DATA MINING-BASED POWER GENERATION FORECAST AT WIND POWER PLANTS

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

BY

MEHMET BARIŞ ÖZKAN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JANUARY 2014

Approval of the thesis:

DATA MINING-BASED POWER GENERATION FORECAST AT WIND POWER PLANTS

submitted by **MEHMET BARIŞ ÖZKAN** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

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

Prof. Dr. Adnan Yazıcı _____
Head of Department, **Computer Engineering**

Assoc. Prof. Dr. Pınar Karagöz _____
Supervisor, **Computer Enginering Department, METU**

Examining Committee Members:

Prof. Dr. Adnan Yazıcı
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Pınar Karagöz
Computer Engineering Department, METU _____

Prof. Dr. İsmail Hakkı Toroslu
Computer Engineering Department, METU _____

Assoc. Prof. Dr. Tolga Can
Computer Engineering Department, METU _____

Dr. Turan Demirci
Marmara Research Center Energy Institute, TÜBİTAK _____

Date: 27/01/2014

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

Name, Last Name: MEHMET BARIŞ ÖZKAN

Signature :

ABSTRACT

DATA MINING-BASED POWER GENERATION FORECAST AT WIND POWER PLANTS

Özkan, Mehmet Barış
M.S., Department of Computer Engineering
Supervisor : Assoc. Prof. Dr. Pınar Karagöz

January 2014, 81 pages

As a result of rapid depletion of non-renewable energy resources, the importance of the efficient utilization of renewable energy sources has increased all over the world and in our country in recent years. Wind has an important role in renewable energy sources with its high potential. However, compared to other renewable energy sources, wind has a spatial and temporal discontinuity characteristic so there is a need for estimating and planning of wind power generation. Wind Power Plants (WPPs) inform their wind power production forecasts for the day-ahead to an energy market and they get profit according to correctness of their declared forecasts. So, the accuracy of estimates of power generation is very important from the economic point of view for WPP owners. In addition, forecasts must be as accurate as possible for efficient and effective administration of energy by electric transmission and distribution operators. Transmission System Operators (TSOs) regulate the energy grid of all country according to energy forecasts. Because of these factors, a reliable wind power

forecast system is crucial for both WPP owners and TSOs.

The accuracy of the wind power estimations is directly proportional to effective use of Numerical Weather Prediction (NWP) data. NWP data have many parameters such as wind speed, wind direction, temperature, pressure, humidity. Data mining methods and models play an important role in order to use these parameters for wind power generation forecasts effectively. The main forecast models in the literature are grouped as physical, statistical and hybrid models. Statistical models are based on constructing a mathematical modeling between past real power data and past NWP data. In this thesis, a new statistical short term (up to 48h) wind power forecast model, namely Statistical Hybrid Wind Power Forecast Technique (SHWIP), which is based on the data mining methodologies, is presented. The main aim of the model is clustering the weather events according to most important NWP parameters for improving the accuracy of the wind power forecasts. It also combines the power forecasts obtained by from three different NWP sources and produces a hybridized final forecast. The model has been verified at Wind Power Monitoring and Forecast System for Turkey (RİTM) since June 2012 and the results of the new model are compared with well-known statistical models and physical models in the literature.

Keywords: Wind Power Forecasting, Data Mining, Numerical Weather Prediction, Dynamic Clustering, K-Means

ÖZ

RÜZGÂR ENERJİ SANTRALLERİNDE VERİ MADENCİLİĞİ TABANLI GÜÇ ÜRETİMİ TAHMİNİ

Özkan, Mehmet Barış
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi : Doç. Dr. Pınar Karagöz

Ocak 2014, 81 sayfa

Son yıllarda tüm dünyada ve ülkemizde yenilenemez enerji kaynaklarının hızla tükenmesi sonucu yenilenebilir enerji kaynaklarından verimli bir şekilde faydalanmanın önemi gittikçe artmıştır. Rüzgâr enerjisi yenilenebilir enerji kaynakları arasında önemli bir yere sahiptir. Buna rağmen diğer yenilenebilir enerji kaynaklarına göre rüzgârın karakteristik olarak alansal ve zamansal olarak süreksizliğe sahip olması nedeniyle rüzgâr güç üretiminin tahminine ve planlamasına ihtiyaç duyulmaktadır. Santraller, rüzgâr güç üretim tahminlerini gün öncesinden enerji borsasına bildirirler ve bu tahminlerinin doğruluğuna göre kar elde ederler. Ayrıca enerjinin elektrik iletim ve dağıtım kurumları tarafından etkin ve verimli bir şekilde yönetilmesi için de tahminlerin mümkün olduğunca doğru olması gerekmektedir. Elektrik Dağıtım Operatörleri tüm ülkenin elektrik hattını bu tahminler doğrultusunda düzenlemektedir. Bu faktörlerden dolayı, güvenilir bir rüzgar gücü tahmin sistemi hem santral sahipleri hem de dağıtım operatörleri

için önemlidir.

Güç tahminlerinin doğruluğu hava tahminlerinden etkin bir şekilde faydalanmakla doğru orantılıdır. Hava tahminleri rüzgâr hızı, rüzgâr yönü, sıcaklık, basınç, nem gibi birçok parametreye sahiptir. Bu parametrelerin rüzgâr güç üretim tahminlerinde etkili bir şekilde kullanılmasında veri madenciliği yöntem ve modelleri önemli bir yere sahiptir. Literaturdeki temel tahmin modelleri fiziksel, istatistiksel ve hibrid olmak üzere gruplandırılır. İstatistiksel modeller geçmiş güç verisi ile geçmiş hava tahmin verisi arasında matematiksel bir modelleme kurma esasına dayanır. Bu tezde İstatistiksel Hibrid Rüzgar Enerjisi Tahmin Yöntemi (SHWIP) isimli veri madenciliği yöntemlerine dayalı, yeni kısa süreli (48 saatlik) tahmin yöntemi sunulmuştur. Modelin ana amacı hava olaylarını en önemli hava parametrelerine göre sınıflandırıp rüzgar güç tahmnilerinin doğruluğunu geliştirmektir. Model aynı zamanda üç farklı hava tahmininden elde edilen güç tahminlerini birleştirip hibrid edilmiş tahminleri oluşturur. Model, Rüzgar Gücü İzleme ve Tahmin Sistemi (RİTM) projesinde Haziran 2012 den beri kullanılmakta ve sonuçları literatürde yer alan çok bilinen istatistiksel ve fiziksel yöntemlerle karşılaştırılmıştır.

Anahtar Kelimeler: Rüzgar Enerjisi Tahmini, Veri Madenciliği, Sayısal Hava Tahminleri, Dinamik Kümeleme, K-Ortalamalar

To my little niece and nephew Öykü & Emre

ACKNOWLEDGMENTS

I would like to express my gratitude to my supervisor Assoc. Prof. Dr. Pınar Karagöz for her guidance, criticism, and her confidence in my abilities and character.

I wish to thank Prof. Dr. Adnan Yazıcı, Prof. Dr. İsmail Hakkı Toroslu and Assoc. Prof. Dr. Tolga Can for their valuable comments during my thesis presentation.

I would like to express my gratefulness to TÜBİTAK MAM Energy Institute manager Dr. Burhan Gültekin for his continuous encouragement during the thesis process.

I would like to thank to Dr. Turan Demirci, Dr. Dilek Küçük and Erman Terciyanlı for their great contributions to my work and help throughout this study.

A special thanks to Serkan Buhan who implemented ANN and SVM forecast models used in the comparison and provided great motivation and guidance during this work.

I would like to express my deepest gratitude to my family for their support, patience, and encouragement throughout my studies.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

RİTM	Wind Power Monitoring and Forecast Center for Turkey
NWP	Numerical Weather Prediction
TSO	Transmission System Operator
WPP	Wind Power Plant
SHWIP	Statistical Hybrid Wind Power Forecast
GFS	Global Forecast System
ECMWF	European Center for Medium range Weather Forecasting
TÜBİTAK	The Scientific and Technological Research Council of Turkey
YEGM	General Directorate of Renewable Energy of Turkey
DMİ	General Directorate of Meteorology of Turkey
TEİAŞ	Turkish Electricity Transmission Company
WAsP	Wind Atlas Analysis and Application Program
CFD	Computational Fluid Dynamics
MOS	Model Output Statistics
ANN	Artificial Neural Networks
SVM	Support Vector Machines
PCA	Principal Component Analysis
ADSL	Asymmetric Digital Subscriber Line
GPRS	General Packet Radio Service
WRF	Weather Research and Forecasting Model
ALADIN	Air Limit Adaptation Dynamic Development International
NMAE	Normalized Mean Absolute Error
NRMSE	Normalized Root Mean Squared Error
ARMA	Autoregressive Moving Average
ARMAX	Autoregressive Moving Average with Exogenous

CHAPTER 1

INTRODUCTION

1.1 Overview

The renewable energy sources has become very popular in the past years due to their beneficial features. High number of studies have been conducted in order to benefit clean energy sources properly in many countries. Also in our country, this issue is noticed by the governmental institutions and energy policies are regulated accordingly. In our country, several studies have been conducted as well in order to benefit from the wind energy in the country [1, 2, 3, 4]. In addition to this, researches for the sun energy is planned to be conducted.

The energy demand of countries has been increasing day by day due to growing technological needs and increasing population of countries. Fossil fuels such as fuel, coal and natural gas are the main energy sources that meet this energy demand. Also in the past years, new nuclear power plants were built in order to obtain more energy. Although these sources are the main energy providers, they have some negative aspects. First of all, they are not healthy energy resources for the humankind. One of the main problems that people try to solve in the last years is the global warming issue. Researchers from all over the world are trying to reduce the negative effects of the global warming and countries spend high amount of money for this research every year. The most important reason for the global warming problem is the excessive usage of non-renewable energy sources [5]. Apart from global warming problem, these sources also cause some other direct dangers. For instance, in the past years, because of the problems in

the nuclear power plants, many people have died especially from the cancer [6]. Apart from the health diseases, obtaining energy from these sources is highly costly process. Governments spend high budgets in order to construct new plants [7]. However, the main challenge in these resources is that they are diminishing day by day since they are not renewable. In addition, countries are regulating their energy policies according to these constraints.

Due to these negative effects, countries have realized the importance of renewable energy sources and they have started to work on new methods in order to benefit from their clean energy sources more effectively in the last years. The main renewable energy sources are hydro, wind, sun and biomass. Recently, new hydro and sun plants and wind farms have been built and countries are encouraging the investors to build new plants. These sources are described as alternative energy sources because they do not have the potential of satisfying the energy needs of all area but they commonly help the fossil fuels and decrease the excessive usage of these sources. In comparison to the fossil fuels, these sources are described as clean energy sources because they do not have any negative effects on the environment. In addition, these sources are renewable and there is not any diminishing problem for these sources as long as the world exists. The other advantage for these sources is the cost issue. Building a new wind power plant or a sun plant is not as expensive as building a new nuclear power plant. Due to this factor, investors consider this developing new area as an opportunity for their economic concerns.

Among all renewable energy sources, wind is one of the most significant types with its rich potential. In Turkey there are nearly 60 Wind Power Plants (WPPs) are in operation with the installed capacity of 2400 MW [8]. This number will reach 10 GW in near future by building new WPPs whose agreements have already been signed by the government and WPP owners [9]. As of our country, there is a target to promote this total capacity to 20 GW in 2023 in the 100th anniversary of the establishment of the Turkey Republic [10]. Generally, WPPs are located in the western part of Turkey as shown in Figure 1.1. [10, 11]. In the Eastern part of Turkey there is not too many WPPs however this region has too much sun potential rather than wind.

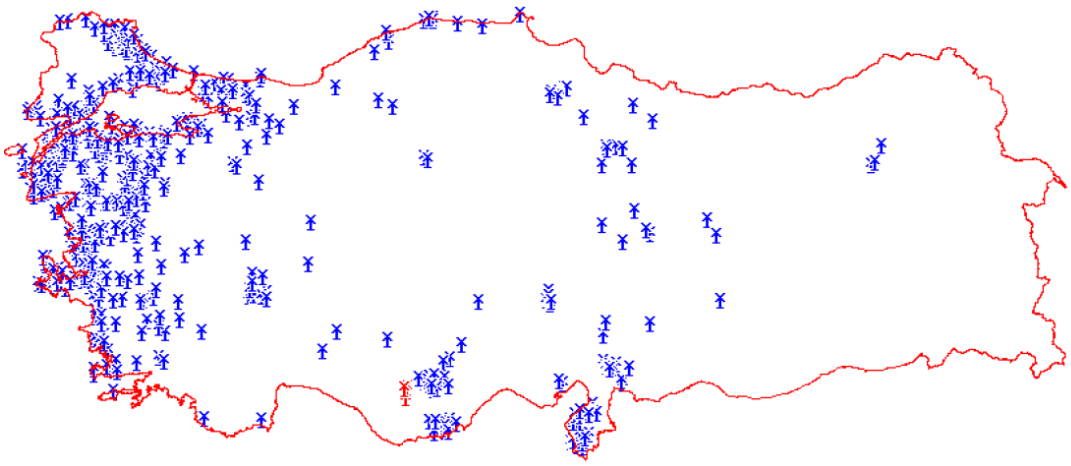

Figure 1.1: Distributions of WPPs in Turkey

When we compare wind with other renewable energy sources it has one basic different different characteristics. Wind energy is so fluctuating and it is not as stable as the other clean energy sources. Because of this factor wind energy has to be under controlled carefully. Generally, countries are building wind power monitoring and forecast centers in order to control this variable energy source.

In [11], general structure of the real time wind power monitoring center of Korea and wind power prediction model are given. In the center, they are using GFS as NWP source and they are producing wind power forecasts based on GFS data. In the center, a physical and a statistical model are combined as a hybrid short term wind power forecast model. In [12], nearly 20000 MW installed wind power capacity in Spain is monitored by The Special Regime Control Center (CECRE) and power forecasts are generated for 48 hours. Apart from short term forecasts, also long term wind power forecasts up to 10 days are produced in the center. The forecasts for the Irish transmission system are verified in [13] for 21 representatives WPPs in the system. In their system, for the day-ahead forecasts, the average NMAE rates for individual WPPs is generally 14 % and they are monitoring nearly 1500 MW installed capacity in the system. Ernst and et.al. give the details of the forecasts of German Transmission System Operation Centres in [14] with 11850 MW installed wind power capacity. In their applications, they are using Artificial Neural Network (ANN) for their

short term wind power forecast models and the results are compared with the persistence model.

In order to control the wind energy in Turkey a new project has been started in 2011. The name of the project is Wind Power Monitoring and Forecast Center for Turkey (RİTM) and project is implemented by TÜBİTAK for General Directorate of Renewable Energy of Turkey (YEGM) [3, 8]. The aim of the center is monitoring the production of the WPPs in the country in real time and generating a reliable forecast system in the center.

The most important issue in a forecast center is the consistency of the wind power forecasts. A reliable forecast system is so crucial for the Transmission System Operators (TSOs) who are responsible from the management of the electricity grids in the country. In Turkey, TEİAŞ manages the flow of the energy sources in the country and it has become one of the stakeholders of the project in 2013. They use the wind power forecasts which are obtained from the RİTM center in order to plan their energy projection in a two days period. Apart from TSOs, the forecasts of the center are so important for the WPP owners and they are used by them. In Turkey there is an energy market, similar to other countries, and WPP owners are declaring their two days forecasts to this market. During these two days, if the given forecasts are consistent with the real production they get huge profit and if there is too difference between them then they pay penalty. Therefore a trusted wind power forecast system is too crucial for the WPP owners from the economic point of view. Currently, wind power forecasts are given to WPPs as a free service but after 2014 YEGM will sell the forecast to WPPs.

In such systems, generally very short term, short term, regional and probabilistic forecasts models are generated. Very short term forecasts are usually up to 6 hours and they are dynamically updated. In our country there is not an energy market for very short term forecasts. In regional forecasting, total forecasts of the regions are obtained by also including to forecasts of WPPs which are not monitored by the system with an up-scaling algorithm. Since some of the WPPs are not monitored by the system the total forecast is not very accurate

but the forecast gives an idea about the whole region. In probabilistic wind power forecasting, the forecasts are given with their confidence intervals. These forecasts are important for the TSOs while deciding on the extreme cases before regulating the energy grid. The most important forecast is the short term wind power forecast. In our country the energy market is for 48 hours but in some countries this duration can be extended to 72 hours. WPPs are declaring their two-day-ahead forecasts to this market and their profits are proportional to the correctness of their forecasts during these two days. In the near future, it is planned that a market for very short term wind power forecasts will be available.

1.1.1 Contributions

The motivation behind this work is constructing a trusted short term wind power forecast model. In this thesis the details of a new data mining based short term wind power forecast model is presented. It is a statistical model and it is based on clustering the weather events in a dynamic way. The model has been operational in RİTM center since June 2012. The performance of the model is compared with some well-known statistical models in the literature and a physical model. In comparison to the other statistical models, the proposed model requires less amount of historical data and it also produces better forecasts even if with only month long training data. The other statistical methods generally need at least one year of historical data in order to produce reliable forecasts. So, the proposed model especially is important for the new established plants with little amount of historical data. In general, the new model has better performance compared to other models. The model performance is compared with two statistical models (ANN and SVM) and a physical model in 14 WPPs which are monitored in RİTM center from start of the project. In 11 of these 14 WPPs, the proposed model has the lowest error rate. Also there was a three month test period for RİTM project which is organized by the client of the project. In this test period of the project, the proposed model is used for the testing performance of the system and it's results are compared with a forecast tool which is developed by another foreign institution. The model has passed the test criterions and after the test results of the project and with a new law, all WPPs in the country

have to participate to this project in the near future. Currently 20 WPPs are monitored by the system however this number will increase to nearly 60 at the end of the 2014.

1.2 Organization of the Thesis

The rest of this thesis is organized as follows.

In Chapter 2, the short term forecast models and related work in the literature are presented. The short term forecast models in the literature generally are classified as Physical, Statistical and Hybrid Models. General structures of these models with real world examples are given in this chapter.

In Chapter 3 of the thesis, background information about the methods used in the application is described. In the application of the model, some algorithms in the literature such as k-means, dynamic clustering are used and their details are presented in this chapter.

The general architecture of the RİTM project is given in Chapter 4. The details of the data used in the forecast application are expressed in this section. In addition, data acquisition and data storage parts of the system are expressed in detail.

In Chapter 5, the proposed new statistical model namely, SHWIP , is described in detail. The model is based on dynamically clustering the NWP data and finding the similar weather events and classifying them. All of the computational details of the model are presented in this section.

The evaluation results of the model are given in Chapter 6. The results are given for the 14 WPPs which are monitored since June 2012. Evaluation results of the model are compared with other well-known forecast models in the literature includin as ANN, SVM and a physical model.

Thesis is concluded with further remarks and possible future works in Chapter 7.

CHAPTER 2

RELATED WORK

The main process in wind power forecasting is transforming the meteorological forecasts obtained from medium-scale quantitative weather forecasting models into power forecasting. All of the models in the literature take NWPs as initial input and produce power forecast of the WPP. Due to this factor, performance of a power forecast model is directly proportional to correctness of the NWP.

In general, the forecast modules also take the wind power curve of the WPP as input. The power curve of a WPP is a 360x251 matrix and it is used to specify the amount of power to be generated at some particular wind speed and directions by the WPP. These curves are constructed by historical data by use of some simulation software. Within the scope of the (RİTM) project, these curves are constructed by WindSim and WAsP software and an example power curve image for a WPP is given in Figure 2.1. [15, 16].

The wind power forecast models in the literature are classified as three groups namely Physical Models, Statistical Models and Hybrid Models [1, 17]. Physical models are based on determining the wind speed at turbine hub height rather than whole plant area. On the other hand, statistical models are based on constructing a mathematical modeling between past historical NWP data and power data. Most of the common statistical models used in the literature for short term wind power forecasting are ANN and SVM. Apart from these two models there are some hybrid models which combine the two approaches. The details of these models are presented in the following sub-sections.

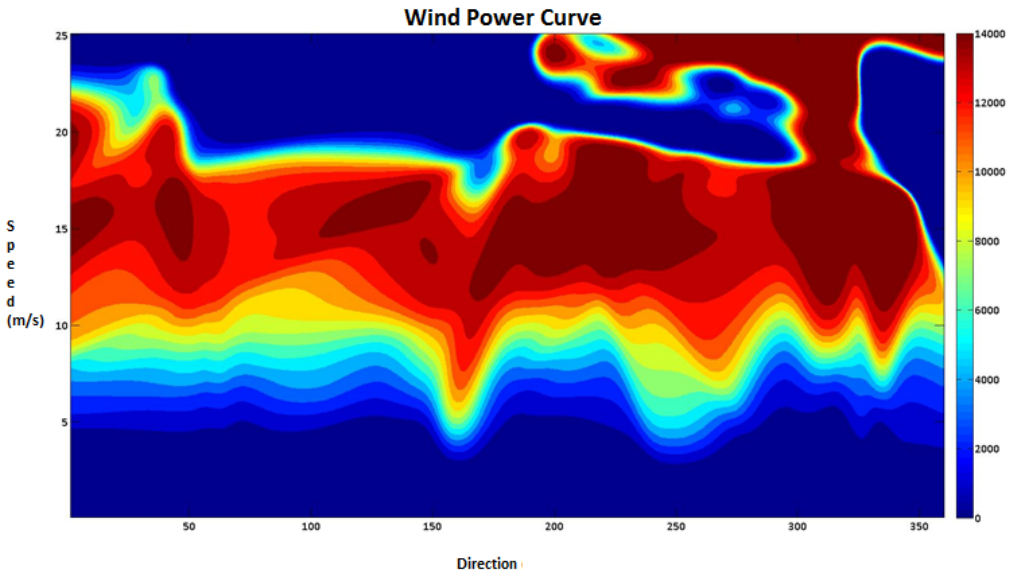

Figure 2.1: Wind Power Curve of a WPP

2.1 Physical Models

Physical models are based on the physical conditions in the WPP area. The aim is reducing the NWP to turbine hub height. Generally, these models take the following data as input:

- Numerical Topography of the WPP area
- Physical conditions of the WPP (turbine location, roughness, obstacles)
- NWP
- Computational Fluid Dynamics (CFD)

General architecture of the Physical Model used in RİTM project is given in Figure 2.2. The WPP area with turbines is modelled in 3D space by using modeling software WindSim and WAsP [15, 16]. Then NWPs are reduced to turbine hub height and estimated wind speed and wind direction values are passed through to power curve in order to produce power forecasts. At this stage methodologies like CFD are applied in order to generate local wind speed

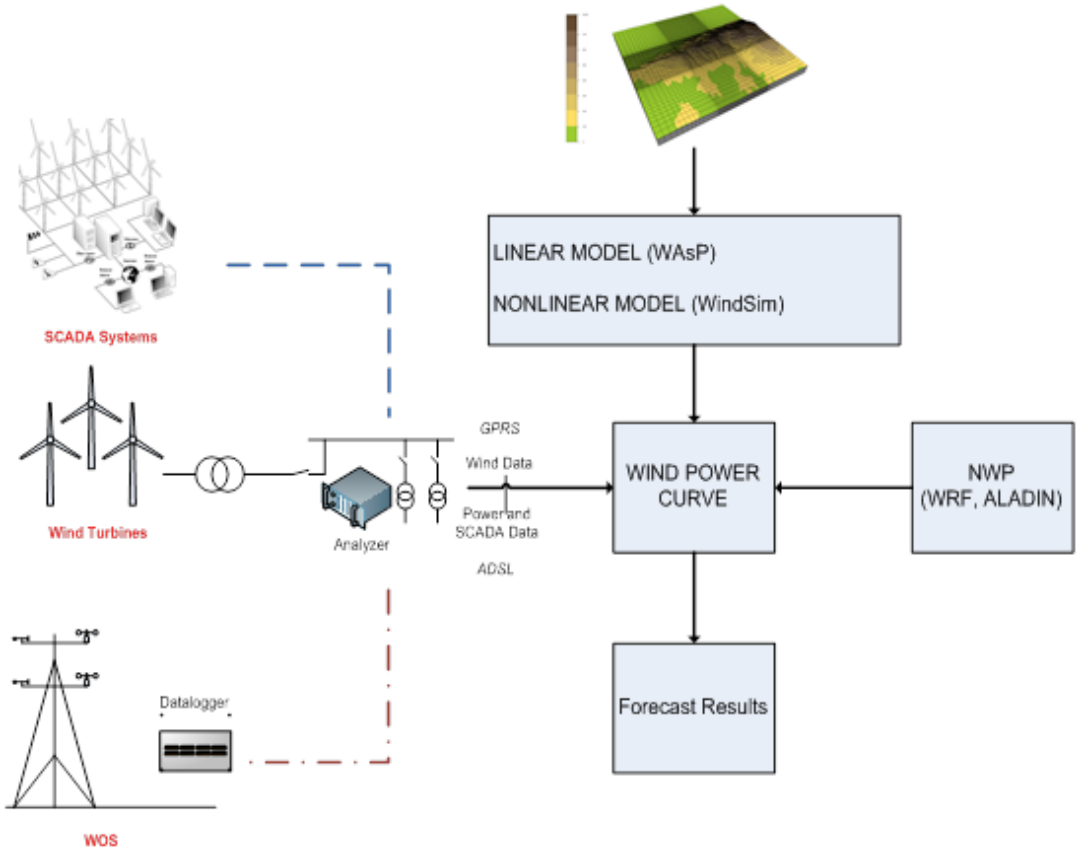

Figure 2.2: Physical Forecast Module Architecture in RİTM project

in turbine area. Also in some models in the literature, as a final step Model Output Statistics (MOS) methods are applied to local wind speed in order to reduce error rate [18].

Two of the famous physical wind power forecast models are the Prediktor [19] and Casandra [20]. Prediktor tool is developed by Landberg in Denmark and it is one of the oldest forecast tools in the literature. It uses the wind speed and wind direction values obtained by the NWP and transform these forecasts to local site area. Finally, these reduced values are used with the power curve of the WPP to obtain the power forecasts [17, 19]. Casandra tool uses GFS as NWP source. However, rather than directly using GFS forecasts, it fixes the NWP according to several physical characters in the WPP region, such as surface pressure, cloud cover and rainwater. It also uses the wind power curve of the WPP obtained by the MOS predicted variables applied on the data [17, 20].

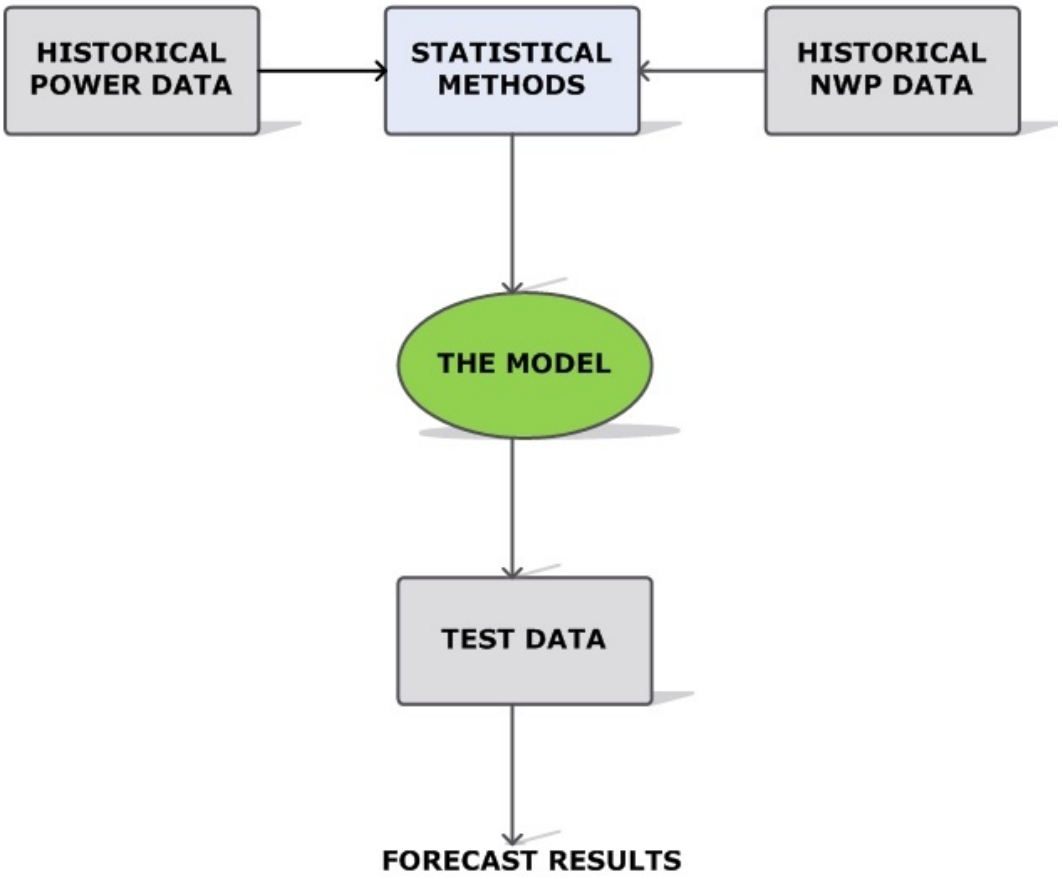

Figure 2.3: General Architecture of the Statistical Models

2.2 Statistical Models

The statistical wind power forecast models are based on extracting the hidden mathematical relation between the NWP, which is the basic input of the wind power forecast models, and the production data. General architecture of these models is presented in Figure 2.3. In the initial stage of such models, the model is trained by the past historical power and NWP data and model is generated. After the model is obtained, the test data (generally 48 hours data for short term wind power forecasts) is given to this model and wind power forecast results are obtained. General statistical models in the literature are presented in the following sub-sections.

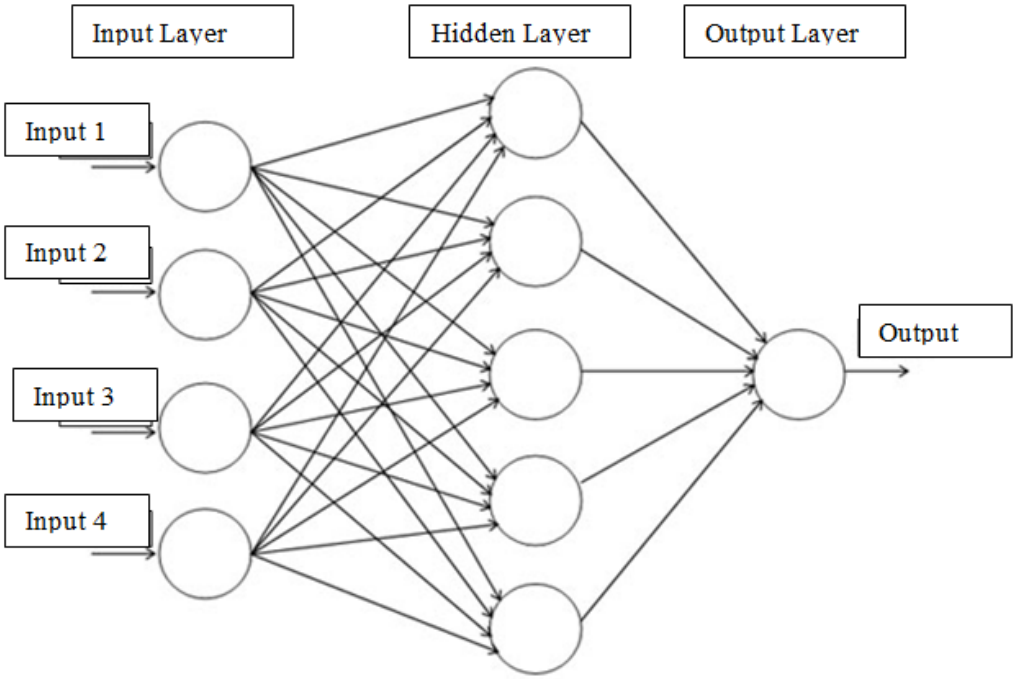

Figure 2.4: General Architecture of the ANN Models

2.2.1 Artifical Neural Networks

Artificial Neural Network (ANN) models are based on the principle that perception obtained through historical data is reflected on forecasting with the logic of neural system. General architecture of the ANN models is given in Figure 2.4. In the input layer, all of the parameters which can affect the correctness are provided as input. Commonly, these parameters are wind speed, wind direction, temperature and humidity. With the selection of appropriate number of neurons in Hidden layer, they perform the weighting compatible with the reactions to system inputs.

One of the famous forecast models in the literature which is based on ANN is the Wind Power Management System which is developed by Ernst and et.al. [14]. This model is used by four German TSOs and in Italy and Austria. The model combines three different NWP sources in the learning process of the model [14, 17]. In their work, Lange and et.al. [21] presented their different multi-level

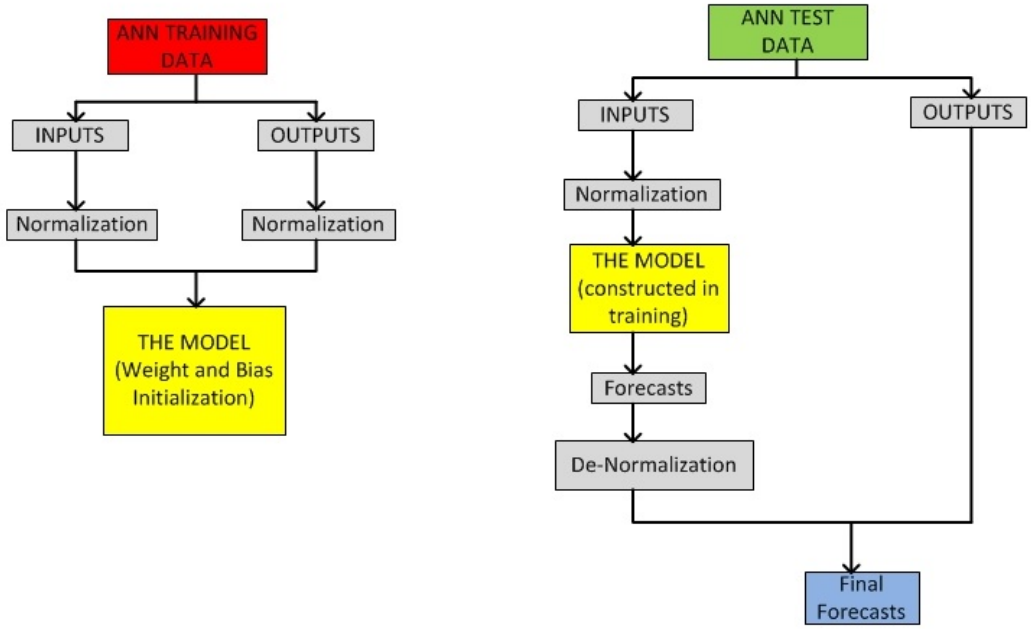

Figure 2.5: Block Diagram for a typical ANN Model

ANN models to compare the error rates between the forecast results of combined NWP source and individual NWP sources. In their paper, Mihai and Gilda Gavrilas present a new ANN model which is based on fuzzy representation of the wind direction parameter [22].

Block diagram of the training and test part for a typical ANN model is given in Figure 2.5. In the training stage, the neurons are determined by applying some normalization process and model is constructed with weight and bias initialization. In the test stage, the forecasts are generated according the model constructed in the train and the values are de-normalized in order to convert to actual values according to WPP maximum capacity.

2.2.2 Support Vector Machines

Support Vector Machines (SVM) model is similar to ANN model but it is a hyper plane solution used especially for classification. General SVM model architecture is represented in Figure 2.6. The problem is transferred into a hyper plane and

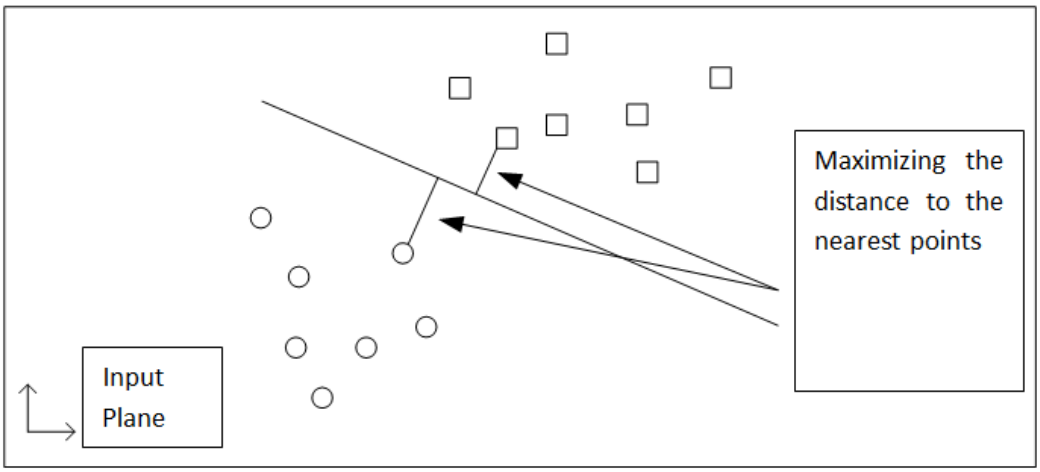

Figure 2.6: General Architecture of the SVM models

solved in a different dimension. It ensures that the different sets are separated considering the maximum margin.

Mathematical modeling behind in a typical SVM model can be expressed as follows:

$$y = wx - b \tag{2.1}$$

Where,

- y= Output vector
- w= Normal value of hyper plane
- x= Input vector
- b= Shifting parameter of the line drawn

In order to separate the data sets via a nonlinear line, Lagrange multipliers and Kernel functions are used [23]. Mathematical statement for this case transforms as follows:

$$y = \sum a_j K(x_j, x) - b \tag{2.2}$$

In this formula, a stands for Lagrange multipliers, K stands for Kernel function measuring the similarity or distance between x and xj.

Training and test block diagram of an SVM model is similar to given diagram in Figure 2.5 for an ANN model. The only difference in the diagram for SVM is, in the training part after the normalization process of the outputs values are converted to a Feature Vector within the range of 0-100. Test procedure is the same as in ANN test diagram presented in Figure 2.5.

Visionpoint is a SVM based wind power forecast model which is used in U.S. and developed by WindLogics [17]. In the model, three different NWP sources are used and they are combined with different weights in order to obtain most suitable forecast data. SVM uses wind speed and wind power generation as inputs of the model and makes a conversion from wind speed to wind power. The model retrains itself in every month. For the short term wind power forecasting, the model's error rates are varying between 12 % to 20 % [17].

In [24], Zeng and Qiao present their SVM based short term wind power forecast model. They apply SVM model both for very short term and short term wind power forecasting and they compare their model with a radial based neural network model and persistence. In both of the models, SVM based model performs better than the other two models.

In [25], Sanz and et.al. are trying to estimate the wind speed for each turbine in the WPP area by using an evolutionary support vector regression algorithm. The model is tested on Spanish wind farms. They are applying statistical downscaling on NWP data obtained by GFS. According to evaluation results their error rates between estimated wind speed and real wind speed values for a typical turbine vary between 1.7 % to 2%. However, they do not give their error rates for wind power forecast values but estimating the wind speed is one of the main concerns all of the wind power forecast models.

2.2.3 Other Statistical Models

Apart from the black-box statistical models such as ANN and SVM, there are some other statistical wind power forecast models and tools in the literature. These models make time series analysis between historical power data and find

the relation between estimated wind speed and wind power.

The Wind Power Prediction Tool is working in such manner [17]. It is developed by Technical University of Denmark and in the model, the power forecasts of a specific wind farm are estimated by the usage of other neighbor WPPs in the same region. The whole area is divided in to sub-areas and each of them assigned to a WPP. Then for each sub- areas wind power forecasts are up scaled and combined in order to obtain the final forecasts which are in 30 min resolution [17, 26].

In [27], Kusiak and et al. present their five different data mining based wind power approaches and compare the performance of the models. Two of the models presented in the paper are based on PCA and k-Nearest Neighbor Search. PCA is used in order to determine the most important wind parameters in forecasting. They transform wind speed and wind direction values to final values by applying feature selection to all wind speed and wind direction parameters and obtain the power forecasts.

They also introduce another data mining based approach which is based on k-NN search. In their model, they are trying to find the most similar past neighbors wind speed data and find a similarity between them. Also in their models they are trying to integrate k-NN model with their ANN based model. According to evaluation results of their works, the ANN based algorithm has better performance compare the other four models. However, integrating their ANN model with k-NN approach increasing their error rates and combined model is producing less accurate and unstable predications. Proposed models are tested with both very short term and short term (up to 84 hours).

2.3 Hybrid Models

Apart from the statistical and physical models, there are some hybrid models in the literature which combine the former two approaches into one model. General structure of such models is provided in Figure 2.7. In these models, apart from the NWP sources also physical characteristics of the terrain such as roughness

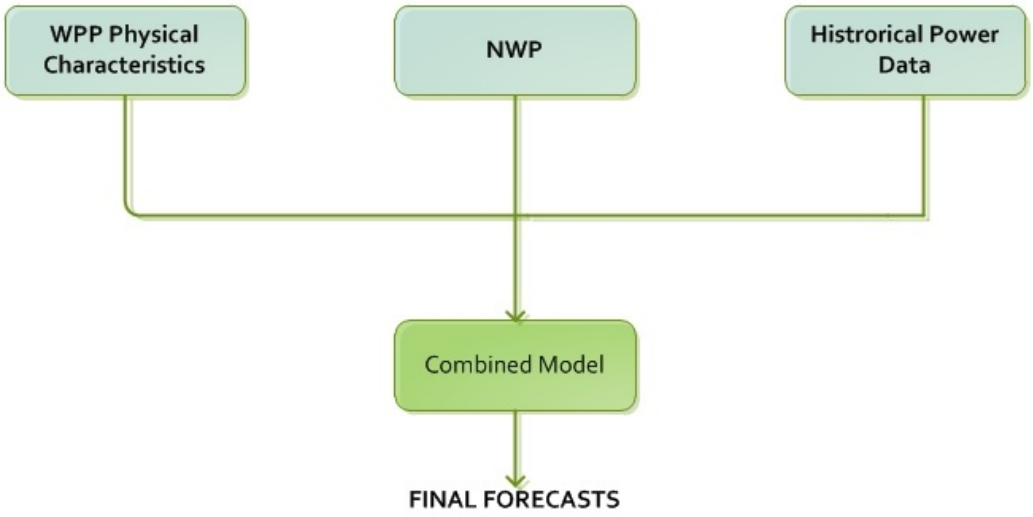

Figure 2.7: General Architecture of the Hybrid models

and obstacles are taken into account and they are combined with the historical power data [17].

Zephyr forecast tool [28] is one of the representative examples for the hybrid models in the literature. The model is extended by use of the Prediktor which is physical model tool and Wind Power Prediction Tool which is a statistical forecast model [17]. The aim of the model is combine both physical and statistical approaches into one model and benefit from the advantages of both approaches. They combined these two approaches in this Zephyr tool in this way: if all historical data is available for the WPP, tool selects the forecasts of the statistical model and if there is no enough data available for the WPP, then the outputs of the physical model are selected by the hybrid forecast tool [28].

CHAPTER 3

BACKGROUND

This chapter presents the background information about the methodologies used in the implementation of the proposed new short term wind power forecast model. The data structure of the most important input of the model, namely NWP, is also described in detail in this chapter.

The main aim of the proposed model is classifying the weather events according to the most important weather forecast parameters. For clustering the weather events, K-means algorithm is used and the details of this algorithm is presented in K-Means sub-section.

One of the most challenging operations in K-means algorithm is deciding the K number. In this proposed model, this number is dynamically determined for each power plant independently and the detail of this operation is given in Dynamic Clustering sub-section.

Weather forecast data has many parameters and before applying the clustering operation among all parameters most crucial ones must be determined. The selection of the most important parameters is performed with PCA and details of it are presented in PCA sub-section.

Finally, in NWP sub-section the details of the weather forecast data are given with their definitions. Within the scope of the project, these sources are taken from three different sources and their properties are also presented in this sub-section

3.1 K-means

Clustering or cluster analysis is an unsupervised approach and it is one of the main methodologies in data analysis applications. K-means clustering algorithm is proposed by MacQueen in 1967 and it is the most common method used in partitioning the data [29]. The aim of the method is partitioning the data in k group and assigning each data to a proper cluster set according to similarity and distance to centroid points.

K-means is one of the oldest algorithms in the computation world and it is generated from the signal processing. The algorithm is easy to implement and general application areas of the algorithm can be listed as:

- Image Processing
- Genetic Algorithm
- Social Networking
- Prediction
- Recommender Systems

Pseudo code of a typical K-means algorithm is presented in Figure 3.1. The input of the algorithm is the data set which has N elements and the output is centroid points of the k cluster set. Firstly, k cluster centroid points are initially selected. In most of the applications, these initial points are selected from the data set directly selecting a particular element. However, in some other applications, these initial points are assigned randomly and selection of the initial centroid points is directly affecting the time complexity of the algorithm. Then, each data is assigned to a cluster according to minimum distance between the cluster centroid point and data itself. After all data is assigned to a cluster, the centroid points are recalculated as the average value of the data points in a particular cluster set. This operation is applied data set continuously until the cluster centroid points do not change. In some other applications, if the data set is too

input : $X = \{x_1, \ldots, x_n\} \in R$

output: $C = \{c_1, \ldots., c_K\} \in R$

Initialize the cluster centroid points

while termination criterion is not met **do**

    for $(i = 1;\ i \leq N;\ i = i + 1)$ **do**

        Assign x_i to the nearest centroid point;

    end

    Recalculate the cluster centers;

    for $(k = 1;\ k \leq K;\ k = k + 1)$ **do**

        Find the set of data points S_k that nearest to centroid c_k

        Calculate c_k as the means of the points in S_k

        $c_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} x_i$;

    end

end

Figure 3.1: Pseudo Code of the K-Means algorithm

big, maximum execution number of this loop is determined by the user and it is terminated directly although centroid points continue to change.

K-means clustering is a NP-Hard problem even for the k=2 and proof is presented in [30] by Dusgupta. The time complexity function of the algorithm can be expressed as follows;

$$O(n.k.d) \tag{3.1}$$

Where n is the number of element in the data set, k is the number of cluster that data will be portioned and d is number of iterations before termination of the process.

One of the weak points of the K-means algorithm is the sensitivity of the method

to outliers and noise in the data. If all data sets have too much outlier element, then the partition of the data may be inaccurate with the usage of K-means. In order to eliminate the outlier problem in the classification, apart from K-medoid algorithm is proposed by Kaufman and Rousseeuw which is also similar to K-means algorithm [31].

3.2 Dynamic Clustering

Although K-means algorithm is one of the most popular methods in portioning the data, it has some weaknesses. The most important issue in such partition algorithms is the deciding on the appropriate number of class and selecting the k correctly. All data sets that K-means applied for portioning have different characteristics. So an appropriate k value for an image processing data cannot be suitable for a recommender system data. Because of that, deciding on the number of k before partitioning data, in other words dynamic clustering the data set, is become one of the popular research are in data mining applications.

In [32], Redmond and Heneghan propose a method based on kd-trees in order to initialize the number of k in k-means algorithm. They use kd-trees to perform the density estimation of data at different points on the data. They test their method for 36 different data sets and compare the results with other algorithms.

Hamerley and Elkan proposed a method namely G-means algorithm in their work [33]. Proposed method starts with k=1 and the number of center points are dynamically growing until the Gaussian distribution of the classes fit for all data sets. They compare their methods with the Pelleg and Moore's X-means algorithm which the k value is determined by the value of Bayesian Information Criterion [34].

The main objective of a good clustering algorithm is minimizing the average squared distance of elements in data from their centroid points and maximizing the distances between centroid points. In order to test the quality of portioning Ray and Turi [35] proposed *validity ratio* coefficient whose formula is stated in Equation 3.4. It is calculated from the *intra* and *inter* distance values whose

equations are stated in Equation 3.2 and Equation 3.3. The average squared distance of the elements are calculated and intra cluster distance variable and minimum distance between two distinct centroid points is assigned to inter cluster variable. The quality variable of the k-means, which is *validity ratio*, is calculated from these two values. The aim is the finding the best number k which makes the validity ratio value to minimum. They applied this test for image processing data over synthetic image files.

$$intra = \frac{1}{N} \sum_{i=1}^{K} \sum_{xec_i} \|x - z_i\|^2 \tag{3.2}$$

$$inter = min(\|z_i - z_j\|^2), i = 1 \; to \; k-1, \; j = i+1 \; to \; k \tag{3.3}$$

$$validity \; ratio = \frac{intra}{inter} \tag{3.4}$$

Main aim of the proposed SHWIP model in this thesis is classifying the weather events in a WPP region. So in the training phase of the model, this will be explained in detail in Chapter 5, the NWP sources are portioned by using K-means algorithm. However, when deciding on the optimal number of cluster number for a particular WPP it is realized that this number is changing from WPP to WPP. Also the optimal cluster number is not always same for a WPP for all training period and it is changeable for different training days. So in the model, this optimal cluster number is calculated for each WPP independently for every training period.

While deciding the optimal cluster number for a WPP in that training period, the validity ratio variable described above is taken into account. First of all for a selected enough NWP data the optimal cluster number for a WPP is investigated. Nearly in all WPPs for the NWP data, the NMAE rates were reasonable when the cluster numbers are between 2-7 and the error rates were deteriorating if we portion the data 8 or more number of clusters. Therefore, in the implementation this issue becomes the constraint of the model. For a WPP in a particular training period, the validity ratio's for cluster number 2 to 7 are

calculated independently and among all values which has the minimum ratio is selected as optimal cluster number that the data will be portioned. The time complexity of the typical K-means algorithm was given in Equation 3.1. In the model this algorithm is executed for 6 times in order to calculate the validity ratios belong to 6 different k values form 2 to 7. Since in the time complexity function the dominant factor is the number of element that will be portioned, executing the k-means 6 times does not reduce the performance of the model too much.

3.3 Principal Component Analysis

Principal Component Analysis is one of the most valuable methods in linear algebra applications. The main aim of the PCA methodologies is dimension reduction on the analysis data without loss any information. Generally, the data that worked on has many parameters and there is a need for compressing the data in order to work more efficiently. PCA methodologies are taking part in this data compression operation. The basic steps of a typical PCA can be listed as below [36]:

- Organizing the data set with use of important parameters
- Calculating the mean for all the parameters and calculating the deviations from the mean values
- Finding a covariance matrix from the standardized data set
- Extracting the eigenvectors and eigenvalues of the covariance matrix
- Converting the data to reduced dimension with the use of selected eigenvectors

There are lots of application areas of PCA such as in image processing, pattern matching and prediction applications. In the proposed wind power forecast model in this thesis, PCA is used in order to reduce the dimension of the NWP data. While clustering the weather events in a plant area, K-means algorithm

was used. However, the NWP data has lots of forecast parameters and it is hard to cluster a data set which has too many parameters. Therefore, by using of the PCA analysis the data set is compressed to a one dimensional structure. While doing this operation, the basic steps described above are applied on the NWP data. Firstly, every parameter in the data set used in wind power forecasting, are standardized and normalized and a Covariance Matrix is calculated from this data sets. After finding the Covariance matrix the most significant eigenvector (not all eigenvectors) of that Matrix extracted and by multiplying this matrix with the data set, the compressed values are obtained. As the final step, this one dimension array data is clustered with use of K-means and the weather events in the WPP are is classified dynamically.

3.4 Linear Regression

Regression analysis is one of the common approaches used in the statistics. Linear Regression is mainly used for modelling to two variables by fitting a linear line between the variables. In order to apply linear regression to data set, there must be a correlation between the explanatory and dependent variable [37].

$$y = a + bX \tag{3.5}$$

A linear regression line equation is given in the Equation 3.5. Y is the dependent variable and X is the explanatory variable. The slope of the line is b and interception value is a.

Linear regression analysis methods are used lots of areas in order to find the hidden relationship between the variables. For instance, in a supermarket data set, the relation between the food and beverage can be extracted by applying linear regression in order to arrange the correlated products in the same place in the market. Also it is used in the future prediction applications frequently. For example, the relation between oil usage and price through the years can be determined by applying linear regression the next years consumption costs may

be estimated.

In the proposed model, linear regression is applied on the wind speed and production data. In the training phase of the proposed model, the best regression line between wind speed and real power is constructed by applying linear regression on these variables. The power value is the dependent variable and wind speed is the explanatory variable as stated in the 3.5. The main aim is the finding the best correlation coefficient which is the slope of the Eqaution 3.5 between the wind speed and production data. For this aim, this correlation coefficient is founded according to calculated NMAE rates in the training period. The detail of the usage of the linear regression is given in detail at the training phase of the model stated in the Chapter 5.

3.5 Numerical Weather Prediction

NWPs are the main data source for a wind power forecast system and in all of the models, the reliability of the wind power forecast models are directly proportional to accuracy of the NWP sources. Within the scope of the project the NWPs are taken from three different sources as listed below:

- ECMWF
- GFS
- DMİ

ECMWF was established on 11 November 1973 by 17 European countries, among which European Medium Range Weather Forecast Center of Turkey was also present. As a result of the research and studies carried out, ECMWF Deterministic Model was set forth and it has become operational as of 1 August 1979 [38].

GFS is developed in 2002. The model, which gives outputs between the resolutions of 35 and 70 km, is run four times a day and generates 196 hour-forecasts. All products are available online free of charge [39].

Table 3.1: A sample GFS forecast data

hour	u0	v0	u1	v1	u5	v5	p	t
1	4536	7090	5987	9479	-3189	16399	879476171	2856154
2	4053	8493	5733	11936	-7097	23507	879294882	2857717
3	5129	8736	6992	12165	-6019	24689	879173554	2858186
4	5289	6680	7182	9384	-5617	21934	879064453	2848498
5	4721	7848	6404	10801	-6697	23189	879060898	2840998
6	3916	8335	5167	11244	-8573	26573	879007187	2842092
7	1965	7778	2178	10285	-9118	29655	878918515	2839904
8	-3270	6171	-4663	7302	1618	33537	878485820	2840373
9	-10904	6090	-12884	7289	-7158	23301	877749296	2844123
10	-9877	10620	-11025	12805	-6138	24373	876967109	2845217
11	-5671	18018	-6465	21531	-943	31946	876418945	2843654
12	-3604	22566	-4267	26393	1586	36075	876048554	2846936
13	-883	20238	-349	23743	7140	32819	875875898	2846467
14	5977	12834	8288	15859	19454	22634	875920390	2833498
15	3703	5176	4925	7309	16334	11328	876182382	2831311

The data obtained from the ECMWF and GFS are given as initial input to WRF model the weather forecasts are obtained from the output of the execution of the WRF model. WRF model is a new generation quantitative weather forecasting model, which can respond to the atmospheric researches as well as the operation predictions [40].

The initial data taken from the DMİ is served as the initial data of the ALADIN model and the weather forecasts are obtained from the execution of this model. ALADIN is a bi-periodical and hydrostatic weather forecast model, which operates with a 4 km horizontal resolution in a field restricted with Cartesian grids. It is first started in November 1990 [41].

A sample fifteen hours NWP data obtained from a GFS for a particular grid point is given in Table 3.1. Every day 48 hours weather forecast data are obtained for all grid points in Turkey and the power forecasts are made by using these weather forecasts data. The data structure stated in Table 3.1 is the same for

ECMWF and DMİ. However, the distance between two grid points in 4 km in DMİ forecasts on the other hand that distance is 6 km in the other two sources.

The parameters and their features can be listed as below:

- u is wind speed x component
- v is wind speed y component
- p is pressure
- t is temperature

The u and v values are the most important parameters and wind speed/direction values are directly calculated from these two parameters. These are taken for six different levels from u0/v0 to u5/v5 and from the ground and distance between levels is 10 m. The u5 and v5 values are more near to turbine level and they are used in the forecasting. The pressure value is in the Pascal unit and the temperature value is in the Kelvin unit which is multiplied with 10000. The details of the using of these parameters in forecasting are described in the Chapter 5.

CHAPTER 4

GENERAL ARCHITECTURE OF THE RİTM SYSTEM

This chapter summarizes the architecture of the wind power monitoring and forecast system of Turkey which is designed in the scope of the Wind Power Monitoring and Forecast Center for Turkey (RİTM) project. The main aim of the project is the large scale integration of the wind power energy to Turkish electricity grid properly. Monitoring issue is most important for the TSOs who manage the energy flow in the country. Forecasting is more crucial for the WPP owners from the economical perspective. In the country, similar to other countries, there is an energy market which WPP owners declare day-ahead forecasts to that institute and their profit directly proportional to accuracy of that forecasts. The main duty of this center is meeting these two expectations continuously.

In order to construct a reliable wind power forecast system various data sources are used and the data coming from these sources must be acquired and stored. In the "Data Acquisition and Data Storage" sub-section the details of the data used in the center is presented.

In the "Wind Power Monitoring Center" sub-section, details about the center, features related to capacity of NWPs the system can handle and properties of server computers in the center are presented.

Finally, in order to monitor all WPPs and follow their forecasts in the center instantly, a map based monitoring and forecast software is designed and it's features are given in "Map Based Monitoring and Forecast Software" sub-section.

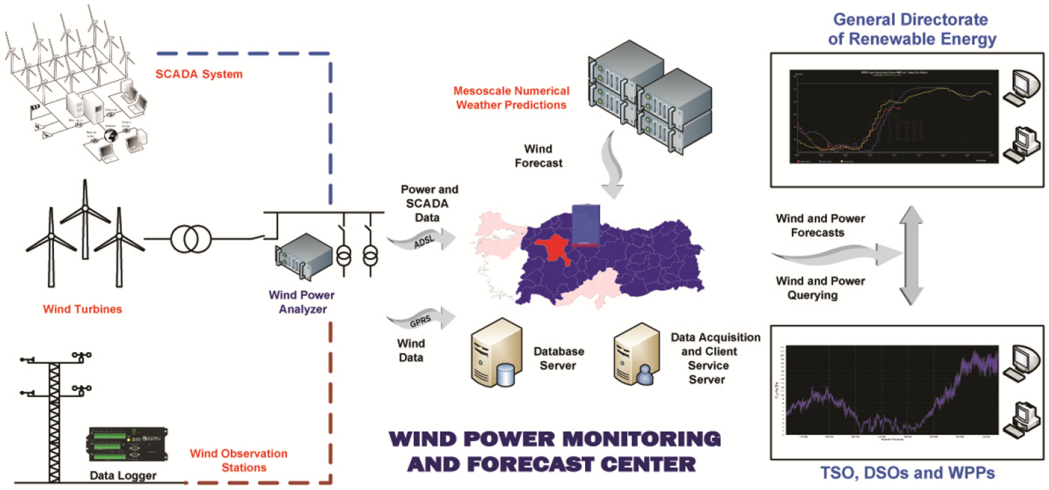

Figure 4.1: The architecture of the RİTM System

4.1 Data Acquisition and Data Storage

In the RİTM center, data needed for a reliable forecast system is collected from various data sources. The general architecture of the system is presented in Figure 4.1 [3]. All of the data coming from different sources are collected and stored in the servers active on RİTM center. By the use of these data sources, various forecast models are running in order to obtain the very short term (up to 6 hours), short term (up to 48 hours) and regional forecasts [1, 2, 4].

The main data sources used in this system can be listed as below and their features are presented in the following sections:

- Wind Power Analyzers
- Wind Masts
- SCADA Systems
- Medium Scale Numerical Weather Forecasts

Figure 4.2: An example Wind Power Analyzer used in the project

4.1.1 Wind Power Analyzers

Wind Power Analyzers are one of the two most important components with NWP sources in this system. In order construct a reliable wind power forecast model, real production data of the WPP must be known especially for constructing a statistical model. An example picture of this analyzer is shown in Figure 4.2. These analyzers are designed by TÜBİTAK for National Power Quality Project in order to monitor the power quality in the country [42].

These analyzers make analyses from two different measurement points of WPP and these points are called as feeders. Analyzers carry out the calculations for electric quantities such as voltage, current, flicker, unbalance and harmonics and these values are sent to center in every 3 second. However, among all power quality parameters, the most crucial one is the power value. The total production value of all turbines are measured from feeders and also power values are sent to center in every 3 second. Especially, these 3 second resolution is much more important in very short term wind power forecasting and it improves the models' performances [4]. ARMAX model is used as very short term forecast method in the center and power data is given as exogenous factor of the model.

As represented in Figure 4.1, the analyzers measure total power data of all turbines and these values are sent to center through ADSL link over network. The power values obtained by each wind power analyzers are stored in the center with the appropriate table served for a particular WPP.

The wind power analyzers have capable of store the measurement data in itself if a problem occurs during the sending values over socket. Devices store the all power quality measurement data in tar files and when the communication problem is solved they send the files to the center. After these files reach the center, they are parsed with the applications run on the servers and data values are written to appropriate tables. Therefore, as long as device does not have a hardware problem, the power data is never lost for a particular WPP.

4.1.2 Wind Masts

Wind Masts are used for similar approaches as Wind Power Analyzers but rather than measuring instant power data of the WPP, they are used for measuring the instant meteorological parameters in the WPP region. An example picture of the typical Wind Mast used in the project is shown in Figure 4.3. With the help of the sensors on it, these devices are measuring the instant wind speed, wind direction, temperature, pressure and humidity in the WPP region.

The communication between Wind Masts and center are provided over GPRS. The Data Logger in the devices are collecting the data from the sensors and producing text files including ten minutes average measurements values. These files are sent to application server in the center for each Wind masts. A data transfer application periodically checks these files weather new measurement data has come or not and these files are parsed and the values are written to the appropriate tables for each WPP. Among all 20 WPPs in the system only 7 of them have Wind Masts and instant meteorological data values are stored for only these 7 WPPs [3]. Later, these values are dynamically monitored by the usage of the client software such as Map Based Monitoring and Forecast Software.

Figure 4.3: An example picture from Wind Masts used in the project

The measurement data taken from the Wind Masts are not directly used in wind power algorithms since they do not serve future meteorological predictions as NWP sources. Instead of giving meteorological future predictions in the area, these masts measure the instant situation and give an idea about the accuracy of the meteorological estimations. The values are used for finding a correlation between NWPs and actual meteorological values but the obtained results are not always compatible. The reason behind this result may be the installation place of the Wind Masts. The most suitable places for wind in the WPP area is installed with wind turbine so Wind Mast places generally are not representative for all WPP area and measurement values are not too accurate.

In other examples in the literature, generally the values obtained from Wind Masts are used in the construction of the power curves of the WPP. In RİTM system, the similar approach is used while using the meteorological data obtained from the Wind Masts. Some of the WPP's power curves are constructed from the Wind Masts data and they are compared with the current power curves in order improve the performance of the forecasts.

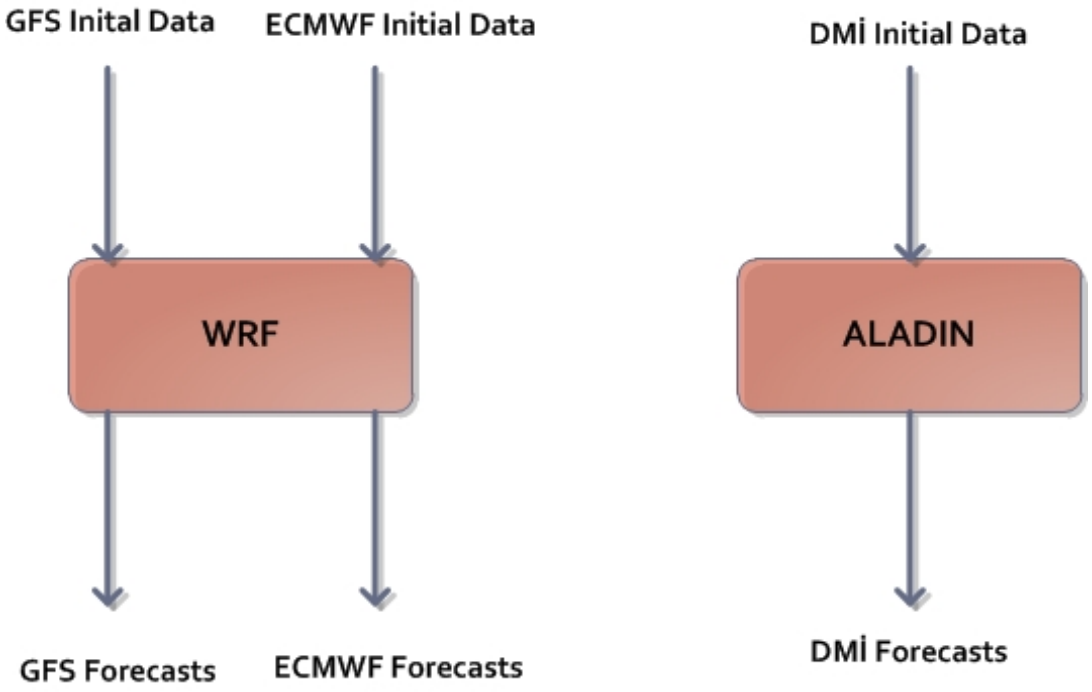

Figure 4.4: Medium Scale Weather Forecasts in RİTM project

4.1.3 Scada Systems

SCADA Systems of the WPPs are used to obtain the availability/status, wind speed and wind direction values of the each individual turbine independently. This information is stored in files at the SCADA Systems and they are sent to application server periodically. However due to WPPs high security concerns, they do not prefer to share the data in their SCADA systems. In the project only two of the WPPs are allowed to fetch data from their SCADA systems. Since the data is not for all WPPs, the data obtained from the SCADA systems are not used in the forecast models.

4.1.4 Medium Scale Numerical Weather Forecasts

Medium Scale Numerical Weather Forecasts are the most important input of the forecast models and the details of the data structure of the NWPs are described in detail in Section 3.4. In the center, the Numerical Weather Forecasts are

Figure 4.5: A panoromic view from the Center

downloaded from GFS, ECMWF and DMİ and the values are stored in the database servers. Then GFS and ECMWF are served as initial input of the WRF model and DMİ is served as initial input of the ALADIN forecast model as shown in Figure 4.4. With the help of applications on the application server, the models are executed independently and three different weather forecasts are obtained. These forecasts cover all Turkey area and grid resolution is 6 km in GFS and ECMWF forecasts and 4 km in DMİ forecasts.

4.2 Wind Power Monitoring Center

Wind Power Monitoring Center is located at the YEGM building and all of the servers are set up in this center. The servers are working 7/24 for getting and processing wind data dynamically. The center is divided into two parts as Study and Monitoring room and Server room. In the study room, a DLP video wall system is set up and from this screen the situations in all WPPs in the system can be monitored by using client software. A panoramic picture from the center study room is shown in Figure 4.5.

Table 4.1: WPPs monitored in the RITM project

WPP	Installed Capacity (MW)	Region
WPP1	15 MW	Region of Marmara
WPP2	90 MW	Region of Aegean
WPP3	60 MW	Region of Marmara
WPP4	35 MW	Region of Marmara
WPP5	10.2 MW	Region of Marmara
WPP6	14.9 MW	Region of Marmara
WPP7	39.2 MW	Region of Marmara
WPP8	135 MW	Region of Mediterranean
WPP9	140.1 MW	Region of Aegean
WPP10	36 MW	Region of Mediterranean
WPP11	30 MW	Region of Aegean
WPP12	35 MW	Region of Aegean
WPP13	57.5 MW	Region of Mediterranean
WPP14	12 MW	Region of Aegean
WPP15	39 MW	Region of Marmara
WPP16	60 MW	Region of Marmara
WPP17	78.2 MW	Region of Aegean
WPP18	40 MW	Region of Black Sea
WPP19	39 MW	Region of Black Sea
WPP20	72 MW	Region of Central Anatolia

The list of the monitored WPPs in the RİTM center as of December 2013 is given in Table 4.1. The plants from WPP15 to WPP20 are started to be monitored recently and the other plants are monitored since 2011. In the center, for the plants from five different regions of Turkey, the power forecasts are produced every day.

Apart from the study room, a server room is located in the center. In the server room five servers are working continuously. The servers and their works can be listed as below:

- 2 Database Servers with one is served as a backup
- 2 Application Servers for recording the data obtained from Wind Power Analyzers, NWPs and Wind Masts
- 1 Web Server for providing services for RİTM web page and Web Based Monitoring and Forecast Software

Database servers, which is the main storage unit of the monitoring center, have 32 GB memory and 10 TB disc area. On the servers, Fedora operating system and PostgreSQL is database management system are used [43]. The application servers with 8 GB memory and 10 TB disc receive data from the wind power analyzers, NWPs and Wind Masts and store the data and feed the monitoring software with this data and the main processes are carried out on them. Currently, the servers handle the operations for 20 WPPs. But it has the capability of serving 200 WPPs at the same time and after the participation of all WPPs in the country to the project it will give service all WPPs at the same time. Apart from the server room in the center another four servers (2 for database, 2 for application) exist in TÜBİTAK building for backup.

4.3 Map Based Monitoring and Forecast Software

Within the scope of the RİTM project, four different client software are designed for different purposes and they can be listed as below [3]:

- Dynamic Monitoring and Querying Software
- Map Based Monitoring and Forecast Software
- Power Quality Monitoring Software
- Web Based Monitoring and Forecast Software

Dynamic Monitoring and Querying Software is used for monitoring the instant wind power produced by the WPP and it is implemented in Java. The values

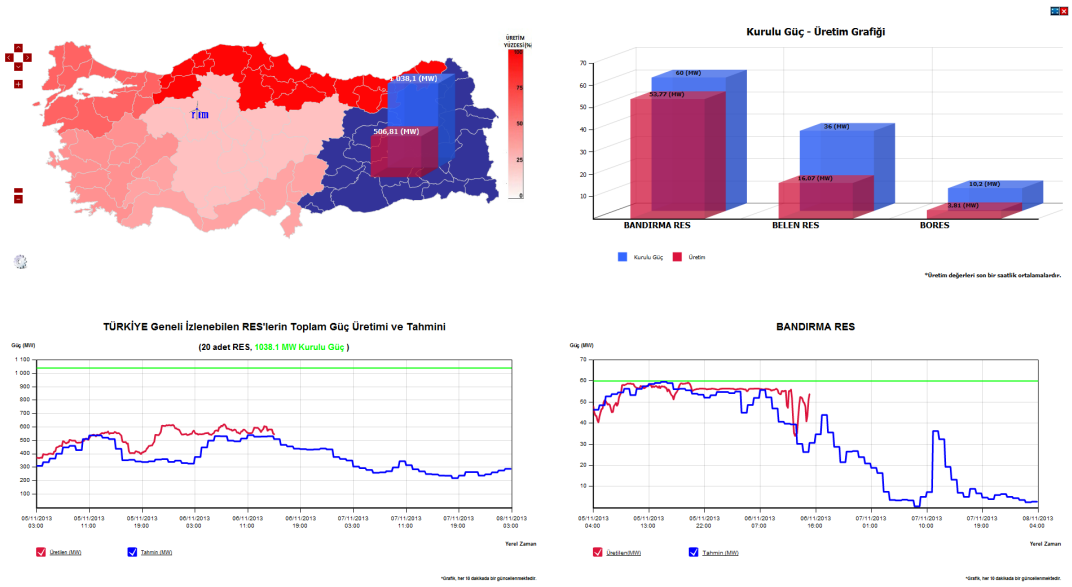

Figure 4.6: Map Based Monitoring and Forecast Software

are updated in every 3 seconds. In addition, this software has capabilities for querying the past days' forecasts and power values, and exporting the query results to save in files. Power Quality Monitoring Software is used for investigating power quality parameters in the WPP such as voltage, current, flicker and harmonics. It is also used for querying the power quality events such as interruption or voltage sag in the WPP. This software is implemented by using C# in .Net environment.

Web Based Monitoring and Forecast Software has similar properties with Dynamic Monitoring and Querying Software but it does not involve report generation. The web pages are Java Server Pages and the software run at Tomcat server [3].

Among these software, the Map Based Monitoring and Forecast Software is working in the RİTM center 7 days 24 hours continously. A view from this software is presented in Figure 4.6. The software is implemented in Java and initially designed for the DLP video walls. It can be also used in the regular personal computers.

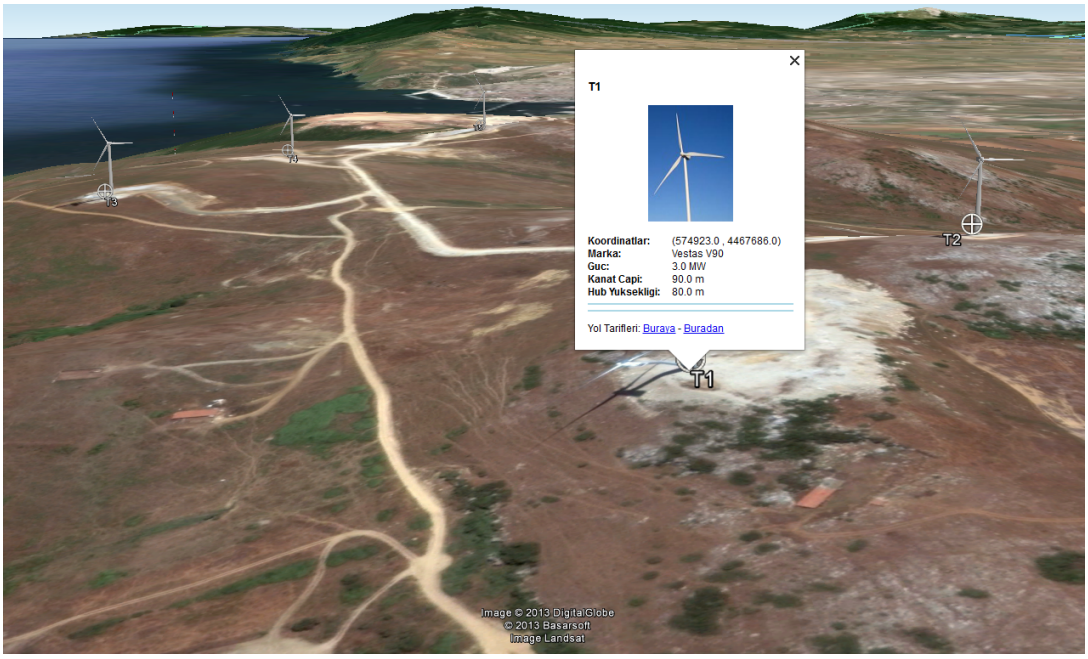

Figure 4.7: Google Earth Integration of the Software

The main screen of the application has four different components. At the upper left panel a country map is located and it has different functionalities. The map has five different regions where the WPPs are located in. These regions' colors are changing according to production ratio of the WPPs in that region in the last ten minutes. If the production percentage is high then the region becomes more red, otherwise it becomes white color. By zooming to a region, the WPPs in that region are shown in the map with their connection status.

If there is a communication problem between the Wind Power Analyzer and the center, the icon representing the WPP becomes red and if there is no problem it becomes green. The map functions are also integrated with Google Earth software [44] and the wind turbines in a WPP region can be viewed in three dimensional view with their technical properties as shown in Figure 4.7 . Also other details about the WPP such as the owner, set up date, total installed capacity and details of the wind mast, if the WPP has one, are accessible.

The second graphic in the upper right part of the main screen shown in Figure 4.6 presents the last one hour total production and installed capacities of the WPPs

in one bar graphic. The WPPs are changing simultaneously by a user specified time period (such as 30 seconds). In the same manner, a bar graphic is located as well in the Turkey map, which represents the current situation throughout the country.

The lower line graphics shown in Figure 4.6 are in similar format. The graphic in the right part shows the production/forecast of a particular WPP in a time series chart for 72 hours period. First 24 hours represent the situation the day before and 48 hours show the day ahead forecasts from the start point of the current day. The real production values are updated every ten minutes with the upcoming new production data. This graphic is also simultaneously changing together with the bar graphics above.

The line graphic for Turkey has the same functionality as the graphic of a particular WPP described above. At the top of this graphic, information is about the total installed capacity, which is being monitored dynamically and the total number of WPPs in the system is given.

CHAPTER 5

PROPOSED TECHNIQUE

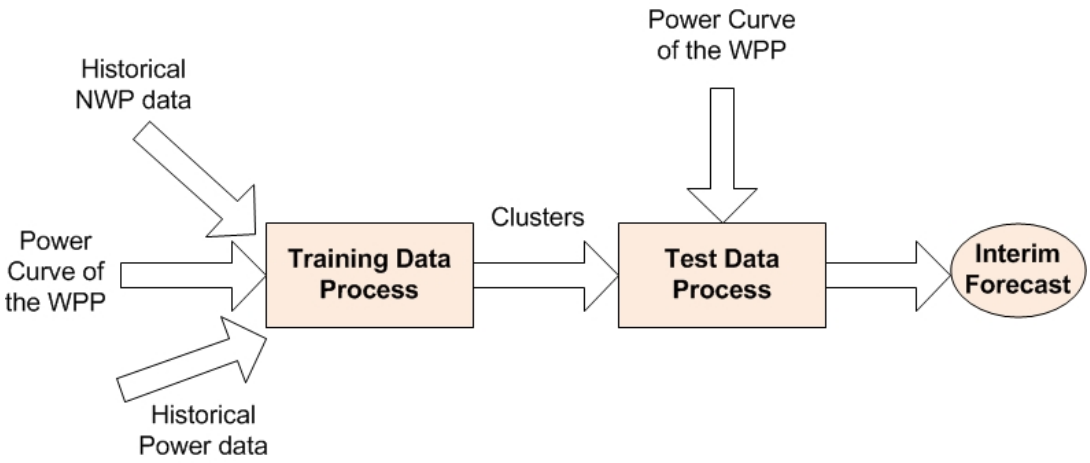

Figure 5.1: Training and Test Process of the Model

5.1 Overview of Statistical Hybrid Wind Power Forecast (SHWIP)

The proposed short term wind power forecast method is based on combining dynamic clustering with linear regression. In the clustering operation, the weather events in the WPP area is classified in a dynamic way and the best correlation linear line between speed and power is founded by linear regression analysis.

The general view of the training and test phase of the model is shown in Figure 5.1 [1]. The main inputs of the training phase is historical power data, historical NWP data and power curve of the WPP and the output is the clusters which are the main input of the test phase.

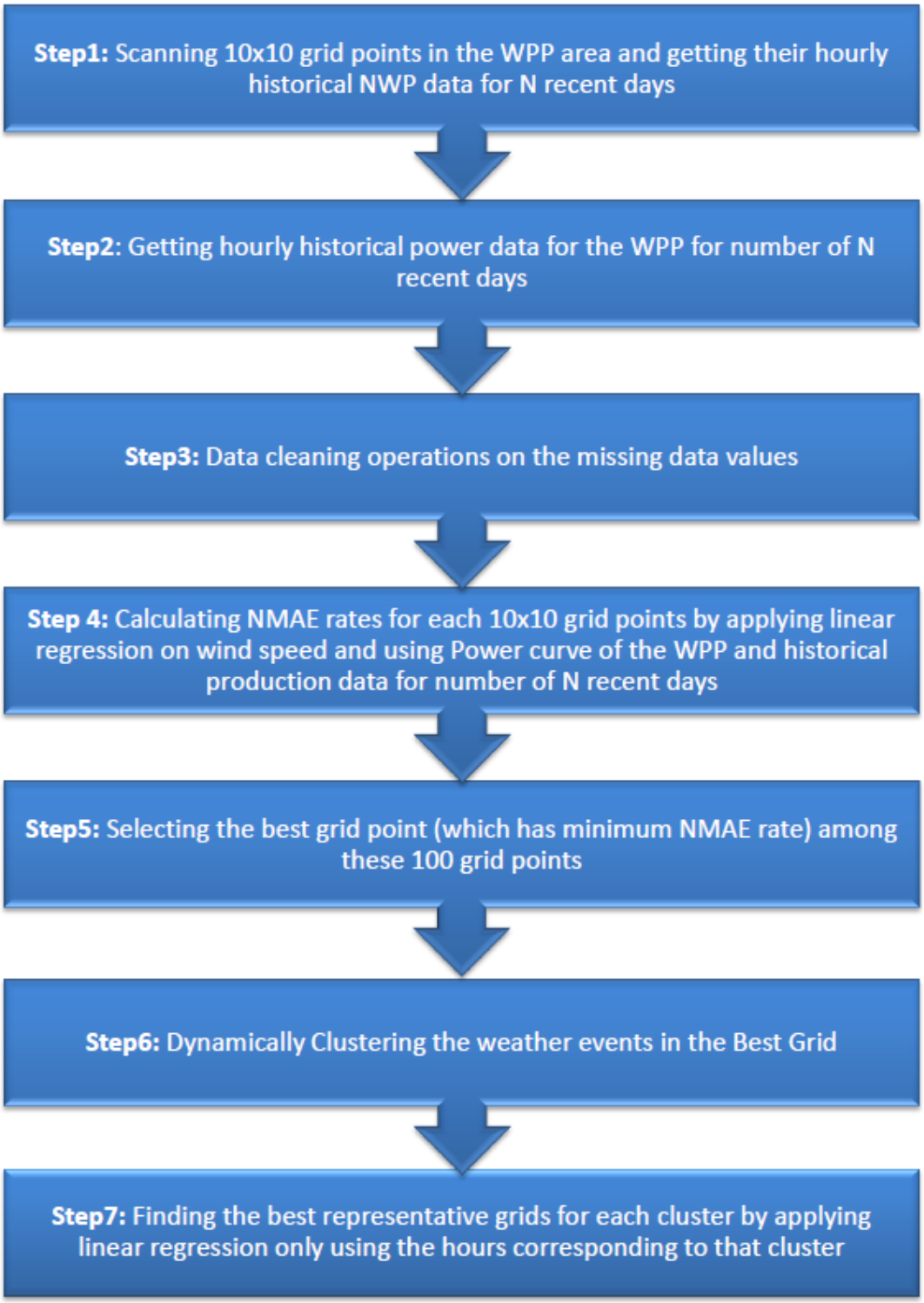

Figure 5.2: Steps of Training Phase

In the test phase, by the use of the clusters with the power curve of the WPP 48 hour interim forecasts for a particular day are obtained. Within the scope of the RİTM project, NWPs are taken from three different sources as stated in Section 3.5 and these interim forecasts are obtained for these sources independently. As the final step, the interim forecasts are combined in to a final forecast with the help of a combination phase. All of these phases are conducted for each WPP independently everyday in order to obtain the forecast results. Then the forecast results for Turkey are obtained by the sum of the WPPs' forecasts in the system. The details of all these phases are stated in the following sub-sections.

5.2 Training Phase in SHWIP

Training Phase is the most significant process in the model where data mining actives are done. The aim of the training phase is constructing the statistical model by using the historical NWP and power data.

This phase can be divided into three sub-section as following:

- Finding the Representative Grids
- Dimension Reduction
- Finding the Optimal Clusters

Main steps of the whole process are stated in Figure 5.2 and all of the parts are described in detail in the following sub-sections.

5.2.1 Finding the Representative Grid

In the first step, for each WPP, 10x10 grid points in the WPP area are scanned where WPP reference coordinate constitutes the center of the area as shown in Figure 5.3. The distance between the two grid points is 4 km for DMİ data and 6 km for GFS and ECMWF data. Generally, the WPPs' areas are not larger than 10 km^2 so scanning a 10x10 grid points (1600 km^2 area for DMİ, 3600 km^2

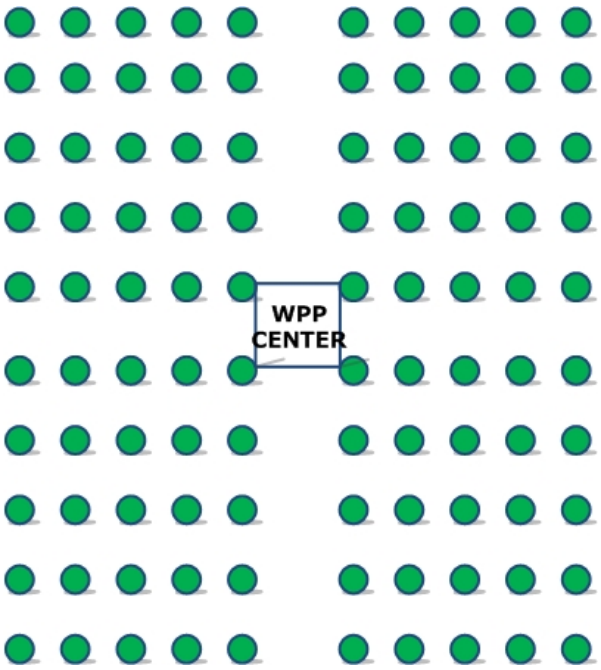

Figure 5.3: Scanned Grid Points in the WPP Area

area for GFS and ECMWF) is enough for the model. For all these grid points, their hourly historical NWP data for N number of recent day is obtained. In the proposed model, this number of day period is chosen as 90 days. Because 90 days include a time period for a season and it is optimal for model finding clusters from the nearest one season period.

In the second step, similar to the first step, the power production values for the WPP's are taken from the database. The values are in 10 minutes resolution. Since the NWP values are obtained hourly, the power data values are converted to hourly data by applying interpolation to 10 min interval values in order to be compatible with the NWP data. These values are also constructed for N recent days. The power values are used in order to calculate the Normalized Mean Absolute Error (NMAE), as given in Equation 5.3 rates of the grid points and determining the performance of the grid's during the training period.

In the NWP data, there may be missing hour values where the initial meteorology data does not exist for the model. In addition, for some periods, if there is a connection problem between the power analyzer in the WPP and center, there

may be missing power values in the training region. Therefore, as given Step 3 of Figure 5.2, a data cleaning operation has been conducted for the NWP and power data in order to eliminate missing data and to work with only complete instances. Missing hours' data is flagged with a dummy value and these hours are not used while constructing the model.

$$s = \sqrt{u^2 + v^2} \tag{5.1}$$

$$d = (\arctan(u/v) \times 180)/\Pi + 180 \tag{5.2}$$

$$NMAE = \frac{\frac{\sum_{i=1}^{N} |y_i - x_i|}{N}}{C} \times 100 \tag{5.3}$$

In Step 4, for each grid points in the WPP ares, their NMAE rates are calculated by applying linear regression on wind speed (s). As stated in Table 3.1, NWP data in the models has parameters such as u (wind speed x component), v (wind speed y component), p (pressure), t (temperature). The wind speed (s) and wind direction (d) values are calculated as given in the Equation 5.1 and 5.2.

As stated before, the power curve of a WPP is a matrix in the dimension of 360x251 that associates wind direction and wind speed to an estimated wind power. By applying linear regression on wind speed (s) values with initial correlation coefficient as 0.5, the NMAE rates are calculated up to coefficient becomes 2.0 with 0.1 increment interval. For a grid, among all NMAE rates in that grid, the minimum rate and its coefficient are determined and it is set as the NMAE of the grid . In the Equation 5.3, x_i is the real power, y_i is the power forecast at i^{th} hour, C is the installed capacity of the WPP and N is the total number of hours processed in the training phase.

After calculating the minimum NMAE rates for each grid point, among all points which has the minimum error rate is selected as the initial representative grid point of the WPP in Step 5. Generally, this point becomes one of the grid point which close to WPP center. However, for some days where the weather

events suddenly change, this representative grid point can be far away from the WPP center point. The NWP's of this point is used for the dynamic clustering operation of the weather data and this grid point is also used in the "Test phase" of the model. Therefore, the latitude and longitude information of this grid point is saved in order to be used later in the "Test phase".

5.2.2 Dimension Reduction

The crucial point in the Figure 5.2 is the sixth step where the dynamic clustering operation is applied. After selecting the best grid points in Step5, the NWP of this point is grouped in order to determine different weather situations in the WPP region.

$$M = \begin{pmatrix} u_1 & v_1 & p_1 \\ u_2 & v_2 & p_2 \\ u_3 & v_3 & p_3 \\ . & . & . \\ . & . & . \\ u_N & v_N & p_N \end{pmatrix} \tag{5.4}$$

$$C = M^{\top} \times M \tag{5.5}$$

$$X = M \times E \tag{5.6}$$

In order to cluster the weather events, the most important weather parameters (u, v and p) are selected from the NWP data. Since the temperature values do not suddenly change as shown in the Table 3.1, it is not taken into consideration while grouping the weather events.

As stated before, K-means clustering algorithm is used while clustering data set. However, since data set has three different parameters (u,v,p), firstly PCA is applied to data set in order to compress the data without loss of any information.

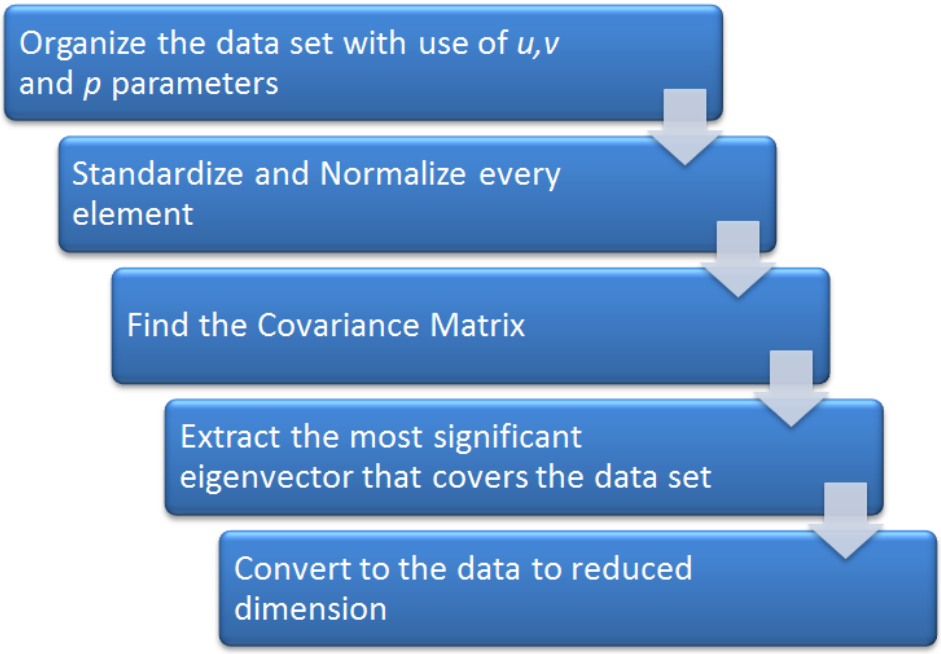

Figure 5.4: PCA in the proposed model

The basic steps of a typical PCA analysis was stated in Section 3.3 and for the NWP data the process shown in Figure 5.4 is applied. First of all data is formed with most significant parameters u,v and p values. Then, for all parameters, mean value and deviation from the mean are calculated and every element is standardized and normalized by the use of these values. At the end of this process, the M matrix stated in Equation 5.4 is formed. This matrix in the dimension of Nx3 where N is the total number of hour processed in the training region (evaluation results are obtained for N=90x24). After that the Covariance Matrix (C) is calculated from the M matrix and it's transpose as given Equation 5.5 and it's in the dimension of 3x3. The most significant eigenvector (E) of this Covariance Matrix is extracted which is in the dimension of 3x1 in order to form the compress data. Finally, from matrix M and the most significant eigenvector matrix (E), the final compressed data matrix X is calculated as stated in Equation 5.6. As a result, at the end of this process, the effects of the principal components u, v and p for each hour are obtained independently without loss any information.

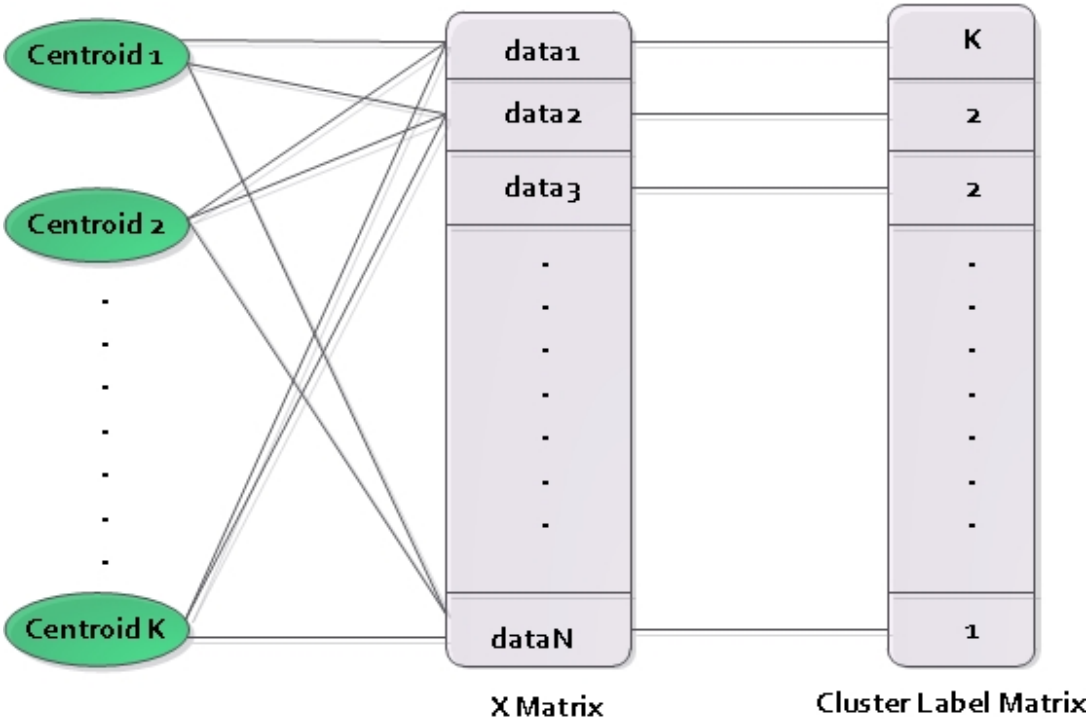

Figure 5.5: Assigning each hour to cluster set

5.2.3 Finding the Optimal Clusters

K-means clustering algorithm is applied on the X matrix to determine the clusters. Initially, the first centroid point is initialized to the minimum element in the data set and the last centroid point is assigned to the highest element in the set. The other cluster centroids are also assigned similar proportional to the first and last cluster centroids. Therefore, at the and of the process, cluster centroid points are ordered from the smallest to the largest. While deciding on the number of clusters (K) dynamic clustering is applied on the data set as stated in Section 3.2. K-means is applied on the data set for k=2 to k=7 and the *validity ratios* [1] described Section 3.2 are calculated for each clustering. Among all *validity ratio* values, the maximum value is selected as the k number. Finally, K-means is applied with the selected k value. This operation is conducted for each WPP independently.

[1] validity ratio=intra/inter where *intra* is the total distances of the each point to its cluster centroid point and *inter* is the minimum distance between cluster centers

The final step of the training period given in the Figure 5.2 involves finding the best representative grid points for each cluster set. To this aim, firstly a "Cluster Label Matrix" is constructed as shown in Figure 5.5. Every element in the X matrix is assigned to a cluster set according the Euclidean distance of data points to cluster centroid points. It is determined according to the minimum Euclidian distance.

The best representative grid point for each cluster set is determined by after the construction of the Cluster Label Matrix. For each cluster set, the steps from 1 to 5 described in Figure 5.2 are applied and best representative grid points are determined. These grid points are used in the "Test phase" of the model.

As the result of the Training phase, the outputs can be listed as below:

- Initial best grid point latitude/longitude information
- The most significant eigenvector E matrix founded in the PCA
- Cluster centroid values
- Each cluster's best representative grid point latitude/longitude information
- Each cluster's best correlation coefficients founded by linear regression for wind speed

These values are kept in the database in order to be used in the Test phase of the model.

5.3 Test Phase in SHWIP

In this phase, WPP power forecasts are obtained for 48 hour by using the outputs of the training phase. In the server, training phase is executed every night and the model is reconstructed with the new data. Test phase is executed every morning when NWP data for the sources (DMİ, GFS, ECMWF) is ready in the database. The steps of the Test phase is shown in Figure 5.6.

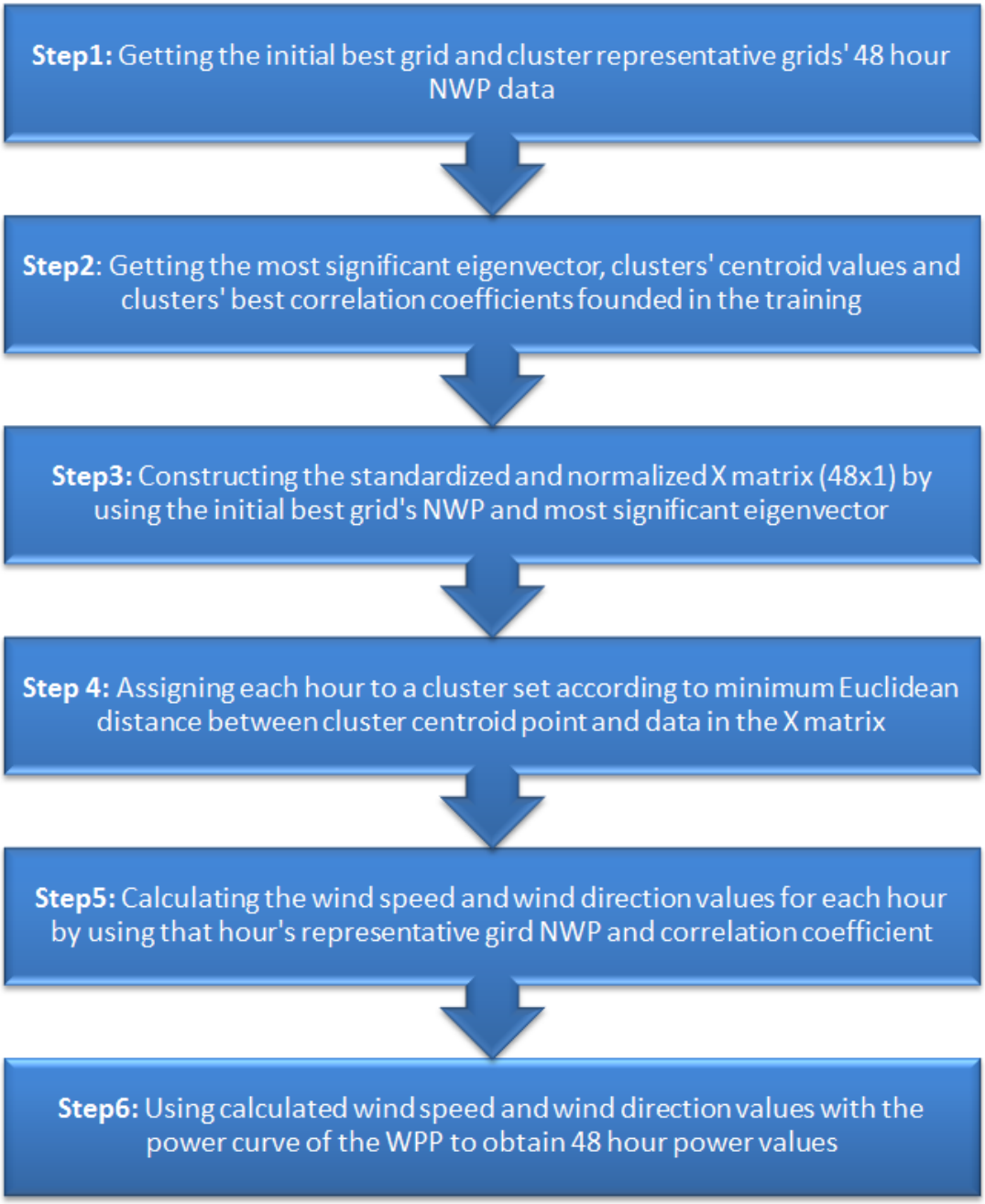

Figure 5.6: Steps of Test Phase

In the first step, the 48 hour NWP data of the initial best grid point and cluster representative grid points are obtained from the database. These values are taken according to latitude and longitude information of the grids saved in the Training phase.

In Step2, the most significant eigenvector, clusters' centroid values and clusters' best correlation coefficients are found in the training phase by applying linear regression to wind speed taken from the database. The most significant eigenvector is used to compress the initial best grid's 48 hour NWP data to 48x1 dimension and cluster's centroid values are used in order to determine the each hour's cluster label. Correlation coefficients are used in order to calculate wind speed estimation. The calculated wind speed and wind direction values are used with power curve of the WPP to obtain power forecasts.

In Step3, similar to the training phase, a compressed X data matrix (48x1) is constructed from the initial best grid's NWP values and the most significant eigenvector. As in Equation 5.4, 48x3 M matrix is constructed from u,v and p values of the initial best grid the point and then the compressed X matrix (48x1) is obtained from this M matrix and the most significant eigenvector as in Equation 5.6. This matrix is used for clustering the each hour in the period.

As in the process shown in Figure 5.5, each hour is assigned to a cluster set in Step4. According to Euclidean minimum distance between the data in the X matrix and cluster centroid values, each hour is assigned to a cluster set. At the end of this step, a cluster label matrix in the dimension of 48x1 is determined that contains the information of that each hour similar to which weather event.

In Step5, each hour wind speed and wind direction estimation are calculated by using the that hour's representive grid point NWP data. Firstly, from u an v values, wind speed (s) and wind direction (d) values are calculated as given in the Equation 5.1 and 5.2. Then the wind speed values are multiplied with the correlation coefficients founded in the training phase. This operation is conducted for every 48 hour independently and wind speed and wind direction predictions are calculated.

In the final step, calculated wind speed and wind direction values are passed to the power curve of the WPP and 48 hour power forecasts of WPP are constructed.

Training and Test phases are executed for each WPP independently. At the end

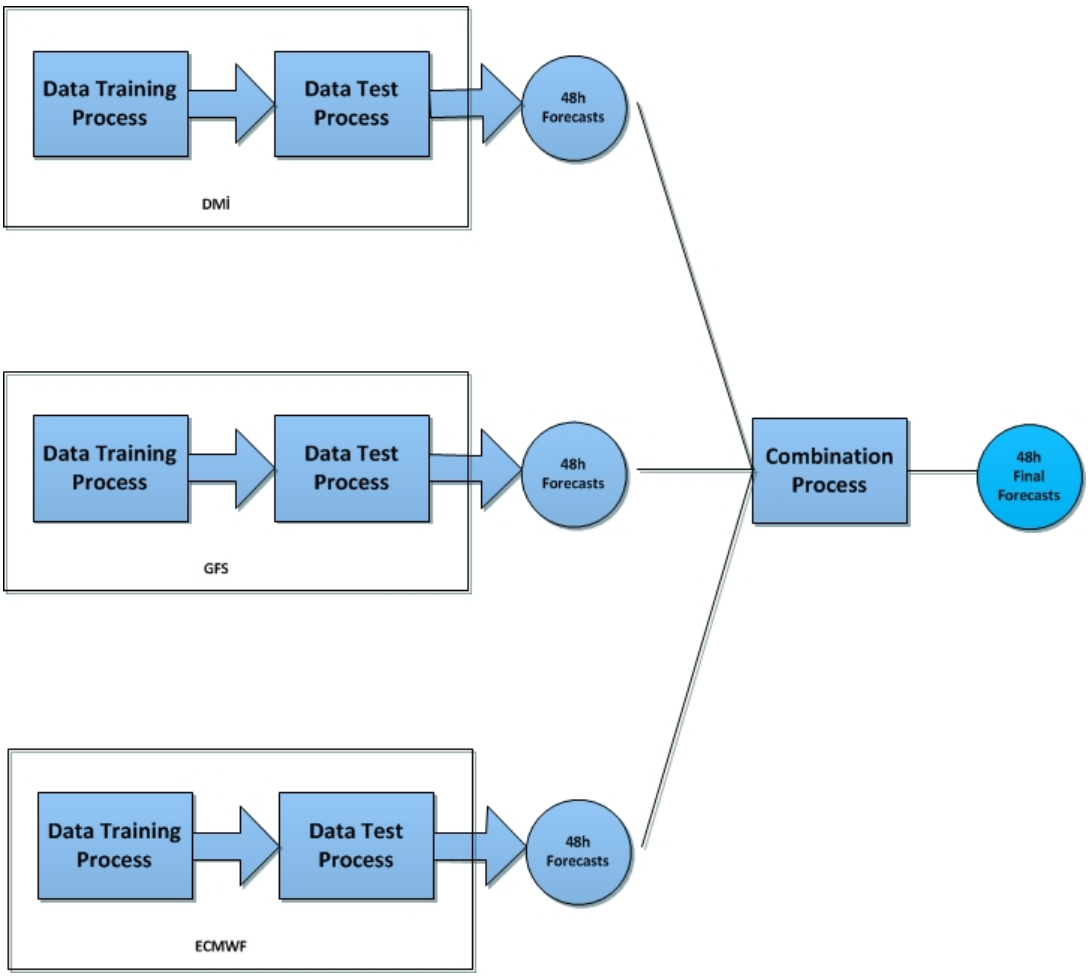

Figure 5.7: Combination structure of the SHWPF model

of this phase, for each WPP, 48 hour day ahead power forecasts are obtained. Additionally, these two phases are conducted for three different NWP (DMİ, GFS, ECMWF) sources independently. Therefore, at the end of the Test phase, for a WPP, three different 48 hour power forecast tuples are generated. These power forecast tuples are used in the "Combination phase" in order to obtain the final power forecasts.

5.4 Combination Phase in SHWIP

Combination phase is the last process of the proposed SHWIP model. The aim of the Combination phase is obtaining a better final power forecast from the

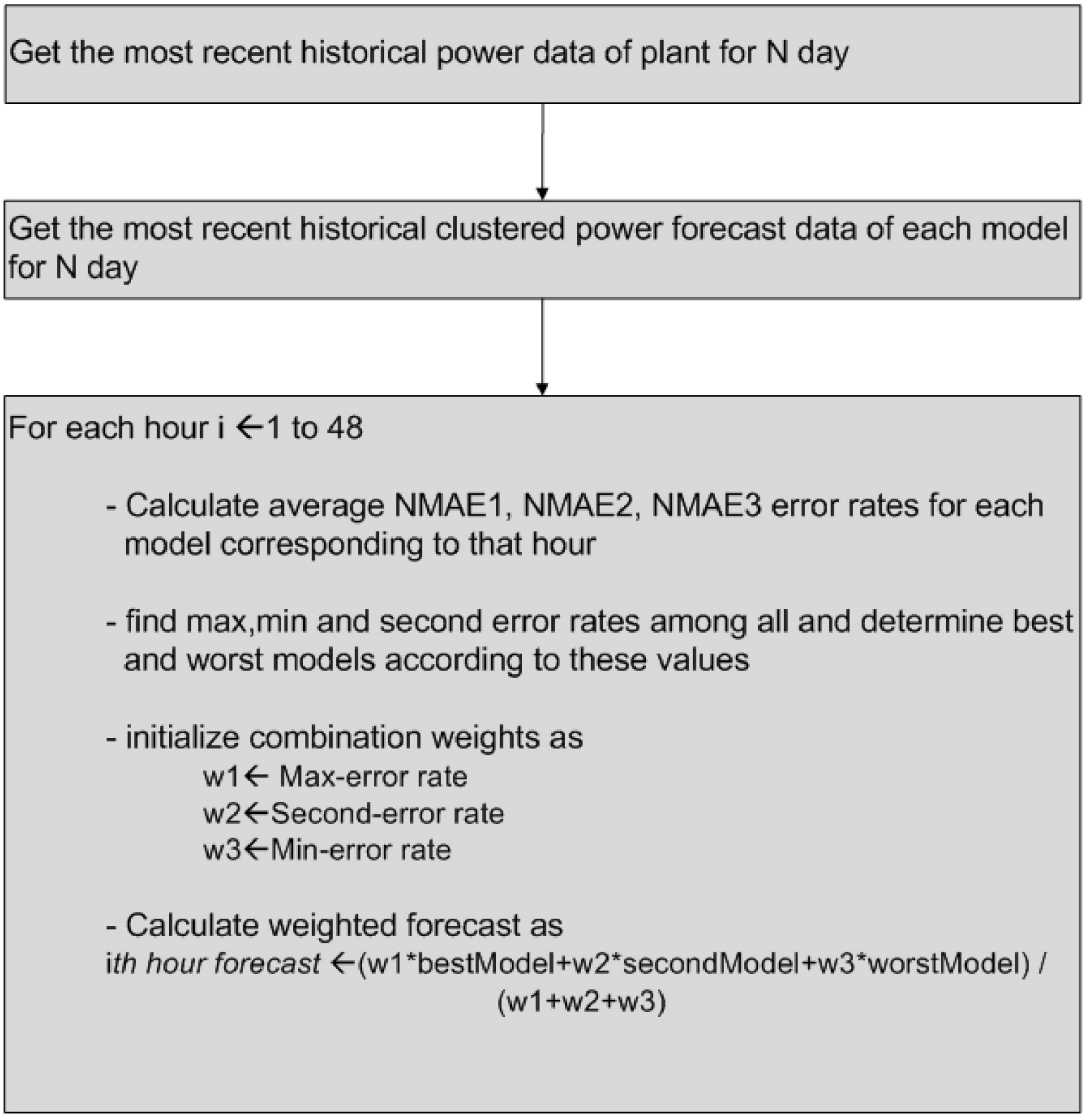

Figure 5.8: Combination algorithm

individual power forecasts obtained from the each NWP source. The general structure is shown in Figure 5.7 [2]. Training and Test phases are executed for each NWP sources independently and at the end of these phases three 48 hour power forecast tuples are obtained for the DMİ, GFS and ECMWF sources. These power forecast tuples are combined into a one final 48 hour power forecast with a combination algorithm in the Combination phase of the SHWIP model. This combination process is conducted for each WPP respectively every day after the execution of the training and test phase for each model independently.

Power forecast tuples are combined with the combination algorithm as stated in Figure 5.8 [1]. Hourly forecasts are combined into a final forecast by weighted average of the models according to historical performance at that hour. The model which has the lowest average Normalized Mean Absolute Error (NMAE) rate for that hour is weighted with maximum error rate and the worst model in that hour is weighted with minimum error rate. This operation is conducted for each hour independently. At the end of this process, the three 48-hour pure clustered forecasts tuples are hybridized into a one final forecast. In the proposed SHWIP model, the N value in Figure 5.8 is selected as 30 and the forecasts are combined according to last one month performances.

The best improvements in the proposed model are obtained from the Combination phase. For example, assume that for a WPP, each NWP's interim forecast NMAE error rates are 14.34 %, 14.46 % and 14.25 %, respectively on the average of six months test period. However, the NMAE rate of the combined forecast is 12.65 % for the same WPP in the same test period. However, combining the power forecasts has also some negative aspects such as losing the ramps in the forecasts and obtaining more smoothed time series. The details of the results are presented in Chapter 6.

CHAPTER 6

EXPERIMENTAL RESULTS

This chapter presents the evaluation results of the proposed SHWIP model. The results are obtained for the time period between the June 2012 to end of the December 2012 for seven months period. The results are given for 14 WPPs that were in the system from the start of the RİTM project. Due to the agreement between WPPs, their real names are not given in the tables and their names are stated as WPPx. Performance results are obtained from the hourly forecast and power data in this time period. Models are compared between them according to average NMAE, RMSE and BIAS results during this seven months time period.

In the "Dynamic Clustering and Discussions" sub-section, the improvements obtained from the dynamic clustering process is presented for each NWP source. Generally, dynamic cluster results are better than a fixed number cluster result for all WPPs.

The comparison results of the final forecasts of proposed model with well known models' forecasts in the literature are presented in the "Combination Results and Discussions" sub-section. Final forecasts of the model are constructed after the Combination phase. The proposed SHWIP model is compared with ANN, SVM and Physical model which also run in the RİTM center every day. Also models' bias analysis results are presented in this sub-section.

Finally, the statistical models' performance comparison results for different amount of training data are given in the "Training Data-Performance" sub-section.

Table 6.1: Cluster number change for a sample WPP

Training Date	WPP1		
	DMİ	GFS	ECMWF
1 June 2012	4	3	3
15 June 2012	7	3	2
1 July 2012	3	3	2
15 July 2012	4	6	2
1 August 2012	6	3	3
15 August 2012	4	3	3
1 September 2012	4	2	5
15 September 2012	7	2	4
1 October 2012	5	5	3
15 October 2012	2	6	3
1 November 2012	3	7	4
15 November 2012	4	3	3
1 December 2012	5	4	3
15 December 2012	7	4	3

6.1 Dynamic Clustering Results and Discussions

This section presents the comparison of the Dynamic Clustering results and fixed cluster numbers for each NWP sources. As stated before in Section 1, in the proposed model, the number of clusters is dynamically determined in the training phase for each WPP with the constraint that the size is between 2 and 7. For example, optimal cluster number change table for a sample WPP for different training days is given in Table 6.1. For this WPP, the cluster numbers for DMİ varies from day to day and data is clustered different numbers of partition from 2 to 7. On the other hand, the situation is more stable for the GFS and ECMWF data and generally optimal cluster numbers are varying between 2 and 4. The graphical representation of the cluster numbers for WPP1 in different training

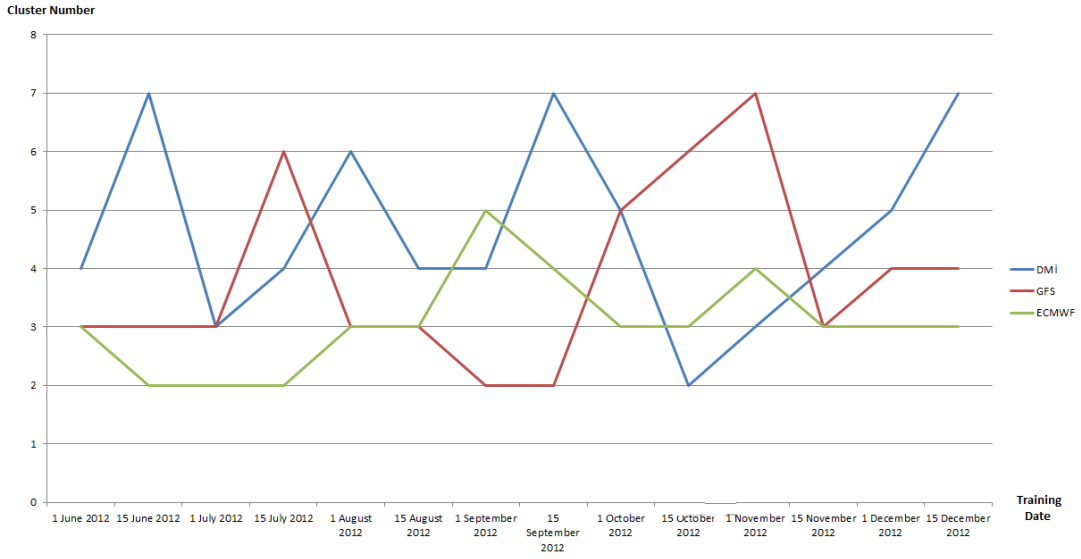

Figure 6.1: Cluster Number Change Graphic for WPP1

dates are represented in Figure 6.1. From this graphic also it is seen that, the cluster numbers are more fluctuating for DMİ forecast data, on the other hand they are more static for the ECMWF forecast data. This factor is related with the differences between the weather forecasts for certain periods. Especially, for the Spring days, the weather forecasts may be too different from each other and due to this factor cluster numbers can vary for different NWP sources.

The reason behind applying a constraint on the selection of number of clusters between 2 and 7 is that the NMAE rates were deteriorating when we portion the data more than 7 clusters and the error rates for cluster number between 2 and 7 were close to each other and they were applicable. So, in the proposed model, the optimal cluster number for the NWP data is determined in this cluster number interval. The test results for each different cluster number and dynamic cluster number for DMİ data is presented in Table 6.2. In this seven months test period, the dynamic cluster NMAE rates are lower than for a particular cluster number for all 14 WPPs. One of the best improvements is observed in WPP8 and it is the second largest WPP in the system. Generally, the NMAE rates are for cluster numbers between 8 and 15 are higher than the error rates obtained from the cluster numbers which are lower than 8.

Table 6.2: Dynamic Clustering DMİ NMAE rates percentage (%)

WPP	2 Cluster	3 Cluster	4 Cluster	5 Cluster	6 Cluster	7 Cluster	8 Cluster	9 Cluster	10 Cluster	15 Cluster	Dynamic
WPP1	14.90	14.87	14.45	14.75	14.55	14.68	14.71	15.00	14.94	15.18	**14.21**
WPP2	12.36	12.50	12.22	12.34	12.47	12.35	12.38	12.42	12.59	12.79	**11.67**
WPP3	14.80	15.19	14.59	14.56	14.81	15.01	15.09	15.09	15.02	15.75	**14.34**
WPP4	17.69	18.15	17.50	17.60	17.57	17.12	17.70	17.66	17.99	17.74	**16.96**
WPP5	13.10	12.99	13.33	13.15	13.18	13.61	13.33	13.27	13.39	13.71	**12.97**
WPP6	12.11	12.54	12.14	11.94	12.09	12.27	12.32	12.42	12.26	12.65	**11.92**
WPP7	13.05	13.13	12.94	12.99	13.04	12.99	13.06	13.02	13.03	13.34	**12.39**
WPP8	11.34	11.43	11.16	11.19	11.35	11.38	11.59	11.31	11.51	11.68	**10.87**
WPP9	11.28	11.31	11.22	11.33	11.41	11.79	11.62	11.56	11.87	11.99	**11.11**
WPP10	16.96	16.42	17.42	17.26	17.39	17.39	17.80	16.84	17.60	18.16	**16.40**
WPP11	18.88	19.38	19.07	19.08	19.19	19.33	19.79	19.32	19.54	19.90	**18.34**
WPP12	15.58	14.94	15.37	15.62	15.60	15.79	15.76	15.86	15.57	16.33	**14.89**
WPP13	13.87	13.34	14.07	14.19	14.23	14.29	14.00	13.80	14.10	13.94	**13.21**
WPP14	14.30	13.82	14.22	14.79	14.76	14.75	14.89	15.04	15.02	16.43	**13.82**

Table 6.3: Dynamic Clustering GFS NMAE rates percentage (%)

WPP	2 Cluster	3 Cluster	4 Cluster	5 Cluster	6 Cluster	7 Cluster	8 Cluster	9 Cluster	10 Cluster	15 Cluster	Dynamic
WPP1	13.60	14.06	13.81	13.75	13.98	13.95	13.72	13.79	13.82	13.68	**13.35**
WPP2	11.10	11.10	11.12	11.11	11.18	11.38	11.45	11.36	11.37	11.34	**10.79**
WPP3	14.80	14.75	15.30	15.15	15.19	15.03	14.94	14.72	14.98	15.05	**14.46**
WPP4	16.45	16.54	16.91	17.18	16.94	17.19	17.22	17.07	16.98	17.13	**16.27**
WPP5	14.18	**13.75**	14.49	14.38	14.39	14.80	14.73	14.52	14.57	14.73	14.04
WPP6	12.21	12.81	12.40	12.30	12.51	12.49	12.59	12.59	12.49	12.57	**12.04**
WPP7	11.96	11.75	11.80	11.95	11.97	12.22	12.27	12.21	12.37	12.65	**11.47**
WPP8	11.78	11.25	11.49	11.52	11.40	11.43	11.54	11.86	11.54	11.55	**10.80**
WPP9	**9.96**	10.22	10.74	10.96	10.88	10.94	10.56	10.78	10.75	10.67	10.12
WPP10	14.52	14.79	14.75	14.73	14.93	14.82	14.79	14.94	14.65	15.46	**14.27**
WPP11	19.01	17.65	18.62	18.00	18.02	18.25	18.49	18.37	18.70	18.54	**17.30**
WPP12	15.78	16.03	16.62	16.44	16.33	16.12	16.52	16.56	16.58	17.03	**15.55**
WPP13	15.09	**14.37**	15.67	15.53	15.56	15.42	15.29	15.66	15.35	15.69	14.42
WPP14	13.53	13.11	13.36	13.40	13.86	13.83	13.84	14.26	14.25	14.15	**12.85**

Table 6.4: Dynamic Clustering ECMWF NMAE rates percentage (%)

WPP	2 Cluster	3 Cluster	4 Cluster	5 Cluster	6 Cluster	7 Cluster	8 Cluster	9 Cluster	10 Cluster	15 Cluster	Dynamic
WPP1	12.73	12.55	12.70	12.84	12.63	12.58	12.75	13.09	12.81	12.72	**12.02**
WPP2	10.27	10.67	10.72	10.58	10.53	10.54	10.45	10.77	10.68	10.70	**10.11**
WPP3	14.43	15.03	15.14	14.74	14.75	14.81	14.69	14.66	15.07	14.85	**14.25**
WPP4	16.36	16.13	16.52	16.58	16.67	16.77	16.49	16.53	16.49	17.22	**15.72**
WPP5	12.58	13.10	12.65	12.84	12.85	12.89	12.74	12.89	12.67	12.92	**12.43**
WPP6	11.57	11.92	11.78	11.67	11.75	11.92	11.78	11.74	11.79	12.09	**11.04**
WPP7	11.76	11.94	12.24	12.05	12.07	11.93	11.78	11.82	11.93	11.85	**11.23**
WPP8	10.44	10.04	10.89	11.06	10.87	10.88	10.63	10.67	10.64	10.72	**9.81**
WPP9	10.44	10.96	10.87	10.95	10.77	10.79	10.72	10.89	10.73	10.66	**10.11**
WPP10	13.30	13.57	13.35	13.68	13.90	13.76	13.91	13.88	14.23	14.61	**13.15**
WPP11	17.48	17.31	17.39	17.75	18.13	18.53	18.20	18.08	17.96	18.18	**16.57**
WPP12	18.23	17.78	18.00	18.64	18.41	17.76	17.83	18.02	18.12	18.29	**17.14**
WPP13	13.59	13.16	14.20	14.25	14.05	13.75	13.89	14.30	14.08	14.99	**12.75**
WPP14	12.95	12.51	13.16	13.11	13.25	13.18	13.36	13.42	13.58	14.16	**11.99**

For the GFS source, the obtained power forecast error rates in the same test period are given in the Table 6.3. In 11 WPPs, the dynamic clustering results are better than all other fixed number of cluster forecast error rates. For WPP5 and WPP13, 3 cluster power forecast and for WPP9, 2 cluster power forecast error rates are lower than the dynamic clustering power forecast error rates. However, the difference is not too much and in all of the other WPPs dynamic clustering results improve the performance of the forecasts. Similar to DMİ results, most of the improvement in the dynamic clustering operation is obtained in the WPP8. As stated before, WPP8 is the second largest WPP in the system according to installed capacity and it is the largest WPP according to area size. Since, the area of that plant is larger than the others, clustering the weather situations in the region and making the WPP's forecasts by using different grid points NWP values is improving the power forecasts positively.

The results for the ECMWF source are presented in the Table 6.4. In the ECMWF data, as in the DMİ source, the dynamic clustering power forecast results are better than the fixed sized clustering in all of the 14 WPPs. The highest improvements are obtained for the WPP11 and WPP12. Especially, in all of the results for the three sources, the worst error rates are seen for WPP11. This situation is more related with the wind farm's physical characteristics.

One month sample real production and dynamic clustering power forecast line graphics for ECMW data for three different WPPs are shown in the Figure 6.2, 6.3 and 6.4. The sample time series graphics are given for the WPP8 which has the lowest error rate, WPP1 which has the average error rate and WPP12 which has the maximum error rate. Graphics are drawn by using the data obtained from the 1 June 2012 and 1 July 2012. WPP8 is in the south part of Turkey on the other hand WPP1 and WPP12 in the western region of Turkey. Additionally, WPP8's installed capacity is 135 MW and WPP1's and WPP12's installed capacities are 12 and 30 MW, respectively.

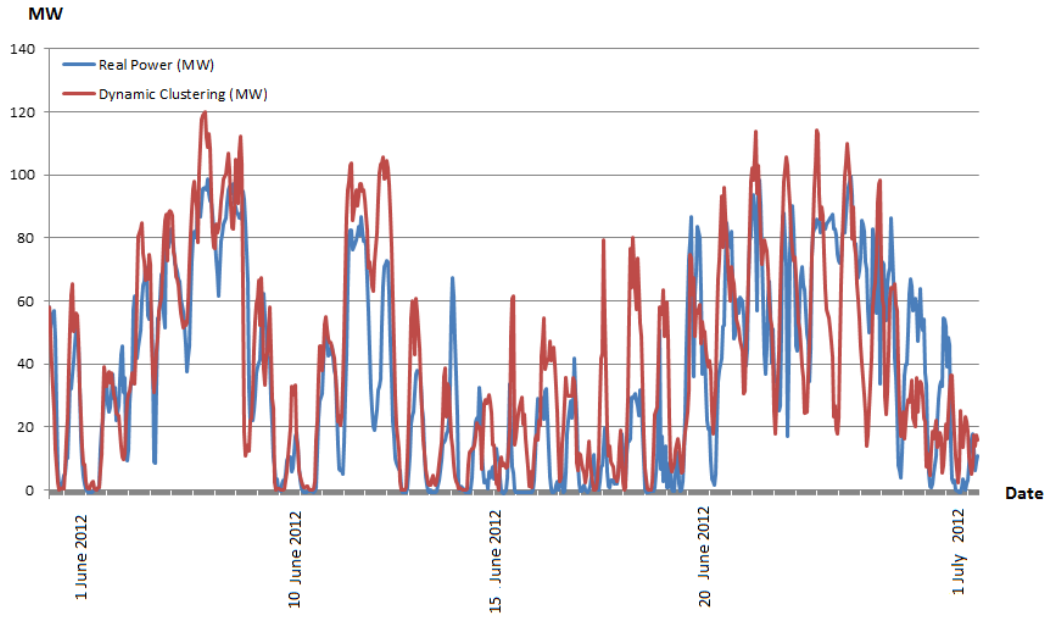

Figure 6.2: Power and Dynamic Clustering Forecast for WPP8

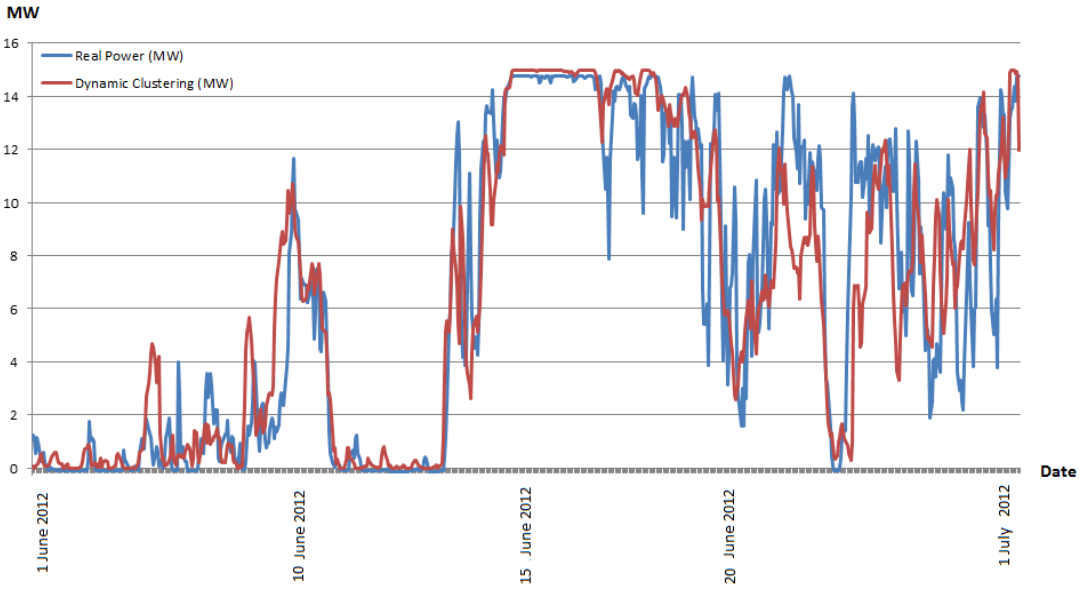

Figure 6.3: Power and Dynamic Clustering Forecast for WPP1

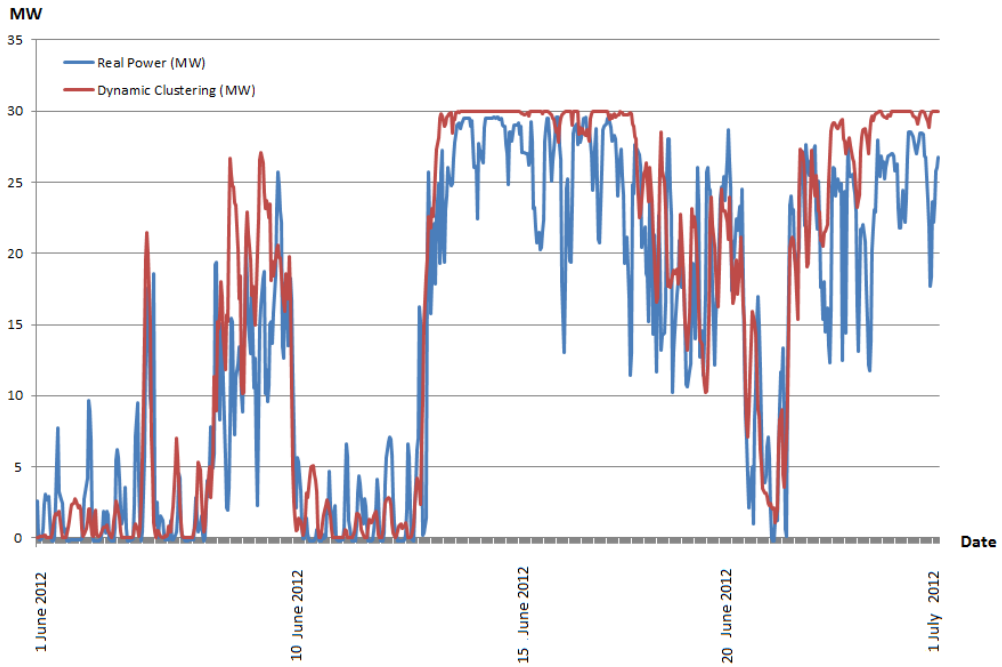

Figure 6.4: Power and Dynamic Clustering Forecast for WPP12

6.2 Combination Results and Discussions

This section presents the combined power forecast results which are obtained from the last phase (Combination Phase) of the proposed SHWIP model. As stated before in Chapter 5, Training and Test phases are applied for each different NWP sources independently every day and as the result of these phases, three 48 hour wind power forecast tuples are formed. These three 48 hour forecast tuples are combined to 48 hour final forecasts with a combination algorithm whose details are given in the Section 5.3.

The results of combined forecasts with Dynamic Clustering results for each NWP sources are presented in the Table 6.5. The results are given in terms of average NMAE rates obtained during the same seven month test periods. As stated before, the best improvements in the model are obtained from the Combination Phase. As seen in Table 6.5, in all of the WPPs, the combined wind power forecast results have lower error rates than the dynamic clustering power forecasts.

Table 6.5: Combination and Dynamic Clustering Results (In terms of NMAE %)

WPP	DMİ Cluster	GFS Cluster	ECMWF Cluster	Combined
WPP1	14.21 %	13.35 %	12.02 %	**11.66** %
WPP2	11.67 %	10.79 %	10.11 %	**9.78** %
WPP3	14.34 %	14.46 %	14.25 %	**13.52** %
WPP4	16.96 %	16.27 %	15.72 %	**14.04** %
WPP5	12.97 %	13.04 %	12.43 %	**12.24** %
WPP6	11.92 %	12.04 %	11.04 %	**10.43** %
WPP7	12.39 %	11.47 %	11.23 %	**10.53** %
WPP8	10.87 %	10.80 %	9.81 %	**8.90**%
WPP9	11.11 %	10.12 %	10.11 %	**9.14** %
WPP10	16.40 %	14.27 %	13.15 %	**12.62** %
WPP11	18.34 %	17.30 %	16.57 %	**16.32** %
WPP12	14.89 %	15.55 %	17.14 %	**11.8** %
WPP13	13.21 %	14.42 %	12.75 %	**11.21** %
WPP14	13.82 %	12.85 %	11.99 %	**11.21** %

According to results presented in Table 6.5, in all of the 14 WPPs, the combined forecasts results are better than a particular dynamic clustering wind power forecast. The most of the improvement is seen in WPP12 with the average 11.80 % NMAE rate. In this plant, the dyanmic clustering error rates are 14.89 %, 15.55 % and 17.14 % respectively. The worst improvement is obtained from the WPP5 which is the smallest WPP in the system according to installed capacities with its 10.2 MW capacity. This plant is near to the cost and does not have very complex physical characteristic. Due to this factor, the NWPs for this WPP are similar in all three sources and their error rates are similar to each other. Since a particular NWP source is not obviously superior then the others, the combined results are also near to dynamic clustering results. Two sample one week real power and forecast graphics obtained from the results of WPP12 and WPP5 are presented in Figure 6.5 and 6.6. The graphics are drawn for the 15-22 Aug 2012 time period.

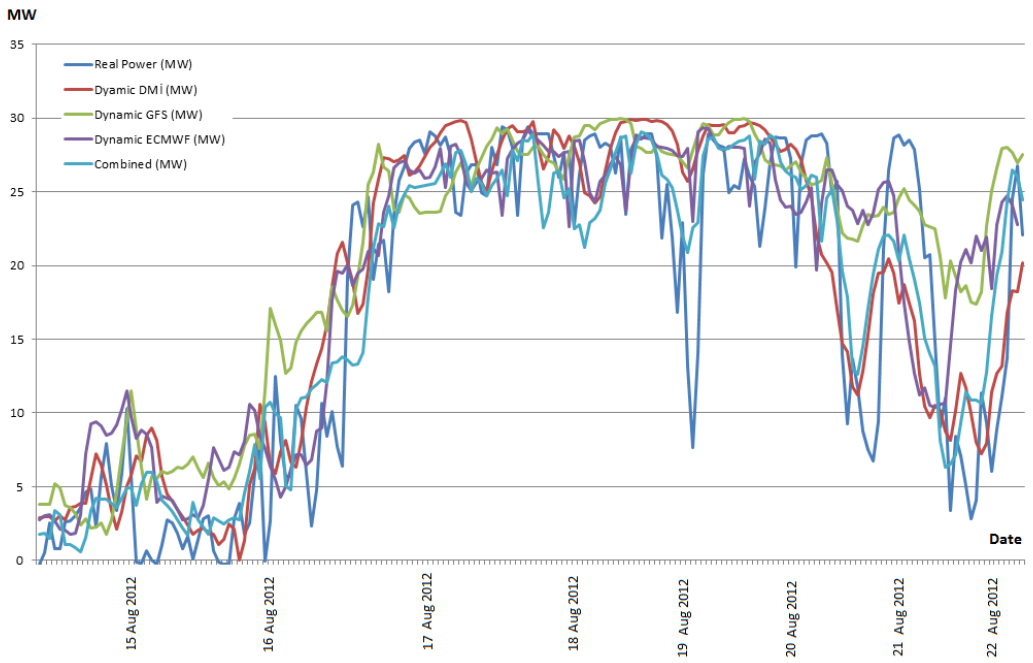

Figure 6.5: Dynamic Clustering and Combined Forecast for WPP12

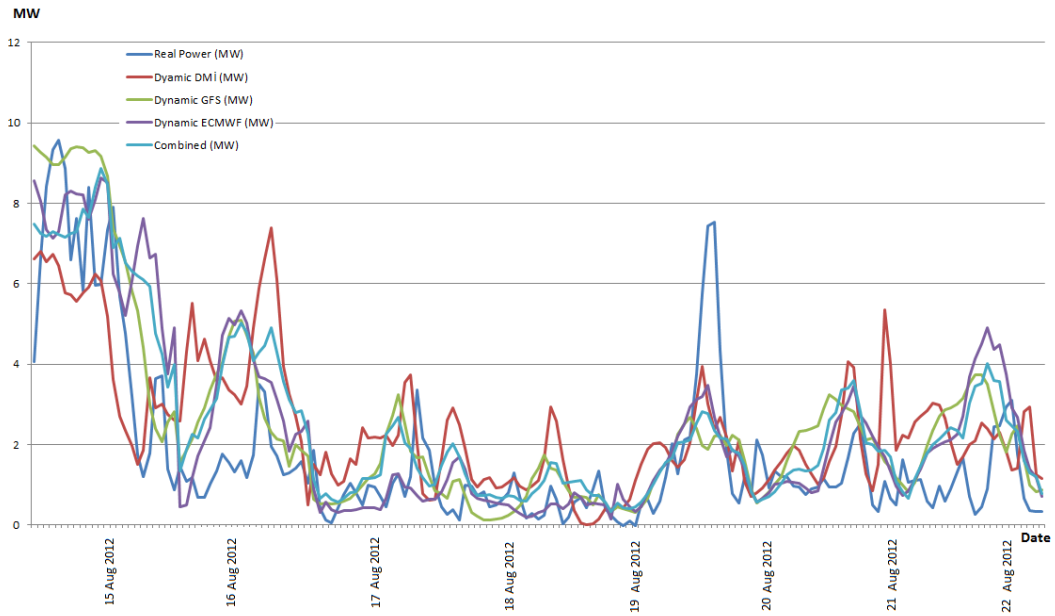

Figure 6.6: Dynamic Clustering and Combined Forecast for WPP5

One of the measurement technic to visualize the performance of a model is creating histogram graphics to see the error distributions. Error distributions histogram graphic for a pilot WPP in the system is given in the Figure 6.7 [2]. Most of the hourly errors are in 0-5 % percentage interval for the WPPs. On the other hand, there are a few hours which error rates reach to 50% percentages. However, this situations are very rare and they do not increase the overall error rate of the WPP too much.

The forecasts throughout the country are obtained from the sum of the forecasts of individual WPPs in the system. In the same test period, NMAE rate for all Turkey is measured as 5.74 % [2]. This number is less than for an individual WPP error rates as expected since it is obtained from all of the WPPs in the system. In addititon, generally in such monitoring centers, if the monitored installed capacity increases then the total error rate reduces. So, it is predicted that after the all WPPs in Turkey are monitored from the system, this error rate may reduce the 4-5 % interval. Similar results are obtained from the forecasts of all Turkey whose error distribution histogram graphics is shown in Figure 6.8. According to histogram graphics, the hourly error rates are well distributed in the model where the error values mostly close to 0 and there are no errors with ratio greater than 30 % for all Turkey.

Although combining the wind power forecasts is improving the overall performance of the system on average, it also has some disadvantages. The final forecasts become a more smoothed time series graphics after the combination operation and it may cause some unintended consequences. Especially, if the ramps (sudden ups and downs in the power) are determined correctly from one of the forecast source, after the combination this information is lost. On the other hand, if the ramps are determined by one of the NWP forecast source wrongly, in the combination step this incorrect situation is fixed and the combined forecasts become more near to real production. As a result, in the overall case combining the forecasts is improving the system performance but it is not successful on determining the ramps on the power and this issue is one of the other research ares in wind power forecasting [45].

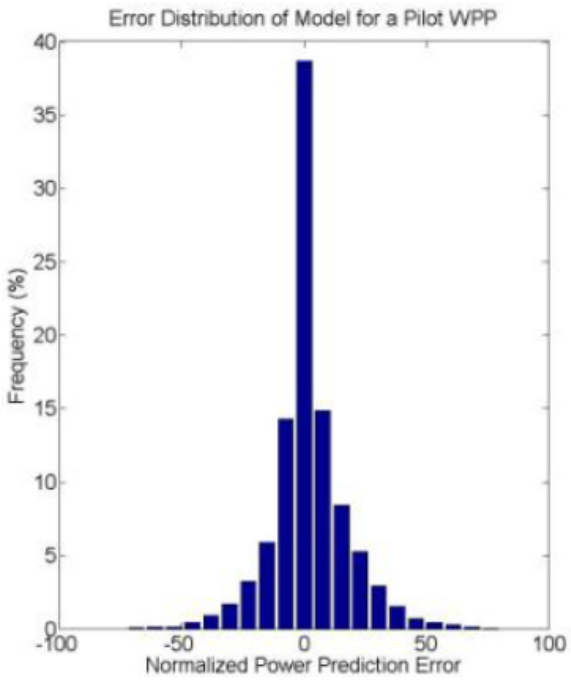

Figure 6.7: Error distribution for a pilot WPP

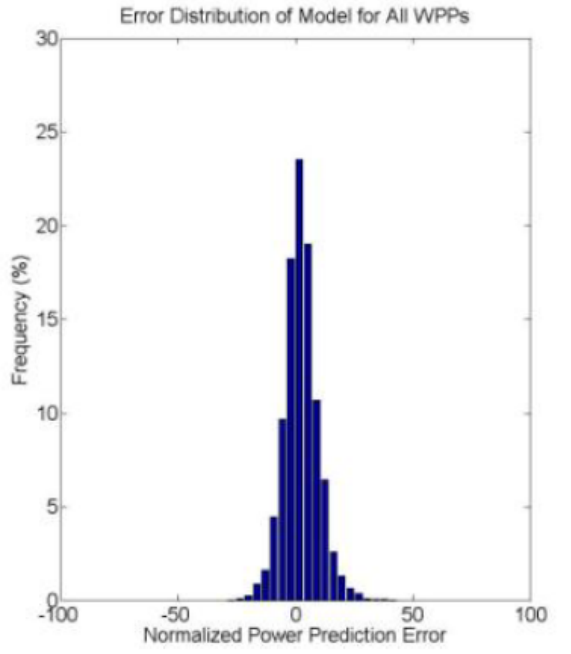

Figure 6.8: Error Distributions for all WPPs in Turkey

6.3 Comparison of the Model with Other Models

This section compares the proposed forecast model (SHWIP) with other two well-known statistical models (ANN and SVM) in the literature and a physical model. Comparisons are made according to NMAE, NRMSE and normalized BIAS results obtained during the described test period. The equations for the comparison measurements are given in the Equations 6.1, 6.2 and 6.3. In these formulas x_i is real power and y_i is the power estimations in the i^{th} hour, C is capacity of the WPP and N is the total number of hours.

$$NMAE = \frac{\frac{\sum_{i=1}^{N} |y_i - x_i|}{N}}{C} \times 100 \tag{6.1}$$

$$NRMSE = \frac{\sqrt{\frac{\sum_{i=1}^{N} (y_i - x_i)^2}{N}}}{C} \times 100 \tag{6.2}$$

$$BIAS = \frac{\frac{\sum_{i=1}^{N} y_i - x_i}{N}}{C} \times 100 \tag{6.3}$$

Both of the ANN and SVM models are used as regression type model for power generation prediction of each WPP. In all of the WPPs, one year historical wind speed and wind direction data are used for training the network and constructing the mathematical modelling. Then these trained models are used for online wind power generation forecast on NWP test data. A well known and fast library for ANN, which is Fast Artificial Neural Network Library (FANN), is used for the implementation of the ANN model [46]. Similarly, the second statistical model SVM is used as regression type model which can be called Support Vector Regression (SVR) for power generation forecasts. In order to implement the SVM model, Library for Support Vector Machines (LIBSVM) is used [47]. A physical model is implemented by using the NWPs of the four nearest grid points to WPP center point. Each model has different characteristics, therefore, in some of the WPPs, the differences between the error rates increase due to WPPs physical characteristics.

Table 6.6: Evaluation Results of Models (In terms of NMAE %)

WPP	ANN	SVM	Physical	SHWIP
WPP1	12.52 %	12.60 %	12.43 %	**11.66** %
WPP2	**9.33** %	9.37 %	10.26 %	9.78 %
WPP3	13.10 %	**13.01** %	17.27 %	13.52 %
WPP4	15.03 %	15.43 %	16.27 %	**14.04** %
WPP5	13.43 %	13.37 %	12.51 %	**12.24** %
WPP6	11.19 %	10.50 %	10.45 %	**10.43** %
WPP7	11.53 %	10.79 %	12.24 %	**10.53** %
WPP8	9.67 %	9.40 %	13.09 %	**8.90** %
WPP9	8.76 %	**8.61** %	12.50 %	9.14 %
WPP10	14.82 %	14.53 %	17.95 %	**12.62** %
WPP11	17.08 %	16.48 %	19.03 %	**16.32** %
WPP12	13.18 %	12.12 %	13.27 %	**11.8** %
WPP13	15.48 %	14.32 %	14.58 %	**11.21** %
WPP14	11.76 %	11.58 %	14.22 %	**11.21** %

The NMAE rates calculated for four different power forecast models for each WPP during the seven months test period are presented in Table 6.6. According to obtained error rates, in all of the WPPs except for WPP2, WPP3 and WPP9, the proposed SHWIP model has the lowest error rates. In WPP2, ANN error rate and in WPP3 and WPP9, SVM error rates are lower than the proposed model. The SHWIP model has the lowest error rates which are lest than 10 % in WPP2, WPP8 and WPP9. These three plants are the three biggest WPPs in the RİTM system according to installed capacities. Both of ANN and SVM model have the highest error rates when compared to the other models in WPP10 and WPP13. These two plants are in the southern part of Turkey with WPP8 and the plant areas in this region have rough geographic characteristics compared to other regions. Similar to ANN and SVM, physical forecast results are not also better in this region plants compare to other plants due to same reason.

Table 6.7: Evaluation Results of Models (In terms of NRMSE %)

WPP	ANN	SVM	Physical	SHWIP
WPP1	18.51 %	18.29 %	18.00 %	**17.98** %
WPP2	**14.36** %	14.37 %	15.32 %	14.73 %
WPP3	18.89 %	**18.57** %	22.63 %	19.48 %
WPP4	20.53 %	**20.46** %	25.03 %	21.95 %
WPP5	19.14 %	19.18 %	18.15 %	**17.76** %
WPP6	16.26 %	15.47 %	15.39 %	**15.35** %
WPP7	16.18 %	15.72 %	17.00 %	**15.41**%
WPP8	13.67 %	13.48 %	19.49 %	**13.10** %
WPP9	**13.65** %	13.49 %	18.43 %	13.98 %
WPP10	18.34 %	18.09 %	24.43 %	**17.54** %
WPP11	23.68 %	**23.06** %	26.43 %	23.96 %
WPP12	17.88 %	17.13 %	19.12 %	**17.00** %
WPP13	19.95 %	19.70 %	19.02 %	**15.82** %
WPP14	18.21 %	18.13 %	20.11 %	**16.89** %

Another measurement method for the error rates is the NRMSE whose results are presented in the Table 6.7. In the RMSE calculation, since the errors are squared before they are averaged, it gives high weight to large errors. Due to this reason, the NRMSE error rates are higher than the NMAE rates in all of the models.

Similar to NMAE results, the proposed SHWIP model has lower error rates in 9 of 14 WPPs. In all of the five WPP which SHWIP has high error rates, both ANN and SVM have lower error rates than SHWIP. Similar to NMAE results, ANN and SVM error rates are also high in the southern plants and these models do not work very well in this region. Physical model has higher error rate than the SHWIP model in all of the WPPs. In addition to that, Physical model has lower NRMSE rate compare to both ANN and SVM in 4 of 14 WPPs. All of the models have the highest error rates in WPP11.

Table 6.8: Evaluation Results of Models (In terms of Normalized BIAS %)

WPP	ANN	SVM	Physical	SHWIP
WPP1	0.17 %	-0.14 %	1.09 %	-1.48 %
WPP2	3.28 %	-4.24 %	2.96 %	-1.97 %
WPP3	1.62 %	-1.61 %	11.94 %	-1.41 %
WPP4	0.86 %	-0.80 %	8.88 %	-0.91 %
WPP5	0.06 %	-0.04 %	0.23 %	-0.69 %
WPP6	0.17 %	-0.01 %	-0.79 %	-2.81 %
WPP7	1.41 %	-0.53 %	4.38 %	-2.09 %
WPP8	11.49 %	-2.24 %	1.43 %	-2.64 %
WPP9	-6.46 %	-4.56 %	0.23 %	-2.98 %
WPP10	-1.47 %	-0.34 %	3.69 %	-3.65 %
WPP11	-0.32 %	-0.14 %	-2.52 %	-3.92 %
WPP12	-0.36 %	0.82 %	-0.1 %	-4.08 %
WPP13	4.21 %	-8.39 %	1.95 %	-2.61 %
WPP14	-0.45 %	0.23 %	1.21 %	-2.25 %

BIAS is another measurement method considered in the wind power forecasting. It is calculated directly by substraction of expected value from the observation value and it gives an idea about the trend of the forecasts. The BIAS results calculated for each WPP are given in Table 6.8.

According the results, the proposed SHWIP model has negative bias values in all of the 14 WPPs. This situation is more related with the combination process. Although the final combined forecasts become more accurate in the average, it loses the information about sudden rises in the forecasts some times and this situation causes the negative bias in the model.

In all of the other models, the bias values are fluctuating and there is no prominent trend in the calculated values. This situation shows that these models have inconsistency for the WPPs located in the different regions of Turkey.

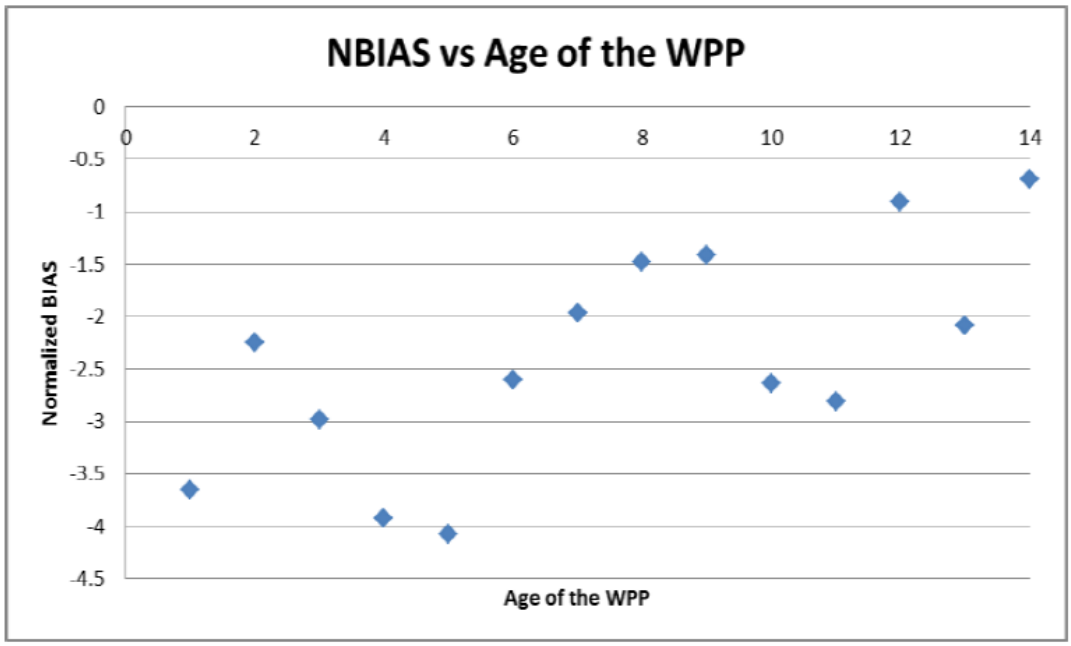

Figure 6.9: The relation between Normalized BIAS and Age of the WPPs

Several plausible results can be deduced from the bias values. Firstly, the proposed SHWIP model has negative biases for all WPPs and this manifests the consistency of the proposed forecast model. Additionally, the bias values have an interesting correlation with the ages of the WPPs, as shown in Figure 6.9 where vertical axis denotes the WPP's normalized BIAS values and X axis denotes the age values of the WPPs [2]. The ages of the considered WPPs vary from 2 to 7 years. Generally, bias values are lower for the older WPPs and higher for the newer ones. This situation is mostly related to the turbine availability in the WPPs. In older WPPs, usually all turbines are in operation and as a result of this, the model has a better learning process. On the other hand, at newer WPPs, not all turbines are generally available at the same time which makes learning the power characteristic of such WPPs harder.

The BIAS in the model must be eliminated after the combination process. The elimination can be done in the proposed SHWIP model easier than the other models because all of the WPPs have negative BIAS in the model. This situation is another advantage of the proposed model to the other models.

Table 6.9: p-value Test Results between Models

WPP	SHWIP vs ANN	SHWIP vs SVM	SHWIP vs Physical
WPP1	0.02	0.06	0.96
WPP2	0.12	0.07	0.007
WPP3	0.001	0.001	0.001
WPP4	0.318	0.02	0.001
WPP5	0.001	0.001	0.84
WPP6	0.001	0.14	0.27
WPP7	0.001	0.011	0.001
WPP8	0.001	0.001	0.001
WPP9	0.001	0.001	0.001
WPP10	0.001	0.001	0.001
WPP11	0.001	0.012	0.03
WPP12	0.001	0.04	0.001
WPP13	0.001	0.001	0.001
WPP14	0.001	0.06	0.001

Another common method used in statistic for hypothesis testing is t-Test[48]. This test compares the means of two groups and it decides on whether the difference between two group of data is statistically significant or not. At the end of the test a p-value is obtained and if the obtained p-value is below 0.05 these two groups are accepted statistically different from each other [48]. The p-values obtained from the 14 WPPs are presented in the Table 6.9. The proposed SHWIP model is tested for each model separately by using first 1000 hourly error rate values of each model. Generally, in lots of the WPPs the obtained p-Values are statistically significant. In the WPP5 and WPP6, the plant areas are not complex since they are near to sea level. Due to these reason at these plants the p-values for the Physical Model is high and there is not too much difference in the forecasts in these plants. On the other hand, for some other WPPs which the plant are is more complicated such as WPP8, WPP10 and WPP13 the p-values are statistically significant and the forecasts are too different from each other.

6.4 Experiments to Evaluate the Effect of Training Data Size on the Accuracy

This section presents the performance results of the statistical models with respect to different training day periods. Since physical model does not have training phase, it is not included in this comparison. The results are obtained for four different training day periods as 30, 90, 180 and 360 days and given in the Table 6.10.

The best advantage of the proposed SHWIP model over the other models is the need for less amount of training data. As shown in Table 6.10, the performance results are similar for SHWIP model in 90, 180 and 360 training day periods. The results for the 30 day training period is worse than the other training periods however, 30 day training is also applicable. In the model, the optimal training day is chosen as 90 day since it covers one season data and the computation cost is also lower than compare to 180 and 360 days.

On the other hand, the performance of the ANN and SVM is directly proportional to amount of training data. These models best perform in the 360 days training data and their performance are deteriorating considerably if less amount of training data is used.

In the real world, we may not find one year training data for some of the WPPs especially for the new established ones. At this situation, ANN and SVM models can not produce reliable forecasts for such WPPs. However, the proposed SHWIP model produces reliable forecasts even if we have only one week historical power data for the WPP. For instance, in the RİTM center, 6 new WPPs are added to system at the end of 2013 and for these plants ANN and SVM models did not produce reliable forecasts and currently they are not running for these plants. On the other hand, the forecasts for these plants are produced by the SHWIP model by using one week training data. Although, the results were not as good as that of 90 day training data, they were also applicable and may be used by the WPP owners and TSOs.

Table 6.10: Error Rates of Models for Different Training Data Amount (In terms of NMAE %)

WPP	30 DAYS			90 DAYS			180 DAYS			360 DAYS		
	ANN	SVM	SHWIP	ANN	SVM	SHWIP	ANN	SVM	SHWIP	ANN	SVM	SHWIP
WPP1	14.62 %	13.64%	**12.25** %	13.49 %	13.24 %	**11.66** %	13.93 %	13.42%	**11.54** %	12.52 %	12.60 %	**11.46** %
WPP2	12.90 %	11.50%	**10.94** %	10.79 %	10.19 %	**9.78** %	10.18 %	10.14%	**9.91** %	**9.33** %	9.37 %	10.04 %
WPP3	16.72 %	15.58%	**14.71** %	14.66 %	14.50 %	**13.52** %	14.73 %	14.64%	**13.75** %	13.10 %	**13.01** %	13.78 %
WPP4	17.49 %	17.24%	**14.95** %	16.23 %	16.00%	**14.04** %	16.18 %	16.10%	**14.06** %	15.03 %	15.43 %	**13.91** %
WPP5	19.43 %	14.63%	**13.31** %	15.59 %	14.49%	**12.24** %	14.29 %	14.33%	**12.48** %	13.43 %	13.37 %	**12.45** %
WPP6	13.88 %	12.82%	**11.54** %	11.53 %	12.26%	**10.43** %	11.24 %	12.22%	**10.55** %	11.19 %	**10.50** %	10.67 %
WPP7	15.18 %	13.05%	**11.88** %	12.23 %	12.89%	**10.53** %	12.12 %	12.31%	**10.40** %	11.53 %	10.79 %	**10.47** %
WPP8	14.55 %	17.25%	**10.24** %	11.49 %	10.97 %	**8.90** %	10.74 %	10.92%	**9.08** %	9.67 %	9.40 %	**9.14** %
WPP9	17.61 %	17.06%	**10.58** %	13.27 %	13.14 %	**9.14** %	12.72 %	12.83%	**9.17** %	8.76 %	**8.61** %	9.25 %
WPP10	19.42 %	24.03%	**13.27** %	11.01 %	21.92%	**12.62** %	18.13 %	19.28%	**12.58** %	14.32 %	14.03 %	**12.45** %
WPP11	19.64 %	19.02%	**17.25** %	18.81 %	19.58%	**16.32** %	18.83 %	19.70%	**16.15** %	17.08 %	16.48 %	**16.02** %
WPP12	18.08 %	14.10%	**12.72** %	12.58 %	12.30%	**11.80** %	12.12 %	12.53%	**11.89** %	13.18 %	12.12 %	**11.78** %
WPP13	18.37 %	18.17%	**12.05** %	17.58 %	17.85%	**11.21** %	13.62 %	14.05%	**11.35** %	15.48 %	14.32 %	**11.45** %
WPP14	15.95 %	13.07%	**12.26** %	13.88 %	14.16%	**11.21** %	12.46 %	12.68%	**11.18** %	11.76 %	11.58 %	**11.14** %

CHAPTER 7

CONCLUSION AND FUTURE WORK

Wind energy has become one of the most significant alternative energy sources in the last years and the predicting the power production of the WPPs correctly is one of the main concerns of the WPPs owners and grid operators. In this thesis, a new method, namely Statistical Hybrid Wind Power Forecast (SHWIP) is presented. The short term forecast results of the proposed model is in operation at RİTM center almost for one year. Considering the evaluation results, the proposed model performs considerably better than some other well known statistical models and a physical model in the literature.

The main superiority of the method to other statistical models is usage of less historical data in the construction of the model. Generally other statistical models such as ANN and SVM need at least one or two years historical data in order to construct the mathematical model. In the proposed method, generally three months historical data is used in the training phase. However, it also produces acceptable forecasts even if one month data is used. So, especially in the new WPPs or the WPPs that we do not have historical data, the proposed method has an important role in the production of the wind power forecasting.

For the future work, the effects of other clustering methodologies other than k-means may be investigated. Although it seems that using another clustering way does not change the performance of the forecasts too much, in some WPPs which have too much outlier forecast data, it may be beneficial to use other clustering algorithm such as k-medoids. Although during the test period that the results are given, the proposed model produces lower error rates on average,

in some days of the test period, the other models achieve better performance rates compared to new model. Therefore, a good combination process for all physical, ANN, SVM and the proposed SHWIP model may also improve the overall performance of the forecasts for WPPs in Turkey. In addition, if the three interim forecasts are too different from each other for a specific day, then the combined forecast become a more smoothed series and it does not determine the ramps in the forecast properly. Therefore, in order to specify the ramps in the WPP another approach also must be included to current combination algorithm. Also producing the power forecasts for each wind turbine independently may improve the overall results and this factor may be investigated as a future study. At this thesis, this study is not conducted because WPPs did not provide their power data for each turbine so proposed model is constructed from the total production data of the all turbines.

REFERENCES

[1] **Ozkan, Mehmet Baris**, Dilek Küçük, Erman Terciyanlı, Serkan Buhan, Turan Demirci, and Pinar Karagoz. A data mining-based wind power forecasting method: Results for wind power plants in turkey. In *Data Warehousing and Knowledge Discovery*, pages 268–276. Springer, 2013.

[2] **Ozkan, Mehmet Baris**, Erman Terciyanlı, Dilek Küçük, Serkan Buhan, Turan Demirci, Ceyhun Yildiz, and Mustafa Günindi. Verification of a real time wind power monitoring and forecast system for turkey. In *Proceedings of the IET Renewable Power Generation Conference. Beijing, China*. 2013.

[3] Erman Terciyanli, Turan Demirci, Dilek Kucuk, Serkan Buhan, **Ozkan, Mehmet Baris**, Ceyda Er Koksoy, Erkan Koc, Ceren Kahraman, Ali Burhan Haliloglu, Tugba Demir, et al. The architecture of a large-scale wind power monitoring and forecast system. In *Power Engineering, Energy and Electrical Drives (POWERENG), 2013 Fourth International Conference on*, pages 1162–1167. IEEE, 2013.

[4] Serkan Buhan, Erman Terciyanli, **Ozkan, Mehmet Baris**, Dilek Kucuk, Ali Burhan Haliloglu, Turan Demirci, Hilal Tuna, and Hakan Unsal. Verification of a very short term wind power forecasting algorithm for turkish transmission grid. In *Power Engineering, Energy and Electrical Drives (POWERENG), 2013 Fourth International Conference on*, pages 1217–1221. IEEE, 2013.

[5] David Satterthwaite. Cities' contribution to global warming: notes on the allocation of greenhouse gas emissions. *Environment and Urbanization*, 20(2):539–549, 2008.

[6] Claudia Spix, Sven Schmiedel, Peter Kaatsch, Renate Schulze-Rath, and Maria Blettner. Case–control study on childhood cancer in the vicinity of nuclear power plants in germany 1980–2003. *European Journal of Cancer*, 44(2):275–284, 2008.

[7] Ioannis N Kessides. Nuclear power: Understanding the economic risks and uncertainties. *Energy Policy*, 38(8):3849–3864, 2010.

[8] Ritm web page. `http://www.ritm.gov.tr/root/index_eng.php`. Last accessed date: 2013-12-24.

[9] Seyit Ahmet Akdağ and Önder Güler. Evaluation of wind energy investment interest and electricity generation cost analysis for turkey. *Applied Energy*, 87(8):2574–2580, 2010.

[10] Yegm web page. `http://www.eie.gov.tr/`. Last accessed date: 2013-09-10.

[11] Young-Mi Lee. Real-time wind power prediction system based on smartgrid in jeju, korea. *The Journal of International Council on Electrical Engineering*, 2(2):194–200, 2012.

[12] Juan Ma Rodríguez, Olivia Alonso, Miguel Duvison, and Tomás Domíngez. The integration of renewable energy and the system operation: The special regime control centre (cecre) in spain. In *Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE*, pages 1–6. IEEE, 2008.

[13] Steven Lang and Eamon McKeogh. Verification of wind power forecasts provided in real-time to the irish transmission system operator. In *Power and Energy Society General Meeting, 2010 IEEE*, pages 1–6. IEEE, 2010.

[14] Bernhard Ernst, Kurt Rohrig, and René Jursa. Online-monitoring and prediction of wind power in german transmission system operation centres. In *Proceedings of the First IEA Joint Action Symposium on Wind Forecasting Techniques, Norrköping, Sweden*, pages 125–145, 2002.

[15] Windsim web page. `http://www.windsim.com/`. Last accessed date: 2013-10-10.

[16] Wasp web page. `http://www.wasp.dk/`. Last accessed date: 2013-10-10.

[17] C Monteiro, R Bessa, V Miranda, A Botterud, J Wang, G Conzelmann, et al. Wind power forecasting: state-of-the-art 2009. Technical report, Argonne National Laboratory (ANL), 2009.

[18] Alan Russell. Computational Fluid Dynamics Modeling of Atmospheric Flow Applied to Wind Energy Reseaarch. Master's thesis, Boise State University, 2009.

[19] Lars Landberg and Simon J Watson. Short-term prediction of local wind conditions. *Boundary-Layer Meteorology*, 70(1-2):171–195, 1994.

[20] MA Gaertner, C Gallardo, C Tejeda, N Martínez, S Calabria, N Martínez, and B Fernández. The casandra project: results of wind power 72-h range daily operational forecasting in spain. In *Proceedings of the European Wind Energy Conference EWEC'03*, pages 16–19, 2003.

[21] Ue Cali, B Lange, J Dobschinski, M Kurt, C Moehrlen, and B Ernst. Artificial neural network based wind power forecasting using a multi-model approach. In *Proceedings of the 7th International Workshop on Large-Scale Integration of Wind Power and on Transmission Networks for Offshore Wind Farms, Madrid (ES)*, 2008.

[22] Mihai Gavrilas and Gilda Gavrilas. An enhanced ann wind power forecast model based on a fuzzy representation of wind direction. In *Neural Network Applications in Electrical Engineering (NEUREL), 2010 10th Symposium on*, pages 31–36. IEEE, 2010.

[23] Marti A. Hearst, ST Dumais, E Osman, John Platt, and Bernhard Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28, 1998.

[24] Jianwu Zeng and Wei Qiao. Support vector machine-based short-term wind power forecasting. In *Power Systems Conference and Exposition (PSCE), 2011 IEEE/PES*, pages 1–8. IEEE, 2011.

[25] Sancho Salcedo-Sanz, Ángel M Pérez-Bellido, Antonio Portilla-Figueras, Luis Prieto, et al. Short term wind speed prediction based on evolutionary support vector regression algorithms. *Expert Systems with Applications*, 38(4):4052–4057, 2011.

[26] Henrik Aa Nielsen, Torben S Nielsen, Alfred K Joensen, Henrik Madsen, and Jan Holst. Tracking time-varying-coefficient functions. *International Journal of Adaptive Control and Signal Processing*, 14(8):813–828, 2000.

[27] Andrew Kusiak, Haiyang Zheng, and Zhe Song. Wind farm power prediction: a data-mining approach. *Wind Energy*, 12(3):275–293, 2009.

[28] Gregor Giebel, Lars Landberg, Torben Skov Nielsen, and Henrik Madsen. The zephyr-project: The next generation prediction system. In *Proc. of the 2001 European Wind Energy Conference, EWEC'01, Copenhagen, Denmark*, pages 777–780, 2001.

[29] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, page 14. California, USA, 1967.

[30] Sanjoy Dasgupta. *The hardness of k-means clustering.* Department of Computer Science and Engineering, University of California, San Diego, 2008.

[31] Leonard Kaufman and Peter J Rousseeuw. *Finding groups in data: an introduction to cluster analysis*, volume 344. Wiley. com, 2009.

[32] Stephen J Redmond and Conor Heneghan. A method for initialising the k-means clustering algorithm using kd-trees. *Pattern recognition letters*, 28(8):965–973, 2007.

[33] Greg Hamerly and Charles Elkan. Learning the k in k means. *Advances in neural information processing systems*, 16:281, 2004.

[34] Dan Pelleg, Andrew W Moore, et al. X-means: Extending k-means with efficient estimation of the number of clusters. In *ICML*, pages 727–734, 2000.

[35] Siddheswar Ray and Rose H Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pages 137–143, 1999.

[36] Matthias Lange. *Analysis of the uncertainty of wind power predictions*. PhD thesis, Universität Oldenburg, 2003.

[37] Tian Pau Chang. Wind speed and power density analyses based on mixture weibull and maximum entropy distributions. *International Journal of Applied Science and Engineering*, 8(1):39–46, 2010.

[38] Ecmwf web page. `http://www.ecmwf.int/`. Last accessed date: 2013-11-01.

[39] Gfs web page. `http://www.emc.ncep.noaa.gov/index.php?branch=GFS`. Last accessed date: 2013-11-01.

[40] Wrf web page. `http://www.wrf-model.org/index.php`. Last accessed date: 2013-11-01.

[41] A Horányi, I Ihász, and G Radnóti. Arpege/aladin: A numerical weather prediction model for central-europe with the participation of the hungarian meteorological service. *Idojárás*, 100(4):277–301, 1996.

[42] National power quality project web page. `http://www.guckalitesi.gen.tr/en/root/index.php`. Last accessed date: 2013-09-15.

[43] Postgresql web page. `http://www.postgresql.org/`. Last accessed date: 2013-11-15.

[44] Google earth web page. `http://www.google.com/earth/`. Last accessed date: 2013-11-15.

[45] C Ferreira, J Gama, L Matias, A Botterud, and J Wang. A survey on wind power ramp forecasting. Technical report, Argonne National Laboratory (ANL), 2011.

[46] Fast artificial neural network library (fann) web page. http://leenissen.dk/fann/wp/. Last accessed date: 2013-11-15.

[47] Library for support vector machines (libsvm) web page. http://www.csie.ntu.edu.tw/cjlin/libsvm/. Last accessed date: 2013-11-15.

[48] Pierre Baldi and Anthony D Long. A bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, 2001.