THREE LAYER WAVELET BASED MODELING FOR RIVER FLOW

SIRAJ MUHAMMED PANDHIANI

UNIVERSITI TEKNOLOGI MALAYSIA
THREE LAYER WAVELET BASED MODELING FOR RIVER FLOW

SIRAJ MUHAMMED PANDHIANI

A thesis submitted in fulfilment of the requirements for the award of the degree of Doctor of Philosophy (Mathematics)

Faculty of Science
Universiti Teknologi Malaysia

NOVEMBER 2016
DEDICATION

Specially dedicated to my beloved late parents, my family and my friends for their patience, support, prayers, encouragement, and blessings.
ACKNOWLEDGEMENT

First of all, I would like to praise and thank almighty Allah who enabled me to complete my doctorate. I thank to almighty Allah for making my dream come true. The day that I dreamt of has finally come and I am graduating my Phd.

I feel honored of being supervised by Dr. Ani Bin Shabri. This thesis would not have been completed without his help, support and guidance. I would like to offer my sincerest gratitude and thanks to my supervisor who has supported me throughout my PhD studies with his patience and knowledge. One simply could not wish for a better and friendlier supervisor.

I would like to express my thanks to UTM for providing me with the support needed to complete my thesis. A special thanks from me to Dr. Maher Al – Ghanem, Managing Director Jubail University College, for providing me with the opportunity and support to achieve this milestone.

I would like to thank my late parents for all their love and encouragement. They raised me with love and supported me in all my pursuits. They have been a constant source of inspiration throughout my life. I will forever be indebted to my parents for their support and encouragement.

The last but not the least heartfelt acknowledgment must go to my wife and kids. Their encouragement and patience has helped me massively throughout this period.
ABSTRACT

All existing methods regarding time series forecasting have always been challenged by the continuous climatic change taking place in the world. These climatic changes influence many unpredictable indefinite factors. This alarming situation requires a robust forecasting method that could efficiently work with incomplete and multivariate data. Most of the existing methods tend to trap into local minimum or encounter over fitting problems that mostly lead to an inappropriate outcome. The complexity of data regarding time series forecasting does not allow any one single method to yield results suitable in all situations as claimed by most researchers. To deal with the problem, a technique that uses hybrid models has also been devised and tested. The applied hybrid methods did bring some improvement compared to the individual model performance. However, most of these available hybrid models exploit univariate data that requires huge historical data to achieve precise forecasting results. Therefore, this study introduces a new hybrid model based on three layered architecture: Least Square Support Vector Machine (LSSVM), Discrete Wavelet Transform (DWT), correlation (R) and Kernel Principle Components Analyses (KPCA). The three-staged architecture of the proposed hybrid model includes Wavelet-LSSVM and Wavelet-KPCA-LSSVM enabling the model to present itself as a well-established alternative application to predict the future of river flow. The proposed model has been applied to four different data sets of time series, taking into account different time series behavior and data scale. The performance of the proposed model is compared against the existing individual models and then a comparison is also drawn with the existing hybrid models. The results of WKP-LSSVM obtained from Coefficient of Efficiency (CE) performance measuring methods confirmed that proposed model has encouraging data of 0.98%, 0.99%, 0.94% and 0.99% for Jhelum River, Chenab River, Bernam River and Tualang River, respectively. It is more robust for all datasets regardless of the sample sizes and data behavior. These results are further verified using diverse data sets in order to check the stability and adaptability. The results have demonstrated that the proposed hybrid model is a better alternative tool for time series forecasting. The proposed hybrid model proves to be one of the best available solutions considering the time series forecasting issues.
ABSTRAK

Semua kaedah ramalan siri masa yang sedia ada sentiasa dicabar oleh perubahan iklim berterusan yang berlaku di dunia. Perubahan iklim yang berlaku ini dipengaruhi oleh banyak faktor yang tidak menentu. Keadaan yang membimbangkan ini memerlukan satu kaedah ramalan yang teguh yang boleh disesuaikan dengan data tak lengkap dan multivariat. Kebanyakan kaedah-kaedah yang sedia ada cenderung untuk terjebak dalam masalah minimum tempatan atau terlebih suaiian yang membawa kepada hasil yang tidak sesuai. Kerumitan data siri masa ramalan tidak membolehkan satu kaedah tunggal untuk menghasilkan keputusan yang sesuai dalam semua keadaan seperti yang didakwa oleh kebanyakan penyelidik. Untuk mengatasi masalah ini, satu teknik berdasarkan model hibrid telah dicipta dan diuji. Kaedah hibrid yang digunakan telah membawa beberapa penambahbaikan berbanding dengan prestasi model tunggal. Walau bagaimanapun, kebanyakan model hibrid didapati mengeksploitasi data univariat yang memerlukan data sejarah yang besar untuk mencapai keputusan ramalan yang tepat. Oleh itu, kajian ini memperkenalkan model hibrid baharu yang berdasarkan seni bina tiga lapis: kuasa dua terkecil mesin vektor sokongan (LSSVM), transformasi Wavelet diskrit (DWT), korelasi (R) dan analisis komponen inti prinsipal (KPCA). Seni bina tiga lapis model hibrid yang dicadangkan termasuk model Wavelet-LSSVM dan Wavelet-KPCA-LSSVM sebagai model alternatif yang mantap untuk meramalkan masa depan aliran sungai. Model yang dicadangkan ini telah digunakan ke atas empat set data siri masa yang berbeza, dengan mengambil kira ciri-ciri siri masa dan skala data yang berbeza. Prestasi model yang dicadangkan dibandingkan dengan model tunggal yang sedia ada dan kemudian perbandingan juga dilakukan dengan model hibrid yang sedia ada. Keputusan WKP-LSSVM yang diperoleh daripada kaedah pengukuran prestasi kecekapan pekali (CE) mengesahkan bahawa model yang dicadangkan mempunyai data yang menggalakkan iaitu masing-masing 0.98%, 0.99%, 0.94% dan 0.99% bagi Sungai Jhelum, Sungai Chenab, Sungai Bernam dan Sungai Tualang. Ia adalah lebih kukuh untuk semua set data tanpa mengira saiz sampel dan tingkah laku data. Keputusan ini selanjutnya disahkan menggunakan pelbagai set data untuk memeriksa kestabilan dan penyesuaian. Keputusan telah menunjukkan bahawa model hibrid yang dicadangkan adalah satu model alternatif yang lebih baik untuk ramalan siri masa. Model hibrid yang dicadangkan membuktikannya sebagai salah satu penyelesaian terbaik yang ada dalam mempertimbangkan isu ramalan siri masa.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>CHAPTER</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>DECLARATION</td>
<td>ii</td>
<td></td>
</tr>
<tr>
<td>DEDICATION</td>
<td>iii</td>
<td></td>
</tr>
<tr>
<td>ACKNOWLEDGEMENT</td>
<td>iv</td>
<td></td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>v</td>
<td></td>
</tr>
<tr>
<td>ABSTRAK</td>
<td>vi</td>
<td></td>
</tr>
<tr>
<td>TABLE OF CONTENTS</td>
<td>vii</td>
<td></td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xii</td>
<td></td>
</tr>
<tr>
<td>LIST OF ABBREVIATION</td>
<td>xiv</td>
<td></td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td>xv</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1</td>
<td>Overview</td>
<td>1</td>
</tr>
<tr>
<td>1.2</td>
<td>Problem Statement</td>
<td>5</td>
</tr>
<tr>
<td>1.3</td>
<td>Objectives</td>
<td>8</td>
</tr>
<tr>
<td>1.4</td>
<td>Scope of Study</td>
<td>8</td>
</tr>
<tr>
<td>1.5</td>
<td>Significances of the Study</td>
<td>9</td>
</tr>
<tr>
<td>1.6</td>
<td>Thesis Contribution</td>
<td>10</td>
</tr>
<tr>
<td>1.7</td>
<td>Thesis Structure and Organization</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>LITERATURE REVIEW</td>
<td>13</td>
</tr>
<tr>
<td>2.1</td>
<td>Introduction</td>
<td>13</td>
</tr>
<tr>
<td>2.2</td>
<td>Existing Methods</td>
<td>14</td>
</tr>
<tr>
<td>2.2.1</td>
<td>Artificial Neural Network Model (ANN)</td>
<td>15</td>
</tr>
<tr>
<td>2.2.1.1</td>
<td>Advantages and disadvantages</td>
<td>17</td>
</tr>
</tbody>
</table>
2.2.2 Adaptive Neuro Fuzzy Inference System (ANFIS) 17
  2.2.2.1 Advantages and disadvantages 18
2.2.3 Autoregressive Moving Averages Models (ARMA) 20
  2.2.3.1 Advantages and disadvantages 21
2.2.4 The Autoregressive Integrated Moving Average (ARIMA) Models 22
  2.2.4.1 Advantages and disadvantages 23
2.2.5 Support Vector Machines (SVMs) 23
  2.2.5.1 Advantages and disadvantages 24
2.2.6 The Hybrid Models 25
2.2.7 The Disadvantages Found with A.I Based Techniques 30

2.3 Hybrid Methodology Models 31
2.4 Measuring Performance of Forecasting Model 51
2.5 Summary 52

3 METHODOLOGY 53
3.1 Introduction 53
3.2 Data Description 54
3.3 Description of Schematics Diagram 54
  3.3.1 Least Square Vector Machine (LSSVM) 55
    3.3.1.1 Construction of LSSVM Model 58
  3.3.2 Decomposing the Input Data through DWT 62
    3.3.2.1 Advantages of Wavelet Analysis 66
  3.3.3 Kernel Principal Component Analysis 66
  3.3.4 Fitting Hybrid Model 71
3.4 The Framework of Study 72
3.5 Accuracy 77
3.6 Summary 79

4 SINGLE LSSVM AND PROPOSED HYBRID WAVELET_LSSVM MODELS 80
4.1 Introduction 80
4.2 Study Area and Data 83
4.2.1 Non-stationary and non-linearity 89
4.3 Appropriateness and Meaningfulness of the Data 93
4.3.1 Relevance and Meaningfulness of LSSVM to the Data 94
4.3.2 Appropriateness of DWT and LSSVM to the Data 97
4.4 Comparison of Results and Discussion 111
4.5 Summary 113

5 PROPOSED HYBRID MODELS WAVELET-KPCA-LSSVM 115
5.1 Introduction 115
5.2 Relevance and Meaningfulness of Wavelet, KPCA and LSSVM to the Data (WKPLSSVM) 117
5.3 Summary 127

6 COMPARSION AND DISCUSSION 128
6.1 Introduction 128
6.2 Comparison 128
6.3 Discussions 138
6.3.1 Discussion on Single Model (LSSVM) 138
6.3.2 Discussion on Wavelet-LSSVM (WLSSVM) 139
6.3.3 Discussion on Wavelet-Based Model Combined with Feature Extraction Technique
KPCA-LSSVM WKPLSSVM) 141
6.4 Comparison of Models and Measures of Evaluation for Streamflow River Data 143
6.5 Conclusions 146
6.6 Contributions 146
6.7 Recommendations for Future Study 149

REFERENCES 150
Appendices A1-C 174 - 204
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>AI models and wavelet transform review papers used in hydrology</td>
<td>33</td>
</tr>
<tr>
<td>2.2</td>
<td>Hybrid model surveyed papers for Continuous and Discrete wavelet-AI methods were used to predict hydrological variable</td>
<td>34</td>
</tr>
<tr>
<td>2.3</td>
<td>Details of the surveyed papers, including year of publication, authors, where hybrid wavelet–AI methods were used to predict hydrological variables</td>
<td>50</td>
</tr>
<tr>
<td>4.1</td>
<td>Pertinent information for four watersheds data</td>
<td>88</td>
</tr>
<tr>
<td>4.2</td>
<td>ADF test for four watersheds data</td>
<td>92</td>
</tr>
<tr>
<td>4.3</td>
<td>BDS test for four watersheds data</td>
<td>92</td>
</tr>
<tr>
<td>4.4</td>
<td>The Input structure of the Models for Forecasting</td>
<td>93</td>
</tr>
<tr>
<td>4.5</td>
<td>Forecasting Performance indicators of LSSVM for Pakistan and Malaysia Rivers</td>
<td>96</td>
</tr>
<tr>
<td>4.6</td>
<td>The correlation coefficients between each of sub-time series and original monthly streamflow data</td>
<td>103</td>
</tr>
<tr>
<td>4.7</td>
<td>Forecasting Performance Indicators for Wavelet (db1) decomposition 3 + LSSVM (WLSSVM) for Pakistan and Malaysia Rivers</td>
<td>104</td>
</tr>
<tr>
<td>4.8</td>
<td>Forecasting Performance Indicators for Wavelet (db2) decomposition 3 + LSSVM (WLSSVM) for Pakistan and Malaysia Rivers</td>
<td>105</td>
</tr>
<tr>
<td>4.9</td>
<td>Forecasting Performance Indicators for Wavelet (db3) decomposition 3 + LSSVM (WLSSVM) for Pakistan and Malaysia Rivers</td>
<td>107</td>
</tr>
</tbody>
</table>
4.10 Forecasting Performance Indicators for Wavelet (db4) decomposition 3 + LSSVM (WLSSVM) for Pakistan and Malaysia Rivers

4.11 Forecasting Performance Indicates for Wavelet (db5) decomposition 3 + LSSVM (WLSSVM) for Pakistan and Malaysia Rivers

4.12 The performance results LSSVM and WLSSVM approach during training period

5.1 Forecasting Performance indicators of WKPLSSVM for Jehlum River of Pakistan \( (\sigma = 1000) \)

5.2 Forecasting Performance indicators of WKPLSSVM for Chanab River of Pakistan \( (\sigma = 1000) \)

5.3 Forecasting Performance indicators of WKPLSSVM for Bernam River of Malaysia \( (\sigma = 1000) \)

5.4 Forecasting Performance indicators of WKPLSSVM for Tualang River of Malaysia \( (\sigma = 1000) \)

6.1 Comparative Performances of All Models for Jehlum Rivers

6.2 Comparative Performances of All Models for Chanab River

6.3 Comparative Performances of All Models for Bernam River

6.4 Comparative Performances of All Models for Tualang River
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE NO.</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>General Structure of the ANN Model</td>
<td>16</td>
</tr>
<tr>
<td>2.2</td>
<td>General Structure of GRNN Model</td>
<td>16</td>
</tr>
<tr>
<td>2.3</td>
<td>The General Structure of the Fuzzy Inference System</td>
<td>18</td>
</tr>
<tr>
<td>2.4</td>
<td>General Structure of ANFIS</td>
<td>20</td>
</tr>
<tr>
<td>3.1</td>
<td>Schematic Diagrams for Goodness of Fit for Each Model</td>
<td>53</td>
</tr>
<tr>
<td>3.2</td>
<td>Flow Chart of Predict Method with LSSVM</td>
<td>58</td>
</tr>
<tr>
<td>3.3</td>
<td>Schematic Diagram for Analysis</td>
<td>65</td>
</tr>
<tr>
<td>3.4</td>
<td>Schematic Diagram for Synthesis</td>
<td>65</td>
</tr>
<tr>
<td>3.5</td>
<td>Structure KPCA from PCA</td>
<td>67</td>
</tr>
<tr>
<td>3.6</td>
<td>The structure of the WKPLLSSVM</td>
<td>72</td>
</tr>
<tr>
<td>3.7</td>
<td>The Framework Structure for WAVELET-LSSVM</td>
<td>73</td>
</tr>
<tr>
<td>3.8</td>
<td>The Framework Structure for WAVELET-KPCA-LSSVM</td>
<td>74</td>
</tr>
<tr>
<td>3.9</td>
<td>The Framework of Study</td>
<td>76</td>
</tr>
<tr>
<td>4.1</td>
<td>Location map of the study area Pakistan Rivers (Jhelam and Chanab Rivers)</td>
<td>84</td>
</tr>
<tr>
<td>4.2</td>
<td>Location map of the study area Malaysia River (Bernam River)</td>
<td>85</td>
</tr>
<tr>
<td>4.3</td>
<td>Location map of the study area Malaysia River (Tualang River)</td>
<td>85</td>
</tr>
<tr>
<td>4.4</td>
<td>Monthly River Flow for Jhelum River</td>
<td>86</td>
</tr>
<tr>
<td>4.5</td>
<td>Monthly River Flow for Chanab River</td>
<td>86</td>
</tr>
<tr>
<td>4.6</td>
<td>Monthly River Flow for Bernam River</td>
<td>87</td>
</tr>
<tr>
<td>4.7</td>
<td>Monthly River Flow for Tualang River</td>
<td>87</td>
</tr>
<tr>
<td>4.8</td>
<td>Decomposed wavelet sub-series components (Ds) of streamflow data of Jhelum River</td>
<td>99</td>
</tr>
<tr>
<td>4.9</td>
<td>Decomposed wavelet sub-series components (Ds) of streamflow data of Chenab River</td>
<td>100</td>
</tr>
</tbody>
</table>
4.10 Decomposed wavelet sub-series components (Ds) of streamflow data of Bernam River

4.11 Decomposed wavelet sub-series components (Ds) of streamflow data of Tualang River

6.1 Comparison Hydrograph for LSSVM-WLSSVM for Jehlum River

6.2 Comparison Hydrograph for LSSVM-WKPLSSVM for Jehlum River

6.3 Comparison Hydrograph for LSSVM-WLSSVM-WKPLSSVM for Jehlum River

6.4 Comparison Hydrograph for LSSVM-WLSSVM for Chanab River

6.5 Comparison Hydrograph for LSSVM-WKPLSSVM for Chanab River

6.6 Comparison Hydrograph for LSSVM-WLSSVM-WKPLSSVM for Chanab River

6.7 Comparison Hydrograph for LSSVM-WLSSVM for Bernam River

6.8 Comparison Hydrograph for LSSVM-WKPLSSVM for Bernam River

6.9 Comparison Hydrograph for LSSVM-WLSSVM-WKPLSSVM for Bernam River

6.10 Comparison Hydrograph for LSSVM-WLSSVM for Tualang River

6.11 Comparison Hydrograph for LSSVM-WKPLSSVM for Tualang River

6.12 Comparison Hydrograph for LSSVM-WLSSVM-WKPLSSVM for Tualang River

6.13 Accuracy data of Jhelum River, Chanab River, Bernam River and Tualang River using CE
LIST OF ABBREVIATIONS

ACF -- Autocorrelation Function
AIC -- Akaike Information Criterion
ANN -- Artificial Neural Networks
ARMA -- Autoregressive Moving Average
ARIMA -- Autoregressive Integrated Moving Average
DWT -- Discrete Wavelet Transform
FT -- Fourier Transform
FFT -- Fast Fourier Transform
LSSVM -- Least Square Support Vector Machine
WLSSVM -- Wavelet-Least Square Support Vector Machine
WPLSSVM -- Wavelet-Principal Component Analysis - Least Square Support Vector Machine
WKPLSSVM -- Wavelet-Kernel Principal Component Analysis - Least Square Support Vector Machine
MAE -- Mean Absolute Error
MSE -- Mean Squared Errors
RMSE -- Root Mean Square Error
PACF -- Partial Autocorrelation Function
PCA -- Principal Components Analysis
CE -- Coefficient of Efficiency
RMSE -- Root Mean Squared Error
SOM -- Self Organizing Map
SSE -- Sum of Squared Error
SVM -- Support Vector Machine
WA -- Wavelets Analysis
WT -- Wavelet Transform
## LIST OF APPENDICES

<table>
<thead>
<tr>
<th>APPENDIX</th>
<th>TITLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>WKPLSSVM for Jehlum River ($\sigma = 2000$)</td>
<td>174</td>
</tr>
<tr>
<td>A2</td>
<td>WKPLSSVM for Jehlum River ($\sigma = 3000$)</td>
<td>175</td>
</tr>
<tr>
<td>A3</td>
<td>WKPLSSVM for Chanab River ($\sigma = 2000$)</td>
<td>176</td>
</tr>
<tr>
<td>A4</td>
<td>WKPLSSVM for Chanab River ($\sigma = 3000$)</td>
<td>177</td>
</tr>
<tr>
<td>A5</td>
<td>WKPLSSVM for Bernam River ($\sigma = 2000$)</td>
<td>178</td>
</tr>
<tr>
<td>A6</td>
<td>WKPLSSVM for Bernam River ($\sigma = 3000$)</td>
<td>179</td>
</tr>
<tr>
<td>A7</td>
<td>WKPLSSVM for Tualang River ($\sigma = 2000$)</td>
<td>180</td>
</tr>
<tr>
<td>A8</td>
<td>WKPLSSVM for Tualang River ($\sigma = 3000$)</td>
<td>181</td>
</tr>
<tr>
<td>B1</td>
<td>Graphs for WLSSVM with db1, db2, db3, db4 and db5</td>
<td>182</td>
</tr>
<tr>
<td>B2</td>
<td>Graphs for WKPLSSVM with db1, db2, db3, db4 and db5</td>
<td>193</td>
</tr>
<tr>
<td>C</td>
<td>Research Articles</td>
<td>204</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

1.1 Overview

River flow forecasting is unarguably considered crucial in providing the information required for the design and operation of river systems. While rivers are essentially related with the streams, there is a dire demand to have reliable mechanisms in place that can deliver a unswerving estimate of the water coming into a stream. This estimation is likely to contribute significantly in resolving a number of hydraulic issues such as the depth of flow, flow velocity, and forces from flowing water on a surface or at hydraulic structures. This could also help in planning for dams in order to balance natural flow of water or for streamlining many other hydraulic structures. Thereby, the availability of comprehensive records of rainfall and other climatic data, which could be used to obtain stream flow data, initiated the practice of rainfall-runoff modeling. Understandably, reliable information on current and future water availability is essential to properly manage the limited water resources and flood moderation. Authorities in water sector cannot allocate water resources optimally for water demands like agricultural, industrial, domestic, hydropower generation and environmental maintenance, unless they are equipped with a reliable forecasting of river flow.

Researchers are keen to develop and investigate various types of hydrological models to attain better management of water scarcity and also to minimize the risk of any potential flooding. Water resources planning and management requires output from these hydrological studies. This output is mainly available in the form of
estimation or forecasting of the magnitude of hydrological variables like precipitation, stream flow and groundwater levels using historical data. This data then is used by water management authorities in many of their activities such as designing flood protection works for urban areas and agricultural land and assessing how much water may be extracted from a river for water supply or irrigation. Referring back to these hydrological models, it has been observed that they can be classified as follows: Knowledge-Driven Modeling and Data-Driven Modeling. The knowledge-driven modeling (also known as physically based model) draws heavily on mathematics (differential equations and finite-difference approximations) in order to calculate the catchment characteristics variables such as severity & period of rainfall events, size, shape, slope and storage of the catchment, etc. The proponents of this method generally hypothesize that more accurate forecasts could be achieved if catchment characteristics variables are also included and combined with water flow data in order to reach a more precise and accurate water estimation. While it may be likely that different combinations of flow and catchment characteristics variables would better the forecast results, in practice especially in developing countries like Malaysia and Pakistan, such information is often either unavailable or difficult to acquire. Besides, it could be an extremely complicated physical process mainly due to ‘... the data collection of multiple inputs and parameters, which vary in space and time ...’ (Akhtar et al., 2009). The mathematical models used under this approach include rainfall-runoff models and stream flow models. The former uses both climatic and hydrological data and the latter relies only on hydrological data.

The second approach is the data-driven modeling which is largely based on pulling out and re-using information that is subliminally present in the hydrological data without having to consider the physical laws that lie beneath the rainfall-runoff processes. In river flow forecasting applications, data driven modeling uses historical river flow time series data and this method has increasingly become more popular due to its ‘... rapid development times and minimum information requirements...’ (Kisi, 2009). Although it may not have the ability to provide physical analysis and discernment of the catchment processes but it is able to provide relatively accurate flow forecasts. Computer science and statistics have improved the data driven modeling approaches for discovering patterns found in water resources time series
data. Much effort has been devoted over the past several decades to the development and improvement of time series prediction models. In general, the stochastic models such as Autoregressive Integrated Moving Average (ARIMA) have been widely used for hydrologic time series forecasting. The popularity of ARIMA mainly relies on Jenkins methodology, forecasting capabilities and richness of information on time-related changes.

Another example data driven model are Adaptive Neuro-Fuzzy Inference System (ANFIS), Artificial Neural Network (ANN) and Fuzzy Logic (FL) have been known as potentially useful methods in modeling time-series hydrologic problems. What distinguishes these models from any of the ones based on knowledge – driven approach is their ability to address flood forecasting problems more precisely and accurately where usually the main concern is to minimize the flood damage (Lim and Young, 2008).

However models like ARIMA are essentially linear models that use stationary data with little or no capacity to capture non-stationary and non-linear data (Otache et al., 2011).

Then there are some models like Artificial Neural Network (ANN) and Support Vector Machine (SVM) that has the capacity to capture non-linear data. Both methods are machine-learning methods that find a wide range of applications both in the field of engineering and social sciences. For the SVM a large amount of computation time will be involved when SVM is applied for solving large-size problem (Cao & Tay, 2003).

The network structure of ANN is hard to determine and usually done by using a trial and error approach. Though ANN and SVM methods were good and produced encouraging results, but this combination is computationally expensive and depends on complex quadric programing (QP) (Mellit et al., 2013; Ismail al et., 2010).
Also for both methods, large amount of computational time will be involved when these methods are applied for large-size problem. These deficiencies have been overcome by the Least Square Support Vector Machine (LSSVM), which solves linear equation instead of a quadratic programming problem. Therefore, the LSSVM has this computational advantage over the other AI methods (Wang and Hu, 2005).

Then there is this method called Discrete Wavelet Analysis (DWT), best known tool for data analysis. The contribution of modeling hydrological resources can be seen in the last few years (Kisi, 2010). These include meteorological pollution simulation, open channel wake flows analysis and ground water level time series modeling (Sowski et al., 2007). Recently, wavelet theory have been introduced in the field of hydrology, wavelet models, mainly due to their natural ability to analyze a signal in time and frequency domains, are becoming a general choice for researchers addressing issues related to hydrological models.

Kernel Principal Component Analysis (KPCA) is another important component was found, which uses a kernel function to map the data in the input space and compute the principle components in a feature space. Due to the wide range of research in kernel methods, KPCA is gradually assuming an important position for modeling non-linear data. KPCA is a non-linear version of PCA, which uses a kernel function to map the data in the input space and compute the principal components in a feature space. KPCA is acknowledged for its ability to produce nonlinear PCs but at the same time, the method is unable to directly reconstruct the data in feature space. Now it is being frequently relied upon for integrating nonlinear transformations.

Therefore, in order to account for the deficiencies found in various models especially in the above mentioned three methods, a hybrid approach, which is hereby referred to as WKPLSSVM, has been adopted in this study for the monthly river flow data in Pakistan and Malaysia. In this proposed three – staged structure, it has also been shown that Wavelet, Kernel Principle Components Analysis (KPCA) and Least Square Support Vector Machine (LSSVM) are highly compatible with each.
For the purpose, the above methods have been used in combination with each other at different stages of the research to obtain more accurate and reliable results.

1.2 Problem Statement

The forecasting arguable remains somewhat a troublesome task for hydrologists who distinguish its crucial part in environmental, water assets management and in water-related disaster control. This process remains highly complex for non-stationary, hydrological and hydro-climatologic features. In the recent times the huge rise in the amount of scientific approaches have been observed. They have applied ‘data-based’ or ‘data-driven’ approaches to hydrologic modelling and forecasting. These modelling methods include the mathematical equation, which were taken from analysis of parallel inputs and output time series (Solomatine and Ostfeld, 2008). The connection between the system states could be defined by such models with variables (i.e. input, internal and output variables), where only a few number of assumptions are considered with respect to the physical performance of the system. The more suited examples of data-driven models are the rating curves, the unit hydrograph, the various statistical (i.e. Linear Regression; LR, multi-linear regression models, Auto Regressive Integrated Moving Average; ARIMA) and the machine learning models (Solomatine and Ostfeld, 2008).

The orthodox black box time series models such as ARIMA, ARIMA with exogenous input (ARIMAX) and Multiple Linear Regression (MLR) are linear in nature and assume stationary dataset. These models remain incapable to handle the non-stationarity and non-linearity for hydrological processes. Which forced the researchers to use the soft-computing models (AI). The AI based models possess a slight edge over statistical-based models such as ANN.

The broader applications of AI methods include Fuzzy, NNs, and SVM models, which are mostly used in different areas of hydrology. Since the appearance of AI based models remain very active in hydrology, the efficient performance of these techniques such as data-driven models has been observed and respectively
published over a wide range of hydrological processes (e.g., precipitation, streamflow, rainfall–runoff, sediment load, groundwater, drought, snowmelt, evapotranspiration, water quality, etc.). The prominent researchers involved in this research area for the last decade and so with the number of publications significantly. The success of these application can be observed by these successful applications for hydrological process modeling (e.g., stream-flow, rainfall–runoff, sediment, groundwater, water quality). All these applications use ANN, Fuzzy, and SVM. Notwithstanding the flexibility and usefulness of AI-based methods in modeling hydrological processes. These AI-based models do possess some drawbacks with highly non-stationary responses, which may differ at wide scale of frequencies. In such cases the ‘seasonality’, a lack of input/output data preprocessing, may not count the AI models to handle non-stationary data with suitability.

Hence, these AI models in all their different application forms also have their own shortcomings and disadvantages. For example, ANN often suffers from local minima and over-fitting, while other soft-computing models, such as SVM, including ANN, are sensitive to parameter selection (Wang et al., 2008). As a result, researchers made an attempt to move away from the application of one stage mathematical or computational models and turned to various hybrid approaches (two stage or three stage structures). It was believed and assumed that hybrid models which combine data preprocessing schemes with AI techniques can play an important role. For example, (Kisi, 2009) and (Sang, 2013) combined wavelets with ANNs to predict the stream flow time series. Their underlying assumption was to use wavelets as a preprocessing technique in order to decompose data so that the issue of non-linearity can be addressed. Wavelet change joined with ANN as information preprocessing strategy can be seen to accomplish higher demonstrating exactness and consistency in various lead time ahead. The wavelet changed information help in enhancing the model execution by catching supportive data on different determination levels. Due to the aforementioned favorable circumstances of wavelet change, it has been found that the hybridization of wavelet change with other AI models like SVM, ANN, ANFIS, straight models, and so forth., enhanced the outcomes altogether than the single consistent model (Prahlada and Deka, 2011). To a large extent, the technique was successful, but unfortunately the effectiveness of wavelets was affected because hydrological time series has noises and show complex
characteristics due to uncertainty of the environment (Sang et al. 2009). Similarly, the wavelet MLR model is not facilitated with automatic updating and hence is not able to adapt to the changing river discharge patterns effectively (Kisi, 2008). In addition, a major drawback of wavelet transform for direction prediction is that the input variables lie in a high-dimensional feature space depending on the number of sub-time series components.

To account for the deficiencies in the all the above mentioned one stage or multi stage hybrid models, this research proposes wavelets based three stage (Wavelets+KPCA+LSSVM) forecasting structure for modeling river flow in Pakistan and Malaysia. First, it is proposed to decompose data using wavelets and then use KPCA for reduced dimensionality, and de-noising of data. (Lee et al., 2004). Once the data is ready to be trained after the sequential application of these two methods, it is preferred to use LSSVM instead of ANN and LR in the present study. ANN and LR has shown some modeling errors like over fitting. On the other hand, LSSVM is considered to be a better data trainer for non-linear data (Ismail et al., 2010). Therefore, the proposed 3-stage model is expected to show more accurate and precise modeling.

Given the afore mentioned limitations of one stage and two stage models, the present study aims to address the following issues related with hydrological time series.

i. How to design a three-layer architecture model based on wavelet as the decomposition method with KPCA technique for dimensional reduction or feature extraction and combined with LSSVM?

ii. Will the proposed Wavelet-KPCA-LSSVM improve the modeling accuracy and at the same time outperform other models?

iii. As the Benchmark LSSVM and Wavelet-LSSVM are employed in other modeling area, can Benchmark LSSVM and Wavelet-LSSVM be employed in the river flow modeling?
1.3 Objectives

In view of the above-mentioned problems, this study intended to propose the three-stage-architecture model based on Wavelet, KPCA and LSSVM to predict monthly stream flow data in Pakistan and Malaysia.

The objectives of the proposed hybrid model are:

i. To develop a hybrid model this is the combination of two independent techniques, i.e. Discrete Wavelets Transform (DWT) with Least Squared Support Vector Machine (LSSVM) for river flow.

ii. To design and develop a model based on Wavelet-KPCA-LSSVM, which combines decomposition, data pre-processing and forecasting techniques for river flow forecasting.

iii. To compare the performance of the hybrid models LSSVM with WLSSVM and WKPLSSVM and the benchmarked individual model LSSVM.

1.4 Scope of Study

The scope of this study covers the procedure of data-driven modelling, which involves analysis of problem, data collection, data pre-processing, model selection, model identification, and evaluation. Which includes:

i. The research focused on proposing a new method for time series forecasting of WAVELT-KPCA-LSSVM, which combines the decomposition technique with KPCA as the data pre-processing technique and LSSVM as a forecasting tool.

ii. Real time series data of monthly river flows are taken from Pakistan and Malaysia from four different rivers that are selected as the case studies.

iii. Radial basis function is selected as the kernel function for LSSVM models.
iv. The newly obtained data set from KPCA are set within two-cut-off values, which are from 70% to 95%.

v. The performance measurement for accuracy prediction is based on the standard statistical performance evaluation such as mean percentage error (MAE), root mean squared error (RMSE) and Nash-Sutcliffe coefficient efficiency (CE). RMSE and MAE are the most widely used performance evaluation criteria and the same will be used in this research.

1.5 Significance of the Study

Model building for the river flow is very significant because the heavy river flow can become the reason to problems (i.e. flooding and erosion). On other end the low river flow restricts the supply of water even for domestic use. In the regard the industrial and hydroelectric power are required to generate more. This study reviews the effectiveness of the proposed model as an alternative tool for model building. The designed research method attempts the decomposition technique with Wavelet and the data pre-processing technique with the help of KPCA model. As the original data are decomposed into numerous signals. In the regards the KPCA is apply to dimensional reduction, and the newly obtained data are then used to know the river flow.

As this study is to provide the accuracy with precision of time with respect to stream flow value based on past time series data. The proposed model requires the Wavelet-KPCA-LSSVM to understand the monthly river flow in Pakistan and Malaysia in order to produce a better result. This helps to provide a better understanding of the trend of the river flow in Pakistan and Malaysia.
1.6 Thesis Contribution

The present research proposed a model that can use non-linear set of data and produce an estimated value. In real life situation the collected data is the result of some non-linear process. The conventional models for addressing hydrological modelling problem were found to be good in estimating values when the input data is linear. Whereas, there performance and credibility comes under question when the input data is non-linear.

In this regard a hybrid model will be proposed for meeting this objective. This hybrid model is prepared by combining three independent techniques, i.e. Discrete Wavelets Transform (DWT), Kernel Principal Component analysis (KPCA) and Least Squared Support Vector Machine (LSSVM). The role of DWT is to decompose the original data at three levels, given as input of KPCA. Then KPCA is used to minimize the dimensionality of high dimensional input vectors and finally the predicted values are obtained by LSSVM. The significance of the hybrid model is that it is time efficient and readily suits for the environments which are time critical. The proposed model is regarded as a prototype for an early warning model. The prime purpose of proposed work is to generate warnings well before time in order to help and create opportunities to save valuable human lives and assets as well.

1.7 Thesis Structure and Organization

This section gives a brief outline of this thesis. Its contexts principally comprise six chapters, each of which is summarized as follows:

Chapter 1: Introduction

This chapter present the research background, identifies research problems, defines the aim and scope of study, describe research methodology, and outlines chapters in the thesis.
Chapter 2: Literature Review

This chapter presents past and current studies in the research area pertaining time series and forecasting models including individual and hybrid models, and evaluating the advantages and disadvantages of the existing solutions. The selected techniques used in the proposed hybrid nonlinear-linear models are discussed.

Chapter 3: Research Methodology

This chapter describes the operational framework of the study detailing of each step in this research i.e, design and development, and also testing and validation of the models.

Chapter 4: Single LSSVM and Proposed Hybrid Wavelet_LSSVM Models

The chapter begins with an overview of hydrological modelling more generally and then focuses on univariate forecasting for monthly streamflow series using data-driven models coupled with data pre-processing techniques. The main focus of this chapter to investigate model performance of single model LSSVM and hybrid model WLSSVM. In the end of the chapter to compare LSSVM and WLSSVM models with each other. Their comparative performance are evaluated using three statistical test; MAE, RMSE, and CE.

Chapter 5: Proposed Hybrid Models Wavelet-KPCA-LSSVM

Chapter 5 details out the explanation on development and integration of the hybrid models Wavelet-LSSVM using KPCA. In the first part of chapter is proposed three-stage-architecture of WKPLSSVM model for river flow forecasting. The statistical performances used in the study are also described in details.


Nalley, D., Adamowski, J., Khalil, B. 2012. Using discrete wavelet transforms to analyze trends in streamflow and precipitation in Quebec and Ontario (1954-


