PERFORMANCE COMPARISON OF FILTERING METHODS ON MODELLING AND FORECASTING TOTAL PRECIPITATION AMOUNT

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

BY

ECEM ÜNAL

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

JUNE 2019

Approval of the thesis:

PERFORMANCE COMPARISON OF FILTERING METHODS ON MODELLING AND FORECASTING TOTAL PRECIPITATION AMOUNT

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

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

Prof. Dr. Ayşen Dener Akkaya
Head of Department, **Statistics** ________________

Assoc. Prof. Dr. Ceylan Yozgatlıgil
Supervisor, **Statistics, METU** ________________

Assist. Prof. Dr. Serdar Neslihanoğlu
Co-Supervisor, **Statistics, Eskişehir Osmangazi University** ________________

Examining Committee Members:

Prof. Dr. İnci Batmaz
Statistics, METU ________________

Assoc. Prof. Dr. Ceylan Yozgatlıgil
Statistics, METU ________________

Assist. Prof. Dr. Serdar Neslihanoğlu
Statistics, Eskişehir Osmangazi University ________________

Assist. Prof. Dr. Fulya Gökalp Yavuz
Statistics, METU ________________

Assist. Prof. Dr. Ceyda Yazıcı
Mathematics, TED University ________________

Date: 28.06.2019

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, Surname: Ecem Ünal

Signature:

ABSTRACT

PERFORMANCE COMPARISON OF FILTERING METHODS ON MODELLING AND FORECASTING TOTAL PRECIPITATION AMOUNT

Ünal, Ecem
Master of Science, Statistics
Supervisor: Assoc. Prof. Dr. Ceylan Yozgatlıgil
Co-Supervisor: Assist. Prof. Dr. Serdar Neslihanoğlu

June 2019, 91 pages

The performance of condensed water vapour in the atmosphere observed as precipitation on the earth surface with the consequence of gravity. It is hard to observe and measure the amount and concentration of total precipitation with its all types changing over time. This difficulty can be explained by the association between the changing amount of precipitation and the variability in the climate with its both causes and consequences. As a result of these factors, modelling and forecasting of monthly total precipitation series is a difficult procedure because of being highly parametrized and varied nature of data. To predict and forecast total precipitation, filtering methods are suggested as an alternative in the literature. Therefore, this study focus on the comparison of modelling and forecasting performances of different types of filtering methods on monthly total precipitation series. To do this, the Kalman Filter method is preferred in order to predict and forecast the naturally uncontrollable outcomes. The Kalman Filter is an algorithm for the estimation of the unobservable true state of the system, which is conducted by incorporation with the models of the system and noisy measurements of parameters. For this purpose, we used the monthly precipitation series of Muğla, Konya and Ordu stations from 1950 to 2010. The regions have been selected in terms of the amount of precipitation as moderate, scarce and abundant regions. The results of modelling and forecasting performance comparison will be a guide for the choice of best performing method for further work related to the precipitation.

Keywords: Hybrid Model, Kalman Filter, Precipitation, State Space Model

ÖZ

TOPLAM YAĞIŞ MİKTARININ MODELLENMESİ VE ÖNGÖRÜLERİN ELDE EDİLMESİNDE FİLTRELEME YÖNTEMLERİNİN BAŞARIM KARŞILAŞTIRMASI

Ünal, Ecem
Yüksek Lisans, İstatistik
Tez Danışmanı: Doç. Dr. Ceylan Yozgatlıgil
Ortak Tez Danışmanı: Dr. Öğr. Üyesi Serdar Neslihanoğlu

Haziran 2019, 91 sayfa

Atmosferde yoğunlaşmış halde bulunan su buharı yerçekiminin de etkisiyle yeryüzünde yağış olarak gözlemlenir. Bütün yağış türlerini de kapsayarak toplam yağış miktarını ve konsantrasyonunu gözlemlemek ve ölçmek oldukça zor bir yöntemdir. Elde edilen yağış miktarındaki değişiklik ile neden ve sonuçlarıyla birlikte iklim yapısındaki çeşitlilik arasındaki ilişki yağış miktarını ölçmeyi zorlaştıran bir etmen olarak gösterilebilir. Zamanla değişen yağış miktarını etkileyen bazı doğal faktörler nedeniyle aylık toplam yağış miktarı verisi çok parametreli ve yüksek varyanslı bir veri olduğu için doğru ve hassas modellemek zordur. Böyle bir durumda iyi öngörüleri sağlayan en iyi modele ulaşmak için bazı filtreleme yöntemlerinin kullanılması seçenek bir yol olabilir. Bu amaçla aylık toplam yağış verisinin modellenmesi ve öngörülerin elde edilmesinin sonuçlarının farklı filtreleme yöntemlerinin başarımları açısından değerlendirilmesi çalışma açısından önemli olacaktır. Çalışmanın temel amacı, aylık yağış verisini modellerken tahmin evresinde gözlemlenen belirsizliğin en aza indirgenmesidir. Bu amaca ulaşmak için tercih edilen filtreleme yöntemi Kalman Filtreleme yöntemi olmuştur. Kalman filtreleme tekniği, aylık toplam yağış miktarı verisinin model parametrelerinin çıkarımları ve sistem durum değişkenlerinin tahmininin yapılmasında kullanılan bir yöntemdir. Bu noktadan bakıldığında, Kalman filtreleme yönteminin, farklı parametrelere sahip aylık yağış verisinin modellenmesine ve tahmin edilmesine farklı bir yön verebileceği düşünülmektedir. Bu nedenle, 1950 ve 2010 yılları arasında Muğla, Konya ve Ordu istasyonlarında gözlemlenen aylık yağış verilerinin çalışmada kullanılması uygun görülmüştür. Seçilen istasyonlar alınan yağış miktarına göre ortalama yağış alan, az yağış alan ve çok yağış alan bölgeler olarak sınıflandırılmıştır. Modelleme ve tahmin etme başarımları karşılaştırıldığında, elde edilen sonuçlar uygulamada en iyi başarımı veren yöntemin seçilebilmesi için ilerideki çalışmalara ışık tutacaktır.

Anahtar Kelimeler: Durum Uzay Modeli, Hibrit Model, Kalman Filtresi, Yağış

To my family

ACKNOWLEDGEMENTS

Foremost, I would like to express my sincere gratitude to my supervisor, Assoc. Prof. Ceylan Yozgatlıgil and co-advisor Assist. Prof. Serdar Neslihanoglu for their support, patience, and understanding along with this study. They always guided me in the right direction to make my thesis stronger. For the whole time, with all the unlimited support and assistance, they were always more than a thesis advisor for me.

Besides my advisors, I would like to present my special thanks to my thesis evaluation jury members, Prof. İnci Batmaz, Assist. Prof. Fulya Gökalp Yavuz and Assist. Prof. Ceyda Yazıcı, for their time and attention. Without their participation, my thesis could not have been successfully finalized.

I am also very grateful for the endless support of my colleagues from International Students Office Aysun Güneş, Erdem Ceydilek, Toygun Karahasanoğlu, and Tülay Akal and also Celal Oğuz Özdemir.

My sincere thanks also go to my brother-in-law Anıl Kılınç. He always encouraged me to work with a big ambition and convinced me that I would do my best in this study.

Last but not least, I am always indebted to my family for their endless and unrequited support and love. Especially my lovely twin-sister, Didem Ünal, lived that hard moments together with me and never left me alone in this challenging journey. She never gave up believing in me not just for the time during this study but also for our whole lives. Finally, I would like to thank my mother and father Kezban and Çetin Ünal. I am aware that they always backed me up, and they will always be next to me where I am. They grew up me, and without them, I could not be able to do any piece of this study. I will be owed to my family for my whole life.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

One of the most common problems of the world is the remarkable changes observed in the climate. The effects of these changes on earth and human-beings cannot be ignored in the sense of creating a habitable future. There are lots of parameters listed which are changing the usual structure of climate, day by day. One of the most known is the global warming which has effect on climate, directly. Especially, instability in the elapsed time for the passing period of climate makes the earth out of balance. The seasons take shorter or longer times than the previous years, relatively. In this circumstances, to expect the usual returns of the seasons lose its meaning at all. At this point, some issues come to the scene such as warmer weather conditions than the expected in winter months, more arid summers, the amount of precipitation in every type less or sometimes more than expected. There may be no solution to make the warmer or colder weather as they should be in practice but if the amount of precipitation is predicted, it will help making lives easier and the world more livable. The plans of these type future events are crucial in the currrent life circle. The precipitation mentioned here is not just consisting of rain but also snow, hail, dew, rime etc. The prediction of such kind of unseasonable and changeable things in the earth with an unstable climate requires some struggle. Especially, when the global warming has combined with the variability in the nature itself, predicting the amount of precipitation with its all types will be a tough process. However, the best prediction has come up with gains for the earth especially for the countries. A good mechanism for the prediction of precipitation will be a guide for agricultural actions, give a way to the engineering activities, enhance the river basin hydrology and water resources, develop even the countries' economics. To be able to predict the amount of precipitation with all the effective parameters on it is needed to make some arrangements or editing on agricultural activities, plan engineering processes or be prepared for the conditions caused by the severe amount of precipitation. Having a good prediction model of the amount of precipitation for the area of agriculture is very important. It provides to make innovations on the existent irrigation systems. How these innovations carried out can be exampled as making new methods for the application procedures of irrigation systems or changing the design of the systems according to the regime of the precipitation in that region. Apart from the assistive side of prediction and forecast of the amount of precipitation for the irrigation systems, it helps planning the seeding and cropping mechanism according to the updated information of precipitation. The regions that are predicted to get much more precipitation may be chosen to grow highly precipitation resistant seeds. Likewise, a seeding which needs a little water to grow up should be planted in regions that receive a little precipitation. These seeding and cropping actions made according to the amount of receiving precipitation in that regions profit from the time and man power that are wasted by wrong seeding system. Knowledge of the amount of precipitation also enables selecting the right agricultural equipment to handle the severe precipitation types (Keefer, 2003). Indeed, the importance of predicting and forecasting the amount of precipitation for the agricultural activities is related to the power of production. Having a perfect prediction means a strong agricultural system which improves the countries' economies from every respects. A country with a strong agricultural economy can raise its own products in proper areas which makes the country powerful in exportation. This directly saves the country from being dependent on importation.

The significance of predicting the amount of precipitation also shows itself in the area of energy. The most popular field of energy associated with the precipitation is hydroelectric power which is a source of electricity. The generation for this type of electricity is processed in large dams (Harting, 2010). In accordance, the precipitation zones are taken into consideration while selecting the right and fertile areas for the construction of plants to produce the energy from the hydroelectrical power. Large dams should be constructed in regions being taking precipitation efficiently so that the maximum energy is reached. The plants are not constructed only in the precipitation zones but also in the drainage basins. There is also a relation between the amount of precipitation and the basins in such a way that the river basins are aimed to constitute in the regions that are known to receive more precipitation. The reason of construction these basins in such regions also prevent the freshet cases which result in due to the severe amount of precipitation. As a result, with the prediction of the amount of precipitation, proper drainage basins are set up and with the source of water gained from the river and by the amount of precipitation naturally, the hydroelectrical energy is produced.

At this point, it should not be ignored that a good prediction of the amount of precipitation also prevents the huge natural disasters. People can take precautions with the prior knowledge of the extreme amount of precipitation.

Another contribution of predicting the precipitation can be considered on tourism areas. People make their vacation plans in accordance with the meteorological estimations. A good prediction of the amount of precipitation is significant as much as a good estimation for weather. The good prediction for precipitation beneficial for the vacationist which results in an increase in the number plannings of holiday. It will have a direct effect on the economy of the country.

As it was clearly understood, a good prediction mechanism for the precipitation has comprehensive effects on daily life and long run plannings for countries with different meanings. Hence, this study is commenced in order to give a direction for solving the problems of all these areas. Since the precipitation has come up with various parameters in it, it will be important to decompose that parameters from the precipitation itself. It means that the significance of the predicting the amount of precipitation is coming from the idea that all the parameters affecting the observed amount of precipitation should be taken into account and the prediction mechanism should solve the problem with all of these parameters. While the useful parts of these parameters are processed in the prediction algorithm, the impractical parts are leaved out the mechanism by filtering them. In this study, it is tried to find the best prediction model for the amount of precipitation with a filtering method. To use the filtering method while predicting the amount of precipitation is a rare method used in the literature. One of these filtering methods is a Kalman Filter method. It is a well-known method especially in predicting the location of devices but as far as our knowledge, it is almost an untried method in the area of predicting the amount of precipitation. The working mechanism of the Kalman Filter for the prediction starts with the prediction and then the filtering step follows the prediction. After that the smoothing part is applied as a last step since the Kalman Filter is taught to be given very accurate results in the smoothing. In this study, the precipitation which is predicted by the Kalman Filter is the monthly total precipitation amount. Three different stations in Turkey are preferred for the application of the Kalman Filter method which are chosen in terms of getting average, low and high amount of precipitation listed as Muğla, Konya and Ordu, respectively. The main purpose of doing these three applications is to observe the performance of the Kalman Filter for different situations and whether the accuracy of the results of the method will be affected by the amount of precipitation or not. After seeing the results of the Kalman Filter method on the training of the precipitation model, forecasting is processed for each station to compare the forecast values being predicted by the Kalman Filter with the original values of precipitation in the test set. This study intends to succeed in prediction of precipitation with Kalman Filter. Our motivation by applying Kalman Filter to the precipitation series is to predict the total amount of monthly precipitation while it is affected by unobservable variables considered as parameters of nature. The aim is to reach the most accurate values of total amount of precipitation by filtering the noisy measurements. The modelling of the total amount of precipitation has some struggles because of having variety series itself. The most common problem encountered in the process of predicting the amount of precipitation has been stated as getting high Mean Squared Error (MSE) values. In this study, we have suggested a method by implementing the Kalman Filter to overcome the issue of having high MSE values while modelling the precipitation. Hence, the novelty tried to be proposed with this study is to gain a new perspective to the studies on precipitation modelling by using a filtering method via Kalman Filter. This study also shows the performance comparison of the standard OLS, ARIMA and the Kalman Filter estimation for the prediction of the precipitation amount. This comparison will be a guide for the future works from the aspect of Kalman Filter performance.

Following this chapter, a literature review about the Kalman Filter and the precipitation prediction models has been conducted in Chapter 2. After that Chapter 3 has discussed the methodology which gives information about the models and techniques used in this study. In connection with the methodology part, the empirical analysis has been applied to the precipitation series and the results have been discussed in Chapter 4. As a final, inferences from the analysis of data have been shared and some discussions have been made for the future studies in Chapter 5.

CHAPTER 2

LITERATURE REVIEW

The history of Kalman Filter applications is not dated back to very time because Rudolf Emil Kalman has produced this method in 1960 to solve the navigation problem of Apollo project which is a human spaceflight program carried out by NASA. After Rudolf Kalman has used the method for the navigation problem, various studies have been conducted by the application of Kalman Filter (Hun *et al.*, 2016). Therefore, in the literature, the most recent applications of Kalman Filter can be seen in various areas.

Using Kalman Filter in the control of complex dynamic systems may be the leading application area. Aircraft, ships, spacecraft and satellites can be given as examples of complex dynamic systems. People generally do not have control on such dynamic systems, and the Kalman Filter predicts the outcome of these dynamic systems such as the flow of rivers during flood conditions, the trajectories of celestial bodies, or the prices of traded commodities (Grewal, 2011). Meinhold *et al.* (1983) used the Kalman Filter method for the application of tracking a satellite's orbit around the earth. Westmore *et al.* (1991) has studied for the relative position estimation based on locations taken from images obtained from an end-point mounted camera by using Kalman Filter. Farrell *et al.* (2000) implemented Kalman Filter for GPS-aided SINU system by adopting the global positioning system (GPS) as an external source to minimize the error of inertial navigation system (INS).

Nair *et al.* (2016) used Kalman Filter application in 2-D tracking of airborne vehicles. With constant velocity and deceleration, tracking of airborne vehicle is prepared to use the Kalman Filter application.

Deep *et al.* (2018) tried to estimate the position of a GPS with the Kalman Filter method. With the observed GPS measurements, the Kalman Filter method is applied to find better estimates for the position.

The studies of Kalman Filter application in the area of tracking objects or navigation are very common. Another commonly application area of Kalman Filter is the economics. Pasricha (2006) mentioned some Kalman Filter applications in economics such as a model for the Demand for International Reserves as an example of Modeling Regime Changes, Exchange Rate Risk Premia.

Menke (2012) used Kalman Filter in three different applications in economics. The first one was the estimation of the output gap time series of Brazilian economy. The second application was creating a model on the Brazilian exchange rate. The final application was about the Brazilian financial market. While the model estimations are different in the study, the principle of Kalman Filter application was the same in three of them.

Neslihanoglu (2014) used the Kalman Filter algorithm in modelling time-varying systematic covariance risk in a Two-Moment capital asset pricing model for financial time series of developed and emerging markets in both univariate and multivariate contexts.

Wu *et al.* (2016) proposed a Kalman Filter based algorithm for solving the economic dispatch problem by minimizing the cost.

Apart from these studies about tracking, navigation or economics, the Kalman Filter method was used in a very different area which is traffic management. Antoniou *et al.* (2010) published a study about the Kalman Filter applications for traffic management. They are interested in the topic of on-line calibration of traffic simulation models and formulate the real-time OD (origin-destination) estimation and prediction problem as a state–space model and solve it using a Kalman Filter algorithm.

While the applications of Kalman Filter on precipitation amount prediction, there are not so many of them exist. The studies about predicting the precipitation by using Kalman Filter method are very recent in the literature.

Asemota *et al.* (2016) conducted a study on modelling seasonal behavior of rainfall in North Nigeria. They used monthly rainfall data collected from 1981 to 2013 in order to pave the way for new agricultural plannings in North Nigeria with the state space models via the Kalman Filter. The state space models in their study are constructed as local level model with stochastic seasonal modelling and the local level model with deterministic seasonal modelling. In order to get more accurate results, they also used the Kalman Smoothing as we did in our study.

Zulfi *et al*. (2018) published a study about the development rainfall forecasting by using Kalman Filter. ARIMA and Kalman Filter methods were compared for the performance in forecasting of rainfall in their study. The in-sample data collected from 2005 to 2015 was divided into clusters using a k-means algorithm, and Kalman Filter method was applied for modelling and forecasting in each cluster. At the end, the study concluded that performance of the Kalman Filter was better than the ARIMA model for forecasting of rainfall.

Maşazade *et al.* (2019) focused on the amount of rainfall is estimated by the Kalman Filter with radar reflectivity measurements. The amount of rainfall obtained from the automatic weather observation stations was assumed to be the unknown state vector, and the radar reflectivity values were used in the measurement model. The aim for applying the Kalman Filter was to model the true rainfall amounts.

In literature, some techniques rather than Kalman Filter have been used for the prediction precipitation models.

Sigrist *et al.* (2011) proposed a study to predict the short term rainfall by using a hierarchical Bayesian model for spatio-temporal data. They used a model combining 3 different forecasts observed from past precipitation observations.

Kotowski and Kazmierczak (2013) conducted a study based on probabilistic models of maximum precipitation. They assumed the interval precipitation amounts criterion to isolate the intensive rainfalls and found maximum precipitation models for Wroclaw for the time period from 1960 to 2009.

Gaikwad *et al.* (2013) conducted a study for the prediction of precipitation models with two approaches empirical method and dynamic method, respectively.

Abdul-Aziz *et al.* (2013) published a study for observing the pattern of rainfall in Ghana with its both low and extreme variabilities by using seasonal ARIMA.

In another study, to predict the amount of rainfall in Sylhet, Bari *et al.* (2015) used seasonal ARIMA model based on Box and Jenkins method which directed them to the most effective model giving the best predictions.

Yozgatlıgil and Turkes (2018) modelled monthly maximum precipitation amounts by using a distributional and time series analysis approach in their study. They have found that the performance of time series model is better than the probabilistic approach which uses the extreme value theory.

Recently, lots of methods have been proposed to model the precipitation (Du *et al.*, 2017, Aagesen *et al.*, 2018).

Liu *et al* (2019) created a Markov chain model by using the data collected from Beijing from 1951 to 2013 to predict the amount of precipitation in 2014 and 2015.

The idea behind this study is conducting the precipitation prediction models with a different filtering technique. The main motivation of this study is the non-existence of the high performance precipitation prediction models for Turkey as far as our research. Although the number of studies are very limited in the literature until time being, with the accurate results of the Kalman Filter, it is taught to be one of the most preferred method for predicting the amount of precipitation.

CHAPTER 3

METHODOLOGY

In this chapter, some background information about the statistical methodology used in the study have been stated. Firstly, the models that are included in the study have been summarized and some specifications of state space models have been examined. After that, the working principle of Kalman Filter and smoother has been clarified with the estimation methods. Finally, a brief information has been placed about logistic regression at the end of the methodology part.

3.1. Models

The mentioned models in this part is consisting of precipitation models, state space models, logistic regression and ARIMA models, respectively.

3.1.1. Precipitation Models

The performance of condensed water vapor in the atmosphere is observed as precipitation on the earth surface with the consequence of gravity. It will be hard to observe and measure the amount and concentration of total precipitation with its all types over time. This difficulty can be explained by the association between the changing amount of precipitation and the variability in the climate with its both causes and consequences (EE *et al.*, 2017). In addition to the effects of consequences of variability in the climate on the total precipitation amount, some natural causes can be listed as factors which have effects on total precipitation. Those factors can be considered as different parameters of nature that can be both results of climate change and effects on total precipitation for a different length of time. According to these different parameters of nature the observed amount of precipitation changes. Furthermore, some different mechanisms such as the rate of humidity observed temperature or cloudiness may affect the time, duration or intensity of the precipitation.

As a result of these factors, an accurate and precise modelling of total precipitation series is a difficult procedure to achieve because of being highly parametrized and highly varied nature of data. While modelling total precipitation, it is important to take the maximum and minimum values of those parameters into account in order to get the most efficient structure of the model. To exert dominance on the different factors effective on total precipitation making easy to understand the structure of the data is important to reach the best model with good forecasts.

To model, the monthly collected precipitation data is taught to shed a light on the agricultural and engineering applications (Keefer, 2003). The variety in precipitation and its natural parameters accords with the issue of applying the right agricultural policies, making the possible energy production units and determining the settlement of dams in a specific region. Hence, being able to understand the structure and nature of precipitation with its all parameters is vital in order to establish a model and enhance it (Trenberth *et al.*, 2003).

3.1.2. State Space Models

A state space representation includes all of the cases of the interest known as a dynamic linear model (Shumway *et al.*, 2016). The non-stationary and time-varying systems are described very well by the state space model consisting of state or transition and measurement or observation equations, respectively. This part of methodology section is primarily taken from (Neslihanoglu, 2014). The state space can be considered as a tool that mixes the observed and unobserved variables. The state or transition equation includes the unobserved state variable α_t while the observation or measurement equation involves both observed variables known as measurements y_t and unobserved state variable α_t. The evolution in the unobserved state variable over time is described in the state equation. The relation between the observed variables y_t and unobserved state variable α_t is defined in the measurement equation. In general the observed variables y_t come up with an error.

The state space model with both state and measurement equations is shown in the form of equations (3.1) and (3.2) for $t = 1, \dots, n$;

$$Y_t = A_t\alpha_t + \varepsilon_t, \tag{3.1}$$

$$\alpha_t = \Phi_t\alpha_{t-1} + w_t. \tag{3.2}$$

The equation 3.1 is the measurement equation showing the dependence of observed variable y_t to the unobserved state variable α_t. The term A_t, a *(q x p)* vector, is describing how the unobserved state variable α_t, a *(p x1)* vector, is turning to the measurements y_t, a *(qx1)* vector, for each time *t*. Here, the observed variables y_t come up with an error term $\mathcal{E}_t$ followed in equation (3.3).

$$\varepsilon_t \sim N(0, H), \tag{3.3}$$

In the definition (3.3), the error term a *(q x 1)* vector is independent and identically normally distributed from *t = 1,..., n* where the variance matrix of it is a *(q x q)* matrix called as *H.*

The equation (3.2) is the state equation reflecting the relation between the unobserved state variable α at time *t* and at time *t-1*. The ϕ_t, a *(p x p)* vector is the speed or transition parameter explains the relation between the unobserved state variable at different times. The $w_{t,}$ a *(p x 1)* vector, is independent and identically distributed error term with a normal distribution shown in equation (3.4).

$$w_t \sim N(0, Q). \tag{3.4}$$

In the definition (3.4), the error term w_t is distributed normally with a zero mean and *(pxp)* vector of matrix $Q_{t.}$ The aforementioned matrices A , ϕ, H and Q are called as system matrices. It is assumed that the A matrix is known and ϕ, H and Q matrices estimated from the given data are constant over time.

The first assumption should be satisfied in the linear Gaussian state space models is that the initial state vector α_0 which is usually a random variable distributed as Gaussian with mean μ_0 and variance Σ_0 shown in equations (3.5), (3.6) and (3.7).

$$E(\alpha_0) \ = \ \mu_0, \tag{3.5}$$

$$Cov(\alpha_0) \ = E\big[\big((\alpha_0 - \mu_0)(\alpha_0 - \mu_0)^T\big)\big] \ = \ \Sigma_0, \tag{3.6}$$

$$\alpha_0 \sim N(\mu_0, \Sigma_0). \tag{3.7}$$

The second assumption should be satisfied for linear Gaussian state space models is that there is zero correlation between the error terms of measurement and state equations ε_t and w_t . To put it differently, ε_t and w_t are independent of each other (Neslihanoglu, 2014).

3.1.2.1. State Space Model Specifications

As it is mentioned before, the state space models is consisting of two equations which are measurement and transition equations. In the measurement equation, it can be seen that how the measurements are changing according to the unobserved vector of α_t over time. This measurement equation can be taught as a time-varying coefficient regression model. The time-varying parameter here is α_t. In the transition equation, the change in the unobserved vector α_t over time is shown. Different models based on the Kalman Filter idea appear to define the evolution of α_t in the state equation (Neslihanoglu, 2014).

3.1.2.1.1. Random Coefficient

In random coefficient models, the state equation contains a term $\hat{\alpha}$ which is the mean value of $\alpha_0, \alpha_1, \dots, \alpha_n$ shown as follows in equation (3.8).

$$\alpha_t = \hat{\alpha} + w_t. \tag{3.8}$$

Here, there is no correlation between the stationary sequences of $\alpha_0, \alpha_1, \dots, \alpha_n$ which is formed with constant mean and variance.

3.1.2.1.2. Random Walk

The state equation is written by assuming a first order random walk model as in equation (3.9).

$$\alpha_t = \alpha_{t-1} + w_t. \tag{3.9}$$

Here, there is autocorrelation between the nonstationary sequence of $\alpha_0, \alpha_1, \dots, \alpha_n$ where the variance of α_t is increasing by time *t*.

3.1.2.1.3. Mean Reverting

The state equation is represented by the following model to make the sequence stationary.

$$\alpha_t - \hat{\alpha} = \Phi(\alpha_{t-1} - \hat{\alpha}) + w_t. \tag{3.10}$$

The coefficients of the diagonal elements of the speed or transition matrix ϕ should be less than 1 in order to have a stationary sequence of $\alpha_0, \alpha_1, \ldots, \alpha_n$.

These three specifications will be used with the names of Kalman Filter Random Coefficient (KFRC), Kalman Filter Random Walk (KFRW) and Kalman Filter Mean Reverting (KFMR) in this study.

3.1.2.2. Estimation of State Space Model Parameter via Kalman Filter Algorithm

Kalman Filter is a recursive process that works by updating an estimate of the unobserved state variable with the consecution function of observed variables. The name of the filter is arising from a study of Rudolph E. Kalman about finding a recursive solution for the linear filtering problem of discrete data (Welch *et al.*, 1995).

The algorithm behind the Kalman Filter is to process the data with the optimal recursive (Ribeiro, 2000). The word “optimal” represents the best estimate based on the idea of minimizing the state error while estimating it (Ribeiro, 2004). The word “recursive” means that the past data has not to be stored by the Kalman Filter and by taking a new measurement it is reevaluated every time by the filter. In short, Kalman Filter uses the state vector α at time t and measurements at time $t+1$ to evaluate the state vector α at time $t+1$ (Gasana, 2012).

It is stated as the best filter since Kalman Filter minimizes the mean square error of the estimated parameters (Gasana, 2012). Therefore, the Kalman Filter is the optimal MMSE state estimator (Shimkin, 2016). The estimation process of a state of the system is performed by a filter according to the available measurements in the system (Ribeiro, 2004). It is actually a state estimation process.

In a very simple way, the working principle of the filter starts with solving the imperfect information has an error, noisy and uncertainty is consisting of initial assumptions. While taking the useful parts of information, the desired states are reached by filtering the noise and uncertainty (Rudy *et al.*, 2011). That is to say, the observed data is used in a Kalman Filter to get the optimal estimate of the system state (Susmel, 2013).

The basic idea behind the Kalman Filter is stated as follows in equations (3.11) and (3.12)

$$Y_t = A_t\alpha_t + \varepsilon_t \qquad \varepsilon_t \sim N(0, H), \tag{3.11}$$

$$\alpha_t = \Phi_t\alpha_{t-1} + w_t \qquad w_t \sim N(0, Q). \tag{3.12}$$

A_t is the observation matrix. Y_t's are the observed measurements. α_t is the unobserved state vector and ϕ is the transition or speed parameter. The measurement and transition equations have come up with error terms ε_t and w_t which are normally distributed with 0 mean and variances H and Q, respectively.

The formulations above are in the structure of the state space model because the linear state space systems are operable for the mechanism of Kalman Filter.

3.1.2.2.1. Kalman Filter Algorithm

When the filter comes to the process, the mechanism changes. If it is scheduled as a structure of steps, it will be followed as;

Step 1: The current estimate of state vector α_t and the initial values of ϕ together with the error terms $\mathcal{E}_t$ and w_t in measurement and state equations start the process. After this, the predicted estimate of α is calculated by the filter for the next time $t+1$.

Step 2: The predicted measurement value y is estimated for the next time as y_{t+1} by the filter putting the calculated α_{t+1} in the measurement equation and using the observed value y_{t+1} which is already known.

Step 3: With the observed value y_{t+1} at time $t+1$, the predicted error which is the difference between observed and predicted measurements is calculated.

Step 4: The adjustment of the α_t prediction is done by the model by allowing part of the prediction to feed through in the adjusted α_t. Then, the process is starting again from Step 1 by using the adjusted α_t as α_{t+1}.

Identification of the unknown parameters' values which minimizes the prediction error is done by solving MLE (Maximum Likelihood Estimation) recursively (Renzi-Ricci, 2016).

The algorithm is actually carried out in 3 steps stated as prediction, filtering and smoothing, respectively.

The prediction part includes the calculations of the estimate of the state vector at time t and the error covariance vector as follows in equations (3.13) and (3.14).

$$\hat{\alpha}_{t+1|t} = \hat{\Phi}_{t|t-1}\, \alpha_{t|t-1}. \tag{3.13}$$

$$\Sigma_{t+1|t} = \Phi_t \Sigma_{t|t-1} {\Phi_t}^T + Q_t. \tag{3.14}$$

In the update part, innovation and innovation covariance, Kalman gain, updated state estimate and updated error covariance are calculated.

Innovation:

$$I_t = Y_t - A_t \hat{\alpha}_{t+1|t}. \tag{3.15}$$

Innovation covariance:

$$S_t = A_t \Sigma_{t+1|t} {A_t}^T H_t. \tag{3.16}$$

Kalman Gain:

$$K_t = \Sigma_{t+1|t} {A_t}^T {S_t}^{-1}. \tag{3.17}$$

Updated state estimate:

$$\hat{\alpha}_{t+1|t+1} = \hat{\alpha}_{t+1|t} + K_t I_t. \tag{3.18}$$

Updated error covariance:

$$\Sigma_{t+1|t+1} = (I_d - K_t A_t) \Sigma_{t+1|t}. \tag{3.19}$$

Kalman Filter is a one-sided filter that predicts state vector α with using past and current values of a variable of interest y.

Kalman Smoother is a two-sided filter estimates the state vector α by using all observed values of the variable of interest y (Mikusheva, 2007).

3.1.2.2.2. Kalman Filter and Smoother Algorithm

Since the problem that Kalman Filter tries to solve is estimating the state of a process controlled by the linear stochastic difference equation, the very simple algorithm of Kalman Filter is based on the following state space model which has already stated before in section 3.1.2.2 equations (3.11) and (3.12) (Welch *et al*., 1995);

$$Y_t = A_t\alpha_t + \varepsilon_t \qquad \varepsilon_t \sim N(0, H), \tag{3.20}$$

$$\alpha_t = \Phi_t\alpha_{t-1} + w_t \qquad w_t \sim N(0, Q). \tag{3.21}$$

The process and measurement noises are represented by the random variables $\mathcal{E}_t$ and w_t which are assumed to be independent of each other, white, and normally distributed as stated in equations (3.3) and (3.4). The process noise covariance matrix H and measurement noise covariance matrix Q in Kalman Filter equations are assumed to be constant over time (Welch *et al*., 1995). Same as with the covariance matrices of the system, the A and ϕ matrices are also constant over time.

The main purpose of this analysis is stated as estimating the unobserved state variable α_t at time t with the given information of $Y_n = \{Y_1;\ Y_2; \dots; Y_n\}$ at time n. Having two different time points as t and n arises 3 different situations according to these time points. The first situation occurs when $(t > n)$ called as “prediction”. This “prediction” issue which is an apriori type of estimation tries to provide information about the quantity of interest at some time $(t + n)$ by using data available up to and including time *t-1*. The second situation is observed when $(t = n)$ called as “filtering”. Filtering issue compromises the excluding of information about a quantity of interest at time *t*, by using data available up to and including time *t*. The final situation shows up when $(t < n)$ called as “smoothing”. Smoothing part is an a posteriori type of estimation in which data available after the time of interest are used for the estimation (Neslihanoglu, 2014).

As a result, the algorithm of the Kalman Filter in our study is actually consisting of 3 steps which are prediction, filtering and smoothing, respectively.

To start the process, the conditional mean and variance of the state vector α_t up to time n are defined as following equations.

$$\alpha_t^n = E(\alpha_t|Y_n), \tag{3.22}$$

$$P_t^n = Var(\alpha_t|Y_n). \tag{3.23}$$

The forward recursion steps of the Kalman Filter and Smoother algorithm with the initial conditions of these conditional mean and variance represented by α_0^0 and P_0^0 used in the prediction and filtering processes characterized as follows (Neslihanoglu, 2014).

$$\alpha_0^0 = \mu_0, \tag{3.24}$$

$$P_0^0 = \Sigma_0. \tag{3.25}$$

The process begins with the prediction of the state α_t as in equation (3.26).

$$\alpha_{t|t-1} = \Phi\, \alpha_{t-1|t-1}, \tag{3.26}$$

Then, the state covariance matrix represented by P is updated according to the formula in equation (3.27).

$$P_{t|t-1} = \Phi\, P_{t-1|t-1} \Phi^T + Q. \tag{3.27}$$

These two statements are set in the prediction process. The term "innovation" or called as the residual which is reflecting the discrepancy between the predicted measurement $A_t\alpha_{t|t-1}$and the actual measurement Y_t and the covariance of it are set in the filtering part stated in the following equations.

Innovation:

$$I_t = Y_t - A_t\alpha_{t|t-1}, \tag{3.28}$$

Innovation covariance:

$$S_t = A_t\Sigma_{t|t-1}{A_t}^T + H. \tag{3.29}$$

Then, Kalman gain equation is stated by using the innovations created in the previous step.

Kalman Gain:

$$K_t = P_{t|t-1}{A_t}^T{S_t}^{-1}. \tag{3.30}$$

The Kalman gain term represented by K_t minimizes the a posteriori error covariance is calculated by substituting the updated state estimate shown as equation (3.31) below into the a posteriori estimate error equation and then substituting that into the updated error covariance represented by equation (3.32) (Welch *et al.*, 1995).

Updated or a posteriori state estimate:

$$\alpha_{t|t} = \alpha_{t|t-1} + K_tI_t. \tag{3.31}$$

Updated error covariance:

$$P_{t|t} = P_{t|t-1} - K_tA_tP_{t|t-1}. \tag{3.32}$$

The equations from (3.24) to (3.32) are cycled for each time t. Here, the process that includes equations (3.26) and (3.27) is the prediction part of the system. The forward recursion in equations (3.28) through (3.32) is called the Kalman Filter. Hereby, the prediction and filtering steps of the process is done.

Until this time, the past and current observations Y_t are used to predict the α_t by Kalman Filter. It is actually needed for the computation of likelihood. However, when the issue is to estimate the α_t, the whole data should be used to predict α_t. This situation is occurred when $(t < n)$ and the application of backward recursion here called as Kalman Smoother (Mikusheva, 2007).

Since the previous part ends with the updated state estimate $\alpha_{n|n}$ and updated error covariance $P_{n|n}$, the initials of the current step will be those values calculated from the Kalman Filter at time $(t = n)$. The working principle of the Kalman Smoother starts by setting the smoothed state and smoothed error variance shown in equations (3.33) and (3.34) respectively at time *t* equal to *n* and continues until *t* is equal to 1.

$$\alpha_{t-1}^{n} = \alpha_{t-1}^{t-1} + J_{t-1}(\alpha_t^n - \alpha_t^{t-1}), \tag{3.33}$$

$$P_{t-1}^{n} = P_{t-1}^{t-1} + J_{t-1}(P_t^n - P_t^{t-1}){J_{t-1}}^T, \tag{3.34}$$

$$J_{t-1} = P_{t-1}^{t-1}\Phi^T[P_t^{t-1}]^{-1}. \tag{3.35}$$

As a summary, the Kalman Filter is a recursive process which is running forward. On the other hand, the Kalman smoother is a process running backward (Mikusheva, 2007). In filtering step, the state α_t is reached recursively moving forward by keeping the values $\alpha_{t|t-1}, \alpha_{t|t}, P_{t|t-1}, P_{t|t}$ from the time *t* is equal to 1 until *n*. After that, Kalman Smoother is applied with the backwards movement until the state at time *t* which is the desired value of estimate (Neslihanoglu, 2014). Starting from $(t = n)$ and repeating the smoothing equations, $\alpha_{n|n}, \alpha_{n-1|n}, \dots, \alpha_{1|n}$ are estimated (Mikusheva, 2007).

3.1.2.2.3. Estimation of Parameters Process via Kalman Filter

In the beginning of the calculations, the system matrices *H*, ϕ and *Q* and the initial values which are the mean of the state vector μ_0 and variance of the state vector Σ_0 are assumed to be known. However; the system matrices *H*, ϕ and *Q* may depend on a vector of unknown hyper parameters stated as Θ. In such cases, the main idea is estimating those unknown parameters by maximum likelihood estimation method.

The derivation of the Maximum Likelihood function is based on the assumptions that are previously mentioned as;

- The initial value of unknown state vector α_0 is normally distributed with known mean μ_0 and variance $\Sigma_{0.}$
- The measurement error term represented by $\mathcal{E}$ has a Normal distribution with 0 mean and variance matrix *H*.
- The transition error term represented by *w* has a Normal distribution with 0 mean and variance matrix *Q*.
- There is no correlation between the error terms $\{\varepsilon_t\}$ and$\{w_t\}$.

Under the satisfied assumptions, the likelihood function represented by *L* is generated in equation (3.36).

$$L_Y(\Theta) = P(Y_1, Y_2, \dots, Y_n; \Theta) \; = \; P(Y_1; \Theta) \prod_{t=2}^{n} P(Y_t|Y_{t-1};\; \Theta)\,. \tag{3.36}$$

According to the normal procedure of Log likelihood function, the next step is taking the natural logarithm of the *L* function. The multiplication turns into summation when we take its logarithm and it is stated in equation (3.37).

$$log\big(L_Y(\Theta)\big) = \sum_{t=1}^{n} log\big(P(Y_t|Y_{t-1};\; \Theta)\big). \tag{3.37}$$

According to the state space models in equations (3.1) and (3.2), the conditional mean and the conditional variance of the Gaussian Y_t are calculated in equations (3.38) and (3.39).

$$E(Y_t|Y_{t-1};\ \Theta) \ = \ A_t \alpha_{t|t-1}, \tag{3.38}$$

$$Var(Y_t|Y_{t-1};\ \Theta) \ = \ \Sigma_t. \tag{3.39}$$

When these expectation and variance values are substituted in the conditional density of Y_t, it has the form like in equation (3.40).

$$P(Y_t|Y_{t-1};\ \Theta) = \frac{1}{2\Pi^{q}/_{2}} \ |\Sigma_t(\Theta)|^{-1/_{2}} \ exp\left(\frac{-1}{2} I_t(\Theta)^T \Sigma_t(\Theta)^{-1} I_t(\Theta)\right), \tag{3.40}$$

where it is the innovation term. When we substitute the above equation into the equation (3.37) which is the logged equation, the final presentation of the equation will be like as follows in equation (3.41).

$$log\left(L_y(\Theta)\right) = \ -\frac{nq}{2} \ log(2\Pi) - \frac{1}{2}\sum_{t=1}^{n} log|\Sigma_t(\Theta)| - \frac{1}{2}\sum_{t=1}^{n} I_t(\Theta)^T \Sigma_t(\Theta)^{-1} I_t(\Theta). \tag{3.41}$$

The update of the unknown parameter vector Θ by maximizing the log likelihood function is mostly done with Newton-Raphson method. The algorithm of the update is stated in the study of Shumway *et al.* (2006) by steps;

1) The initial values are selected for the unknown parameters vector $\Theta^{(0)}$.
2) The Kalman Filter is run by using these initial values $\Theta^{(0)}$. As like in the first step of Kalman Filter, the innovations I_t and the variance of innovations S_t are set and used to calculate the log likelihood function $log\left(L_y(\Theta)\right)$.

3) The Newton- Raphson algorithm is processed to observe the updated estimates of Θ to obtain the new estimates stated as $\Theta^{(1)}$.
4) $\Theta^{(j+1)}$, $\Theta^{(j)}$ and innovations, variance of innovations are obtained by repeating the steps 2 and 3.
5) The algorithm ends when the values of $\Theta^{(j+1)}$and $\Theta^{(j)}$ are different from each other, or the values of $L_Y(\Theta^{(j+1)})$ and $L_Y(\Theta^{(j)})$ are different from each other, by less than a predetermined small amount (Neslihanoglu, 2014).

3.1.3. Logistic Regression

In some linear regression models, the dependent variable can be classified as successes or failures. In other words, the values may come from binomial distribution. This kind of regression models are called Generalized Linear Models. The logistic regression is one type of it (Rundel, 2013). A Generalized Linear Model is constructed in equation (3.42).

$$\eta_i = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon. \tag{3.42}$$

It is made up with two functions which are a link function that shows the relation between the mean of dependent variable and the linear predictor and a variance function that shows the relation between the variance of dependent variable and the mean (Turner, 2008).

Link function:

$$g(\mu_i) = \eta_i. \tag{3.43}$$

Variance Function:

$$Var(Y_i) = \Phi Var(\mu). \tag{3.44}$$

The logistic regression is used to build models for binary categorical variables as a dependent variable. The independent variables can be numerical or categorical as well. Since the dependent variable is coming from binomial distribution, the probability of success represented by p is modeled in logistic regression (Rundel, 2013). In other words, the success probability of Y given X is modeled as (Tibshirani, 2014);

$$P(Y = 1|X) = \frac{exp(\beta x)}{1 + exp(\beta x)}. \tag{3.45}$$

Rearrange the equation (3.46);

$$log\left(\frac{P(x)}{1 - P(x)}\right) = \beta^i X. \tag{3.46}$$

The left-hand side of the equation (3.46) is called as log odds or logit of $P(x)$. In this structure, it can be said that one unit increase in the independent variable $x_i$will cause a β^ichange in the log odds while keeping other predictors fixed (Tibshirani, 2014).

If we make the equation (3.47) like;

$$\left(\frac{P(x)}{1 - P(x)}\right) = e^{\beta^i X}. \tag{3.47}$$

The interpretation of the coefficients will be made as one unit increase in the independent variable x_i changes the odds by $e^{\beta i}$ while keeping other predictors fixed (Tibshirani, 2014).

To put it simply, there is a binary response variable Y which will be modeled based on the conditional probability $P(Y = 1|X = x)$ as a function of x and the parameters are estimated by maximum likelihood estimation (Shalizi, 2019).

3.1.4. ARIMA Models

The stochastic process developing in time $t = 1, \ldots n$ create time series shown as $\{y_t\}_{-\infty}^{+\infty}$. Since they are observed sequentially in time, the observations are dependent to each other. The aim of time series analysis is to see the structure of time-dependent variables while making a forecast for the future observations (Yozgatlıgil, METU OpenCourseWare, 2011).

The first assumption satisfied in time series analysis called as stationarity means that the variability in the behavior of the values is constant over time. The auto covariance and auto correlation function (ACF) between y_t and y_{t-k} for a stationary series shown in equations (3.48) and (3.49).

$$\gamma_k = Cov(y_t, y_{t-k}) = E[(y_t - \mu)(y_{t-k} - \mu)], \tag{3.48}$$

$$\rho_k = Corr(y_t, y_{t-k}) = \frac{\gamma_k}{\gamma_0}. \tag{3.49}$$

The stationary process $\{y_t\}$ which can be written in the form of linear combination of the sequence of uncorrelated white noise called as Moving Average time series shown in equation (3.50).

$$Y_t = \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{p-1} + \varepsilon_t. \tag{3.50}$$

The different time series models which are invertible can be re-expresses by each other, then it creates Autoregressive representation of time series shown in equation (3.51).

$$Y_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{q-1}. \tag{3.51}$$

In equation (3.50), it is called moving average process with order *q* denoted by *MA(q)* and in equation (3.51), Y_t is called as autoregressive series with order *p* denoted by *AR(p).* In both equations, ε_t is white noise.

The series Y_t which is observed as a combination of both autoregressive and moving average processes called as Autoregressive Moving Average series denoted by *ARMA (p, q)* in equation (3.52).

$$Y_t = \varphi_1 Y_{t-1} + \cdots + \varphi_p Y_{p-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{q-1}. \tag{3.52}$$

The differencing is a very important term used while dealing with the non-stationary time series to make them stationary. At this point, an autoregressive moving average series which needs to be differenced to be made stationary is called as Integrated Autoregressive Moving Average series denoted by *ARIMA (p, d, q)* in which the difference is shown by *d*. The cases in which the integrated autoregressive moving-average series have had strong seasonal characteristics as showing same structure after a regular time interval called Seasonal ARIMA shown in equation (3.53).

$$\Phi(L^s)\varphi(L)(1-L)^d(1-L^s)^D Y_t = \Theta(L^s)\theta(L)\varepsilon_t. \tag{3.53}$$

It is denoted by *ARIMA (p, d, q) x (P, D, Q)*$_s$ where the seasonality is represented by *s* (Ihaka, 2005).

An ARIMA model is constructed by either applying regular time series analysis or using an automated algorithm. For the regular time series analysis, decisions of being stationary, having trend or seasonality should be made and best model should be selected according to the smallest information criterion. In the automated algorithm, *auto.arima()* function in *RStudio* suggests the best ARIMA model with the smallest information criterion (Hyndman & Athanasopoulos, 2018).

3.2. Model Evaluation Criteria

The accuracy for the out-of-sample forecasts have been measured by looking the MAE and MSE values. The MAE measures the absolute value of the differences between the forecasted and original values. The MSE operates as taking the squared root of the mean value of squared errors which treats these errors like they are large but infrequent. When the difference between MAE and MSE increases this means that the error size is more consistent. The small values of forecast evaluation criteria means that the out of sample forecasts are estimated properly and the model constructed at the end is a significant model (Neslihanoglu, 2014). The equations for the MAE and MSE criteria shown in equations (3.54) and (3.55).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_{i-}\,\widehat{y_i}\,|, \tag{3.54}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y_i})^2\,. \tag{3.55}$$

CHAPTER 4

DATA ANALYSIS

In this chapter, the empirical part of study has been conducted. After giving a brief information, the data description and application sections for each station have been mentioned, respectively.

4.1. Introduction

In this part, the monthly total precipitation amount, monthly average temperature, relative humidity, and cloudiness series from three different regions are obtained from the Turkish Meteorology Services between 1950 and 2010 to predict and forecast precipitation amount applying the Kalman Filter by using *RStudio*. These regions are abundant, moderate and scarce in terms of getting the precipitation. This selection provided us to see the performance of our applications for different cases. The models conducted are actually based on the idea of linear regression. The observed precipitation amount is the response variable in the regression model where monthly average temperature, relative humidity, and cloudiness series are chosen as predictors. Since these three independent variables may be considered as being in a close relation, there may be a doubt of multicollinearity. However, according to the results of VIF (Variance Inflation Vector) found as 2.1810, 1.3206, 2.1751 for each predictor, there is no evidence for multicollinearity. As it is mentioned in the methodology part, the regression covers prediction, filtering and smoothing, respectively. The predicted, filtered and smoothed results found for precipitation at the end are obtained by applying Kalman Filter. The aim by using Kalman Filter is to see how close the smoothed variables of precipitation observed at the end to the actual precipitation values used at the beginning while it is affected by the other independent variables. The most simple regression model is constructed in equation (4.1).

$$Y_t = \alpha + \beta_{it} X_{it} + \varepsilon_t, \qquad i = 1,2,3. \tag{4.1}$$

However, the above model is needed to be in the form of a state space model shown in equations (3.1) and (3.2) in order to apply Kalman Filter. As it is stated before there are different state space model specifications in section 3.1.2. One of which can be associated with our model is the mean reverting specification of state space model which is called a Kalman Filter Mean Reverting (KFMR) here. The mean reversion idea here can be taught as returning to the average value at the end. The idea for the measurement equation will be the same but the transition equation will converted into the form of Kalman Filter Mean Reverting followed by equations (4.2) and (4.3).

$$Y_t = \alpha + \beta_{it} X_{it} + \varepsilon_{it} \qquad \varepsilon_{it} \sim N(0, H), \tag{4.2}$$

Measurement model

$$\beta_{it} = \bar{\beta}_i + \Phi_i\left(\beta_{it} - \bar{\beta}_i\right) + w_{it} \qquad w_{it} \sim N(0, Q). \tag{4.3}$$

Transition model

Y_t is the observed values which is related with the unobserved state variables β_t. These β_t's are the independent variables in our first linear regression model. The effect of these variables on the amount of precipitation is modelled in the equation (4.2). In the transition model, which is designed in terms of the Kalman Filter Mean Reverting, how the unobserved state variables changes over time by depending on its mean value is examined. The error terms for measurement and transition equations should be normally distributed with 0 mean and variances H and Q, respectively.

Note that the whole mechanism for the application of Kalman Filter will be based on these models in this study. However, the application part is done with two different approaches.

The first series itself is used for the application to obtain the predicted, filtered and smoothed values for precipitation. Some negative values are observed for the total amount of precipitation, which is impossible, at the end of those calculations especially for smoothed values. To overcome this, it is tried to make some transformation on the series. However, the transformation did not solve the problem. Those negative values again observed. At that point, it is decided to count those negative values as zero because they were already very close to the zero. In other words, for the first application part with the actual data, the negative values obtained for the predicted, filtered and smoothed values are representing the zero values for precipitation. This means that there is no expectation for precipitation on that day.

The second data being hybrid data is created from the actual data to handle negative values observed in the smoothing part by using the logistic regression method. The below steps are followed for generating the hybrid model;

- The precipitation amount series are arranged as 0 if amount is 0, otherwise 1.
- A logistic regression model is fitted by using explanatory variables.
- The estimated 0's fixed as 0 and the estimated 1's are put in Kalman Filter procedure as precipitation amount.
- The accuracy measures and forecasts for future observations are obtained.

The algorithm of Kalman Filter is implemented for the series obtained these two approaches and final results are observed. After that, a forecast technique is carried out to see the performance of the Kalman Filter for both datasets as Rolling Window Forecast technique. The main goal is to make this forecasting method is comparing the performances of Kalman Filter estimates and Ordinary Least Square estimates and seeing which one gives better about predictions for precipitation.

4.2. Analysis for Muğla

In this part, a brief information about the structure of Muğla station data and results of application will be given.

4.2.1. Data Description for Muğla Station

The series is composed of monthly total precipitation amount of Muğla from 1950 to 2010 including three independent variables, namely average temperature, relative humidity, and cloudiness series. Because Muğla is receiving precipitation on average whole year compared to the rest of the regions in Turkey, it is also chosen as a representative of a moderate region in terms of precipitation intake. The data is provided by the Turkish Meteorological Service.

The descriptive statistics of the Muğla station data are shown in the following table;

Table 4.1. Descriptive Statistics for Muğla Station

	Precipitation	**Temperature**	**Relative Humidity**	**Cloudiness**
Minimum	0.00	2.50	29.10	0.10
1st Quartile	11.10	7.90	51.38	1.50
Median	55.85	14.10	64.80	3.50
Mean	96.72	14.97	62.47	3.39
3rd Quartile	143.40	22.20	73.83	5.00
Maximum	645.30	29.00	89.40	7.90

As figured in the Table 4.1, the minimum observation for the precipitation is 0 mm. because the days without any precipitation is represented with 0 mm. The mean value of the precipitation is 96.72 mm, which is a moderate value when the whole data is taken into consideration with its minimums and maximums. If the observations for temperature value are considered, the temperature never lies under zero and never exceeds 29 ^{0}C. The average temperature for the whole data collected from 1950 to 2012 is obtained as 14.97 ^{0}C which is an indicator of a moderate zone in terms of weather conditions. Because of the location of Muğla which is a city by sea, humidity is observed as a result of maritime climate. The observed minimum relative humidity value of 29.10 is the evidence of this seaside effect. Even the average value of relative humidity is around 65. The cloudiness in the Muğla city is mostly at low levels.

If the time series plot of Muğla station is examined, firstly, it should be checked whether there is a trend pattern or not. Trend can be observed in two ways. If the graph is going up (down) by time, it is called as positive (negative) secular trend. It is needed to have enough data to detect the pattern of trend. Since our series is consisting of 732 observation, it can be easily understood from the time series plot of precipitation values.

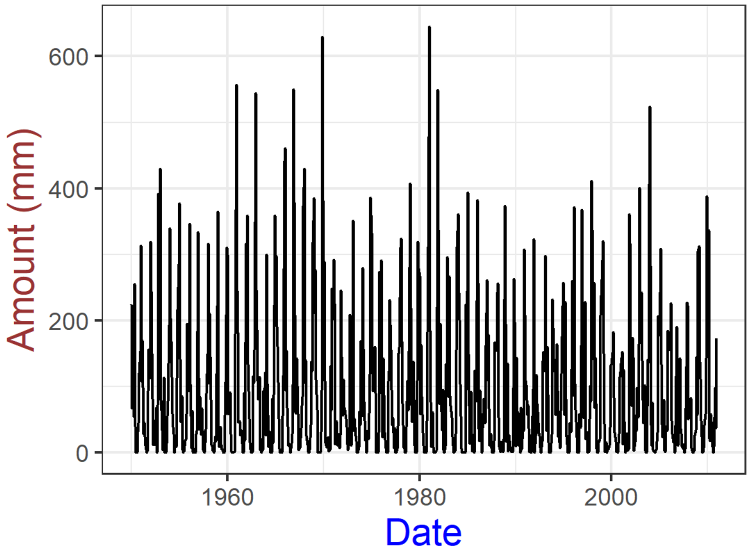

Figure 4.1. Time Series Plot of the Total Precipitation Amount of Muğla Station from 1950 to 2010

According to the Figure 4.1, there is no clue for the pattern of trend, hence stationarity test was applied and it is need that the series is stationary. After deciding about trend, the variation is taken into consideration. The variation is actually coming from its first basic idea which shows the peaks and troughs. Those are the high and low points in the data. Figure 4.1 helps us to see where the times series precipitation data goes up and goes down along the 62 years. Seasonality means regular peaks and troughs which happen at the same time each year. There are regular ups and downs in specific periods of each year which is seen in Figure 4.1. Therefore, it can be said that the precipitation data is a seasonal time series.

4.2.2. Application for Muğla Station

The applications of the actual model and hybrid model have been examined respectively in this section.

4.2.2.1. Actual Model

Since our series has “0”s for the precipitation values, some negative values in both prediction, filtering and smoothing processes are obtained. Those negative values are counted as zero as it is stated before in this application procedure for the data itself. We define our linear regression model in a very simple way in equation (4.4).

$$\begin{aligned} Precipitation_t = \ & \alpha + \ \beta_1 (Temperature)_t + \ \beta_2 (Relative\ Humidity)_t \\ & + \beta_3 (Cloudiness)_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H). \end{aligned} \tag{4.4}$$

The time varying version of regression model (equation 4.4) into the state space form as in equation (4.5).

$$Precipitation_t = \alpha + \beta_{1t}(Temperature)_t + \beta_{2t}(Relative\ Humidity)_t + \beta_{3t}(Cloudiness)_t + \varepsilon_t, \quad \varepsilon_t \sim N(0, H) \tag{4.5}$$

In accordance with equation (3.10), the state equations can be written as;

$$\begin{aligned} \beta_{1t} &= \bar{\beta}_1 + \Phi_1(\beta_{1t} - \bar{\beta}_1) + w_{1t} \quad & w_{1t} \sim N(0, Q_1), \\ \beta_{2t} &= \bar{\beta}_2 + \Phi_2(\beta_{2t} - \bar{\beta}_2) + w_{2t} \quad & w_{2t} \sim N(0, Q_2), \\ \beta_{3t} &= \bar{\beta}_3 + \Phi_3(\beta_{3t} - \bar{\beta}_3) + w_{3t} \quad & w_{3t} \sim N(0, Q_3). \end{aligned} \tag{4.6}$$

with priors

$$\beta_{10} \sim N(\mu_{\beta_1}, \Sigma_{\beta_1}), \quad \beta_{20} \sim N(\mu_{\beta_2}, \Sigma_{\beta_2}), \quad \beta_{30} \sim N(\mu_{\beta_3}, \Sigma_{\beta_3}). \tag{4.7}$$

The estimation of the parameters of the distributions has been made from the data in the estimation part. Here, the intercept of the regression is represented by α and the slopes of the regression are defined as β_{1t}, β_{2t} and β_{3t} estimated by MLE as $\hat{\alpha}, \widehat{\beta_{1t}}, \widehat{\beta_{2t}}$ and $\widehat{\beta_{3t}}$, respectively. Note that, these parameters estimation process is discussed in section 3.1.2.2.1.

For the model selection, the MAE and the MSE values are obtained for each step prediction, filtering and smoothing respectively. Inside each algorithm, *R*^{2} and *adjusted R*^{2} values are generated. At the end, $\overline{Y_t}$ values for prediction, filtering and smoothing parts calculated and compared with the original precipitation values.

The smoothed precipitation values achieved at the end are compared with the original precipitation values in order to see how we close to the actual observations by applying the KFMR to predict the monthly precipitation data. The plots below are carried out by dividing the data into 4 sequential groups in order to see the two lines smoothed and original precipitation values clearly. The relation is shown in Figures 4.2., 4.3, 4.4 and 4.5.

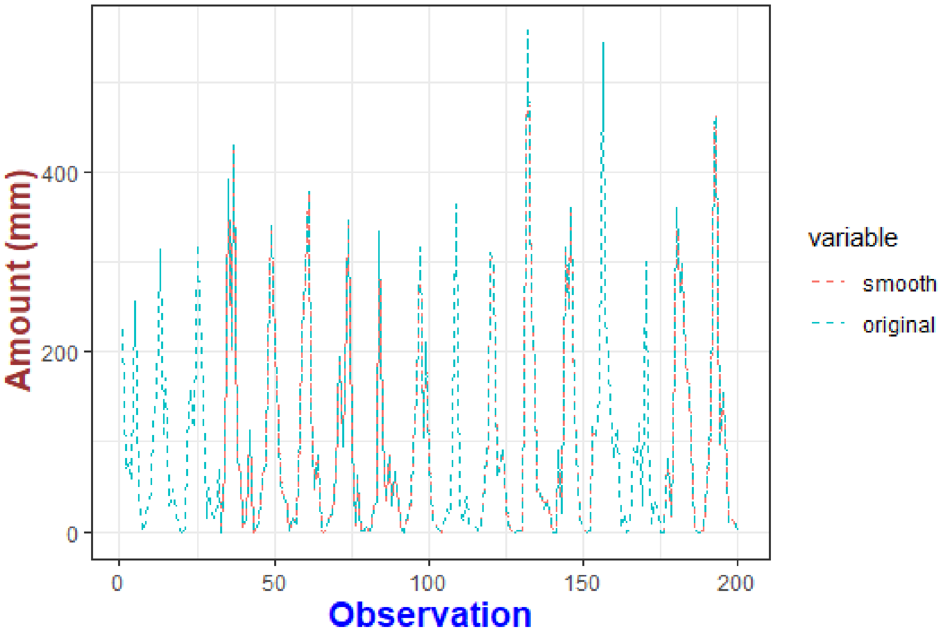

Figure 4.2. Smoothed vs Original Precipitation Values Muğla I

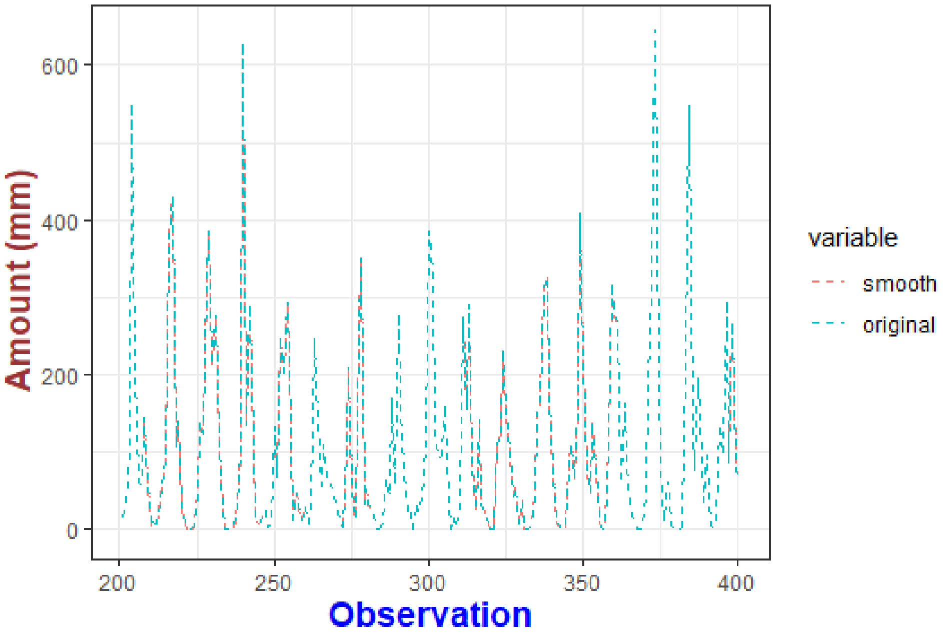

Figure 4.3. Smoothed vs Original Precipitation Values Muğla II

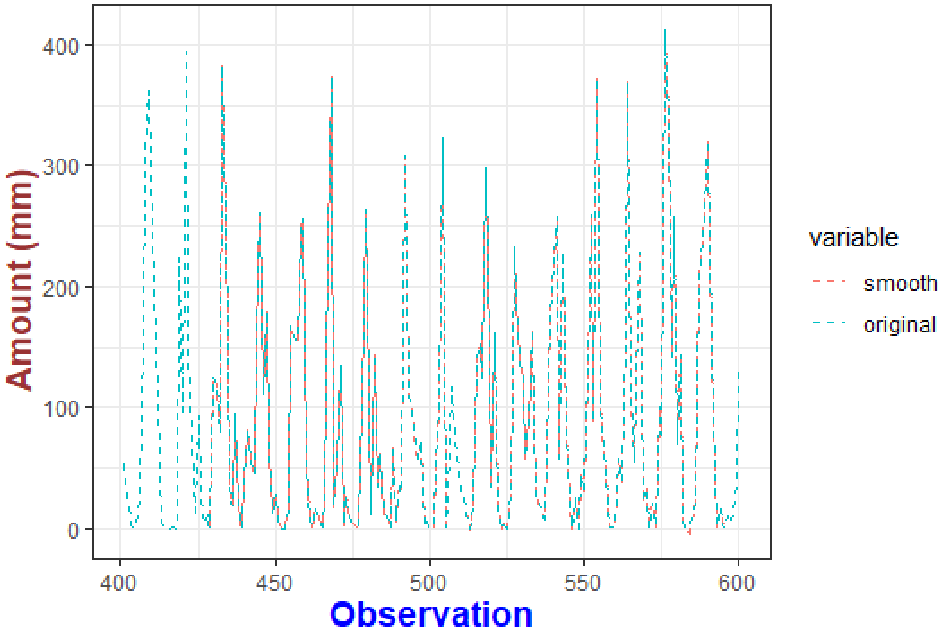

Figure 4.4. Smoothed vs Original Precipitation Values Muğla III

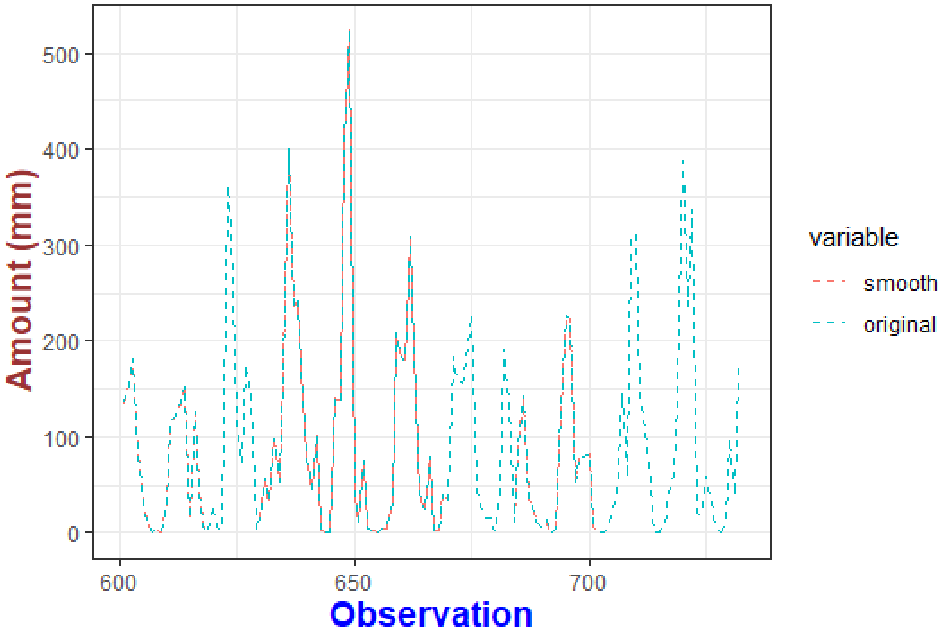

Figure 4.5. Smoothed vs Original Precipitation Values Muğla IV

The red line represents the original values of precipitation and the blue line represents the smoothed values of precipitation. It can be clearly seen that there is a perfect match between the original variables and smoothed variables. As it is seen that the smoothed values for precipitation is acting almost same as the original precipitation values.

When focusing on a specific time period in order to see the relation up too close, Figure 4.6 is generated based on the smoothed and original observations of precipitation between 1970 and 1971.

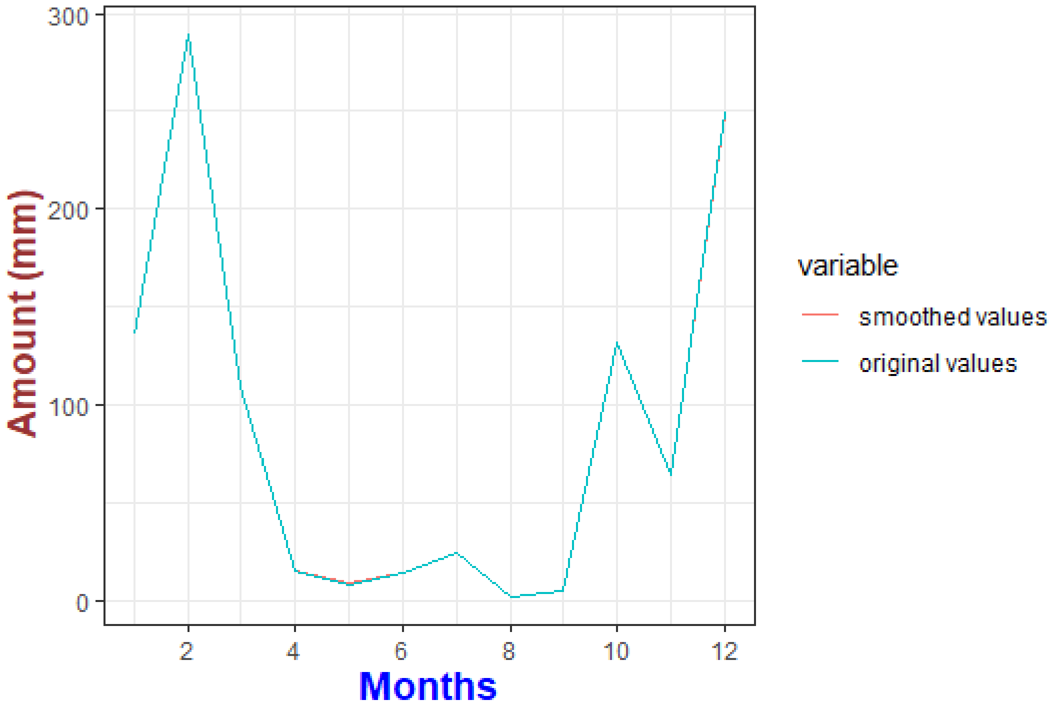

Figure 4.6. Smoothed vs Original Precipitation Values of Muğla for 1-Year Time Period

Even if the discrimination of red line and blue line is hard in the Figures 4.2, 4.3, 4.4 and 4.5 for smoothed and original values comparison, the intervals or points can easily detected from the Figure 4.6 that the red line is differentiating from the blue line. According to the Figure 4.6, the smoothed values of precipitation in each month of year again matched with the original ones although some insignificant differences observed especially around mid of the year and around September. On that period, the smoothed values can be seen as moving with little deviations than the original values. However, the significance of catching this perfect integration is valid in every period of the mentioned year.

The meaning of this perfect match between smoothed predictions of precipitation values and original precipitation values is that to use the KFMR for predicting the monthly precipitation which is affected by some unobserved variables or noises gives very accurate results. Since the data is a seasonal time series data, it might be taught that prediction of precipitation at the specific periods of year according to the region is a simple procedure. However, Kalman Filter is not just good at predicting the values of precipitation in that specific time periods. It also very well worked for whole time periods among the years. This strong closeness between the smoothed values by Kalman Filter and original values are shown briefly in Table 4.2.

Table 4.2. Smoothed vs Original Precipitation Values for Muğla

Smoothed precipitation values	**Original precipitation values**
224.3998	224.7000
67.2703	67.2000
77.7153	77.7000
55.0009	54.9000
255.5676	255.7000
40.1895	40.1000
-0.3102	0.0000
-0.2966	0.0000
31.2793	31.4000
44.9046	44.8000

Very small differences are just observed for the decimal points of two precipitation values.

The difference between the original precipitation values and the smoothed precipitation values has construct the residual term and the standardized residual term. The first 10 residual and standardized residual values are listed in Table 4.3 and all the residuals are plotted in Figure 4.7.

Table 4.3. Residuals and Standardized Residuals for Muğla

Residuals	Standardized Residuals
0.3002	0.0031
-0.0703	-0.0009
-0.0153	-0.0001
-0.1009	-0.0010
0.1323	0.0014
-0.0895	-0.0030
0.3102	0.0533
0.2966	0.0506
0.1206	0.0053
-0.1046	-0.0019

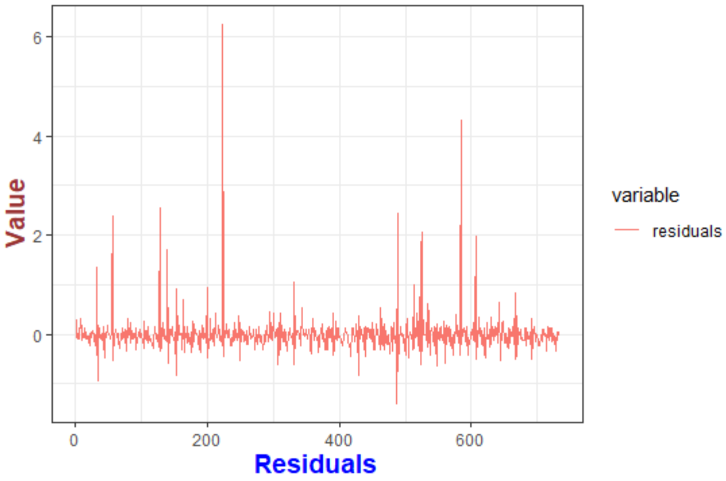

Figure 4.7. Residuals for Muğla

In the most ideal scenario, residuals should be small and unstructured. What the meaning of small and unstructured residual is that the variation of the dependent variable has been explained successfully in the model. In this study, the residuals and standardized residuals for the smoothed precipitation values are very small which makes our model significant.

These very accurate results direct us to see the MAE and MSE values for the model constructed with the smoothed precipitation values. The MAE and MSE values are figured out in Table 4.4.

Table 4.4. MAE and MSE Values for Smoothed Precipitation Values of Muğla

SMAE	**SMSE**
0.1871	0.1743
SMAE: Smoothing MAE	
SMSE: Smoothing MSE	

The models with small MSE and MAE values are the proper models to use (Hyndman & Athanasopoulos, 2018). For model selection, MSE and MAE values are compared between the models and the model with small MAE and MSE is chosen at the end. According to the Table 4.4, MAE result for the model with smoothed precipitation values is 0.187144 and MSE result for same model is 0.174367. Both MSE and MAE values for the model created by smoothed variables are very small which support the idea that the model constructed with smoothed precipitation values and β values is a meaningful model and the existent perfect match between the smoothed and original precipitation variables is significant.

The significance of the model constructed by smoothed variables in terms of the relations between the precipitation and independent variables temperature, relative humidity and cloudiness is controlled by looking into R^2 and a*djusted* R^2 values. These values describes how much of the change in dependent variable can be explained by the change in independent variables. In other words, it defines the accuracy of the model shown in Table 4.5.

Table 4.5. R^2 and Adjusted R^2 Values for Muğla

SadjR^2	**SR^2**
0.9999	0.9999

The smoothed *adjusted R^2* and *R^2* are very close to the 1. This means that 99% of the change in the amount of smoothed precipitation can be explained by the explanatory variables. The model consisting of smoothed variables is a very reasonable model in terms of the relationship between the dependent and independent variables. Kalman Filter again proves its strength in predicting the amount of precipitation for the regression model.

These perfect results give a way to the study as making a forecast on precipitation. Rolling window forecast technique is used for forecasting the precipitation values. The procedure followed in rolling window forecasting technique is given in the following algorithm.

Step 1. A starting value is taken as 500. In other words, the model is estimated with the first 500 observations to forecast the observation 501.

Step 2. The observation 501 is included in the estimation sample and the model is estimated again with 501 observations to forecast the observation 502.

Step 3. The process is repeated until a forecast for all 32 out of sample observations is reached. The first 10 forecast values and the original values are stated in the Table 4.6.

Table 4.6. Forecast and Original Precipitation Values of Muğla

Forecast values	Original values	Difference
7.2127	7.2000	0.01270
0.0832	0.1000	-0.0167
-0.0321	0.0000	-0.0321
2.5870	2.6000	-0.0129
26.7200	26.7000	0.0200
34.8128	34.8000	0.0128
145.1987	145.2000	-0.0012
57.6206	57.6000	0.0206
304.9838	305.000	-0.0161
312.1912	312.2000	-0.0087

It can be seen from the Table 4.6, the forecast algorithm is worked very well because the differences shown in the third column of Table 4.6 are very small. The plot of the forecast and original values of precipitation.

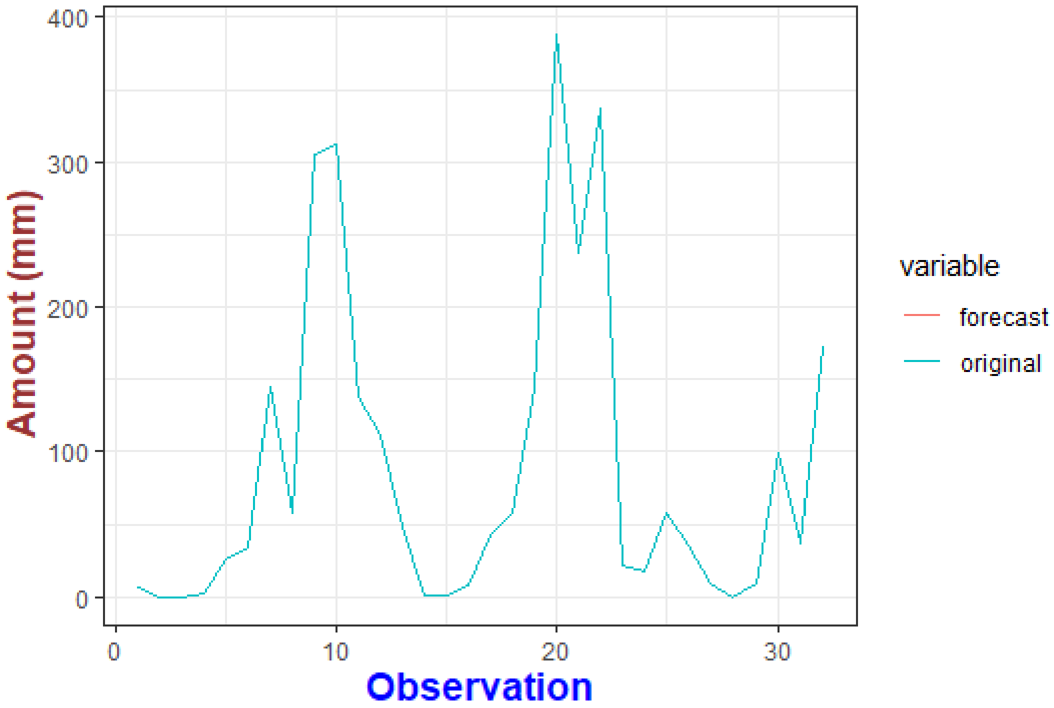

Figure 4.8. Forecasting Precipitation Values of Muğla

The red line is for the 32 forecast values and the blue line is for the 32 original values. Our forecasting mechanism is good at forecasting the precipitation values.

To prove the performance of applied forecast algorithm, the MSE and MAE values are also calculated and found as in Table 4.7.

Table 4.7. MAE and MSE Results for Forecast Values of Muğla

MAE	**MSE**
0.0132	0.0001

The MAE and MSE values for the forecast application are very small which are reasonable.

After concluding the forecast calculations and interpreting the visuals, the performance of the Kalman Filter (KFMR) estimation is needed to compare with the performance of OLS estimation and Seasonal ARIMA model for predicting the precipitation. The Seasonal ARIMA model fitted by using the *auto.arima( )* function in RStudio. The logic behind the auto arima algorithm is to have a combination of the unit root tests and minimum information criteria values of AIC and BIC (Hyndman & Athanasopoulos, 2018). In other words, auto arima suggests a model with the smallest information criterion. After using *auto.arima( ),* the Seasonal ARIMA model observed as SARIMA *(1, 0, 0) x (1, 1, 0)* by checking the residual diagnostics.

The forecast results have been added to the comparison in order to see the performances of all techniques on prediction. The comparison is shown in the Figure 4.9.

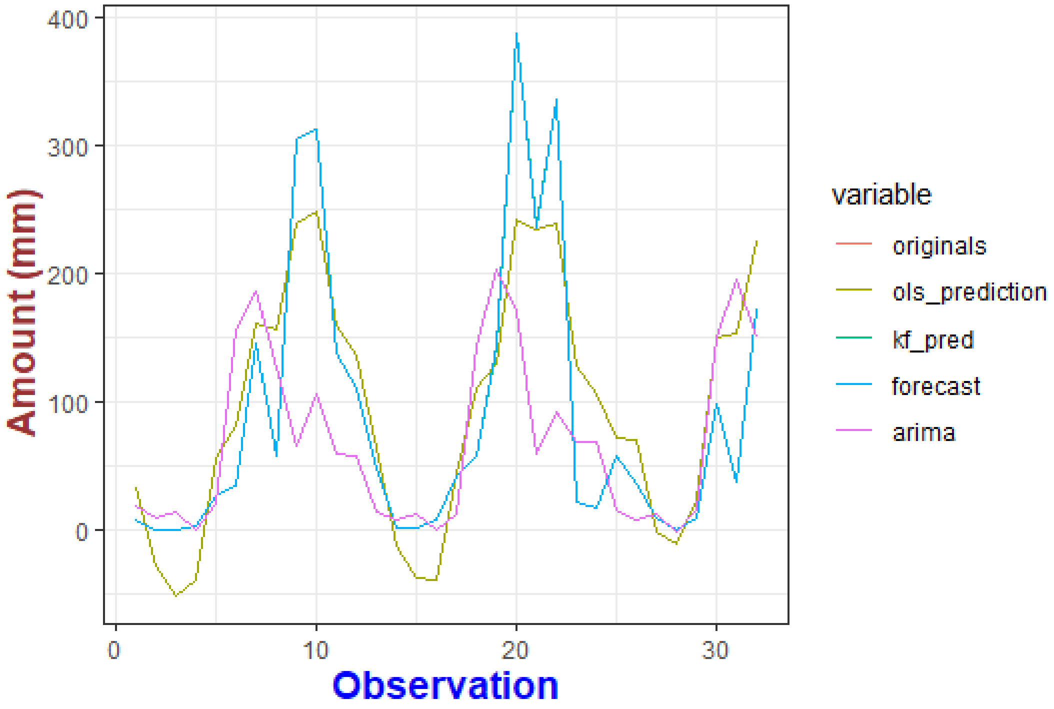

Figure 4.9. Prediction Comparison for Muğla

Since the Kalman Filter predictions for precipitation have perfect match with the original values of precipitation as same as with the forecast values, they are put in the same line in Figure 4.9. Original precipitation values, Kalman Filter predictions and forecast values are represented by the blue line all together. The green line is for the OLS predictions of precipitation and the purple line represents the predicted precipitation values by ARIMA. When OLS estimation technique is implemented to the data, it has not worked as well as Kalman Filter which can be seen obviously from the Figure 4.9. The peaks and troughs have been predicted in the right direction with pattern of the original precipitation values by the OLS. However, especially the points that precipitation have get its maximum and minimum values in some specified periods, the OLS predictions have-not caught that points. The ups and downs in those points have not been clarified with the OLS estimation technique. Although there is similarity on the dispersion pattern of predictions for the OLS and ARIMA, even the

OLS has been working better than the ARIMA. The predictions of precipitation values have not caught the original values in the maximum and minimum points in ARIMA as well. Therefore, Kalman Filter has outperformed the OLS and ARIMA for predicting the amount of monthly total precipitation.

4.2.2.2. Hybrid Model

Since the previous analysis to actual data has given some negative values for the predicted, filtered and smoothed precipitation, those values have been counted as zero. However, in this part of analysis, it has been tried to make some transformation on precipitation observations. The whole process has been repeated but the results again contained the negative values in prediction, filtering and smoothing steps. At this point, making a transformation to handle the negative values was meaningless because the small values creates big differences in the models and the error terms are expected to increase. Hence, a hybrid model was considered to apply. As it is mentioned in Section 4.1, the procedure followed to create a hybrid model is starting with the arrangement of precipitation amount series as 0 if the amount is 0, otherwise 1. With this obtained binary independent variables, a logistic regression is fitted with the same predictors; temperature, relative humidity and cloudiness. Those estimated 0 values are fixed as 0, and that estimated 1 values are used in the KFMR using the original precipitation amount series. The whole KFMR application is applied with the new data to obtain the accuracy measures and forecasts. The new hybrid model is again structured as a state space model form. The only difference here that the binary response variable consisting just 1's. The accuracy results which has been obtained from the logistic regression is as follows in Table 4.8.

Table 4.8. Logistic Regression Confusion Matrix for Muğla Hybrid Model

Accuracy	0.9235
95% CI	(0.8913, 0.9486)
No Information Rate	0.8934
P-Value [Acc > NIR]	0.0332
Kappa	0.4810
Mcnemar's Test P-Value	0.0003
Sensitivity	0.3846
Specificity	0.9877
Pos Pred Value	0.7894
Neg Pred Value	0.9308
Prevalence	0.1065
Detection Rate	0.0409
Detection Prevalence	0.0519
Balanced Accuracy	0.6861

The accuracy which is a measurement showing that the classification is true is 92% in the Table 4.8. This is enough to say that our classification by arranging 0’s as itself and the other precipitation amounts as 1 is an accurate technique. Sensitivity means to predict the true values as true. It is 38% a little bit small but the specificity which means identifying the negatives correctly is very high as 98%.

After checking the logistic regression accuracy measurements, the hybrid model has been prepared for the Kalman Filter algorithm. The transition model has been again turned into the Kalman Filter Mean Reverting (KFMR) as in equations (4.4), (4.5) and (4.6).

The smoothed values of precipitation in the model have been compared visually with the original precipitation values in Figures 4.10, 4.11, 4.12 and 4.13. Since having a large sample size decreases the significance of the pattern in visual, the whole data have not been directly used as it is in the plot. The plots are showing the pattern of these two vector of values monthly.

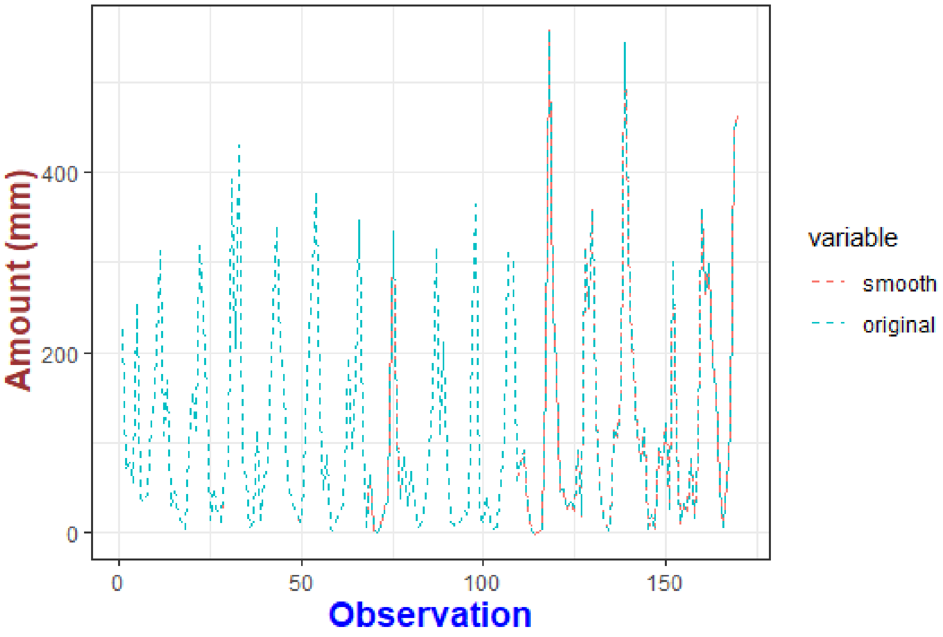

Figure 4.10. Smoothed vs Original Precipitation Values Muğla Hybrid Model I

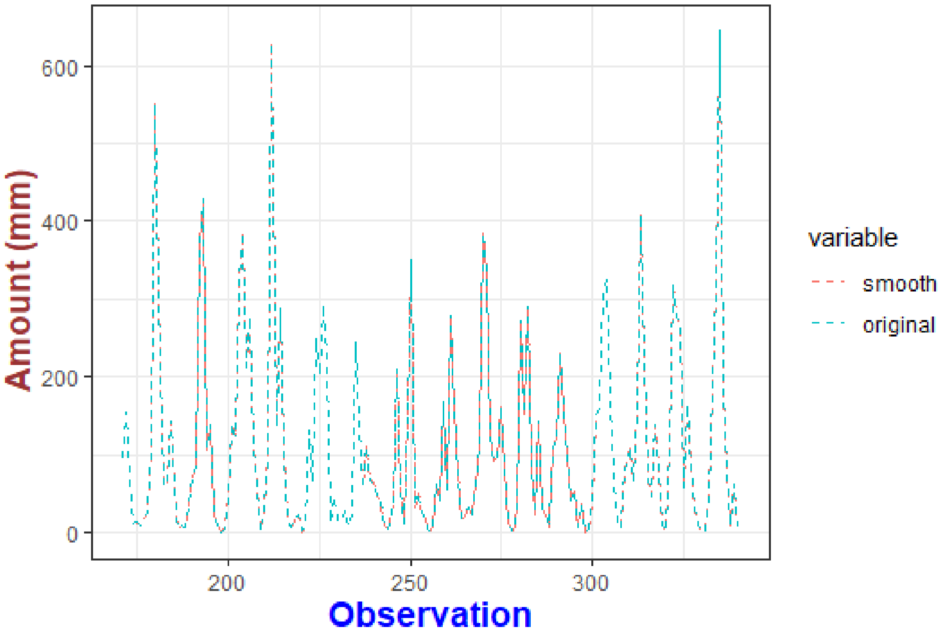

Figure 4.11. Smoothed vs Original Precipitation Values Muğla Hybrid Model II

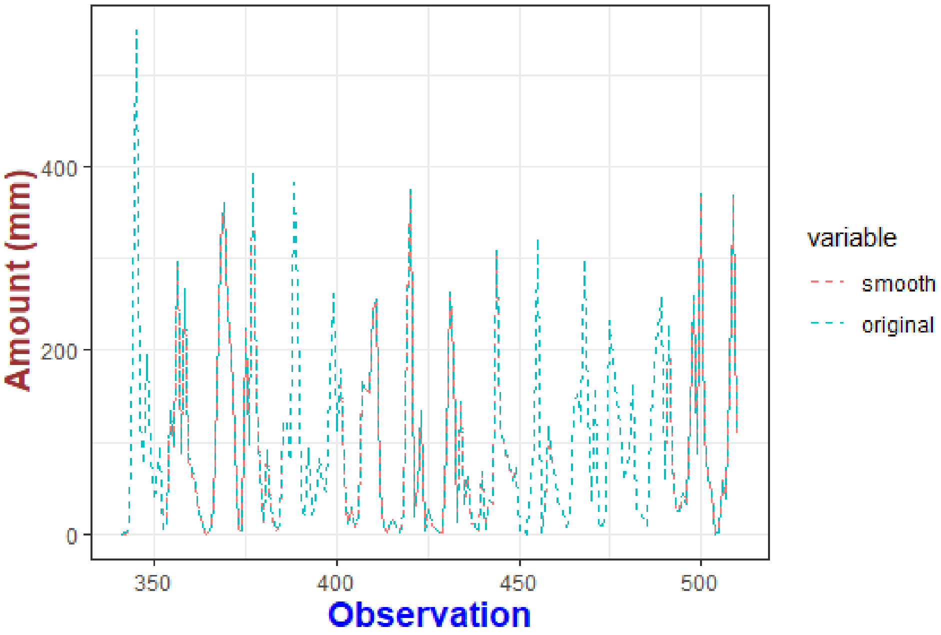

Figure 4.12. Smoothed vs Original Precipitation Values Muğla Hybrid Model III

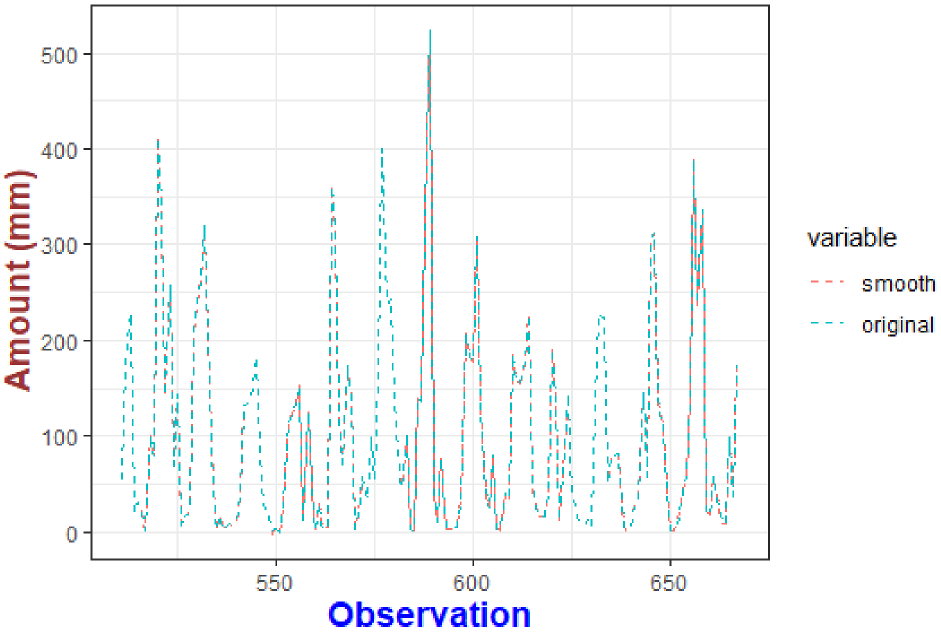

Figure 4.13. Smoothed vs Original Precipitation Values Muğla Hybrid Model IV

The perfect accord between the smoothed variables of precipitation represented by blue line and the original precipitation variables represented by red line has been proved again for the hybrid model. Although our data have been turned into a hybrid model, Kalman Filter has shown its performance as well as in the hybrid model. This performance is shown in Figure 4.14 by focusing a specific time period in details.

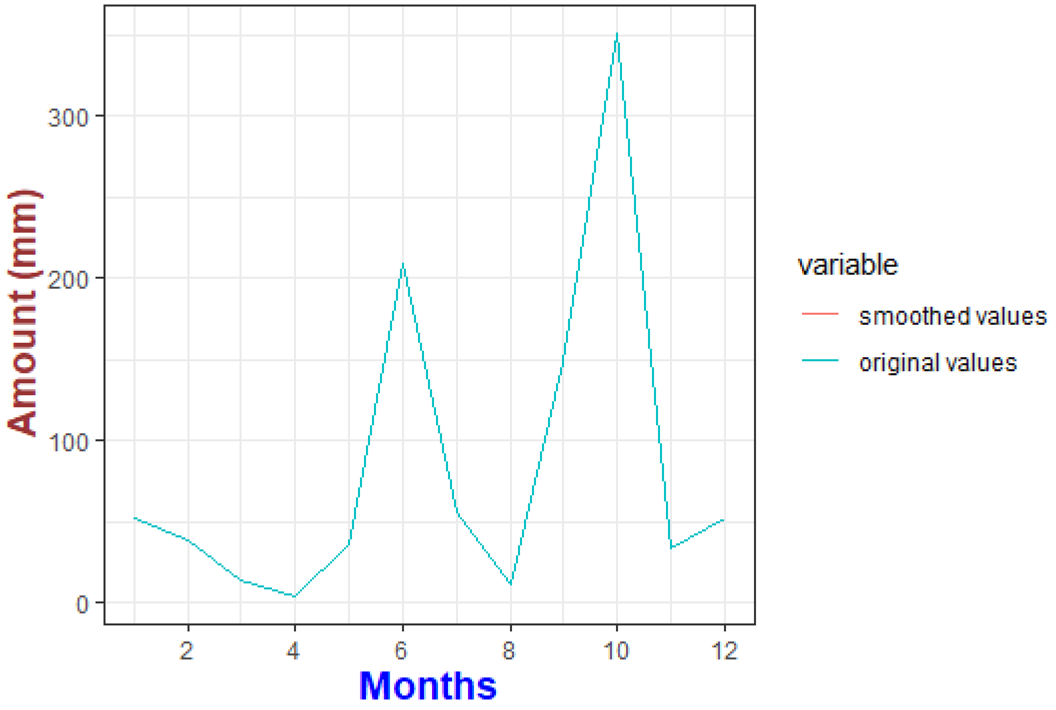

Figure 4.14. Smoothed vs Original Precipitation Values of Muğla Hybrid Model for 1-Year Time Period

Figure 4.14 is constructed with the observations from 1970 to 1971 in order to see the differences for 12 month period. Again, the lines representing the original and smoothed values are perfectly matched but the insignificant small differences can be seen in the smoothed variables around August.

By estimating the state vector in KFMR, amount of monthly precipitation is successfully smoothed. To handle the noises in the data has been achieved by Kalman Filter. As a result, these smoothed values of precipitation have been detected as so close to the originals in every period of data. A small part of these values shown in the figures above has been placed in Table 4.9.

Table 4.9. Smoothed vs Original Precipitation Values for Muğla Hybrid Model

Smoothed precipitation values	**Original precipitation values**
224.4326	224.7000
67.2860	67.2000
77.7255	77.7000
55.0213	54.9000
255.5615	255.7000
40.1579	40.1000
31.2142	31.4000
44.9252	44.8000
120.1770	120.2000
153.2621	153.2000

The reason why the red and blue lines represents the original and smoothed values of precipitation overlap directly inferred from the insignificant differences in Table 4.9.

The residuals defined as the difference between original and smoothed values of precipitation have briefly shown in Table 4.10 and Figure 4.15.

Table 4.10. Residuals and Standardized Residuals for Muğla Hybrid Model

Residuals	**Standardized Residuals**
0.2673	0.0028
-0.0860	-0.0011
-0.0255	-0.0002
-0.1213	-0.0012
0.1384	0.0014
-0.0579	-0.0019
0.1857	0.0082
-0.1252	-0.0022
0.0229	0.0002
-0.0621	-0.0004

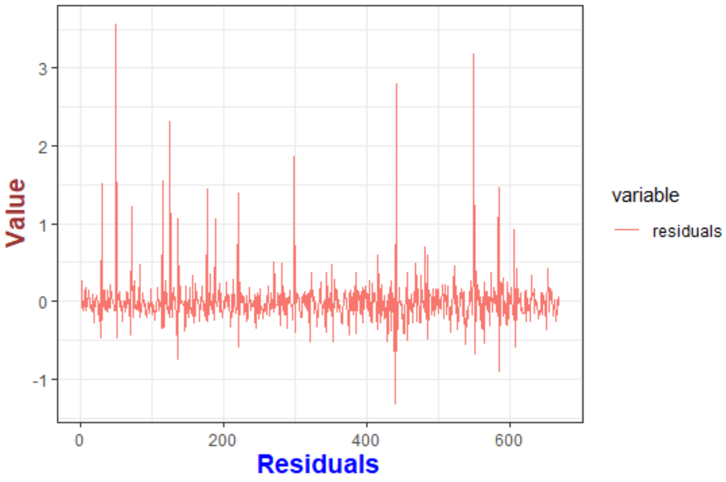

Figure 4.15. Residuals for Muğla Hybrid Model

According to the Figure 4.15, the residuals around 0 have not shown a structure. Having these small numbers for the residuals and standardized residuals means that Kalman Filter has been operated properly for the prediction of monthly amount of precipitation for the hybrid model.

The remarkable performance of Kalman Filter in the hybrid model have need to be supported by the significant model evaluation criteria values which are MAE and MSE. Even though MAE and MSE values have been obtained for prediction, filtering and smoothing steps, the considered results of MAE and MSE have been the ones calculated in the smoothing step. The results are as given in the following table;

Table 4.11. MAE and MSE Values for Smoothed Precipitation Values of Muğla Hybrid Model

SMAE	**SMSE**
0.1810	0.1220
SMAE: Smoothing MAE	
SMSE: Smoothing MSE	

Smoothing MAE and MSE values have been interpreted as significant as being model evaluation criteria since the smaller MAE and MSE values means the better model. When they have been compared to the ones in the previous application for city of Muğla, the MSE value for the hybrid model has been performed 30% better than the one in the previous model.

The model evaluation criteria have directed us to check the *R^2* and *adjusted R^2* values of the model to see how much of the variability in the smoothed precipitation values in the model can be explained by the temperature, relative humidity and cloudiness. The values are represented in Table 4.12.

Table 4.12. R^2 and Adjusted R^2 Values for Muğla Hybrid Model

SadjR^2	**SR^2**
0.9999	0.9999

From Table 4.12, the 99% variability in the smoothed precipitation values can be explained by the independent variables for the hybrid model. This is actually a very extreme value for an accuracy measure; however, this can be counted as proof of the power of Kalman Filter in precipitation modelling.

After giving the details about the results of Kalman Filter algorithm on the hybrid data, the same forecasting method have been preferred to use which is rolling window forecast technique. The same number of sample size have been chosen at the beginning for the estimation procedure to forecast the observation 501. The algorithm goes like this until 32 forecast values have been observed. The head 10 forecast values with the originals are given in Table 4.13.

Table 4.13. Forecast and Original Precipitation Values of Muğla Hybrid Model

Forecast values	Original values	Difference
79.8221	79.8000	0.0221
82.4157	82.4000	0.0157
7.2090	7.2000	0.0090
0.0824	0.1000	-0.0175
2.5795	2.6000	-0.0204
26.7168	26.7000	0.0168
34.8115	34.8000	0.0115
145.1999	145.2000	0.0001
57.6182	57.6000	0.0182
304.9850	305.0000	-0.0149

As it can be understood from the small differences between the forecasted values and original values, the forecast technique has given significant values for the hybrid model as well. This idea has been also provided in Figure 4.16.

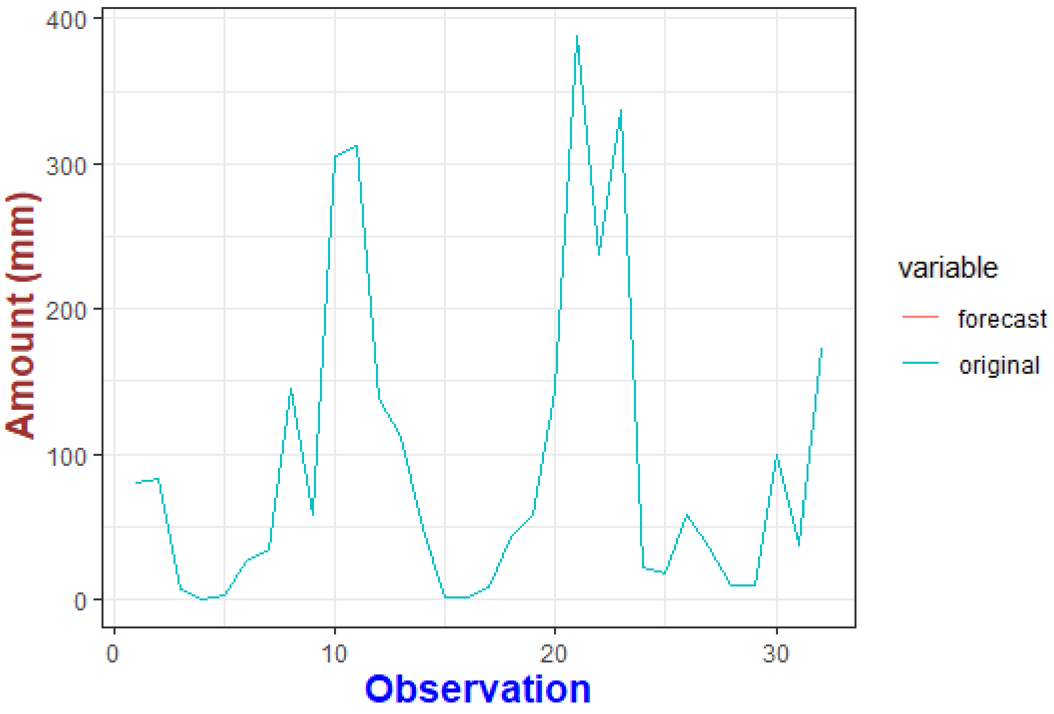

Figure 4.16. Forecasting Precipitation Values of Muğla Hybrid Model

In all peaks and troughs, the forecasting mechanism for the precipitation have worked properly as expected. The MAE and MSE values of forecasting have also supported the proper working mechanism of our forecast in hybrid model as stated in Table 4.14.

Table 4.14. MAE and MSE Results for Forecast Values of Muğla Hybrid Model

MAE	**MSE**
0.0132	0.0002

The MAE values are very close to each other; however, the performance of MSE in forecasting for the hybrid model is 30% higher than the previous model.

At the end of analysis, the performances of OLS, ARIMA and Kalman Filter estimations for the monthly precipitation data have examined. The ARIMA model has been again fitted with the *auto.arima( )* and found as ARIMA(3, 0, 2). The forecast results have been included the plot as well seen in Figure 4.17.

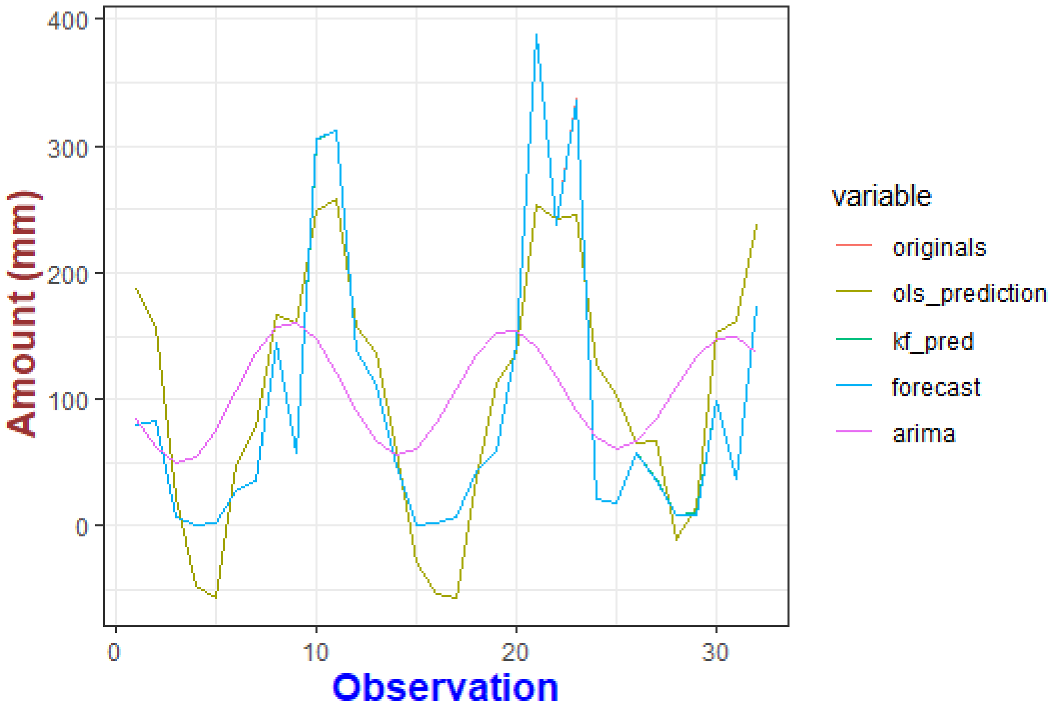

Figure 4.17. Prediction Comparison for Muğla Hybrid Model

The Kalman Filter is again better than the OLS and ARIMA methods for predicting the precipitation for the hybrid model. Since the original precipitation values have been very close to the smoothed precipitation values found by KFMR, they cannot seen directly. The overlapped lines of original precipitation values, Kalman Filter predictions and forecast values are the proof of the good performance of our method. However, the OLS estimation and ARIMA method have not worked as good as Kalman Filter as expected for the hybrid model. Although the directions of the increases and decreases are estimated truly by the OLS estimation, the amount of that increases and decreases has not been caught properly. In other words, the ups and downs within an increasing pattern are not able to see clearly in the OLS estimation. Yet, the Kalman Filter predictions has included all movements and patterns of the original precipitation values. Unlike OLS estimation, ARIMA method has not even shown the up and down movements and it has just a smooth line for the precipitation values. Hence, OLS and ARIMA methods have been overwhelmed by Kalman Filter method for the prediction of the amount of monthly precipitation.

4.3. Analysis for Konya

In the following parts, description of Konya data will be made and application results will be shared, respectively.

4.3.1. Data Description for Konya

The data used in this part of analysis has been taken from the city of Konya which is a 61-yeared data collected from 1950 to 2010. The sample size for Konya data has been again chosen as 732 and the same independent variables have been preferred to use in the analysis as temperature, relative humidity and cloudiness. The aim of making another Kalman Filter application to a different data is wondering the performance of Kalman Filter on predicting the monthly precipitation in a region in which the amount of precipitation has been observed very rare. That is to say Konya can be categorized as a scarce region in terms of getting precipitation. According to the data taken from the Turkish State Meteorological Service the days that Konya have taken precipitation in a year is 22% which is enough to named Konya as droughty. Although Konya is an arid area in terms of the amount of precipitation, the days without any kind of precipitation have been too few in our data. In other words, the zero values in the precipitation vector of our data have been observed around 10 observations. Therefore, only the Kalman Filter method has been applied directly to the data and another application to hybrid data with logistic regression has not been considered as necessarily. Since the two applications in Muğla station have already given close results, to make the hybrid model application has been evaluated as non-essential. After making decision about reducing the Kalman Filter application just for the actual data, some descriptive statistics have been given for Konya data in Table 4.15.

Table 4.15. Descriptive Statistics for Konya

	Precipitation	**Temperature**	**Relative Humidity**	**Cloudiness**
Minimum	0.000	-7.700	25.500	0.200
1st Quartile	14.200	2.913	52.250	2.500
Median	30.890	10.300	62.200	4.300
Mean	34.810	10.246	62.630	4.154
3rd Quartile	50.120	17.800	73.380	5.700
Maximum	157.500	26.300	89.600	8.300

The minimum value has been observed as 0 for the amount of precipitation as expected for the region. The observed maximum value for the precipitation in Konya is 157.50 mm which is the 25% of the observed maximum precipitation value of Muğla. The average value of precipitation for the 61-yeared time period in Konya is 34.81 which is the 30% of the average precipitation value observed in Konya. This difference has supported our argument about Konya being a scarce region in terms of precipitation.

While the temperature has never been observed under zero in Muğla, the minimum temperature for Konya has been seen as -7.70 ^{0}C. Even the Konya is near the seaside cities, the relative humidity has not been measured as much as Muğla and the number of cloudy days has seemed very rare.

With the knowledge of descriptive statistics, the time series plot of the precipitation data has been examined in Figure 4.18.

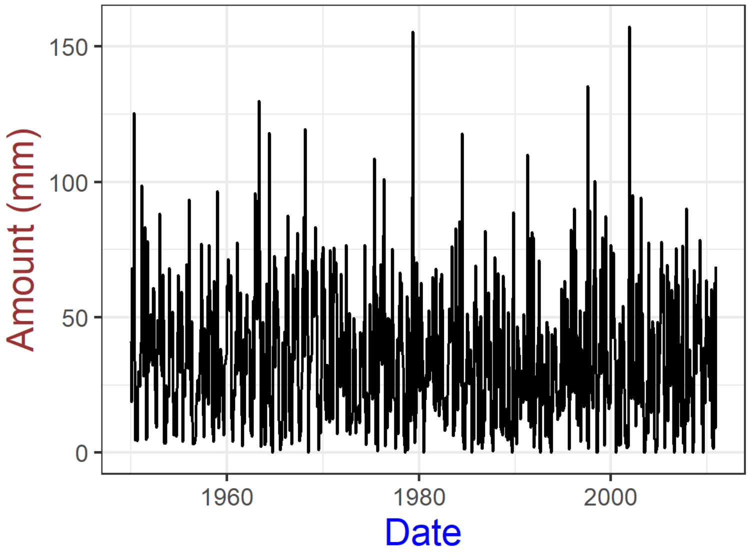

Figure 4.18. Time Series Plot of the Total Precipitation Amount of Konya Station from 1950 to 2010

According to the Figure 4.16, there is enough evidence to say that the series has shown no trend pattern. After deciding about trend, the variation is taken into consideration. The high and low points in the observation of precipitation have been clearly detected from the plot. The observed maximum precipitation values have reached 500 mm and the minimum values for precipitation have been approached to 0. Those are the high and low points in the data. When seasonality is considered for the data, there has been definitely seasonal pattern in the observations. The regular ups and downs in specific periods of each year have generated a pattern for seasonality.

4.3.2. Application for Konya Station

Since the data for Konya has not included zero values while collecting the precipitation observations as much as Muğla data, as it has been mentioned before, the hybrid model application will not be considered in the application part of study. The negative values obtained for the smoothed values of precipitation has been counted as zero since the data has already had a few zero values in any case. In other words, in those days, it can be taught as there were no precipitation in any type.

The precipitation regression model has been constructed for Konya station as a state space model form like in equations (4.4), (4.5) and (4.6)

Since the data is collected as monthly for 61 years and the sample size is large, the plot of smoothed and original values of precipitation has been cut into 4 for seeing more clearly in Figures 4.19, 4.20, 4.21 and 4.22.

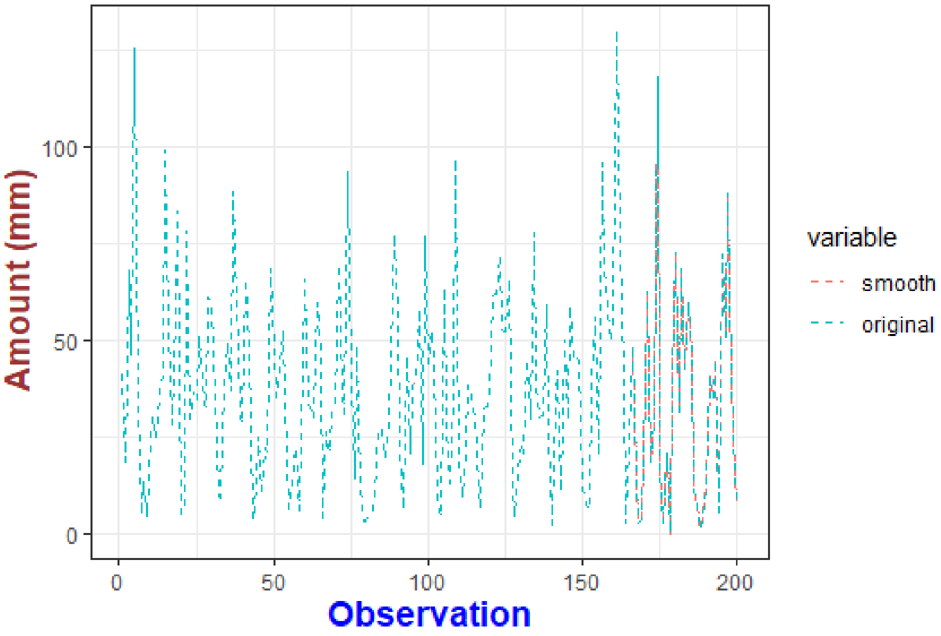

Figure 4.19. Smoothed vs Original Precipitation Values Konya I

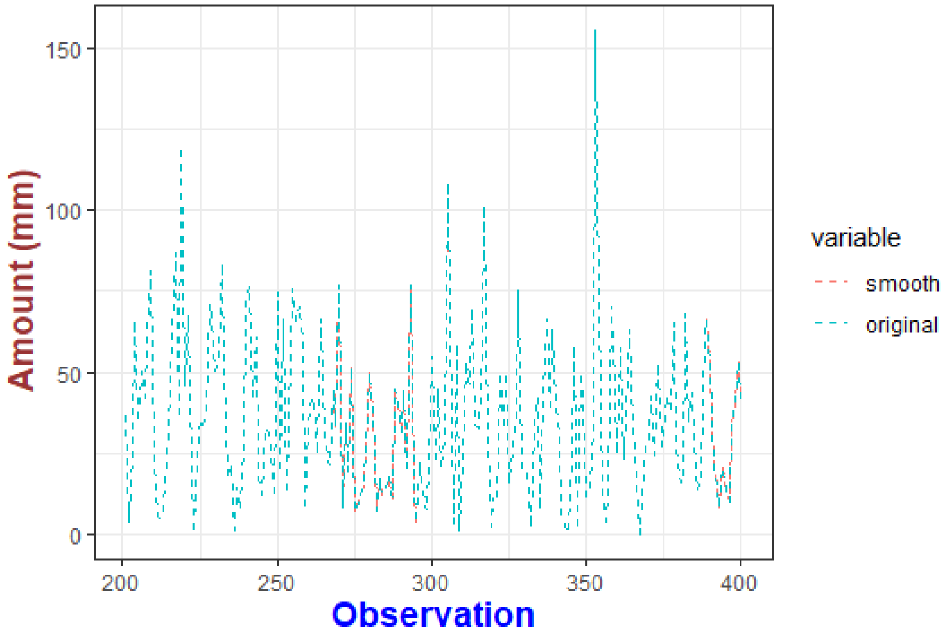

Figure 4.20. Smoothed vs Original Precipitation Values Konya II

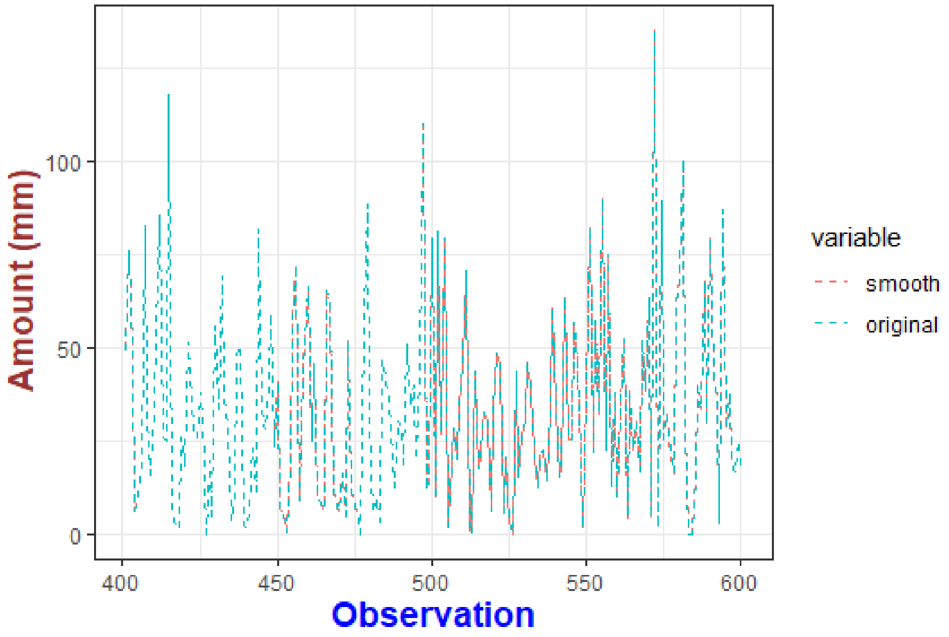

Figure 4.21. Smoothed vs Original Precipitation Values Konya III

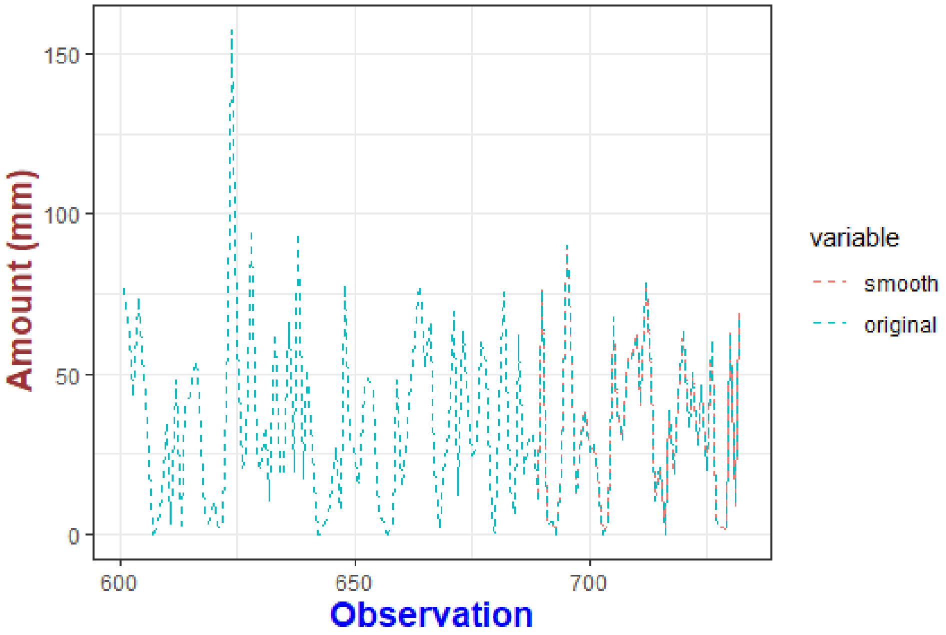

Figure 4.22. Smoothed vs Original Precipitation Values Konya IV

As well as in the Muğla station, the smoothed values of predicted precipitation values by the Kalman Filter method have seen very close to the original values of precipitation values. The line for original values represented by blue and the line for smoothed values represented by red have been stratified in the plot. Even if there are small differences between those values, they can be seen in a very detailed plot which is constructed based on the idea of focusing the years between 1970 and 1971 as in Figure 4.23.

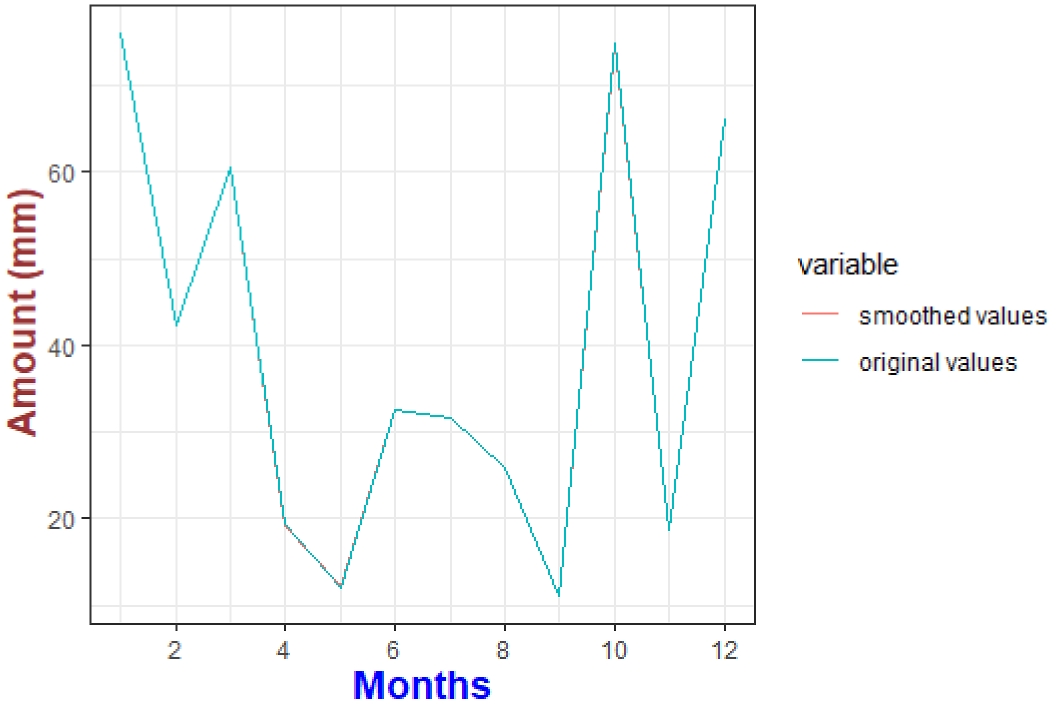

Figure 4.23. Smoothed vs Original Precipitation Values of Konya for 1-Year Time Period

The smoothed values represented by red line again shows the same movements with original variables. In the mid of year (around June), very small differences can be observed in the smoothed variables of Kalman Filter predictions; however, those differences have been counted as insignificant for the performance of Kalman Filter.

The difference between smoothed and original precipitation values can be seen directly from Table 4.16.

Table 4.16. Smoothed vs Original Precipitation Values for Konya

Smoothed precipitation values	Original precipitation values
41.0794	41.0801
18.6914	18.6919
68.1933	68.1985
45.7696	45.6652
125.5439	125.5271
33.7987	33.8120
4.6648	4.6532
13.3546	13.3637
4.2437	4.2500
24.8185	24.8183

It has been deducted from the table that the smoothed values predicted by Kalman Filter algorithm are very close to the original ones. Because of those fractional differences, the two lines for the smoothed and original variables of precipitation have seen as one line. As it has mentioned in Muğla application, those differences between the original variables and fitted smoothed variables of precipitation called as residuals. The residuals and standardized residuals have been obtained to show the differences even if they are very small in Table 4.17 for the first 10 observations and Figure 4.24 for all the residuals.

Table 4.17. Residuals and Standardized Residuals for Konya

Residuals	Standardized Residuals
0.0006	0.0000
0.0004	0.0000
0.0051	0.0002
-0.1040	-0.0045
-0.0170	-0.0006
-0.0132	0.0007
-0.0116	-0.0006
0.0090	0.0005
0.0062	0.0003
-0.0002	0.0000

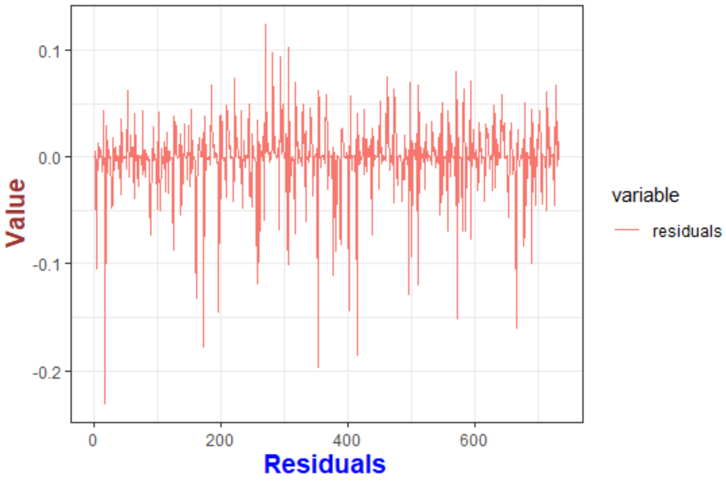

Figure 4.24. Residuals for Konya

In Figure 4.24, all the residuals are significant because of being close to the 0. It can be counted as a proof of one line in the abovementioned plots which shows the perfect match between the original and smoothed values of precipitation.

The assessments about the predicted smoothing values of precipitation have directed the study to focus on making inferences about the model which is fitted by using that smoothed values of precipitation obtained by the Kalman Filter. Firstly, MAE and MSE values have been preferred for making inferences about the model. The MAE and MSE model selection criteria for model constructed by smoothed values of precipitation have given in Table 4.18.

Table 4.18. MAE and MSE Values for Smoothed Precipitation Values of Muğla

SMAE	**SMSE**
0.0196	0.0012
SMAE: Smoothing MAE	
SMSE: Smoothing MSE	

Here, the smallest results that encountered until this part of study for the model evaluation criteria. The working mechanism of MAE and MSE is behind the idea that having the smallest values means the preferable model. Therefore, the fitted model constructed with the smoothed predictions of precipitation as a response variable obtained by the application of Kalman Filter is an accurate model based on the small values of MAE and MSE in Table 4.18 as 0.019649 and 0.0012, respectively. The precision of the fitted model has also been proved by the R^2 and *adjusted* R^2 values shown in Table 4.19.

Table 4.19. R^2 and Adjusted R^2 Values for Konya

SadjR^2	**SR^2**
0.9999	0.9999

Both *adjusted* R^2 and R^2 values are almost 1 which means that 99% change in the smoothed predictions of precipitation values can be explained by the state variables temperature, relative humidity and cloudiness. Since the higher R^2 value means the better fitted model, the model constructed with the smoothed predictions of precipitation values obtained by the application of Kalman Filter algorithm considered as a better model in the mean of explaining the variability in the response.

This part has been finished by the assessment of model evaluation criteria and the rolling window forecasting technique has been applied to Konya data as well as in Muğla data. Since the sample sizes are same in two application stations, the same algorithm has been processed and 32 forecast values have been calculated for the precipitation values shown 10 of them in Table 4.20.

Table 4.20. Forecast and Original Precipitation Values of Konya

Forecast values	Original values	Difference
27.9063	27.9000	0.0063
13.4334	13.4000	0.0334
0.0315	0.0000	0.0315
3.5272	3.5000	0.0272
67.5284	67.6000	-0.0715
38.6547	38.7000	-0.0452
29.3071	29.3000	0.0071
54.7093	54.7000	0.0093
54.4973	54.5000	-0.0026
62.2078	62.2000	0.0078

The forecasted values for precipitation have again seen very close to the original ones. This relation has been figurated as well in Figure 4.25.

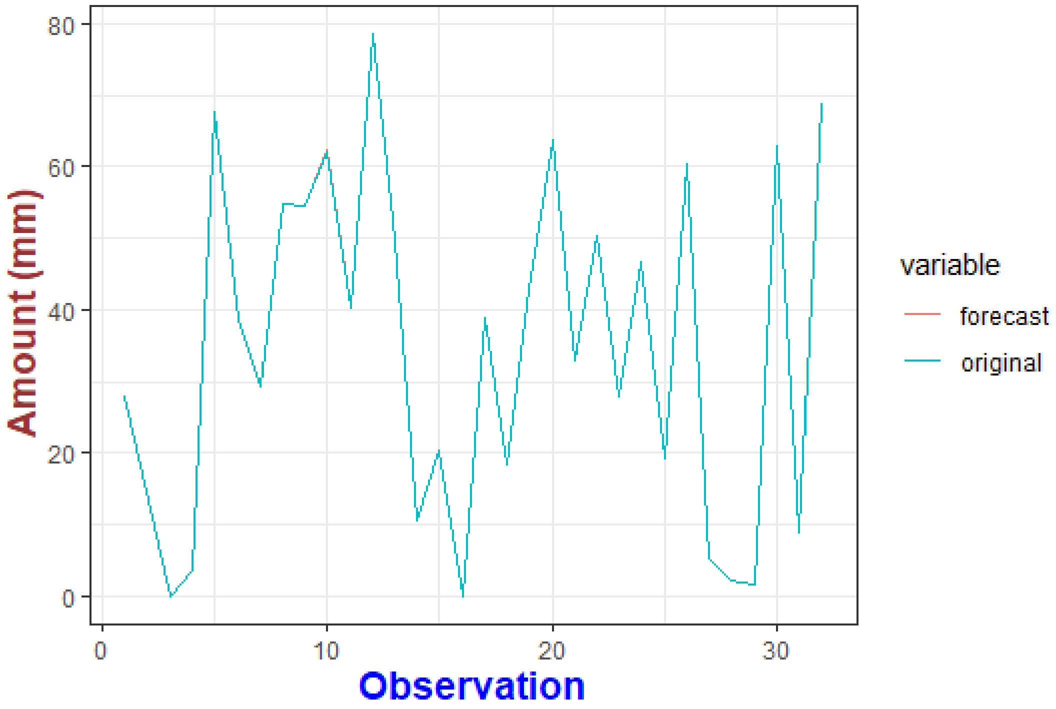

Figure 4.25. Forecasting Precipitation Values of Konya

The rolling window forecast technique has given very accurate forecast result for precipitation. All movements in the last 32 observations have been clearly fulfilled by the forecast values of precipitation shown in Figure 4.25. The assistive part on the proper results of forecast has been included the MAE and MSE values in Table 4.21.

Table 4.21. MAE and MSE Results for Forecast Values of Konya

MAE	MSE
0.0285	0.0012

Since the MAE and MSE values measure the average magnitude of the errors in the forecasts, it is important to have small values of them. The values 0.0285 and 0.00126 are small enough to say that the magnitude of errors are considerably close to zero.

After all the results of predicted precipitation values, the most accurate way is to compare the performances of all techniques in the following plot.

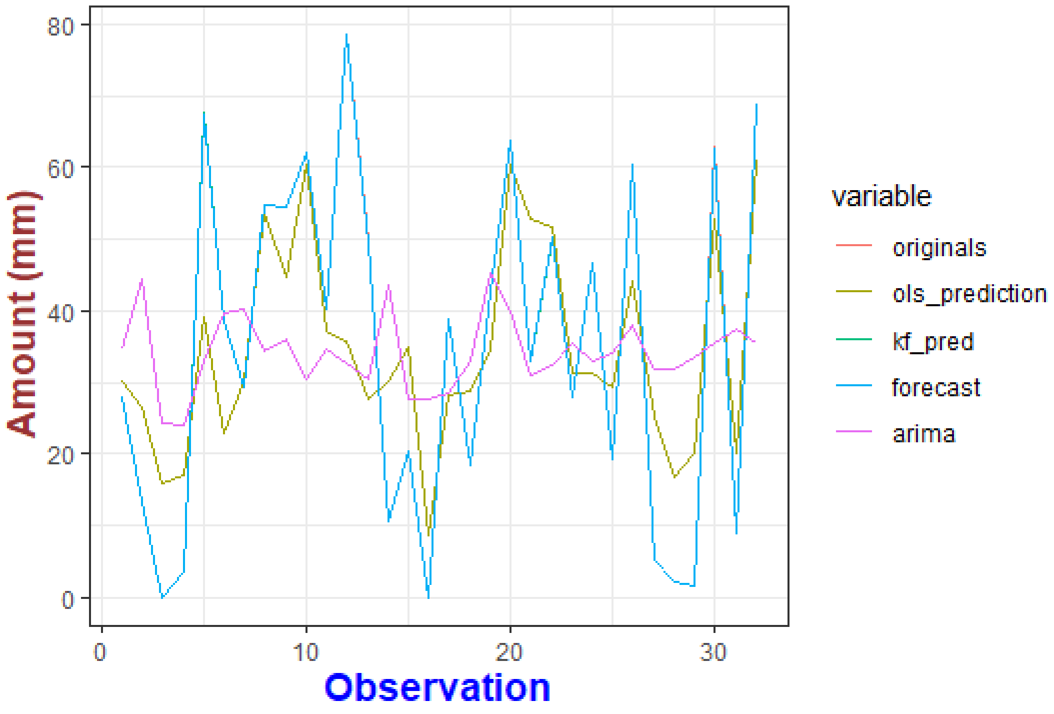

Figure 4.26. Prediction Comparison for Konya

In Figure 4.26, the green line represents the OLS estimation and purple line represents the seasonal ARIMA *(1, 0, 0) x (2, 0, 0)* model which is observed with the usage of *auto.arima()* for the prediction values of precipitation. The overlapped line represented by blue is for the Kalman Filter predictions, forecast values and the original values, respectively. OLS estimation has been considered as inaccurate in predicting the amount of precipitation. Even though the movements of that predicted values have been in the same direction of the original precipitation values, they have never caught the exact values of precipitation especially for the end points. In the ARIMA model, there is a pattern especially around the average values of precipitation. Furthermore, the ARIMA predictions have shown opposite movements of OLS estimation and Kalman Filter method. When the prediction of precipitation values have been increasing in OLS and Kalman Filter method, there is a decreasing pattern in ARIMA method. The Kalman Filter mechanism for the prediction of precipitation has worked very well not only in the direction but also in the endpoints than the OLS and ARIMA methods. The forecast values obtained by following the Kalman Filter algorithm have also performed with very small deviations from the original precipitation values.

4.4. Analysis for Ordu

In this part, the properties of Ordu data will be summarized and after that application results will be placed.

4.4.1. Data Description for Ordu

The data in the last part of application has been collected from 1950 to 2010 from the Ordu city with a sample size of 732. In the construction of the model, the temperature, relative humidity and cloudiness have chosen as independent variables which are taught to have effect on the total amount of monthly precipitation. The region of Ordu is included is a popular area with taking the precipitation. It is known that the observed amount of precipitation even in the driest spell of the year has been too much in Ordu. In fact, Ordu is an abundant region in terms of precipitation. According to the information taken from the Turkish Meteorological Service, the number of days with precipitation has been recorded almost 45% in a year. The aim of the Kalman Filter application in this part is to see the performance of the algorithm in such a region which has very high precipitation regime. Since most of the methods have had some challenges in predicting the amount of precipitation in such regions, it is important to see whether the Kalman Filter is functional in the prediction of precipitation or not. In the data taken from Ordu station, the zero values have never been observed from 1950 to 2010. This means that there was no day with any type of precipitation. This information has shown that there is no need for the application on a hybrid model. The Kalman Filter algorithm will be implemented directly to the original data. Firstly, the descriptive statistics about the data have been given in Table 4.22.

Table 4.22. Descriptive Statistics for Ordu

	Precipitation	**Temperature**	**Relative Humidity**	**Cloudiness**
Minimum	0.30	2.80	54.20	2.20
1st Quartile	49.94	8.82	69.81	5.20
Median	76.94	13.88	73.90	6.10
Mean	86.22	14.14	73.31	6.04
3rd Quartile	112.33	19.74	77.00	6.90
Maximum	269.40	26.60	90.30	9.20

The minimum value observed for the precipitation has never reached to zero even if it has been very close to it. The maximum value of precipitation is the highest value among three stations which has been recorded as 269.40 mm. The average amount of monthly precipitation over 61 years is almost 90 mm. The minimum temperature has never been under zero all time and the average temperature can be counted as moderate relatively. Being in the Black Sea region of Turkey and placed near the seaside have made the Ordu city humid. The humidity for such a long period has always followed at higher rates in Ordu around 70's. The existence of clouds has been mentioned as a return of getting any of precipitation whole year. Based on the knowledge about the precipitation regime of Ordu and the descriptive statistics of the series, the time series plot has been drawn in Figure 4.27.

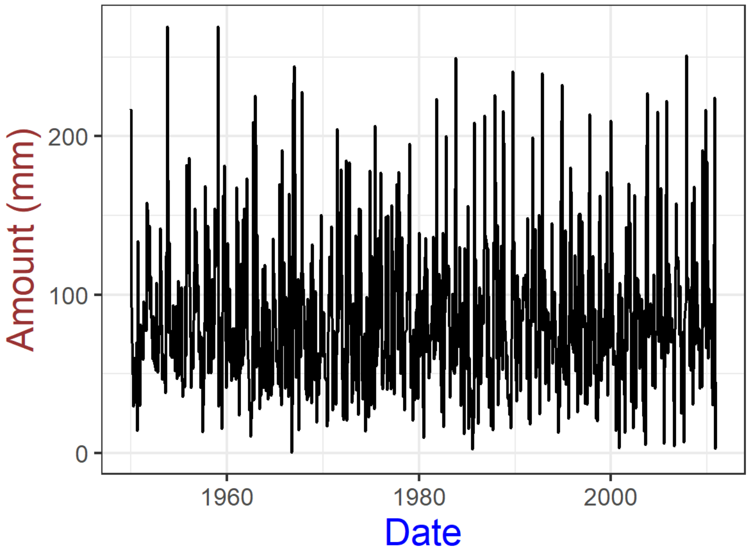

Figure 4.27. Time Series Plot of the Total Precipitation Amount of Ordu Station from 1950 to 2010

According to the Figure 4.27, it has shown that the mean of precipitation values is stationary around a constant value. It has not changed over time in the plot. There has been again a seasonal pattern. There is no evidence of trend pattern like Muğla and Konya stations. The maximum values for the amount of precipitation have observed the points very close to the 600's as expected for an abundant region in terms of precipitation.

4.4.2. Application for Ordu Station

Since the data itself has not included any zero values for precipitation, there won't be negative values for the prediction of them. Therefore, the direct values for prediction of precipitation has been calculated at the end without conducting a hybrid model.

The procedure followed for the Kalman Filter application for predicting the amount of precipitation will be the same with the previous applications in Muğla and Konya stations. The precipitation regression model has been constructed for Ordu station as a state space model form like in equations (4.4), (4.5) and (4.6). The results for the smoothed values of prediction for precipitation obtained by Kalman Filter have shown in a plot with the original precipitation values in Figures 4.28, 4.29, 4.30 and 4.31.

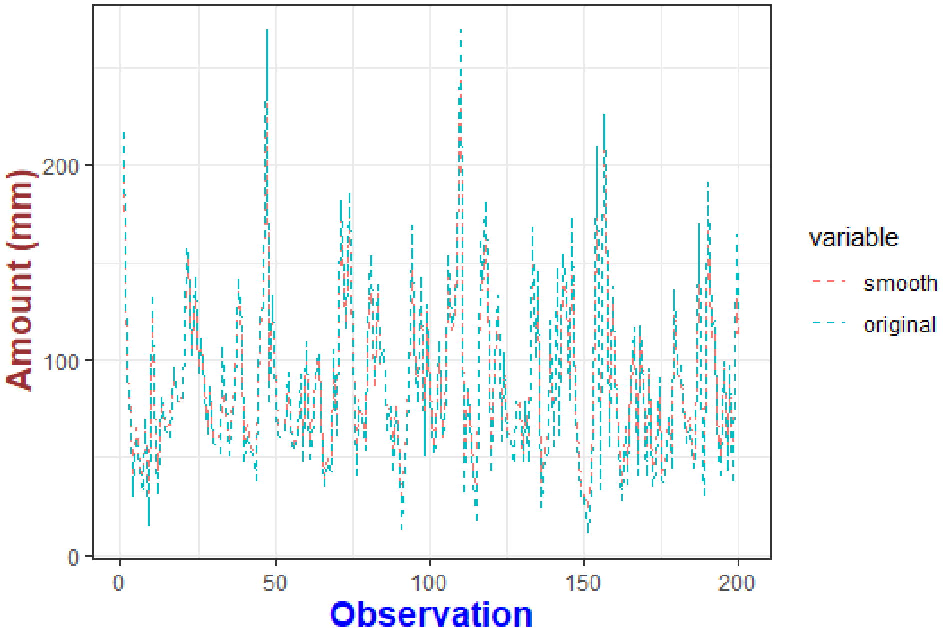

Figure 4.28. Smoothed vs Original Precipitation Values Ordu I

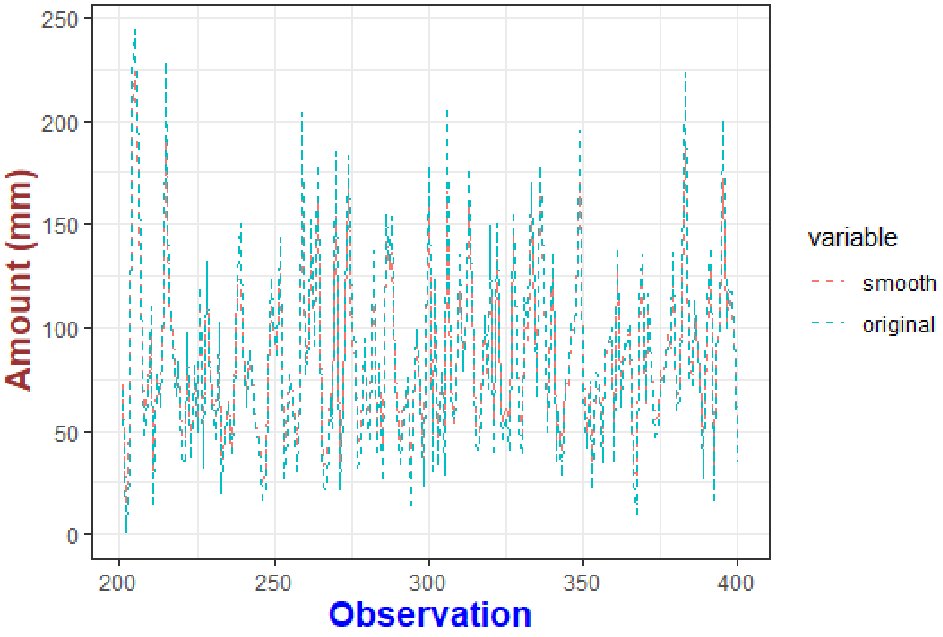

Figure 4.29. Smoothed vs Original Precipitation Values Ordu II

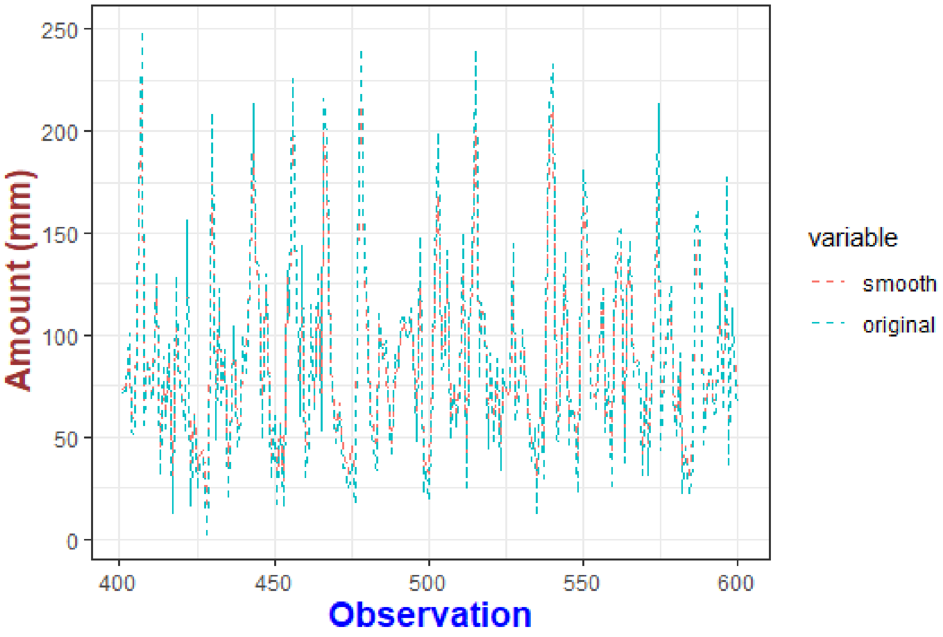

Figure 4.30. Smoothed vs Original Precipitation Values Ordu III

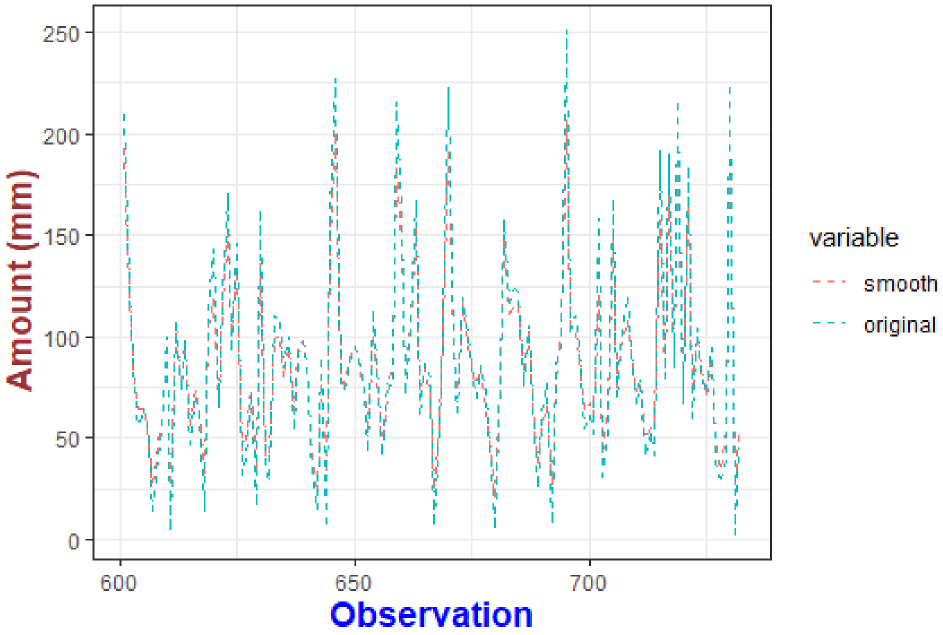

Figure 4.31. Smoothed vs Original Precipitation Values Ordu IV

.

The blue line represents the original precipitation values and the red line represents the smoothed form of predicted precipitation values by Kalman Filter. The situation for Ordu data is not same as with the situations in Muğla and Konya stations. In details, the differences between those two stations have seen clearly in this time even if the sample size is large and the plots have not been so legible. At some time points, red line which represents the smoothed values have shown itself. The detailed version for comparison of smoothed and original precipitation values has been plotted in Figure 4.32 from 1970 to 1971.

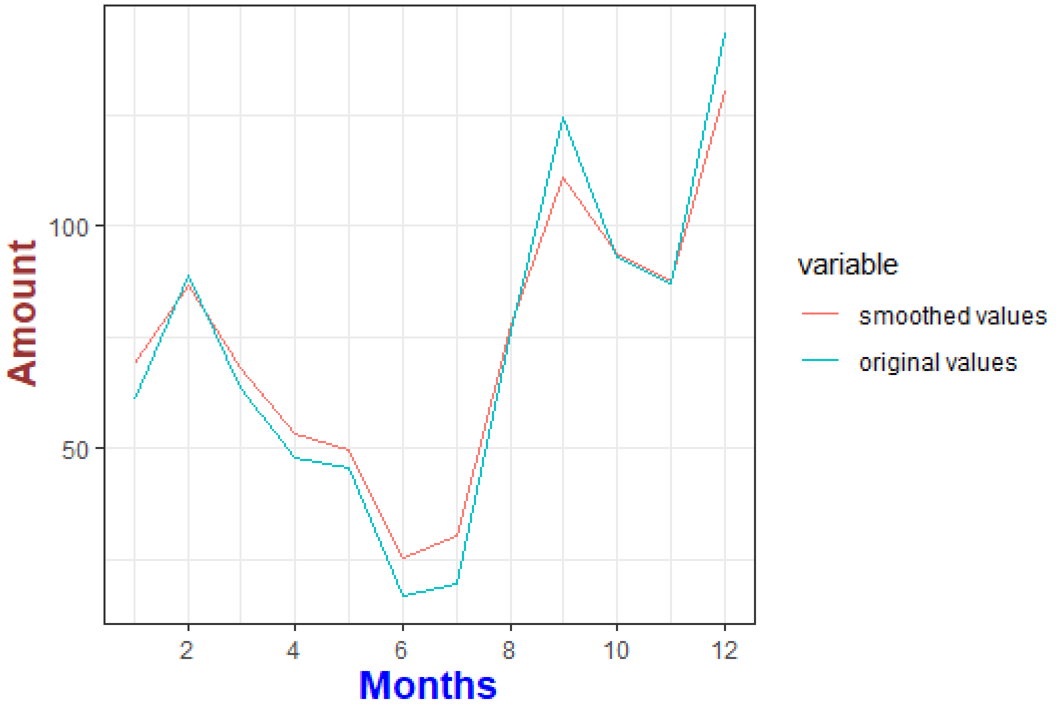

Figure 4.32. Smoothed vs Original Precipitation Values of Ordu for 1-Year Time Period

Even it has been observed differences between these values, in most of the points, the smoothed forms of predicted values of precipitation by Kalman Filter are able to catch the original ones. The magnitude of the differences can be classified as insignificant which can be proved with just first 10 observations of smoothed form of predicted and original values of precipitation in Table 4.23.

Table 4.23. Smoothed vs Original Precipitation Values for Ordu

Smoothed precipitation values	Original precipitation values
201.3365	217.0370
92.4520	92.3430
73.0330	69.4480
39.9169	29.4430
64.7675	60.3720
45.8494	45.9040
39.7047	32.1150
63.6303	70.1760
31.3831	14.0570
121.4772	133.880

The smoothed form of predicted precipitation values has been observed both above and below the original values. However; the difference has never acrossed 10. It is very significant to predict the amount of precipitation for such a region which gets every type of precipitation whole year. In brief, even though the Kalman Filter has not caught the original values exactly, it has been very close to them. The working mechanism of Kalman Filter in prediction of precipitation cannot be ignored especially in such a region. That differences have shown as being residuals and standardized residuals briefly in Table 4.24 and all of them in Figure 4.33.

Table 4.24. Residuals and Standardized Residuals for Ordu

Residuals	Standardized Residuals
15.7004	0.2828
-0.1090	-0.0023
-3.5849	-0.0682
-10.4739	-0.2224
-4.3955	-0.0846
0.0545	0.0014
-7.5897	-0.1895
6.5456	0.1641
-17.3261	-0.4177
12.4027	0.2636

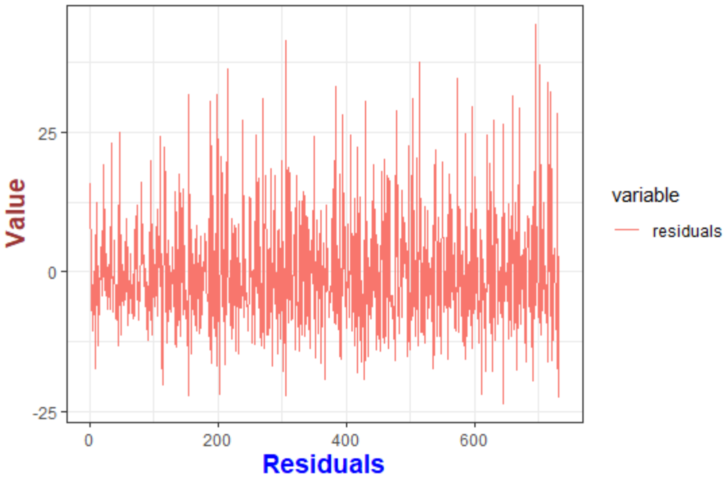

Figure 4.33. Residuals for Ordu

Although they have not been observed as small as like in the Muğla and Konya data, the residuals cannot be counted as having high values. In fact, a nonstructural visualization of the residuals in Figure 4.33 has promoted the strength of Kalman Filter method. Most of the residuals have been observed around 0 and just some of them has been measured more than the expected. However, those values do not decrease the accuracy of Kalman Filter algorithm in Ordu data. Yozgatlıgil and Turkes (2018) has applied extreme value approach to predict the amount of precipitation in Black Sea sub region of Turkey. Since Ordu station is in the same region, the MSE values have been compared. The minimum MSE value observed as 48.66 in their study. While the monthly total precipitation amount has been used in this study, they modelled the monthly maximum precipitation amount by using location parameter. At this point, the performance of Kalman Filter is evaluated as an improvement for the modelling of precipitation. A better performance from the Kalman Filter is expected when it is applied to monthly maximum precipitation amount.

The model inferences have been made according to the MAE and MSE values calculated for the model constructed with the smoothed form of prediction values of precipitation in Table 4.25.

Table 4.25. MAE and MSE Values for Smoothed Precipitation Values of Ordu

SMAE	**SMSE**
8.6693	125.4195
SMAE: Smoothing MAE	
SMSE: Smoothing MSE	

The values in Table 4.25 are the largest values that are faced with until this time for the MAE and MSE values. Actually, since the Kalman Filter has not worked in Ordu data for the prediction of precipitation as good as like in Konya and Muğla data, the values for MAE and MSE values were expected. Even if these values have been observed higher than the ones in previous applications that does not mean the Kalman Filter is not considerable to apply a region getting high amount of precipitation. For the differences obtained in Table 4.24, these MAE and MSE values have been acceptable because the observed R^2 and *adjusted* R^2 values for the explanation of the variability in the response variable are very remarkable shown in Table 4.26.

Table 4.26. R^2 and Adjusted R^2 Values for Ordu

SadjR^2	**SR^2**
0.9493	0.9495

This means that 94% change in the response variable can be explained by the variation of independent variables in the model. In other words, 94% of the variability in the smoothed form of predicted values for precipitation can be easily explained by the temperature, relative humidity and cloudiness. These very high values of R^2 and *adjusted* R^2 cannot be ignored because a 94% explanation in the variability of a model have been taught very accurate in statistics. Hence, even the performance of Kalman Filter decreases for Ordu data, it still works properly.

When the rolling window forecast technique has been applied with the same procedure, the forecast values and the actual values of precipitation seem to be very close to each other stated in Table 4.27.

Table 4.27. Forecast and Original Precipitation Values of Ordu

Forecast values	Original values	Difference
52.1067	52.1000	0.0067
158.0566	158.1000	-0.0433
30.6200	30.6000	0.0200
53.2090	53.2000	0.0090
167.9679	168.0000	-0.0320
68.6117	68.6000	0.0117
102.4917	102.5000	-0.0082
120.3793	120.4000	-0.0206
94.1070	94.1000	0.0070
65.3235	65.3000	0.0235

That is to say, our forecasting technique applied subsequently to the Kalman Filter procedure have again shown a very good performance for precipitation series. The plot of the forecast for the last 32 observations has been drawn in Figure 4.34.

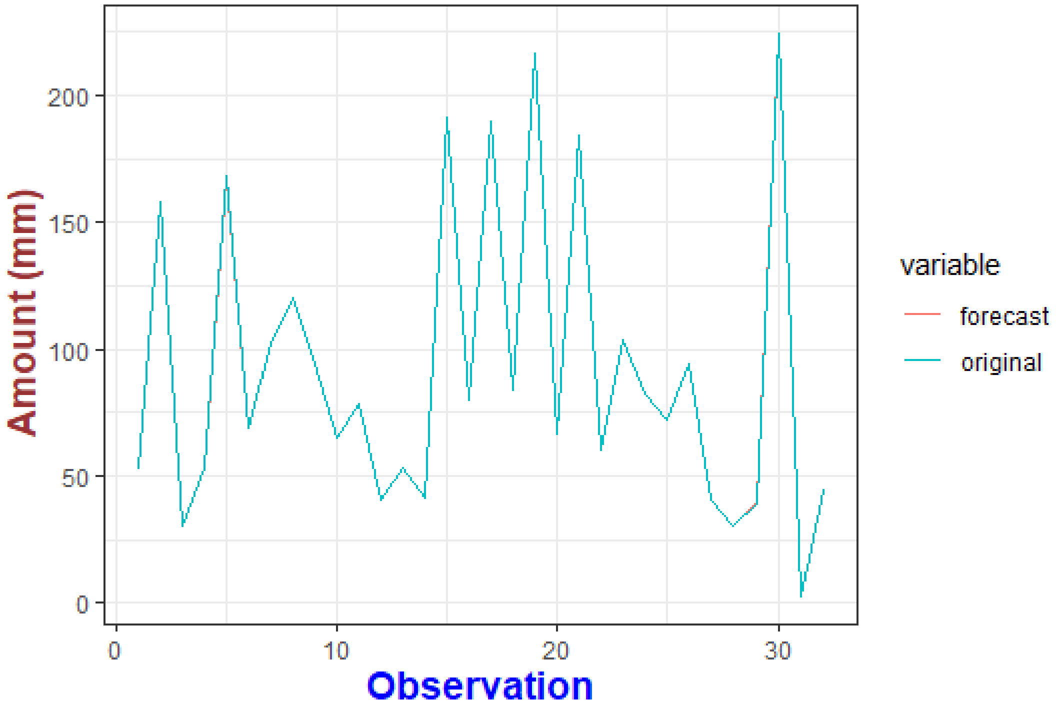

Figure 4.34. Forecasting Precipitation Values of Ordu

Since the plot of forecasted and original values of precipitation has just one line, it can be said that the precipitation values have been forecasted very well. The MAE and MSE results have get along with the deduction made in Figure 4.30 shown in Table 4.28.

Table 4.28. MAE and MSE Results for Forecast Values of Ordu

MAE	MSE
0.0195	0.0006

The significant values of MAE and MSE have shown that the magnitude of the errors in forecasting values are small which makes the forecast more reliable.

Finally, the performance comparison of Kalman Filter, OLS and seasonal ARIMA *(0, 0, 0) x (2, 0, 0)* conducted with *auto.arima()* by checking the model diagnostics and the forecast results for precipitation have been examined in Figure 4.35.

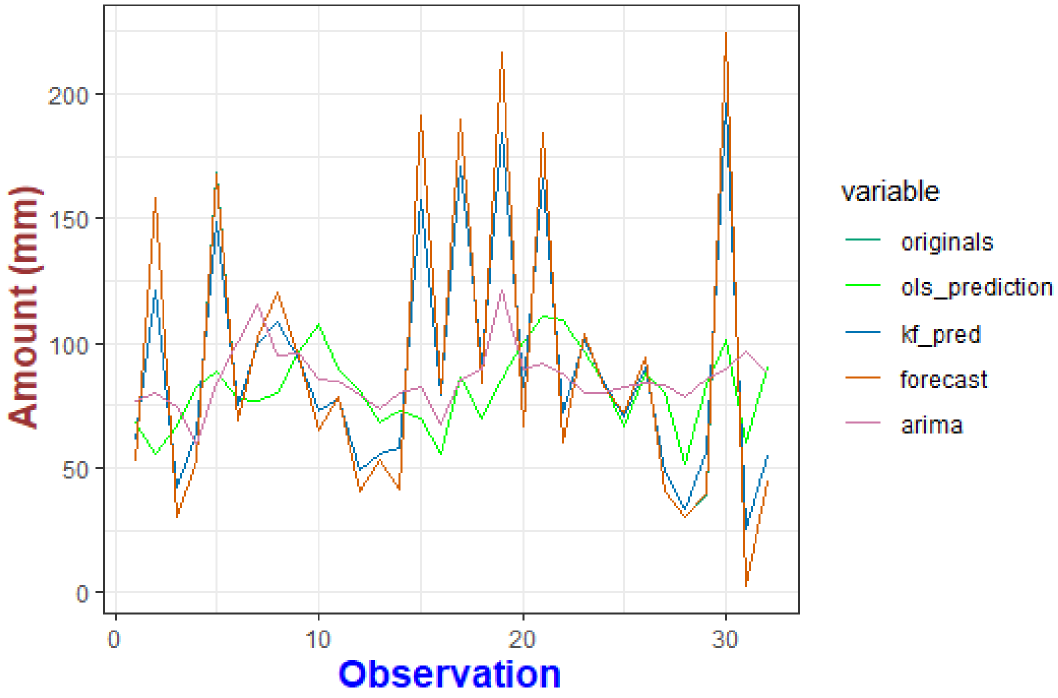

Figure 4.35. Prediction Comparison for Ordu

The forecast values represented by the red line and the original values represented by dark green line have observed together in one line as expected from the interpretations made before. The most important thing in Figure 4.35 is to be able see the difference between the performances of OLS estimation, ARIMA method and Kalman Filter algorithm in prediction of amount of precipitation. The predicted values obtained by the application of Kalman Filter represented by blue line have seen very close to the original values of precipitation. The movements in the Kalman Filter predictions are all in the same direction with the original ones. The up and down points have been almost caught by the Kalman Filter predictions. Since the structure of two values has appeared almost the same, the Kalman Filter algorithm have done its duty successfully in the mean of precipitation prediction. On the other hand, the OLS technique could not estimate even the directions of the movements of original precipitation values. While original values have had an increasing tendency, the OLS predictions have not followed the same structure in some parts. Like OLS, the performance of ARIMA method has fallen behind the Kalman Filter for predicting the precipitation. It could not catch the increases and decreases in the same way. To sum up, Kalman Filter has shown a better performance than the OLS estimation and ARIMA technique to predict the amount of precipitation in a region which is abundant in terms of precipitation.

CHAPTER 5

CONCLUSION AND DISCUSSION

The issue of predicting the precipitation is a challenging process at all since various parameters of nature involved in the procedure that are directly effective on precipitation. These parameters have effect not only on the expected amount of precipitation but also on the exact time or place of precipitation. Although it is hard to predict the amount of precipitation with these parameters, Kalman Filter method offers accurate results in prediction. In other words, to predict the amount of precipitation with its all noisy measurements is aimed with the application of Kalman Filter in this study. With the knowledge of Kalman Filter's accuracy in prediction, in this study, the monthly precipitation values have been predicted by filtering based on a state space model. The parameters which are taught as deterministic factors for the amount of precipitation chosen as temperature, relative humidity and cloudiness. With three different applications of Kalman Filter, this study aims to see the performances of filtering in different regions classified in terms of the amount of precipitation received. First station chosen as a moderate one in terms of getting precipitation is Muğla. Second one is preferred from a scarce region by precipitation is Konya. The last application region is selected from an abundant region for the precipitation is Ordu. This study has discussed the issue that how strong Kalman Filter algorithm is in the prediction of different amount of precipitation. Besides the filtering and prediction processes of the KFMR, with the smoothing part of algorithm, the most accurate values for the predicted precipitation values have been observed. KFMR has been applied especially into different series to measure the performance of it. Furthermore, the performance of Kalman Filter has been tested for the series with missing values like in hybrid model. Kalman Filter application has given same accurate results in the hybrid model as well. According to the results in all applications, Kalman Filter have given very accurate results for the smoothed values of predicted precipitation amount. Since using the Kalman Filter in predicting the amount of precipitation is not a common method, this study will bring in new perspectives to the literature with all the accurate results. The smoothed values of precipitation have been calculated very close to the original precipitation observations. It is supported by the precision of the constructed regression models with the smoothed values of predicted precipitation which have R^2 and *adjusted* R^2 values very close to 1. In the subsidiary part of application, it has been tried to estimate the amount of precipitation not only with KFMR but also OLS and ARIMA. The results have also promoted the idea of receiving very accurate results for predicted values of precipitation with the application of Kalman Filter. Although ARIMA is one of the most common techniques used for the prediction and forecast of time series models, Kalman Filter has outperformed the ARIMA in predicting the amount of precipitation for all the applications. Moreover, OLS estimation have get closed for the ups and downs of original precipitation values, however; it has not caught the actual values one by one. In other words, OLS estimation has been also failed against the Kalman Filter method for different regions in terms of getting amount of precipitation. According to the results of three application stations, the performance KFMR has shown better working mechanism than the OLS and ARIMA models in predicting the amount of precipitation. Therefore, it has been concluded that having too many zeros in data itself like Muğla station does not affect the performance of Kalman Filter at all. Furthermore, the insignificant number of zero values in the series has not affected the performance of Kalman Filter as in Konya station. Likewise, getting severe amount of precipitation does not cause a decrease in the performance of Kalman Filter. Actually, it is important to have very accurate results in such regions because it is an unusual case. To expect accurate results from Kalman Filter in a moderate region can be considerable but for an extreme case such as Konya and Ordu, these prediction results can be categorized as statistically significant. Even if Kalman Filter has not shown the same perfect performance in Ordu station as same as with Muğla and Konya stations due to being an abundant region in terms of precipitation, the smoothed values of predicted precipitation observations have been still considered as remarkable statistically. Nevertheless there are differences between smoothed and original values, they can be ignored when the total amount of precipitation in the region is taken into consideration. In conclusion, Kalman Filter has been preferred with the knowledge of being a strong method for the prediction of precipitation, and it can be said that it has handled with all the struggles which are coming from the origin of data. The performance of filter has changed according to the application areas; however, it has never gone behind the OLS and ARIMA in prediction. With the very accurate results of smoothed values of predicted precipitation, Kalman Filter has shown its noticeable performance for different application stations. In short, KFMR has reached the goal of predicting the amount of monthly precipitation with small error terms despite the tight conditions coming from the origin of precipitation data and preferred regions for the application. After having these results, it might be discussed that whether Kalman Filter shows the same performance in hourly or daily precipitation data or not. That kind of a study would be significant in terms of being a guide for preventing some natural disasters caused by sudden floods.

REFERENCES

Aagesen, L., Miao, J., Allison, J. E., & Aubry, S. (2018). Prediction of Precipitation Strengthening in the Commercial Mg Alloy AZ91 Using Dislocation Dynamics. *Metallurgical and Materials Transactions A.*

Abdul-Aziz, A. R., Anokye, M., Kwame, A., Munyakazi, L., & Nsowah-Nuamah, N. N. (2013). Modeling and Forecasting Rainfall Pattern in Ghana as a Seasonal Arima Process:. *International Journal of Humanities and Social Science* .

Antoniou, C., Ben-Akiva, M., & Koutsopoulos, H. N. (2010). Kalman Filter Applications for Traffic Management. In *Kalman Filter.* Croatia: Intech.

Asemota, O. J., Bamanga, M. A., & Alaribe, O. J. (2016). Modelling Seasonal Behaviour of Rainfall in Northeast Nigeria. A State Space Approach. *International Journal of Statistics and Applications*, *6(4)*(203-222).

Bari, S. H., Rahman, M. T., Shourov, M. M., & Ray, S. (2015). Forecasting Monthly Precipitation in Sylhet City Using ARIMA Model. *Civil and Environmental Research*.

Deep, A., Mittal, M., & Mittal, V. (2018). Application of Kalman Filter in GPS Position Estimation. *IEEE 8th Power India International Conference (PIICON).* Kurukshetra.

Du, J., Liu, Y., Yu, Y., & Yan, W. (2017). A Prediction of Precipitation Data Based on Support Vector Machine and Particle Swarm Optimization (PSO-SVM) Algorithms. *Algorithms*.

EE, E., GN, C., & Nzoiwu, C. (2017). Analysis of Precipitation Concentration Index (PCI) for Awka Urban Area, Nigeria. *Hydrology: Current Research*, *08*.

Farrell, J. A., Givargis, T. D., & Barth, M. J. (2000). Real-Time Differential Carrier Phase GPS-Aided INS. *IEEE Transactions on Control Systems Technology*.

Gaikwad, G. P., & Nikam, V. B. (2013). Different Rainfall Prediction Models And General Data Mining Rainfall Prediction Model. *International Journal of Engineering Research & Technology, 2*(7).

Gasana, E. (2012). *Intorduction to Kalman Filtering.*

Grewal, M. S. (2011). Kalman Filtering. In *International Enclylopedia of Statistical Science.* Springer Berlin Heidelberg.

Harting, C. (2010, November 28). Rainfall as an Energy Source. Retrieved from http://large.stanford.edu/courses/2010/ph240/harting2/

Hun, L. C., Yeng, O. L., Sze, L. T., & Chet, K. V. (2016). Kalman Filtering and Its Real-time Applications. In *Real-time Systems.* Rijeka: IntechOpen.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and Practice.* Melbourne, Australia: OTexts.

Ihaka, R. (2005). *Time Series Analysis Lecture Notes.* The University of Auckland: Retrieved from https://www.stat.auckland.ac.nz/~ihaka/726/notes.pdf

Keefer, T. O. (2003). Precipitation Simulation Models. In *Encyclopedia of Water Science.* Tuscon, Arizona, United States of America.

Kotowski, A., & Kazmierczak, B. (2013). Probabilistic Models of Maximum Precipitation for Designing Sewerage. *Journal of Hydrometeorology*.

Liu, T., Xlao-Hua, Y., Xue, Q., & Song, F. (2019). Application of the Weighted Markov Model in Precipitation Forecast in Beijing. *International Conference on Information Technology, Electrical and Electronic Engineering.* Sanya.

Maşazade, E., Bakır, A. K., & Kırcı, P. (2019). A Kalman filter application for rainfall estimation using radar reflectivity. *Turkish Journal of Electrical Engineering & Computer Sciences, 27:1198 - 1212.*

Meinhold, R. J., & Singpurwalla, N. D. (1983). Understanding the Kalman Filter. *The American Statistician, 37*(2), 123-127.

Menke, H. W. (2012). *Introduction to the Kalman: Application in Economics.*

Mikusheva, A. (2007). *MIT OpenCourseWare.* Retrieved from 14.384 Time Series Analysis: https://dspace.mit.edu/bitstream/handle/1721.1/46343/14-384Fall-2007/OcwWeb/Economics/14-384Fall-2007/CourseHome/index.htm?sequence=1&isAllowed=y

Nair, N. R., Sudheesh, P., & Jayakumar, M. (2016). 2-D Airborne Vehicle Tracking Using Kalman Filter. *International Conference on Circuit, Power and Computing Technologies.* Nagercoil.

Neslihanoglu, S. (2014). *Validating and Extending the Two-Moment Capital Asset Pricing Model for Financial Time Series.*

Pasricha, G. K. (2006). Kalman Filter and its Economic Applications. Retrieved 01 10, 2019, from https://mpra.ub.uni-muenchen.de/22734/

Renzi-Ricci, G. (2016). Estimating Equity Betas: What can a Time-Varying Approach Add? Nera Economic Consulting. Retrieved February 12, 2019, from https://www.nera.com/content/dam/nera/publications/2016/PUB_Estimating_Equity_Betas_0916.pdf

Ribeiro, M. I. (2000). *Introduction to Kalman Filtering: A Set of Two Lectures.*

Ribeiro, M. I. (2004). *Kalman and Extended Kalman Filters: Concept, Derivation and Properties.* Lisboa: Institute for Systems and Robotics.

Rudy, M. B., Salguero, R. A., & Holappa, K. A. (2011). A Kalman Filtering Tutorial for Undergraduate Students. *International Journal of Computer Science & Engineering Survey, 8*(1).

Rundel, C. (2013). *Logistic Regression.*

Shalizi, C. R. (2019). Logistic Regression. In *Advanced Data Analysis from an Elementary Point of View.* Cambridge University Press.

Shimkin, N. (2016). *Estimation and Identification in Dynamical Systems (048825).* Technion-Israel Institute of Technology, The Erna and Andrew Viterbi Department of Electrical Engineering.

Shumway, R. H., & Stoffer, D. S. (2016). State Space Models. In *Time Series Analysis and Its Applications with R Examples* (pp. 287-295). New York: Springer.

Sigrist, F., Künsch, H., & Stahel, A. W. (2011). A dynamic nonstationary spatio-temporal model for short term prediction. *The Annals of Applied Statistics*.

Susmel, R. (2013). *Bauer College of Business.* Retrieved from Econometrics II: Quantitative Methods in Finance II: https://www.bauer.uh.edu/rsusmel/phd/ec2-8.pdf

Tibshirani, R. (2014). *Advanced Method for Data Analysis.*

Trenberth, K. E., Dai, A. R., & Parsons, D. (2003). The Changing Character of Precipitation. *Bulletin of the American Meteorological Society, 84*(9).

Turner, H. (2008). *Introduction to Generalized Linear Models.*

Welch, G., & Bishop, G. (1995). *An Introduction to Kalman Filter.*

Welch, G., & Bishop, G. (1995). *An Introduction to Kalman Filter.* New York: University of North Carolina at Chapel Hill Chapel Hill.

Westmore, D. B., & Wilson, W. J. (1991). Direct Dynamic Control of a Robot Using an End-Point Mounted Camere and Kalman Filter Position Estimation. *International Conference on Robotics and Automation.* Sacramento: IEEE.

Wu, Y., Liu, G., Guo, X., Shi, Y., & Xie, L. (2016). A Self-adaptive Chaos and Kalman Filter-based Particle Swarm. In *Soft Computing.* Springer.

Yozgatlıgil, C. (2011). *METU OpenCourseWare.* Applied Time Series Analysis: Retrieved from http://ocw.metu.edu.tr/course/view.php?id=145#section-4

Yozgatlıgil, C., & Türkeş, M. (2018). Extreme Value Analysis and Forecasting of Maximum Precipitation Amounts in the Western Black Sea Subregion of Turkey. *International Journal of Climatology*.

Zulfi, M., Hasan, M., & Purnomo, K. D. (2018). The Development Rainfall Forecasting Using Kalman Filter. *Journal of Physics: Conference Series, 1008(1)*(012006).