MODELING PAN EVAPORATION BY HYBRID WAVELET TRANSFORM SUPPORT VECTOR MACHINES

Thesis

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

LEELADHAR PAMMAR

DEPARTMENT OF APPLIED MECHANICS AND HYDRAULICS

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE - 575025

OCTOBER, 2016

MODELING PAN EVAPORATION BY HYBRID WAVELET TRANSFORM SUPPORT VECTOR MACHINES

Thesis

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

LEELADHAR PAMMAR

(Register No.:121156AM12P02)

DEPARTMENT OF APPLIED MECHANICS AND HYDRAULICS

NATIONAL INSTITUTE OF TECHNOLOGY KARNATAKA

SURATHKAL, MANGALORE - 575025

OCTOBER, 2016

DECLARATION

(by the Ph.D Research Scholar)

I hereby *declare* that the Research Thesis entitled "**Modeling Pan Evaporation by Hybrid Wavelet Transform Support Vector Machines**", which is being submitted to the **National Institute of Technology Karnataka, Surathkal,** in partial fulfillment of the requirements for the award of the Degree of Doctor of Philosophy in **Civil Engineering**, is a *bonafide report of the work carried out by me*. The material contained in this report has not been submitted to any University or Institution for the award of any degree.

LEELADHAR PAMMAR

(Register Number – 121156AM12P02)

(Register Number, Name and Signature of the student)

Department of Applied Mechanics and Hydraulics

Place: NITK, SURATHKAL

Date: OCTOBER, 2016

CERTIFICATE

This is to *certify* that Research Thesis entitled "**Modeling Pan Evaporation by Hybrid Wavelet Transform Support Vector Machines**" submitted by **LEELADHAR PAMMAR** (Register Number: 121156AM12P02) as the record of the research work carried out by him, is *accepted as* Research Thesis *submission* in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy.

Dr. Paresh Chandra Deka
Research Guide
Associate Professor
Dept. of Applied Mechanics and Hydraulics
(Name and Signature with date and seal)

Dr. G. S. Dwarakish
Chairman - DRPC
(Name and Signature with date and seal)

DEDICATED

TO MY PARENTS

ACKNOWLEDGEMENT

It is my pleasure to express profound gratitude and indebtedness towards my research supervisor **Dr. Paresh Chandra Deka**, Associate Professor, Department of Applied Mechanics and Hydraulics for his continued inspiration, motivation, support, discussions and great patience throughout this research work, which made this study possible. It is valuable experience to learn many aspects from him as a good teacher. I admire among his other qualities, kindness and balanced approach towards success and failure, his scientific foresight and excellent knowledge have been crucial to the accomplishment of this work: who managed nicely to spare valuable time for guidance, valuable suggestions and excellent supervision of my research work. I consider myself privileged for having had the opportunity to conduct research in the areas of soft computing techniques under his able supervision.

I am greatly indebted to Research Progress Appraisal Committee (RPAC) members, Dr. S. Shrihari, Department of Civil Engineering and Dr. Raj Mohan B, Department of Chemical Engineering, for their critical evaluation, constructive comments and suggestions during the progress of the work helped me to improve the quality of the work.

I also extend my heartfelt thanks to Dr. G. S. Dwarakish, Head, Department of Applied Mechanics and Hydraulics and Chairman D.R.P.C for his continuous support, encouragement for providing all the necessary departmental facilities during my research period.

I gratefully acknowledge Dr. Lakshman Nandagiri, Department of Applied Mechanics and Hydraulics, for his support and providing access to the pan evaporation data collected from I.M.D Pune during the course of my work. I am thankful to all the faculty members, department of Applied Mechanics and Hydraulics, NITK Surathkal, for timely suggestions and reviewing the work during the progress seminars.

I am grateful to the management of Nitte University, Trustee Sri. Vinay Hegde, Principal Dr. N Niranjan Chiplunkar, Vice Princiapal’s Dr. I. R. Mithantaya and Dr. Shrinivas Rao, Dean Research Dr. Sudesh Bekal and Dr. Srinath Shetty, head of department Civil Engineering for their support and encouragement to carry out research. I am also thankful to all the faculty members and non-teaching staff of Department of Civil Engineering for their care and support during my Ph.D work.

I take this opportunity to express thanks to my research mates Sujay Raghavendra, Karthik, Amit Patil, Surendra H J, Dr. Manjunath K B for their valuable support and rendering my stay in the NITK Campus more than wonderful.

I also acknowledge the help and support provided by non-teaching staff, Sri.Jagadish B, Sri Balakrishna, Sri. Ananda Devadiga, Sri. Gopala krishna, Sri. Padmanabha Acharya, Mr. Harish Salian, Mr. Harish D and Mrs. Pratima Prakash for their support and help during research work.

The Inspiration and support given by the other fellow research Scholars of the Department of Applied Mechanics and Hydraulics have also been much appreciated.

The love, support, patience and encouragement from my father Mr. Shankar Pammar, mother Mrs. Parvati Pammar, brother Mr. Yashavant Pammar, sister Ms. Rohini Pammar and wife Mrs. Shilpa Pammar, prepared me to start my research and submit this work in time. Finally, I would like to thank the Almighty God for blessing me with Good Health, ability to work hard and guiding me to every success in life.

LEELADHAR PAMMAR

Place: NITK Surathkal

Date: 24-10-2016

ABSTRACT

Estimates of evaporation from open water bodies have gained a lot of importance and are crucial for several environment related functional systems, particular field related applications in water resources and ecology. Accurate estimations of evaporations would help these functional systems to overcome the crisis. The acute shortage of fresh water created serious issues in climatic zones like arid and semi-arid places.

The creative and novel approaches are being employed in the estimation of evaporation. In this thesis, it is attempted to address the complex pan evaporation (PE) phenomenon by hybridizing Discrete Wavelet Transform (DWT) and Support Vector Machines (SVM). The superiority of DWT relies on its multi-resolution potential at various scales and SVM capable of establishing a rapid and accurate relationship between input-output patterns. Two stations, namely Bajpe (humid) and Bangalore (Semi- arid) located in the state of Karnataka, India are chosen for model development and to ensure efficiency of developed models.

The model development begins SVM regression with kernel functions, namely Polynomial, Radial basis function (RBF) and Pearson VII function based kernel (PUK). The novel Gamma test (GT) was used to decide the best input output combination. Parameter optimization was carried by Grid search. The developed models showed better estimations of pan evaporation, but exhibited some limitations with non-linearity, sparse and noisy data. These limitations forced data pre-prcessing technique to get introduced via two mother functions, namely Daubechies (DB) and Haar.Fine tuning of various levels and orders were carried out. DB with order 3 and level 4 produced optimum results with SVR models. Overall, this hybrid combination shows promising potential to provide optimal solutions for problems arising out of pan evaporation estimation.

Keywords: Pan evaporation, Support vector machines, Wavelet transform, Kernel functions, Daubechies wavelet, Time series.

CHAPTER 1

INTRODUCTION

1.1 GENERAL

Estimation of evaporation amount is very essential for monitoring, survey and management of water resources. Despite this significance, Evaporation is regarded as one of the least satisfactorily explained element of the hydrological cycle (Brutsaert 1982; Jackson 1985; Sudheer et al. 2002). The complex interactions between the components of the land-plant-atmosphere system makes evaporation estimations quite difficult (Singh and Xu 1997). The water molecules keeps exchanging between the atmosphere and the land on a continuous basis. As per hydrological definition, the evaporation is limited to the net rate of water, which is transferred from the land to the atmosphere. This transformation in the state requires an exchange of approximately 2510 J per each gram of water evaporate (Keskin et al. 2004). The complex phenomenon of evaporation from the open pan as well as surface is influenced by various meteorological parameters such as rainfall, temperature, relative humidity, wind speed and sunshine hours.

It is necessary to understand the various forms of evaporation and distinguish between potential evaporation, actual evaporation and pan evaporation. (Brutsaert and Parlange 1998) gave a clear description between these forms of evaporation. Author described potential evaporation is based on energy availability or in other words the availability of moisture does not limit the potential evaporation. Average evaporation considered from a large region is actual evaporation. Actual evaporation is commonly less than the potential evaporation because moisture availability may be a limiting factor. Pan evaporation is the estimates of evaporation observed from an evaporation pan. Pan evaporation does not depend upon moisture content so it can be considered to be same as an estimate of potential evaporation.

The climate change triggered as a result of global warming, certainly has made a remarkable impact on evaporation and subsequently affecting the water availability. The change in evaporation trends due to changing climatic conditions must be considered seriously to avoid crisis. As it could assist in quantifying, the potential impacts of climate change on evaporation. Studies made on evaporation trends across many climatic zones around the globe have been reported, but conclusions made in the studies vary significantly. Increasing Pan evaporation trends were observed in Bet Dagan, Israel (Cohen et al. 2002) and also in Phoenix, Arizona and northeast Brazil (Silva 2004). However, the decreasing trends in pan evaporation were reported by Chattopadhyay and Hulme (1997) in India, in Australia (Roderick et al. 2007), in the United States (Peterson et al. 1995). Higher evaporation rate creates a more arid environment while the downward trend of evaporation results in a more humid environment. Studies conducted by Ogolo (2011) on decennial trend analysis on pan evaporation for three decades (1970-2002) for four regions reported that, a general downward trend observed in all the regions in the first decade (1970-1979). Followed by an upward trend of PE was observed for the rest decades for all the regions including the average trends in Nigeria. There exist scope for further studies on trend analysis of pan evaporation using advanced approaches which may provide more clear conclusions that helps for efficient planning and managing the available sources of water.

The demand for water is ever increasing and resources are depleted very rapidly. In addition to this, water loss due to evapotranspiration is not directly available. This creates utmost importance in considering the evaporation phase of the hydrological cycle in the study of consumptive use of water. In this perspective, it is very much essential to know the rate and amount of evaporation from water surfaces for assessing the value of natural water bodies. This will help to manage the available resources towards usages such as municipal and industrial water supply, irrigation, condenser cooling water, hydroelectric power, navigation and recreation. A lot of investments in water saving measures are made in water resources management. Managers are finding ways to reduce inefficiencies in water supply systems, including factors due to evaporation of water from reservoirs. Climate change is also expected to result in an increase in evaporation, and ultimately a reduction in the water yield (Adeloye et al. 1999). Various attempts are made in connection to this to reduce evaporation from storages have been developed (Martinez et al. 2006). Therefore, the estimation of reservoir evaporation is a vital and based on that design and ongoing operations related to water supply reservoir can be planned.

The estimation of the water loss by evaporation is very vital for monitoring, survey and management of available water resources, design of irrigation and drainage systems and irrigation scheduling at a farm scale as well as at catchment scale (Martinez et al. 2006; Gundekar et al. 2007). Especially in arid regions with considerably low rainfall, these estimations are very crucial to plan and avail the limited water available for longer periods.

1.2 APPROACHES TO ESTIMATION OF EVAPORATION

Understanding the rate of evaporation of surface water resources is essential for precise management of the water balance. However, the task of evaporation estimation is difficult to measure experimentally over large water surfaces; several techniques and models have been suggested and used in the past for its determination. There are two ways to estimate evaporation from the free water surface. One based on meteorological parameters and the other is direct field measurements (Zhang et al. 2004; Fu et al. 2009).

One of the simplest approaches to measure the water evaporation in the field is with the help of evaporation pan, which indicates the combined measurement effect of complex meteorological interactions such as temperature, solar radiation, humidity, and wind speed. Although pan evaporation (PE) may not fully represent lake evaporation, it has been found as proportional to actual evaporation on moist surfaces, such as lakes or irrigated fields (Kahler and Brutsaert 2006). With the popularity of pan evaporation several studies are made with pan evaporation using various designs of pans across many climatic zones throughout the world. Motivated by satisfactory results shown by pan evaporations, several attempts have been made to use these data to estimate actual evaporation even in non-moist environments. Linacre (1994) conducted a study on evaporation trends across the space and over different decades. These attempts need to be refined to a finer degree, if these estimations are to be used while projecting and exploiting the open water sources. This is very crucial depending upon the climatic zones because of non-uniform water availability.

Indirect methods of estimating evaporation consist of water budget methods, mass transfer, and energy budget. Many researchers have attempted to estimate the evaporation using the indirect methods based on climatic variables, but these methods require a large amount of data which is not easily accessible (Burman 1976; Rosenberry et al. 2007).

Due to the nonlinearity associated with the evaporation process, it is difficult to provide accurate estimations of evaporation based on the conventional empirical approaches. Because of that, some researchers emphasize the estimation of accurate evaporation in the research field using modeling techniques (Kim et al. 2012; Nourani and Fard 2012; Jothiprakash and Kote 2011; Singh and Xu 1997; Bruton et al. 2000; Kumar and Tiwari 2012). For efficient use of data available for estimation of evaporation, data modeling becomes essential. The modeling of estimating evaporation from surface reservoirs is a very active field of study and continuous improvement thrust may be the current focus of study. If model predictions are to be used for regulatory purposes, research or design, then the modeling effort should be scientifically sound, robust, and defensible (Engel et al. 2007).

Several data driven modeling approaches have been utilized for estimation of evaporation which is discussed in the following chapter literature review. Few of the modeling evaporation include single soft computing approaches (Keskin et al. 2004; Moghaddamnia et al. 2008; Deswal and Pal 2008; Keskin et al. 2009) and approaches consisting of two or more techniques (Eslamian et al. 2008; Kasiviswanathan et al. 2009; Sanikhani and Kisi 2012). Several drawbacks of the neural networks and ANFIS were identified, specifically the complexity of their implementation, risk of over-fitting, and degraded performance with sparse data, have favored the use of Support vector machine (SVM) in a variety of applications. The study conducted by Deswal and Pal (2012) suggests the usefulness of SVM’s algorithm technique in modeling the pan evaporation from reservoirs and reported satisfactory results. SVM's are capable of producing accurate and robust classification results, even when input data are non-monotone and non-linearly separable. So they can help to evaluate more relevant information in a convenient way. Input vectors of SVM are quite flexible; hence various other influential factors (such as temperature, relative humidity, and wind speed) can be easily incorporated into the model (Moghaddamnia et al. 2009). This creates opportunity for employing SVM in modeling complex and nonlinear hydrology feature to obtain better results.

In modeling hydrologic time series, sometimes signals are highly non-linear and exhibit seasonal irregularity. Under such circumstances, SVM alone may not be able to cope with non-linear data, if preprocessing of input and output data is not performed. In this context, wavelet transform may be utilized for data preprocessing. The wavelet transform is a strong mathematical signal processing tool with the ability of analyzing nonlinear and non-stationary data. It can produce both time and frequency information with a higher resolution. This provides an opportunity for better potential techniques such as SVM's to blend with wavelet transform to match up with the growing demands.

The recent literature shows the application of hybrid models towards the estimation of hydrological features such as evapo-transpiration (Kaheil et al. 2008), temperature (Liu et al. 2012), precipitation (Kisi and Cimen 2012). This provides an opportunity for employing proven techniques such as SVM's to blend with the other data pre-processing techniques such as wavelet transform to match up with the growing demands.

1.3 PROBLEM DEFINITION

The process of evaporation is very much complex and non-linear in nature with respect to the meteorological parameter which influences the evaporation. The estimation of evaporation is very essential for various field related activities. The applied methodologies over evaporation estimation have resulted in varied conclusions. Evaporation pans are normally used as one of the direct methods to estimate evaporation.

However, there are many errors associated with measurement of pan evaporation such as debris in water, animal activities around the pan, exposure of the pan and measurement of water depth in the pan. In spite of these factors, pan evaporation estimations yielded better performance in comparison to other conventional methods. The works carried out on pan evaporation have shown varied trends with different climatic conditions. Data modeling techniques found superior to conventional approaches applied on pan evaporation. In data modeling approach, several techniques have been adopted to improvise the level of accuracy.

Recently a novel machine learning technique, SVM has received increasing attention due to their remarkable generalization performance. It is based on the structured risk minimization principle, which seeks to minimize an upper bound of the generalization error rather than the empirical error commonly implemented in neural networks. However, these regression techniques are although powerful, they suffer with the nonlinear and non-stationary data. Data pre-processing techniques such as wavelet transform can minimize noise and unwanted information so that regression technique can understand pattern in a better way. This combination of hybridization may prove to be successful in evaporation estimation.

In this perspective, this study attempts to investigate the performance of hybrid model formed with wavelet transform - support vector machine for modeling daily pan evaporation developed for the data recorded at two different climatic stations representing humid and semi-arid conditions. The hybrid model performance is compared with the performance of conventional support vector machine (SVM) models. This provides an opportunity to judge the model efficiency to provide accurate estimations of pan evaporation under varied climatic conditions, which may assist further research in the estimation of pan evaporation.

1.4 ORGANIZATION OF THE THESIS

This thesis comprises of five chapters, the overview of each chapter is presented below.

Chapter 1 Introduction: In this chapter the relevant information regarding current scenario of evaporation and its estimation approaches is provided. The rising demands of accurate estimation and appropriate approaches employed are highlighted in the order of their occurrence. The overview of research scope of the problem identified in the domain of evaporation and the suitability of the approach employed to address it is presented.

Chapter 2 Literature Review: This chapter represents the research carried out in the field of evaporation estimation. The accuracy of estimation varies among the different approaches attempted. The summary highlights the major works carried out on estimation of evaporation. The research objectives are set based upon the research gap in the works conducted in the past and novelty.

Chapter 3 Study area and Methodology: In this chapter, the study area chosen are presented along with the model development methodology adopted. The statistical analysis of data collected, the attributes of pan evaporation chosen for the study are discussed in the initial part of this chapter. The next part discusses the basics of the techniques adopted in the research, i.e support vector machine (SVM) and discrete wavelet transform (DWT) along with the flow chart of model development. The importance of suitability of kernel functions and mother wavelet functions to improvise the accuracy of pan evaporation estimation are described. The usefulness of advanced tools such as Gamma test Grid search is explained. Statistical indices used for the performance evaluation of developed models are presented.

Chapter 4 Results and Discussion: This is the core chapter of the thesis, which discusses the analysis and interpretation carried out on the performance of developed model results via individual SVM and hybrid DWT-SVR adopted in the current research. The interpretation reveals the usefulness of techniques employed in model development and ranks them according to their superiority of performance.

Chapter 5 Summary and Conclusion: This chapter represents the summary of research work carried out, research findings, and conclusions drawn. The chapter ends with suggested directions for future work, and limitations of the present work.

CHAPTER 2

REVIEW OF LITERATURE

2.1 INTRODUCTION

With the rising demands of saving water resources, evaporation estimations have caught the attention of researchers for many decades. The accurate estimates of evaporation are very useful to retain the water bodies for a long period. Researchers are finding a reliable method of estimation of pan evaporation. It is also important because of its key role in the management of available water resources in the varied climatic regions. Accuracy enhancement in evaporation estimation from open water bodies is rapidly increasing for the past decades. These estimations would assist many field related activities in hydrology, irrigation management, water resources applications. Vast research has been carried out in the domain of evaporation with various backgrounds as per their requirements. As usual thirst of researchers in this evaporation field is not fulfilled and studies are being carried to understand it even better than so far.

2.2 LITERATURE REVIEW

In the following sections, the research carried out in the field of evaporation and associated elements of hydrological cycle based on various approaches using conventional approaches, soft computing methods and advanced modelling using SVM and wavelet transform are highlighted.

2.2.1 Using conventional methods

Various approaches are attempted in the estimation of evaporation from an open water body, also known as lake evaporation. They are basically categorized under two heads, direct and indirect methods. These methods are developed on the empirical computations based on various attributes influencing evaporation. Some of important methods include water budget method, energy budget method, mass-transfer approach, the Penman method, eddy correlation method, combination equation and the pan coefficient method (Dingman 1994). Due to various complexity involved in estimations of evaporation through empirical equations developed on climatic factors, pan evaporation found its usefulness in the field. The difficulty experienced in the indirect methods is requirement of data, which are not easily available, particularly in developing countries (Burt et al. 2005). Pan evaporation received much of the attention of researchers due to the fact that, it combines the accumulated effects of all the climatic parameters. However, their study on pan estimations could not account for the combined effect of all the meteorological parameters on evaporation loss (Finch and Calver 2008). Similar studies on pan estimation reported the same (Singh et al. 1981; Senapati et al. 1985; Chandra et al. 1988; Bhakar and Singh 2004; and Kadhane and Tatewar 2006). The studies were reported on indirect methods of estimation of evaporation. In the increasing order of complexity and data availability, include temperature-based equations (Thornthwaite 1948); radiation-based approximations (Smith et al. 1998); humidity-based empirical equations (Romanenko 1961); combined formulae, which include humidity and wind speed (Penman 1948); or more intensive computations of an energy balance at the evaporation surface (McKenzie and Craig 2001).

Burt et al. (2005) conducted a review on evaporation and discussed the measurement techniques along with estimates of evaporation over different climatic conditions. They identified the major factors that influence the evaporation and quantified their significance. The authors faced challenges in obtaining evaporation data that included contrasting climatic conditions, initial moisture, and soil type. They suggested considering quality control concerns while carrying out research in the domain. They also indicated that the field instruments like Lysimeter is quite sensitive to site specific conditions. The authors also found that, evaporation is not given much of importance in the discussion of evapotranspiration and suggested the researchers consider evaporation seriously to avoid issues that may arise in future.

Lowe et al. (2009), demonstrated to quantify the uncertainties associated with the pan coefficient method in estimates of reservoir evaporation. The authors suggested locating the stations to place pans to be very close to reservoirs, which subsequently enhance the estimates of evaporation. They also recommended employing sophisticated geostatistical techniques to estimate the pan evaporation at the reservoir location. These precise methods help to overcome the uncertainty associated with spatial transposition factors and generate a near closer pan coefficient.

Beside this, trends in evaporation have been analyzed by many researchers for many regions, resulted in varied conclusions. Pan evaporation with increasing have been reported in Israel (Cohen et al. 2002) and also in northeast Brazil (Silva 2004). In contradictory to these reports the decreasing trends in pan evaporation were documented by Chattopadhyay and Hulme (1997) in India; Peterson et al. (1995) in the United States, the former Soviet Union, Europe and Siberia; Roderick and Farquhar (2004) in Australia; Li et al. (2013) in China, and by Tebakari et al. (2005) in Thailand. Increasing trend was also observed with the estimations of long-term series of potential evaporation data for Oxford, UK indicates an increasing trend (Burt and Shahgedanova 1998).

Ogolo (2011), carried out on the trend of pan evaporation in 4 different climatic regions covering about 21 tropical stations in Nigeria. The influence of the change in some meteorological variables on the observed trend in PE was also investigated; an upward trend of PE was established for all the regions. Decennial trend analysis for three decades (1970 to 2002) for all the regions was carried out. There was a general downward trend observed in all the regions in the first decade (1970-1979). This occurrence was coincidental with the widely reported global solar dimming. However, an upward trend of PE was observed for the rest two respective decades for all the regions including the average trends in Nigeria.

2.2.2 Using soft computing techniques and data modeling

The drawbacks associated with the conventional methods provide scope for new approaches. Attempts were then made in the field of pan evaporation using soft computing and data modeling techniques. Keskin et al. (2004) identified the difficulties associated with a pan used for the direct measurements. The authors explored fuzzy logic models as alternative approaches to conventional evaporation estimations.

Considering the potential of NF models to estimate pan evaporation, the work was extended to model evapotranspiration using ANFIS. Results showed ANFIS is capable to produce better estimation against previous conventional methods (Kisi and Ozturk 2007).

Deswal and Pal (2008), carried out study on estimation of evaporation losses over a reservoir using the artificial neural network modeling approach. They found the usefulness of neural network based modeling technique over the simple linear regression and multiple linear regression approach in accurate prediction of the evaporation. The paper concludes that, all the parameter combination is the vital reason behind the loss of reservoir water due to evaporation.

Many researchers used ANN based modeling technique on evaporation (Jain et al. 2008; Shiri et al. 2011; Shirsath and Singh 2009; Goyal et al. 2014; Shirgure 2013; Malik and Kumar 2015; Lowe et al. 2009; Finch and Calver 2008; Kişi 2013). The researchers found that, the model accuracy can be fine-tuned with the different input parameters used while developing models. The prediction accuracy of ANN was superior to linear regression. A similar approach was attempted by Rahimikhoob (2009), also upheld ANN over empirical approaches.

Piri et al. (2009), evaluated the potential of ANN to estimate evaporation in a hot and dry region. They found ANN can model the evaporation even under such contrasting regions. Although results were site specific, the integrated ANN and autoregressive with exogenous inputs enhances the performance over the traditional ANN. Gamma test was also tried in determining the input to output combination.

Shirsath and Singh (2010), estimated daily pan evaporation using ANN and multiple linear regression (MLR) models. They also compared the results with Penman, Priestley-Taylor and Stephens and Stewart models. The comparison made on the developed models indicated better agreement between the ANN estimations and measurements of daily pan evaporation than other models.

Kisi (2006), improvised the soft computing model approach by adopting Neuro-Fuzzy (NF) to model daily pan evaporation using meteorological variables. The study demonstrated that NF technique is capable of modeling daily evaporation. They made a comparison between Neuro fuzzy, ANN and Stephens–Stewart (SS) methods. Subsequently found NF and ANN both working well.

Keskin et al. (2009), found the limitation associated with the probabilistic, statistical, and stochastic approaches is the large quantity of data availability for the modeling purposes. They suggested these approaches are not practically suitable for local evaporation studies. The results revealed ANFIS approach is superior to fuzzy sets in modeling the evaporation process.

Sivapragasam et al. (2009), carried out work on evaporation-seepage losses using Genetic Programming (GP) modeling approach for Reservoir Water Balance in semi-arid Regions along with the Penman equation. They found similar performance among GP and Penman equation. They also concluded that GP is superior when any other losses are included in the estimation.

A similar approach was made by Kasiviswanathan et al. (2009) on Estimation of monthly evaporation using GP and Thornthwaite method. They concluded that, GP is having good potential to address the non-linearity associated with the data by considering the prominent and important parameters which influence the evaporation process the most. Thornthwaite model found to underestimate the evaporation may be due to the empirical nature.

Sanikhani and Kisi (2012), investigates the ability of two different Adaptive Neuro-Fuzzy Inference Systems (ANFIS) including grid partitioning and subtractive clustering (SC), in modeling daily pan evaporation (Epan). Comparison of results indicates that both ANFIS-GP and ANFIS-SC are superior to the MNLR (Multi Non-Linear Regression), ANN, Stephens-Stewart (SS) and Penman methods in modeling Epan.

Guven and Kisi (2011), conducted a study on daily pan evaporation modeling using linear genetic programming (LGP), which is an extension of genetic programming (GP) technique. The LGP estimates were compared with those of the Gene-expression programming (GEP), which is another section of GP, multilayer perceptrons (MLP), radial basis neural networks (RBNN), generalized regression neural networks (GRNN) and Stephens–Stewart (SS) models. Comparisons revealed that LGP model performance is superior to rest other models developed in the study from the available climatic data.

Kasiviswanathan et al. (2011), explores the utility of an evolutionary based data driven modeling approach, Genetic Programming (GP) to model the evaporation process. A monthly averaged climatic data of temperature and rainfall was used. The performance of the GP model is compared with Thornthwaite method, and results from the study indicate that the GP performed better than the Thornthwaite method.

Malik and Kumar (2015), conducted a study on simulating daily pan evaporation using ANN, CANFIS and MLR models. It was observed with generated results that, ANN model with six input variables working well in comparison to the CANFIS and MLR models.

Within the last few years, drawbacks experienced in modeling evaporation found their solutions with utilization of support vector machines in various domains including hydrology and civil engineering (Dibike et al. 2001; Pal and Goel 2006; Pal 2006). With the superior generalization capability of kernel estimations SVM is working well in comparison to neural network approach.

Moghaddamnia et al. (2008), worked on evaporation estimations, using SVM's technique to produce the accurate estimation of evaporation in the Chahnimeh reservoirs of Zabol in the southeast of Iran. The Gamma Test (GT) was introduced in this paper for the first time in modeling one of the key hydrological components: evaporation estimation modeling.

Kim et al. (2012), conducted a study to develop and apply the neural network models to estimate daily pan evaporation (PE) for different climatic zones such as temperate and arid climatic zones, Republic of Korea and Iran. Three kinds of the neural network models, namely multilayer perceptron-neural networks model (MLPNNM), generalized regression neural networks model (GRNNM), and support vector machine-neural networks model (SVM-NNM), were used to estimate daily PE. The paper concludes that superior performance was found with SVM-NNM model and reported that all of the available climatic data are needed for estimating daily PE for different climatic zones, such as temperate and arid climate studied at Republic of Korea and Iran.

Eslamian et al. (2008), attempted to estimate the monthly pan evaporation using ANN and SVM. They found that SVM produced better accuracy and fast computation over ANN.

Deswal and Pal (2008), carried out a study on pan evaporation estimations using Support Vector Machines (SVM's) based approach. The influence of different input combinations was investigated with the performance of SVM's compared with multiple linear regression. The results indicated superiority of SVM in providing accurate estimations of pan evaporation.

In the last few years, drawbacks experienced in modeling hydrological features such as evaporation, evapotranspiration, rainfall using various Artificial Intelligence (AI) were better addressed by SVM's (Kumar et al. 2007; Jain et al. 2008; Tabari et al. 2012). It is found to be working well in comparison to neural networks and other similar techniques.

The pre-processing of raw data in terms of analyzing variations, periodicities, trends in time series has received much attention of researchers in recent years in various fields (Smith et al. 1998; Chou and Wang 2003). The processed data eliminates the unwanted information and noise from the signals so that regression modeling techniques can understand pattern in a better way. Studies also showed that, decomposition of inter-decadal and inter-annual components of rainfall data were made possible by wavelet transform (Lu 2002) and scaling delineation of Precipitation and Evaporation (Ping and Yu 2014). DWT was also tried for decomposition of unit hydrograph (Chou and Wang 2002).

There were few literatures regarding hybridization of wavelet transform and other data driven techniques in evaporation modeling (Abghari et al. 2012). The wavelet method is very robust due to the fact that it does not possess any potentially erroneous assumption or parametric testing method. Another advantage of the wavelet method is that wavelet variance decomposition allows users to study different investing pattern in different time scales independently(Wang and Ding 2003).

In wavelet analysis the selection of the mother wavelet function as well as the decomposition level of the signal plays a very crucial role. Discrete wavelet transform (DWT) are orthogonal wavelets which analyze the nonlinear as well as non-stationary signal in different time signals. Among the families of DWT, Daubechies (DB) wavelets have been widely implemented (Rafie et al. 2009). Another main issue in wavelet analysis is the order of the mother wavelet function, which was selected previously by trial-and-error methods based on intrinsic characteristics of the data in several papers (Tse et al. 2004; Samanta and Balushi 2003; Kar and Mohanty 2006; Saravanan et al. 2008).

Abghari et al. (2012), evaluated different types of mother wavelet functions for finding the performance standards of models in daily pan evaporation prediction in the Lar synoptic station. Comparison of WNN and MLP modeled results indicated that, Mexican hat mother wavelet is accurate in the daily pan evaporation modeling.

McMahon et al. (2013), considered summary of techniques of various approaches to estimate both actual and potential evaporation, reference crop evapotranspiration and pan evaporation. They suggested researchers to use standard meteorological data averaged (or estimated as an average) over a 24 hour's day rather than considering only daylight hours used in the analysis.

Modeling nonlinear signals with various trends, irregularities and noisy data was experienced some limitations with the adoptability of only single techniques. Researchers found hybrid models working better in comparison to single technique model. Several hybridization was tried in the domain of hydrology and water resources (Silva 2014; Espinoza et al. 2005; Shiri and Kisi 2010; Chen et al. 2006).

2.3 SUMMARY OF LITERATURE REVIEW AND RESEARCH OBJECTIVES

The accurate and reliable estimates are very important for effective utilization of water bodies. The literature reports a wide range of methodologies and techniques being adopted across the world to best suit climatic conditions. The thrust is to provide accurate estimates of evaporation to assist the associated field requirements.

Literature has witnessed various approaches attempted in this perspective ranging from conventional methods to advanced data modeling. As discussed in the previous section, phenomenon of evaporation is quite complex and nonlinear due to several interacting climatological factors, such as temperature, humidity, winds speed, bright sunshine hours, rainfall etc. There are a large number of studies in which some hydrological processes are simulated by nonlinear models based on Artificial Neural Networks, Support Vector Machines, Fuzzy Logical system, Local Linear Regression, Multiple linear regression (MLR) and so on. Keskin et al. (2004) examined the potential of the fuzzy logic approach in estimation of daily pan evaporation. In the recent past researchers emphasized on the use of Artificial neural networks (ANN) as they found to be effective tools to model nonlinear systems (Kumar et al. 2002, Sudheer et al. 2003; Shirgure and Rajput 2011). A neural network model is architecture is essentially analogous to the human brain. However, ANN, due to its "black box" nature, immense computational burden, prone to overfitting, and the empirical nature of model development somehow paved way for hydrological modeling using Support Vector Machines (Sujay and Deka 2015).

Based on literature, the SVMs showed their superiority over other data driven modeling approaches. SVMs have shown better generalization ability, the architectures of the SVMs are guaranteed to be unique, and they are trained much more rapidly than other techniques used in the field of hydrological features modeling especially evaporation estimations (Cherkassky and Ma 2004; Han et al. 2007; Raghavendra and Deka, 2014; Reddy and Nair, 2013; Tabari et al. 2012; Yu et al. 2004) .

Data modeling approaches need sufficient and reliable records of data to provide closer estimates to observe. Among the data modeling methods, SVM's showed good potential to understand the input and output pattern and provide accurate and robust classification results even handling non-monotonous and non-linearly separable data. The algorithms used in SVM based kernel transformations linearize data on an implicit basis, so that the accuracy of results does not rely more on the operational expertise of the users, fine tuning for the optimal choice of the linearization function of non-linear input data. SVM's local linear approximation can offer an important support for recognizing the mechanisms linking different financial ratios with the final score of a company. For all of these reasons SVM's showed its superiority among its competitive modeling techniques which can effectively complement the information gained from classical linear classification techniques.

As seen in the literature, the wavelet technique shows its superiority in dealing with non-linear signal approximation and classification. The mother wavelet functions plays key role in mapping the original signal better way and analyze the trends, seasonal irregularities and noise.

It is very vital and valuable to make a distinction between the performance of regression technique such as SVM and hybrid combination including DWT and SVM. The developed hybrid model is expected to produce accurate and reliable predictions of pan evaporation, and overcome the difficulties faced in the data modeling approaches carried. The analysis and interpretation of results may assist concerned departments, including water distribution authority, to form the strategies to resolve issues related to water management.

Based on the literature and scope of research in modeling pan evaporation following objectives are formulated for proposed research.

1. Development of various models for estimation of pan evaporation using wavelet and SVM hybridization in two contrasting climatic regions.
2. Parameter optimization such as various wavelet functions and SVM kernel functions.
3. Performance evaluation of various models and selection of best model for specific site conditions.
4. Compare the hybrid model performance with single SVM model for different input scenarios such as time series and causable variables.

CHAPTER 3

DATA COLLECTION AND METHODOLOGY

This Chapter includes two parts A) Data collection and B) Methodology and model development. In the first part the selection of suitable data stations, the climatic conditions, weather variables identification and location are discussed. The second part discusses the basics and adaptability of techniques employed, i.e conventional Support Vector Regression (SVR) model and a hybrid model of discrete wavelet transform support vector regression is discussed.

3.1 DATA COLLECTION

For this research, two climatically contrasting stations are selected to determine the efficiency of developed models in providing an accurate estimation of evaporation. The meteorological stations selected in the study area Bajpe representing humid climatic condition and Bangalore represent semi-arid condition as per Thronthwaite's classification (Ramachandra et al. 2004); located in Karnataka state of India. The daily recorded weather attributes used in the model building are mean air temperature (T), wind speed (W), rainfall (P), mean relative humidity (Rh), sunshine hours (Sh) and pan evaporation (E). Study stations are shown in **Figure 3.1**.

3.1.1 Study area: Bajpe

The Bajpe meteorological station is located close to Arabian sea in Dakshina Kannada district. The geographic coordinates of this place are 12° 57´ N and 74° 53´ E. The average rainfall is about 3600 mm with an altitude of 103 m above mean sea level. It is around 18 km from the heart of the port city of Mangalore. The weather attributes used in this study consist of seven years daily data recorded during the period of 2000-2006. The first five years (2000–2004) data were used as the training data set and the remaining data i.e (2005-06) were used as testing data set.

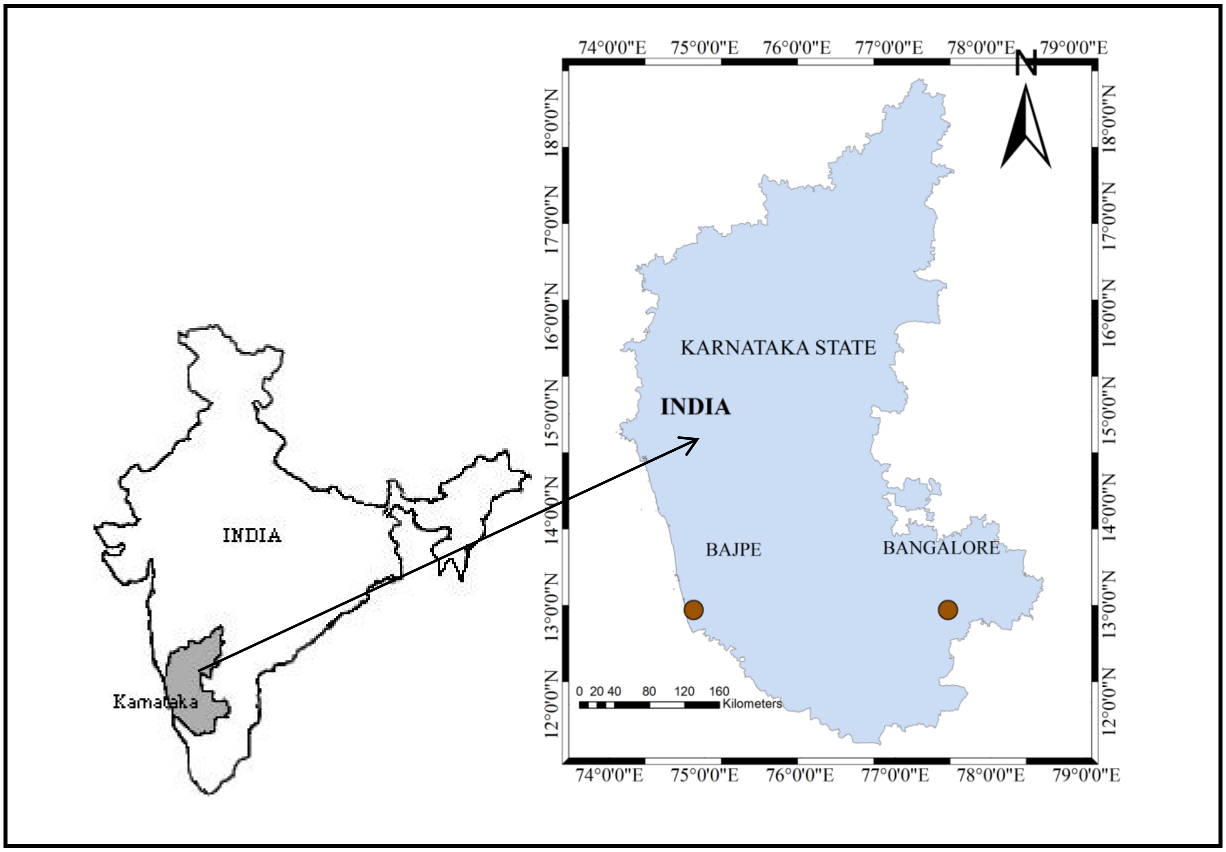


Figure 3.1 Meteorological stations selected in the study area.

3.1.2 Study area: Bangalore

The meteorological station in Bangalore is situated 350 Kms away from the Arabian sea with the geographic coordinates of approximately 13° 39' N and 77° 22' E. The altitude is about 900 m above mean sea level with average annual rainfall 860 mm. Because of its elevation, Bangalore station experiences a pleasant and moderate climate throughout the year. A total of ten years daily data attributes recorded in the period 1975-1984 are used in the study. The first seven year data (1975-1981) were used as the training data set and the remaining data i.e (1982-1984) were used as testing data set. **Table 3.1** shows the statistical analysis of daily recorded data on the station’s chosen.

Table 3.1: Statistical analysis of the daily recorded weather data

Station's With Period	Attribute	Statistical Parameters					
		X_{mean}	X_{max}	X_{min}	S_d	C_v	Correlation with Pan evaporation
Bajpe (2000-2006)	Temperature (°c)	22.14	32.75	11.50	1.63	0.07	0.62
	Wind speed (m/s)	5.10	8.00	4.00	0.52	0.10	-0.13
	Rainfall (mm)	9.10	291.00	0.00	22.00	2.00	-0.54
	Relative Humidity (%)	78.00	100.00	35.00	11.34	0.14	-0.65
	Sunshine Hour's (No's)	5.20	11.90	2.00	3.60	0.64	0.40
	Pan evaporation (mm)	3.90	12.40	0.00	1.78	0.46	1.00
Bangalore (1975-1984)	Temperature (°c)	24.09	31.30	13.55	2.42	0.10	0.58
	Wind speed (m/s)	6.00	9.00	1.00	1.47	0.25	-0.08
	Rainfall (mm)	3.00	136.00	0.00	8.00	3.00	-0.20
	Relative Humidity (%)	66.00	100.00	17.00	15.35	0.23	-0.66
	Sunshine Hour's (No's)	5.30	11.50	0.00	4.01	0.76	0.40
	Pan evaporation (mm)	4.20	10.20	0.00	1.84	0.44	1.00

The **Table 3.1** indicates that, standard deviation and coefficient of variation are higher for Bangalore station data compared to Bajpe station, considering almost all the attributes.

3.2 METHODOLOGY

3.2.1 Support vector machine study review

The Support Vector Machine (SVM) is developed on the basis of statistical learning theory. It is considered to be an approximation implementation of the method of structural risk minimization with a good generalization capability. Advanced algorithm of SVM has been proven to be robust and efficient for classification (Vapnik 1995), regression (Vapnik 1995; Kaheil et al. 2008), forecasting and prediction. The standard SVM algorithm being used currently was proposed by Cortes and Vapnik(Cortes and Vapnik 1995). The beauty of SVM approach is two-fold. The algorithm is regarded as simple to understand, yet it is powerful that the predictive accuracy of this approach overwhelms many other methods, such as Neural Networks, nearest neighbor, and also Decision Tree. The modeling techniques like SVM's have shown their potential to reproduce the unknown relationship exist between a set of input and the output variables of the system. SVM has gained the popularity over other modeling techniques because of their great advantage of minimizing both model complexity and prediction error simultaneously.

SVM's showed their superiority in producing robust and accurate classification outputs on a sound theoretical basis, even while handling non-monotone and non-linearly separable data (Vapnik 1995). The computation rate is also one of the highlighting factors of SVM, they evaluate more relevant information in a convenient and faster way. The speciality of SVM is that the accuracy developed in the model results does not totally depend on the quality of human expertise, judgment for the optimal choice of the linearization function of non-linear input data, since they linearize data on an implicit basis by means of kernel transformation(Vapnik 1995). All these advantages made SVM's as a useful tool for effectively complementing the information gained from classical linear techniques. The general structure of support vector machines is displayed in **Figure 3.2**.

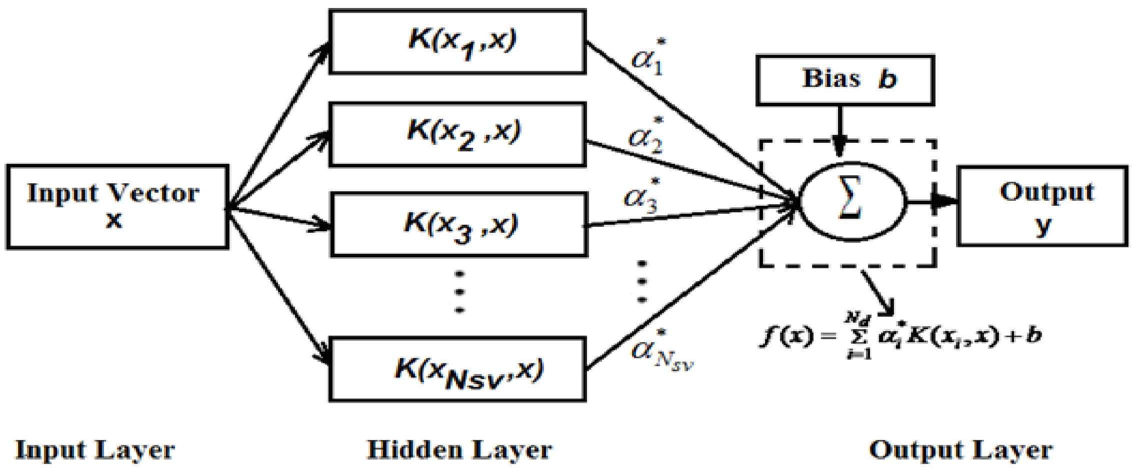
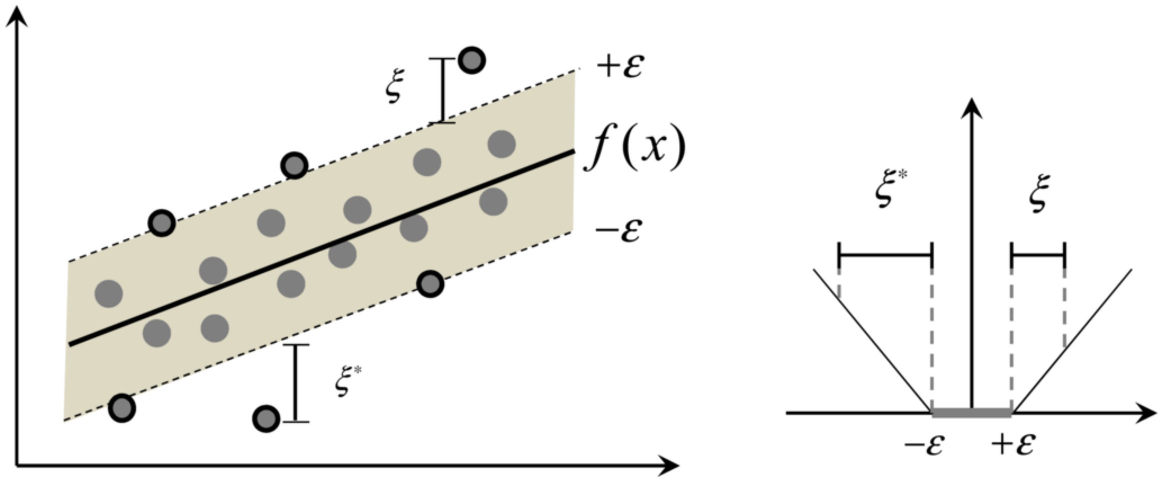


Figure 3.2 General structure of support vector machines

(Tezel and Buyukyildiz 2015)

3.2.2 The architecture of SVM for regression

The neural networks have utilized quadratic loss function i.e. in multilayer perceptrons and radial-basis function networks due to its computational convenience. However, it is found that they are quite sensitive to the presence of the outliers. Neural networks perform poorly when the underlying distribution of the additive noise has a long tail. The advanced SVM's have a solution for this, since they adopt an ε-insensitive loss function as shown in **Figure 3.3**. This makes the generated models to be more robust, i.e. insensitive to small changes.



(a) Data and best fit function in feature space (b) Penalty function

Figure 3.3: ε-insensitive loss function

3.2.3 Selection of kernel

The flexibility of the SVM is provided by the use of kernel functions that implicitly map the data to a higher dimensional feature space. A linear solution in the higher dimensional feature space corresponds to a non-linear solution in the original, lower dimensional input space. This makes SVM a feasible choice for solving a variety of problems in hydrology, which are non-linear in nature (Maity et al. 2010).

The choice of kernel function depends mainly upon the data type i.e which kind of underlying relation needs to be estimate to relate the input data with the desired output property. Because the nature of data is usually unknown, the best mapping function must be determined experimentally which can best fit the most data types. The Kernel function selection in model development is carried out on the basis of optimizing methods such as Cross Validation (CV) or grid search method. In literature trial and error basis is considered the most to select the type of kernel function and tuning of associated hyper parameters (Chappelle et al. 2002; Duan et al. 2003, Kwok 2000). On the basis of trial and error, best kernel function can be selected using above mentioned methods i.e cross validation or grid search. However, these methods are computationally expensive when dealing with a large number of kernel types or parameters. Even the kernel selected by these optimization methods also cannot be guaranteed optimal in some cases.

The selection of a proper kernel function plays vital role in SVM based classification/regression problems. A number of kernels are discussed in the literature (Vapnik 1995), but it has arrived at the decision to choose one which gives the best generalization for a given data set. The literature shows a good number of kernel functions being employed and showed their capabilities to perform, among them recent one proposed is universal Pearson VII function based kernel (PUK) to solve SVM based regression problems (Pal 2006); (Ustun et al. 2006). The paper also suggested that this kernel can be employed as an alternative to the linear, polynomial and radial basis function kernels (Pal 2006).

Polynomial kernels are very much advantageous for increasing the dimensionality as the order of the polynomial defines the smoothness of the function. The polynomial kernel has more hyper parameters than the RBF kernel. As the degree of the polynomial increases, the classification surface becomes more complex.

The **RBF Kernel** is capable to produce efficient interpolation; however it may have some weak points to give longer-range extrapolation.

The **PUK kernel** has the possibility to change easily, by adapting its two parameters, from a Gaussian into a Lorentzian peak shape and more.

The mathematical expressions of three types of SVR based kernel functions employed in this work are displayed in the following equations.

Expression for Polynomial kernel - $K(X,Y) = \exp(X.Y + k)^{d}$ (3.1)

Expression for RBF - $K(X,Y) = \exp(-\gamma \|X - Y\|^{2})$ (3.2)

Expression for PUK - $K(X,Y) = 1/[1 + (2\sqrt{\|X - Y\|}^{2} \sqrt{2^{(1/\omega)}} - 1/\sigma)^{2}]^{\omega}$ (3.3)

Where K-Kernel function, X is input, Y is output, d is exponential degree, γ-Width, σ and ω are known as half width or Pearson width.

3.2.4 Issues related to model parameters

Though SVM has various advantages as listed above, the parameter calibration remains an open issue. The parameters involved and selected by the user are:

1. Parameter C controls the trade-off between the training error and the model complexity. Since SVM map data into high dimensional feature space, C is sensitive to the model performance. Only a good choice of C can provide a good result.

2. Another parameter is ε from the ε-insensitive loss function. ε can be related to the noise of the training data. However the noise of the real world data is usually unknown.

3. Another parameter is σ, the width of the Gaussian kernel. It controls the complexity of the model. The smaller σ is, the more powerful SVM can approximate. The dimension of the feature space of Gaussian kernel is infinitely large. The results of SVM are implicitly provided from the feature space by using the kernel method. These three major free parameters need to be calibrated before SVM can be utilized to its fullest. These parameters must be tuned simultaneously. It is a quite difficult problem of regression and there is, however, no good and efficient method available. It is reported on SVM applications that tuning these parameters is largely a trial and error process(Asefa et al. 2006; Muller et al. 1997; Dibike et al. 2001; Liong and Sivapragasam 2002).

3.2.5 Wavelet transform

A wavelet transform is a strong mathematical signal processing tool, which has the ability to analyze both nonlinear as well as non-stationary data. At the same time it can produce both time and frequency information with a higher resolution, which is considered as the limitation of previous versions of wavelet family such as Fourier transformation (FT). Wavelet theory is discussed thoroughly in Labat et al. (2000) and Mallat (1989).

As a pre-processing tool wavelet transforms provide useful decompositions of the original time series, so that data that has been pre-processed improves the ability of a forecasting model by capturing information on various resolution levels (Adamowski and

Jan 2008). In addition, it has also been found that pre-processing data with wavelet transforms can lead to models that better represent the true features of the underlying system by eliminating noise(Adamowski 2008)

Signals whose frequency content does not change with time are called stationary signals. In other words, the frequency content of stationary signals is not changed in time. In stationary signals it is not necessary to know at what time, frequency components exist, since all frequency components exist at all times.

The wavelet representation addresses the limitation, by adaptivelypartitioning the time-frequency plane, using a range of window sizes. At high frequencies, the wavelet transform gives up some frequency resolution compared to the Fourier transform. **Figure 3.4** shows a representation of the effect of using FT and WT. WT provides multi resolution analysis, i.e. at low scales (high frequency) it gives a better time resolution (represented by compact width of time window as shown in **Figure 3.4**) and poor frequency resolution (represented by the wider width of the scale window as shown in **Figure 3.4**) and at high scales (low frequency) it gives a better frequency resolution and poor time resolution and in actual practice for all the time series signals such information is important. The lower scales (i.e. Compressed wavelet) trace the abrupt change or high frequency of a signal and the higher scales (i.e. Stretched wavelet) trace slowly progressing occurrences or low-frequency component of the signal. The wavelet transform breaks the signal into its wavelets (small wave) which are scaled and shifted versions of the original wavelet so called mother wavelet.

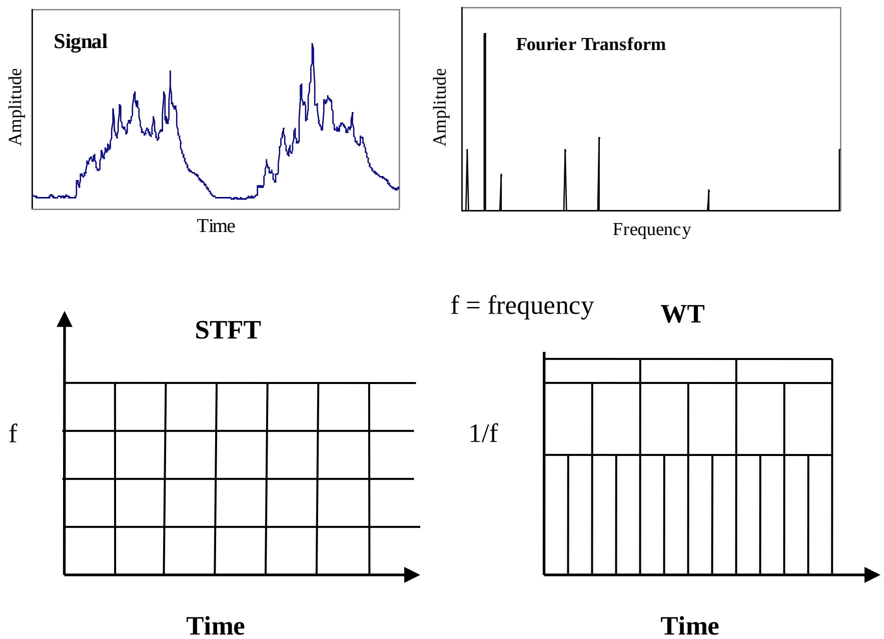


Figure 3.4. Fourier Transform and Wavelet Transformation

The wavelet transform is divided into two types:

1) Continuous wavelet transform (CWT)

2) Discrete wavelet transform (DWT)

3.2.6 Continuous wavelet transform (CWT)

The basic objective of the CWT is to achieve a complete time-scale representation of localized and transient phenomenon occurring at different time scales (Labat, 2008). The Continuous Wavelet Transform (CWT) of a signal *x(t)* is given by the Eqution 3.4.

$$CWT(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} x(t).\psi^{*}\left(\frac{t-b}{a}\right).dt \qquad (3.4)$$

In the above equation, the transformed signal is a function of two variables, *a*and*b*, the scale and translation factor, respectively, of the function ψ *(t)*. * corresponds to complex conjugate (Mallat, 1998). ψ *(t)* is the transforming function, and is called the mother wavelet, which is defined mathematically as

$$\int_{-\infty}^{\infty} \psi(t)dt = 0 \tag{3.5}$$

The term translation is related to the location of the window, as the window is shifted through the signal. This term, obviously, corresponds to time information in the transform domain. The scale parameter is defined as 1/frequency. Low frequencies (high scales) correspond to a global information of a signal (that is usually spans the entire signals), whereas high frequencies (low scales) correspond to a detailed information of a hidden pattern in the signal (that usually lasts a relatively short time).

The CWT is computed by changing the scale of the analysis window, shifting the window in time, multiplying by the signal, and integrating over all times.

The original signal is reconstructed using the inverse wavelet transform as

$$x(t) = \frac{1}{C_{\psi}} \int_{-\infty}^{\infty}\int_{0}^{\infty} \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) CWT(a,b)\, \frac{da.db}{a^2} \tag{3.6}$$

where C_{ψ} is admissibility constant.

The generation of wavelet coefficients for a time series involves five steps (Misiti and Misiti 1996):

i) Given a signal *x(t)* and a wavelet function *Ψ(t),* compare the wavelet to a section at the start of the signal. (**Figure 3.5 a**).

ii) Compute the coefficient, (say *C1,1*; scale = 1, time = 1), which is an indication of the correlation of the wavelet function with the selected section of the signal.

iii) Shift the wavelet to the right (and find coefficient *C1,2*; scale = 1, time = 2) and repeat steps (i) and (ii) until the entire signal is covered. (**Figure 3.5 b**).

iv) Dilate (scale) the wavelet (and find coefficient *C2,1*; scale = 2, time = 1) and repeat steps (i) through (iii). (**Figure 3.5 c**).

v) Repeat steps (i) through (iv) for all scales to obtain coefficients at all scales and at different sections of the original signal.

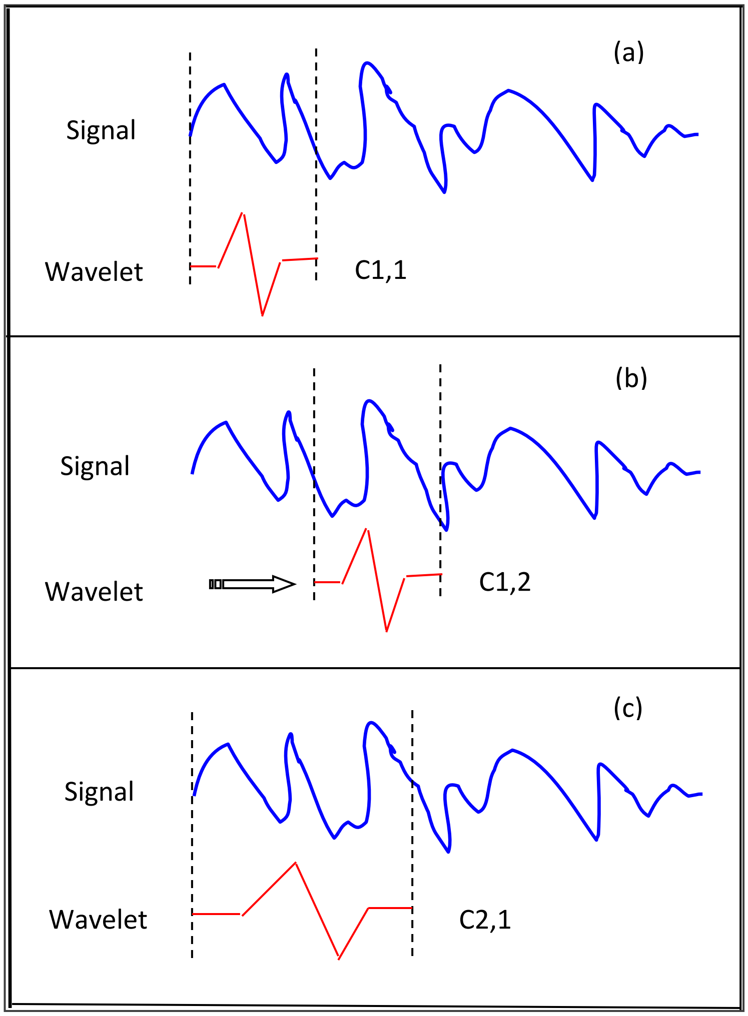


Figure 3.5: Generating wavelet coefficients from a time series

3.2.7 Discrete wavelet transform (DWT)

Usually all hydrological time series data are observed at discrete time interval, rather than continuous time. A discretization of Equation. 3.4 based on the trapezoidal rule is the simplest discretization of the continuous wavelet transform. Calculating the CWT coefficients at every possible scale is a fair amount of work, and it generates a lot of data. CWT produces N^2 coefficients from a data set of length N. Hence redundant information is locked up within the coefficients, which may or may not be a desirable property(Rajaee 2011). If one chooses scales and positions based on the powers of two (dyadic scales and positions) then the analysis will be much more efficient as well as more accurate, which will provide N transform coefficients. This transform is called discrete wavelet, and has the form as

$$\psi_{m,n}(t) = \frac{1}{\sqrt{a_o^m}}\psi\left(\frac{t - nb_o a_o^m}{a_o^m}\right) \tag{3.7}$$

where$m$ and n are integers that control the wavelet dilation and translation, respectively; b_o is the location parameter and must be greater than zero; a_o is a specified fixed dilation step greater 1. From this equation, it can be seen that the translation step $nb_o a_o^m$ depends upon the dilation, a_o^m. The most common and simplest choice for parameters a_o and b_o are 2 and 1 (time steps), respectively. This power of two logarithmic scaling of the translations and dilations is known as the dyadic grid arrangement (Mallat 1999)and is defined as

$$\psi_{m,n}(t) = 2^{-m/2}\psi(2^{-m}t - n) \tag{3.8}$$

For discrete time series, x_t, where x_t occurs at discrete time t, the discrete wavelet transform becomes

$$W_{m,n} = 2^{-m/2}\sum_{i=0}^{N-1}\psi(2^{-m}t - n)x_t \tag{3.9}$$

where$W_{m,n}$ = wavelet coefficient for the discrete wavelet of scale $a = 2^m$ and location $b = 2^m n$. Equaion (3.7) considers a finite time series,x_t, $t = 0, 1, 2, \ldots, N - 1$, and N is an integer power of 2: $N = 2^M$; n is time translation parameter. This gives the range of m and n as, respectively, $0 < n < 2^{M-m} - 1$ and $1 < m < M$. At the largest wavelet scale (i.e. 2^m where $m = M$) only one wavelet is required to cover the time interval, and only one coefficient is produced. At the next scale (2^{m-1}), two wavelets cover the time interval, hence two coefficients are produced, and so on down to $m = 1$. At $m = 1$, the a scale is 2^1, i.e. 2^{M-1} or $N/2$ coefficients are required to describe the signal at this scale. The total number of wavelet coefficients for a discrete time series of length $N = 2^M$ is then $1 + 2 + 4 + 8 + \ldots + 2^{M-1} = N-1$.

In addition to this, a signal smoothed component, $\overline{W}$, is left, which is the signal mean. Thus, a time series of length N is broken into N components, i.e. with zero redundancy. The inverse discrete transform is given by (Mallat 1999):

$$x_t = \overline{W} + \sum_{m=1}^{M} \sum_{n=0}^{2^{M-m}-1} W_{m,n} 2^{-m/2} \psi(2^{-m} t - n) \tag{3.10}$$

or in a simple format as

$$x_t = \overline{W}(t) + \sum_{m=1}^{M} W_m(t) \tag{3.11}$$

where $\overline{W}(t)$ is the approximation sub-signal at level M and $W_m(t)$ are detail sub-signals at levels m = 1, 2,…., M. The detail wavelet coefficients, $W_m(t)$ can capture small features of interpretational value in the data. The residual term $\overline{W}(t)$ represents background information of data.

DWT operates two sets of function viewed as high-pass (wavelet function) and low-pass (scaling function) filters. The original time series are passed through high-pass and low-pass filters (as shown in **Figure 3.6**) and down sampled by two (i.e throwing away every second data point)(Deka and Prahlada 2012). After passing the signal through high pass and low pass filters, detailed (D_1, D_2,…., D_n, which are high frequency components of the original signal) and approximation coefficients (A_1, A_2,….A_n, which are low frequency components of the origininal signal), respectively, are obtained. At any n^{th} decomposition level there will be one series of approximation coefficients at n^{th} level (i.e. A_n) and n series of detailed coefficients (i.e. D_1, D_2,…., D_n), hence there will be total n+1 coefficients and the sum of $A_n + D_1 + D_2 + \ldots\ldots + D_n$ is equal to the original signal x(t).

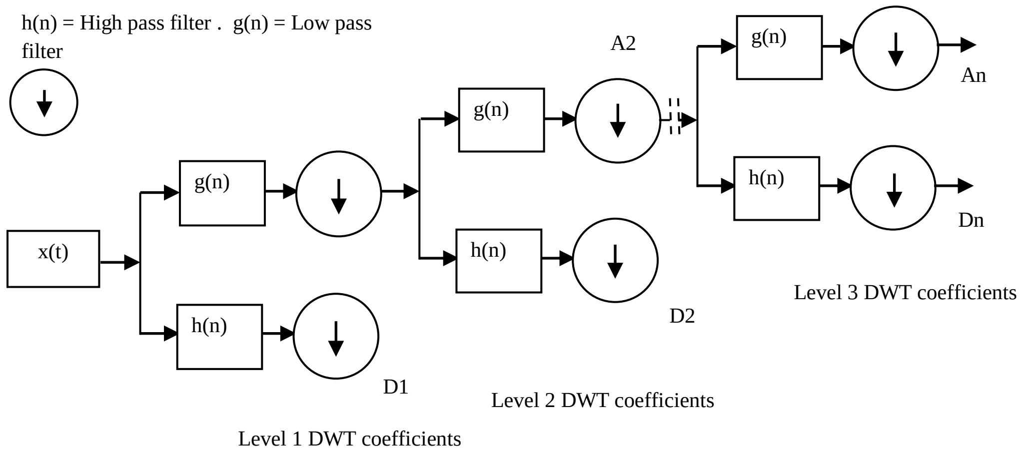


Figure 3.6. Wavelet decomposition tree

The major advantage of using the wavelet method is its robustness since it does not include any potentially erroneous assumption or parametric testing procedure. Apart from that in wavelet method, there exists a wavelet variance decomposition which allows one to study different investing behavior in different time scales independently (Martinez and Gilabert 2009).

This work attempts to come up with a better estimation model that can give satisfactory results over existing methodologies. For this, a signal processing tool called wavelet transformation was introduced to hybridize with the support vector regression and also to ensure the performance comparison was carried out with individual SVR models.

3.3 MODEL DEVELOPMENT FLOW CHART

The methodology adopted in model development and analysis is shown in **Figure 3.7**

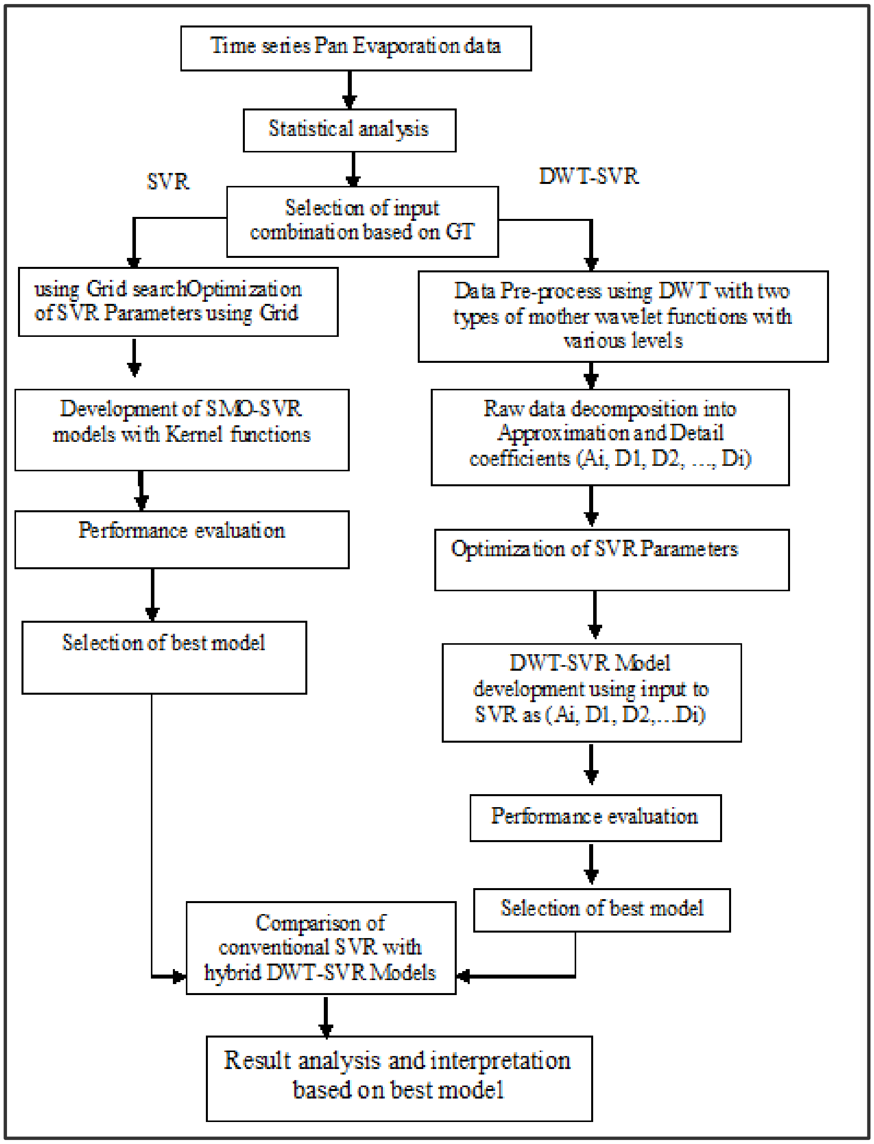


Figure 3.7 Flowchart of model development

3.3.1 Model inputs selection using Gamma test (GT)

The selection of proper combinations of inputs is very essential for finding better input-output patterns. Various techniques are available to decide the best pair of inputs to be used in modeling output. One such technique is Genetic algorithm (GA) to identify possible candidate variables for inclusion in the model. It is as well as conducting a sensitivity analysis, which enables to determine which variables to include in modeling. With advancements in modern computing technology and development of a novel algorithm from the computing science community called the Gamma Test (GT), it has been possible to make significant progresses in tackling these problems (Kaheil et al. 2008). It is achieved by the estimation of variance of the noise Var (r) computed from the raw data using efficient, scalable algorithms. The novel technique of GT enables us to quickly evaluate and estimate the best mean squared error that can be achieved by a smooth model on unseen data for a given selection of inputs, prior to model construction.

Background of Gamma Test: In order to model time series data, we need to construct the model by choosing the past values, up to some number m (often called the embedding dimension) to form the inputs of the model. The output is then the current value of the time series. Thus an embedding of a time series is a selection of past values which are used to predict the current value via a model constructed from the data. A regular embedding takes all past values up to m. An irregular embedding chooses some subset of the m past values, and there are $2^m - 1$ possible irregular embeddings once 'm' is chosen. It was suggested in Judd et.al (1998) that, irregular embeddings may often provide a better model. In this work Gamma test was used in selecting irregular embeddings for time series data.

In a set of input-output data, The GT estimates the minimum mean square error (MSE) that achieve by a smooth model, this estimate is called GT statistic (Γ).

Suppose we have a set of data observations, $\{(x_i, y_i), 1 \le i \le M\}$ that the output y is determined by x input vectors, where $x_i \in R^m$ are vectors confined to some closed bounded set $c \in R^m$ and $y_i \in R^m$ is associated output scalar.

In this method, relationship between input-output can be written as:

$$y = f(x) + r \tag{3.12}$$

Where f and r are a smooth function and a random variable, respectively where r represents noise. The GT is an estimate of the model output variance that cannot be calculated by a smooth data model. The GT is based on the $k^{th}\left(1 \le k \le p\right)$ nearest neighbors $x_{N(i,k)}$ for each vector $x_i\left(1 \le i \le M\right)$.

Specifically, the GT is derived from the delta function of the input vectors:

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M}\left|x_i - x_{N[i,k]}\right|^2 \tag{3.13}$$

Where |…| denotes Euclidean distance, and Gamma function is given as following.

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M}\left|y_{i-}y_{N(i,k)}\right|^2 \tag{3.14}$$

Where $y_{N(i,k)}$ is the corresponding y-value for the kth nearest neighbor of x.

In order to compute Γ a least squares regression line is constructed for the ppoints $\left(\delta_M(k), \gamma_M(k)\right)$.

$$\gamma = A\delta + \Gamma \tag{3.15}$$

The intercept on the vertical axis $(\delta = 0)$ is the Γ value, it can be shown that $\gamma_M(k) \to Var(r)$ in probability $\gamma_M(k) \to 0$.

The GT offers an estimate of the best MSE achievable using a modeling technique for unknown smooth functions of continuous variables (Evans and Jones 2002). The GT is a mathematical algorithm, which reduces volume of model development work and creates guidance for proper needed input data and most important variables before developing model.

The GT is a non parametric method and the results apply regardless of the particular techniques used to subsequently build a model of f. One can standardize the result by considering another term V- ratio, which returns a scale invariant noise estimate between 0 and 1. V-ratiois defined as,

$$V - ratio = \frac{\Gamma}{\sigma^2(y)}$$

where $\sigma^2(y)$ = variance of output y, which allows a judgment to be formed independent of the output range as to how well the output can be modeled by a smooth function. A V-ratio value close to 0 indicates that there is a high degree of predictability of the given output y. The reliability of the Γ statistic can be determined by runninga series of the GT for a definite number of unique data points M, to establish the size of the data set required to produce a stableasymptote. The GT result would avoid the over-fitting of a model beyond the stage where the MSE on the training data is smaller than $Var(r)$ and help one to decide the required data length tobuild a meaningful model.

This technique can be used to find the best embedding dimensions and time lags for time series analysis. This information would help us determine the best input combinations to achieve a particular target output (Cortes and Vapnik 1995). The GT is designed to efficiently solve overtraining problem, as one of the serious weaknesses associated with almost all nonlinear modeling techniques, by giving an estimate of how closely any smooth model could fit the unseen data. In practice, the Gamma test can be achieved through winGamma™ software implementation (Cortes and Vapnik 1995). The Gamma test is known as a tool for non-linear modeling and analysis, which can examine input/output pattern in a numerical data set. More importantly, the Gamma test estimates the part of the output variance which cannot be accounted by any smooth model based on the inputs, even though this model is unknown. This tool is handy also because of its rapid processing of data, especially in large databases which consisting thousands of points of data sets, While a single run of the GT takes a few moments (Jones 2004).

Model Identification using Genetic Algorithm (GA) in GT : An embedding is a selection of inputs chosen from all the possible inputs. In winGamma, an embedding is designated by a string of '1's and '0's called a mask. Thus if there are five inputs the mask then '10111' indicates that all inputs are to be used in the embedding except the second.A useful feature associated with a full embedding or GA search is the Embedding Histogram, which shows the frequency of embeddings with a specific Gamma statistic. GA searches the space of all masks using a Genetic Algorithm (GA) to find good embeddings. The parameters which can be used to control this search are (default values of parameters are given in brackets):

Population Size (100): The size of the population of masks being used throughout the search.

Mutation Rate (0.01):The probability that an individual bit will be mutated during the reproduction process.

Crossover Rate (0.5):The chance of inserting a random length run of bits from a parent maskto a child mask (i.e. the probability that a crossover event occurs during the reproduction process).

Gradient Fitness (0.1) :The weighting in the GA fitness function for masks giving a low gradient in the Gamma Test. Increasing this weighting will place more emphasis on the relative simplicity of the modeling function

Intercept Fitness (0.8) :The weighting in the GA fitness function for masks with a low absolute value of the Gamma statistic. Increasing this weighting will place more emphasison the model accuracy.

Length Fitness (0.1) :The weighting in the GA fitness function for masks with a given numberof '1's. Increasing this weighting will encourage the selection of masks with fewer '1's andthereby place more emphasis on simpler models.

In this work Genetic Algorithm was performed in different dimensions varying the number of inputs to the model, which clearly presented the response of the data model to some different combination of input data sets. Input combinations and gamma test results generated for Bajpe and Bangalore stations are displayed in **Table 3.2 and Table 3.3** respectively.

Table 3.2: Gamma Test Results for Selection of Input Combinations (Bajpe Station)

Input Combination	Gamma Value	Standard error (S.E)	V-Ratio
T	0.116	0.001	0.467
T + W	0.113	0.001	0.455
T + W+ P	0.074	0.005	0.296
T + W + P + Rh	0.073	0.004	0.288
T + W + P + Rh + Sh	0.072	0.002	0.289

The developed model, consisting of a five input and one output set of I/O pairs) was identified as the best structure because it showed low noise level (Γ value), low Gamma value, the low V- ratio value (indicating the existence of a reasonably accurate, smooth model). These values altogether can give a clear indication that it is quite adequate to construct a nonlinear predictive model using around 1826 data points. The combination including all inputs found to show the least gamma value 0.072 and SE 0.002 and selected as best combination.

Table 3.3: Gamma Test Results for Selection of Input Combinations (Bangalore Station)

Input Combination	Gamma Value	Standard error (S.E)	V-Ratio
T	0.145	0.001	0.582
T + W	0.139	0.002	0.557
T + W+ P	0.110	0.003	0.443
T + W + P + Rh	0.077	0.002	0.311
T + W + P + Rh + Sh	0.075	0.003	0.303

The model combination, including a five input and one output set of I/O pairs) was identified as the best structure because it showed low noise level (Γ value), low Gamma value, the low V- ratio value (indicating the existence of a reasonably accurate, smooth model. These values altogether can give a clear indication that it is quite adequate to construct a nonlinear predictive model using around 2555 data points. The combination including all inputs found to show the least gamma value 0.075 and SE 0.003 and selected as best combination.

3.3.2 Importance of parameter optimization

The parameters used for model building influence the effectiveness of the nonlinear SVR. Among them major parameters are the cost constant C, the radius of the insensitive tube ε, and the kernel parameters (Drucker et al. 1999). These parameters are mutually influences each other and hence varying the value of one parameter brings changes in the other linked parameters also.

The parameter C identifies the smoothness/flatness of the approximation function. The smaller the value of C leads to a poor approximation resulting in under-fitting of training data. On the other side greater C value over-fits the training data and sets its objective to minimize only the empirical risk making way for more complex learning.

The Parameter optimization process is very tedious and requires lots of trial and error (Raghavendra. N and Deka 2014). Nevertheless the model will be benefitted to more efficient and reliable. The grid search results obtained using Bajpestation tabulated in **Table 3.4** and same for Bangalore is shown in **Table 3.5**.

The parameter denotes smoothening the complexity of the approximation function and controls the width of the ε-insensitive zone used for fitting the training data. Ultimately the number of support vectors is based on parameter ε and then both the complexity and the generalization capability of the approximation function is dependent upon its value. It also governs the precision of the approximation function. Smaller values of ε lead to more number of support vectors and results in complex learning machine. Greater ε values result in more flat estimates of the regression function. .Since the present study includes various combinations of inputs. The epsilon (ε) values for 0.01 the prime parameters C and γ were optimized.

At SMO-SVR, there are two methods to arrive at optimal parameter values, a grid search and a cross-validation. Grid search attempts to find values of each parameter across the specified search range using geometric steps. Generally grid search needs abundant data for computations hence not economical computationally, as the model is evaluated at various points within the grid for each parameter. If cross-validation parameter selection is employed then V-fold cross-validation is used by the search to estimate the optimal parameters using the error computed from the training data. The derived parameters were later used as inputs for Sequential Minimal Optimization (SMO-Reg) kernel SVR functions for further computations. For this study, a value of ε =0.001 was working well with the desired parameters.

Table 3.4: Optimized Parameters for Combinations of Input Parameters used in SVR models (Bajpe Station)

Input Combinations	No. of Support Vectors	Polynomial Kernel			RBF Kernel			PUK Kernel		
		C	Deg	E	C	γ	ε	C	ω	Σ
T	1627	18.00	2.00	0.01	20.00	2.00	0.01	24.00	5.50	1.14
T+W	1604	14.50	2.00	0.01	5.00	10.00	0.01	8.75	4.80	1.65
T+W+P	1578	16.25	2.00	0.01	1.00	1.00	0.01	3.25	3.90	2.70
T+W+P+Rh	1567	15.65	2.00	0.01	2.00	10.00	0.01	4.10	2.45	3.80
T+W+P+Rh+Sh	1552	3.50	2.00	0.01	3.00	1.00	0.01	2.75	1.95	4.20

Table 3.5: Optimized Parameters for Combinations of Input Parameters used in SVR models (Bangalore Station)

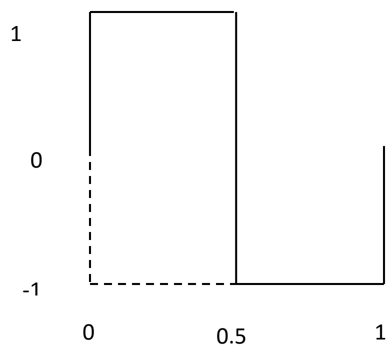
Input Combinations	No. of Support Vectors	Polynomial Kernel			RBF Kernel			PUK Kernel		
		C	Deg	E	C	γ	ε	C	ω	Σ
T	2195	42.00	2.00	0.01	44.00	5.00	0.01	40.00	3.50	1.25
T+W	2056	31.00	2.00	0.01	25.00	7.00	0.01	23.00	3.30	1.95
T+W+P	2102	18.00	2.00	0.01	14.00	10.00	0.01	17.00	4.55	2.60
T+W+P+Rh	2059	11.00	2.00	0.01	9.25	1.00	0.01	10.50	2.75	3.20
T+W+P+Rh+Sh	1976	6.75	2.00	0.01	7.33	3.30	0.01	5.57	2.80	3.60

3.3.3 Mother Wavelets selection

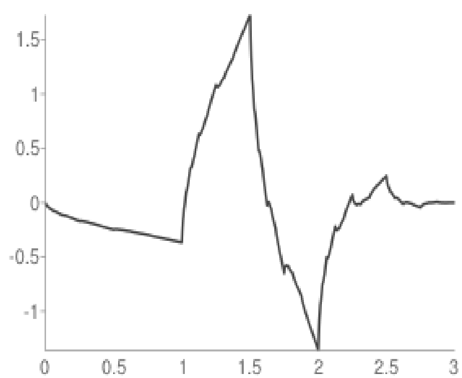
In this study wavelet transform are employed for their capabilities handle nonlinear data via denoising, which can enhance the SVR modeling accuracy. (Torrence and Compo 1998) described wavelet as a tool to analyze the irregularities in dataset within a time series. Because of compact support in which wavelets are defined, wavelet filter banks are well suited to denoise, decompose, nonlinear time series. Fundamental manuals and practical's guide to wavelet analysis were provided by(Torrence and Compo 1998; Witten and Frank 2005; Daubechies 1992; Meyer 1992; Mallat 1999).

The choice of the mother wavelet depends on the nature of data to be analyzed. In this study, evaporation data signals found to exhibit non linearity and seasonal irregularities, and noisy. The data needed to be matched with the irregular mother wavelet pattern. In this perspective, Daubechies (Db) of order 1 (db1) to 5 (db5) has been employed. Daubechies wavelets of order 1 to 5 are shown in Figure 3.8. All Daubechies wavelets of order N (dbN) are asymmetric, orthogonal and biorthogonal. They are compactly supported wavelets with extremal phase and highest number of vanishing moments for a given support width (Misiti, 2010).

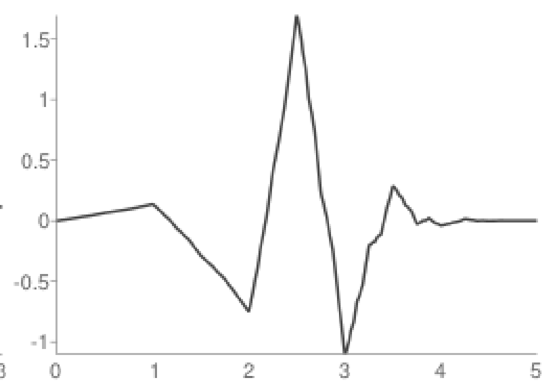
Additional to Daubechies wavelets, the Haar wavelet functions of order 1 to 3 were also employed to make comparison between these two over their potential to pre-process the data. Haar wavelets are the simple, fast, memory efficient and exactly reversible without edge effects that are problem with other wavelet transforms.



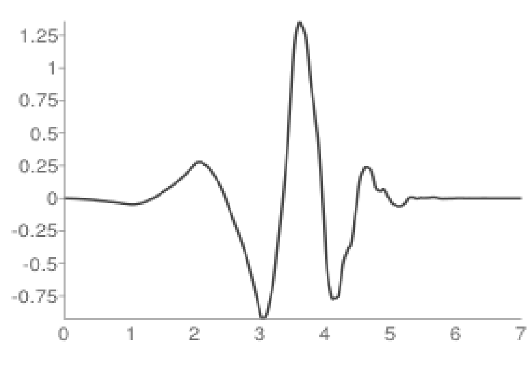
(a) db1



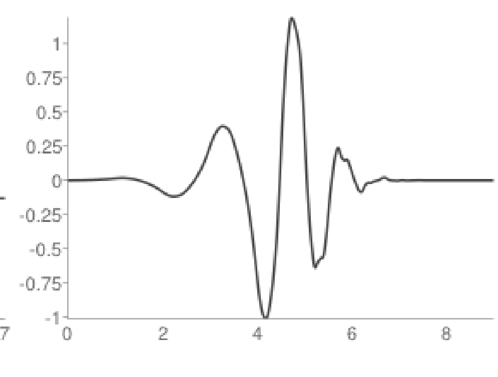
(b) db2



(c) db3



(d) db4



(e) db5

Figure 3.8 Daubechies wavelets of order 1 to 5

The approximation at level 3 and detail coefficients of pan evaporation at Bajpe station and Bangalore station using db4 wavelet at level 3 along with reconstructed signals is shown in **Figure 3.9 and 3.10** respectively.

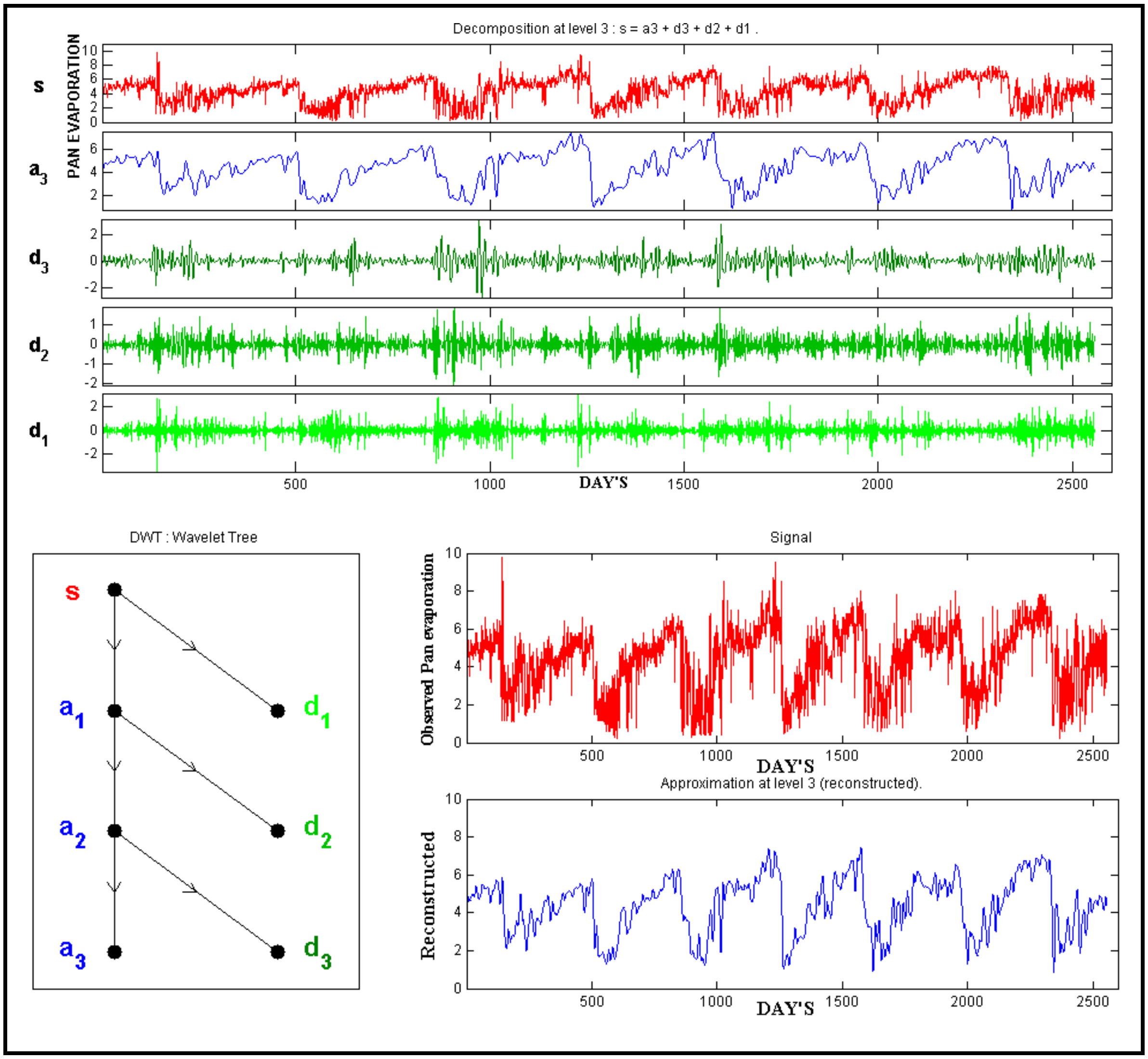


Figure 3.9: Decomposition (Approximation and detail) of observed pan evaporation data signal at level 3 using db4 wavelet for the full length dataset of Bajpe station

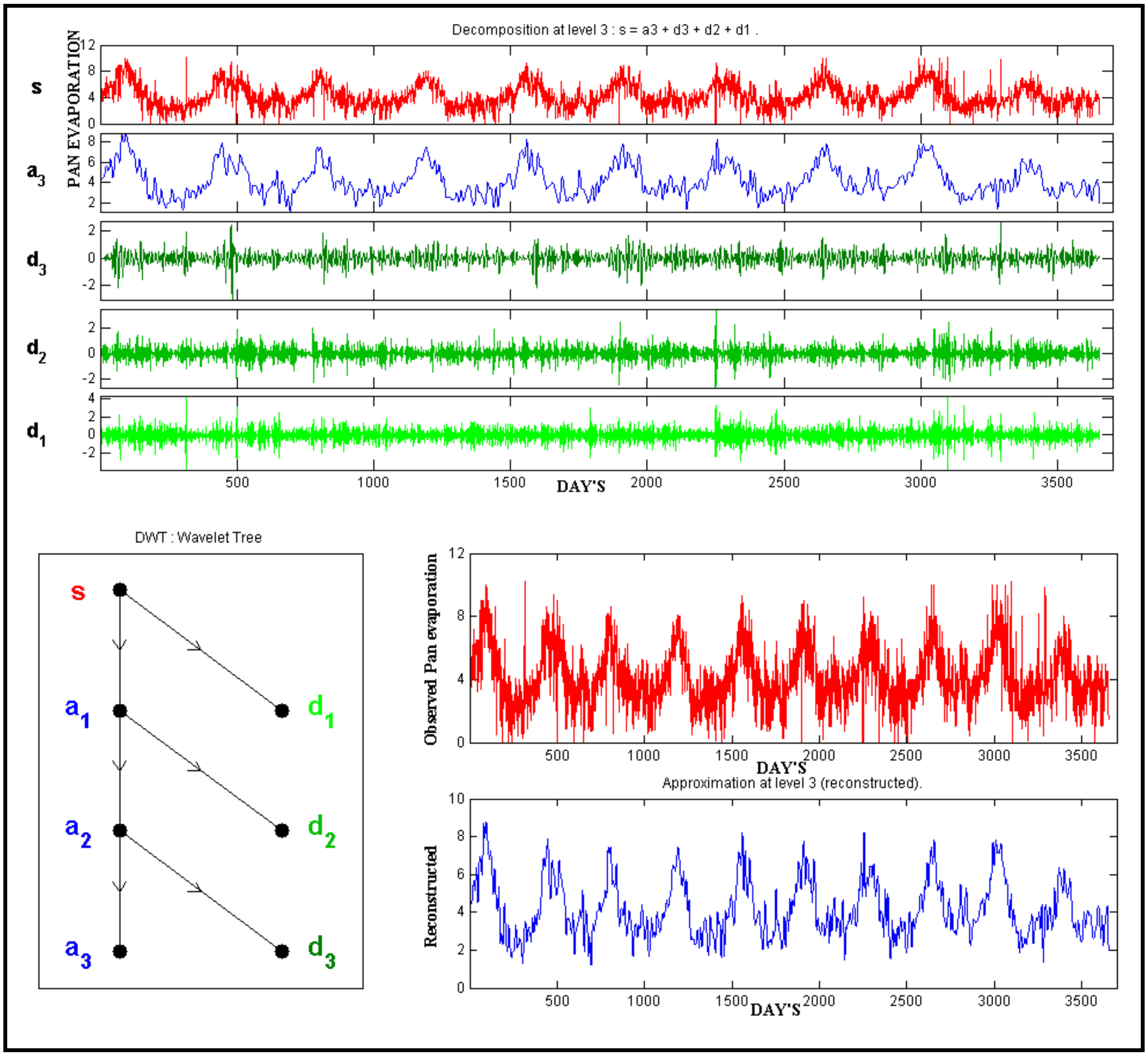


Figure 3.10: Decomposition (Approximation and detail) of observed pan evaporation data signal at level 3 using db4 wavelet for the full length dataset of Bangalore station

The decomposition of raw pan evaporation data results in processing to use it for further regression modeling process (SVM). As mentioned in the previous sections parameter optimization needs to be performed for the DWT processed data before SVM to learn the pattern and develop better estimations of PE. The optimized parameters for DWT-SVR is shown in **Table 3.6 and Table 3.7** for Bajpe station and Bangalore station respectively.

Table3.6: Optimal SVR parameters for DWT-SVR models (Bajpe station)

SMO-SVR RBF	Mother Wavelet Functions							
	DB4 level 1	DB4 level 2	DB4 level 3	DB4 level 4	DB4 level 5	Haar level 3	Haar level 4	Haar level 5
C	7.00	5.50	9.00	14.00	8.75	8.00	10.00	12.00
Gamma	3.75	4.30	5.10	6.45	4.75	3.75	5.25	6.50
Epsilon	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table3.7: Optimal SVR parameters for DWT-SVR models (Bangalore station)

SMO-SVR RBF	Mother Wavelet Functions							
	DB4 level 1	DB4 level 2	DB4 level 3	DB4 level 4	DB4 level 5	Haar level 3	Haar level 4	Haar level 5
C	14	12	11	9.25	6.75	8	9	11.25
Gamma	8.5	6.25	3.1	3.75	1.25	2.25	5.9	6.7
Epsilon	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

3.4 EVALUATION CRITERIA

The performance of the developed models needs to be evaluated qualitatively and quantitatively. The superiority of the models was judged by visual observation and also on the errors indicated by statistical indices. The evaluation of performances of SVR and DWT-SVR models were gauged against statistical indices such as Root mean square error (RMSE), Mean absolute error (MAE), Correlation coefficient (CC) and Nash-Sutcliffe efficiency (NSE). The expressions of the statistical indices used in the study are displayed in the following equations with the notations as: N is the number of observations; X is the observed values; Y is the estimated values; $\overline{X}$ is the mean of observed values; $\overline{Y}$ is the mean of estimated values.

1. $\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(X\text{-}Y)^2}{N}}$ (3.16)

Models with lowest RMSE are considered to be superior models and vice-versa.

2. $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|Y\text{-}X\right|$ (3.17)

Superior models yield lowest MAE and vice-versa.

3. $\text{CC} = \frac{\sum_{i=1}^{N}(X\text{-}\overline{X}).(Y\text{-}\overline{Y})}{\sqrt{\sum_{i=1}^{N}(X\text{-}\overline{X})^2.\sum_{i=1}^{N}(Y\text{-}\overline{Y})^2}}$ (3.18)

The value of a correlation coefficient ranges between -1 and 1. Models with superior performance indicate value close to 1 and vice-versa.

4. $\text{NSE} = 1\text{-}\frac{\sum_{i=1}^{N}(X\text{-}Y)^2}{\sum_{i=1}^{N}(X\text{-}\overline{X})^2}$ (3.19)

The Nash–Sutcliffe model efficiency coefficient is used to assess the estimated capacity of pan evaporation models. Nash–Sutcliffe efficiency ranges between $-\infty$ to 1. An efficiency of 1 indicates best match of modeled pan evaporation data to the observed data and vice-versa.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 INTRODUCTION

In this research, a novel hybrid combination of wavelet transform support vector machine model for daily pan evaporation is proposed. The purpose of this study is to investigate the performance of a hybrid model of discrete wavelet transform support vector regression (DWT-SVR) and conventional single support vector regression models. The study was carried out on two climatically contrasting stations Bajpe and Bangalore as mentioned in the introduction. The stations are spaced 350 Kms with each other representing different climatic zones as stated in the chapter 3. SVM proved its capabilities for both classification and regression related to hydrological problems and wavelet transform known for its generalization capabilities as a data pre-processing technique. The hybrid combination is expected to provide better performance compared to conventional individual techniques. Result analysis is conducted to decide upon the superiority of the techniques. The analysis will suggest robust and capable model to estimate pan evaporation, highlighting the factors responsible for evaporation to occur.

Using meteorological time series evaporation data, statistical analysis was carried out on all the attributes associated with pan evaporation for both the stations (**Refer Table 3.1 and Table 3.2**). The daily time series data of seven years recorded at Bajpe station and are used. Out of which, five years of data utilized in training the models and two years for testing the models. Similarly a total of nine year daily time series data recorded at Bangalore station used in the model building, of which six year data used in training and three years data for testing models. The training and testing data ratio were considered as a thumb rule of 70% training-30% testing combination (Fielding and Bell 1997). The weather variables used in this study consist of mean air temperature (T), wind speed (W), rainfall (P), mean relative humidity (Rh), sunshine hours (Sh) and pan evaporation (E). It is seen over the statistical analyses that attributes and their ranges are varied significantly among the stations and suggests for the usefulness of the study on the climatically contrasting zones. The correlation pattern showed similarity among both the stations, i.e. mean air temperature, wind speed, and sunshine hours showed positive correlation with pan evaporation and remaining such as rainfall, mean relative humidity showed negative correlation with pan evaporation. In that, major affecting factors for evaporation to take place in the order of higher influence are mean air temperature, sunshine hours and wind speed.

The data analysis and input – output combination was carried out using Gamma test over the time series data of both the stations. The various combinations were evaluated to decide the best possible combination.

Once the inputs and output combination was finalized, support vector machine regression technique was employed in model building and analysis without pre-processing of raw data. As discussed in the previous chapters, SMO-SVR methodology was employed with three types of kernels i.e Polynomial, Radial basis function, and PUK kernels which are competitive enough to provide judgmental results. The parameter selection was later conducted using a grid search with various trials and errors to reach desired parameters to enhance model efficiency to produce better results as discussed in previous sections. The combination including all five input parameters with evaporation as output combination was amongst the high performance. All the three kernels used in training and testing, respectively re-confirm that, for evaporation to take place these listed parameters must act united than individual influences.

In the next stage, the wavelet support vector regression (DWT- SVR) models are obtained combining two methods, DWT and SVR. The pre-processing of data was carried out using DWT. The DWT-SVR model is an SVR model, which uses coefficients generated from decomposing original data and recompiling the data to feed in SMO-SVR. For the DWT- SVR model inputs, signals split into a detail and an approximation. The approximation obtained from the first-level is split into new detail and approximation, and this process is repeated. From the basics of DWT, the maximum level of decomposition can be computed using the formula (Nourani et al. 2009; Wang and Ding 2003).

$$L=2^M \tag{4.1}$$

$$\text{Hence, } M = [\ln L / \ln 2] \tag{4.2}$$

Where L is length of data points and M is maximum decomposition level. In the present study the length of the data set is 1826 and 2557 for Bajpe and Bangalore station's respectively. Accordingly, maximum decomposition may occur in DWT up to 10th level for Bajpe training data and 11th for Bangalore training data. However the optimum level of decomposition occurred in order 4 and 3 of Daubechies wavelet mother function. Because of the fact that wavelet transform decomposes only the approximations of the signal, it may cause problems while applying wavelet transform in certain applications where the important information is located in higher frequency components (Shinde et al. 2013).

As discussed in the DWT methodology in the previous section, suitability of mother function for this study was decided on the basis of trial and error method and finally Daubechies wavelet mother function with level 3 was confirmed and derived coefficients from that for further DWT-SVR models. The function of discrete wavelet transformation is to discretize the non-stationary pan evaporation data into stationary sub signals to separate the periodic properties, non-linearity and dependence relationship. These sub signals usually in the form of approximation coefficients (A1, A2.., An) and detail coefficients (D1, D2.., Dn).

For wavelet analysis, Discrete Wavelet Transformation (DWT) is used and Daubechies wavelet order-4 (db4) was selected as a mother wavelet. The selected mother wavelet 'db4' is a simplest wavelet having only 3 wavelet filter coefficients with exact reconstruction possibilities. To get the decomposed wavelet coefficients here, various decomposition levels has been tried (L1 to L5) but only Daubechies wavelet mother function with level 3 was showing better results when fed as inputs to SVM on trial and error basis to enhance the performance.

As discussed in the previous section the parameter selection is very essential to build efficient models. Grid search with various trails and errors conducted to reach desired parameters. The optimized parameters for DWT-SVR are displayed in **Table 3.6 and 3.7** (previous chapter).

4.2 RESULTS WITH SUPPORT VECTOR REGRESSION (SVR)

In the first stage, the SMO-based SVR model is employed for the modeling of pan evaporation using various input combinations. The input parameters identified in statistical analysis of daily time series data recorded at both the stations considered for the study were used. Although gamma test recognizes the better input – output pattern, but justification needs robust and reliable technique such as SVR to be tested with. The model is trained with available daily time series data for five years (2000-04) and tested with two years (2005–2007) of data recorded at Bajpe station and daily time series data of six years (1975-81) and tested with three years (1982–1984) of data recorded at Bangalore station. As discussed in the previous chapter the key factor in the success of SVR lies with kernel functions and optimum parameters to fine tune the kernel functions. However, identification of appropriate kernel function in the advance of model building is not possible. Therefore, trial and error iterations are essential to judge the suitability of kernel function for the data set. The accuracy of kernels based relies on the selection of the model parameters. The best fitting of models depends upon the number of support vectors generated during model building. The generated number of support vectors should not be more than 70% to avoid over-fitting and whereas below 30% lead to under-fitting (Fielding and Bell 1997). The generated results are tabulated in the previous chapter. In this study three types of kernel functions are utilized and their optimal parameters were found out using SMO-SVR (Sequential mean optimization support vector regression) with grid search and CV parameter optimization, as discussed in the methodology chapter. The optimum results computed with SVR are displayed in **Table 4.1 and 4.2**.

4.2.1 Performance evaluation

As disused in the previous chapter, the estimated pan values by SVR models and DWT-SVR models were compared with observed values using four types of performance measures. They are RMSE, MAE, CC, and NSE. The model performance is said to be optimum with lowest RMSE and MAE, whereas CC and NSE are supposed to be closer to the value one.

Different combinations of input variables that influence the pan evaporation the most were tried on training and testing data sets of both the stations.. It is clear from the **Tables 4.1 to 4.4** that, as the number of influential attributes combines together, model superiority increases. The estimation efficiency is enhanced. In the **Tables 4.1 and 4.2** corresponding to training and testing data from Bajpe station, there are five types of input combinations are evaluated employing three types of kernel functions as listed in the previous chapter. Initially with a lone attribute of temperature the model results are very poor, showing higher RMSE, MAE and lower values of CC and NAE. But as the model gets added with more influential attributes, it shows superior performance. Tables also reveal that, the kernel functions played their roles to make model superior and robust. For the all five input combination scenarios, RBF kernel function showed slightly better performance in comparison to the other two kernel functions in both training and testing period modes. It is also found that, PUK showed near similar results to RBF kernel function. Considering the results phase wise, it is observed that, testing phase results are superior to training results for the estimations computed for Bajpe station.

Table 4.1: Statistical indices of SMO-SVR models with combinations of input parameters for Bajpe station. (Training data)

Input Combination	SMO-SVR Polynomial kernel				SMO-SVR RBF kernel				SMO-SVR PUK kernel			
	RMSE (mm)	**MAE (mm)**	**CC**	**NSE**	**RMSE (mm)**	**MAE (mm)**	**CC**	**NSE**	**RMSE (mm)**	**MAE (mm)**	**CC**	**NSE**
T	1.321	0.898	0.692	0.931	1.121	0.912	0.747	0.946	1.142	0.926	0.731	0.940
T + W	1.264	0.778	0.720	0.942	0.995	0.878	0.795	0.959	1.002	0.897	0.776	0.956
T + W+ P	1.147	0.782	0.745	0.953	0.982	0.722	0.815	0.965	0.991	0.825	0.794	0.967
T + W + P + Rh	1.132	0.768	0.781	0.961	0.961	0.692	0.827	0.973	0.963	0.695	0.821	0.970
T + W + P + Rh + Sh	1.032	0.759	0.794	0.967	**0.941**	**0.687**	**0.832**	**0.977**	0.906	0.650	0.845	0.981

Table 4.2: Statistical indices of SMO-SVR models with combinations of input parameters for Bajpe station. (Testing data)

Input Combination	SMO-SVR Polynomial kernel				SMO-SVR RBF kernel				SMO-SVR PUK kernel			
	RMSE (mm)	**MAE (mm)**	**CC**	**NSE**	**RMSE (mm)**	**MAE (mm)**	**CC**	**NSE**	**RMSE (mm)**	**MAE (mm)**	**CC**	**NSE**
T	1.348	1.145	0.611	0.925	1.168	0.929	0.757	0.955	1.142	0.951	0.721	0.951
T + W	1.214	0.926	0.655	0.940	1.112	0.878	0.735	0.959	1.002	0.897	0.731	0.956
T + W+ P	1.151	0.898	0.754	0.951	1.024	0.915	0.785	0.962	1.121	0.922	0.781	0.967
T + W + P + Rh	1.122	0.819	0.785	0.969	0.997	0.793	0.817	0.977	1.005	0.802	0.821	0.973
T + W + P + Rh + Sh	0.997	0.770	0.834	0.972	**0.990**	**0.763**	**0.839**	**0.985**	0.981	0.761	0.838	0.977

Table 4.3: Statistical indices of SMO-SVR models with combinations of input parameters for Bangalore station. (Training data)

Input Combination	SMO-SVR Polynomial kernel				SMO-SVR RBF kernel				SMO-SVR PUK kernel			
	RMSE (mm)	**MAE (mm)**	**CC**	**NSE**	**RMSE (mm)**	**MAE (mm)**	**CC**	**NSE**	**RMSE (mm)**	**MAE (mm)**	**CC**	**NSE**
T	1.485	1.167	0.592	0.948	1.427	1.089	0.632	0.952	1.432	1.096	0.629	0.952
T + W	1.463	1.147	0.610	0.949	1.327	1.041	0.667	0.955	1.400	1.069	0.651	0.954
T + W+ P	1.413	1.088	0.644	0.953	1.249	0.929	0.736	0.963	1.344	1.013	0.686	0.957
T + W + P + Rh	1.108	0.829	0.798	0.971	1.023	0.747	0.831	0.975	1.004	0.724	0.838	0.976
T + W + P + Rh + Sh	1.081	0.808	0.809	0.972	**1.009**	**0.728**	**0.836**	**0.976**	1.060	0.785	0.817	0.973

Table 4.4: Statistical indices of SMO-SVR models with combinations of input parameters for Bangalore station. (Testing data)

Input Combination	SMO-SVR Polynomial kernel				SMO-SVR RBF kernel				SMO-SVR PUK kernel			
	RMSE (mm)	**MAE (mm)**	**CC**	**NSE**	**RMSE (mm)**	**MAE (mm)**	**CC**	**NSE**	**RMSE (mm)**	**MAE (mm)**	**CC**	**NSE**
T	1.519	1.059	0.571	0.943	1.485	0.997	0.600	0.946	1.481	1.007	0.600	0.946
T + W	1.543	1.080	0.563	0.941	1.545	1.049	0.569	0.941	1.595	1.116	0.547	0.937
T + W+ P	1.519	1.028	0.582	0.943	1.497	0.984	0.610	0.945	1.520	1.017	0.593	0.943
T + W + P + Rh	1.379	0.879	0.673	0.953	1.390	0.883	0.668	0.952	1.395	0.881	0.668	0.952
T + W + P + Rh + Sh	1.431	0.912	0.651	0.949	**1.382**	**0.887**	**0.681**	**0.959**	1.389	0.888	0.671	0.952

The modeled results for the Bangalore station training and testing phase are displayed in **Table 4.3** and **Table 4.4.** Once again the combination of all influential parameters affected the pan evaporation the most. Kernel estimated results varied with the different combinations and also with training and testing models. The drawback experienced in this set of model performance is the sudden decline in the performance from training to testing data phase. Taking into account of RBF kernel results for discussion on the matter of such decline, the values of RMSE, MAE, CC, and NAE were 1.009, 0.728, 0.836 and 0.976 respectively in the training phase. The values of RMSE, MAE, CC, and NAE at testing phase were 1.382, 0.887, 0.681, and 0.959 respectively. There is a rise in values of RMSE and MAE and reduced values of CC and NAE confirms such decline. The reasons behind such decline may be due to the variations in the ranges of variables. (**Refer Table 3.1**). It was observed in the statistical analysis of both the station data's, that some individual attributes such as wind speed, rainfall and relative humidity showed much of variations in their ranges. This may be due to seasonal atmospheric variations. The sparseness in data recorded was found more with Bangalore data than, Bajpe data. The individual SVR model has captured these seasonal variations to its limitations.

The computed results confirm the RBF's superiority in modeled estimations in both training and testing phases for the varied climatic stations. Among the other two kernel estimations PUK estimations found near close to RBF estimations. The Polynomial showed much contrasting results with RBF in both training and testing modeled estimations for Bangalore station.

The**Figures 4.1, 4.2,** and **4.3** represent estimation of SVR Polynomial, RBF, and PUK estimations of pan evaporation over observed pan evaporation for Bajpe and Bangalore stations respectively.

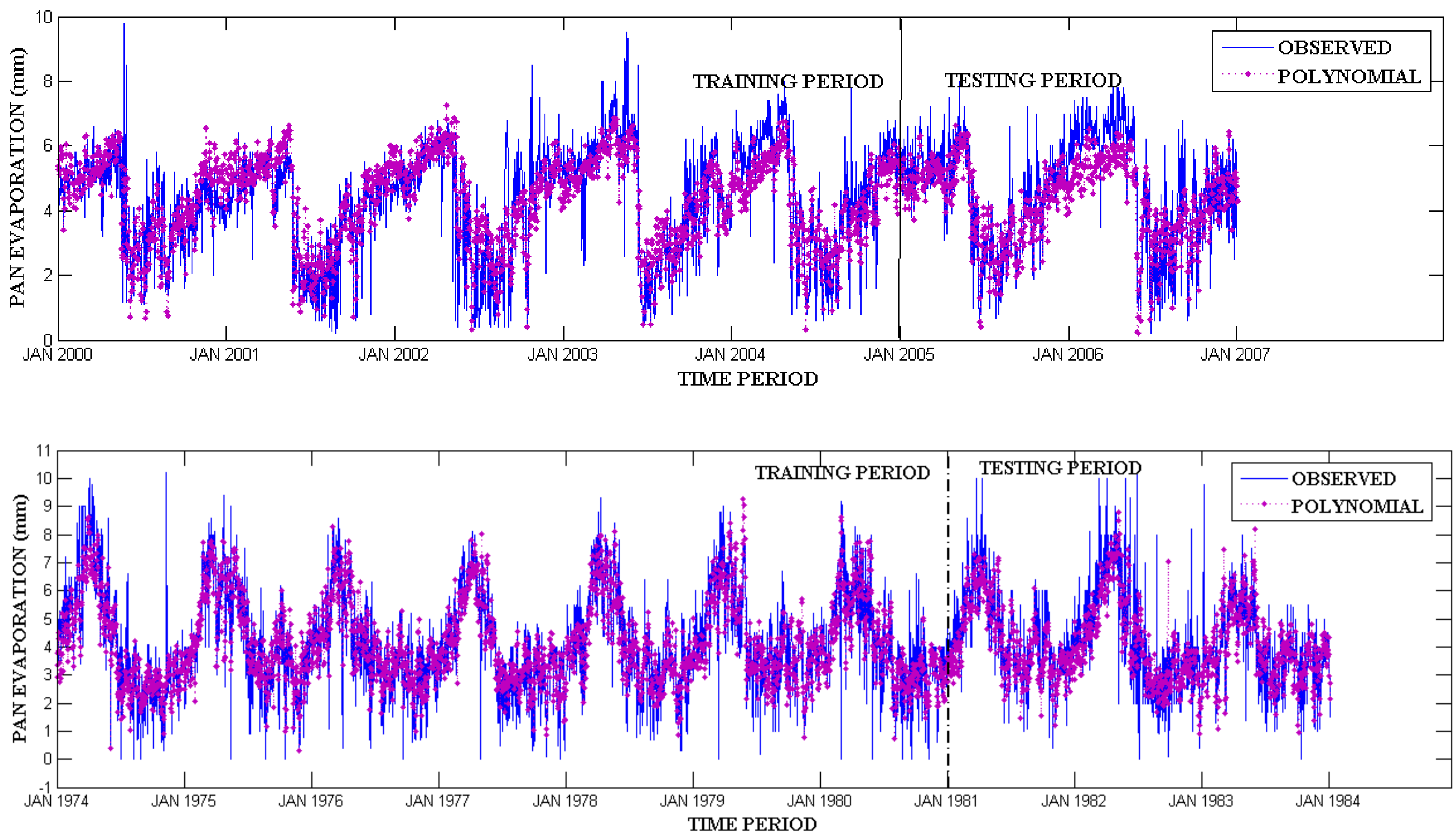


Figure 4.1: Pan evaporation estimation using Polynomial kernel for Bajpe (Top) and Bangalore station (Bottom).

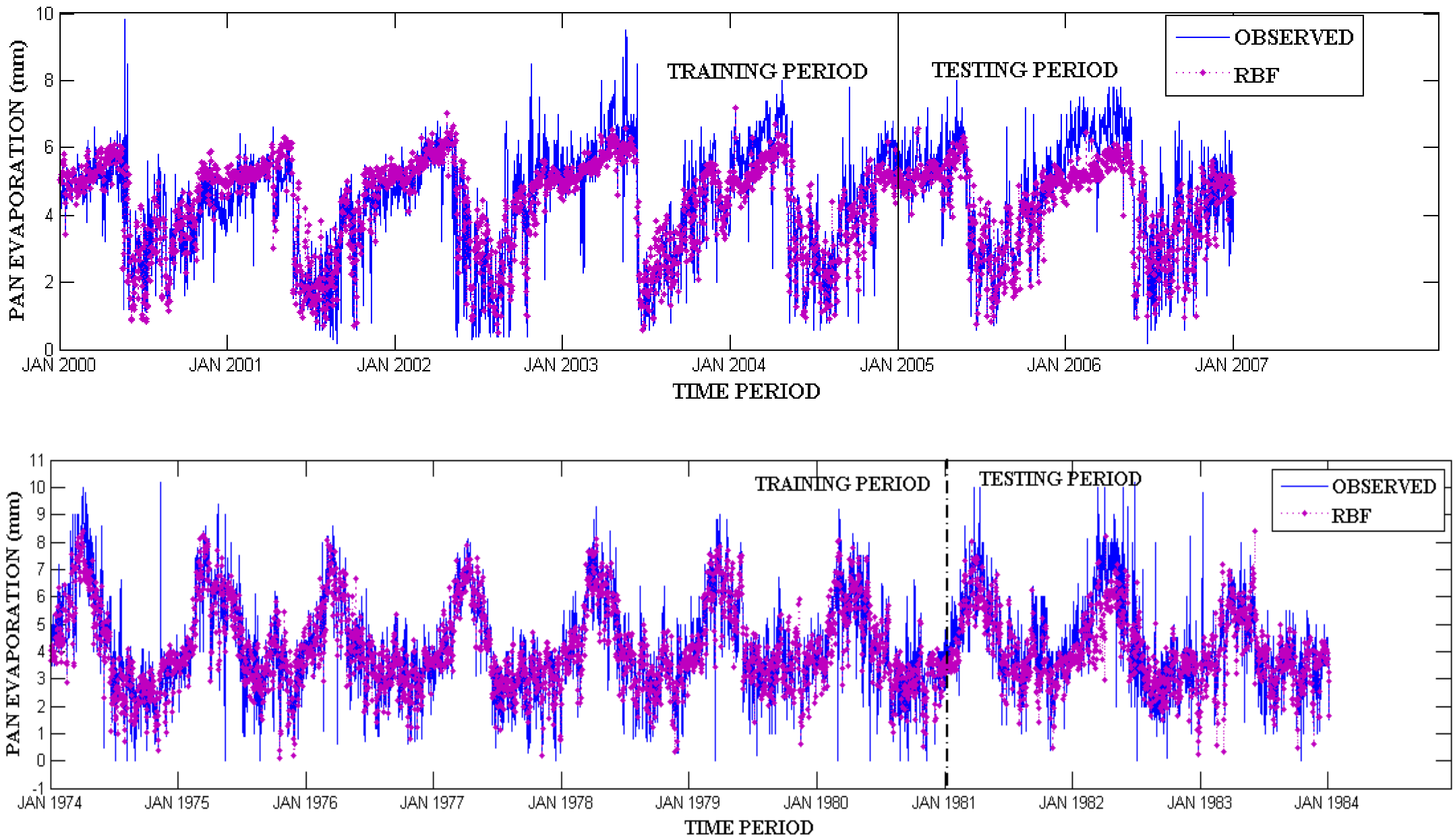


Figure 4.2: Pan evaporation estimation using RBF kernel for Bajpe (Top) and Bangalore station (Bottom).

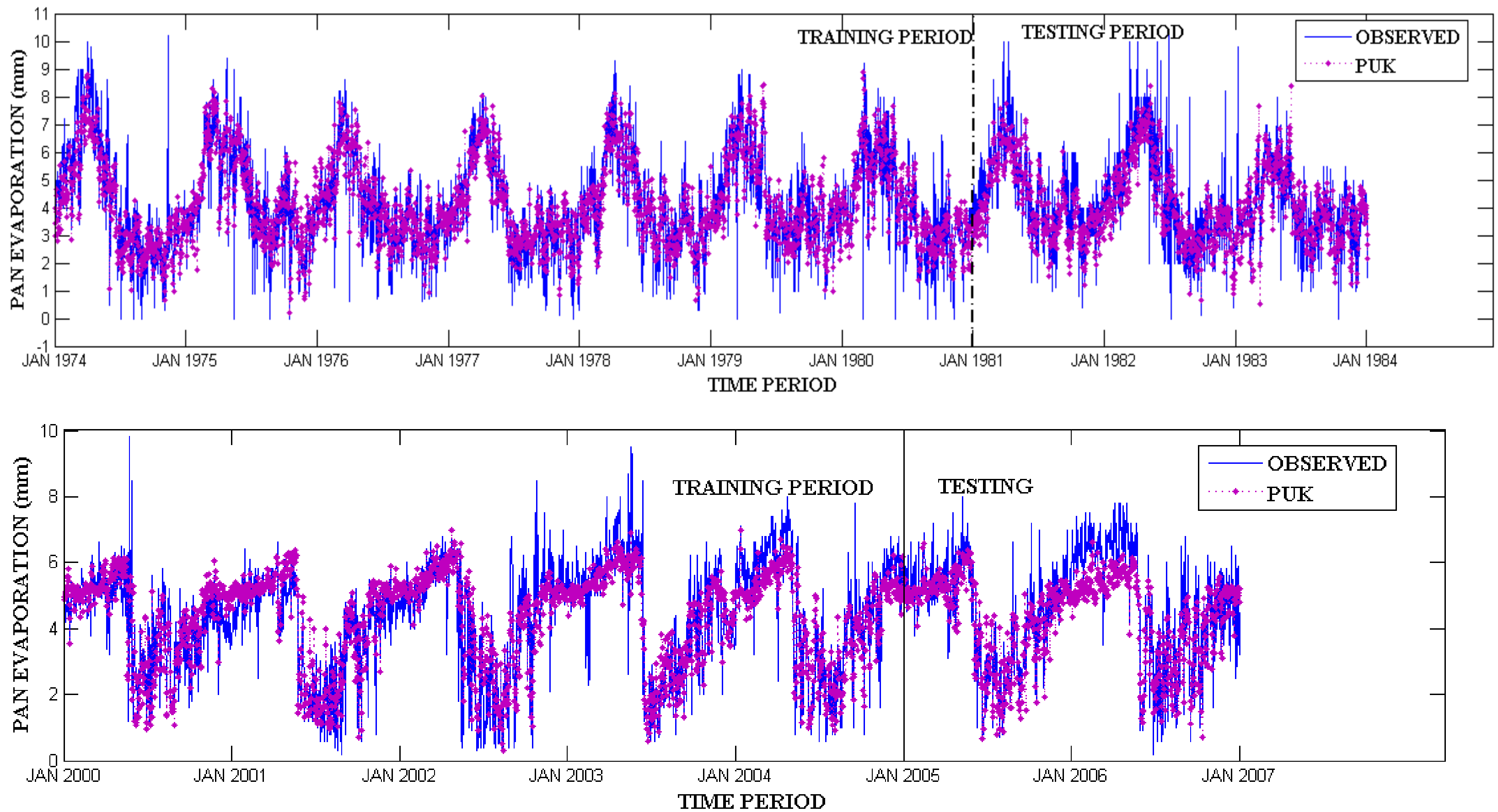


Figure 4.3: Pan evaporation estimation using PUK kernel for Bajpe (Top) and Bangalore station (Bottom).

In the **Figures 4.1, 4.2,** and **4.3**, it is clearly seen that temporal variation of pan evaporation isquite complex. However, SVR based RBF function learns the pattern better and capture the variations better in training as well as in testing phases. At higher pan evaporation value estimation almost matches with the observed pan evaporation. Neglecting the seasonal irregularity, RBF estimations are closer to observed pan evaporation values.

Discussing on these time series plots station wise, Polynomial kernel estimations are poor for the both the stations. It has failed to match the observed points as seen with the modeled results in both the phases as well as for both the stations. As discussed in earlier section RBF kernel estimations for Bajpe station in testing is better than training. This is seen in plots that, low and middle range pan evaporation values at testing phase are closer to observed values. But there is a slight change in the results trend exhibited with the Bangalore station. As RBF estimated values are closer to observed values in the training phase than testing phase. PUK kernel estimation, pattern was near similar to RBF for the estimated results of both the stations, except at few periods where there are noisy data.

Considering the summer season such as April to June in these recorded data period, the modeled pan evaporation are found underestimated compared to observed values during training as well as testing period. During rainy season, July to September, modeled pan evaporation is almost in close agreement or slightly overestimated than observed values. In general this trend is found to be uniform for all the kernel based functions such as RBF, PUK, and Polynomial under SVR. However, it was observed that the edge of superiority for RBF over other two kernel functions of PUK and Polynomial in modeling pan evaporation for Bajpe station.

Similar patterns of performance were observed in Bangalore station for the training period, but kernel estimations found to have some gaps between modeled and observed values.

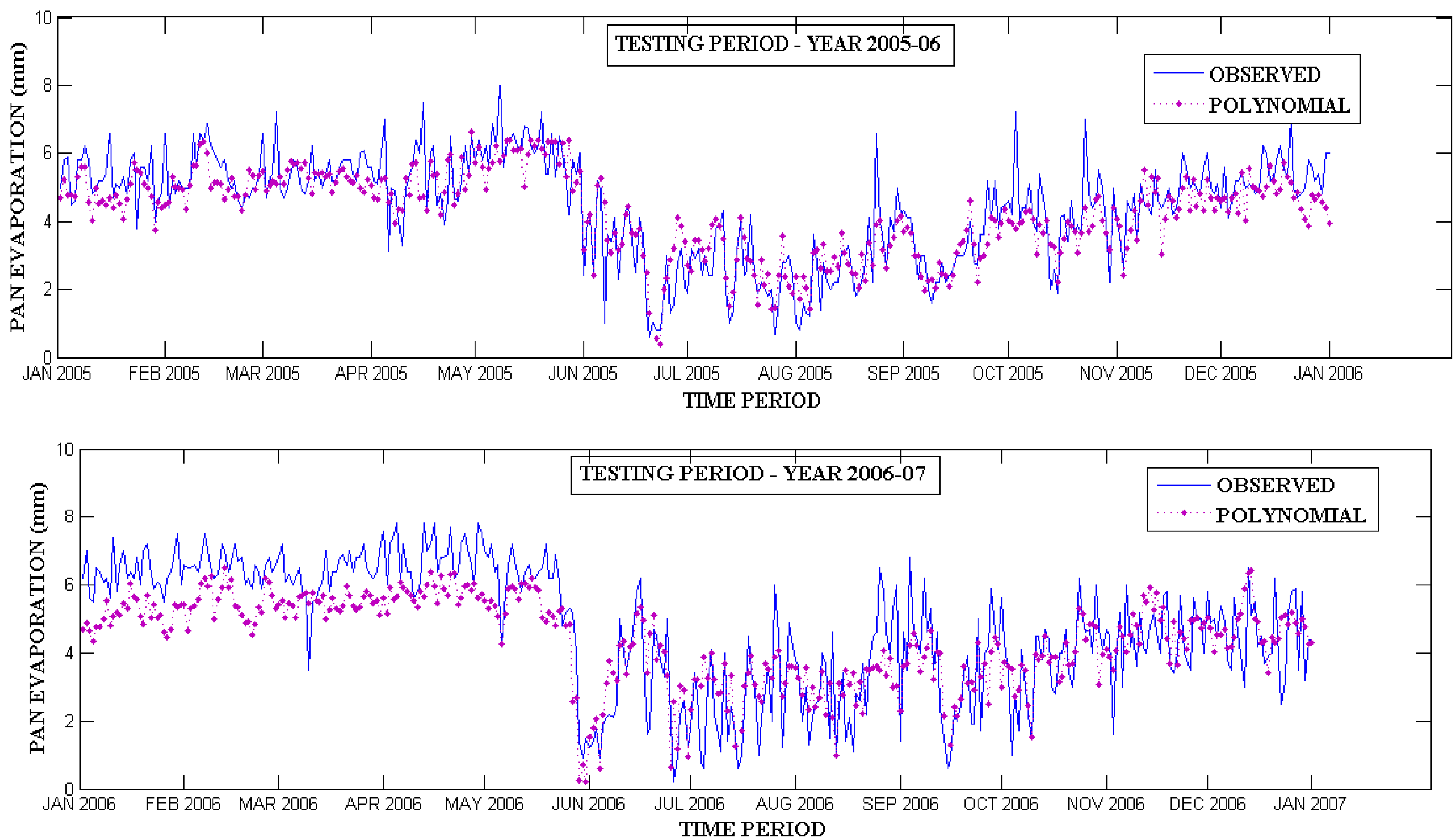


Figure 4.4: Daily pan evaporation estimation using SVR- Polynomial kernel for testing period 2005-06 & 2006-07 of Bajpe station.

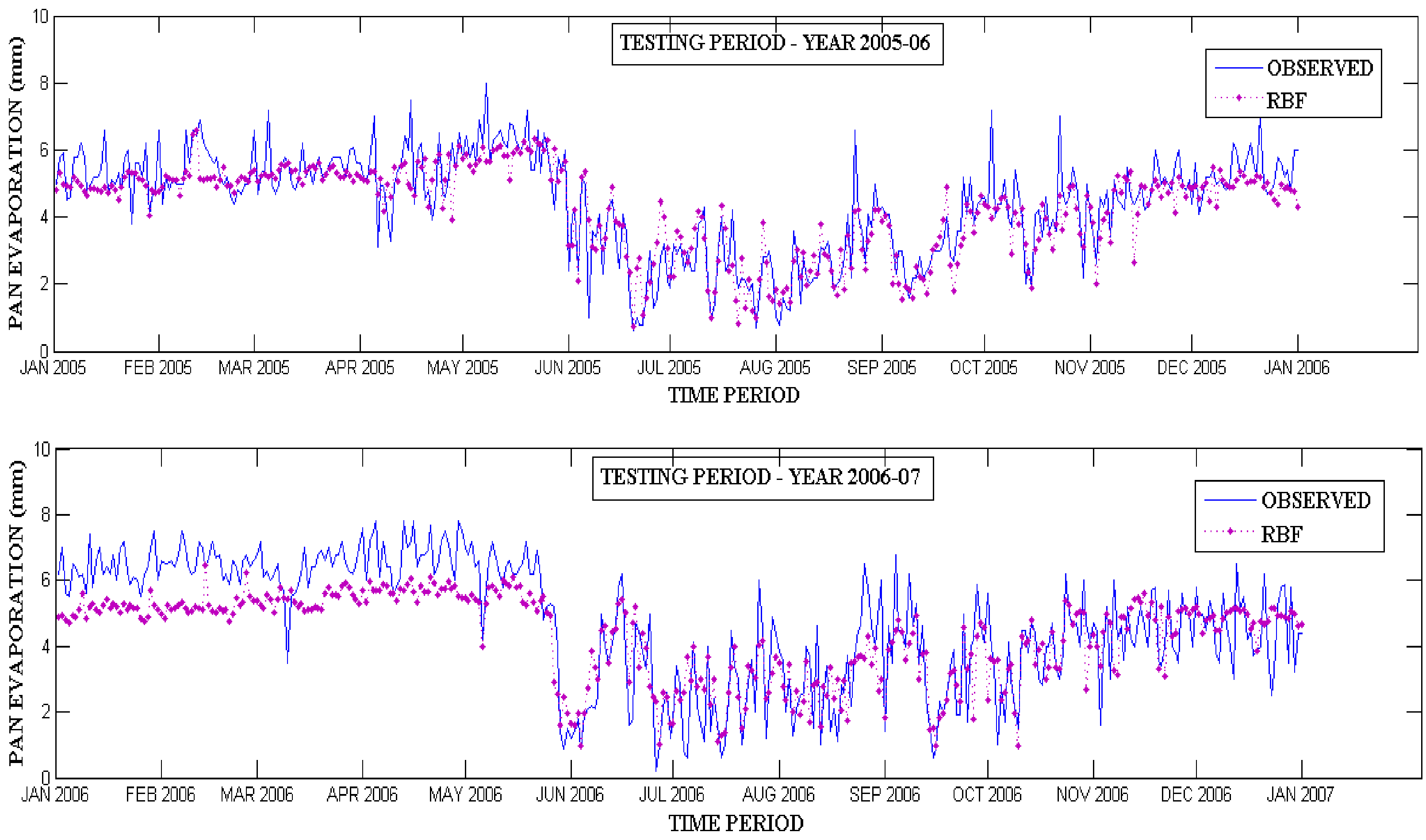


Figure 4.5: Daily pan evaporation estimation using SVR- RBF kernel for testing period 2005-06 & 2006-07 of Bajpe station.

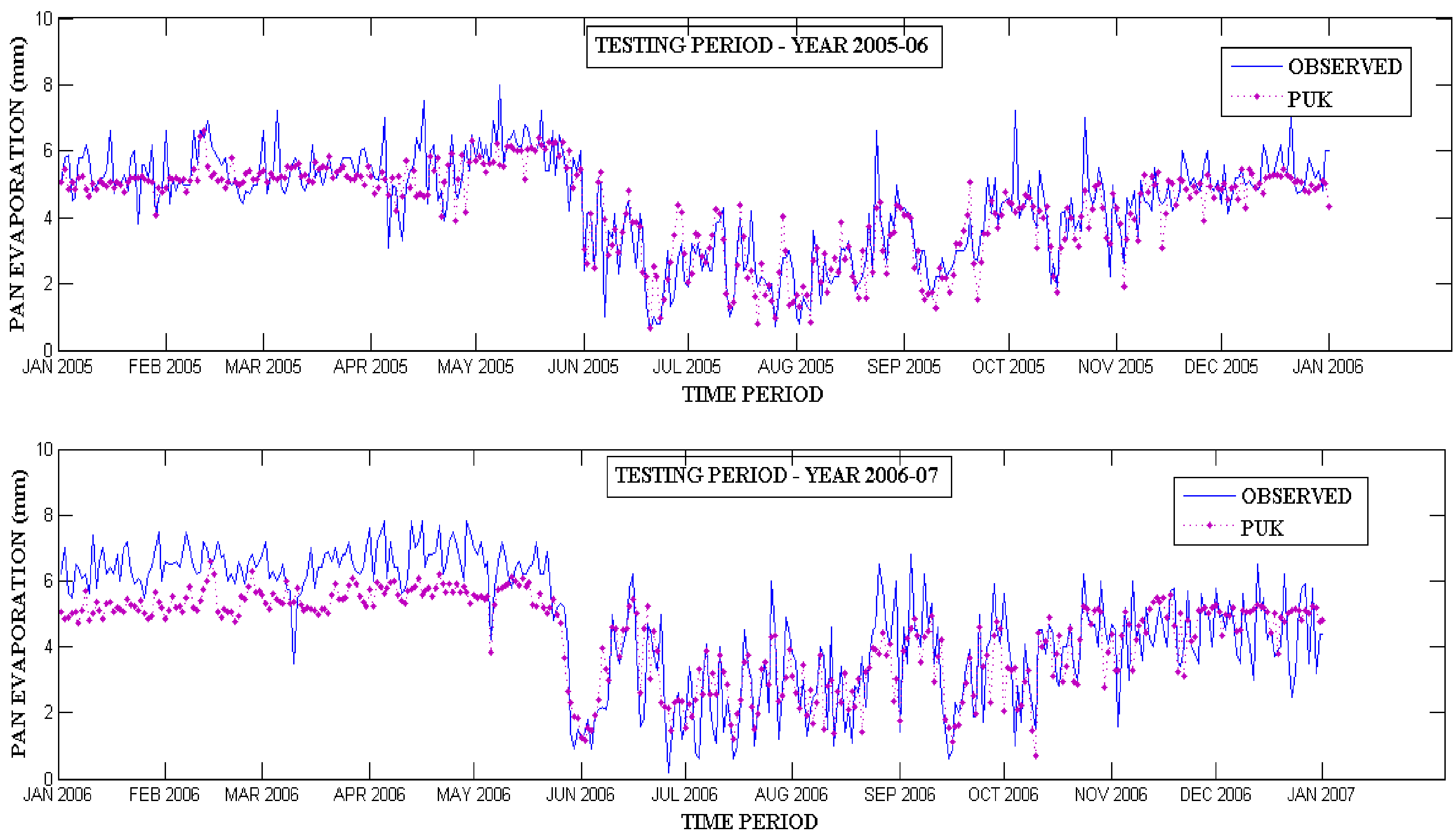


Figure 4.6: Daily pan evaporation estimation using SVR- PUK kernel for testing period 2005-06 & 2006-07 of Bajpe station.

The **Figure 4.4** represents time series plots of estimated pan evaporation using SVR- Polynomial kernel along with the observed pan evaporation for the testing periods of 2005-06 and 2006-07 respectively at station Bajpe. As seen in these plots Polynomial estimations are found to exhibit certain difference with the observed pan evaporation values for almost entire testing periods. Polynomial kernel estimated values are below par for the period 2006-07, particularly during the crucial periods such as January to May when the evaporation will be more. Overall the estimates computed by Polynomial kernels were underestimated the observed values for the majority of the testing periods mentioned.

The **Figure 4.5** shows SVR- RBF kernel estimated pan evaporation against the observed pan evaporation time series plots for the testing periods of 2005-06 and 2006-07 respectively at station Bajpe. For the seasonal periods where the observed pan evaporation is more RBF found to provide fair, accurate estimations particularly for the period 2005-06. Overall the estimated values nearly trace the observed values. But for the time series plot of 2006-07 initial pan evaporation values ranging between 3 and 7 RBF kernel under-estimates the estimated values over observed pan evaporation values during the period January 2006 to June 2006. However, there is improved accuracy in the estimations for the other ranges of pan evaporation values.

In the **Figure 4.6,** SVR- Puk kernel estimated pan evaporation along with the observed pan evaporation time series plots for the testing periods of 2005-06 and 2006-07 respectively at station Bajpe. The Puk estimation found almost closer to the one provide by RBF. As seen previously in RBF kernel plots, PuK also failed to accurately estimate the observed pan evaporation values during the crucial period of January 2006 to June 2006. In continuation,PuK estimation was better for low range pan evaporation values than peak values of pan evaporation.

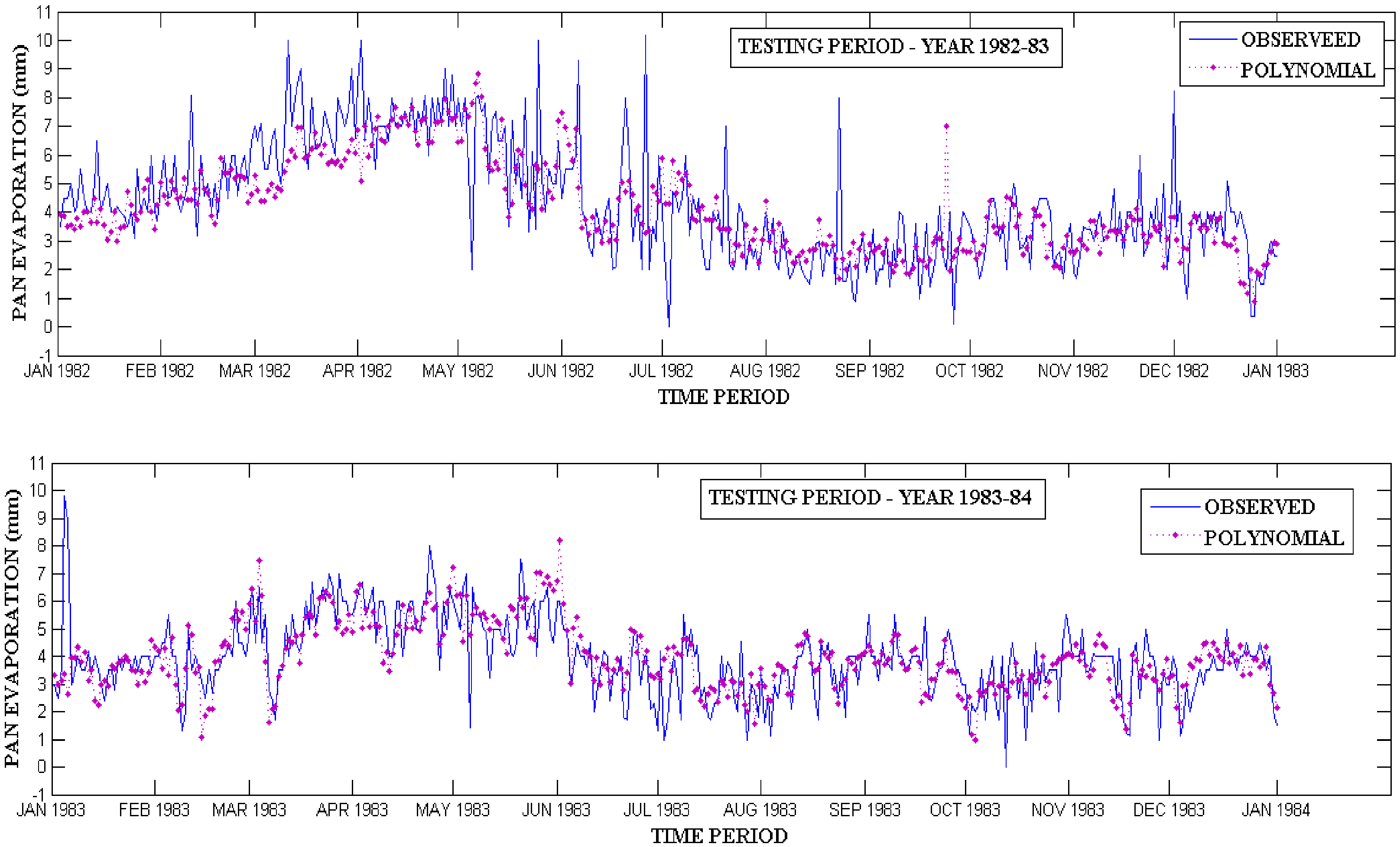


Figure 4.7: Daily pan evaporation estimation using SVR- Polynomial kernel for testing period 1982-83 & 1983-84 of Bangalore station.

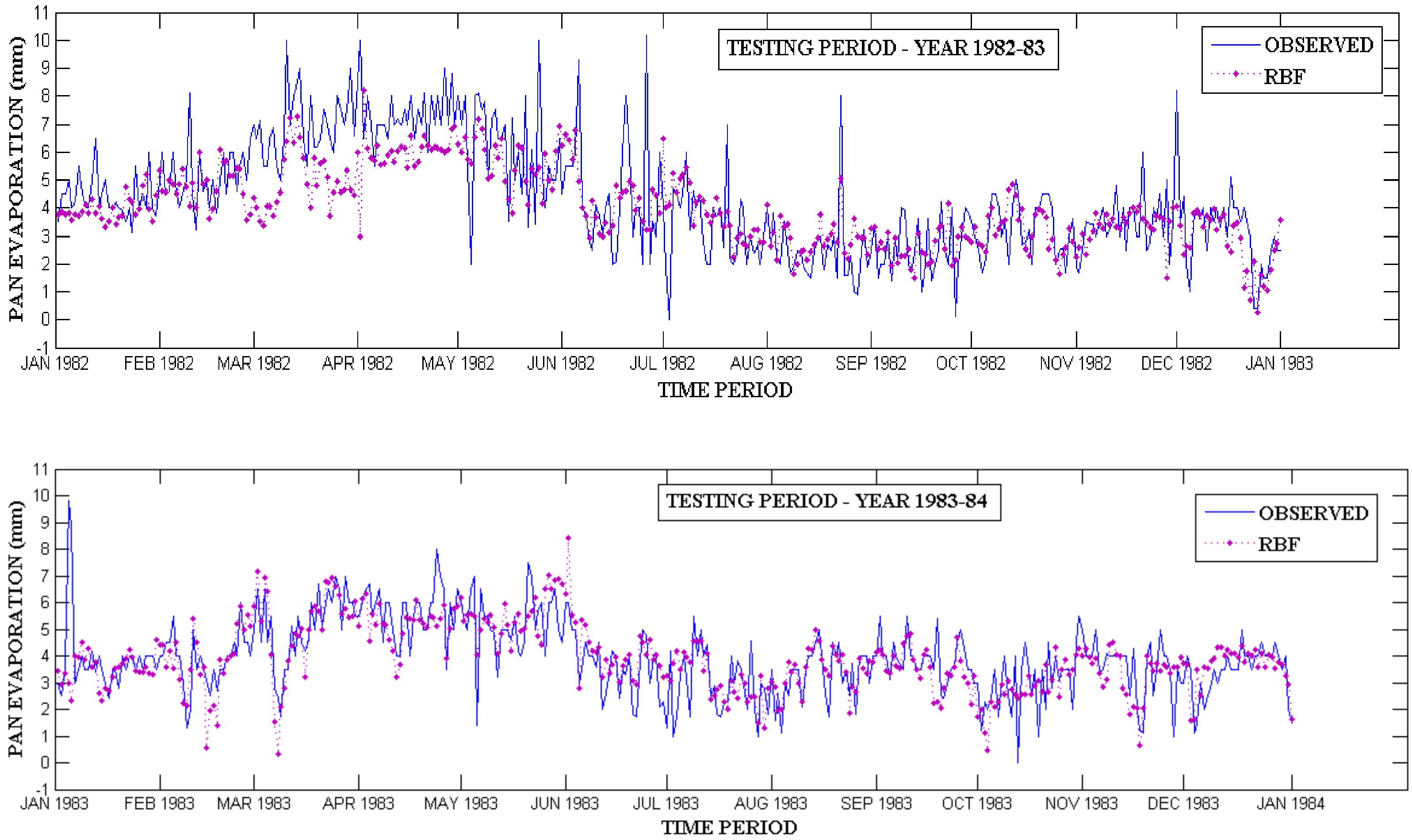


Figure 4.8: Daily pan evaporation estimation using SVR- RBF kernel for testing period 1982-83 & 1983-84 of Bangalore station.

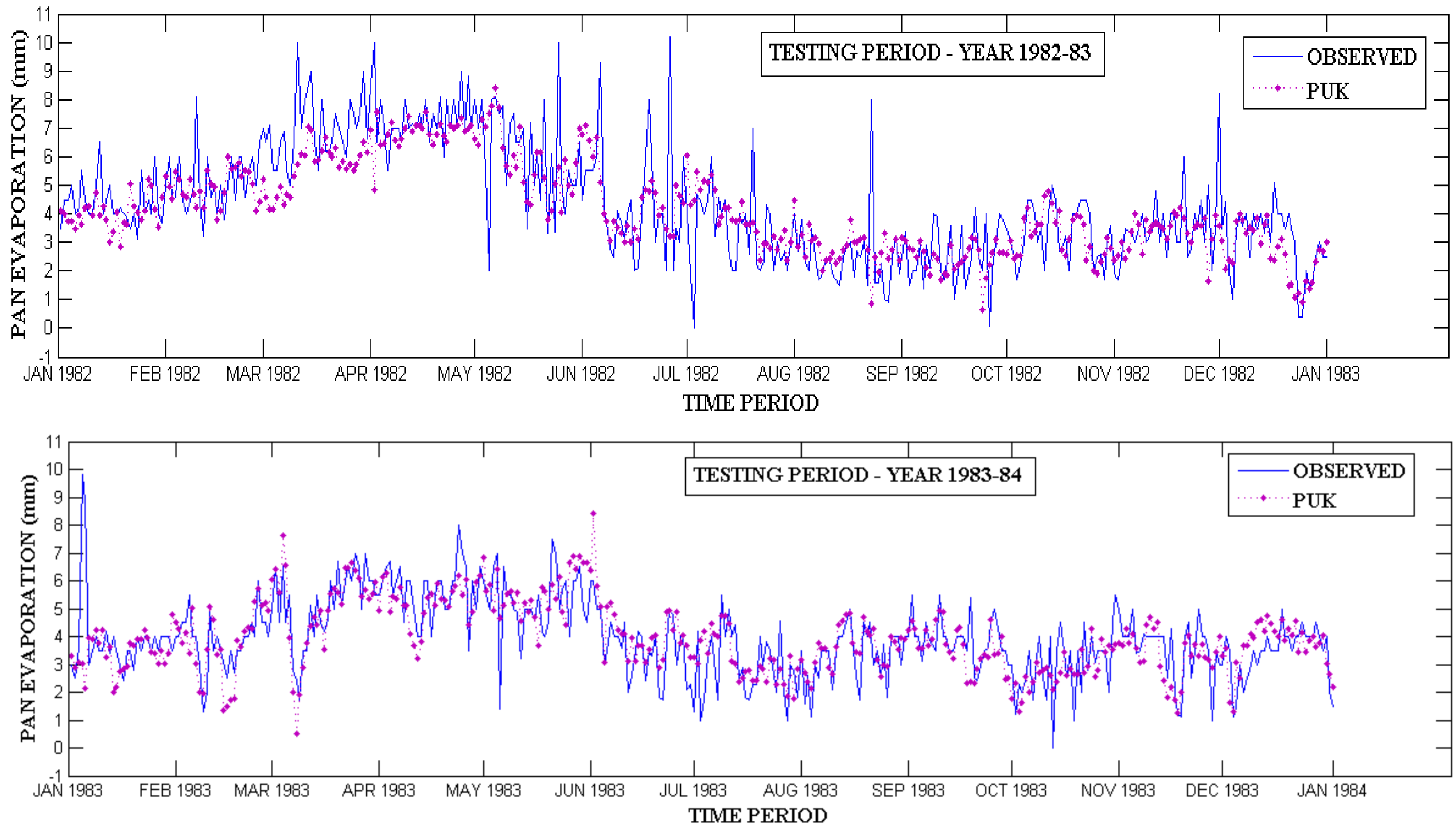


Figure 4.9: Daily pan evaporation estimation using SVR- RBF kernel for testing period 1982-83 & 1983-84 of Bangalore station.

The **Figure 4.7** displays time series plots of estimated pan evaporation using SVR- Polynomial kernel along with the observed pan evaporation for the testing periods of 1982-83 and 1983-84 respectively at station Bangalore. The testing phase estimations provided by all the three SVR kernel estimations for the Bangalore station are not quite satisfactory than Bajpe. However, it is seen in these plots that, improved accuracy in estimations for the crucial periods such as January to May which are very important and essential for various water resources field applications.Taking into considerations of results testing periods wise, performance trend is quite satisfactory for the period 1983-84 than 1982-83. This is because non linearity in the observed data was the minimum for the period 1983-84 than for the period 1982-83.

The **Figure 4.8** shows SVR- RBF kernel estimated pan evaporation against the observed pan evaporation time series plots for the testing periods of 1982-83 and 1983-84 respectively at station Bangalore. RBF estimations were better in comparison to remaining two kernel estimations for the testing periods. The modeled results are almost underestimated as compared to observed values for the majority portions of these testing periods. The estimates are matching the observed pan evaporation for the period 1983-84. The peak value in the periods of 1982-83 was not well captured by RBF estimations.

In the **Figure 4.9**, SVR- Puk kernel estimated pan evaporation along with the observed pan evaporation time series plots for the testing periods of 1982-83 and 1983-84 respectively at station Bangalore. Performance trend is almost comparable to the RBF. The limitations of SVR kernels in capturing highly nonlinear pan evaporation trend is once again displayed in the PUK estimation plots. The PUK estimations exhibit small gaps with observed data for middle and low range pan evaporation values.

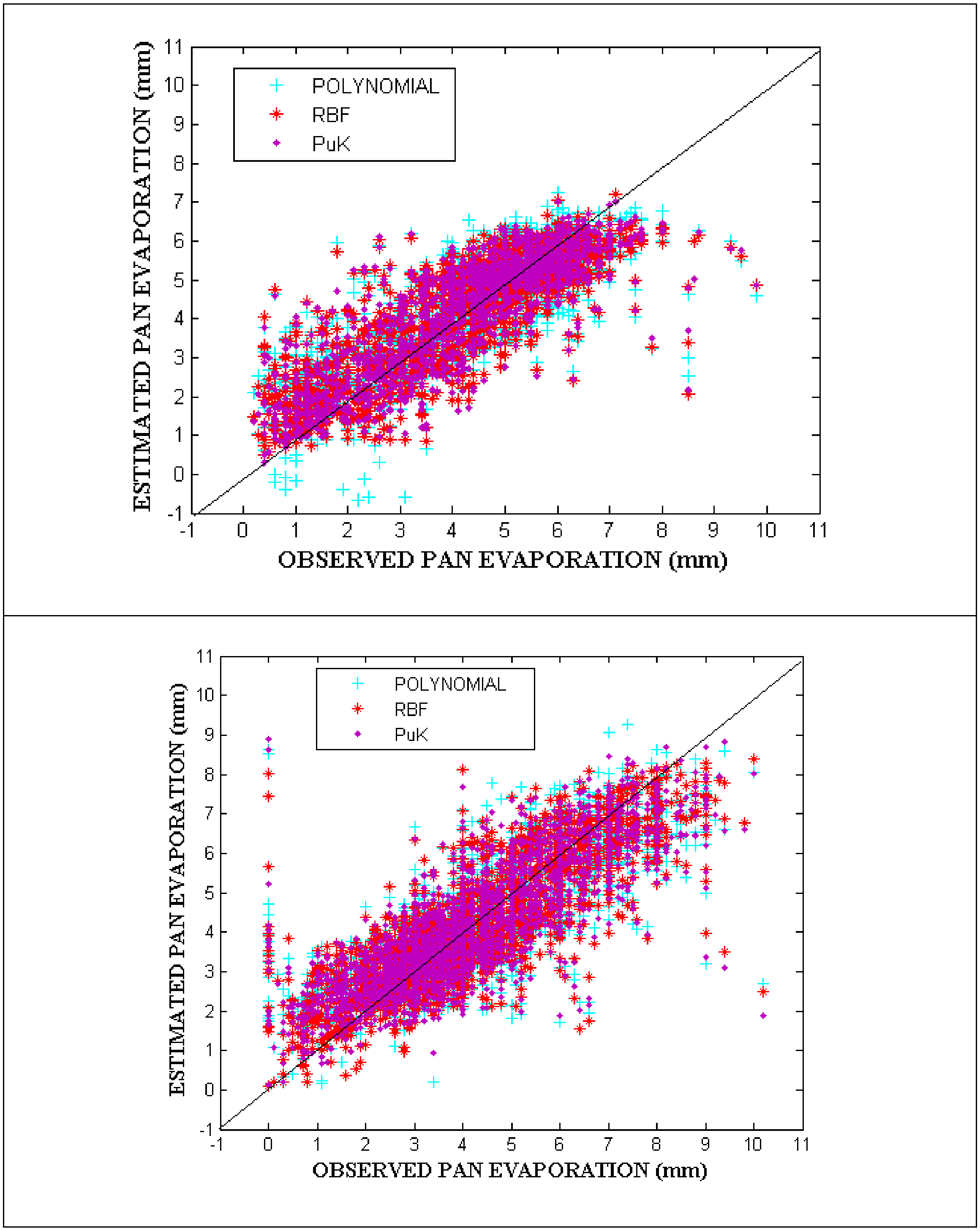


Figure 4.10: Scatter plot between SVR kernels estimated and observed daily pan evaporation training data for Bajpe (top) and Bangalore (bottom) stations.

In the scatter plots shown in **Figure 4.10**, the pan evaporation estimates of RBF seem to be less scattered. The RBF model's estimates are closer to the ideal line than those of the other models, especially for the mean and peak pan evaporation values as observed with both the stations estimates. Although RBF estimates are near closer to observed pan values, Polynomial kernel values with lower range are very much scattered from the ideal line with Bangalore station estimates than Bajpe station estimates. The PUK estimates compete better with RBF and stay as nearest closer to ideal line. It is also seen in these plots that, middle range pan evaporation estimates are much closer to observed values and are less scattered from the dividing line.

As far kernel estimations are concerned, in Bajpe station estimates are nearly seen overestimated the observed values, whereas Bangalore estimations are underestimating the observed values. This indicates that, the variations observed in the statistical range of attributes of these stations have influenced the model estimations. Especially ranges in the attributes of the Bangalore station were most varied.

Due to the larger data set containing daily pan evaporation data, these scatter plots seem to be clumsy to rank superiority of any kernel based function performance. Almost all the model results for training of both the stations are found with the similar performance level, except Polynomial kernel estimations particularly for low and very high range values. Also for both the stations, performance trends are similar as appeared in **Figure 4.10.**

It may be concluded that that all the kernel estimations are near closer to observed values in the training phase. The consistency of model performance in dealing with two contrasting statistical data sets of meteorological attributes was superior and flexible with RBF kernel based estimations and also found robustness capability.

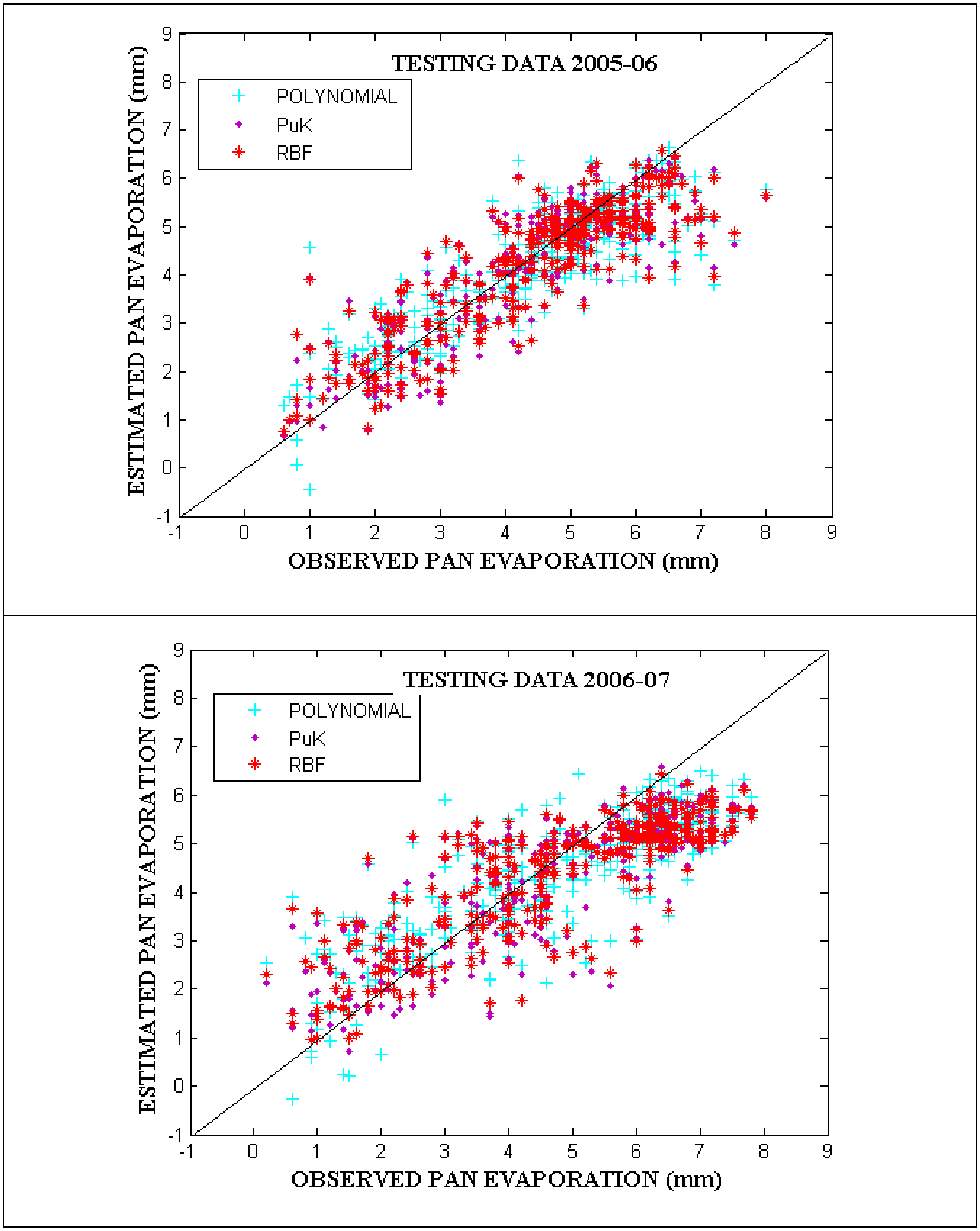
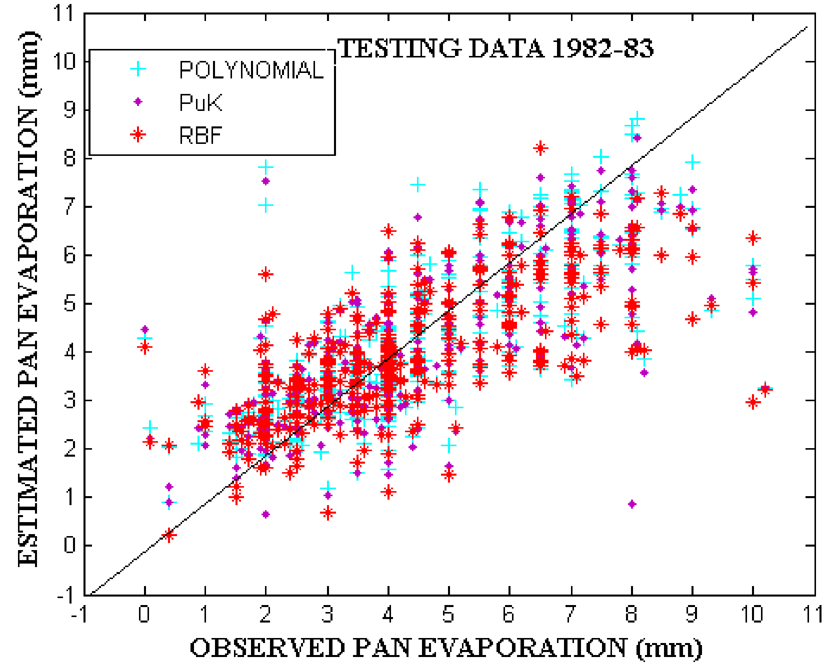


Figure 4.11: Scatter plot between estimated and observed daily pan evaporation testing data for Bajpe station.



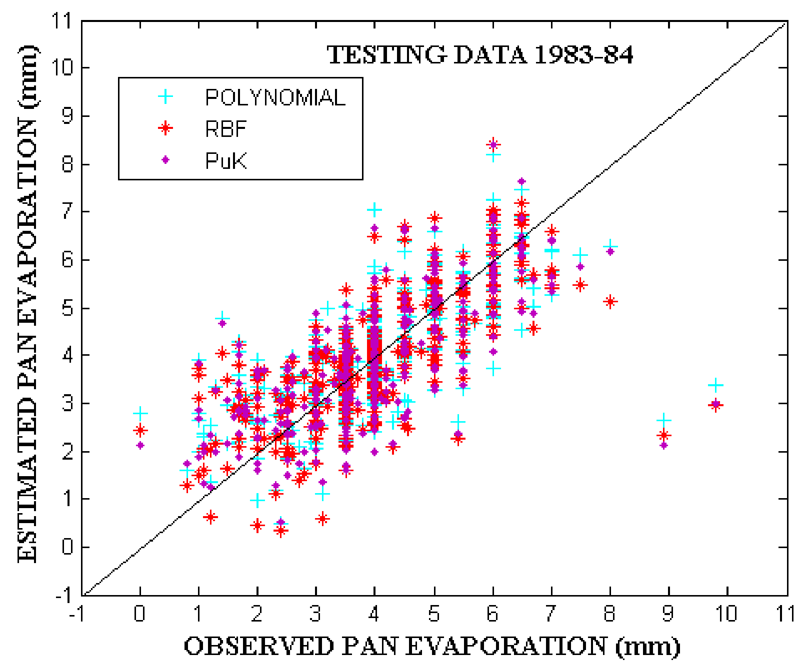


Figure 4.12: Scatter plot between estimated and observed daily pan evaporation testing data for Bangalore station.

The **Figure 4.11** and **4.12** represent scatter plots of testing data of Bajpe and Bangalore. Scatter plots of the training period, which were clumsy could not able to establish the superiority rank of the kernel estimations. In testing period plots the data points are limited. It becomes possible to distinguish the performance of kernel estimations with each other.

Considering the Bajpe station scatter plot of period 2005-06 in **Figure 4.11**, it is observed that, the scattering of estimated points from the ideal line is more for the Polynomial kernel estimated values and most of them found to be underestimated. RBF kernel estimates are close to the ideal line with better accuracy than PUK kernel. PUK estimations are less scattered as compared to polynomial kernel, but found to be underestimated observed values. On an average for middle range pan evaporation values, all three kernels estimated values are in better agreement with observed values. For The range 6 to 9 Polynomial and PUK estimated values scatters widely from the ideal line.

The Scatter plot for the testing period 2006-07 of Bajpe station shows the variable accuracy of these three kernel estimations. The polynomial and PUK estimated values more deviated from the observed points than RBF.

In **Figure 4.12**, it is found that for the testing period of 1982-83 estimated values of Polynomial kernel are more scattered for almost all the range of pan evaporation. RBF estimates are quite good for the range 2 to 4. But as the range exceeds 4, the accuracy decreases and estimated values show more deviation with observed data. Taking consideration of the testing period 1983-84, the estimated values are almost balanced and show equal spacing on either side with respect to the ideal line. This indicates that at some periods, the model has overestimated the observed values and vice versa.

4.3 RESULTS WITH DISCRETE WAVELET TRANSFORM SUPPORT VECTOR REGRESSION (DWT- SVR)

The discrete wavelet support vector regression (DWT- SVR) models are obtained combining two methods, DWT and SVR. The DWT-SVR model is an SVR model, which uses coefficients generated from decomposing original data and recompiling the data to feed in SMO- SVR. For the DWT- SVR model inputs, signals split into a detail and an approximation. The approximation obtained from the first-level is split into new detail and approximation, and this process is repeated. Because of the fact that wavelet transform decomposes only the approximations of the signal, it may cause problems while applying wavelet transform in certain applications where the important information is located in higher frequency components (Shinde et al. 2013).

For wavelet analysis, Discrete Wavelet Transformation (DWT) is used and Daubechies wavelet order-4 (db4) was selected as a mother wavelet. The selected mother wavelet 'db4' is a simplest wavelet having only 3 wavelet filter coefficients with exact reconstruction possibilities. To get the decomposed wavelet coefficients here, various decomposition levels has been tried (L1 to L8) but only Daubechies wavelet mother function with level 3 was showing better results when fed as inputs to SVM on trial and error basis to enhance the performance.

As discussed in the previous section the parameter selection is very essential to build efficient models. A grid search method once again used to obtain desired parameters which then fed to DWT-SVR models building process.

Table 4.5: Statistical indices of DWT-SVR RBF models with combinations of mother wavelets of Bajpe station

Mother Wavelet's	**DWT- SVR RBF Kernel Training period**				**DWT- SVR RBF Kernel Testing period**			
	RMSE (mm)	MAE (mm)	CC	NSE	RMSE (mm)	MAE (mm)	CC	NSE
DB4 level 1	0.568	0.395	0.845	0.969	0.625	0.381	0.867	0.963
DB4 level 2	0.553	0.379	0.899	0.974	0.579	0.385	0.875	0.978
DB4 level 3	**0.448**	**0.363**	**0.924**	**0.982**	**0.525**	**0.410**	**0.953**	**0.991**
DB4 level 4	0.460	0.382	0.884	0.976	0.615	0.422	0.854	0.981
D4 level 5	0.498	0.251	0.870	0.980	0.621	0.357	0.845	0.984
Haar 3	0.451	0.368	0.909	0.992	0.517	0.404	0.934	0.987
Haar 4	0.631	0.388	0.845	0.974	0.691	0.429	0.830	0.989
Haar 5	0.651	0.377	0.885	0.965	0.712	0.466	0.902	0.974

Table 4.6: Statistical indices of DWT-SVR RBF models s with combinations of mother wavelets of Bangalore station

Mother Wavelet's	DWT- SVR RBF Kernel Training period				DWT- SVR RBF Kernel Testing period			
	RMSE (mm)	MAE (mm)	CC	NSE	RMSE (mm)	MAE (mm)	CC	NSE
DB4 level 1	0.375	0.287	0.968	0.996	0.524	0.373	0.937	0.993
DB4 level 2	0.263	0.190	0.983	0.998	0.419	0.282	0.948	0.995
DB4 level 3	**0.240**	**0.166**	**0.985**	**0.998**	**0.334**	**0.223**	**0.959**	**0.997**
DB4 level 4	0.260	0.182	0.981	0.996	0.340	0.512	0.914	0.993
D4 level 5	0.290	0.201	0.975	0.997	0.472	0.335	0.906	0.994
Haar 3	0.253	0.173	0.980	0.996	0.446	0.332	0.957	0.997
Haar 4	0.251	0.177	0.978	0.997	0.449	0.335	0.951	0.995
Haar 5	0.278	0.177	0.976	0.998	0.560	0.374	0.917	0.992

As discussed in the previous chapter, Daubechies mother wavelet functions of level 1 to 5 are employed along with Haar wavelet function of level 3, 4 and 5. From the **Table 4.5** and **4.6**, out of various levels of Daubechies of order 4, level 3 provides optimum result. IN terms of RMSE, level 3 yield lowest value of 0.448 out of other levels of the same mother wavelet function as well as Haar. Also the MAE was 0.363, CC was 0.924 with NSE 0.982 confirms the superiority of DB4 level 3 as seen in the computed result of the training period. The trend of performance continues to be same in testing phase also. At the same time Haar wavelet function also produced better results in training period but failed to show the same performance in testing period.

The best Daubechies mother wavelet functions of order 4 and level 3 was considered as base mother function and coupled with three types of support vector regression kernels to strengthen the work, In the conventional method of SVR, RBF kernel function produced optimum results. From the **Tables 4.5** and **4.6**, RBF kernel once again outperformed the other two competent kernels and produced optimum results. This strengthens the accuracy level of RBF kernel functions of SVR models.

Table 4.7: Statistical indices of DWT-SVR models with combinations of kernel functions of Bajpe station

Mother Wavelet	DWT- SVR POLYNOMIAL KERNEL				DWT- SVR RBF KERNEL				DWT- SVR PUK KERNEL			
	RMSE (mm)	MAE (mm)	CC	NSE	RMSE (mm)	MAE (mm)	CC	NSE	RMSE (mm)	MAE (mm)	CC	NSE
Db4 level 3 (Training Model)	0.443	0.356	0.926	0.978	**0.448**	**0.363**	**0.924**	**0.982**	0.318	0.225	0.962	0.991
Db4 level 3 (Testing Model)	0.554	0.427	0.950	0.984	**0.525**	**0.410**	**0.953**	**0.991**	0.518	0.400	0.921	0.989

Table 4.8: Statistical indices of DWT-SVR models with combinations of kernel functions of Bangalore station

Mother Wavelet	DWT- SVR POLYNOMIAL KERNEL				DWT- SVR RBF KERNEL				DWT- SVR PUK KERNEL			
	RMSE (mm)	MAE (mm)	CC	NSE	RMSE (mm)	MAE (mm)	CC	NSE	RMSE (mm)	MAE (mm)	CC	NSE
Db4 level 3 (Training Model)	0.353	0.269	0.968	0.996	**0.240**	**0.166**	**0.985**	**0.998**	0.243	0.170	0.979	0.996
Db4 level 3 (Testing Model)	0.447	0.321	0.947	0.995	**0.334**	**0.223**	**0.959**	**0.997**	0.505	0.320	0.951	0.993

Considering the Daubechies mother wavelet functions of order 4 and level 3 as base mother function, DWT coupled with three types of support vector regression kernels to strengthen the work. **Table 4.7 and Table 4.8** represent the DWT-SVR modeled results for training and testing periods of Bajpe and Bangalore station respectively.

Comparing the **Table 4.7** with **Table 4.1** related to modeled training results of Bajpe stations, the DWT-SVR results have shown enhanced performance in comparison to conventional SVR models. For SVR models, statistical indices were RMSE 0.941, MAE 0.687, CC 0.832 and NSE 0.977, but DWT-SVR RBF yielded RMSE 0.448, MAE 0.363, CC 0.924 and NSE 0.982 for the training period. This highlights the escalated performance of DWT SVR over SVR modeled pan evaporation estimations. Once again the RBF showed the leading performance in comparison to its competitor kernels. The testing results were even better with respect to training period models.

Similarly, there was highly improvised performance with DWT-SVR than SVR models for the Bangalore station as shown in **Table 4.8**. However, the only change observed with the model performance is that of testing period estimations. The testing period results were slightly down because of the possible seasonal irregularities were not fully captured by DWT-SVR models.

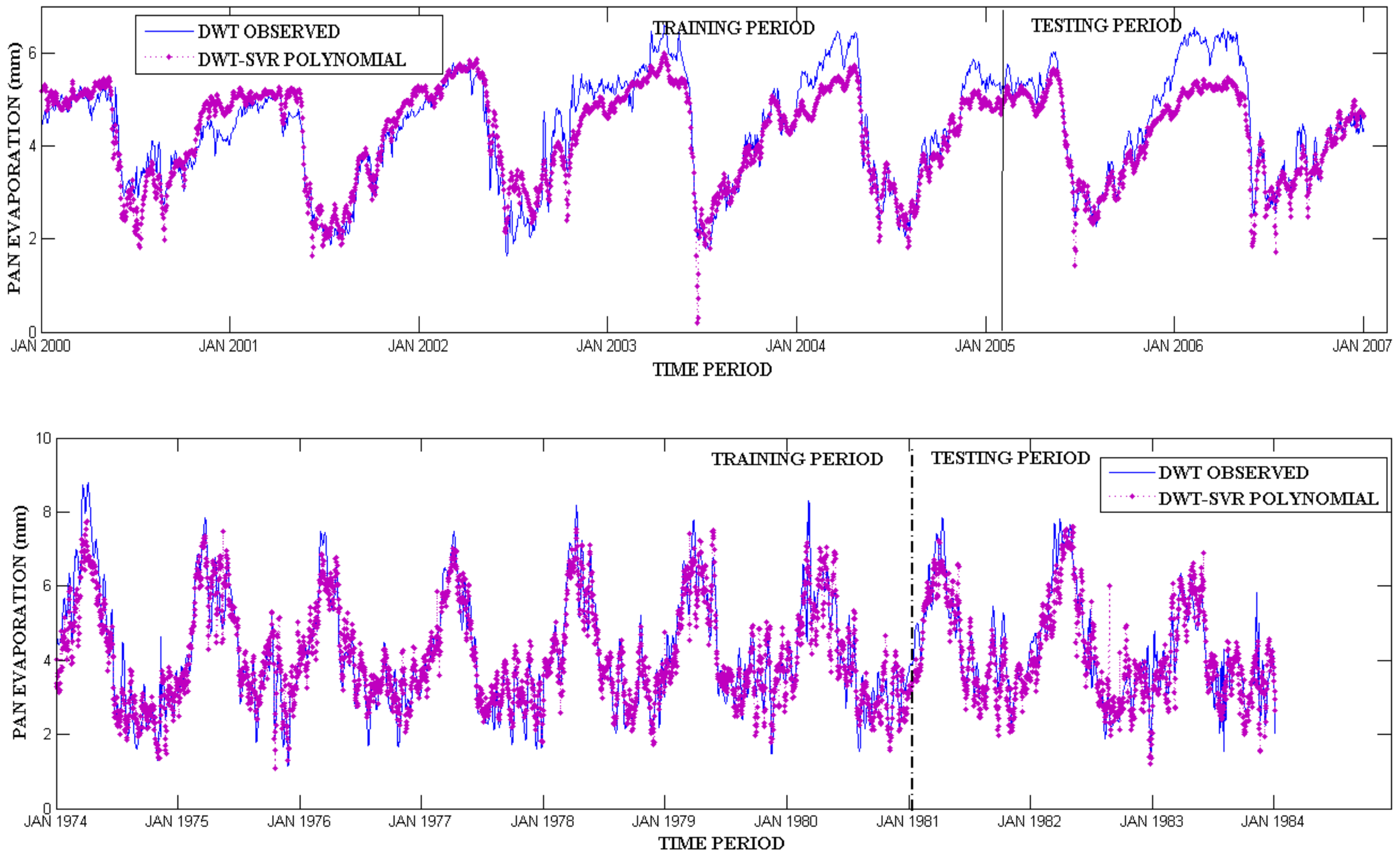


Figure 4.13: Daily pan evaporation estimation using DWT SVR- Polynomial kernel for Bajpe (Top) and Bangalore station (Bottom).

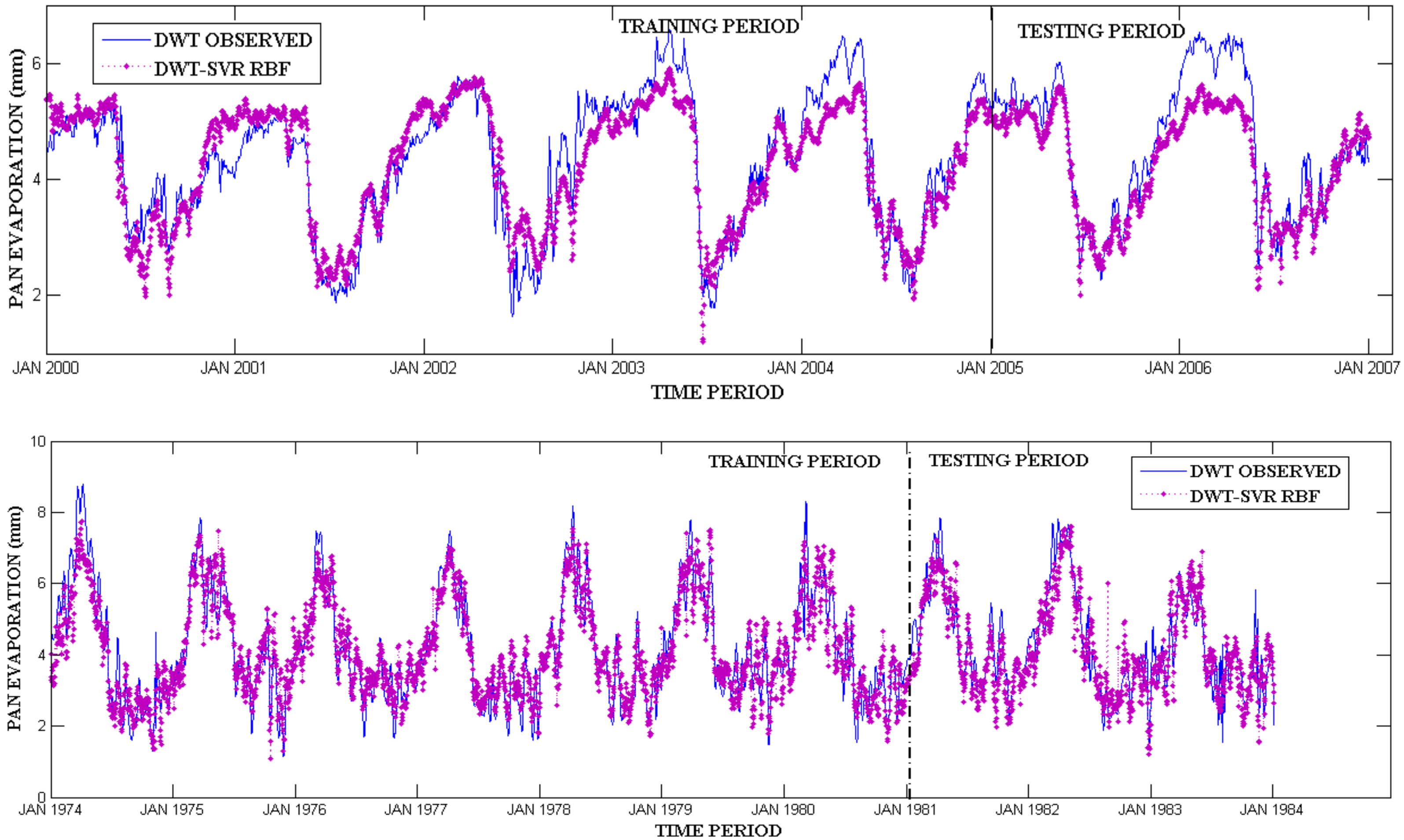


Figure 4.14: Daily pan evaporation estimation using DWT SVR- RBF kernel for Bajpe (Top) and Bangalore station (Bottom).

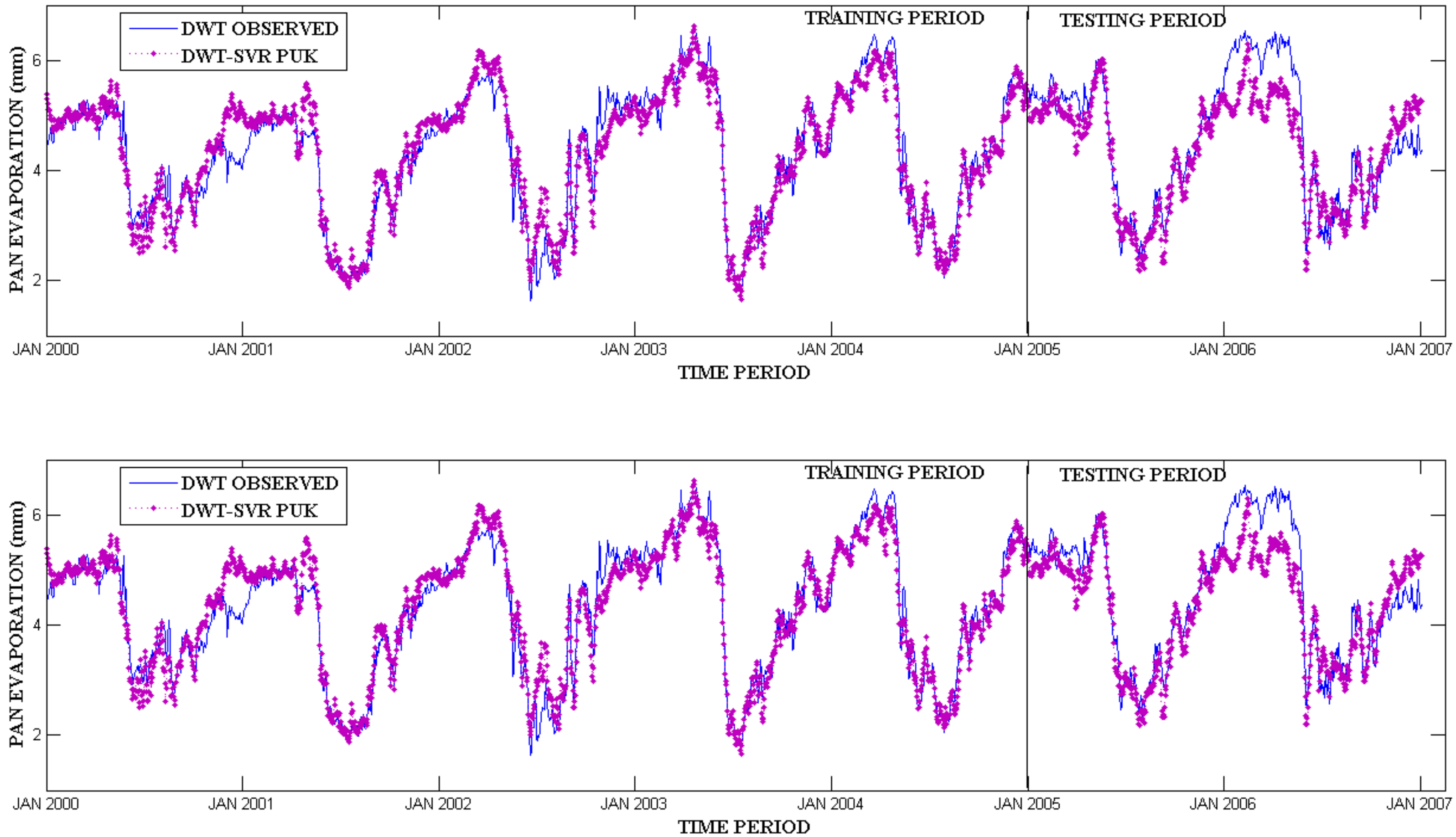


Figure 4.15: Daily pan evaporation estimation using DWT SVR- PUK kernel for Bajpe (Top) and Bangalore station (Bottom).

The results discussed in the **Tables 4.7 and 4.8** are visualized in the **Figure 4.13, 4.14, and 4.15.** As discussed in the previous chapter, the coefficients computed from the original attributes are used in model building. Due to removal of unwanted information from the data signal, model could able to yield better performance. However, it is to be seen in the plots that nonlinear pattern of pan evaporation is unaltered. The estimated pan evaporation results, and then plotted against the coefficients of the original pan evaporation.

Considering DWT-SVR model performance as displayed in time series plots, it is observed that RBF estimations outperform the Polynomial and PUK kernels. The RBF estimations are very accurate for the training and testing phases of Bajpe station. However, there is a slight decline in the testing results of Bangalore station. The discussion over such decline is made in the previous sections.

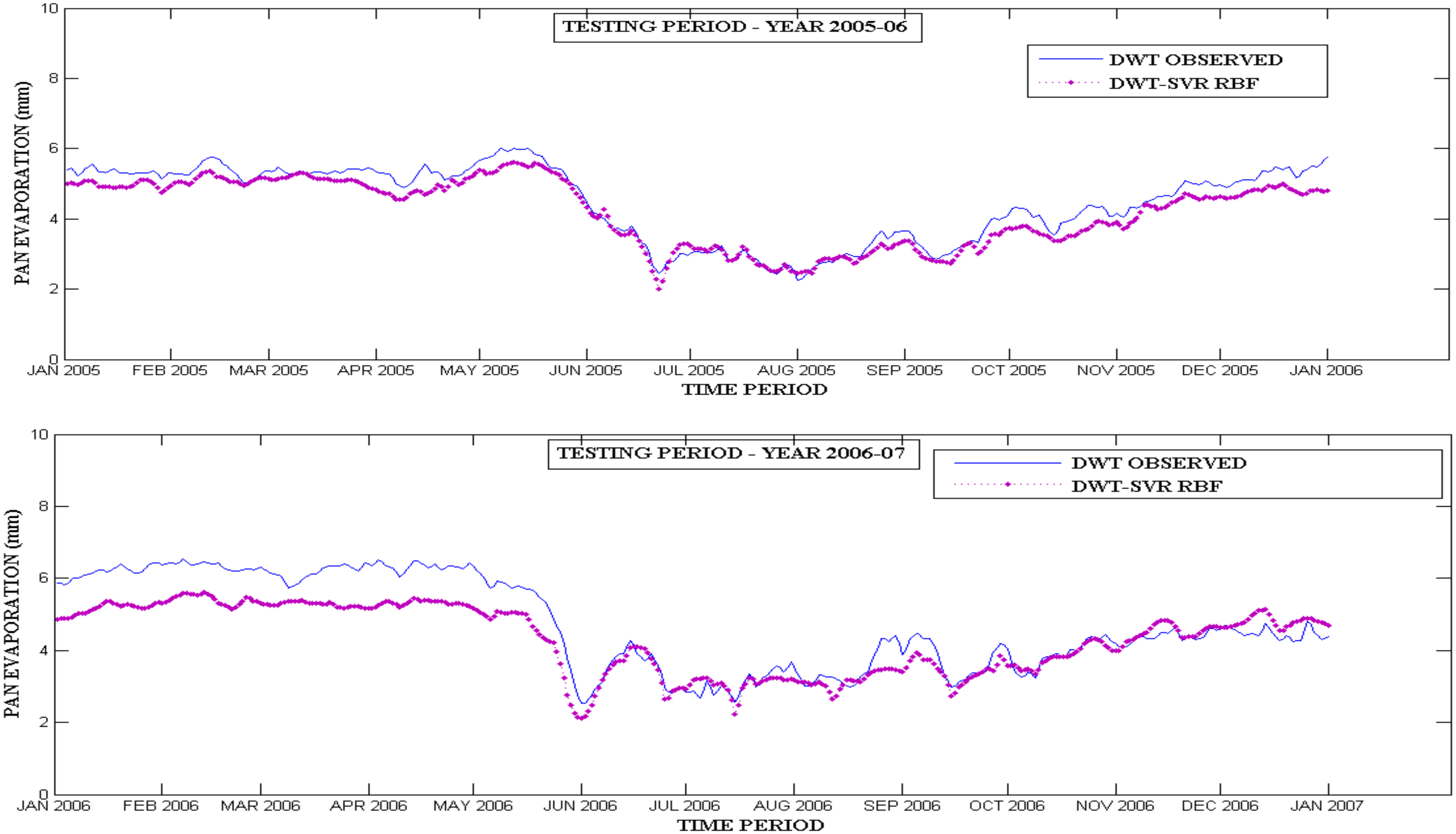


Figure 4.16: Daily pan evaporation estimation using DWT-SVR RBF kernel for testing period 2005-06 & 2006-07 of Bajpe station.

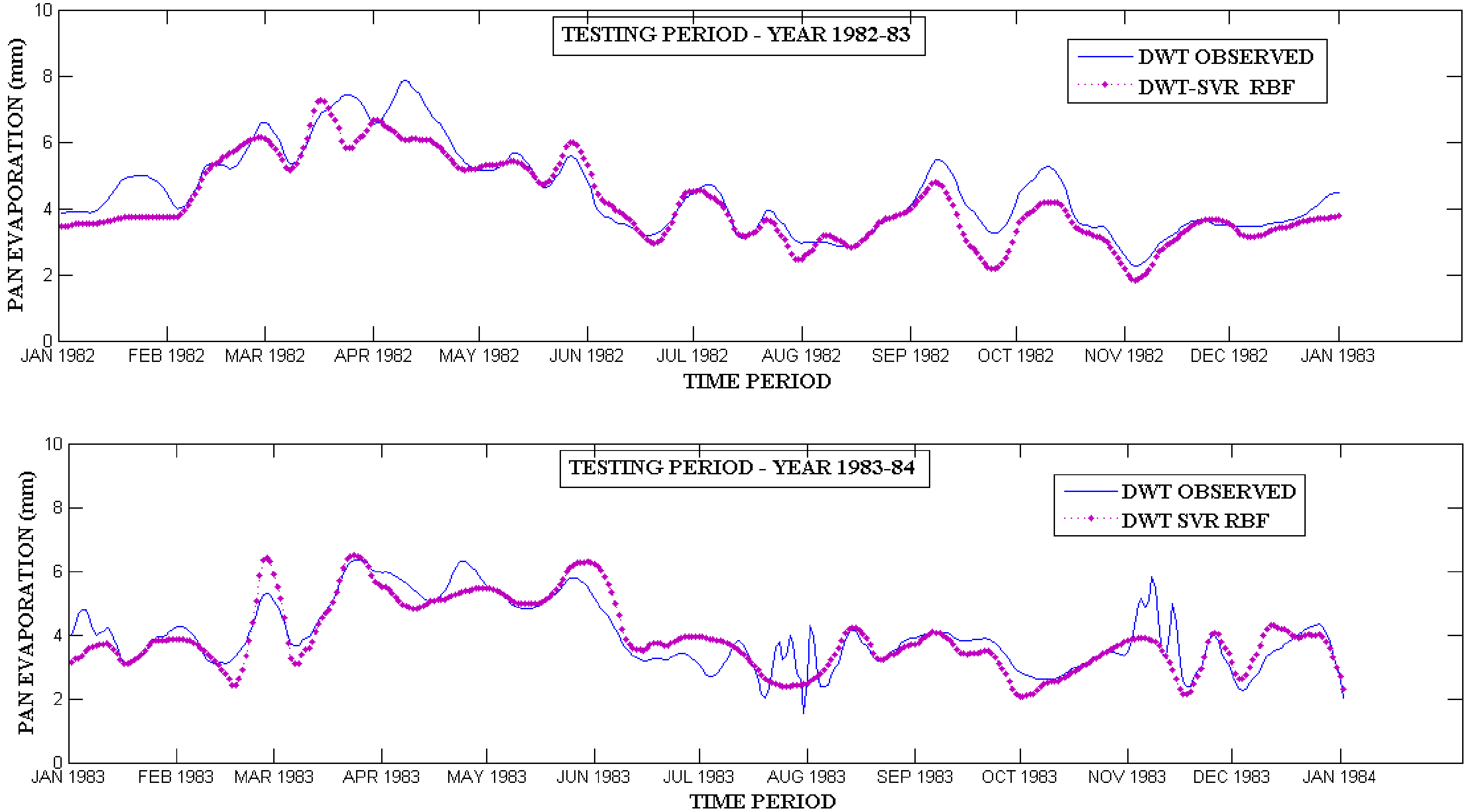


Figure 4.17: Daily pan evaporation estimation using DWT-SVR RBF kernel for testing period 1982-83 & 1983-84 of Bangalore station.

The time series plots of DWT-SVR RBF models developed for testing periods of 2005-06 and 2006-07 at the station Bajpe are displayed in **Figure 4.16**. For the period of 2005-06, the estimated values are very close to processed observed pan evaporation points through the span length. The DWT-SVR RBF model is able to provide accurate estimations for all the seasonal data periods. Considering the model performance for the period 2006-07, the model efficiency has slightly reduced for the period January to May 2006. But nevertheless for the rest of the period model performed better and estimated points are closer to the processed raw data.

The **Figure 4.17** represents the DWT-SVR RBF time series plot for testing periods of 1982-83 and 1983-84. In this figure it is observed that, seasonal variations are mapped closely by the DWT-SVR RBF model. The estimated values exhibit minimum deviations with the processed pan evaporation data for both the testing periods.

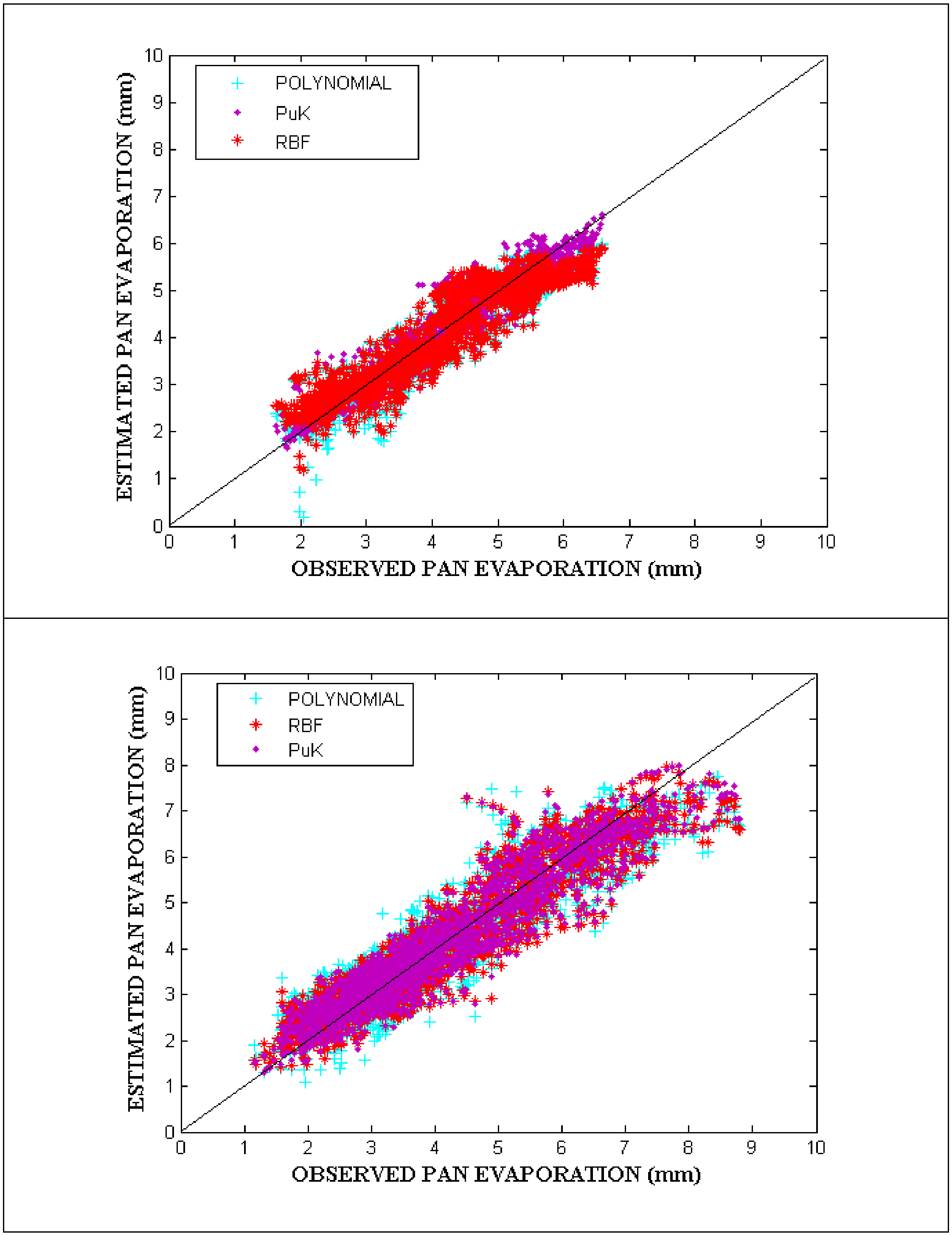


Figure 4.18: Scatter plot between DWT-SVR kernels estimated and observed daily pan evaporation training data for Bajpe (top) and Bangalore (bottom) stations.

Figures 4.18 represents the scatter plots of training data for the stations Bajpe and Bangalore stations. The DB 4-3 mother wavelet with three types of SVR kernel estimated values are plotted in the figures. The training data set for both the stations is longer, however the values are not much scattered and they stick closer to 45° line.

Considering the results station wise, Bajpe station results indicate that all three kernels estimated values are in a good agreement with observed values. The scatter plot seems to be clumsier due to lengthy data set, so it becomes difficult to rank the superiority among the kernel estimations. However, RBF values are closer to observed values. The kernel estimated values seems to be underestimating the observed values. More or less similar pattern of results found with Bangalore station, some differences may be seen in terms of estimations were more balanced in terms of under and overestimated observed values. Over all these scatter plots once again strengthen the estimation accuracy of RBF kernel function in comparison to the remaining two.

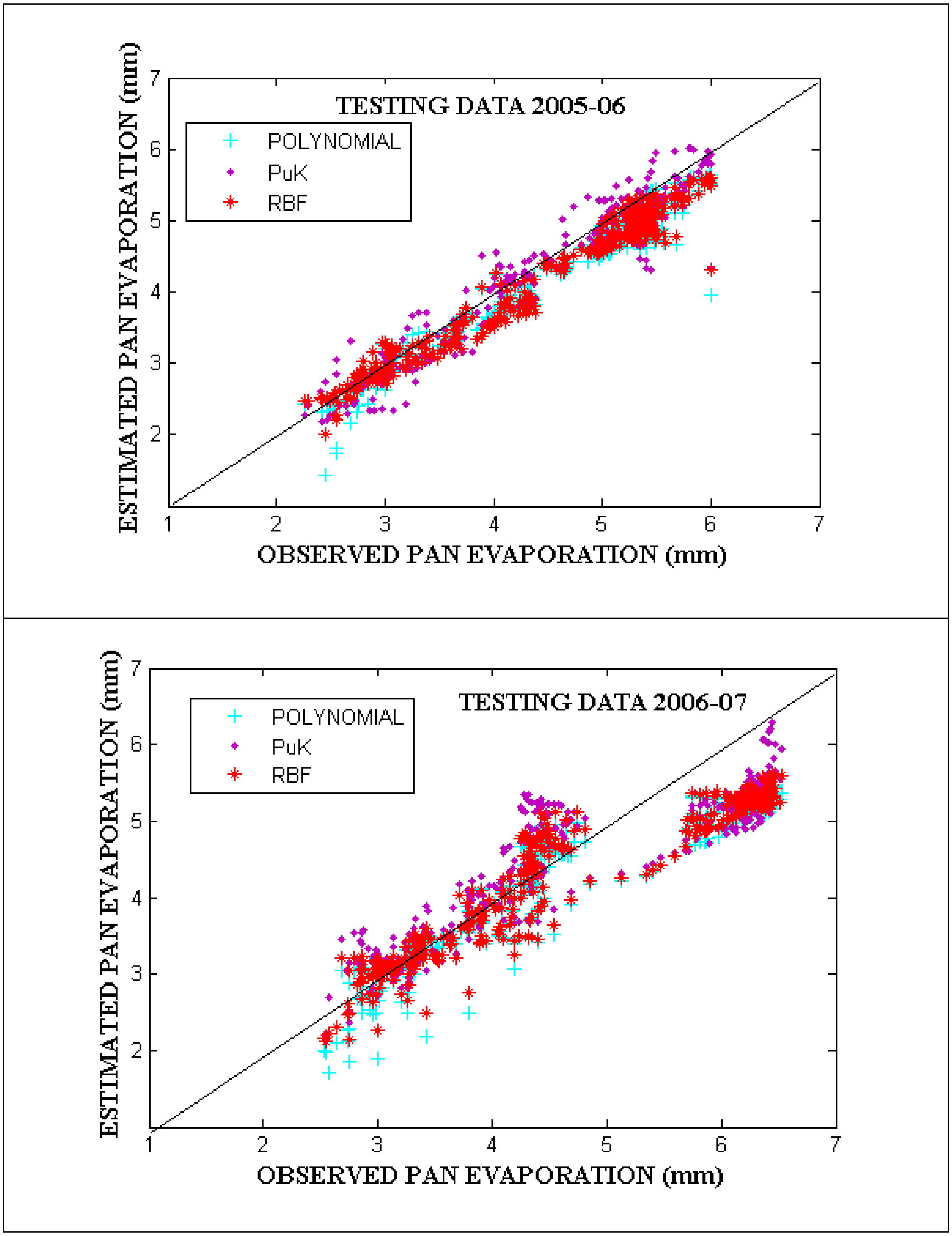


Figure 4.19: Scatter plot between DWT-SVR estimated and observed daily pan evaporation testing data for Bajpe station.

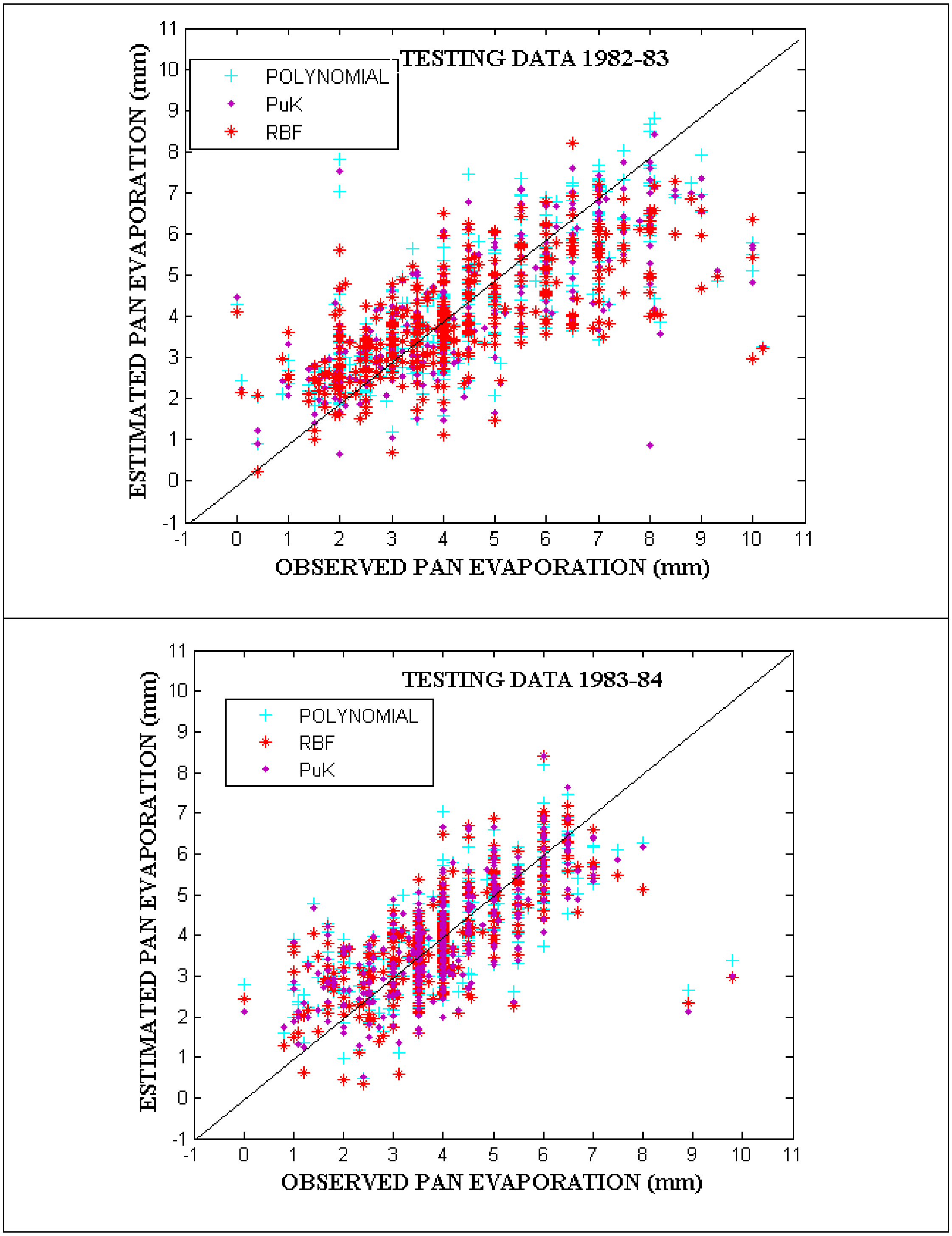


Figure 4.20: Scatter plot between DWT-SVR estimated and observed daily pan evaporation testing data for Bangalore station.

Figures 4.19 and 4.20 represent the scatter plots of testing data for the stations Bajpe and Bangalore stations. The DB 4-3 mother wavelet functions with three types of SVR kernel estimated values are plotted in the figures. Due to a data set limited to one year period only, some distinct observation can be made on these kernel estimations. RBF once again makes its remarks in providing closer values of real data. It has better potential to produce an accurate estimation of pan evaporation

Considering the results station wise, Bajpe station results indicate that all three kernels estimated values were of underestimated during testing periods of the year 2005 to 2006 and 2006 to 2007 as appeared in the figures above. From Bangalore station results, it is seen that estimations were more balanced in terms of under and overestimated observed values during the testing period of the year 1982 to 1983 and 1983 to 1984 as observed in the figures above.

CHAPTER 5

SUMMARY AND CONCLUSIONS

5.1 SUMMARY OF WORK

This thesis attempts to investigate the modeling performance of SVM's on nonlinear time series data of pan evaporation characterized by trends, seasonal irregularities. Usually, SVM models have shown better potential and capable of directly analyzing real-world time series in comparison to several other techniques such as Fuzzy system, ANFIS, ANN, GP and similar other techniques. But, while handling nonlinear and complex pan evaporation data, the model performance will not be that accurate. The pre-processing of data would make it smoother for SVR regression to produce a more accurate estimation of pan evaporation. In view of that hybrid combination of wavelet transform and Support Vector Machines is used in modeling pan evaporation.

The models are developed on the pan evaporation data recorded at two stations which are climatically contrasting to each other. The stations are Bajpe and Bangalore located in the state Karnataka, in India. The former station represents the humid climate and a latter represents semi-arid climate. Model development was carried out using daily recorded attributes of pan evaporation data of seven years of Bajpe and same consisting of ten years data of Bangalore station. The input combinations of attributes used in this research were evaluated using Gamma test. The parameter optimization for SVR kernels were carried out using Grid search.

The sequence of model development work carried out is as follows:

1. Explore the capability of SVR based three kernel functions namely Polynomial, RBF, and PUK to directly model pan evaporation data exhibiting trends, seasonal irregularities, discontinuities and other complex behavior.

2. Understand the limitations of single SVM models and identify the possibilities of improving performance through adding data pre-processing technique such as Wavelet transform.
3. Explore the potential of wavelet transform based technique for data decomposition prior, to be fed as inputs to SVM models.
4. Evaluate the performance of discrete wavelet transform-based processing technique in improvising the accuracy of pan evaporation estimations better than single SVM models.

5.2 MAIN FINDINGS OF THE WORK

1. **Based on evaluation of developed models, SVR kernel based models faced some limitations in modeling pan evaporation data with trends, seasonal patterns, discontinuities and other complex behavior**.

   The SVR modeled pan evaporation results reveals that, overall model performance was low in comparison to hybrid models. SVR estimations were accurate for a certain period of modeled results, but at crucial periods SVR showed in-capabilities to handle the data associated with trend, seasonal irregularities and discontinuities.

2. **The application of wavelet transform as a pre-processing technique results in improved model performance for SVR models.**

   The limitations exposed by the SVR kernels in handling nonlinearity in pan evaporation were better addressed by discrete wavelet transform data pre-processing technique. The irregularities were minimized with decomposition with Daubechies mother wavelet function at level 4 and order 3 to improvise the accuracy of SVR kernel estimations.

3. **Evaluation of hybrid model performance suggests that, Daubechies mother wavelet functions are superior to map the original signal.**

   For wavelet analysis, DWT is used with two different mother wavelet functions such as Daubechies and Haar with various optimum orders and levels. Among them Daubechies wavelet with order-4 (db4) and level 3 found superior. The results of DWT-SVR strengthen this fact by showing higher performance. In general, wavelet transform pre-processing is beneficial for analyzing nonlinear time series, and is robust to the presence of noise, seasonal irregularities and trend patterns, and inhomogeneous variance.

4. **The model performance was robust even under varied climatic conditions.** As one of the set objectives of this research, the model efficiency was tested with two different climatic conditions. Observing the patterns of estimations of both hybrid DWT-SVR models and single SVR models, it is clear that in spite of nonlinear data and irregularities, performance pattern is unaltered. Although the accuracy level of estimation varies among stations, but overall pattern has remained the same.

5.3 CONCLUSIONS DRAWN FROM THE CONDUCTED RESEARCH

Based on the result analysis following conclusions are drawn.

- SVR-RBF results showed superior performance in modeling pan evaporation, the estimated values found very close for the low and medium pan evaporation values.
- In modeling pan evaporation, hybrid models of DWT-SVR is found superior to conventional SVR models.
- DWT-SVR estimated pan evaporation values were more accurate for the humid station i.e Bajpe than semi-arid station selected in this study i.e Bangalore.

5.4 STUDY LIMITATIONS

Some of the limitations and assumptions involved in this study are as follows:

- Modeling and analysis of pan evaporation require qualitative and quantitative data. It was experienced that data length could be even more to provide comprehensive and judgmental discussion on the estimated results which could cover the climatic changes that may take place over the decades.
- Model efficiency in this work tested against only two climatic zones i.e. humid and semi-arid, but conclusions could be drawn even better including more stations covering rest other climatic zones.

5.4 SUGGESTED DIRECTIONS FOR FUTURE WORK

There are several enhancements that can be made in this work:

i) The thesis focused on wavelet transform-based pre-processing for class regression techniques such as SVR models for pan evaporation estimations. Further research can be conducted in this domain with other modeling techniques of Artificial intelligence, using a mixture of time series with homogeneous variance profiles, and time series with discontinuities in the variance across time scales.

ii) The present work can be extended to further level by considering more stations covering other climatic conditions especially arid region.

REFERENCES

Abghari, H., Ahmadi, H., Besharat, S., and Rezaverdinejad, V. (2012). "Prediction of Daily Pan Evaporation using Wavelet Neural Networks." *Water.R*esour.*M*anag, 26(12), 3639–3652.

Adamowski, J. F. (2008). "Development of a short-term river flood forecasting method for snowmelt driven floods based on wavelet and cross-wavelet analysis." *J. Hydrol.*, 353(3-4), 247–266.

Adeloye, A. J., Nawaz, N. R., and Montaseri, M. (1999). "Climate change water resources planning impacts incorporating reservoir surface net evaporation fluxes: A case study." *INT. J. Water. Resour. D*, 15(4), 561–581.

Asefa, T., Kemblowski, M., McKee, M., and Khalil, A. (2006). "Multi-time scale stream flow predictions: The support vector machines approach." *J. Hydrol.*, 318, 7–16.

Bhakar, S. R. and Singh, R. V. (2004). “Effect of Climatic Parameters on Evaporation

and Evapotranspiration.” *J.IND.Water.Soc*, 24(1), 12-18.

Blatter, C. (1998). "Wavelets. A primer." *AK Peters*, Natick, Massachusetts, USA.

Bruton, J. M., McClendon, R. W., and Hoogenboom, G. (2000). "Estimating daily pan evaporation with artificial neural networks." *T.ASCE*, 43(2), 491–496.

Brutsaert, W. (1982). "Evaporation into the Atmosphere. Theory, History, and Applications;" Dordrecht: Holland, D. *Reidel Co*.

Brutsaert, W., and Parlange, M. B. (1998). "Hydrologic cycle explains the evaporation paradox." *Nature*, 396(6706), 30-30.

Burman, Robert D. (1976), “Intercontinental comparison of evaporation estimates.”, *J. Irrig.Drain. Div*, 102(1), 109-118.

Burt, C. M., Mutziger, A. J., Allen, R. G., and Howell, T. A. (2005). "Evaporation Research: Review and Interpretation." *J. Irrig. Drain. Eng.*, 131(1), 37–58.

Burt, T. P., and Shahgedanova, M. (1998). "An historical record of evaporation losses since 1815 calculated using long-term observations from the Radcliffe Meteorological Station, Oxford, England." *J. Hydrol.*, 205(1-2), 101–111.

Chandra, A., Shrikhande, V. J., and Kulshreshta, R. (1988). "Relationship of pan evaporation with meteorological parameters." *J. Indian Water Reso. Soc*, 8(2), 41-44.

Chapelle, O., Vapnik, V., Bousquet, O., and Mukherjee, S. (2002). "Choosing multiple parameters for support vector machines." *MachineLearning*, 46(1-3), 131-159.

Chattopadhyay, N., and Hulme, M. (1997). "Evaporation and potential evapotranspiration in India under conditions of recent and future climate change." *Agr.Forest. Meteorol.*, 87(1), 55–73.

Chen, Y., Yang, B., and Dong, J. (2006). "Time-series prediction using a local linear wavelet neural network." *Neurocomputing*, 69(4-6), 449–465.

Cherkassky, V., and Ma, Y. (2004). "Practical selection of SVM parameters and noise estimation for SVM regression." *Neuralnetworks*, 17(1), 113-126.

Chou, C. M., & Wang, R. Y. (2002). "On-line estimation of unit hydrographs using the wavelet-based LMS algorithm/Estimation en ligne des hydrogrammes unitaires grâce à l'algorithme des moindres carrés moyens à base d'ondelettes." *Hydrol. Sci. J.*, 47(5), 721-738..

Cohen, S., Ianetz, A., and Stanhill, G. (2002). "Evaporative climate changes at Bet Dagan, Israel, 1964-1998." *Agr.Forest. Meteorol.*, 111(2), 83–91.

Cortes, C. and Vapnik, V. (1995). "Support Vector Networks." *MachineLearning*, 20, 273-297.

Daubechies, I. (1992). "Ten Lectures on Wavelets." *IEEE Symposium on ComputerBased Medical Systems* 61.

Daubechies, I. (1993). "Orthonormal bases of compactly supported wavelets II. Variations on a theme." SIAM J.MathAnal, 24(2), 499-519.

Dingman, S. L. (1994). "Physical Hydrology", *Prentice Hall*, Inc., New Jersey,7458.

Deka, P. C., and Prahlada, R. (2012). "Discrete wavelet neural network approach in significant wave height forecasting for multistep lead time." *Ocean.Eng*, 43, 32–42.

Deswal, S., and Pal, M. (2008). "Artificial Neural Network based Modeling of Evaporation Losses in Reservoirs." *Int.J.Math.Phy.Eng.Sci*, 39(3), 177–181.

Deswal, S., and Pal, M. (2008). "Modeling of Pan Evaporation Using Support Vector Machines Algorithm." *ISH J. Hyd.Eng*, 14(1), 104–116.

Dibike, Y. B., Velickov, S., Solomatine, D., and Abbott, M. B. (2001). "Model Induction with Support Vector Machines: Introduction and Applications." *J.Comp.Civ. Eng*.

Dibike, Y. B., Velickov, S., Solomatine, D., & Abbott, M. B. (2001). "Model induction with support vector machines: introduction and applications." *J.Comp.Civ. Eng.*, 15(3), 208-216.

Drucker, H., Wu, D., and Vapnik, V. N. (1999). "Support vector machines for spam categorization." *IEEE.T.Neural.Networ.*, 10, 1048–1054.

Duan, K., Keerthi, S. S., and Poo, A. N. (2003). "Evaluation of simple performance measures for tuning SVM hyperparameters." *Neurocomputing*, 51, 41-59.

Engel, B., Storm, D., White, M., Arnold, J., and Arabi, M. (2007). "A hydrologic/water quality model application protocol." *J.Am.Water.Res.Assoc.*, 43, 1223–1236.

Eslamian, S. S., Gohari, S. A., Biabanaki, M., & Malekian, R. (2008). Estimation of monthly pan evaporation using artificial neural networks and support vector machines. J.Appl.Sci., 8(19), 3497-3502.

Espinoza, F., Minsker, B., and Goldberg, D. (2005). "Adaptive Hybrid Genetic Algorithm for Groundwater Remediation Design." *J.Wat.Res.Planning. Manage.*, *131*(1), 14–24.

Evans, D., and Jones, A. J. (2002). "A proof of the gamma test." *The Royal Society*, 458, 2759-2799.

Fielding, A. H., and Bell, J. F. (1997). "A review of methods for the assessment of prediction errors in conservation presence / absence models." *Env. Con.*, 24(1), 38–49.

Finch, J., and Calver, A. (2008). "Methods for the quantification of evaporation from lakes."

Fu, G., Charles, S. P., and Yu, J. (2009). "A critical overview of pan evaporation trends over the last 50 years." *Clim.Change*, 97, 193–214.

Goyal, M. K., Bharti, B., Quilty, J., Adamowski, J., and Pandey, A. (2014). "Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS." *Exp.Sys.App.*, 41(11), 5267–5276.

Guimaraes Santos, C. A., and Silva, D. (2014). "Daily streamflow forecasting using a wavelet transform and artificial neural network hybrid models." *Hyd.Sci.J.*, 59(2), 312–324.

Gundekar, H. G., Khodke, U. M., Sarkar, S., and Rai, R. K. (2007). "Evaluation of pan coefficient for reference crop evapotranspiration for semi-arid region." *Irr.Sci*, 26(2), 169–175.

Guven, A., & Kişi, Ö. (2011). Daily pan evaporation modeling using linear genetic programming technique. *Irr.Sci*., 29(2), 135-145.

Han, D., Chan, L., and Zhu, N. (2007). "Flood forecasting using support vector machines." *J. hyd.*, 9(4), 267-276.

Jackson, R. D. (1985). "Evaluating evapotranspiration at local and regional scales." *Proc. IEEE*, 73(6), 1086-1096.

Jain, S. K., Nayak, P. C., and Sudheer, K. P. (2008). "Models for estimating evapotranspiration using artificial neural networks, and their physical interpretation." *Hyd.Pro.*, 22, 2225–2234.

Jones, A. (2004). "New tools in non-linear modelling and prediction." *Comp. Man. Sci.*, 1(2), 109–149.

Jothiprakash, V., and Kote, A. S. (2011). "Improving the Performance of Data-Driven Techniques through Data Pre-processing for Modelling Daily Reservoir Inflow." *Hyd.Sci.J.*, 56(1), 168–186.

Judd, K., and Mees, A. (1998). "Embedding as a modeling problem." *Physica D: Nonlinear Phenomena*, 120(3), 273-286.

Kadhane, R. L. and Tatewar, S. P. (2006), "Development of Relationship of Climatic Parameters with Evaporation and Evapotranspiration. "*J. Pradushan, Nirmulan*, 3(1), 29-33.

Kaheil, Y. H., Rosero, E., Gill, M. K., McKee, M., and Bastidas, L. A. (2008). "Downscaling and forecasting of evapotranspiration using a synthetic model of wavelets and support vector machines." *IEEE.Tran.Geo.Rem.*, 46(9), 2692-2707.

Kahler, D. M., and Brutsaert, W. (2006). "Complementary relationship between daily evaporation in the environment and pan evaporation." *Wat.Res.*, *42*(5), 1–9.

Kar, C., and Mohanty, A. R. (2006). "Monitoring gear vibrations through motor current signature analysis and wavelet transform." *Mec.Sys.Sig.Pro.*, 20(1), 158–187.

Kasiviswanathan, K. S., Pandian, R. S. R., Saravanan, S., and Agarwal, A. (2011). "Genetic programming approach on evaporation losses and its effect on climate change for Vaipar Basin." *Int. J.Com.Sci.*, 8, 269–274.

Kasiviswanathan, K. S., Saravanan, S., Agarwal, A., and Sathyamurthi, S. (2009). "Estimation of Monthly Evaporation for Kovilar Reservoir using Genetic Programming and Thornthwaite Method." *Proceedings of the 4th IASME / WSEAS International Conference on Water Resources, Hydraulics and Hydrology (Whh'09)*, 186–190.

Keskin, M. E., Terzi, Ö., and Taylan, D. (2004). "Fuzzy logic model approaches to daily pan evaporation estimation in western Turkey." *Hyd.Sci.J.*, 49(6), 37–41.

Keskin, M. E., Terzi, Ö., & Taylan, D. (2009). Estimating daily pan evaporation using adaptive neural-based fuzzy inference system. Theor.App.Clim., 98(1-2), 79-87.

Kim, S., Shiri, J., and Kisi, O. (2012). "Pan Evaporation Modeling Using Neural Computing Approach for Different Climatic Zones." *Wat.Res.Man.*, 26(11), 3231–3249.

Kisi, O. (2006). "Daily pan evaporation modelling using a neuro-fuzzy computing technique." *J. Hydrol.*, 329(3-4), 636–646.

Kisi, O. (2013). "Evolutionary neural networks for monthly pan evaporation modeling." *J. Hydrol.*, 498, 36–45.

Kisi, O., and Cimen, M. (2012). "Precipitation forecasting by using wavelet-support vector machine conjunction model." *Eng.Appl.Artif.Intel*, 25(4), 783–792. http://doi.org/10.1016/j.engappai.2011.11.003

Kisi, O., and Ozturk, O. (2007). "Adaptive Neurofuzzy Computing Technique for Evapotranspiration Estimation." *J. Irrig. Drain. Eng.*, 133(4), 368–379.

Kumar, D. N., Reddy, M. J., and Maity, R. (2007). "Regional Rainfall Forecasting using Large Scale Climate Teleconnections and Artificial Intelligence Techniques." *J. Int. Sys.*, 16(4), 307–322.

Kumar, D., and Tiwari, A. K. (2012). "Evaporation Estimation Using Artificial Neural Networks and Adaptive Neuro-Fuzzy Inference System Techniques.", J.Meteorology, 8(16), 81–88.

Kwok, J. T. Y. (2000). "The evidence framework applied to support vector machines." IEEE Tran.Neur.Net., 11(5), 1162-1173.

Labat, D., Ababou, R., and Mangin, a. (2000). "Rainfall-runoff relations for karstic springs. Part II: Continuous wavelet and discrete orthogonal multiresolution analyses." *J. Hydrol.*, 238(3-4), 149–178.

Li, Z., Chen, Y., Shen, Y., Liu, Y., & Zhang, S. (2013). "Analysis of changing pan evaporation in the arid region of Northwest China." *Wat.Res.Res.*, 49(4), 2205-2212.

Linacre, E. T. (1994). "Estimating U.S. Class A Pan Evaporation from Few Climate Data." *Wat.Int.*, 19(1), 5–14.

Liong, S. Y., and Sivapragasam, C. (2002). "Flood stage forecasting with support vector machines." *J.Am.Wat.Res.Ass.*, 38, 173–186.

Liu, X., Yuan, S., and Li, L. (2012). "Prediction of temperature time series based on wavelet transform and support vector machine." *Jou.Com.*, 7(8), 1911–1918.

Lu, R. Y. (2002). “Decomposition of interdecadal and interannual components for North China rainfall in rainy season.” Chinese.J.Atm., 26, pp. 611-624.

Lowe, L. D., Webb, J. A., Nathan, R. J., Etchells, T., and Malano, H. M. (2009). "Evaporation from water supply reservoirs: An assessment of uncertainty." *J. Hydrol.*, 376(1-2), 261–274.

Maity, R., Bhagwat, P. P., and Bhatnagar, A. (2010). "Potential of support vector regression for prediction of monthly streamflow using endogenous property." *Hyd. Pro.*, 24(7), 917-923.

Malik, A., and Kumar, A. (2015). "Pan Evaporation Simulation Based on Daily Meteorological Data Using Soft Computing Techniques and Multiple Linear Regression." *Wat.Res.Man.*, 29(6), 1859–1872.

Mallat, S. (1989). "A theory for multiresolution signal decomposition: the waveletrepresentation." *PAMI*, *11*, 674–693.

Mallat, S. (1999). "A Wavelet Tour of Signal Processing." *A Wavelet Tour of Signal Processing*, 20–41.

Martinez, B., and Gilabert, M. (2009). "Vegetation dynamics from NDVI time series analysis using the wavelet transform." *Rem.Sen.Env.*, 113, 1823–1842.

Martinez, J. M. M., Alvarez, V. M., Gonzalez-Real, M. M., and Baille, A. (2006). "A simulation model for predicting hourly pan evaporation from meteorological data." *J. Hydrol.*, 318(1–4), 250–261.

McKenzie, R. S., and Craig, A. R. (2001). "Evaluation of river losses from the Orange River using hydraulic modelling." *J. Hydrol.*, 241, 62–69.

McMahon, T. A., Peel, M. C., Lowe, L., Srikanthan, R., and McVicar, T. R. (2013). "Estimating actual, potential, reference crop and pan evaporation using standard meteorological data: A pragmatic synthesis." *Hyd.Ear.Sys.Sci.*, 17, 1331–1363.

Meyer, Y. (1992). "Wavelets and Applications." *J.Aco.Soc.Ame.*, 92(5), 3023.

Misiti, M., and Misiti, Y. (1996). "Wavelet toolbox. "*The MathWorks Inc.*

Misiti, M., Misiti, Y., Oppenheim, G., and Poggi, J., (2010). ""Wavelet toolbox: for use with MATLAB." *The MathWorks*, Nastic, Mass.

Moghaddamnia, A., Ghafari, M., Piri, J., and Han, D. (2009). "Evaporation Estimation Using Support Vector Machines Technique." *Int.Jou.Eng.App.Sci.*, 5(2), 76–84.

Müller, K., Smola, A., Ratsch, G., Scholkopf, B. J. (1997). "Predicting time series with support vector machines." *Artificial Neural Networks—ICANN'97*, 1327(x), 999–1004.

Nourani, V., Alami, M. T., and Aminfar, M. H. (2009). "A combined neural-wavelet model for prediction of Ligvanchai watershed precipitation." *Eng.App.Art.Int.*, 22, 466–472.

Nourani, V., and Fard, S. M. (2012). "Sensitivity analysis of the artificial neural network outputs in simulation of the evaporation process at different climatologic regimes." *Adv.Eng.Sof.*, 47(1), 127–146.

Ogolo, E. O. (2011). "Regional trend analysis of pan evaporation in Nigeria (1970 to 2000)." J.Geo.Reg.Pla., 4(10), 566.

Pal, M. (2006). Support vector machine-based feature selection for land cover classification: a case study with DAIS hyperspectral data. *Int.J.Rem.Sen*., 27(14), 2877-2894.

Pal, M., and Goel, A. (2006). "Prediction of the end-depth ratio and discharge in semi-circular and circular shaped channels using support vector machines." *Flow Measurement and Instrumentation*, 17(1), 49–57.

Penman, H. L. (1948). "Natural evaporation from open water, hare soil and grass." *Proceedings of the Royal Society of London. Series A: Mathematical and Physical Sciences*, 193, 120–145.

Piri, J., Amin, S., Moghaddamnia, A., Keshavarz, A., Han, D., and Remesan, R. (2009). "Daily pan evaporation modeling in a hot and dry climate.", *J.Hyd.Eng*., 14(8), 803-811.

Ping, J. and Yu, F. (2014), "Scaling Delineation of Precipitation and Evaporation with Wavelet Analysis at Anyang City in North Plain, China", *J.App.Sci.*, 14(16), 1809-1818.

Peterson, T. (1995). "Evaporation losing its strength.", *Nature*, 377, 687-688.

Rafiee, J., Tse, P. W., Harifi, A., and Sadeghi, M. H. (2009). "A novel technique for selecting mother wavelet function using an intelli gent fault diagnosis system." *ExpSys.App.*, 36(3), 4862–4875.

Raghavendra. N, S., and Deka, P. C. (2014). "Support vector machine applications in the field of hydrology: A review." *App.Soft.Com.*, 19, 372–386.

Rahimikhoob, A. (2009). "Estimating daily pan evaporation using artificial neural network in a semi-arid environment." *Theo.App. Clim.*, 98(1-2), 101–105.

Rajaee, T. (2011). "Wavelet and ANN combination model for prediction of daily suspended sediment load in rivers." *Sci.Tot.Env.*, 409(15), 2917–2928.

Ramachandra, T. V., Kamakshi, G., and Shruthi, B. V. (2004). "Bioresource status in Karnataka." *Ren.Sus.Ene.Rev.*, 8, 1–47.

Reddy, R., and Nair, R. R. (2013). "The efficacy of support vector machines (SVM) in robust determination of earthquake early warning magnitudes in central Japan." *J.Ear.Sys.Sci.*, 122(5), 1423-1434.

Roderick, M. L., and Farquhar, G. D. (2004). "Changes in Australian pan evaporation from 1970 to 2002." *Int.J.Clim.*, 24(9), 1077–1090. http://doi.org/10.1002/joc.1061

Roderick, M. L., Rotstayn, L. D., Farquhar, G. D., and Hobbins, M. T. (2007). "On the attribution of changing pan evaporation." *Geo.Res.Letters*, 34(17), 1–6.

Romanenko, V. A. (1961). "Computation of the autumn soil moisture using a universal relationship for a large area." Proc.Ukrainian.Hyd.Res., (Kiev), 3.

Rosenberry, D. O., Winter, T. C., Buso, D. C., and Likens, G. E. (2007). "Comparison of 15 evaporation methods applied to a small mountain lake in the northeastern USA." *J. Hydrol.*, 340 (3-4), 149–166.

Samanta, B., and Al-balushi, K. R. (2003). "Artificial neural network based fault diagnostics of rolling element bearings using time-domain features." *Mech.Sys. Sig.Proc.*, 17(2), 317–328.

Sanikhani, H., Kisi, O., Nikpour, M. R., and Dinpashoh, Y. (2012). "Estimation of daily pan evaporation using two different adaptive neuro-fuzzy computing techniques." Wat.Res.Man., 26(15), 4347-4365.

Saravanan, N., Kumar Siddabattuni, V. N. S., and Ramachandran, K. I. (2008). "A comparative study on classification of features by SVM and PSVM extracted using Morlet wavelet for fault diagnosis of spur bevel gear box." *Exp.Sys.App.s*, 35, 1351–1366.

Seifi, A., and Riahi-Madvar, H. (2012). "Input variable selection in expert systems based on hybrid gamma test-least square support vector machine ANFIS and ANN models." *Provisional chapter*. Intech.

Senapati, P. C., Mishra, N., and Lal, R. (1985). "Relationship between pan evaporation and meteorological parameters at Bhubaneshwar (Orissa)." *J.Indian.Wat.Reso.Soc*, 5, 27-32.

Shinde, S., Nanaware, P., Kudlikar, G., and Nagi, H. (2013). "Sketch Based Image Retrieval System Using Wavelet Transform." *Int.J.Inn.Res.Dev.*, *2*(4), 69-74.

Shirgure, P. S., and Rajput, G. S. (2011). "Evaporation modeling with neural networks–A Research Review." *Int.J.Res.Rev.Soft.Int.Com.*, *1*(2), 37-47.

Shirgure, P. S. (2013). "Review article Evaporation modeling with artificial neural network : A review", *2*, 73–84.

Shiri, J., Dierickx, W., Pour-Ali Baba, A., Neamati, S., and Ghorbani, M. A. (2011). "Estimating daily pan evaporation from climatic data of the State of Illinois, USA using adaptive neuro-fuzzy inference system (ANFIS) and artificial neural network (ANN)." *Hyd.Res.*.

Shiri, J., and Kisi, O. (2010). "Short-term and long-term streamflow forecasting using a wavelet and neuro-fuzzy conjunction model." *J. Hydrol.*, 394(3-4), 486–493.

Shirsath, P. B., and Singh, A. K. (2009). "A Comparative Study of Daily Pan Evaporation Estimation Using ANN, Regression and Climate Based Models." *Wat.Res. Man.*, 24(8), 1571–1581.

Shirsath, P. B., and Singh, A. K. (2010). "A Comparative Study of Daily Pan Evaporation Estimation Using ANN, Regression and Climate Based Models." *Wat.Res. Man.*, 24(8), 1571–1581.

Silva, D. (2004). "On climate variability in Northeast of Brazil." *J.Arid.Env.*, 58(4), 575–596.

Singh, R. V., Chauhan, H. S., and Ali, A. B. M. (1981). "Pan evaporation as related to meteorological parameters." *Journal of agricultural engineering*.

Singh, V. P., and Xu, C.-Y. (1997). "Evaluation and Generalization of 13 Mass-Transfer Equations for Determining Free Water Evaporation." *Hydr. Pro.*, 11, 311–323.

Sivapragasam, C., Vasudevan, G., Maran, J., Bose, C., Kaza, S., and Ganesh, N. (2009). "Modeling evaporation-seepage losses for reservoir water Balance in semi-arid regions." *Wat. Res.Man*. http://doi.org/10.1007/s11269-008-9303-3

Smith, L. C., Turcotte, Ã. D. L., and Isacks, B. L. (1998). "Stream flow characterization and feature detection using a discrete wavelet transform." *Hyd. Pro.*, 12, 233–249.

Sudheer, K. P., Gosain, A. K., Mohana Rangan, D., and Saheb, S. M. (2002). "Modelling evaporation using an artificial neural network algorithm." *Hyd. Pro.*, 16(16), 3189-3202.

Sudheer, K. P., and Jain, S. K. (2003). "Radial basis function neural network for modeling rating curves." *J.Hyd.Eng.*, 8(3), 161-164.

Sujay R. N., & Deka, P. C. (2015). "Forecasting monthly groundwater level fluctuations in coastal aquifers using hybrid Wavelet packet–Support vector regression."*Cog.Eng.*, 2(1), 999414.

Tabari, H., Kisi, O., Ezani, A., and Hosseinzadeh Talaee, P. (2012). "SVM, ANFIS, regression and climate based models for reference evapotranspiration modeling using limited climatic data in a semi-arid highland environment." *J. Hyd.*, 444-445, 78–89.

Tebakari, T., Yoshitani, J., and Suvanpimol, C. (2005). "Time-Space Trend Analysis in Pan Evaporation over Kingdom of Thailand." *J.Hyd.Eng.*, 10, 205–215.

Tezel, G., & Buyukyildiz, M. (2016). Monthly evaporation forecasting using artificial neural networks and support vector machines. Theo.App.Clim., 124(1-2), 69-80.

Thornthwaite, C. W. (1948). "An approach toward a rational classification of climate." Geo.Rev., 38(1), 55-94.

Torrence, C., and Compo, G. P. (1998). "A Practical Guide to Wavelet Analysis." *Bulletin of the American Meteorological Society*, 79(1), 61–78.

Tse, P. W., Yang, W. X., and Tam, H. Y. (2004). "Machine fault diagnosis through an effective exact wavelet analysis." *J. Sound and Vibration*, 277(4-5), 1005–1024.

Turc, L. (1961), "Estimation of irrigation water requirements, potential evapotranspiration: A simple climatic formula evolved up to date.",*Ann. Agron*., 12, 13–49.

Üstün, B., Melssen, W. J., and Buydens, L. M. (2006). "Facilitating the application of Support Vector Regression by using a universal Pearson VII function based kernel." *Chem.Int.Lab.Sys.*, *81*(1), 29-40.

Vapnik, V. N. (1995). "*The Nature of Statistical Learning Theory*,8, Springer.

Wang, W., and Ding, J. (2003). "Wavelet network model and its application to the prediction of hydrology." *Nature and Science*, 1(1), 67-71.

Yu, C. S., Lin, C. J., and Hwang, J. K. (2004). "Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions." *Protein Science*, 13(5), 1402-1406.

Zhang, L., Zhou, W., and Jiao, L. (2004). "Wavelet support vector machine. *IEEE Transactions on Systems, Man, and Cybernetics". Part B, Cybernetics : A Publication of the IEEE Systems, Man, and Cybernetics Society*, 34(1), 34–39.

List of Publications

Journals (Published/Accepted)

1. Leeladhar Pammar., and Deka, P. C. (2015). "Forecasting Daily Pan Evaporation Using Hybrid Model of Wavelet Transform and Support Vector Machines." *Int. J. of Hydrology Science and Technology*, 5(3), 274 – 294
   DOI: 10.1504/IJHST.2015.071354
2. Leeladhar Pammar., and Deka, P. C. (2014). "Prediction of daily pan evaporation using support vector machines." *IJEE*, 7(1), 195-202 (H Index: 4, ISSN: 09745904)
3. Leeladhar Pammar., and Deka, P. C. "Daily Pan Evaporation Modeling in Climatically Contrasting Zones with Hybridization of Wavelet Transform and Support Vector Machines." *Water Resources, Springer.* (Under press).

Conferences

1. Leeladhar Pammar., and Deka, P. C. (2014). "Modeling Pan Evaporation using Support Vector Machines with Parameter Optimization.", *In Proc. International conference on emerging trends in engineering (ICETE 2014).*, NMAM Institute of Technology, Nitte., May 2014, Vol. 1, 110-116.

BIO-DATA

Name	**:**	LEELADHAR PAMMAR
Date of Birth	**:**	15-07-1985
Address	**:**	House No. 697/D-23, Patribasav Nagar, 3rd cross, Bailhongal, Karnataka, India.
Profession	**:**	Asst.Professor, Department of Civil Engineering, NMAM Institute of technology, Nitte.
E-mail	**:**	lohitnitks@gmail.com
Contact Number	**:**	09738919241
Qualification	**:**	• Graduation in B.Tech Civil Engineering from NITK, Surthkal. • M.Tech –Construction Technology and Management, from N.I.T.K Surathkal.
Publications	**:**	03 International Journals 01 International Conference