D., Reichstein, M., Hollinger, D. Y., Richardson, A. D., Barr, A. G., et al. (2007). Comprehensive comparison of gap-filling techniques for eddy covariance net carbon fluxes. *Agricultural and Forest Meteorology, 147*(3-4), 209-232.

Musil, C. M., Warner, C.B., Yobas, P.K. Jones, S.L. (2002) A Comparison of Imputation Techniques for Handling Missing Data. *West J Nurs Res*, 24**:**815-829.

Nelwamondo, F. V., and Marwala, T. (2007). *Handling missing data from heteroskedastic and nonstationary data. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 4491 LNCS (PART 1), pp. 1293-1302

Nelwamondo, F.V., Golding, D., Marwala, T. (2009). A dynamic programming approach to missing data estimation using neural network. *Information Sciences* ,in press.

Nelwamondo, F.V., Mohamed, S., Marwala, T. (2007). A comparison of neural network and expectation maximization techniques. *Current Science* 93 (11), pp. 1514-1521.
noise and to determine embedding parameters. *Phys. Rev*. A, 46, 3111.

Grassberger, P. and Procaccia, I. (1983).Measuring the strangeness of strange attractors. *Physica D*, 9,189-208

Patterson, D.W. (1996). *Artificial Neural Networks*. Prentice Hall.

Paulhus, J.L.H. and Kohler, M.A. (1952). Interpolation of missing precipitation records. *Mon. Weather Rev.* 80, pp. 129–133.

Pegram, G. (1997). Patching rainfall data using regression methods. 3. grouping, patching and outlier detection. *Journal of Hydrology, 198*(1-4), 319-334.

Peterson, T. C., Vose, R., Schmoyer, R., and Razuvaëv, V. (1998). Global historical climatology network (GHCN) quality control of monthly temperature data. *International Journal of Climatology, 18*(11), 1169-1179.

Plaia, A., and Bondì, A. L. (2006). Single imputation method of missing values in environmental pollution data sets. *Atmospheric Environment, 40*(38), 7316-7330.

Potter, T.D. and Colman, B.R. (2003). *Handbook of Weather, Climate, and Water Dynamics, Climate, Physical Meteorology, Weather Systems, and Measurements.* Wiley-Interscience

Ramos-Calzado P., Gómez-Camacho J., Pérez-Bernal F., F. Pita-López M., (2008) A novel approach to precipitation series completion in climatological datasets: Application to Andalusia. *International Journal of Climatology*, Vol. 28 (11), pp. 1525-1534 *Rev. A*, 45, 3403. Reprinted in [Ott et al. (1994)].

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Hoboken, NJ: Wiley.

Sahin, S., and Cigizoglu, H. K. (2010). Homogeneity analysis of turkish meteorological data set. *Hydrological Processes, 24*(8), 981-992.

Schafer J.L.,(1997) *Analysis of Incomplete Multivariate Data*. London: Chapman and Hall / CRC Press.

Schafer. J. L. and Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7 (2), 147-177.

Schneider T., (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate,*Vol. 14, pp. 853–871.

Sehgal, M. S. B., Gondal, I., and Dooley, L. S. (2005). Collateral missing value imputation: A new robust missing value estimation algorithm for microarray data. *Bioinformatics, 21*(10), 2417-2423.

Şen, Z., and Habib, Z. (2000). Spatial analysis of monthly precipitation in turkey. *Theoretical and Applied Climatology, 67*(1-2), 81-96.

Şen, Z., and Habib, Z. (2001a). Monthly spatial rainfall correlation functions and interpretations for turkey. [Fonctions mensuelles de corrélation spatiale de la pluie et interprétations en Turquie] *Hydrological Sciences Journal, 46*(4), 525-535.

Şen, Z., and Habib, Z. (2001b). Spatial rainfall pattern identification by optimum interpolation technique and application for turkey. *Nordic Hydrology, 32*(2), 85-98.

Small, M. (2005). *Applied Nonlinear Time Series Analysis: Applications in Physics, Physiology and Finance*. Nonlinear Science Series A, World Scientific.vol 52.

Smith, K. W., and Aretxabaleta, A. L. (2007). Expectation-maximization analysis of spatial time series. *Nonlinear Processes in Geophysics, 14*(1), 73-77.

Sorjamaa A., Lendasse A., Cornet Y., Deleersnijder Eric. (2009). An improved methodology for filling missing values in spatio-temporal climate dataset - Application to Tanganyika Lake dataset. *Computational Geosciences,* inpress.

Takens, F. (1981). *Detecting Strange Attractors in Turbulence*. Lecture Notes in Math. Vol. 898, Springer, New York.

Tan , M.T., Tian, G.L., Ng, K.W. (2010). *Bayesian Missing Data Problems.* Chapman and Hall/CRC Press.

Tanner, W. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions (Springer Series in Statistics).* Springer; 3rd edition 220 pages.

Tayanç, M., Im, U., Doğruel, M., and Karaca, M. (2009). Climate change in turkey for the last half century. *Climatic Change, 94*(3-4), 483-502.

Teegavarapu, R. S. V., and Chandramouli, V. (2005). Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology, 312*(1-4), 191-206.

Toth E., Brath A., Montanari A., (2000). Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of Hydrology,*Vol. 239, pp. 132–147

Toth, E., Brath, A., and Montanari, A. (2000). Comparison of short-term rainfall prediction models for real-time flood forecasting. *Journal of Hydrology, 239*(1-4), 132-147.

Türkeş, M. (1996). Meteorological drought in Turkey: A historical perspective, 1930-1993, *Drought Network News*, University of Nebraska, 8, 17-21.

Türkeş, M. (1998). Influence of geopotential heights, cyclone frequency and southern oscillation on rainfall variations in Turkey. *Int. Jor. of Climatology*, 18, 649-680.

Türkeş, M. (1999). Vulnerability of Turkey to desertification with respect to precipitation and aridity conditions. *Turkish Journal of Engineering and Environmental Sciences*, 23, 363-380.

Türkeş, M. ve Erlat, E. (2008). Influence of the Arctic Oscillation on the Variability of Winter Mean Temperatures in Turkey. *Theoretical and Applied Climatology*, 92, 75-85.

Türkeş, M., and Erlat, E. (2003). Precipitation changes and variability in turkey linked to the north atlantic oscillation during the period 1930-2000. *International Journal of Climatology, 23*(14), 1771-1796.

Türkeş, M., and Erlat, E. (2005). Climatological responses of winter precipitation in turkey to variability of the north atlantic oscillation during the period 1930-2001. *Theoretical and Applied Climatology, 81*(1-2), 45-69.

Türkeş, M., and Erlat, E. (2005). Climatological responses of winter precipitation in turkey to variability of the north atlantic oscillation during the period 1930-2001. *Theoretical and Applied Climatology, 81*(1-2), 45-69.

Türkeş, M., and Erlat, E. (2009). Winter mean temperature variability in turkey associated with the north atlantic oscillation. *Meteorology and Atmospheric Physics, 105*(3-4), 211-225.

Türkeş, M., Koç, T., and Sariş, F. (2009). Spatiotemporal variability of precipitation total series over turkey. *International Journal of Climatology, 29*(8), 1056-1074.

Tuikkala, J., Elo, L. L., Nevalainen, O. S., and Aittokallio, T. (2008). Missing value imputation improves clustering and interpretation of gene expression microarray data. *BMC Bioinformatics, 9*

Türkeş, M. (1996). Spatial and temporal analysis of annual rainfall variations in turkey. *International Journal of Climatology, 16*(9), 1057-1076.

Türkeş, M. (1999). Vulnerability of turkey to desertification with respect to precipitation and aridity conditions. *Turkish Journal of Engineering and Environmental Sciences, 23*(5), 363-380.

Türkeş, M., and Sümer, U. M. (2004). Spatial and temporal patterns of trends and variability in diurnal temperature ranges of turkey. *Theoretical and Applied Climatology, 77*(3-4), 195-227.

Türkeş, M., Sümer, U. M., and Demir, I. (2002). Re-evaluation of trends and changes in mean, maximum and minimum temperatures of turkey for the period 1929-1999. *International Journal of Climatology, 22*(8), 947-977.

Türkeş, M., Sumer, U. M., and Kilic, G. (1995). Variations and trends in annual mean air temperatures in turkey with respect to climatic variability. *International Journal of Climatology, 15*(5), 557-569.

Türkeş, M., Sümer, U. M., and Kiliç, G. (2002). Persistence and periodicity in the precipitation series of turkey and. *Climate Research, 21*(1), 59-81.

Unal, Y., Kindap, T., and Karaca, M. (2003). Redefining the climate zones of turkey using cluster analysis. *International Journal of Climatology, 23*(9), 1045-1055.

Von Hann, J. (2009). *Handbook of Climatology*. BiblioLife

Weerasinghe, S. (2009). A missing values imputation method for time series data: an efficient method to investigate the health effects of sulphur dioxide levels. *Environmetrics* ,21 (2), pp. 162-172.

WMO., (1983). *Guide to Climatological Practices*., 2nd edn. Word Meteorological Organization: WMO no 100. Secretariat of the World Meteorological Organization: Geneva.

Xia Y., Fabian P., Stohl A., Winterhalter M., (1999a). Forest climatology: Estimation of missing values for Bavaria Germany. *Agricultural and Forest Meteorology*, Vol. 96 (1-3), pp. 131-144.

Xia, Y., Fabian, P., Stohl, A., and Winterhalter, M. (1999b). Forest climatology: Reconstruction of mean climatological data for bavaria, germany. *Agricultural and Forest Meteorology*, *96*(1-3), 117-129.

Young K.C., (1992). A three-way model for interpolating for monthly precipitation values. *Monthly Weather Review*,Vol. 120**.,** pp**.** 2562–2569.

Yucel R.M., He Y., Zaslavsky A.M. (2008). Using calibration to improve rounding in imputation*.* *American Statistician*, Vol. 62 (2)., pp. 125-129.