PROBABILISTIC AND ARTIFICIAL INTELLIGENCE MODELLING OF DROUGHT AND AGRICULTURAL CROP YIELD IN PAKISTAN

A thesis submitted by

Mumtaz Ali

BSc, MSc, M.Phil

For the award of

Doctor of Philosophy

2019

Abstract

Pakistan is a drought-prone agricultural nation where hydro-meteorological imbalance and increasing scarcity of water resources are immensely constraining water, leading to a reduction in agricultural productivity. Rainfall and drought are imperative matters of consideration, both for hydrological and agricultural applications. The naturally chaotic characteristics of meteorological and agronomic variables can exhibit non-linearity and non-stationarity, leading to significant challenges in any model to generate reliable forecasts. Another key challenge that could significantly affect the accuracy and applicability is the selection of the pertinent features. To feasibly emulate the future rainfall, drought and crop yield, probabilistic and artificial intelligence forecasting models are promising tools in the modern era of data science.

The aim of this doctoral thesis is to design new hybridized probabilistic and artificial intelligence forecasting models for rainfall, drought and crop yields within the agricultural hubs in Pakistan. The choice of these study regions is a strategic decision, to focus on precision agriculture given the importance of this particular province in socio-economic activities of Pakistan, a nation where agriculture is considered to be the lifeblood of its people. This thesis is constructed upon three primary objectives to design probabilistic and machine learning models with subsequent robust evaluations of these models by means of statistical score metrics and diagnostic plots.

The first objective establishes the online sequential extreme learning machines (OSELM) model hybridized with the Markov Chain Monte Carlo based copula models and the artificial intelligence-based Bat algorithm was used to develop the MCMC-Cop-

Bat-OS-ELM hybrid model to forecast monthly rainfall. A total of 25 different Markov Chain Monte Carlo based copula models were developed at a first stage. The Bat algorithm was used to select the best copula model incorporated in OSELM algorithm.

The second objective of this research has three distinct milestones. Firstly, an ensemble modelling strategy was designed. An ensemble-ANFIS based uncertainty assessment modeling approach was constructed for medium and long term (i.e., 3-, 6-, 12-months) drought forecasting horizons at three diverse geographic locations. Secondly in this objective, a new ELM based committee of modeling approach is developed for the purpose of short term drought forecasting using multiple climatological inputs. Thirdly, a new multivariate approach utilizing the multivariate empirical mode decomposition (MEMD) algorithm to enable multiple predictor inputs is devised to forecast near-real-time short and long-term drought with kernel ridge regression (KRR) and random forest (RF) models. Feature optimization process was also carried out with the simulated annealing (SA) selection algorithm.

Objective 3 of this thesis consists of two major findings. This involves a new approach for cotton yield prediction with Markov Chain Monte Carlo-based simulation model integrated with the genetic programing algorithm (i.e. GP-MCMC-Cop) using multiple meteorological data in the agricultural zones in Punjab and Sindh province. While the second outcome was to develop a universal two-phase hybrid ACO-OSELM model using OSELM and a feature based input selection ant colony optimization (ACO) algorithm to predict wheat yield within the Punjab province.

In conclusion, this study ascertains the potential applicability of probabilistic and artificial intelligence predictive models in hydrology and crop yields. The major implications of this research thesis is that the current forecasting models developed at multiple horizons are likely to help design and reinforce new scientific tools as well as constructing knowledge-based systems for precision agriculture, climate change adaption policy formulation and major decisions made by agronomists, government and other stakeholders.

Certification of Thesis

This Thesis is entirely the work of ***Mumtaz Ali*** except where otherwise acknowledged. The work is original and has not previously been submitted for any other award, except where acknowledged.

Mumtaz Ali — ***18/04/2019***
PhD Candidate — Date

Endorsements

Dr. Ravinesh C. Deo — ***18/04/2019***
Principal Supervisor — Date

Dr. Nathan J. Downs — ***18/04/2019***
Associate Supervisor — Date

Assoc. Prof. Tek Maraseni — ***18/04/2019***
Associate Supervisor — Date

Students and supervisors signatures of endorsement are held at the University.

Statement of Contribution

The articles produced from this doctoral thesis were a joint contribution of the authors. The articles generated from this doctoral thesis were a joint contribution of the student and supervisory team with a majority of the research led and completed by the student. The field of research is in national priority areas are: 'Agriculture, Land and Farm Management FOR 0701, Environmental Science and Management FOR 0502 and Information and Computing Sciences FOR 08'. The details of the scientific contribution of each author are provided below:

- **Article 1: Mumtaz Ali,** Ravinesh C. Deo, Nathan J. Downs and Tek Maraseni, (2018), "Multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula- Bat algorithm for rainfall forecasting", *Atmospheric Research*. Volume 213 (2018), Pages 450-464. **(Q1: Impact Factor 3.817; SNIP 1.447; 89th percentile).**

  The overall contribution of Mumtaz Ali for this article were (70%) to the establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript. The contribution of Ravinesh C. Deo was 20% to supervised and assisted in scientific methodological development with important technical inputs, editing, and co-authorship of the manuscript. The contribution of Nathan J. Downs (5%) and Tek Maraseni (5%) was included editing and proofreading of the manuscript.

- **Article 2: Mumtaz Ali,** Ravinesh C. Deo, Nathan J. Downs and Tek Maraseni, (2018), "An Ensemble-ANFIS based Uncertainty Assessment Model for Forecasting multi-scalar Standardized Precipitation Index", *Atmospheric Research*. Volume 207 (2018), Pages 155-180. **(Q1: Impact Factor 3.817; SNIP 1.447; 89th percentile).**

  The percentage contributions of Mumtaz Ali for this article were (70%) to the establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript. The contribution of Ravinesh C. Deo was 20% to supervised and assisted in scientific methodological development with important technical inputs, editing, and co-authorship of the manuscript. The contribution of Nathan J. Downs (5%) and Tek Maraseni (5%) was included editing and proofreading of the manuscript.

- **Article 3: Mumtaz Ali,** Ravinesh C. Deo, Nathan J. Downs and Tek Maraseni, (2018), "Multi-stage committee based extreme learning machine model incorporating the influence of climate parameters and seasonality on drought forecasting", *Computers and Electronics in Agriculture*, Volume 152 (2018), Pages 14-165. **(Q1: Impact Factor 2.427; SNIP 1.563; 96th percentile).**

  The contributions of Mumtaz Ali were (70%) to the establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript. The contribution of Ravinesh C. Deo was 20% to

supervised and assisted in scientific methodological development with important technical inputs, editing, and co-authorship of the manuscript. The contribution of Nathan J. Downs (5%) and Tek Maraseni (5%) was included editing and proofreading of the manuscript.

- **Article 4: Mumtaz Ali,** Ravinesh C. Deo, Nathan J. Downs and Tek Maraseni, (2019), "Improving SPI-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridized with simulated annealing and kernel ridge regression algorithms", *Journal of Hydrology*. Volume 576 (2019), Pages 164-186. **(Q1: Impact Factor 4.405; SNIP 1.917; 95th percentile).**

  The percentage contributions for this paper are (Mumtaz Ali 80%), (Ravinesh C. Deo 10%), (Nathan J. Downs 5%) and (Tek Maraseni 5%). Mumtaz Ali contributed to the development of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript. Ravinesh C. Deo was contributed to supervised and assisted in scientific methodological development with important technical inputs, editing, and co-authorship of the manuscript. The contribution of Nathan J. Downs and Tek Maraseni was editing and proofreading of the manuscript.

- **Article 5: Mumtaz Ali,** Ravinesh C. Deo, Nathan J. Downs and Tek Maraseni, (2019), "Cotton yield prediction with Markov Chain Monte Carlo simulation

model integrated with genetic programing: a new hybrid copula-driven approach," *Agricultural and Forest Meteorology*, Volume 263 (2018), Pages 428-448. **(Q1: Impact Factor 4.039; SNIP 1.794; 98th percentile).**

The contributions of Mumtaz Ali were (70%) to the establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript. The contribution of Ravinesh C. Deo was 20% to supervised and assisted in scientific methodological development with important technical inputs, editing, and co-authorship of the manuscript. The contribution of Nathan J. Downs (5%) and Tek Maraseni (5%) was included editing and proofreading of the manuscript.

- **Article 6: Mumtaz Ali,** Ravinesh C. Deo, Tek Maraseni and Nathan J. Downs, (2019), "Two-phase ant colony optimized algorithm integrated with online sequential extreme learning machine to predict wheat yield", *IEEE Access* **(Under 2nd review; Q1: Impact Factor 3.557; SNIP 1.758; 97th percentile).**

The percentage contributions of Mumtaz Ali for this article were (75%) to the establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript. The contribution of Ravinesh C. Deo was 20% to supervised and assisted in scientific methodological development with important technical inputs, editing, and co-authorship of the manuscript. The contribution of Tek Maraseni (5%) was included editing and proofreading of the manuscript.

Additional Articles from Doctoral Research

- **Article 7: Mumtaz Ali,** Ravinesh C. Deo and Tek Maraseni, (2019), "Forecasting long-term precipitation for water resource management: A new multi-step, non-dominated sorting genetic algorithm, integrated with singular valued decomposition and random forest hybrid model, *Stochastic Environmental Research and Risk Assessment* **(Under review; Q1: Impact Factor 2.668; SNIP 1.173; 89th percentile).**

  The contributions of Mumtaz Ali were (75%) to the establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript. The contribution of Ravinesh C. Deo was 20% to supervised and assisted in scientific methodological development with important technical inputs, editing, and co-authorship of the manuscript. The contribution of Tek Maraseni (5%) was included editing and proofreading of the manuscript.

- **Article 8: Mumtaz Ali,** Ravinesh C. Deo and Tek Maraseni, (2019), "Incorporating climate data to estimate standardized precipitation index using data intelligent models, *Water Resource Management* **(Under review; Q1: Impact Factor 2.644; SNIP 1.276; 86th percentile).**

  The percentage contributions of Mumtaz Ali for this article were (75%) to the establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript. The contribution of Ravinesh C. Deo

was 20% to supervised and assisted in scientific methodological development with important technical inputs, editing, and co-authorship of the manuscript. The contribution of Tek Maraseni (5%) was included editing and proofreading of the manuscript.

Book Chapter

- **Article 9: Mumtaz Ali,** Ravinesh C. Deo and Tek Maraseni, (2019) Monthly Rainfall Forecasting using Markov Chain Monte Carlo Simulation Techniques integrated with Statistical Bivariate Copulas, In: Samui P, Bu DT, Chakraborty S, Deo RC. (Eds.) (September 2019), *Handbook of Probabilistic Models.* New York, NY: Elsevier (2019).

  The contributions of Mumtaz Ali were (85%) to the establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript. The contribution of Ravinesh C. Deo was 10% to supervised and assisted in scientific methodological development with important technical inputs, editing, and co-authorship of the manuscript. The contribution of Tek Maraseni (5%) was included editing and proofreading of the manuscript.

- **Article 10: Mumtaz Ali** and Ravinesh C. Deo, (2019) Modelling wheat yield with data-intelligent algorithms: artificial neural network versus genetic programming and minimax probability machine regression, In: Samui P, Bu DT,

Chakraborty S, Deo RC. (Eds.) (September 2019), *Handbook of Probabilistic Models*. New York, NY: Elsevier (2019).

The contributions of Mumtaz Ali were (85%) to the establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript. The contribution of Ravinesh C. Deo was 15% to supervised and assisted in scientific methodological development with important technical inputs, editing, proofreading and co-authorship of the manuscript.

List of Publications

Refereed Journal Articles

1. **Mumtaz Ali,** Ravinesh C. Deo, Nathan J. Downs and Tek Maraseni, (2018), "Multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula- Bat algorithm for rainfall forecasting", *Atmospheric Research*. Volume 213 (2018), Pages 450-464. **(Q1: Impact Factor 3.817; SNIP 1.447, Scopus Rank = 89th percentile).**
2. **Mumtaz Ali,** Ravinesh C. Deo, Nathan J. Downs and Tek Maraseni, (2018), "An Ensemble-ANFIS based Uncertainty Assessment Model for Forecasting multi-scalar Standardized Precipitation Index", *Atmospheric Research*. Volume 207 (2018), Pages 155-180. **(Q1: Impact Factor 3.817; SNIP 1.447, Scopus Rank = 89th percentile).**
3. **Mumtaz Ali,** Ravinesh C. Deo, Nathan J. Downs and Tek Maraseni, (2018), "Multi-stage committee based extreme learning machine model incorporating the influence of climate parameters and seasonality on drought forecasting", *Computers and Electronics in Agriculture.* Volume 152 (2018), Pages 14-165. **(Q1: Impact Factor 2.427; SNIP 1.563, Scopus Rank = 96th percentile).**
4. **Mumtaz Ali,** Ravinesh C. Deo, Nathan J. Downs and Tek Maraseni, (2019), "Improving SPI-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridized with simulated annealing and kernel ridge regression algorithms",

Journal of Hydrology. Volume 576 (2019). Pages 164-184. **(Q1: Impact Factor 4.405; SNIP 1.917, Scopus Rank = 95th percentile).**

5. **Mumtaz Ali,** Ravinesh C. Deo, Nathan J. Downs and Tek Maraseni, (2019), "Cotton yield prediction with Markov Chain Monte Carlo simulation model integrated with genetic programing: a new hybrid copula-driven approach," *Agricultural and Forest Meteorology*, Volume 263 (2018), Pages 428-448. **(Q1: Impact Factor 4.039; SNIP 1.794, Scopus Rank = 98th percentile).**
6. **Mumtaz Ali,** Ravinesh C. Deo, Tek Maraseni and Nathan J. Downs, (2019), "Two-phase ant colony optimized algorithm integrated with online sequential extreme learning machine to predict wheat yield", *IEEE Access* **(Under 2nd minor revision; Q1: Impact Factor 3.557; SNIP 1.758, Scopus Rank = 86th percentile).**
7. **Mumtaz Ali,** Ravinesh C. Deo and Tek Maraseni, (2019), "Forecasting long-term precipitation for water resource management: A new multi-step, non-dominated sorting genetic algorithm, integrated with singular valued decomposition and random forest hybrid model, *Stochastic Environmental Research and Risk Assessment* **(Under review; Q1: Impact Factor 2.668; SNIP 1.173, Scopus Rank = 89th percentile).**
8. **Mumtaz Ali,** Ravinesh C. Deo and Tek Maraseni, (2019), "Incorporating climate data to estimate standardized precipitation index using data intelligent models, Water Resource Management **(Under review; Q1: Impact Factor 2.644; SNIP 1.276, Scopus Rank = 86th percentile).**

Book Chapters

9. **Mumtaz Ali,** Ravinesh C. Deo, Nathan Downs and Tek Maraseni, (2019) Monthly Rainfall Forecasting using Markov Chain Monte Carlo Simulation Techniques integrated with Statistical Bivariate Copulas, In: Samui P, Bu DT, Chakraborty S, Deo RC. (Eds.) (September 2019), *Handbook of Probabilistic Models*. New York, NY: Elsevier (2019).
10. **Mumtaz Ali** and Ravinesh C. Deo, (2019) Modelling wheat yield with data-intelligent algorithms: artificial neural network versus genetic programming and minimax probability machine regression, In: Samui P, Bu DT, Chakraborty S, Deo RC. (Eds.) (September 2019), *Handbook of Probabilistic Models.* New York, NY: Elsevier (2019).

Acknowledgement

Please accept my great appreciation to these individuals and organizations for their whole-hearted assistance, scientific guidance, valuable and useful advice. In particular, I would like to express my gratitude and sincere appreciation to:

- **Dr Ravinesh C. Deo**, Principal Supervisor, and leader of *Advanced Data Analytics*: Environmental Modelling & Simulation Research Group for his motivation and guidance, and for patiently guiding me while I was learning the ropes of advanced probabilistic and artificial intelligence modelling and publishing in high-quality journals.
- **Dr Nathan J. Downs,** Associate Supervisor, to whom I am grateful for sharing valuable suggestions, plentiful comments and criticisms. His critical analysis helped me to figure out practical applications and directions of future research.
- **Associate Professor Tek Maraseni,** Associate Supervisor, for his scientific advice and expertise on manuscript preparation, interpretation of results and useful and critical comments in the journal articles and thesis.
- I am also grateful to University of Southern Queensland (USQ) Office of Research and Graduate Studies for awarding the (USQ-PRS, 2016 – 2019) scholarship to pursue this Ph.D. Without the funding, this research would not have eventuated.
- I would like to take this opportunity to thank all the organizations that provided free-to-access data including Pakistan Meteorological Department, Pakistan (PMD) and Federal Bureau of Statistics Islamabad, Pakistan.

Table of Contents

Abstract ..... **i**

Certification of Thesis ..... **iv**

Statement of Contribution ..... **v**

List of Publications ..... **xii**

Acknowledgement ..... **xv**

List of Figures ..... **xix**

List of Tables ..... **xxvii**

List of Acronyms ..... **xxxiv**

Hybrid Models Notations ..... **xxxvi**

Chapter 1: Introduction **1**

1.1: Background 1

1.2: Statement of the problem 4

1.3: Objectives 8

1.4: Significance of the Research study 11

1.5: Thesis layout 12

Chapter 2: Data and Methodology **17**

2.1: Study area 17

2.1.1: The Province of Punjab ***18***

2.1.2: The Province of Sindh ***19***

2.1.3: The Province of Khyber Pakhtunkhwa (KPK) ***20***

2.2: Data description 21

2.2.1: Meteorological data ***23***

2.2.2: Agricultural crop yield data *23*

2.2.3: Synoptic scale climate indices – various sources *23*

2.2.4: Periodicity (i.e. number of months) factor *24*

2.2.5: Multi-scaler standardized precipitation index (SPI) *24*

2.3: General methodology 24

Chapter 3: Multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm for rainfall forecasting **29**

Chapter 4: An ensemble-ANFIS based uncertainty assessment model for forecasting multi-scalar standardized precipitation index **46**

Chapter 5: Multi-stage committee based extreme learning machine model incorporating the influence of climate parameters and seasonality on drought forecasting **73**

Chapter 6: Improving SPI-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridized with simulated annealing and kernel ridge regression algorithms .. **91**

Chapter 7: Cotton yield prediction with Markov Chain Monte Carlo-based simulation model integrated with genetic programing algorithm: A new hybrid copula driven approach **113**

Chapter 8: Two-phase ant colony optimization algorithm integrated with online sequential extreme learning machine to predict wheat yield **135**

Chapter 9: Conclusion 161

9.1: Synthesis 161

9.2: Novel contributions of the study 167

9.3: Limitations of the current study and recommendations for future research 170

References 172

List of Figures

Chapter 1

Figure 1.1 Schematic view of this PhD research thesis.

Chapter 2

Figure 2.1 Famous tourist areas are Shangrila Lake, Cold Desert, Skardu, Nattar valley and Neelum valley.

Figure 2.1 Map of Pakistan showing the provinces of Punjab, Sindh and Khyber Pakhtunkhwa (KPK).

Chapter 3 (Objective 1, Article 1 – Published, *Atmospheric Research*, vol. 213 (2018), Pages. 459–464)

Figure 1 Map of the agricultural regions in Pakistan showing the sites considered for the development of the multi-stage hybrid MCMC-Cop-Bat-OS-ELM model.

Figure 2 Partial autocorrelation function (PACF) of historical monthly rainfall time-series in the model's training phase for Faisalabad, Multan, and Jhelum stations used in this study. The green dashed lines denote the statistically significant boundary at the 95% confidence interval.

Figure 3 Flow chart of the proposed multi-stage, hybrid MCMC based copulas integrated with Bat algorithm and OS-ELM model.

Figure 4 Scatterplot of the forecasted and observed rainfall (mm) in the testing phase for the proposed hybrid MCMC-Cop-Bat-OS-ELM, MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models using the 1-month (t – 1) significantly lagged data including the coefficient of determination (r2) and a linear fit inserted in each panel for study zones (a) Faisalabad, (b) Jhelum and (c) Multan.

Figure 5: Empirical cumulative distribution function (ECDF) of the forecasted error |FE| (mm) generated by the proposed hybrid MCMC-Cop-Bat-OS-ELM model versus its counterpart models in the testing period for Faisalabad, Jhelum and Multan stations.

Figure 6 Box-plots of the forecasted error |FE| (mm) in testing phase for the proposed hybrid MCMC-Cop-Bat-OS-ELM model with counterpart models of monthly forecasted rainfall for (a) Faisalabad, (b) Jhelum and (c) Multan.

Figure 7 Taylor diagram showing the correlation coefficient between the observed and forecasted rainfall and standard deviation for the proposed hybrid multi-stage MCMC-Cop-Bat-OS-Elm model in comparison with MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models for the station (a) Faisalabad, (b) Jhelum, and (c) Multan.

Chapter 4 (Objective 2, Article 1 – Published, *Atmospheric Research*, vol. 207 (2018), Pages. 155–180)

Figure 1 Basic structure of: (a) Ensemble-ANFIS, (b) MPMR and (c) M5 Tree models.

Figure 2 The 1 and 3 month Standardized Precipitation Index (SPI) with rainfall data for drought periods (Sept 2009 to Dec 2010) at Islamabad station, Pakistan.

Figure 3 Map of the study sites.

Figure 4 Partial autocorrelation function (PACF) of historical SPI time-series based on 3, 6, and 12 month scale for stations: (a) D. I. Khan, (b) Islamabad, (c) Nawabshah. The blue line denotes the statistically significant boundary at the 95% confidence interval.

Figure 5 Ensemble-ANFIS model results attained from 10-fold simulations, analyzed in terms of the maximum and minimum absolute forecasting

error (AFE) (i.e., bars) and the average of 10 model ensemble forecasts (i.e., green) compared with observed SPI for 3, 6, and 12 month scale data for testing months. (a) D. I. Khan, (b) Islamabad, (c) Nawabshah.

Figure 6 (LEFT) Times-series of Ensemble-ANFIS vs. M5Tree and MPMR models-generated absolute forecasting error (AFE). (RIGHT) Scatterplot of the forecasted and observed SPI in the testing phase based on 3, 6 and 12 month scales (a) D. I. Khan, (b) Islamabad, and (c) Nawabshah. For each scatterplot, the least square fitting line and its respective correlation coefficient is shown.

Figure 7 Cumulative frequency of errors generated by the ensemble-ANFIS vs. M5Tree and MPMR model- based on the 3, 6 and 12 month SPI forecasts. (a) D. I. Khan, (b) Islamabad, (c) Nawabshah. Note that the percentage is shown in the respective error bracket.

Figure 8 Boxplot of the distribution of absolute forecasting error (AFE) generated by the ensemble-ANFIS vs. M5 Tree and the MPMR model 3, 6 and 12 month scale SPI forecasts. (a) D. I. Khan, (b) Islamabad, (c) Nawabshah.

Chapter 5 (Objective 2, Article 2 – Published, *Computers and Electronics in Agriculture*, vol. 152 (2018), Pages 14-165)

Figure 1 Schematic view of the (a) ELM and (b) PSO-ANFIS models.

Figure 2 Monthly standardized precipitation index showing drought characteristics and the correspondence with precipitation (PCN) data of the drought period (May 1991 to Mar 1992) for Multan Station. The unit of PCN is millimeter (mm).

Figure 3 Map of the study regions Islamabad, Multan and D. I. Khan in Pakistan.

Figure 4 Schematic of the proposed multi-stage, Comm-ELM mode (blue) *vs.* Comm-PSO-ANFIS (green) and Comm-MLR (green) models.

Figure 5 Scatterplot of the forecasted (SPIFOR) and observed (SPIOBS) data in the testing phase using the multi-stage, Comm-ELM vs. Comm-PSO-ANFIS and Comm-MLR models with periodicity (i.e. month) as an input parameter with the coefficient of determination (r2) inserted in each panel for study zones (a) Islamabad, (b) Multan and (c) Dera Ismail Khan.

Figure 6 Empirical cumulative distribution function (ECDF) of the forecast error, |FE| in the testing period using Comm-ELM vs. Comm-PSO-ANFIS and Comm-MLR models for the stations (a) Islamabad, (b) Multan and (c) Dera Ismail Khan.

Figure 7 Taylor diagram showing the correlation coefficient between observed and forecasted SPI and standard deviation of the proposed Comm-ELM vs. Comm-PSO-ANFIS and Comm-MLR models for the stations (a) Islamabad, (b) Multan and (c) Dera Ismail Khan.

Figure 8 Box-plots of forecasted error |FE| of Comm-ELM vs. Comm-PSO-ANFIS and Comm-MLR models forecasted monthly SPI index with periodicity (left) in comparison with non-periodicity (right) for Islamabad, Multan and Dera Ismail Khan.

Figure 9 Polar plots shows the monthly average values of |FE| generated from the Comm-ELM vs. Comm-PSO-ANFIS and Comm-MLR models in forecasting SPI for Site 1: Islamabad, Site 2: Multan and Site 3: D. I. Khan.

Chapter 6 (Objective 2, Article 3 – Published, *Journal of Hydrology*)

Figure 1 Map of the selected study locations in Pakistan.

Figure 2 Schematic structure of the proposed multi-phase (MEMD-SA-KRR) model integrating multivariate empirical mode decomposition (MEMD)

at phase 1 and simulated annealing (SA) at phase 2 with kernel ridge regression (KRR) model at phase 4.

Figure 3 Empirical cumulative distribution function (ECDF) of forecasted error |FE| 1-month (SP1), 3-month (SPI3), 6-month (SPI6) and 12-month (SPI12) generated by the proposed multi-phase MEMD-SA-KRR vs. MEMD-SA-RF, Standalone KRR and Standalone RF models for (a): Faisalabad, (b): Islamabad, and (c): Jhelum.

Figure 4 Box-plots of forecasted error |FE| of 1-month (SP1), 3-month (SPI3), 6-month (SPI6) and 12-month (SPI12) in testing period generated by multi-phase MEMD-SA-RF vs. MEMD-SA-RF, Standalone KRR and Standalone RF models for (a): Faisalabad, (b): Islamabad, and (c): Jhelum.

Figure 5 Cumulative frequency of 1-month (SP1), 3-month (SPI3), 6-month (SPI6) and 12-month (SPI12) generated by MEMD-SA-KRR vs. MEMD-SA-RF, standalone KRR and standalone RF models of |FE| error for (a): Faisalabad, (b): Islamabad, and (c): Jhelum.

Figure 6 Taylor plots depicting the predictive skill of MEMD-SA-RF vs. MEMD-SA-RF, Standalone KRR and Standalone RF models for (a): Faisalabad, (b): Islamabad, and (c): Jhelum in the testing period of 1-month (SP1), 3-month (SPI3), 6-month (SPI6) and 12-month (SPI12).

Figure 7 The 1-month, 3-month, 6-month and 12-month observed vs. forecasted SPI in the testing period generated by all four models for (a): Faisalabad, (b): Islamabad, and (c): Jhelum. Least square regression line with the coefficient of determination (r2) is shown.

Chapter 7 (Objective 3, Article 1 – Published, *Agricultural and Forest Meteorology*, vol. 263 (2018), Pages 428-448)

Figure 1 Map of the selected study region in Pakistan.

Figure 2 A flow chart of the hybrid genetic programming algorithm integrated with a Markov Chain Monte Carlo based copula model.

Figure 3 Marginal distribution of average climate (temperature, rainfall, humidity), GP based predicted cotton yield (Cpred) and observed cotton yield (Cobs) with Kendall's tau (red) for (a) Multan, (c) Nawabshah and (c) Faisalabad station.

Figure 4 Empirical cumulative probability distribution (ECP) of average (temperature, rainfall, humidity), GP based predicted cotton yield (Cpred) and observed cotton yield (Cobs) with Kendall' tau (fit) for (a) Multan station, (b) Nawabshah station and (c) Faisalabad station.

Figure 5 Joint and marginal distribution of uniformly distributed average (temperature, rainfall, humidity), GP based predicted cotton yield (Cpred) and observed cotton yield (Cobs) for (a) Multan station, (b) Nawabshah station and (c) Faisalabad station.

Figure 6 Dependence structure of the GP-MCMC-copula based prediction cotton yield, MCMC based copula with average temperature, rainfall, humidity climate parameters versus the observed cotton yield for (a) Multan, (b) Nawabshah and (c) Faisalabad. *Note: Both the GP based predicted cotton yield, climate parameters (x-axis) and the cotton yield (y-axis) are presented in the probability space. Blue lines present the copula isolines and green circles show the observed normalized cotton yield.*

Figure 7 Posterior distributions of the GP copulas derived by MCMC simulation within a Bayesian framework for the stations (a) Multan, (b) Nawabshah and (c) Faisalabad. Blue asterisks show the copula parameter derived by the local optimization while the green circles show the parameters derived by theoretical optimization. The red bins are the MCMC-derived parameters and the aqua (□) crosses show the maximum likelihood

parameter of the MCMC. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

Figure 8 Predicted (C_{pred}) and observed (C_{obs}) cotton yield generated by GP-MCMC copulas against MCMC copulas and GP models in the seasons of the testing period for the stations (a) Multan, (b) Nawabshah and (c) Faisalabad. Note that the bars shows the absolute prediction errors, PE = $|C_{obs} - C_{pred}|$.

Figure 9 Scatterplot of predicted (Cpred) and observed (Cobs) cotton yield using the GP-MCMC-copula, MCMC-copula and standalone GP models, with the coefficient of determination (r^2) inserted in each panel of study zones. (a) Multan (b) Nawabshah and (c) Faisalabad.

Figure 10 Box-plots of absolute prediction error (APE, kg/ha) of GP-MCMC based copula models vs. MCMC based copula models and the standalone GP model in predicted cotton yield for the stations for (a) Multan (b) Nawabshah, and (c) Faisalabad.

Chapter 8 (Objective 3, Article 2 – Submitted, under revision: *IEEE Access*)

Figure 1 Map of the study region. (a) Provinces of Pakistan. (b) Districts of Punjab where the present study was undertaken. (c) Selected training sites in red and the corresponding test site in yellow. Note that the sites shown in green have 'no available wheat yield data' and those in blue were not selected by the Ant Colony Optimisation algorithm.

Figure 2 Bar graphs of the root mean squared error (RMSE) encountered by the Ant Colony Optimisation algorithm in the selection of training study sites for each testing study site: Site 1: Rahimyar Khan, Site 2: D. G. Khan, Site 3: Kasur, Site 4: Sialkot, Site 5: Rawalpindi, and Site 6: Jhang.

Figure 3 Time series of the annual wheat yield data for the training stations selected by the Ant Colony Optimisation algorithm for each testing study

site. Site 1: Rahimyar Khan, Site 2: D. G. Khan, Site 3: Kasur, Site 4: Sialkot, Site 5: Rawalpindi, and Site 6: Jhang.

Figure 4 Partial autocorrelation function correlation coefficient (PACF) of the historical annual wheat yield time series for each testing study site: Site 1: Rahimyar Khan, Site 2: D. G. Khan, Site 3: Kasur, Site 4: Sialkot, Site 5: Rawalpindi, and Site 6: Jhang.

Figure 5 Flow chart of the proposed hybrid two-phase Ant Colony Optimization algorithm integrated with Online Sequential Extreme Learning Machine (OSELM) model.

Figure 6 Scatterplots of the predicted (Wpred) and observed wheat yield (Wobs) (kgha-1) in the testing phase of the ACO-OSELM vs. ACO-ELM and ACO-RF models including the coefficient of determination (r2) and a linear fit inserted in each panel for the tested study zones.

Figure 7 Box-plots of the prediction error |PE| (kgha-1) of ACO-OSELM vs. ACO-ELM and ACO-RF models between the predicted and observed wheat yield for Site 1: Rahimyar Khan, Site 2: D. G. Khan, Site 3: Kasur, Site 4: Sialkot, Site 5: Rawalpindi, and Site 6: Jhang.

Figure 8 Empirical cumulative distribution function (ECDF) of the prediction error, |PE| (kgha-1) for the testing stations using ACO-OSELM vs. ACO-ELM and ACO-RF models.

Figure 9 Polar plots showing the prediction error |PE| in each year generated from the ACO-OSELM vs. ACO-ELM and ACO-RF models in predicting wheat yield for Site 1: Rahimyar Khan, Site 2: D. G. Khan, Site 3: Kasur, Site 4: Sialkot, Site 5: Rawalpindi and Site 6: Jhang.

List of Tables

Chapter 2

Table 2.1 Details of all data used in this PhD research thesis.

Chapter 3 (Objective 1, Article 1 – Published, *Atmospheric Research*, vol. 213 (2018), Pages. 459–464)

Table 1 Descriptive statistics of the study sites.

Table 2 The performance evaluation of the proposed hybrid MCMC-Cop-Bat-OS-ELM in respect to the hybrid MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models in the training period based on the mean square error (MSE) and correlation coefficient (r).

Table 3 The selected Markov Chain Monte Carlo (MCMC) based copula models ranked by the Bat algorithm. Note that a total of 25 MCMC-copula models were used here.

Table 4 Local and MCMC estimated copula parameters with the 95% confidence of interval (CI).

Table 5 Evaluation of the MCMC based copulas models constructed with model data significant at a lag (t – 1) using the Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC) and Maximum Likelihood (Max_L) criterion for each station.

Table 6 Evaluation of hybridized multi-stage MCMC-Cop-Bat-OS-ELM vs. MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models using root mean square error (RMSE; mm) and mean absolute error (MAE; mm), correlation coefficient (r), Willmott index (WI), and Nash-Sutcliffe coefficient (NSE) in the testing period. The best model is boldfaced (red).

Table 7: Geographic comparison of hybridized multi-stage MCMC-Cop-Bat-OS-ELM *vs.* MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF model using relative root mean squared error (*RRMSE; %*), relative mean absolute error (*RMAE; %*) and the Legates & McCabe's Index (*LM*). The best model is boldfaced (red).

Chapter 4 (Objective 2, Article 1 – Published, *Atmospheric Research*, vol. 207 (2018), Pages. 155–180)

Table 1 Descriptive statistics of the study sites' geographic, hydrologic and drought characteristics.

Table 2 Training performance of 10-member ('ensemble-ANFIS') model with correlation coefficient (r) and mean square error (MSE). Note: The inputs are based on statistically significant lagged data, SPI (t – 1).

Table 3 Training performance of M5 Tree and MPMR models. Note: The inputs are based on statistically significant lagged data, SPI (t – 1).

Table 4 Ensemble-ANFIS model evaluated in testing phase with Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Correlation Coefficient (r), Willmott Index (WI), Nash-Sutcliff (NSE) and Legates & McCabes Index (LM) for 3, 6, and 12 month forecasts. The best model is boldfaced (blue).

Table 5 Comparison of 10-member average of the Ensemble-ANFIS vs. the M5 Tree and MPMR-based SPI forecasting performance. Note that the best model is boldfaced (blue).

Table 6 The performance of the ensemble-ANIS model in all station for quantifying drought, as measured by the actual difference ( ) between the forecasted and the observed properties of drought over the testing period; (Severity, S ≡ accumulated negative SPI after the onset of

drought is detected (SPI<0); Intensity, I ≡ minimum value of the SPI; Duration, D ≡ sum of the consecutive months in which drought status is sustained for all stations. Note that the average model is boldfaced (blue).

Table 7: Ninety five percentage of confidence band using ensemble-ANFIS model for low, and upper forecasted multi-scale drought duration property for the stations (a) Dera Ismail Khan, (b) Islamabad, and (c) Nawabshah. Note than Per = Percentile, $\Delta = SPI_{For} - SPI_{Obs}$, and FDP (m ± e) = Forecasted Duration Property (mean ± error).

Chapter 5 (Objective 2, Article 2 – Published, *Computers and Electronics in Agriculture,* vol. 152 (2018), Pages 14-165)

Table 1 Descriptive statistics of the study sites' geographic, drought and hydrologic characteristics over the study period (1981 to 2015).

Table 2 Training performance of Comm-ELM vs. Comm-PSO-ANFIS and Comm-MLR models with correlation coefficient (r) and mean square error (MSE). [Definition of the model inputs as stated in Section 3.1 are: T = mean monthly temperature; PCN = Precipitation; H = mean relative humidity; SOI = Southern Oscillation Index and MP = monthly cycle or periodicity].

Table 3 Influence of the model input combinations applied for forecasting of monthly SPI using the Comm-ELM vs. Comm-PSO-ANFIS and Comm-MLR models measured by root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (r). Note that the best model is boldfaced (blue).

Table 4 Evaluation of the Comm-ELM vs. Comm-PSO-ANFIS and Comm-MLR models using Willmott's index (WI), Nash-Sutcliffe (ENS) and Legates-McCabe's (LM) agreement, for (a) Islamabad,

(b) Multan; and (c) Dera Ismail Khan. Note that the best model is boldfaced (blue).

Table 5 Geographic comparison of the study regions using Comm-ELM vs. Comm-PSO-ANFIS and Comm-MLR models using relative root mean squared error (RRMSE, %) and the relative mean absolute error (RMAE, %) computed within the test sites. Note that the best model is boldfaced (blue).

Chapter 6 (Objective 2, Article 3 – Published, *Journal of Hydrology*)

Table 1 Summary of the basic statistics of the meteorological predictors, large scale climate mode indices and target variable of the study locations. Further, std. denotes standard deviation, skew represents skewness, and kurt indicates kurtosis.

Table 2 Design parameters involved to decomposed IMFs and residuals for training and testing period in each study site using multivariate empirical mode decomposition (MEMD) method.

Table 3 Design parameters involving in the selected IMFs for training period for each study site using simulate annealing (SA) algorithm. The number of total IMFs is also given.

Table 4 Performance of training period using multivariate empirical mode decomposition model hybridized with simulated annealing and Kernel ridge regression (i.e. MEMD-SA-KRR) model vs. MEMD-SA-RF, Standalone KRR and Standalone RF models in terms of r and MSE. The choice of Kernel types were: polynomial, linear and Gaussian.

Table 5 Multi-scale analysis of testing period using the MEMD-SA-KRR vs. MEMD-SA-RF, Standalone KRR, and Standalone RF models

measured by RMSE, MAE and r. The optimum model is blue bold faced.

Table 6 Multi-scale analysis in testing period of MEMD-SA-KRR *vs.* MEMD-SA-RF, Standalone KRR and Standalone RF models using E_{WI}, E_{NS} and E_{LM}. Note that the best model is boldfaced (blue).

Table 7 Geographic evaluation of the MEMD-SA-KRR vs. MEMD-SA-RF, Standalone KRR and Standalone RF models using relative percentage error (RPE, %). Note that the best model is boldfaced (blue).

Chapter 7 (Objective 3, Article 1 – Published, Agricultural and Forest Meteorology, vol. 263 (2018), Pages 428-448)

Table 1 Descriptive statistics of the study sites' geographic, crop and hydrologic characteristics.

Table 2 Local and Markov-Chain Monte Carlo (MCMC) copula parameters with 95% confidence interval (CI) for genetic programming (i.e., GP-MCMC) based copula model and Markov-Chain Monte Carlo (MCMC) based copula model for predicting cotton yield. The parameters (para) with respective the CI is boldfaced (blue) and all models inputs are: T=temperature, R=rainfall, H=humidity.

Table 3 Testing performance of the genetic programming-based Markov Chain Copula Model (GP-MCMC) vs. the MCMC based copula and a standalone GP model in in terms of the Root Mean Squared Error (RMSE, kg/ha), Mean Absolute Error (MAE, kg/ha), Correlation Coefficient (r), Nash-Sutcliff (NSE), Willmott Index (WI), including Akiake information criterion (AIC), Bayesian information criterion (BIC) and Maximum-likelihood (MaxL) for the prediction of cotton yield. The optimal model is boldfaced (blue).

Table 4 Evaluation of the GP-MCMC vs. the MCMC and a standalone GP model using the relative root mean squared error (RRMSE), relative mean absolute error (RMAE) and Legates & McCabe's Index (LM). The optimal model is boldfaced (blue).

Table 5 Geographical comparison showing the performance skills of the GP-MCMC vs. the MCMC based best copula models and a standalone GP model using the relative root mean squared error (RRMSE), relative mean absolute error (RMAE) and the Legates & McCabe's Index (LM). Note that the best model is boldfaced (blue).

Chapter 8 (Objective 3, Article 2 – Submitted, under revision: *IEEE Access*)

Table 1 Geographic properties and wheat yield statistics of the study sites for Punjab, Pakistan.

Table 2 Selected training stations using Ant Colony Optimization (ACO) algorithm with the correlation coefficient (r) for each training station against the testing station.

Table 3 Training data points (in terms of selected training sites) and testing data point for each testing site using ACO algorithm with skewness and kurtosis of training and testing data.

Table 4 Training performance of two-phase hybrid ACO-OSELM *vs.* ACO-ELM and ACO-RF models with correlation coefficient (*r*) and root mean squared error (*RMSE, kgha*$^{-1}$).

Table 5 Testing performance of ACO-OSELM vs. ACO-ELM and ACO-RF models measured by root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (r).

Table 6 The performance of ACO-OSELM vs. ACO-ELM and ACO-RF models using Willmott's index (WI), Nash-Sutcliffe (NSE) and Legates-McCabe's (LM) agreement, for Site 1: Rahimyar Khan, Site 2: D. G.

Khan, Site 3: Kasur, Site 4: Sialkot, Site 5: Rawalpindi and Site 6: Jhang. Note that the best model is boldfaced (blue).

Table 7 Geographic comparison of the accuracy of the ACO-OSELM vs. ACO-ELM and ACO-RF models in terms of relative root mean squared error (RRMSE, %) and the relative mean absolute error (RMAE, %) computed within the test sites. . Note that the best model is boldfaced (blue).

List of Acronyms

PMD	Pakistan Meteorological Department
KPK	Khyber Pakhtunkhwa
D. I. Khan	Dera Ismail Khan
PBS	Pakistan Bureau of Statistics
GDP	Gross Domestic Product
IPCC	International Panel for Climate Change
SOI	Southern Oscillation Index
SST	Sea Surface Temperatures
PDO	Pacific Decadal Oscillation
IOD	Indian Ocean Dipole
EMI	El-Nino southern oscillation Modoki index
SAM	Southern Annular Mode
JISAO	Joint Institute of the Study of the Atmosphere and Ocean
BMA	Bureau of Meteorology, Australia
JAMSTEC	Japan Agency for Marine-Earth Science
BAS	British Antarctic Survey
SPI	Standardized Precipitation Index
PACF	Partial Auto-Correlation Function
MCMC-Cop	Markov Chain Monte Carlo based copula
OSELM	Online Sequential Extreme Learning Machine
ELM	Extreme Learning Machine
RF	Random Forest
ANFIS	Adaptive Neuro Fuzzy Inference System
M5Tree	M5 Tree Model
MPMR	Minimax Probability Machine Regression
ANN	Artificial Neural Network
PSO-ANFIS	Particle Swarm Optimization based ANFIS
MLR	Multiple Linear Regression

EMD	Empirical Mode Decomposition
EEMD	Ensemble Empirical Mode Decomposition
CEEMD	Complete Ensemble Empirical Mode Decomposition
MEMD	Multivariate Empirical Mode Decomposition
SA	Simulated Annealing
KRR	Kernel Ridge Regression
IMF	Intrinsic Mode Function
GP	Genetic Programming
ACO	Ant Colony Optimization
$\boldsymbol{NS_E}$	Nash–Sutcliffe Efficiency
FE	Forecasting Error
LM	Legates-McCabe's Index
MAE	Mean Absolute Error
RMAE	Relative Mean Absolute Error
MSE	Mean Squared Error
r	Pearson's Correlation coefficient
RMSE	Root-Mean-Square-Error
RRMSE	Relative Root-Mean-Square Error
WI	Willmott's Index of Agreement
$\boldsymbol{Max_L}$	Likelihood value
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CI	Confidence of Interval

Hybrid Models Notations

MCMC-Cop-Bat-OS-ELM	Online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm
MCMC-Cop-Bat-ELM	Extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm
MCMC-Cop-Bat-RF	Randome forest integrated with Markov Chain Monte Carlo copula-Bat algorithm
E-ANFIS	Ensemble-ANFIS
Comm-ELM	ELM based committee model
Comm-PSO-ANFIS	PSO-ANFIS based committee model
Comm-MLR	MLR based committee model
MEMD-SA-KRR	Multivariate emperical mode decomposition integrated with simulated annealing and Kernel ridge regression model
MEMD-SA-RF	Multivariate emperical mode decomposition integrated with simulated annealing and random forest model
GP-MCMC-Cop	Genetic programming integrated with Markov Chain Monte Carlo copula
ACO-OSELM	Ant colony optimzation model coupled with online sequential extreme learning machine
ACO-ELM	Ant colony optimzation model coupled extreme learning machine
ACO-RF	Ant colony optimzation model coupled with random forest

Chapter 1

Introduction

1.1: Background

The recent variability in long-term climate (e.g., seasonal) and short-term (weather patterns) due to natural variability and anthropogenic factors has significant impacts on increased vulnerability of agricultural crop yield and water resources. The chaotic behaviours of these climatological events induce nonlinearity and non-stationarity within the crop yields. Extreme weather events such as excess rainfall, droughts, hail, heatwaves and extreme temperatures have often caused significant impacts on the crop yield, even in high agricultural yield and technological countries. Since the inter-annual climate variations explain a third of global crop yield variability, it caused huge challenges to the global food security in ensuring sufficient food for the increasing population in the 21st century. Therefore, it is urgent for improvements in the understanding of climate risks on the crop yield in order to minimize the climate-related impact and support agricultural and water resource managers in strategy development and decision-making to avoid any probable catastrophes.

Notably, a significant change in the rainfall events can generally affect economic growth, particularly in developing countries (Odusola and Abidoye, 2015). Extreme precipitation can also have severe impacts on the world climate (Kundzewicz et al., 2006). Anthropogenic and naturally-induced aberrations in regional-scale rainfall can directly affect the agricultural sector since rainfall plays a vital role in both the growth and the production of crops (Maraseni et al., 2012). Additionally, it also brings major water-related disasters (Barredo, 2007) such as the shortage of rainfall on the long run can lead to water scarcity (Langridge et al., 2006;

Vörösmarty et al., 2010) while excessive amounts of rainfall can cause flooding and damage to human and wildlife health, infrastructure and the economy (Bhalme and Mooley, 1980).

Prolonged precipitation/rainfall deficits with a series of dry spell epochs cause meteorological drought events. Drought is characterized as a climatological menace that can occur in arid, semi-arid, or tropical rainforest zones (Keyantash and Dracup, 2002; Vicente-Serrano, 2016; Wilhite et al., 2000). Drought events can last from short to long period ranging from one month to four years as recent climate change significantly affects rainfall patterns (Vicente-Serrano, 2016). Drought severely disturbs water resources, agriculture crops, energy supply and industrial sectors, and it is a growing concern (Deo et al., 2009; IPCC, 2012; McAlpine et al., 2007; Yaseen et al., 2018b). Long-term droughts significantly pose challenges to groundwater reservoirs and cause significant water scarcity (Cai and Cowan, 2008) and the related socio-economic costs (Dijk et al., 2013; Wittwer et al., 2002).

Climate variability is one of the most important factors affecting year-to-year crop production and subsequent revenues (Deo et al., 2009; IPCC, 2012; McAlpine et al., 2007; Yaseen et al., 2018). In particular, extreme weather events such as droughts, hail, heatwaves and extreme temperatures have often caused significant impacts on the crop yield, even in high agricultural yield and technology countries. Since the inter-annual climate variations explain a third of global crop yield variability, it caused huge challenges to the global food security in ensuring sufficient food for the increasing population in the 21st century. Therefore, it is urgent for improvements in the understanding of climate risks on the crop yield in order to minimize the climate-related impact and support agricultural managers in strategy development and decision-making.

The ability to forecast rainfall and drought events and crop yields under climate change in an accurate manner, particularly in agricultural belt regions, can increase the ability of stakeholders to formulate better water planning and resource management decisions. Hence, the forecasted rainfall and drought information are important in managing hydrological and agricultural drought events, and designing early warning systems. Precise and reliable future information on rainfall would assist in constructing of prudent and timely procedures and techniques for optimal distribution and utilization of water for industrial, hydro-electricity generation and recreation. In addition, the advanced or projected knowledge of the rainfall, drought, and crop yield at micro-scale would allow farmers and farm managers to make proactive sustainable decisions for efficient irrigation, grazing, water quality monitoring, yield predictions (Gill et al., 2006) and be wary of seasonal cropping. This information has the potential of being cascaded into the design of knowledge-based systems for monitoring water resources, flood events and empowers precision agriculture.

Recent advances in computational capacity have allowed for application of the probabilistic and machine learning based predictive models in many areas. The full dependence structure between climate variables and crop yield captured by the copula approach will provide better appreciate climate risks and impacts and useful information to agricultural and climate modeler in terms of managing climate risks. While the data-driven models extract pertinent predictive features from historical data sets. Since forecasting is an important aspect of climatological, hydrological and agricultural sustainability, it is an open area of research. Largely, a systematic layered improvement has been the key element in technological evolutions and is the way to develop newer models for hydrological and agricultural applications as well. Therefore, new and advanced predictive models hybridized with copula and feature optimization

and multi-resolution analysis approaches are being explored in this study to forecast rainfall and drought as well as agricultural crop yield prediction within Pakistan's agricultural hubs.

1.2: Statement of the problem

Anthropogenic and naturally-induced anomalies in regional-scale rainfall can directly affect the agricultural crops (Maraseni et al., 2012), water scarcity (Langridge et al., 2006; Vörösmarty et al., 2010) while excessive amounts of rainfall can cause flooding and damage to human and wildlife health, infrastructure and the economy (Bhalme and Mooley, 1980). Drought is a socio-economic hazard poses severe threats to groundwater reservoirs, leading to the scarcity of water, crop failure, disturbed habitats and loss of social or recreational opportunity (Deo et al., 2015; Mpelasoka et al., 2008; Riebsame et al., 1991; Wilhite et al., 2000). The occurrence of drought leads to consequences for runoff that affects stream flow in agricultural sectors (Cai and Cowan, 2008) with substantial economic costs (Dijk et al., 2013; Wittwer et al., 2002).

Pakistan is an agricultural nation, among top ten countries suffering from global warming (Pachauri et al., 2014). An abrupt change in rainfall trend over Pakistan has been observed in the last few years (Aamir and Hassan, 2018) which leads to major flooding events that severely damage the economy including infrastructure and agricultural crops (News, 2010). The estimated damage in the 2010 event to infrastructure was approximately 4 billion US dollars whereas the damage in the agricultural sector amounted to about 500 million US dollars (Hicks and Burton, 2010). Excessive rainfall events in 2010 and 2017 caused damage costs of approximately 43 billion US dollars (NOAA, 2017; WMO, 2017) and 500 billion US dollars, respectively (Mansoor, 2010; Tarakzai, 2010). Equally, drought events have been a major contributing factor towards reduced agricultural crop yields and significant reductions of the gross domestic product of Pakistan. Thus, for facilitating prudent strategies for water resources management and

mitigation of drought impacts on agriculture and its repercussions, it is imperative for hydrologists, agriculturalists, and resource planners to develop effective modelling and prediction techniques for rainfall and drought events.

As Pakistan is an agriculture nation and agriculture is known to contribute to about 21% of the county's GDP (Sarwar, 2014). Important crops grown in Pakistan are wheat, rice, cotton, sugarcane, maize, different vegetables and fruits. Cotton is an integral commodity for the economic development of Pakistan as the nation is highly dependent on the cotton industry and its related textile sector due to which the cotton crop has been given a principal status in the country. In the past, the prediction of cotton yields has been based primarily on the effect of climate change with the adoption of traditional approaches (Ayaz et al., 2015; Hina Ali, 2013) for a large area, either for a whole province, or national region, but not for a small locality.

In the past, Pakistan has faced significant crises of wheat supply, particularly in the period of 2012-2013, which occurred due to the failure of the province of Punjab, to meet its target production value. A plausible reason for this deficit was attributed to the poor agricultural planning and inaccurate estimations to satisfy the national grain needs (Bokhari, 2013). A report published in the Express Tribune (Sajjad, 2017) indicates that, similar to the past experience, Pakistan is likely to further face wheat shortages into the future. In 2005, the actual yield in Pakistan was relatively low compared to the predicted yield, and as such, poor estimations have moderated the market price and prompted the government to export the grains from the international market (Dorosh and Salam, 2008; Niaz, 2014).

Due to such uncertainties that directly have a detrimental impact on income and food security for the already staggering economies of developing Pakistan, the government and

policymakers require improved forecast models to facilitate them to estimate the potential reductions and associated food security risks due to a shortage of wheat yield. This justifies the pivotal role of advance probabilistic and artificial intelligence models that can predict more accurately at a micro scale which can provide help for decision-making in precision agriculture and farming systems.

Robust probabilistic and artificial intelligence predictive models with better accuracies could serve as suitable alternatives for forecasting rainfall, drought and crop yields. However, the foremost and critical issues of selecting the non-redundant (and most important) input data remain a problem of interest for forecasters. This is because the use of irrelevant inputs can add unnecessary challenges in the model execution and consequently increases the model complexity whilst reducing the model's forecasting accuracy (Hejazi and Cai, 2009; Maier et al., 2010).

Additionally, rainfall trends, drought events and crop yields and the interrelated climatic/hydrological inputs exhibit a complex temporal behavior with non-stationarity features (e.g., trends, seasonal variations, periodicity and jumps in time-series) that can affect the preciseness of data-driven models (Adamowski and Chan, 2011; Adamowski et al., 2012). The copula models which can perform a careful assessment of the dependence structure in terms of probability can be applied to ameliorate this issue in predictive model development. The hybridization of probabilistic and machine learning models can certainly improve the accuracy utilizing the probabilistic predictive features.

The ensemble based uncertainty assessment modelling approach can also enabled uncertainty between multi-models to be rationalized more efficiently, leading to a reduction in forecast error caused by stochasticity in drought behaviours. In this research thesis, a 10-member

ensemble-ANFIS modelling approach has been trailed. Model combinations are also very uncommon in climatological applications and have been overlooked in environmental applications (Baker and Ellison, 2008).

In this thesis, a new model combination based on "*The wisdom of crowds*" philosophy is developed. The notion is to extract the pertinent information simulated by the standalone expert models and generate a collective forecast. The conventional model combinations required simple averaging of forecasts from various models. However, the weaknesses of combinations based on simple averaging is that the overall model performance is compromised by the worst performing model(s). On the other hand, the committee based models approach could overcome the inherent drawbacks of individual standalone models, building on the aptness, and subsequently surpassing the individual performances (Barzegar et al., 2018; Rostami et al., 2014). In this study, committee based extreme learning machine (Comm-ELM) model is developed and evaluated.

Owing to the variability in climate-based for a drought model, a suite of multi-resolution analytical tools can be useful to extract embedded features in a non-static time series signal that are related to a drought variable, and thus, they may help to improve an existing data intelligent model. The multivariate empirical mode decomposition (MEMD) is an advance generalized form of EMD and CEEMD which demarcates multivariate inputs to performs accurate investigation of composite and nonlinear procedures (Rehman and Mandic, 2009). Additionally the MEMD fixes the mode alignment problems arise in the joint analysis within a multi-dimensional data (Looney and Mandic, 2009). Due to data dependent and self-adaptive nature (Alvanitopoulos *et al*., 2014), the MEMD is useful to extract relevant features without any loss of information and temporally preserve the physical structure of the inputs (Wu *et al*., 2011).

Another advantage of MEMD methods is that, it overcomes the non-stationarity and non-linearity problem via decomposition of the original time series.

The climate change is high influence on agricultural crops yields. Another hybrid copula based machine learning algorithm is developed utilized the several meteorological input variables. There is a statistically significant relationship with historical agricultural crop yield estimation and for this purpose, a feature based input selection artificial intelligence model is developed that conditioned to search for the suitable, statistically relevant data sites. Overall, this research thesis intends to address issues of appropriate input selection, non-linearity and non-stationarity of the input data in forecasting the rainfall and drought as well as crop yield prediction within the agricultural hubs in Pakistan.

1.3: Objectives

The primary aim of this research, presented as a collection of journal papers, was to develop a set of high-precision hybrid probabilistic and machine learning models for hydrological purposes (rainfall and drought forecasts) and agricultural crop yield prediction to analyze the inter-association between climate variables and crop yields.

Therefore, this PhD thesis, presented as a collection of 6 Quartile 1 (Q1) papers, has adopted various copula functions within a statistical modelling framework and the relevant data intelligent models to achieve following specific objectives:

Objectives 1: Develop Rainfall Forecasting Model

1. To develop hybridized online sequential extreme learning machine (OSELM) model with Markov Chain Monte Carlo based copula models and Bat algorithm to forecast monthly

rainfall. Several (i.e. 25) Markov Chain Monte Carlo based copula models were developed at first. The Bat algorithm was used to select the best four copula models that are incorporated in OSELM, ELM and RF models. The preciseness of the hybrid model was validated in respect to the ELM and RF models.

This work has been published in *Atmospheric Research* journal (Scopus Quartile 1).

Objectives 1: Design Drought Forecasting Model

This objective has three main findings in developing drought high precision forecasting models. These findings are discussed below:

1. Develop and explore an ensemble-ANFIS based uncertainty assessment modeling approach for medium and long term (3-, 6-, 12-months) drought forecasting. Applying 10-member simulations, ensemble-ANFIS model was validated for its ability to forecast severity (S), duration (D) and intensity (I) of drought. This enabled uncertainty between multi-models to be rationalized more efficiently, leading to a reduction in forecast error caused by stochasticity in drought behaviours. The results are benchmarked with the M5 Model Tree and Minimax Probability Machine Regressions (MPMR).

   This work has been published in *Atmospheric Research* journal (Scopus Quartile 1).
2. Develop and explore a new committee of modeling approach for short term (monthly) drought forecasting using multiple climatological inputs. Committee of modeling is a model combination technique, which is uncommon in climatological studies. In this study, the ELM-based committee was investigated and bench marked PSO-ANFIS-based committee and MLR-based committee models.

This work has been published in *Computers and Electronics in Agriculture* journal (Scopus Quartile 1).

3. Devise a new multivariate approach to MEMD modeling to allow for utilization of multiple predictor inputs. This new multivariate MEMD modeling technique has been developed and evaluated to forecast near-real-time short and long term drought with KRR and RF models. Feature optimization was carried out with the simulated annealing (SA) selection algorithm.

   This work has been published in *Journal of Hydrology* (Scopus Quartile 1).

Objective 3: Agriculture Crop Yield Prediction Modelling

It is clearly seen that this objective has two key outcomes. The first outcome indicates the use of multiple meteorological variables by developing a hybrid model to predict cotton yield. On the other hand, the second outcome is based on the development of a universal wheat yield data intelligent model utilizing wheat yield data at district level. These findings are following:

1. Cotton yield prediction with Markov Chain Monte Carlo-based simulation model integrated with genetic programing algorithm using multiple meteorological data of rainfall, temperature and humidity. Several different types of GP-MCMC-copula models were developed, each with the well-known copula families (i.e., Gaussian, student t, and Clayton, Gumble Frank and Fischer-Hinzmann functions) to screen and utilize an optimal cotton yield forecast model for the present study region.

   This work has been published in *Agricultural and Forest Meteorology* journal (Scopus Quartile 1).

2. To develop a two-phase hybrid OSELM, ELM and RF models using feature based input selection ant colony optimization (ACO) algorithm to predict wheat yield. The ACO algorithm is conditioned to search for the suitable, statistically relevant data sites for the model's training, and the corresponding testing sites by virtue of a feature selection strategy utilizing a total of 27 agricultural counties' datasets in the agro-ecological zones in Punjab province in Pakistan. The developed model can be explored as a decision-support tenet for crop yield estimation in regions where a statistically significant relationship with historical agricultural crop is well-established.

   This work has been reviewed once and is under second review in *IEEE Access* (Scopus Quartile 1).

1.4: Significance of the Research study

The findings of this research are significant as Pakistan is an agriculture nation. griculture in Pakistan is contributing about 21% of the county's GDP (Sarwar, 2014). Cotton is one of the cash crops which is an integral commodity for the economic development of Pakistan as the nation is highly dependent on the cotton industry and its related textile sector due to which the cotton crop has been given a principal status in the country. Cotton crop is grown from May-August as an industrial crop in 15% of the nation's available land area producing 15 million bales during 2014-15 (Reporter, 2015). Pakistan is fourth largest cotton grower, third largest exporter and fourth largest consumer (Banuri, 1998). In 2013, 1.6 million farmers (out of a total of 5 million in all sectors) engaged in cotton farming, growing more than 3 million hectares (Banuri, 1998; Reporter, 2015). On the other hand, wheat accounts for 2.6% of Pakistan's GDP and 12.5% to the GDP of the agronomy sector (Survey, 2012). According to United Nations Food and Agriculture Organization, Pakistan was placed in the eighth position as a global wheat

producer from 2007 to 2009 (FAO, 2013). Wheat accounts for nearly 36% of the total cropped area, 30% of the value added by major crops and 76% of the total production of food grains. The developed hybrid probabilistic and artificial intelligence models are vital for policy makers and governments for better future planning in relation to trade, development policies. Moreover, these modelling strategies can provide timely information for rapid decision-making during the growing season. Additionally, it can be used as early warning decision support systems for food scarcity, water scarcity due to high increase in population.

1.5: Thesis layout

The schematic diagram illustrating the overview Thesis is shown in Figure 1.1. It clearly outlines the graphical abstract for understanding and the need for reliable and precise forecasting tool for rainfall, drought and crop yield. This thesis is organized into nine distinct chapters, as follows:

Chapter 1

This chapter presents the introductory background and the statement of problem pertaining to the research and presents the objectives of this study.

Chapter 2

Chapter 2 describes the study area; data and general methodology used in this study and sets the scene for the following chapters. This Chapter provides general view points while the specific study area, data, and methods are presented in the respective chapters.

Chapter 3

This chapter is presented as a published journal article in the journal, *Atmospheric Research* (**https://doi.org/10.1016/j.atmosres.2018.07.005**). It is devoted to the establishment of Markov chain Monte Carlo based copula modelling hybridized with Bat

algorithm and online sequential extreme learning machine (MCMC-Cop-Bat-OS-ELM model for rainfall forecasting. It outlines the issues with traditional approaches, model development and outcomes with respect to comparative MCMC-Cop-Bat based extreme learning machine (MCMC-Cop-Bat-ELM) and MCMC-Cop-Bat based random forest (MCMC-Cop-Bat-RF). The modelling process involves the development of 25 members of MCMC based copula models. The selection of best MCMC based copula model was done by Bat algorithm and finally incorporated those selected member of MCMC-copula models in the OSELM model to forecast the rainfall. Chapter 3 addresses the first research objective of this study.

Chapter 4

This chapter is presented as a published article in the journal, *Atmospheric Research* (**https://doi.org/10.1016/j.atmosres.2018.02.024**). This chapter describes the application of ensemble modelling strategy using adaptive-neuro fuzzy inference system (ANFIS) approach for drought forecasting. Chapter 4 is the first outcome in response to the second research objective of this study whereby medium and long term drought is forecasted using newly developed a 10-member ensemble-ANFIS based uncertainty assessment model in comparison with M5tree and mini-max probability machine learning regression (MPMR) models. Applying 10-member simulations, ensemble-ANFIS model was validated for its ability to forecast severity (S), duration (D) and intensity (I) of drought. This enabled uncertainty between multi-models to be rationalized more efficiently, leading to a reduction in forecast error caused by stochasticity in drought behaviours.

Chapter 5

This chapter is presented as a published journal article in the journal, *Computers and Electronics in Agriculture* (**https://doi.org/10.1016/j.compag.2018.07.013**). This chapter is devoted to the application of committee modelling approach for short term drought forecasting. The monthly drought is forecasted using this innovative committee of models based on extreme learning machine (Comm-ELM) in comparison with Comm-PSO-ANFIS and Comm-MLR models. It outlines the model development and performances of these committee models. Chapter 5 captures the second outcome of research objective 2 of this study.

Chapter 6

This chapter is presented as published article in the top ranking *Journal of Hydrology* **(https://doi.org/10.1016/j.jhydrol.2019.06.032)**. The main theme of this chapter is to present the development of a novel multivariate empirical mode decomposition (MEMD) forecasting technique for 1-month, 3-month, 6-month and 12-month drought forecasting. The short, medium and long term drought is forecasted using MEMD simulated annealing (SA) with feature selection. It outlines the development of the novel MEMD-SA-KRR model and its performances with respect to a comparative MEMD-SA-RF) and the standalone KRR and RR models. This chapter targeted the issue and gaps in outcome 1 and 2 of objective 2. Chapter 6 turns out the third milestone of the second research objective of this study.

Chapter 7

This chapter is presented as a published journal article in the journal, *Agricultural and Forest Meteorology* (**https://doi.org/10.1016/j.agrformet.2018.09.002**). Cotton yield prediction with Markov Chain Monte Carlo-based simulation model integrated with

genetic programing algorithm using multiple meteorological data of rainfall, temperature and humidity. Several different types of GP-MCMC-copula models were developed, each with the well-known copula families (i.e., Gaussian, student t, and Clayton, Gumble Frank and Fischer-Hinzmann functions) to screen and utilize an optimal cotton yield forecast model for the present study region. Chapter 7 is the first finding in response to the third research objective.

Chapter 8

This chapter is presented as a submitted manuscript (under review) in *IEEE Access* journal. A two-phase hybrid OSELM model using feature based input selection ant colony optimization (ACO) algorithm to predict wheat yield. The ACO algorithm is conditioned to search for the suitable, statistically relevant data sites for the model's training, and the corresponding testing sites by virtue of a feature selection strategy utilizing a total of 27 agricultural counties' datasets in the agro-ecological zones in Punjab province in Pakistan. The developed model can be explored as a decision-support tenet for crop yield estimation in regions where a statistically significant relationship with historical agricultural crop is well-established. It outlines the model development (ACO-OSELM) and the outcomes benchmarked against comparative ACO-ELM and ACO-RF models. Chapter 8 is the second outcome in response to the third research objective.

Chapter 9

This chapter presents the synthesis of the study with concluding remarks, limitations, and recommendations for future works.

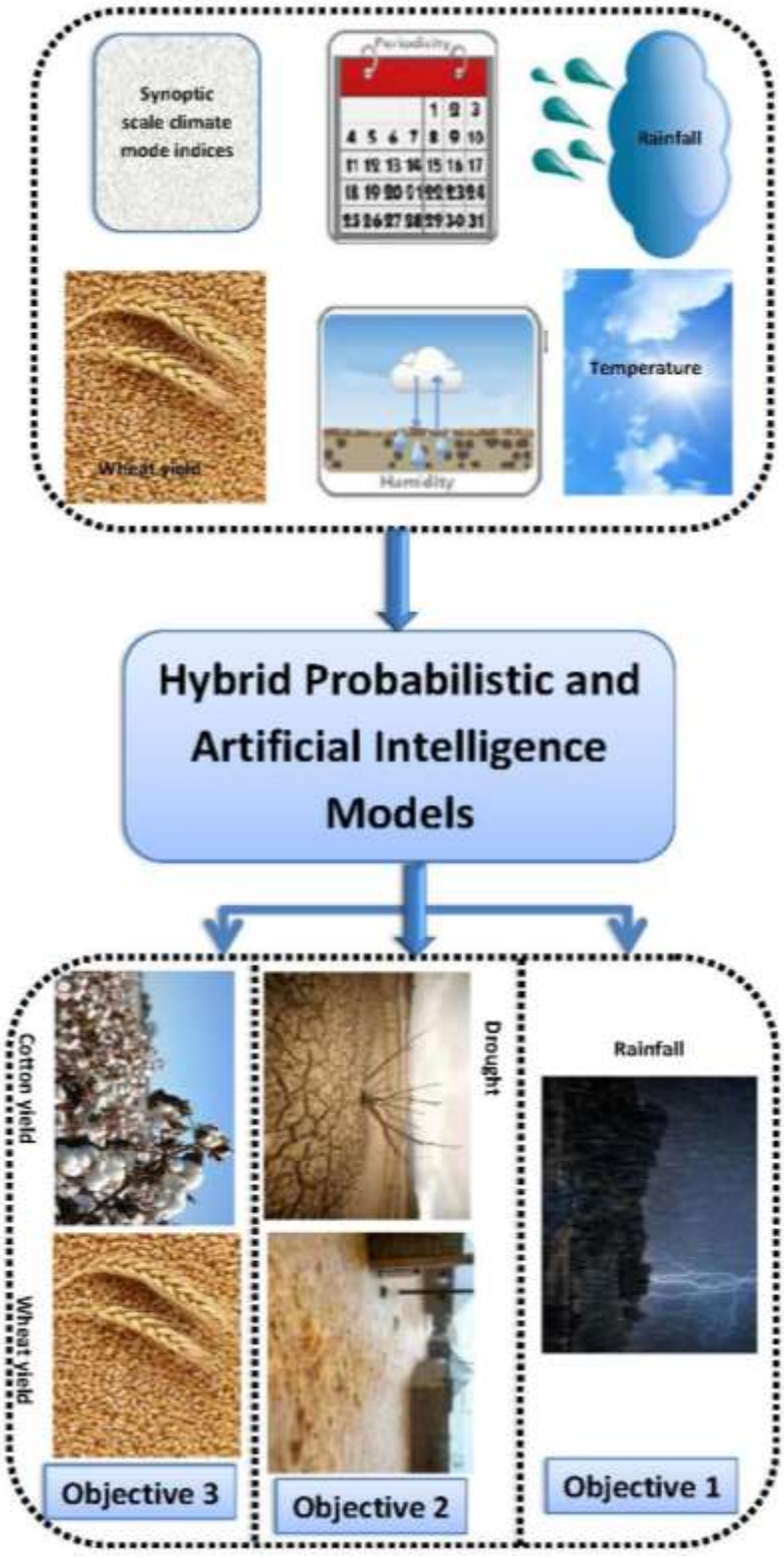


Figure 1.1: Schematic view of thesis.

Chapter 2

Data and Methodology

This chapter provides an overview of the study locations adopted in this research Thesis in developing the hybrid probabilistic and artificial intelligence predictive models. Different sites within the study region were selected to achieve each objective, which is described in detail in each of the chapters. The description of data used, length of data and limitations if any, are also presented. This chapter also introduces a brief account of methodology, while specific model development techniques have been described in respective chapters. The description of the study area is given next. This is followed by the data used and the general procedure used in this work for hybrid probabilistic and artificial intelligence models development.

2.1: Study area

The study areas focused in this thesis are the rich agricultural zones in three provinces of Punjab, Sindh and Khyber Pakhtunkhwa (KPK) located in Pakistan. Pakistan is an agriculture nation and agriculture contributes about 21% of national GDP. Important crops grown in Pakistan are wheat, rice, cotton, sugarcane, maize, different vegetables and fruits. The drought and floods have become a frequent phenomenon in Pakistan. Pakistan hosts five major deserts which were historic forests. Pakistan is also one of the best destinations for tourists and an estimated 1.1 million foreign tourists in 2011 contributing 351 million dollar to the economy. Some of the most beautiful and attractive tourist spots are:

Figure 2.1: Famous tourist areas are Shangrila Lake, Cold Desert, Skardu, Nattar valley and Neelum valley.

The importance of these provinces and selected study sites are discussed below in detail:

2.1.1: The Province of Punjab

Punjab is the second largest province by area of approximately 205,344 km^2 with a population of approximately 110 million (PBS, 2017). The capital city of Punjab is Lahore which is a historical, cultural, economic and industrial hub (PBS, 2017). Punjab is an industrialized province that contributes 24% to the province's GDP (strategy, 2018). It is also one of South Asia's most urbanized regions with approximately 40% of people living in urban areas(strategy, 2018). Punjab has total 36 districts that highly dependent on agriculture.

It has tropical wet and dry climate experiencing extreme weather conditions with foggy winter. The temperature varies between -2^{o}C to 45^{o}C. The landscape in Punjab mostly consists on alluvial plains surrounded by Indus, Jhelum, Chenab, Ravi and Sutlaj rivers. Agricultural sectors in Punjab province play a vital role in the economy with contributions ranging from 56.1% to 61.5% (strategy, 2018). Further, extensive irrigation systems make this region a rich agricultural hub. The major crops are cotton and wheat whereas other important crops include rice, sugarcane, millet, corn, oilseeds, pulses, vegetables, and fruits. Livestock and poultry production are also common. Punjab contributes approximately 76% to annual food grain production in Pakistan. Cotton and rice are considered to be important cash crops that contribute substantially to the national exchequer. Small and medium farming strategies are more practicing for the purpose of gaining independency in agriculture sector. Considering the region as a major agricultural belt, the development of hybrid probabilistic and artificial intelligence models for rainfall, drought and crop yield prediction is an interesting research endeavor.

2.1.2: The Province of Sindh

Sindh is located in the southeast of Pakistan, and is one of the third largest province by area and second largest province by population. According to the 2017 census, Sindh had approximately 47.9 million populations. Karachi is the capital of Sindh province which is the most populous city in Pakistan. Sindh is divided into 29 districts. The contribution of Sindh's province to country GDP is between 30% - 32%. The contribution in agriculture sector varies from 21.4% - 27.7% whereas manufacturing sector shares from 36.7% - 46.5%. Sindh has a much diversified economy varying from industry and finance to extensive agriculture sector.

The climate varies from tropical to subtropical which is hot in the summer and mild to warm in winter. The temperatures in summer normally rise above 46 °C while drops to 2 °C

during winter in the northern and higher elevated areas. The average annual rainfall is approximately 177.8 millimeter. The region's shortage of rainfall is compensated by the inundation of the melting of Himalayan snow and rainfall in the monsoon season. Agriculture crops such as cotton, rice, wheat, sugar cane, dates, bananas, and mangoes are very important in Sindh. Pakistan is the world's 4th largest Mango producer mostly grown in Sindh and Punjab provinces.

2.1.3: The Province of Khyber Pakhtunkhwa (KPK)

The province of KPK located in the northwest of Pakistan. It is the third largest province by population (40.5 millions) and third largest in economy as well (10.5% of Pakistan) whiles the smallest by area (101,741 km^2). The capital of KPK is Peshawar which is one of the oldest cities in the world. The northern areas in KPK are cold and snowy with heavy rainfall in winter. The KPK province is a central hub for tourism due to it snow-capped peaks and lush green beauty (Report, 2013).

The climate of KPK differs hugely for an area of its size that includes almost all of the climates encompassing found in Pakistan. Dera Ismail Khan is declared on of the hottest places in South Asia (NWFP, 2010). Agriculture is an important sector in KPK and highly depends on cash crops. The main cash crops include wheat, maize, tobacco, rice; sugar beets, sugarcane, as well as fruits are grown in the province. The part of the economy that KPK dominates is forestry, where its contribution has normally varies from of 34.9% to 81%, with an average of 61.56%. A Billion Tress Tsunami project was initiated in KPK province in 2014 to tackle the challenge of global warming.

Figure 3 shows these provinces and their geographic locations in Pakistan.

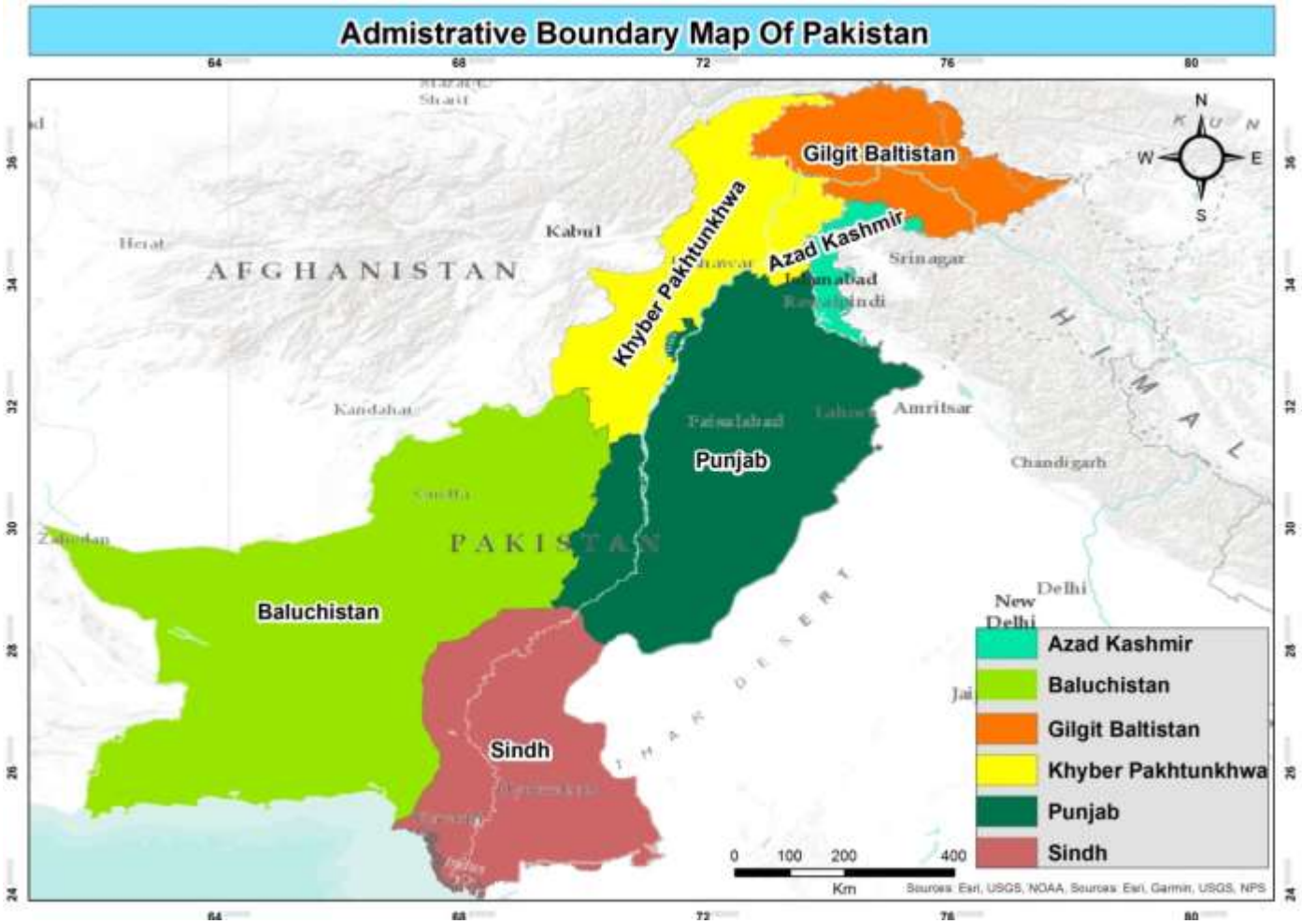


Figure 2.2: Map of Pakistan showing the provinces of Punjab, Sindh and KPK.

2.2: Data description

A variety of data sources was utilized in developing hybrid probabilistic and artificial intelligence forecasting models for rainfall, drought and agricultural crop yield. In a concise way, Table 2.1 describes the data used with respective sources and other relevant details in achieving each objective.

Table 2.1 Details of all data used in this research.

		Data used	Source	Study period	Forecast horizon
OBJECTIVE 1	Paper 1 (Chapter 3)	Predictors: Antecedent Lags of rainfall Target: Rainfall	Pakistan Meteorological Department (PMD)	January 1981 to December 2015	Monthly
OBJECTIVE 2	Paper 1 (Chapter 4)	Predictors: Rainfall and antecedent Lags of drought index (i.e. SPI) Target: SPl	Pakistan Meteorological Department (PMD)	January 1981 to December 2015	3-, 6- and 12 month
	Paper 2 (Chapter 5)	Predictors: Meteorological Variables and SOI index Target: SPl	Pakistan Meteorological Department (PMD), Bureau of Meteorology Australia (BMA)	January 1981 to December 2015	1-month
	Paper 3 (Chapter 6)	Predictors: Meteorological Variables and Synoptic scale climate mode indices Target: SPl	PMD, BMA and various other sources	January 1981 to December 2015	1-, 3-, 6- and 12 month
OBJECTIVE 3	Paper 1 (Chapter 7)	Predictors: Meteorological Variables Target: Cotton yield	PMD and Federal Bureau of Statistics, Pakistan	January 1981 to December 2013	Yearly (seasonal)
	Paper 2 (Chapter 8)	Predictors: Antecedent Lags of wheat yield Target: Wheat yield	Federal Bureau of Statistics, Pakistan	January 1981 to December 2013	Yearly (seasonal)

2.2.1: Meteorological data

The meteorological data were obtained from the Pakistan Meteorological Department, Pakistan (PMD, 2016). The study period is 1981 to 2015. The predictor inputs comprised of mean monthly rainfall (mm), mean monthly temperature (°C) and mean monthly humidity. The missing values of monthly rainfall were substituted by average of the respective time-averaged value from the climatological period because the rainfall for those months is not available.

2.2.2: Agricultural crop yield data

The agricultural crop yield data was sourced from the Federal Bureau of Statistics (Economic wing), Islamabad, Pakistan and Agriculture Marketing Information Service, Directorate of Agriculture (Economics & Marketing) Punjab, Lahore Pakistan (Districts, 2008; Service, 2012; Service, 2014). The area and production data of crop estimates were supplied by the provincial Crop Reporting Services and compiled by the Economic Wing of the devolved Ministry of Food and Agriculture and later by the Federal Bureau of Statistics. Crop yield in the year 2009 was not available in the acquired dataset. To overcome this situation, the average value of all the cotton yield data from 1981-2013 was substituted for the missing 2009 data. The acquired crop yield data were in the units of toons and hectares which are not the standard units. Therefore, the first task was to convert the yield data in the standard unit (kg/ha).

2.2.3: Synoptic scale climate indices – various sources

The monthly synoptic scale climate indices are southern oscillation index (SOI), sea surface temperatures (Nino3SST, Nino3.4SST, Nino4SST), pacific decadal oscillation (PDO), Indian ocean dipole (IOD), El-Nino southern oscillation Modoki index (EMI), southern annular mode (SAM) were sourced from National Climate Prediction Centre (Nicholls, 2004; SST, 2018), Joint Institute of the Study of the Atmosphere and Ocean (JISAO, 2018), Bureau of

Meteorology, Australia (BMA, 2018), Japan Agency for Marine-Earth Science (JAMSTEC, 2018) and from the British Antarctic Survey (BAS, 2018). Among these indices, the sea surface temperatures (SSTs) are the most important ones as they indicate climate variability, while the other indices (i.e., Pacific Decadal Oscillation (PDO), the Indian Ocean Dipole (IOD) and El Nino Modoki Index) are contingent upon them.

2.2.4: Periodicity (i.e. number of months) factor

The periodicity factor also plays a vital role in performance of the data intelligent models. In our case, it improves the accuracy of the hybrid machine learning models.

2.2.5: Multi-scaler standardized precipitation index (SPI)

The SPI quantifies the wet and dry scenarios based on statistical probability theory. The multi-scaler SPI index was computed by utilizing the approach in (McKee *et al.*, 1993).

2.3: General methodology

Prior to develop hybrid probabilistic and machine leaning models, data quality checking phase is necessary. A calendar averaging technique was applied to replace all missing data during this phase. The meteorological and crop yield data and the interrelated atmospheric parameters, as well as the climatic indices, naturally display stochastic behavior. In addition, the inputs are in the different set of units or are dimensionless. As a result, appropriate scaling or normalization is required to avoid the dominance of inputs with large numeric ranges that in turn may undermine the effects of lower range values. Normalization also brings the data to a common scale. The data are normalized between [0, 1] and due to invertible nature of the normalization, the results is not be affected (Hsu *et al*., 2003). The normalization is carried out following Eq. (1) to handle large variation in the data (Hsu *et al*., 2003).

$$\Theta_{norm} = \frac{\left(\Theta - \Theta_{\min}\right)}{\left(\Theta_{\max} - \Theta_{\min}\right)} \quad (1)$$

In Eq. (1), Θ represents the input/output, $\Theta_{\min}$ is the minimum value, $\Theta_{\max}$ is the maximum value of the data and Θ_{norm} is the corresponding normalized numeric value.

In this research, various forecasting models are considered for an evaluation of their preciseness in emulating rainfall, drought and crop yield since a robust modelling approach is necessary. The models range from the well-known Markov Chain Monte Carlo based copula (MCMC-copula), online sequential extreme learning machine (OSELM), extreme learning machine (ELM), random forest (RF), Kernel ridge regression (KRR), ensemble based adaptive neuro fuzzy inference system (ensemble-ANFIS), mini-max probability machine regression (MPMR), M5 Tree, particle swarm optimization based ANFIS (PSO-ANFIS), multiple linear regression (MLR) and genetic programming (GP) are adopted.

Copulas are powerful mathematical tools that have the ability to connect two or more time-independent variables (Nelsen, 2003) that provide a systematic way of observing the causal dependent structure which generates a basis for constructing families of bivariate (multivariate) distributions (Fischer and Hinzmann, 2006). ELM is a state-of-the-art data intelligent model developed by Huang et al. (2006) used for the purpose of designing a Single Layer Feedforward Neural Network (SLFN). ELM is relatively faster, and thus computationally efficient compared with other traditional learning algorithms (Rajesh and Prakash, 2011). As a variant of the classical ELM model, the online sequential extreme learning algorithm (OSELM) operates in two learning phases) i.e., initialization and a sequential learning phase. The M5 Model tree and random forest are regression tree based algorithms. However, the main difference is that the M5

Tree model is based on a single regression tree while the RF model uses an ensemble of regression trees with bootstrap-aggregation technique (Breiman, 1996; Mitchell, 1997; Quinlan, 1992). KRR model is based on kernels and a ridge regression approach (Zhang *et al*., 2013), which is used to deal with over-fitting in the regression using regularization and the kernel technique to capture non-linear relationships (You *et al*., 2018). ANFIS was introduced by (Jang, 1993) which is an improved ANN technique that is fundamentally identical to the fuzzy inference systems (FIS) model, yet utilizing the merits of both ANN and FIS designed on a common paradigm. MPMR, a non-linear probabilistic machine regression model that has the capability of maximizing the least probability within the interval of true regression of the objective function (Strohmann and Grudic, 2003). To improve the versatility of ANFIS model, particle swarm optimization (PSO) technique (Çavdar, 2016) to tune the ANFIS parameters. MLR is a generalized form of the simple regression model from single to multiple predictors where the objective is to deduce a model that can exhibit the maximum deviations in the predictor data to evaluate their corresponding regression coefficients equation (Draper and Smith, 1981; Montgomery et al., 2012). Genetic programming is a heuristic evolutionary algorithm which has the potential to offer solutions of any form without the user specifying the problem (Koza, 1992).

In order to handle the non-stationarity features within the inputs, data pre-processing via proper multi-resolution analysis tool is necessary. Hence, hybridized models with advanced self-adaptive multi-resolution tools including multivariate empirical mode decomposition (MEMD) are adopted. In addition, new approaches are developed and explored including a committee of models.

Appropriate input selection is imperative for input dimension reduction and improves the model performances. The optimization by means of feature selection approaches also has its own advantages and disadvantages and therefore many algorithms were explored including the linear partial-auto correlation function (PACF), Bat algorithm, simulated annealing (SA) algorithm, ant colony optimization (ACO) algorithm. In addition to the standalone approaches, the specific hybrid models developed in this study include:

1. Three hybrid models, MCMC-Cop-Bat-OS-ELM, MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF for monthly rainfall forecasting. PACF was utilized for determination of significant lags. Bat algorithm was utilized for selecting best copula model.
2. The 10-member ensemble based ANFIS model was designed to forecast medium and long term drought forecasting. PACF was utilized for determination of significant lags.
3. A novel ELM based committee of models (Comm-ELM), Comm-PSO-ANFIS and Comm-MLR for short term (monthly) drought forecasting with ELM, PSO-ANFIS and MLR as the underlying expert models.
4. New multivariate empirical mode decomposition (MEMD) based models to address non-stationarity within multiple predictor inputs in MEMD transformation were developed. Feature optimization was achieved using the simulated annealing algorithm leading to hybridized MEMD-SA-KRR and MEMD-SA-RF models.
5. The genetic programming (GP) was hybridized with several MCMC based copula models but GP-MCMC-Clayton copula was the most responsive hybrid model to predict cotton yield.

6. Three hybrid models, ACO-OSELM, ACO-ELM and ACO-RF for wheat yield prediction. ACO algorithm was utilized for selecting best training sites in relation to testing site. PACF was utilized for determination of significant lags.

For model evaluations, a diverse range of statistical metrics were used including the Pearson's correlation coefficient (r), mean squared error (MSE), root-mean-square-error (RMSE), mean absolute error (MAE),Willmott's Index (WI), Nash–Sutcliffe Efficiency (ENS), the Legates-McCabe's index (LM), Likelihood value (Max_L), Akaike Information Criterion (AIC), The Bayesian Information Criterion (BIC) and confidence of interval (CI). In addition to the use of numerical assessment metrics, diagnostic plots including box plots, scatter diagram, histogram, time series plot, polar plot and Taylor plots are also utilized for a robust evaluation. Relative measures (i.e., relative root-mean-square-error (RRMSE), and mean absolute error (MAE)) are also used for model comparisons at geographically distinct sites.

The mathematical realizations of each model, feature optimization techniques, multi-resolution analysis tools and the model evaluation metric, as well as the specific model development procedures, are described in the respective chapters.

Chapter 3

Multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm for rainfall forecasting

Foreword

This chapter is an exact copy of the published article in *Atmospheric Research* journal (Vol. 213, Pages 450-464).

It describes the hybridization of the probabilistic and machine learning models. The Markov Chain Monte Carlo based copula (MCMC-cop) models integrate with bio-inspired feature selection based Bat algorithm and online extreme learning machine (OS-ELM) model for rainfall forecasting. Extreme learning machine (ELM) and random forest (RF) models were also hybridized in this study. The MCMC-Cop modeling approach establishes the probabilistic forecast of rainfall using 25 different types of copulas. The Bat algorithm then select the best possible forecast of MCMC-Cop models that are later utilize in OS-ELM model to forecast the final rainfall for the selected regions of study.

The newly designed hybrid models MCMC-Cop-Bat-OS-ELM is compared against MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF for monthly rainfall forecasting at the three rich agricultural zones in Punjab province, Pakistan. The performance of MCMC-Cop-Bat-OS-ELM model was better than the other comparative models.

Atmospheric Research 213 (2018) 450–464

Contents lists available at ScienceDirect

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmosres

Multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm for rainfall forecasting

Mumtaz Ali*, Ravinesh C. Deo*, Nathan J. Downs, Tek Maraseni

School of Agricultural, Computational and Environmental Sciences, International Centre for Applied Climate Sciences, Institute of Agriculture and Environment, University of Southern Queensland Springfield, QLD 4300, Australia

ARTICLE INFO

Keywords:
Rainfall forecasting
Markov chain Monte Carlo simulation
Copulas
Bat algorithm
OS-ELM
ELM
And RF

ABSTRACT

To ameliorate agricultural impacts due to persistent drought-risks by promoting sustainable utilization and pre-planning of water resources, accurate rainfall forecasting models, addressing the dynamic nature of drought phenomenon, is crucial. In this paper, a multi-stage probabilistic machine learning model is designed and evaluated for forecasting monthly rainfall. The multi-stage hybrid MCMC-Cop-Bat-OS-ELM model utilizing online-sequential extreme learning machines integrated with Markov Chain Monte Carlo (MCMC) based bivariate-copula and the Bat algorithm is employed to incorporate significant antecedent rainfall (t–1) as the model's predictor in the training phase. After computing the partial autocorrelation function (PACF) at the first stage, twenty-five MCMC based copulas (i.e., Gaussian, t, Clayton, Gumble, Frank and Fischer-Hinzmann etc.) are adopted to determine the dependence of antecedent month's rainfall with the current and future rainfall at the second stage of the model design. Bat algorithm is applied to sort the optimal MCMC-copula model by a feature selection strategy at the third stage. At the fourth stage, PACF's of the optimal MCMC-copula model are computed to couple the output with OS-ELM algorithm to forecast future rainfall values in an independent test dataset. As a benchmarking process, standalone extreme learning machine (ELM) and random forest (RF) is also integrated with MCMC based copulas and the Bat algorithm, yielding a hybrid MCMC-Cop-Bat-ELM and a MCMC-Cop-Bat-RF models. The proposed multi-stage hybrid model is tested in agricultural belt region in Faisalabad, Jhelum and Multan, located in Pakistan. The testing performance of all three hybridized models, according to robust statistical error metrics, is satisfactory in comparison to the standalone counterparts, however the multi-stage, hybridized MCMC-Cop-Bat-OS-ELM model is found to be a superior tool for forecasting monthly rainfall. This multi-stage probabilistic learning model can be explored as a pertinent decision-support tool for agricultural water resources management in arid and semi-arid regions where a statistically significant relationship with antecedent rainfall exists.

1. Introduction

Anthropogenic and naturally-induced anomalies in regional-scale rainfall can directly affect the agricultural sector since rainfall plays a vital role in both the growth and the production of crops (Maraseni et al., 2012; Nguyen-Huy et al., 2018). The effect is not only restricted to the agricultural sector but it also brings major water-related disasters (Barredo, 2007) such as the shortage of rainfall on the long run is leading to drought events (Palmer, 1965). This can lead to water scarcity (Langridge et al., 2006; Vörösmarty et al., 2010) while excessive amounts of rainfall can cause flooding and damage to human and wildlife health, infrastructure and the economy (Bhalme and Mooley, 1980). The economy of Pakistan, a nation that is still in its developing phase, has also been severely damaged due to major flooding events, including the damage to infrastructure and agricultural crops (News, 2010). The estimated damage in the 2010 event to infrastructure was approximately 4 billion US dollars whereas the damage in the agricultural sector amounted to about 500 million US dollars (Hicks and Burton, 2010). The total economic damage was considerably large, totaling to approximately 43 billion US dollars in 2010 (Mansoor, 2010; Tarakzai, 2010). Equally, drought events (Ali et al., 2018) have been a major contributing factor towards reduced agricultural yields and significant reductions of the gross domestic product of Pakistan. Further, prolonged decline of adequate rainfall can

* Corresponding authors.
E-mail addresses: Mumtaz.Alli@usq.edu.au (M. Ali), ravinesh.deo@usq.edu.au (R.C. Deo), nathan.downs@usq.edu.au (N.J. Downs), tek.maraseni@usq.edu.au (T. Maraseni).

https://doi.org/10.1016/j.atmosres.2018.07.005
Received 3 April 2018; Received in revised form 5 June 2018; Accepted 3 July 2018
Available online 05 July 2018

cause a fall in hydraulic heads having severe consequences for crop irrigation from wells due to changes in the properties of groundwater reservoirs (Santos et al., 2014). Therefore, the ability to forecast rainfall in an accurate manner, particularly in agricultural belt regions, can increase the ability of stakeholders to formulate better water planning and resource management decisions.

Data-intelligent models, particularly developed for local (e.g., farm) scales, have the ability to utilize past data, and hence may offer a viable and reasonably accurate solution to drought disaster management through a projection of future rainfall (Luk et al., 2001). The study of Chiew et al. (1998) developed data-intelligent, predictive models for rainfall forecasting using an empirical method, whereas Sharma (2000) developed a nonparametric probabilistic model to forecast seasonal to inter-annual rainfall in Australia. Burlando et al. (1993) used an autoregressive moving average (ARMA) model for short-term rainfall forecasting in the USA whereas Hung et al. (2009) applied an artificial neural network (ANN) model for rainfall forecasting in Thailand and Lin et al. (2009) forecasted hourly rainfall using support vector machines for Taiwan. Yaseen et al. (2017) developed a rainfall forecasting model using the novel hybrid intelligent model based adaptive neuro fuzzy inference system(ANFIS) integrated with Firefly algorithm (FFA) for Pahang river catchment located in the Malaysian Peninsula, Mason (1998) forecasted seasonal rainfall of South Africa using a nonlinear discriminant analysis model while Nguyen-Huy et al. (2017) developed a novel copula-statistical rainfall forecasting model in Australia's agro-ecological zones. Accurate rainfall forecasting is a significant challenge for Pakistan due to high variation in seasonal, annual and inter-annual rainfalls, exacerbated by climate change.

Despite the need, only a few studies on rainfall forecasting, particularly at local or regional scales, have been carried out in Pakistan. For example, the study of Salma et al. (2012) forecasted rainfall trends in different climatic zones of Pakistan utilizing the autoregressive integrated moving average (ARIMA) model. Archer and Fowler (2008) applied meteorological data to forecast seasonal runoff on the River Jhelum, Pakistan on the basis of multiple linear regression models. Reale et al. (2012) forecasted an extreme rainfall event (in the Indus River Valley, Pakistan, 2010) with a global data assimilation and forecasting model. Faisal and Gaffar (2012) utilized the Thiessen polygon method of weighted rainfall forecast in Pakistan, whereas the study of Ahasan and Khan (2013) simulated flood producing rainfall events in 2010 over north-west Pakistan using weather research and a forecasting model. These studies have provided immensely useful information to various stakeholder, revealing the capability of data-driven models to generate acceptably accurate rainfall forecasts where only historical datasets were applied to construct the forecast model.

The aforementioned studies (Ahasan and Khan, 2013; Archer and Fowler, 2008; Faisal and Gaffar, 2012; Reale et al., 2012; Salma et al., 2012) focused in Pakistan indicate that rainfall forecasting has been mostly based on statistically-based models. In addition to this, a majority of these studies have been conducted to forecast seasonal rainfall using several different datasets. Moreover, there is a limitation of applying advanced data-intelligent models (considering significantly nonlinear behavior of rainfall and its predicators) for accurate forecasting at a micro (or landscape) scale, which can provide help in decision-making for a better management of water resources and flood modelling in the future aimed to reducing the overall risk. For example accurate forecasting is beneficial at catchment scale for agro-forestry applications (Terêncio et al., 2018; Terêncio et al., 2017). Accurate rainfall forecasting can have several economic benefits, for example, a realistic forecast of heavy rainfall could allow airline dispatches to rout their flights in a timely manner (Graham, 2002). In addition to this, a more accurate rainfall forecasting tool might enable appropriate decision about flooding, crop sowing and harvesting and managing of water resources (Graham, 2002; Jones et al., 2000; Toth et al., 2000). To address these issues, there is an apparent need for data intelligent models to forecast rainfall more accurately than the currently statistically-based (i.e., regression) approaches that have various data distribution or linearity assumptions.

In this study, for the first time, a multi-stage online sequential extreme learning machine (OS-ELM) model integrated with Markov Chain Monte Carlo (MCMC) based copulas and the Bat algorithm is developed, denoted as the “MCMC-Cop-Bat-OS-ELM model”. For the purpose of comparison, the standalone extreme learning machine (ELM) without any hybridization and the random forest (RF) models are also developed. The proposed multi-stage MCMC-Cop-Bat-OS-ELM model is tested for rainfall forecasting in three agricultural districts: Faisalabad, Multan, and Jhelum located in Pakistan. The novelty of this study is therefore, to design and apply the newly proposed multi-stage, hybrid MCMC-Cop-Bat-OS-ELM model for rainfall forecasting in Pakistan, a developing nation where accurate predictions are likely to promulgate significant benefits to agriculture, climate adaptation and decision-making in the water resources sector.

To test the applicability of the proposed multi-stage MCMC-Cop-Bat-OS-ELM model, this study fulfils four objectives: (1) To develop a probabilistic MCMC based copula model integrated with the Bat algorithm in order to determine the optimal MCMC-copula model; (2) To incorporate the selected optimal MCMC-copula model based on the Bat algorithm in the OS-ELM model to develop a multi-stage MCMC-Cop-Bat-OS-ELM hybrid prediction tool; (3) To incorporate the significant antecedent lagged rainfall to effectively forecast the current and future rainfall in the consequent month; and (4) To validate the forecasting ability of the proposed hybrid MCMC-Cop-Bat-OS-ELM model for rainfall forecasting in Pakistan.

The literature on accurate rainfall forecasting shows that several approaches were adopted using data intelligent models.

2. Previous work

Accurate rainfall forecasting provides a key role in agriculture, water resources and early flooding warning systems (Yaseen et al., 2018, 2017, 2016). Ortiz-García et al. (2014) used support vector classifiers to forecast rainfall in Spain using meteorological variables and observational data to forecast rainfall in Spain using support vector classifiers in comparison with multi-layer perceptron, extreme learning machine, decision trees and K-nearest neighbor model. Kashiwao et al. (2017) developed a neural network-based local rainfall prediction system in Japan using different climatological data but their findings were not significant as Japan Meteorological Agency. Their findings were based on different precipitation levels (volume of rainfall) at different times (hours). Sánchez-Monedero et al. (2014) modelled rainfall utilizing hierarchical nominal–ordinal support vector classifier in Spain. Their work was based to forecast the occurrence and amount of rainfall using meteorological variables. Abbot and Marohasy (2014) applied artificial neural network to forecast monthly rainfall in Australia utilizing climate indices and they found Inter-decadal Pacific Oscillation has significant influence on rainfall.

Literature on rainfall forecasting using different data intelligent models around the globe is widely available. For example; Villafuerte et al. (2014) analyzed long-term trends and variability of rainfall extremes in the Philippines where the results were showing that inter-annual variations in extreme precipitation indices are influenced greatly by the El Niño–Southern Oscillation Moazami et al. (2014) designed Gaussian copula methods to estimate rainfall in Iran using satellite data where the results were slightly improved by 35.42% (RMSE) and 36.66% (r). Terêncio et al. (2017) developed a model based on flexible weights by incorporating Multi Criteria Analysis and Geographic Information Systems for allocating rainwater in Portugal on the basis of physical, socio-economic and ecologic variables. In another study, Terêncio et al. (2018) conducted a study on the balance of rainwater between sustainability and storage capacity using dam wall height as evaluation parameter. Several other rainfall studies have been developed based on statistical or machine learning models (Bellu et al.,

2016; Dai et al., 2014; Darji et al., 2015; Hardwinarto and Aipassa, 2015; He et al., 2013; Khedun et al., 2014). In this work, a multi-stage probabilistic machine learning model (MCMC-Cop-Bat-OS-ELM) is designed which utilizes MCMC based copula in combination with a Bat algorithm and OS-ELM model which is tested in different geographic locations in Pakistan.

3. Theoretical framework

3.1. Online sequential extreme learning machine (OS-ELM)

ELM is a state-of-the-art data intelligent model developed by Huang et al. (2006) used for the purpose of designing a Single Layer Feed-forward Neural Network (SLFN). ELM is relatively faster, and thus computationally efficient compared with other traditional learning algorithms (Rajesh and Prakash, 2011; Deo and Şahin, 2015; Deo et al., 2017). The SLFN with M hidden nodes of N arbitrary inputs $(x_k, y_k) \in \Gamma^n \times \Gamma^n$ with an activation function $f(.)$ can be mathematically formulated as:

$$\sum_{i=1}^{M} \rho_i f(x_k,; c_i, w_i) = y_k \tag{1}$$

Where $k = 1, 2, \cdots, N$, $c_i \in \Gamma$is the bias of *ith* node which is assigned randomly whereas $w_i \in \Gamma$ is a random input weight vector. The function $g(x_k,; c_{i,\ w_i})$ denotes the output corresponding to the *ith* hidden node with respect to inputx_k. Therefore Eq. (1) reduces to the following form:

$$\mathrm{H}\beta = Y \tag{2}$$

$$\text{WhereH} = \begin{bmatrix} f(x_1,; c_1,, w_1) & \cdots & f(x_1,; c_M,, w_M) \\ \vdots & \cdots & \vdots \\ f(x_N,; c_1,, w_1) & \cdots & f(x_N,; c_M,, w_M) \end{bmatrix}_{N\times M},$$

$\beta = ({\beta_1}^T {\beta_2}^T, \cdots, {\beta_L}^T)^T{}_{m\times M}$ and $Y = ({t_1}^T {t_2}^T, \cdots, {t_L}^T)^T{}_{m\times M}$. The least square solution of the linear systems provides the following output weight:

$$\beta = \mathrm{H}^{+} Y \tag{3}$$

Where H^{+}represents the Moore–Penrose generalized inverse ofH. The SLFNs with random input weight selection effectively acquire distinct training examples with minimum chance of error (Huang, 2003; Tamura and Tateishi, 1997).

The traditional ELM uses all N-samples of data for training purposes but in real applications, this data may use chunk-by-chunk (or one-by-one) because the learning process is time consuming in ELM which requires each time new training data.

As a variant of the classical ELM model, the online sequential extreme learning algorithm (OS-ELM) operates in two learning phases) i.e., initialization and a sequential learning phase. In OS-ELM, the matrix H_0 in the initialization phase is filled which is later used in the learning phase. In the initialization phase, the random weights and biases are assigned to the small chunk of initial training data to compute the hidden layer output matrix. The sequential learning phase is then commenced either on a one-by-one or chunk-by-chunk basis and the used data is not allowed to be used again. For more details on OS-ELM, the readers are referred to a number of previous studies (e.g., (Lan et al., 2009; Liang et al., 2006; Yadav et al., 2016)).

3.2. Markov chain Monte Carlo (MCMC) based statistical copula model

In this study we hybridize the OS-ELM algorithm with a copulas, a set of powerful mathematical tools that have the ability to connect two or more time-independent variables (Nelsen, 2003). A copula function is basically a mathematical function that is defined from $I^2(F, G)$ to $I(H)$ such that $[F(x), G(y), H(x,y)]$ is a point in $I^3$with $I \in [0,1]$ and X, $Y$are continuous random variables with distribution functions $F(x) = P(X \le x)$ and $G(y) = P(Y \le y)$, and $H(x,y) = P(X \le x, Y \le y)$ is a function that describes their joint distribution.

In this paper, we utilize the 25 different types of copulas to improve the performance of the OS-ELM model. The primary copulas can be written mathematically as follows:

I. the Gaussian copula (Li et al., 2013), expressed as:

$$\int_{-\infty}^{\phi^{-1}(a)} \int_{-\infty}^{\phi^{-1}(b)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left(\frac{2\theta xy - x^2 - y^2}{2(1-\theta^2)}\right) dxdy,\ \theta \in [-1,1] \tag{4}$$

II. a t-copula (Li et al., 2013), formulated as:

$$\int_{-\infty}^{t_{\theta_2}^{-1}(a)} \int_{-\infty}^{t_{\theta_2}^{-1}(b)} \frac{\Gamma\left(\left(\theta_2+2\right)/2\right)}{\Gamma\left(\theta_2/2\right)\pi\theta_2\sqrt{1-{\theta_1}^2}} \left(1 + \frac{x^2 - 2\theta_1 xy + y^2}{\theta_2}\right)^{\left(\theta_2+2\right)/2} dxdy,\ \theta_1 \in [-1,1],\ \theta_2 \in (0,\infty) \tag{5}$$

III. a Clayton copula (Clayton, 1978), written as:

$$\max(a^{-\theta} + b^{-\theta} - 1, 0)^{-1/\theta},\ \theta_2 \in [-1,\infty)\backslash 0 \tag{6}$$

IV. a Frank copula (Li et al., 2013), which may be defined according to the following mathematical formulation:

$$-\frac{1}{\theta}\ln\left(1 + \frac{(\exp(-\theta a) - 1)(\exp(-\theta b) - 1)}{\exp(-\theta) - 1}\right),\ \theta \in \mathbb{R}\backslash 0 \tag{7}$$

V. a Gumble copula (Li et al., 2013), expressed as:

$$\exp\left(-(((-\ln(a))^{\theta}) + (-\ln(b))^{\theta})^{\frac{1}{\theta}}\right),\ \theta \in [1,\infty) \tag{8}$$

VI. a Fischer-Hinzmann copula (Fischer and Hinzmann, 2006), given as:

$$[\theta_1(\min(a,b))^{\theta_2} + (1-\theta_1)(ab)^{\theta_2}]^{\frac{1}{\theta_2}},\ \theta_1 \in [0,1],\ \theta_2 \in \mathbb{R} \tag{9}$$

The remaining 19 different types of copulas used in this paper have been discussed in previous studies (Mojtaba Sadegh and AghaKou, 2017). In all types of copula-based models an unknown process κ links observation $\widetilde{Y}$to parameters θ^*in the modelling inference analysis (Mojtaba Sadegh and AghaKou, 2017) and can be given through the following equation.

$$\widetilde{Y} = \kappa(\theta^*) + \xi \tag{10}$$

Where ξindicates a vector of measurement errors. The vector $e = \widetilde{Y} - Y$is called the error residual and $e = \{e_1, e_2, \cdots, e_n\}$where n is the number observations that include the effects of model structural errors (Mojtaba Sadegh and AghaKou, 2017). Bayesian analysis is going to be carried for model inference and uncertainty quantification purposes because Bayesian analysis quantifies uncertainty with a probability distribution (Mojtaba Sadegh and AghaKou, 2017).

Bayes' law attributes all modelling uncertainties to the parameters and estimates the posterior distribution of model parameters by the following equations (Mojtaba Sadegh and AghaKou, 2017):

$$p_{(\theta \mid \widetilde{Y})} = \frac{p(\theta)p_{(\widetilde{Y} \mid \theta)}}{p(\widetilde{Y})} \tag{11}$$

Where $p(\theta)$and $p_{(\theta \mid \widetilde{Y})}$ defines prior and posterior distribution of parameters, respectively. Further, $p_{(\widetilde{Y} \mid \theta)} \simeq L_{(\theta \mid \widetilde{Y})}$ denotes the likelihood given as,

$$L_{(\theta \mid \widetilde{Y})} = \frac{n}{2}\ln\frac{\sum_{i=1}^{n}[\widetilde{y_i} - y_i(\theta)]^2}{n} \tag{12}$$

To solve Eq. (12) analytically and numerically, a Markov Chain Monte Carlo (MCMC) simulation technique will be adopted to sample from the posterior distribution. For more details, readers are referred to

(Mojtaba Sadegh and AghaKou, 2017).

3.3. Bat algorithm

The bio-inspired Bat algorithm is used to rank the best MCMC based copula models in this paper where the performance of the OS-ELM algorithm can be improved with a copula-driven approach. The Bat algorithm originally introduced by Yang (2010) is mainly based on the echolocation behavior of micro bats. In this ranking algorithm, each bat is encoded with a velocity, frequency and location at an integration in a space (solution space) (Fister et al., 2014). The location is then considered the desired solution of the problem in the Bat algorithm. Mathematically, the modified location $x^t{}_j$ with velocity $\nu^t{}_j$ can be expressed as:

$$f_j = f_{\min} + (f_{\max} - f_{\min})\beta \tag{13}$$

$$\nu^t{}_j = \nu^{t-1}{}_j + (x^{t-1}{}_j - x_*)f_j \tag{14}$$

Where x_* is the current location and $\beta \in [0,1]$ is a vector from a uniform distribution.

$$x^t{}_j = x^{t-1}{}_j + \nu^t{}_j \tag{15}$$

Further, the loudness and pulse emission rated can be given as:

$$\Phi_j{}^{t+1} = \alpha\Phi_j{}^t \tag{16}$$

$$t_j{}^{t+1} = r_j{}^0(1 - \exp(-\gamma t)) \tag{17}$$

Where $0 < \alpha < 1$ and $\gamma > 0$. At first stage, the initial population with parameters is generated to determine the best solution in the population.

In the next stage of modelling, the bats are then moved to update the rules in the search space to improve the best solution using random walks. Next, the best solution is evaluated by conditional archiving to update the current best solution. Detailed literature on the Bat algorithm can be seen in (Cai et al., 2014; Fister et al., 2014; Fister Jr et al., 2013; Yang, 2011; Yang and He, 2013).

3.4. Random forest model

To assertain it a robust forecast rool, the skill of multi-stage hybridized MCMC-Cop-Bat-OS-ELM model is benchmarked with a random forest (RF) equivalent data-intelligent model. It is noteworthy that the ensemble learning strategies such as bootstraping and bagging generates classifiers and aggregates the results in the form of decision trees (Breiman, 1996; Schapire et al., 1998). Therefore, the random forest (RF) model is basically a baggging approach with an additional layer of randomness in the prediction process (Breiman, 1996). Each node is split using randomly chosen best subsets of predictors that perform very well and are robust against the overfitting (Breiman, 2001).

The strategy of RF can be presented in the following steps.

Step 1: Construct n_{trees} of bootstrapping from the input data where n is the number of trees.

Step 2: Grow an unpruned regression tree by randomly sample m_{try} of the predictors to select best split among the variables.

Step 3: Aggregate the predictions of n_{trees} to predict new data.

Detailed analysis on RF can be seen in (Breiman, 2001; Liaw and Wiener, 2002; Robert et al., 1998; Segal, 2004).

4. Materials and method

4.1. Rainfall data

In this paper, we use the rainfall data obtained from the Pakistan Meteorological Department, Pakistan for the year 1981 to 2015 (PMD, 2016) for the selected regions, Faisalabad, Multan, Jhelum in Punjab, as shown in Fig. 1.

To evaluate the versatility of the multi-stage, hybridized MCMC-Cop-Bat-OS-ELM model for rainfall forecasting in Pakistan's agricultural belt, the study sites were chosen carefully to ensure that they were broadly representative of the diverse climatic conditions. The first site, Faisalabad, is classified as desert with average annual rainfall of 375 mm, and it is situated in the rolling flat plains of northeast Punjab (Table 1) (Servey, 2016). It is especially noted that Faisalabad is a major producer of wheat, rice, cotton, sugarcane, maize, vegetables and fruits. The next study site, Multan is located in the southern part of Punjab province. At this site, the climate is generally arid with relatively hot summers, and an average temperature of 42.3 °C. The third study site, Multan bears some of the most extreme temperatures in the country as well as the coldest winters (Department, 2010; PMD, 2016). It is noteworthy that Cotton, Citrus, and Mango are heavily reliant upon the availability and sustainable management of water received through rainfall and it is the major economic crop in Multan. Jhelum, on the other hand, is located in the Pothohar Plateau of the Punjab province of Pakistan where agriculture in the district Jhelum is strongly dependent on rainfall. Major crops such wheat, pulses, bajra, maize, rice, fruits and vegetables are known commodities that support the livelihood of the Pakistani people.

4.2. Development and validation of the proposed forecasting model

Historical rainfall data were used to develop the proposed mutli-stage MCMC-Cop-Bat-OS-ELM model. The original rainfall data with statistically significant lagged values at $(t-1)$ as the input predictor was employed in the first stage of model development. The proposed MCMC-Cop-Bat-OS-ELM model was developed under the MATLAB environment on a Pentium 4 2.93 GHz dual core Central Processing Unit. The development and validation of the multi-stage MCMC-Cop-Bat-OS-ELM model can be described in the following four stages:

Stage 1: The statistically significant lags were calculated from the original rainfall data using the partial autocorrelation function (PACF) that can be seen in Fig. 2.

Stage 2: After incorporating the significant lag at $(t-1)$ as the input predictor, the MCMC simulation technique adopts the global and local optimization appraoch to determine the best copula parameters. Twenty-five different types of MCMC based bivariate copulas were employed at this stage. The MCMC algorithm begins with a random search to find the optimum copula parameters adapting the Shuffled Complex Evolution algorithm (Duan et al., 1993). The sample selection acquires the highest likelihood value as the initial point for a Markov chain in intelligent prior sampling. The jump direction is then diversified by the adaptive metropolis algorithm (Haario et al., 2001) to enhance the search. The MCMC based copula models are then moved in the next stage for ranking purposes.

Stage 3: The bio-inspired Bat algorithm is applied to rank the MCMC based copula models and determine the best copula models using a feature selection strategy. Some predefined parameters need to be defined at this stage. The number of bats in this research were 100 with 250 iterations where a predefined minmum and maximum frequency (20 and 50 respectively) were used.

Stage 4: Again, the PACF was used to calcuate the statistical significant lag of the selected MCMC based copula models. The signifcant lags of each selected MCMC copula models based on the Bat algortihm are then incporated as input predictors in the OS-ELM model the for final rainfall forecasting. Different acitvation functions were tested to determine the best acitvation function. The optimal radial base function (rbf) was found when the number of hidden neurons was 60 and block size set to 1 in the development of MCMC-Cop-Bat-OS-ELM. For comparison purpose, extreme learning machine (ELM) and random forest (RF) models were evaluated (Fig. 3).

Following (Deo et al., 2017), data were partitioned into 70% and 30% for training and testing purposes. The normalization of data was accomplished by statistical rules to overcome the numerical difficulties

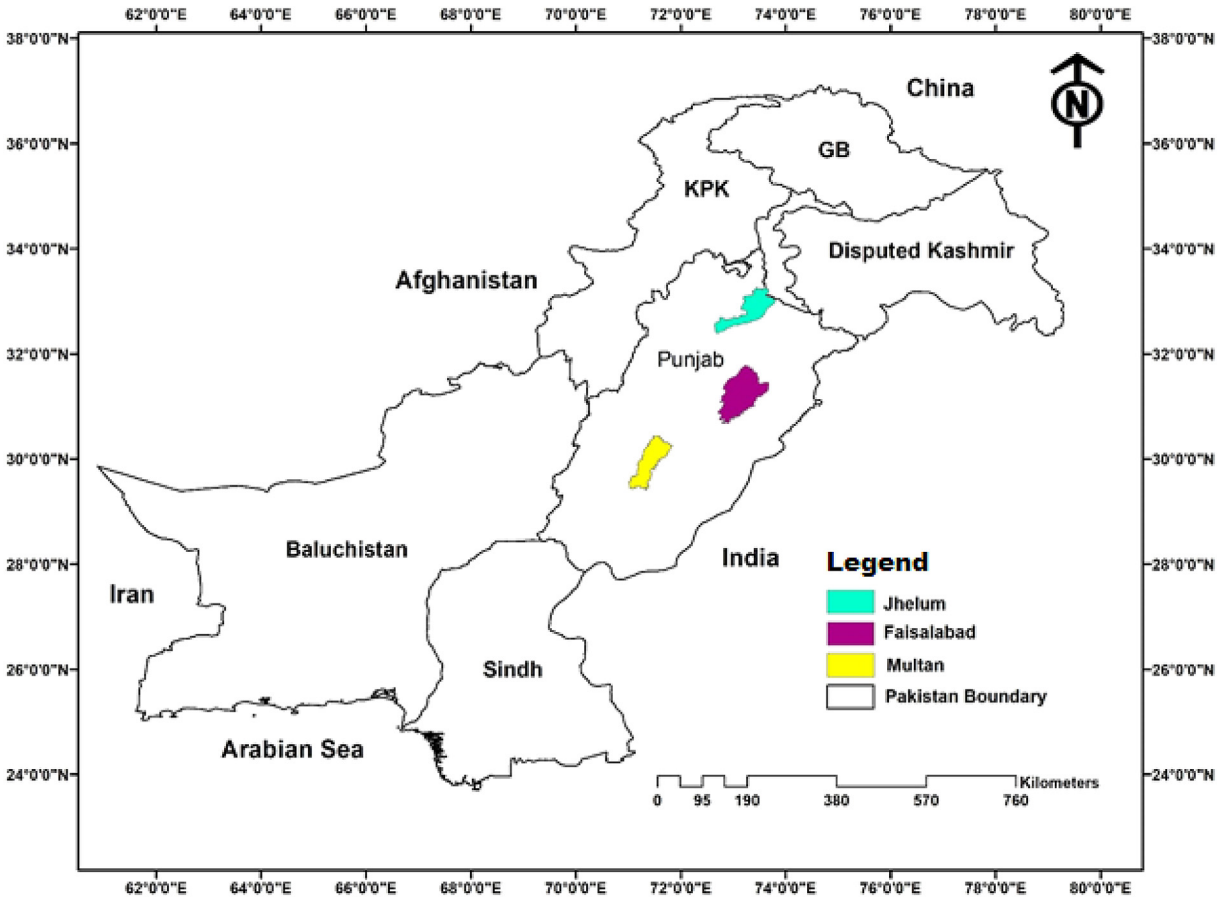


Fig. 1. Map of the agricultural regions in Pakistan showing the sites considered for the development of the multi-stage hybrid MCMC-Cop-Bat-OS-ELM model.

Table 1
Descriptive statistics of the study sites.

Station	Geographic characteristics			Rainfall (mm)					
	Longitude	Latitude	Elevation (m)	Mean	Std.	Min	Max	Skewness	Kurtosis
Faisalabad	73.08^0	31.42^0	184	35.40	49.28	0.10	435.3	3.04	14.01
Jhelum	73.72^0	32.92^0	234	75.66	92.25	0.20	648.6	2.32	7.07
Multan	71.47^o	30.19^0	129	21.70	30.44	0.10	217.3	3.17	13.11

caused by the data features, patterns and fluctuations (Hsu et al., 2003) using:

$$x_{norm} = \frac{(x - x_{\min})}{(x_{\max} - x_{\min})} \tag{18}$$

where x describes any datum point of input or output variable, $x_{\min}$is the minimum value of the whole dataset, $x_{\max}$is the maximum value and x_{norm}denotes the normalized datum point. The correlation coefficient '*r*', in combination with the mean squared error, *MSE* was applied to investigate the performance of the MCMC-Cop-Bat-OS-ELM with other counterpart models in the training period (Table 2).

The magnitudes of *r* and *MSE* attained in training of the hybrid MCMC-Cop-Bat-OS-ELM model for monthly rainfall forecasting at Faisalabad were seen to be: (r = 0.984, *MSE* = 381.40 mm). Equivalent metrics for Jhelum were found to be: (r = 0.992, *MSE* = 568.69 mm) and finally for Multan were: (r = 0.990, *MSE* = 48.13 mm). For comparison, the MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF model were also studied.

The magnitudes of these assessment metrics can be seen in Table 2. Overall, the training performance of the MCMC-Cop-Bat-OS-ELM model was high for all of the study regions. It is thus envisaged that the MCMC-Cop-Bat-OS-ELM model testing performance, as seen later, is relatively accurate for rainfall forecasting at these tested sites.

4.3. Model performance criteria

In this paper, we have applied different types of assessment tools in model evaluation phase. The mathematical formulations of all these assessment metrics are given as follows:

I. The Likelihood value (Max_L) (Thyer et al., 2009), calculated as:

$$\max_L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\tilde{\sigma}^2 - \frac{1}{2}\tilde{\sigma}^{-2}\sum_{i=1}^{n}[R_{obs,i} - R_{for,i}]^2 \tag{19}$$

II. The Akaike Information Criterion (AIC) (Akaike, 1974), given by:

$$AIC = 2D + n.\ln - \left(\frac{\sum_{i=1}^{n}[R_{obs,i} - R_{for,i}]^2}{n}\right) - 2CS \tag{20}$$

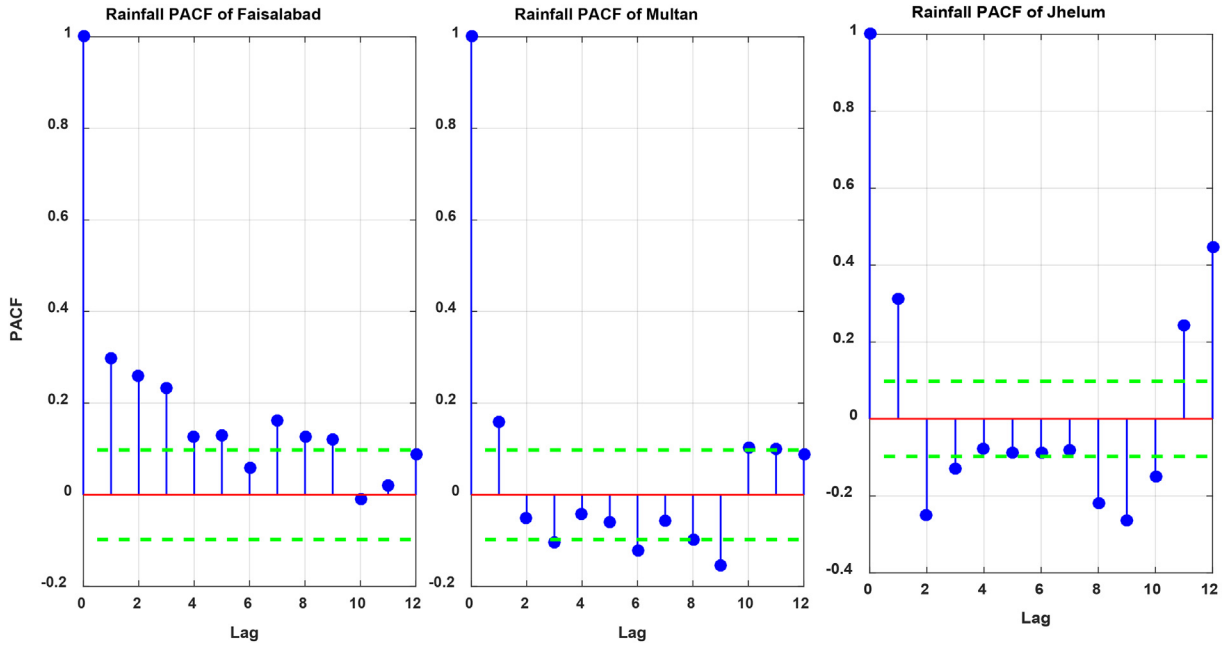


Fig. 2. Partial autocorrelation function (PACF) of historical monthly rainfall time-series in the model's training phase for Faisalabad, Multan, and Jhelum stations used in this study. The green dashed lines denote the statistically significant boundary at the 95% confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

III. The Bayesian Information Criterion (BIC) (Schwarz, 1978), is given by:

$$BIC = D.\ln + n.\ln - \left(\frac{\sum_{i=1}^{n}[R_{obs,i} - R_{for,i}]^2}{n}\right) - 2CS \tag{21}$$

IV. Correlation coefficient (r) (Dawson et al., 2007), expressed as:

$$r = \left(\frac{\sum_{i=1}^{N}(R_{obs,i} - \overline{R}_{obs,i})(R_{for,i} - \overline{R}_{for,i})}{\sqrt{\sum_{i=1}^{N}(R_{obs,i} - \overline{R}_{obs,i})^2}\sqrt{\sum_{i=1}^{N}(R_{for,i} - \overline{R}_{for,i})^2}}\right) \tag{22}$$

V. Willmott's Index (*WI*) Willmott (1981), expressed as:

$$WI = 1 - \left[\frac{\sum_{i=1}^{N}(R_{for,i} - R_{for,i})^2}{\sum_{i=1}^{N}(|R_{for,i} - \overline{R}_{for,i}| + |R_{for,i} - \overline{R}_{for,i}|)^2}\right],\quad 0 \le d \le 1 \tag{23}$$

VI. Nash-Sutcliffe coefficient (NS_E) (1970), expressed as:

$$NS_E = 1 - \left[\frac{\sum_{i=1}^{N}(R_{obs,i} - R_{for,i})^2}{\sum_{i=1}^{N}(R_{for,i} - \overline{R}_{for,i})^2}\right] \tag{24}$$

VII. Root mean square error (*RMSE*) is expressed as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(SPI_{FOR,i} - SPI_{OBS,i})^2 \tag{25}$$

VIII. Root mean square error (*RMSE*) (Dawson et al., 2007), expressed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_{for,i} - R_{obs,i})^2} \tag{26}$$

IX. Mean absolute error (*MAE*) (Dawson et al., 2007), expressed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|(R_{for,i} - R_{obs,i})| \tag{27}$$

X. Legates-McCabe's Index (*LM*) (Legates and McCabe, 1999), expressed as:

$$LM = 1 - \left[\frac{\sum_{i=1}^{N}|R_{for,i} - R_{obs,i}|}{\sum_{i=1}^{N}|R_{obs,i} - \overline{R}_{obs,i}|}\right] \tag{28}$$

XI. Relative root mean square error (*RRMSE*,%) (Legates and McCabe, 1999), expressed as:

$$RRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(R_{for,i} - R_{obs,i})^2}}{\frac{1}{N}\sum_{i=1}^{N}(R_{obs,i})} \times 100 \tag{29}$$

XII. Relative mean absolute percentage error (*RMAE*; %) (Legates and McCabe, 1999), expressed as

$$RMAE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{(R_{for,i} - R_{obs,i})}{R_{obs,i}}\right| \times 100 \tag{30}$$

where $R_{obs,\_i}$ and $R_{for,\_i}$ are the observed and forecasted i^{th} value of rainfall R, $\overline{R}_{obs,i}$ and $\overline{R}_{for,i}$ are the observed and forecasted mean of R, and N is the number of tested data points.

It is important to clarify that $\tilde{\sigma} = \frac{\sum_{i=1}^{n}[\tilde{y}_i - y_i(\theta)]^2}{n}$ and CS is a constant, Eqs. (19)–(21) are used to assess the fitting of copula models whereas Eqs. (22)–(27) show the performance accuracy of the proposed MCMC-Cop-Bat-OS-ELM model while Eqs. (28)–(30) are utilized for

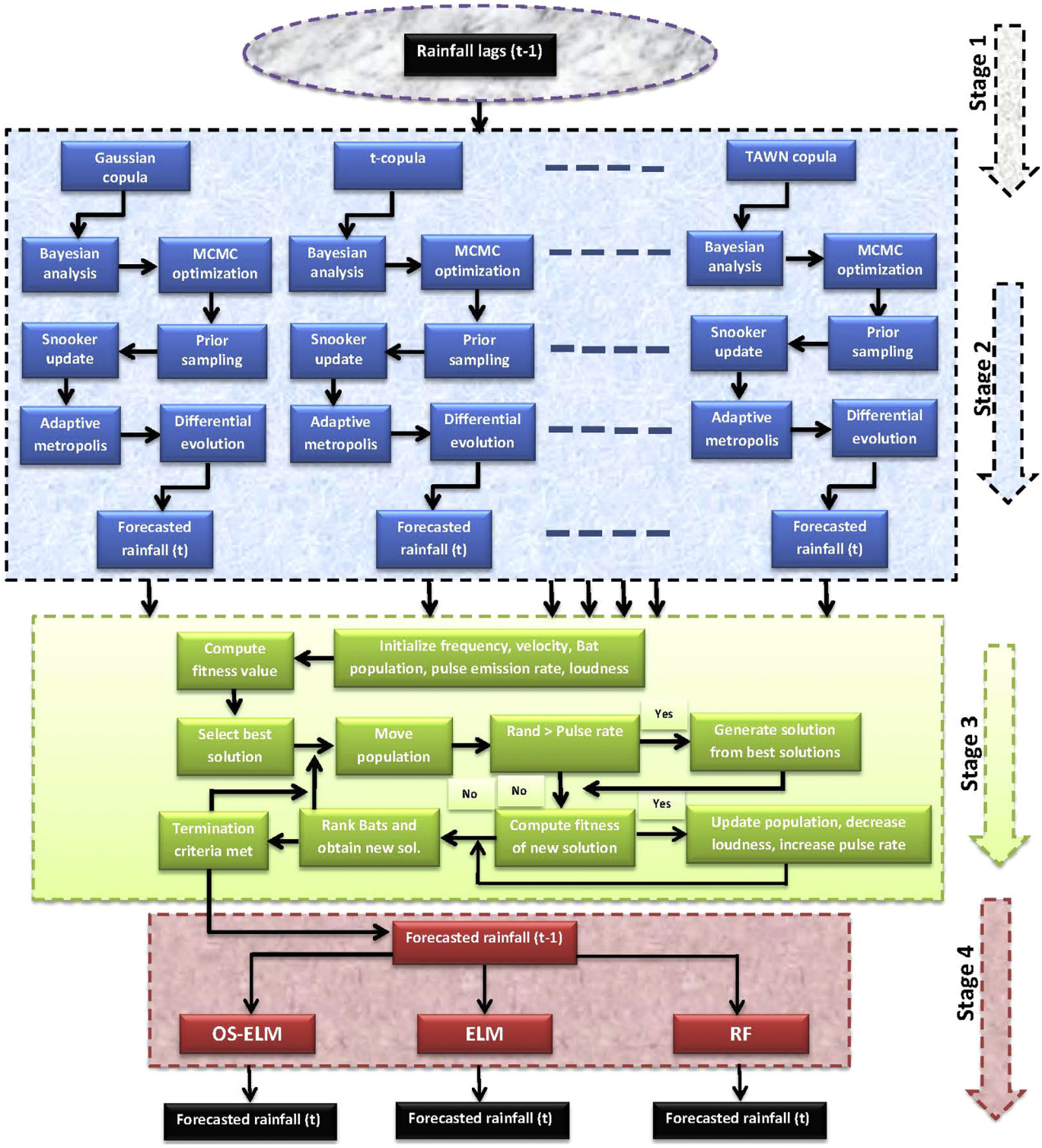


Fig. 3. Flow chart of the proposed multi-stage, hybrid MCMC based copulas integrated with Bat algorithm and OS-ELM model.

geographical comparison of different sites. For further details about these performance metrics, readers are to consult other research works (e.g., (Dawson et al., 2007; Ertekin and Yaldiz, 2000; Legates and McCabe, 1999; Mohammadi et al., 2015; Nash and Sutcliffe, 1970; Willmott, 1981, 1982, 1984; Willmott et al., 2012)).

5. Results

The results of the MCMC-Cop-Bat-OS-ELM with the comparative models, MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF have been evaluated based on the above criterion (Eqs. (19)–(30)). Table 3 shows the selected MCMC-copula model using the Bat algorithm on the basis of feature selection. Out of a total of twenty-five tested MCMC-copula models, seven were selected to be the best by the Bat algorithm for Faisalabad, eight for Jhelum and seven for Multan station. The accuracy error of the Bat algorithm for Multan station is approximately 0.042 between the cost and objective function during the selection of MCMC based copula models. Similarly, this error for Jhelum and Multan

Table 2
The performance evaluation of the proposed hybrid MCMC-Cop-Bat-OS-ELM in respect to the hybrid MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models in the training period based on the mean square error (MSE) and correlation coefficient (r).

Model	Faisalabad		Jhelum		Multan	
	r	*MSE* (mm)	r	*MSE* (mm)	r	*MSE* (mm)
MCMC-Cop-Bat-OS-ELM	0.984	381.40	0.992	568.69	0.990	48.13
MCMC-Cop-Bat-ELM	0.974	635.05	0.972	1956.24	0.984	80.13
MCMC-Cop-Bat-RF	0.938	1741.66	0.917	8424.31	0.903	549.14

station is 0.044 and 0.052 respectively.

Table 4 shows the values of Local ($\alpha_i, i = 1,2,3$) and MCMC ($\beta_i, i = 1,2,3$) parameters with 95% confidence interval (CI) of MCMC based bivariate copula models. The MCMC based MCMC-Gumble copula attained these values ($\alpha_1 \approx 1.48$, $\beta_1 \approx 1.48$, CI([1.46 1.49])) whereas MCMC- Joe copula ($\alpha_1 \approx 1.81$, $\beta_1 \approx 1.81$, CI([1.77 1.85])) and MCMC-Cubic ($\alpha_1 \approx 2.00$, $\beta_1 \approx 2.00$, CI([1.52 1.99])) for Faisalabad (see, Table 4). Similarly, Table 4 shows the values of these metrics for Multan and Jhelum stations. Since the values of CI for all selected MCMC based copula models are positive, the fitting of copulas are correct and results are significant.

Table 5 presents the preciseness of selected MCMC based copula models in terms of *AIC, BIC* and Max_L. The MCMC based F-H copula in Faisalabad station attained the highest values of $AIC \approx -3961.3$, $BIC \approx -3953.2$, $Max_L \approx 1982.7$, followed by the MCMC-Gumble copula where, $AIC \approx -3909.1$, $BIC \approx -3905.1$, $Max_L \approx 1955.6$ and MCMC based on a Cuadras-Auge copula with $AIC \approx -3897.9$, $BIC \approx -3893.9$, $Max_L \approx 1950.0$. For Multan, the best MCMC copula appeared to be BB5 which obtained $AIC \approx -3890.8$, $BIC \approx -3882.7$, $Max_L \approx 1947.4$, followed by MCMC based BB1 copula ($AIC \approx -3887.9$, $BIC \approx -3879.8$, $Max_L \approx 1945.9$) and MCMC-Frank-copula ($AIC \approx -3887.4$, $BIC \approx -3883.4$, $Max_L \approx 1944.7$). For Jhelum station, the MCMC based Roch-Alegre copula attained high performance accuracy ($AIC \approx -4083.7$, $BIC \approx -4075.6$, $Max_L \approx 2043.9$) followed by Joe ($AIC \approx -4007.2$, $BIC \approx -4003.2$, $Max_L \approx 2004.6$) and Fischer-Hinzmann ($AIC \approx -4002.0$, $BIC \approx -3993.9$, $Max_L \approx 2003.0$). Overall, Jhelum attained highest accuracy in terms of *AIC, BIC* and the Max_L criterion followed by Faisalabad and Multan.

Fig. 4(a-c) displays a scatterplot showing the goodness-of-fit. Its correlation coefficient r is also shown to depict the extent of agreement between forecasted and observed rainfall. The MCMC-Cop-Bat-OS- ELM model is clearly better than MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF in terms of r^2 (MCMC-Cop-Bat-OS-ELM ≈ 0.972, MCMC-Cop-Bat-ELM ≈ 0.940, MCMC-Cop-Bat-RF ≈ 0.908) for Faisalabad station. Again, the proposed MCMC-Cop-Bat-OS-ELM model is more accurate for Multan, and Jhelum stations in terms of the achieved r^2 as compared to MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models. Overall, the proposed multi-stage MCMC-Cop-Bat-OS-ELM model convincingly outperforms the comparison models for all the study regions, confirmed by attaining the larger r^2-value.

The percentage of the empirical cumulative distribution function (ECDF) was plotted at each station for different forecasting abilities in Fig. 5. According to this figure, the MCMC-Cop-Bat-OS-ELM method was slightly better than MCMC-Cop-Bat-ELM for all stations, and both models were superior to the MCMC-Cop-Bat-RF model. Based on the percentage of errors in the bracket (0 to ± 300 mm) for the Faisalabad Jhelum and Multan station, Fig. 5 clearly confirms that the MCMC-Cop-Bat-OS-ELM method was the most responsive model in forecasting rainfall data.

Table 6 presents the preciseness of the proposed multi-stage MCMC-Cop-Bat-OS-ELM model in comparison with the MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models, evaluated for all stations in terms of *RMSE, MAE, r,* NS_E and *WI*. The proposed MCMC-Cop-Bat-OS-ELM model in Faisalabad station attained the lowest values of $RMSE \approx 16.59$ mm, $MAE \approx 12.83$ mm while the highest values of $r \approx 0.986$, $NS_E \approx 0.971$ and $WI \approx 0.972$ were followed by MCMC-Cop-Bat-ELM ($RMSE \approx 24.89$ mm, $MAE \approx 18.64$ mm, $r \approx 0.968$, $NS_E \approx 0.9344$ and $WI \approx 0.935$) and the MCMC-COP-Bat-RF ($RMSE \approx 29.65$ mm, $MAE \approx 19.43$ mm, $r \approx 0.953$, $NS_E \approx 0.899$ and $WI \approx 0.900$). For Multan and Jhelum stations, again the MCMC-Cop-Bat-OS-ELM appeared to the best followed by MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models.

The proposed MCMC-Cop-Bat-OS-ELM shows good accuracy in comparison with the other two counterparts. Multan attained the highest accuracy in terms of lowest *RMSE*, and *MAE* values and the highest *r, NSE* and *WI* agreements followed by Faisalabad, and Jhelum.

Fig. 6 compares boxplots of the MCMC-Cop-Bat-OS-ELM model with MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models for each station. The outliers specified by + in every boxplot represent the extreme magnitudes of the forecasting error |FE| within the testing period (months) along with their upper quartile, median and lower quartile values. The distributed |FE| is justified by these boxplots showing a much lesser spread was achieved by the MCMC-Cop-Bat-OS-ELM model for Faisalabad followed by the MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models.

The proposed MCMC-Cop-Bat-OS-ELM model again achieved a good accuracy in terms of |FE| for Jhelum station in relation to the counterpart models. Similarly, the MCMC-Cop-Bat-OS-ELM model performed well for Multan station in rainfall forecasting followed by the MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models. By observing Fig. 6, the accuracy of the MCMC-Cop-Bat-OS-ELM model for all stations appeared to be better than the other models.

Table 7 shows a geographical comparison of the proposed multi-stage MCMC-Cop-Bat-OS-ELM model using relative root mean squared error (*RRMSE*), relative mean absolute error (*RMAE*) and Legates & McCabe's Index (*LM*) for the different locations (Faisalabad, Jhelum, and Multan). Jhelum appeared to be the most accurate station in forecasting rainfall by MCMC-Cop-Bat-OS-ELM ($LM \approx 0.881$, $RRMSE \approx 11.27\%$, $RMAE \approx 40.12\%$), MCMC-Cop-Bat-ELM ($LM \approx 0.769$, $RRMSE \approx 21.56\%$, $RMAE \approx 132.50\%$) and MCMC-Cop-Bat-RF ($LM \approx 0.584$, $RRMSE \approx 40.87\%$, $RMAE \approx 303.60\%$) followed by Multan and Faisalabad respectively. In terms of site-averaged performance, the MCMC-Cop-Bat-OS-ELM model was found to yield the highest Legates-McCabe's agreement and lowest relative percentage errors (*RRMSE, RMAE*).

Fig. 7(a-c) is a Taylor diagram. This diagram provides a more concrete and conclusive argument about the statistical summary of how well the forecasted rainfall matched with the observed rainfall in terms of their correlation. The similarity between forecasted and observed

Table 3
The selected Markov Chain Monte Carlo (MCMC) based copula models ranked by the Bat algorithm. Note that a total of 25 MCMC-copula models were used here.

Station	Selected copula models								Total selected model	*MSE* (mm)
Faisalabad	Gumbel	Joe	FGM	Cuadras-Auge	Cubic	Burr	F-H	TAWN	7	0.042
Multan	Frank	Joe	Burr	M-O	F-H	BB1	BB5		8	0.044
Jhelum	Joe	Cuadras-Auge	Cubic	Burr	F-H	Roch-Alegre	BB1		7	0.052

Table 4
Local ($\alpha_i, i = 1,2,3$) and MCMC ($\beta_i, i = 1,2,3$) estimated copula parameters with the 95% confidence of interval (CI).

Faisalabad									
Copula	α_1	α_2	α_3	β_1	β_2	β_3	95% CI (α_1,β_1)	95% CI, (α_2,β_2)	95% CI, (α_3,β_3)
Gumbel	1.38			1.48			[1.46 1.49]		
Joe	1.81			1.81			[1.77 1.85]		
FGM	1.00			1.00			[0.99 1.00]		
Cuadras-Auge	0.48			0.48			[0.47 0.48]		
Cubic	2.00			2.00			[1.52 1.99]		
Burr	1.11			1.11			[1.06 1.16]		
Fischer-Hinzmann	0.43	0.61		0.43	0.61		[0.41 0.44]	[0.49 0.73]	
Multan									
Frank	1.02			1.12			[1.06 1.18]		
Joe	1.24			1.24			[1.22 1.25]		
Burr	3.54			3.54			[3.35 3.75]		
Marshal-Olkin	12.73	23.93		0.31	0.16		[0.15 0.33]	[0.14 0.30]	
Fischer-Hinzmann	0.26	−1.14		0.26	−1.16		[0.23 0.29]	[−1.94 -0.67]	
BB1	28.52	31.80		0.02	1.14		[0.00 0.04]	[1.11 1.15]	
BB5	1.00	0.39		1.00	0.38		[1.00 1.15]	[0.01 0.38]	
TAWN	1.00	0.92	1.1534	1.00	0.98	1.1464	[0.50 0.99]	[0.45 0.98]	[1.14 1.31]
Jhelum									
Joe	1.42			1.42			[1.40 1.44]		
Cuadras-Auge	0.31			0.31			[0.29 0.32]		
Cubic	0.75			0.75			[0.11 1.37]		
Burr	2.09			2.09			[2.01 2.15]		
Fischer-Hinzmann	0.52	−3.64		0.52	−3.68		[0.49 0.53]	[−4.02–3.25]	
Roch-Alegre	0.35	1.50		0.35	1.49		[0.26 0.41]	[1.46 1.53]	
BB1	0.00	1.23		0.00	1.23		[0.00 0.00]	[1.22 1.24]	

rainfall is quantified in terms of their correlation and standard deviations.

For Faisalabad station, the correlation of the MCMC-Cop-Bat-OS-ELM model with observations was about 0.98, followed by MCMC-Cop-Bat-ELM ≈ 0.96 and MCMC-Cop-Bat-RF ≈ 0.95. The MCMC-Cop-Bat-OS-ELM model was closer to the observed rainfall as its correlation is about 0.991 compared to MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF for Jhelum station. Similarly the MCMC-Cop-Bat-OS-ELM again appeared to be the best model for Multan station because its correlation lies within close neighbourhood of observed rainfall data. Overall, the correlation of the MCMC-Cop-Bat-OS-ELM model was closer to the observed rainfall compared to the other models.

6. Discussion: limitations and opportunity for further research

Accurate rainfall forecasting can complement and facilitate better planning of water management (Terêncio et al., 2018; Ali et al., 2018; Yaseen et al., 2016, 2017, 2018). (Terêncio et al., 2018). Furthermore, accurate predictions of rainfall can reduce water-related natural disasters (Barredo, 2007; Langridge et al., 2006; Palmer, 1965; Vörösmarty et al., 2010), and potential impacts upon wildlife health, infrastructure and the economy (Bhalme and Mooley, 1980). The total economic damage done by rainfall in Pakistan, the case study in this paper, can be considerable, totaling more than 40 billion US dollars in 2010 (Mansoor, 2010; Tarakzai, 2010). Advance data-driven models can offer a viable and reasonably accurate solution to disaster management through a projection of the future trends in rainfall utilizing past datasets (Luk et al., 2001).

This paper has designed a new hybrid multi-stage MCMC-Cop-Bat-OS-ELM model that uses historical significant rainfall lags (t – 1) as a predictor to forecast the current and future rainfall within the agricultural region of Pakistan. The proposed hybrid multi-stage MCMC-Cop-Bat-OS-ELM model is compared with the MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models to evaluate its predictive ability. The designed MCMC-Cop-Bat-OS-ELM model performed reasonably well because of the improved features estimated by the distribution of MCMC copula models (Mojtaba Sadegh and AghaKou, 2017). The MCMC based copula used Bayesian analysis centred on likelihood function estimate

Table 5
Evaluation of the MCMC based copulas models constructed with model data significant at a lag (t – 1) using the Akaike Information Criterion (*AIC*), Bayesian Information Criterion (*BIC*) and Maximum Likelihood (Max_l) criterion for each station.

Faisalabad								
Criteria	Gumbel	Joe	FGM	Cuadras-Auge	Cubic	Burr	F-H	
AIC	−3909.1	−3490.9	−3452.3	−3897.9	−2566.7	−3571.8	−3961.3	
BIC	−3905.1	−3486.9	−3448.3	−3893.9	−2562.6	−3567.8	−3953.2	
Max_L	1955.6	1746.5	1727.2	1950.0	1284.3	1786.9	1982.7	
Multan								
	Frank	Joe	Burr	M-O	F-H	BB1	BB5	TAWN
AIC	−3887.4	−3820.9	−3874.5	−3787.1	−3802.5	−3887.9	−3890.8	−3884.6
BIC	−3883.4	−3816.9	−3870.5	−3779.0	−3794.5	−3879.8	−3882.7	−3872.5
Max_L	1944.7	1911.5	1938.3	1895.6	1903.3	1945.9	1947.4	1945.3
Jhelum								
	Joe	Cuadras-Auge	Cubic	Burr	F-H	Roch-Alegre	BB1	
AIC	−4007.2	−3669.3	−2945.6	−3985.2	−4002.0	−4083.7	−3798.3	
BIC	−4003.2	−3665.3	−2941.6	−3981.2	−3993.9	−4075.6	−3790.2	
Max_L	2004.6	1835.7	1473.8	1993.6	2003.0	2043.9	1901.2	

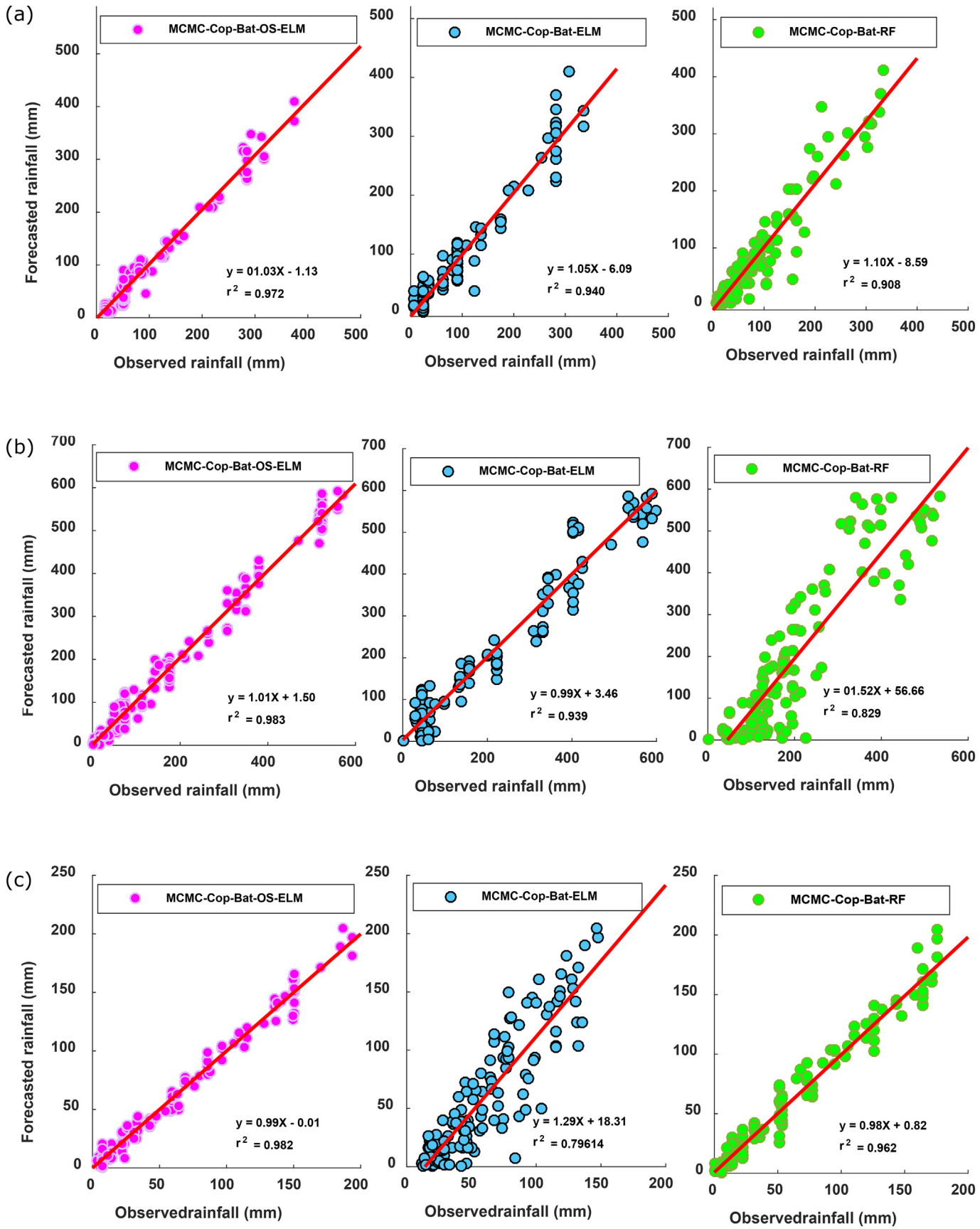


Fig. 4. Scatterplot of the forecasted and observed rainfall (mm) in the testing phase for the proposed hybrid MCMC-Cop-Bat-OS-ELM, MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models using the 1-month (t – 1) significantly lagged data including the coefficient of determination (r^2) and a linear fit inserted in each panel for study zones (a) Faisalabad, (b) Jhelum and (c) Multan.

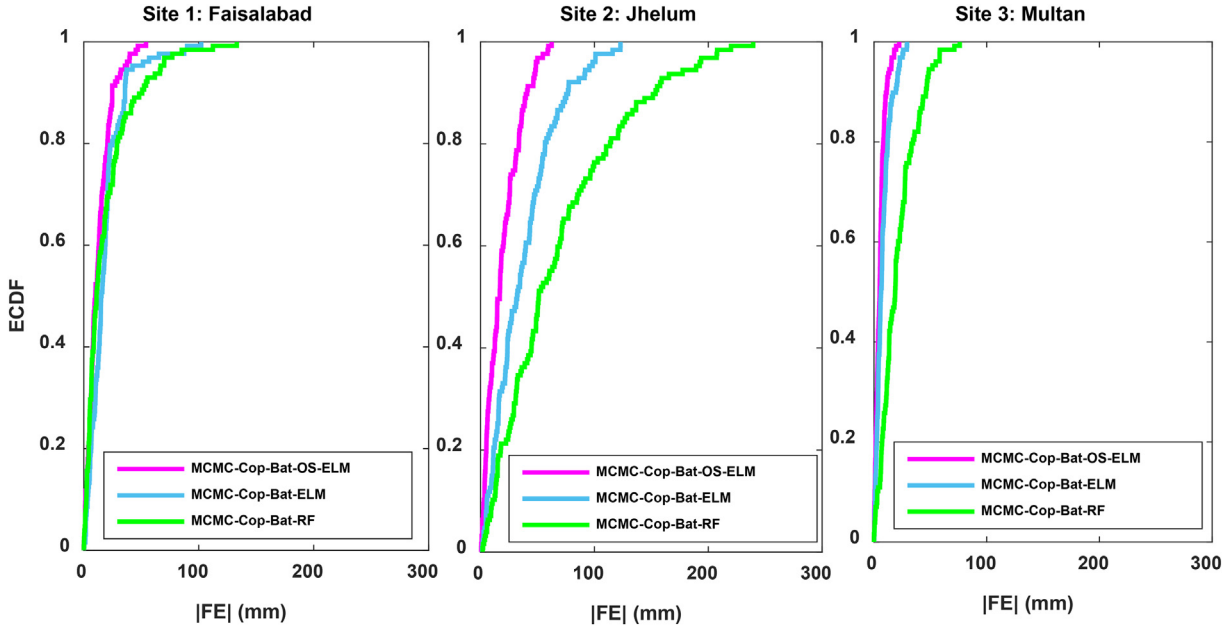


Fig. 5. Empirical cumulative distribution function (ECDF) of the forecasted error |FE| (mm) generated by the proposed hybrid MCMC-Cop-Bat-OS-ELM model versus its counterpart models in the testing period for Faisalabad, Jhelum and Multan stations.

Table 6
Evaluation of hybridized multi-stage MCMC-Cop-Bat-OS-ELM vs. MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models using root mean square error *(RMSE; mm)* and mean absolute error *(MAE; mm)*, correlation coefficient (*r*), Willmott index *(WI)*, and Nash-Sutcliffe coefficient *(NS_E)* in the testing period. The best model is boldfaced (red).

Model	*RMSE* (mm)	*MAE* (mm)	*r*	NS_E	*WI*
Faisalabad					
MCMC-Cop-Bat-OS-ELM	16.59	12.83	0.986	0.971	0.972
MCMC-Cop-Bat-ELM	24.89	18.64	0.968	0.9344	0.935
MCMC-Cop-Bat-RF	29.65	19.43	0.953	0.899	0.900
Multan					
MCMC-Cop-Bat-OS-ELM	6.99	5.46	0.991	0.983	0.988
MCMC-Cop-Bat-ELM	10.28	7.91	0.981	0.963	0.974
MCMC-Cop-Bat-RF	26.83	21.47	0.891	0.752	0.786
Jhelum					
MCMC-Cop-Bat-OS-ELM	24.37	19.16	0.992	0.989	0.989
MCMC-Cop-Bat-ELM	46.63	37.31	0.969	0.938	0.958
MCMC-Cop-Bat-RF	87.50	68.30	0.907	0.787	0.807

copula parameters with underlying uncertainties (Mojtaba Sadegh and AghaKou, 2017).

In this paper, the Bat algorithm ranked the MCMC-copula models to select the optimal bivariate copula. The frequency of the Bat algorithm was tuned continuously which used echolocation to find the optimum MCMC-copula (Yang and He, 2013). Moreover, the automatic zooming in Bat algorithm searched for promising optimal MCMC-copulas at a faster convergence rate (Yang and He, 2013). The ability of parameter control in the Bat algorithm provides optimal MCMC-copula models (Yang and He, 2013). Finally the online- sequential learning strategy (Yadav et al., 2016) adopted in OS-ELM model to further optimize the forecasting accuracy using the selected MCMC-Cop-Bat models.

The proposed MCMC-Cop-Bat-OS-ELM with its counterpart models was successfully evaluated to yield high accuracy based on *RRMSE* and *RMAE* metrics respectively with a reasonably large statistical correlation of Legates-McCabe's ($LM \approx 0.881$) between the forecasted and observed rainfall for Jhelum station and similarly for other stations (Table 7). The proposed MCMC-Cop-Bat-OS-ELM is a suitably optimized model because of the selected optimum MCMC-copulas based on the Bat optimization algorithm as well as tested several activation functions in combination with hidden neurons for a more advance fast learning algorithm, OS-ELM. Therefore, the proposed MCMC-Cop-Bat- OS-ELM model can be used to forecast rainfall in the future where the forecasting of rainfall will likely become even more important due to increasing demand for water resources and agricultural growth in terms of irrigation and where such large datasets are available for model training.

This study can be extended to other locations where rainfall data is available to provide accurate forecasting of future rainfall. Furthermore, the present study can also be extended by an alternative method for rainfall prediction within a probabalistic framework using a vine copula approach and the multi-stage MCMC-Cop-Bat-OS-ELM model can be compared accordingly (Nguyen-Huy et al., 2017). Due to the superior and accurate performance, it is to be noted that the proposed MCMC-Cop-Bat-OS-ELM model may be a successful predictor of other related data sets such as crop yield estimation, droughts events, etc. This may be of great interest to government policy makers and agricultural engineers to avoid the possibility of inaccurate estimation in the future (Akhtar, 2014; Niaz, 2014).

Due to the aforementioned qualities of the proposed multi-stage MCMC-Cop-Bat-OS-ELM model, it is possible to apply it to the forecasting of other climatological parameters such as temperature, humidity, wind speed, solar radiation, and soil moisture. As Pakistan is currently suffering from global warming (Abubakar, 2017), heatwave and drought forecasting may be another possible avenue for future work using these models. Another possible study may involve use of several types of climatological parameters and indices to forecast rainfall. The rainfall data of several different stations may be used for training to develop a future universal rainfall forecasting model for Pakistan.

In terms of model optimization for rainfall forecasting, the hybridization of different models has the potential to generate better estimates than standalone models. Therefore, the proposed models could be optimized with ensemble methods (Ali et al., 2018; Dietterich, 2002; Lei and Wan, 2012; Yun et al., 2008) to achieve more accurate results and confidence intervals, improving on the current level of uncertainty and information (e.g., error bars). The ANFIS algorithms which are very powerful can be another optimization method to be considered in this regard (Jang, 1993; Yaseen et al., 2017). Moreover, the more advanced models such as the Ensemble methods (Dietterich, 2002), Particle

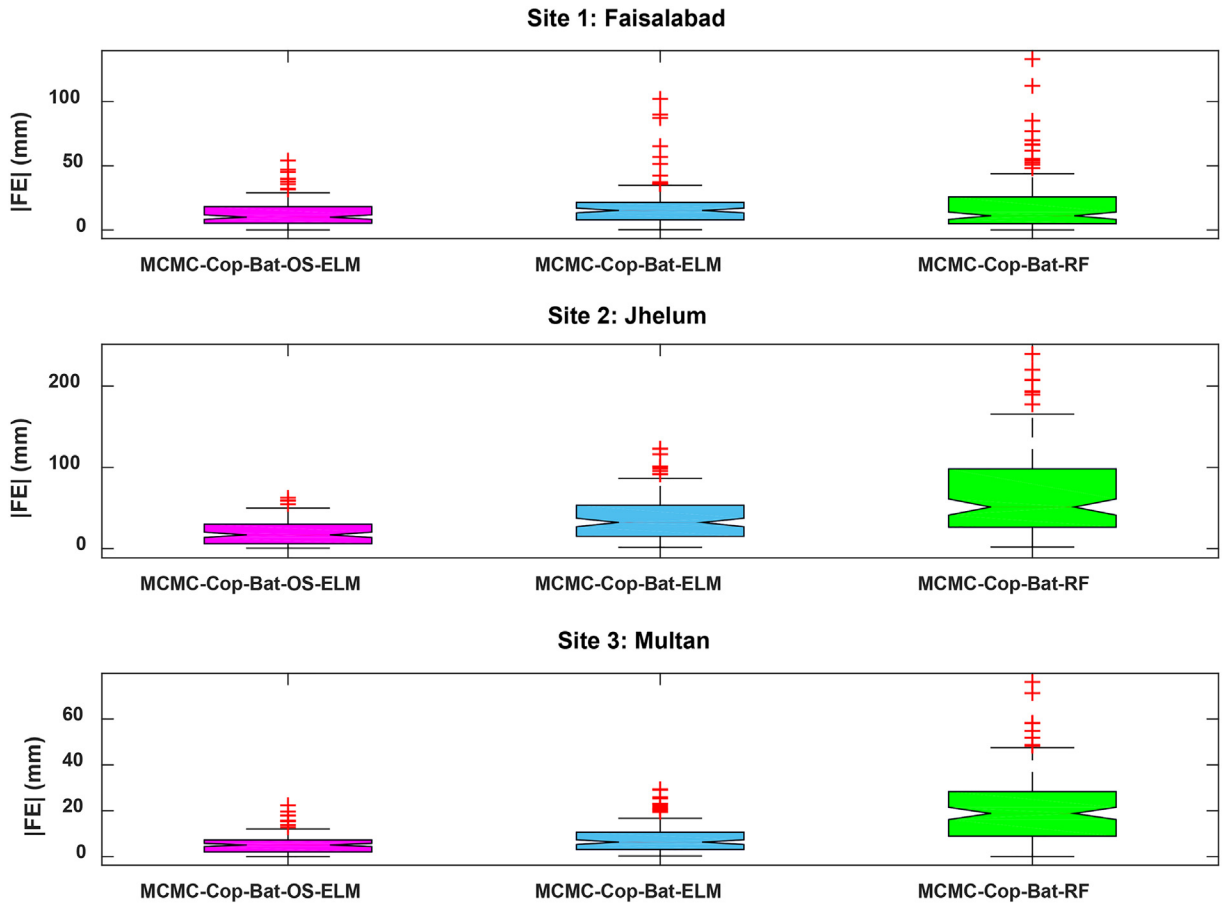


Fig. 6. Box-plots of the forecasted error |FE| (mm) in testing phase for the proposed hybrid MCMC-Cop-Bat-OS-ELM model with counterpart models of monthly forecasted rainfall for (a) Faisalabad, (b) Jhelum and (c) Multan.

Swarm Optimization (Chen and Yu, 2005; Zhisheng, 2010), Genetic algorithms (Davis, 1991), chaos theory (Briggs and Peat, 1989) etc. coupled with traditional and even the newly explored vine copulas (Nelsen, 2003; Nguyen-Huy et al., 2017; Nguyen-Huy et al., 2018) may be able to generate good results due to their optimization ability.

The regression models (Draper and Smith, 2014) may be coupled with copulas in a way to optimize the forecasting ability of possible hybrid copula models. A more generalized framework of hybridizing copulas with generalized mixed models (Draper and Smith, 2014) to develop MCMC-copula mixed models for rainfall is currently being prepared in upcoming work. Autoregressive fractionally integrated moving average (Ling and Li, 1997) based copula (ARFIMA-copula) and least square support vector machine-based copula (LSSVM-copula) models may also be used to forecast rainfall. As the standard statistical approaches avoid the hurdle of model uncertainty that leads to over-confident inferences and more risky decisions, Bayesian model averaging (BMA) techniques (Hoeting et al., 1998; Raftery et al., 2005) have the ability to model uncertainty for accurate predictions. Therefore, BMA techniques provide yet another option to be used to model uncertainty in rainfall that is produced due to several factors such as missing climate data, extreme weather conditions and the likely influence of climate change.

Multi-resolution tools like frequency resolution may also be applied to broaden the scope of this study. In this regard, maximum overlap discrete wavelet transformation (Holschneider, 1988; Khalighi et al., 2011; Prasad et al., 2017), empirical mode decomposition EMD (Al-Musaylh et al., 2018; Rilling et al., 2003), and singular value decomposition (De Lathauwer et al., 1994) have the potential to be utilized for improved predictions. Other options could be the use of feature selection techniques (Guyon and Elisseeff, 2003; Jain and Zongker, 1997; Salcedo-Sanz et al., 2018) where several variables may also be used to screen the optimal predictors to forecast rainfall with a greater accuracy. While the present study has not utilized these ancillary tools, the relatively good accuracy of the OS-ELM model does provide an avenue for further independent studies.

Table 7
Geographic comparison of hybridized multi-stage MCMC-Cop-Bat-OS-ELM vs. MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF model using relative root mean squared error (*RRMSE; %*), relative mean absolute error (*RMAE; %*) and the Legates & McCabe's Index (*LM*). The best model is boldfaced (red).

Station	MCMC-Cop-Bat-OS-ELM			MCMC-Cop-Bat-ELM			MCMC-Cop-Bat-RF		
	LM	*RRMSE*, %	*RMAE*, %	*LM*	*RRMSE*, %	*RMAE*, %	*LM*	*RRMSE*, %	*RMAE*, %
Faisalabad	0.827	17.76	46.26	0.749	26.64	81.05	0.723	31.93	54.97
Multan	0.878	11.36	49.91	0.823	16.70	41.72	0.530	44.06	245.64
Jhelum	0.881	11.27	40.12	0.769	21.56	132.50	0.584	40.87	303.60

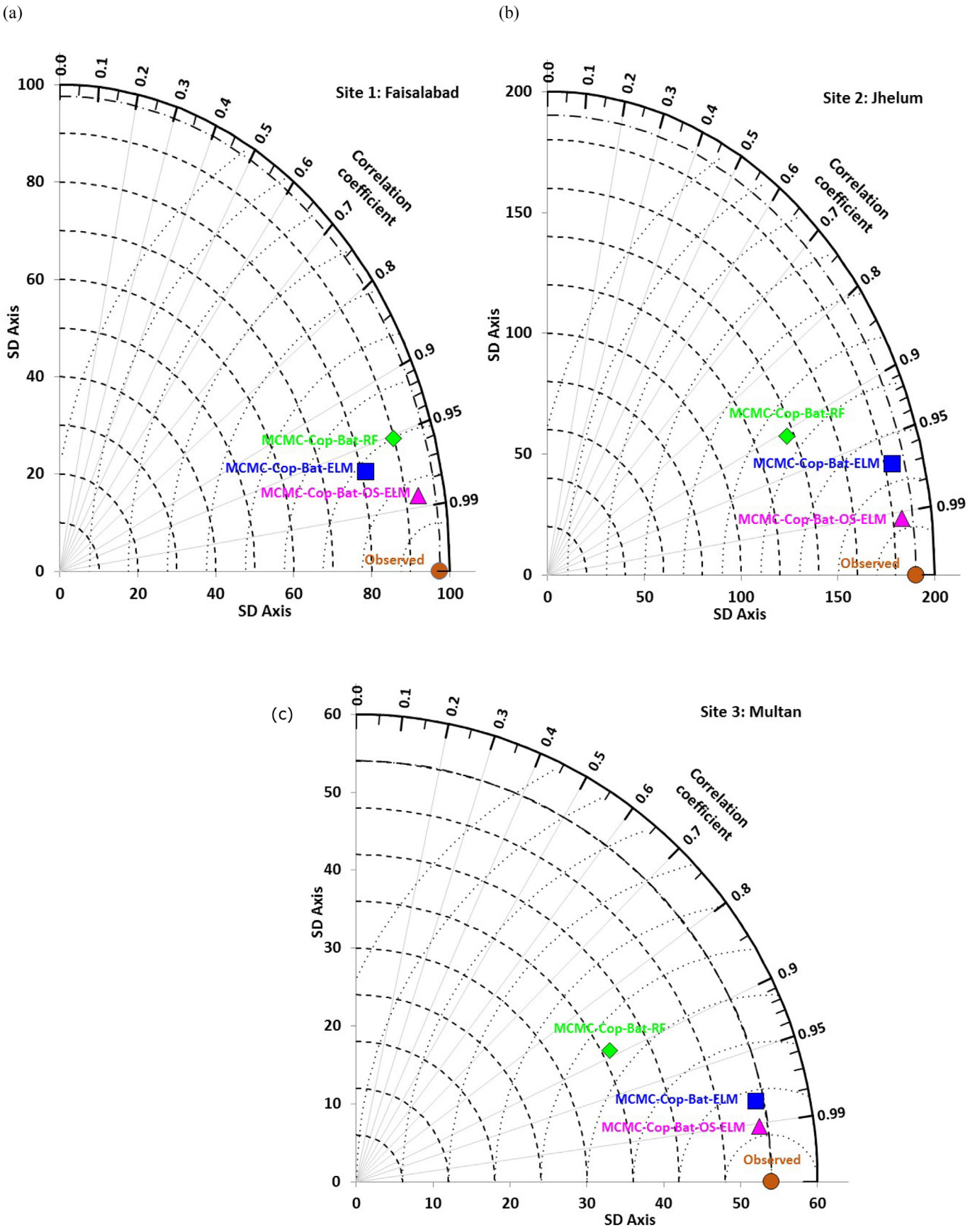


Fig. 7. Taylor diagram showing the correlation coefficient between the observed and forecasted rainfall and standard deviation for the proposed hybrid multi-stage MCMC-Cop-Bat-OS-Elm model in comparison with MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models for the station (a) Faisalabad, (b) Jhelum, and (c) Multan.

7. Conclusion

For the first time, this paper has developed a hybrid multi-stage MCMC- Cop-Bat-OS-ELM model using the significant antecedent lags of monthly rainfall as predictor variables to forecast future rainfall for different geographical sites in Pakistan. The rainfall data from 1981 to 2015 for a total of three stations were used to develop the proposed multi-stage MCMC- Cop-Bat-OS-ELM model in order to achieve a high level of accuracy. Further, several types of evaluation criterion were adopted to assess the performance of the proposed model.

In the multi-stage MCMC- Cop-Bat-OS-ELM model, after incorporating the input rainfall data, the MCMC algorithm adopted global optimization as well as a local optimization technique to find the best copula parameters that helped in developing 25 MCMC-copula models. These MCMC based copula models were then ranked by the bio-inspired Bat algorithm to extract the best MCMC-copula models using a feature selection strategy. Finally, the statistically significant lags of selected MCMC based copula models were utilized in the OS-ELM model as predictors to generate the final forecast. Further, ELM and RF models were used for comparison purposes. Evidently, the performance of all the proposed models was better but the MCMC-Cop-Bat-OS-ELM model was found to be superior among others (see, Tables 6 and 7) as evident by its low relative forecasting errors and high performance metrics. The forecasting errors by the best MCMC-Cop-Bat-OS-ELM model for the Jhelum station were $RMSE \approx 24.37$ mm whereas the MAE was 19.16 mm, respectively. While the performance metrics were ($r \approx 0.992$, $NS_E \approx 0.989$, $WI \approx 0.989$).

By assessing the performance of the multi-stage MCMC-Cop-Bat-OS-ELM model for Jhelum station using the most advanced normalized metrics of Legates-McCabe's, the MCMC-Cop-Bat-OS-ELM model was found to have the highest agreement. The obtained LM agreement values between the forecasted and observed rainfall for Jhelum station were $LM \approx 0.881$ whereas the relative percentage errors $RRMSE$ and $RMAE$ were 11.27%, 40.12%. The proposed multi–stage MCMC-Cop-Bat-OS-ELM model also generated convincingly better results for Faisalabad and Multan stations.

This study provides a baseline relevant with other models being potentially utilized to forecast rainfall and other climatological parameters more accurately in future studies. The proposed multi-stage hybrid MCMC-Cop-Bat-OS-ELM model can be applied to other areas such as agricultural crop yield prediction, drought forecasting and so on. Further to this, agricultural policy makers in Pakistan can use this model for prior rainfall forecasting that will assist in the optimal management of crop sowing, harvesting and irrigation. Moreover, accurate future rainfall forecasting can warn the government and impacted stakeholders prior to significant flooding and the allocation of water resources for future.

Acknowledgement

This research utilized rainfall data acquired from Pakistan Meteorological Department, Pakistan, which are duly acknowledged. This study was supported by University of Southern Queensland USQPRS (2017-2019) Office of Graduate Studies Postgraduate Research Scholarship (2017–2019) awarded to the first author. We thank both reviewers and the Editor-in-Chief for their constructive comments that has improved the clarity of the final paper.

References

Abbot, J., Marohasy, J., 2014. Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. Atmos. Res. 138, 166–178.
Abubakar, S.M., 2017. Pakistan 7th most Vulnerable Country to Climate Change. DAWN.
Ahasan, M., Khan, A., 2013. Simulation of a flood producing rainfall event of 29 July 2010 over north-west Pakistan using WRF-ARW model. Nat. Hazards 69 (1), 351–363.
Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (6), 716–723.
Akhtar, I.U.H., 2014. Pakistan Needs a New Crop Forecasting System.
Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2018. An ensemble-ANFIS based uncertainty assessment model for forecasting multi-scalar standardized precipitation index. Atmos. Res. 207, 155–180.
Al-Musaylh, M.S., Deo, R.C., Adamowski, J.F., Li, Y., 2018. Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting. Appl. Energy 217, 422–439. http://dx.doi.org/10.1016/j.apenergy.2018.02.140.
Archer, D.R., Fowler, H.J., 2008. Using meteorological data to forecast seasonal runoff on the river Jhelum, Pakistan. J. Hydrol. 361 (1), 10–23.
Barredo, J.I., 2007. Major flood disasters in Europe: 1950–2005. Nat. Hazards 42 (1), 125–148.
Bellu, A., Fernandes, L.F.S., Cortes, R.M., Pacheco, F.A., 2016. A framework model for the dimensioning and allocation of a detention basin system: the case of a flood-prone mountainous watershed. J. Hydrol. 533, 567–580.
Bhalme, H.N., Mooley, D.A., 1980. Large-scale droughts/floods and monsoon circulation. Mon. Weather Rev. 108 (8), 1197–1211.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.
Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.
Briggs, J., Peat, F.D., 1989. Turbulent Mirror: An Illustrated Guide to Chaos Theory and the Science of Wholeness. HarperCollins Publishers.
Burlando, P., Rosso, R., Cadavid, L.G., Salas, J.D., 1993. Forecasting of short-term rainfall using ARMA models. J. Hydrol. 144 (1–4), 193–211.
Cai, X., Wang, L., Kang, Q., Wu, Q., 2014. Bat algorithm with Gaussian walk. Int. J. Bio-Inspired Computation. 6 (3), 166–174.
Chen, G.-c., Yu, J.-s., 2005. Particle swarm optimization algorithm. Information and Control-Shenyang 34 (3), 318.
Chiew, F.H., Piechota, T.C., Dracup, J.A., McMahon, T.A., 1998. El Nino/southern oscillation and Australian rainfall, streamflow and drought: links and potential for forecasting. J. Hydrol. 204 (1), 138–149.
Clayton, D.G., 1978. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65 (1), 141–151.
Dai, Q., Han, D., Rico-Ramirez, M.A., Islam, T., 2014. Modelling radar-rainfall estimation uncertainties using elliptical and Archimedean copulas with different marginal distributions. Hydrol. Sci. J. 59 (11), 1992–2008.
Darji, M.P., Dabhi, V.K., Prajapati, H.B., 2015. Rainfall forecasting using neural network: a survey. In: Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in. IEEE, pp. 706–713.
Davis, L., 1991. Handbook of Genetic Algorithms.
Dawson, C.W., Abrahart, R.J., See, L.M., 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ. Model Softw. 22 (7), 1034–1052.
De Lathauwer, L., De Moor, B., Vandewalle, J., by Higher-Order, B.S.S, 1994. Singular value decomposition. Proc. EUSIPCO-94, Edinburgh, Scotland, UK, 175–178.
Deo, R.C., Şahin, M., 2015. Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia. Atmos. Res. 153, 512–525.
Deo, R.C., Wen, X., Qi, F., 2017. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Appl. Energy 168, 568–593.
Department, P.M, 2010. Dry Weather Predicted in the Country during Friday/Monday.
Dietterich, T.G., 2002. Ensemble learning. In: The Handbook of Brain Theory and Neural Networks. Vol. 2. pp. 110–125.
Draper, N.R., Smith, H., 2014. Applied Regression Analysis. John Wiley & Sons.
Duan, Q., Gupta, V.K., Sorooshian, S., 1993. Shuffled complex evolution approach for effective and efficient global minimization. J. Optim. Theory Appl. 76 (3), 501–521.
Ertekin, C., Yaldiz, O., 2000. Comparison of some existing models for estimating global solar radiation for Antalya (Turkey). Energy Convers. Manag. 41 (4), 311–330.
Faisal, N., Gaffar, A., 2012. Development of Pakistan's new area weighted rainfall using Thiessen polygon method. Pakistan J. Meteorol. (Pakistan).
Fischer, M.J., Hinzmann, G., 2006. A new class of copulas with tail dependence and a generalized tail dependence estimator. Diskussionspapiere//Friedrich-Alexander-Universität Erlangen-Nürnberg. In: Lehrstuhl für Statistik und Ökonometrie.
Fister Jr., I., Fister, D., Yang, X.-S., 2013. A hybrid bat algorithm. arXiv preprint arXiv:1303.6310.
Fister, I., Yang, X.-S., Fong, S., Zhuang, Y., 2014. Bat algorithm: recent advances. computational intelligence and informatics (CINTI). In: 2014 IEEE 15th international symposium on. IEEE, pp. 163–167.
Graham, S., 2002. Weather Forecasting through the Ages: Feature Articles.
Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3 (Mar), 1157–1182.
Haario, H., Saksman, E., Tamminen, J., 2001. An adaptive metropolis algorithm. Bernoulli 7 (2), 223–242.
Hardwinarto, S., Aipassa, M., 2015. Rainfall monthly prediction based on artificial neural network: a case study in Tenggarong Station, East Kalimantan-Indonesia. Procedia Computer Science. 59, 142–151.
He, S., Raghavan, S.V., Nguyen, N.S., Liong, S.Y., 2013. Ensemble rainfall forecasting with numerical weather prediction and radar-based nowcasting models. Hydrol. Process. 27 (11), 1560–1571.
Hicks, M.J., Burton, M.L., 2010. Preliminary Damage Estimates for Pakistani Flood Events. Center for Business and Economic Research, Ball State University.
Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1998. Bayesian model averaging.

In: Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, pp. 77–83.
Holschneider, M., 1988. On the wavelet transformation of fractal objects. J. Stat. Phys. 50 (5), 963–993.
Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A Practical Guide to Support Vector Classification.
Huang, G.-B., 2003. Learning capability and storage capacity of two-hidden-layer feedforward networks. IEEE Trans. Neural Netw. 14 (2), 274–281.
Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2006. Extreme learning machine: theory and applications. Neurocomputing 70 (1), 489–501.
Hung, N.Q., Babel, M.S., Weesakul, S., Tripathi, N., 2009. An artificial neural network model for rainfall forecasting in Bangkok, Thailand. Hydrol. Earth Syst. Sci. 13 (8), 1413–1425.
Jain, A., Zongker, D., 1997. Feature selection: evaluation, application, and small sample performance. IEEE Trans. Pattern Anal. Mach. Intell. 19 (2), 153–158.
Jang, J.-S., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Transactions on Systems, Man, and Cybernetics 23 (3), 665–685.
Jones, J.W., Hansen, J.W., Royce, F.S., Messina, C.D., 2000. Potential benefits of climate forecasting to agriculture. Agric. Ecosyst. Environ. 82 (1–3), 169–184.
Kashiwao, T., et al., 2017. A neural network-based local rainfall prediction system using meteorological data on the internet: a case study using data from the Japan meteorological agency. Appl. Soft Comput. 56, 317–330.
Khalighi, S., Sousa, T., Oliveira, D., Pires, G., Nunes, U., 2011. Efficient feature selection for sleep staging based on maximal overlap discrete wavelet transform and SVM. In: Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, IEEE, pp. 3306–3309.
Khedun, C.P., Mishra, A.K., Singh, V.P., Giardino, J.R., 2014. A copula-based precipitation forecasting model: investigating the interdecadal modulation of ENSO's impacts on monthly precipitation. Water Resour. Res. 50 (1), 580–600.
Lan, Y., Soh, Y.C., Huang, G.-B., 2009. Ensemble of online sequential extreme learning machine. Neurocomputing 72 (13–15), 3391–3395.
Langridge, R., Christian-Smith, J., Lohse, K., 2006. Access and resilience: analyzing the construction of social resilience to the threat of water scarcity. Ecol. Soc. 11 (2).
Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241.
Lei, K.S., Wan, F., 2012. Applying ensemble learning techniques to ANFIS for air pollution index prediction in Macau. In: International Symposium on Neural Networks. Springer, pp. 509–516.
Li, C., Singh, V.P., Mishra, A.K., 2013. A bivariate mixed distribution with a heavy-tailed component and its application to single-site daily rainfall simulation. Water Resour. Res. 49 (2), 767–789.
Liang, N.-Y., Huang, G.-B., Saratchandran, P., Sundararajan, N., 2006. A fast and accurate online sequential learning algorithm for feedforward networks. IEEE Trans. Neural Netw. 17 (6), 1411–1423.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R News. 2 (3), 18–22.
Lin, G.F., Chen, G.R., Wu, M.C., Chou, Y.C., 2009. Effective forecasting of hourly typhoon rainfall using support vector machines. Water Resour. Res. 45 (8).
Ling, S., Li, W., 1997. On fractionally integrated autoregressive moving-average time series models with conditional heteroscedasticity. J. Am. Stat. Assoc. 92 (439), 1184–1194.
Luk, K.C., Ball, J.E., Sharma, A., 2001. An application of artificial neural networks for rainfall forecasting. Math. Comput. Model. 33 (6–7), 683–693.
Mansoor, H., 2010. Pakistan Evacuates Thousands in Flooded South.
Maraseni, T.N., Mushtaq, S., Reardon-Smith, K., 2012. Integrated analysis for a carbon- and water-constrained future: an assessment of drip irrigation in a lettuce production system in eastern Australia. J. Environ. Manag. 111, 220–226.
Mason, S., 1998. Seasonal forecasting of south African rainfall using a non–linear discriminant analysis model. Int. J. Climatol. 18 (2), 147–164.
Moazami, S., Golian, S., Kavianpour, M.R., Hong, Y., 2014. Uncertainty analysis of bias from satellite rainfall estimates using copula method. Atmos. Res. 137, 145–166.
Mohammadi, K., et al., 2015. A new hybrid support vector machine–wavelet transform approach for estimation of horizontal global solar radiation. Energy Convers. Manag. 92, 162–171.
Mojtaba Sadegh, E.R., AghaKou, Amir, 2017. Multivariate copula analysis toolbox (MvCAT): describing dependence and underlying uncertainty using a Bayesian framework. Water Resour. Res. 53 (6), 17.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. 10 (3), 282–290.
Nelsen, R.B., 2003. Properties and applications of copulas: A brief survey. In: Dhaene, J., Kolev, N., Morettin, P.A. (Eds.), Proceedings of the First Brazilian Conference on Statistical Modeling in Insurance and Finance. University Press USP, Sao Paulo, pp. 10–28.
News, D., 2010. Floods to Hit Economic Growth: Finance Ministry. Dawn News.
Nguyen-Huy, T., Deo, R.C., An-Vo, D.-A., Mushtaq, S., Khan, S., 2017. Copula-statistical precipitation forecasting model in Australia's agro-ecological zones. Agric. Water Manag. 191, 153–172.
Nguyen-Huy, T., Deo, R.C., Mushtaq, S., An-Vo, D.-A., Khan, S., 2018. Modeling the joint influence of multiple synoptic-scale, climate mode indices on Australian wheat yield using a vine copula-based approach. Eur. J. Agron. 98, 65–81.
Niaz, M.S., 2014. Wheat Policy—A Success or Failure. DAWN Newspaper.
Ortiz-García, E., Salcedo-Sanz, S., Casanova-Mateo, C., 2014. Accurate precipitation prediction with support vector classifiers: a study including novel predictive variables and observational data. Atmos. Res. 139, 128–136.
Palmer, W.C., 1965. Meteorological drought. US Department of Commerce, Weather Bureau, Washington, DC.
PMD, 2016. Pakistan meteorological department, Pakistan.
Prasad, R., Deo, R.C., Li, Y., Maraseni, T., 2017. Input Selection and Performance Optimization of ANN-Based Streamflow Forecasts in a Drought-Prone Murray Darling Basin Using IIS and MODWT Algorithm. Atmospheric Research.
Raftery, A.E., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133 (5), 1155–1174.
Rajesh, R., Prakash, J.S., 2011. Extreme learning machines-a review and state-of-the-art. International Journal of Wisdom based Computing. 1 (1), 35–49.
Reale, O., Lau, K., Susskind, J., Rosenberg, R., 2012. AIRS impact on analysis and forecast of an extreme rainfall event (Indus River Valley, Pakistan, 2010) with a global data assimilation and forecast system. J. Geophys. Res. 117 (D8).
Rilling, G., Flandrin, P., Goncalves, P., 2003. On empirical mode decomposition and its algorithms. In: IEEE-EURASIP workshop on nonlinear signal and image processing. IEEER, Grado, Italy, pp. 8–11.
Robert, B., Yoav, F., Peter, B., Sun, L.W., 1998. Boosting the margin: a new explanation for the effectiveness of voting methods. Ann. Stat. 26 (5), 1651–1686.
Salcedo-Sanz, S., Deo, R.C., Cornejo-Bueno, L., Camacho-Gómez, C., Ghimire, S., 2018. An efficient neuro-evolutionary hybrid modelling mechanism for the estimation of daily global solar radiation in the Sunshine State of Australia. Appl. Energy 209, 79–94.
Salma, S., Shah, M., Rehman, S., 2012. Rainfall trends in different climate zones of Pakistan. Pakistan Journal of Meteorology. 9 (17).
Sánchez-Monedero, J., Salcedo-Sanz, S., Gutiérrez, P.A., Casanova-Mateo, C., Hervás-Martínez, C., 2014. Simultaneous modelling of rainfall occurrence and amount using a hierarchical nominal–ordinal support vector classifier. Eng. Appl. Artif. Intell. 34, 199–207.
Santos, R., Fernandes, L.S., Moura, J., Pereira, M., Pacheco, F., 2014. The impact of climate change, human interference, scale and modeling uncertainties on the estimation of aquifer properties and river flow components. J. Hydrol. 519, 1297–1314.
Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S., 1998. Boosting the margin: a new explanation for the effectiveness of voting methods. Ann. Stat. 1651–1686.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.
Segal, M.R., 2004. Machine Learning Benchmarks and Random Forest Regression.
Servey, B., 2016. Asian Urban Information of Kobe.
Sharma, A., 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: part 3—a nonparametric probabilistic forecast model. J. Hydrol. 239 (1), 249–258.
Tamura, S.i., Tateishi, M., 1997. Capabilities of a four-layered feedforward neural network: four layers versus three. IEEE Trans. Neural Netw. 8 (2), 251–255.
Tarakzai, S., 2010. Pakistan Battles Economic Pain of Floods. Jakarta Globe.
Terêncio, D., Fernandes, L.S., Cortes, R., Pacheco, F., 2017. Improved framework model to allocate optimal rainwater harvesting sites in small watersheds for agro-forestry uses. J. Hydrol. 550, 318–330.
Terêncio, D., Fernandes, L.S., Cortes, R., Moura, J., Pacheco, F., 2018. Rainwater harvesting in catchments for agro-forestry uses: a study focused on the balance between sustainability values and storage capacity. Sci. Total Environ. 613, 1079–1092.
Thyer, M., et al., 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: a case study using Bayesian total error analysis. Water Resour. Res. 45 (12).
Toth, E., Brath, A., Montanari, A., 2000. Comparison of short-term rainfall prediction models for real-time flood forecasting. J. Hydrol. 239 (1), 132–147.
Villafuerte II, M.Q., et al., 2014. Long-term trends and variability of rainfall extremes in the Philippines. Atmos. Res. 137, 1–13.
Vörösmarty, C.J., et al., 2010. Global threats to human water security and river biodiversity. Nature 467 (7315), 555–561.
Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2 (2), 184–194.
Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 63 (11), 1309–1313.
Willmott, C.J., 1984. On the evaluation of model performance in physical geography. In: Spatial statistics and models. Springer, pp. 443–460.
Willmott, C.J., Robeson, S.M., Matsuura, K., 2012. A refined index of model performance. Int. J. Climatol. 32 (13), 2088–2094.
Yadav, B., Ch, S., Mathur, S., Adamowski, J., 2016. Discharge forecasting using an online sequential extreme learning machine (OS-ELM) model: a case study in Neckar River, Germany. Measurement 92, 433–445.
Yang, X.-S., 2010. Nature-Inspired Metaheuristic Algorithms. Luniver press.
Yang, X.-S., 2011. Bat algorithm for multi-objective optimisation. International Journal of Bio-Inspired Computation. 3 (5), 267–274.
Yang, X.-S., He, X., 2013. Bat algorithm: literature review and applications. International Journal of Bio-Inspired Computation. 5 (3), 141–149.
Yaseen, Z.M., 2016. Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq. J. Hydrol. 542, 603–614.
Yaseen, Z.M., et al., 2017. Rainfall pattern forecasting using novel hybrid intelligent model based ANFIS-FFA. Water Resour. Manag. 32 (1), 105–122.
Yaseen, Z.M., 2018. Application of the hybrid artificial neural network coupled with rolling mechanism and Grey Model algorithms for streamflow forecasting over multiple time horizons. Water Resour. Manag. 1–17.
Yun, Z., et al., 2008. RBF neural network and ANFIS-based short-term load forecasting approach in real-time price environment. IEEE Trans. Power Syst. 23 (3), 853–858.
Zhisheng, Z., 2010. Quantum-behaved particle swarm optimization algorithm for economic load dispatch of power system. Expert Syst. Appl. 37 (2), 1800–1803.

Chapter 4

An ensemble-ANFIS based uncertainty assessment model for forecasting multi-scalar standardized precipitation index

Foreword

This chapter is an exact copy of the published article in *Atmospheric Research* journal (Vol. 207, Pages 155-180).

Ensemble techniques are supervised learning algorithms that are used to minimize the uncertainty and to produce more reliable and consistent predictions. Further to rainfall forecasting (Chapter 3), forecasting of another important hydrological variable, the drought, is undertaken in this chapter with the employment of ensemble modelling technique. An ensemble adaptive neuro-fuzzy inference system (ensemble-ANFIS) based uncertainty assessment modeling approach is developed for medium and long term (3-, 6-, 12-months) drought forecasting. Applying 10-member simulations, ensemble-ANFIS model was validated for its ability to forecast severity (S), duration (D) and intensity (I) of drought. This enabled uncertainty between multi-models to be rationalized more efficiently, leading to a reduction in forecast error caused by stochasticity in drought behaviours.

The results of ensemble-ANFIS model are benchmarked with M5Model Tree and Minimax Probability Machine Regression (MPMR) at three candidate sites in Pakistan. The accuracy of ensemble-ANFIS model is better than M5Tree and MPMR models.

Atmospheric Research 207 (2018) 155–180

Contents lists available at ScienceDirect

Atmospheric Research

journal homepage: www.elsevier.com/locate/atmosres

An ensemble-ANFIS based uncertainty assessment model for forecasting multi-scalar standardized precipitation index

Mumtaz Ali*, Ravinesh C. Deo*, Nathan J. Downs, Tek Maraseni

School of Agricultural, Computational and Environmental Sciences, Institute of Agriculture and Environment, University of Southern Queensland, Springfield, QLD 4300, Australia

ARTICLE INFO

Keywords:
Standardized precipitation index
Drought forecasting
Ensemble based adaptive neuro fuzzy inference system
M5 tree
Minimax probability machine regression

ABSTRACT

Forecasting drought by means of the World Meteorological Organization-approved Standardized Precipitation Index (SPI) is considered to be a fundamental task to support socio-economic initiatives and effectively mitigating the climate-risk. This study aims to develop a robust drought modelling strategy to forecast multi-scalar SPI in drought-rich regions of Pakistan where statistically significant lagged combinations of antecedent SPI are used to forecast future SPI. With ensemble-Adaptive Neuro Fuzzy Inference System ('ensemble-ANFIS') executed via a 10-fold cross-validation procedure, a model is constructed by randomly partitioned input-target data. Resulting in 10-member ensemble-ANFIS outputs, judged by mean square error and correlation coefficient in the training period, the optimal forecasts are attained by the averaged simulations, and the model is benchmarked with M5 Model Tree and Minimax Probability Machine Regression (MPMR). The results show the proposed ensemble-ANFIS model's preciseness was notably better (in terms of the root mean square and mean absolute error including the Willmott's, Nash-Sutcliffe and Legates McCabe's index) for the 6- and 12- month compared to the 3-month forecasts as verified by the largest error proportions that registered in smallest error band. Applying 10-member simulations, ensemble-ANFIS model was validated for its ability to forecast severity (S), duration (D) and intensity (I) of drought (including the error bound). This enabled uncertainty between multi-models to be rationalized more efficiently, leading to a reduction in forecast error caused by stochasticity in drought behaviours. Through cross-validations at diverse sites, a geographic signature in modelled uncertainties was also calculated. Considering the superiority of ensemble-ANFIS approach and its ability to generate uncertainty-based information, the study advocates the versatility of a multi-model approach for drought-risk forecasting and its prime importance for estimating drought properties over confidence intervals to generate better information for strategic decision-making.

1. Introduction

Drought is considered a slow onset natural hazard that can last for a prolonged dry period in natural climate cycles, and is usually restricted to arid, semi-arid, desert or rain-forest regions (Keyantash and Dracup, 2002; Vicente-Serrano, 2016; Wilhite et al., 2000). The occurrence of drought can happen in any climate zone. Environmental parameters such as high surface temperature and winds, low humidity and the timing and characteristics of rainfall are greatly influenced by drought occurrences. Distribution of rain days in crop growing seasons, intensity and duration of rainfall, as well as onsets and terminations are important factors that influence the duration and severity of drought. The occurrence of drought leads to consequences for runoff that affects stream flow in agricultural sectors (Cai and Cowan, 2008) with substantial economic costs (Dijk et al., 2013; Wittwer et al., 2002). Advanced intelligent data models can effectively manage the drought-risk with better economic returns (Koehn, 2015; Qureshi et al., 2016; Timbal and Hendon, 2011; Williams et al., 2015). Therefore, modelling approaches have been adopted to forecast drought behaviour, utilizing hydrological models (Brown et al., 2015), Markov chain (Rahmat et al., 2016), Bayesian space–time models (Crimp et al., 2015) and recently, data-intelligent models (Abbot and Marohasy, 2012; Abbot and Marohasy, 2014; Dayal et al., 2016a; Dayal et al., 2016b; Dayal et al., 2017; Deo et al., 2017; Deo and Şahin, 2015a; Deo and Şahin, 2015b; Deo and Şahin, 2016; Deo et al., 2016b). Moreover, the study of Rahmat et al. (2016), the study of Deo et al. (2017) and Dayal et al. (2016a) have utilized Markov chain and multiple intelligent data models, respectively, to forecast the standard precipitation index (SPI), a primary indicator of drought status.

Drought is a socio-economic hazard, happening on a year-to-year

* Corresponding authors.
E-mail addresses: Mumtaz.Alli@usq.edu.au (M. Ali), ravinesh.deo@usq.edu.au (R.C. Deo), nathan.downs@usq.edu.au (N.J. Downs), tek.maraseni@usq.edu.au (T. Maraseni).

https://doi.org/10.1016/j.atmosres.2018.02.024
Received 6 November 2017; Received in revised form 22 January 2018; Accepted 28 February 2018
Available online 02 March 2018

and a season-to-season basis with severe threats to groundwater reservoirs, leading to the scarcity of water, crop failure, disturbed habitats and loss of social or recreational opportunity (Deo et al., 2015; Mpelasoka et al., 2008; Riebsame et al., 1991; Wilhite et al., 2000). The impacts of drought are extremely aggravated in drought-prone regions which suffer even warmer and drier conditions. The severe scarcity of water resources due to rapid growth of population and an expansion of agricultural, energy and industrial sectors is a growing concern (IPCC, 2012; McAlpine et al., 2007). Hence, the challenges posed by drought, particularly prolonged drought events, can influence hydrologists, agriculturalists and natural resource managers in everyday strategic decision tasks (Bates et al., 2008; Deo et al., 2016a; Mishra and Singh, 2010; Mishra and Singh, 2011; Wilhite and Hayes, 1998). To tackle drought-related issues, effective strategies that aim to tackle the impact of current, and predict the occurrence of future drought events are imperative.

Drought indices are used to monitor and forecast drought (Mishra and Singh, 2010; Mishra and Singh, 2011). Different indices have been developed based on their appropriateness in a given region. Palmer Drought Severity Index (PDSI), designed by Palmer (1965), measures the dryness based on precipitation and temperature. PDSI is an effective index in determining long-term drought, extending over several months, but it is not that effective for describing short-term events. Other complexities of the PDSI include the estimation of soil moisture amounts and the 'hard-to-calibrate' drought parameters, including evapotranspiration and recharge rates in climatically diverse global regions. Responding to the drawbacks associated with the PDSI, researchers have developed the Crop Moisture Index (CMI), particularly for agricultural drought events (Palmer, 1968). CMI enumerates the rank of the precipitation to compute positive and negative precipitation anomalies. CMI is a weekly based index capable of determining the short-term drought impact on agriculture. However, it has a weakness for long-term drought, as improvements in the short-term may be insufficient to offset the long-term issues. Byun and Wilhite (1999) introduced the Effective Drought Index (EDI) that was developed on daily precipitation data. As also advocated by the study of Deo and Şahin (2015a, 2015b), the EDI can be considered as a relatively good index for operational monitoring of meteorological and agricultural drought, although the consideration of precipitation alone does not take into account other environmental parameters that also cause drought (e.g., the impact of temperature).

In this paper we focus on the SPI, an index introduced by McKee et al. (1993) as a global drought monitoring tool that was based on the Lincoln Declaration of drought by the World Meteorological Organization (Hayes et al., 2011). National Meteorological and Hydrological Services in USA and elsewhere have adopted the SPI to characterize meteorological drought (Cancelliere et al., 2007; Choubin et al., 2016; Hayes et al., 1999; Jalalkamali et al., 2015; McKee et al., 1993; Svoboda et al., 2012; Yuan and Zhou, 2004), with a recent study of Deo et al. (2017) that has also applied this index for drought modelling in Australia. SPI is a powerful index because: (1) it examines the water shortage situation that construct a statistical distribution of rainfall ranging from 1 to 48 months (enabling short and long-term assessments); (2) it generates a normalized standard metric of rainfall surpluses/deficits in relation to a benchmark climatology (Hayes et al., 1999; McKee et al., 1993; Yuan and Zhou, 2004) (enabling its comparison in geographically diverse regions); (3) it has been explored for drought mitigation studies in diverse climatic regions (Almedeij, 2016; Choubin et al., 2016; Svoboda et al., 2012); and (4) SPI has the potential to represent short and long-term drought in a probabilistic fashion on multiple timescales. These broad representations and usages of SPI makes it possible to examine soil moisture levels that express the precipitation anomalies on a comparatively short timescale with hydrological reservoirs to replicate long-term anomalies (Svoboda et al., 2012). Due to these features, SPI has become an ideal metric for management of not only hydrological but also agricultural drought events (Guttman, 1999).

Intelligent models used to forecast drought based on SPI have been explored. An SPI-based methodology was proposed by Cancelliere et al. (2007) to evaluate the probabilities of drought transition in Sicily, Italy. Jalalkamali et al, (2015) in Yazd, Iran conducted a study to forecast SPI using multilayer perceptron artificial neural network (MLP ANN), adaptive neuro-fuzzy inference systems (ANFIS), support vector machine (SVM), and autoregressive integrated moving average (ARIMA) multivariate models. In another study, ANFIS and ANN Wavelet models were used by Shirmohammadi et al. (2013) to forecast SPI for Azerbaijan. An SPI-based forecasting study was done by Santos et al. (2009) using an ANN model for San Francisco. Moreira et al. (2015) have utilized satellite-based image data with climate indices (North Atlantic and Southern Oscillation Index) for drought forecasting with ANN. A non-parametric model was developed by Cancelliere et al. (2006) to forecast SPI for Sicily and Belayneh in Italy. An ANN model in comparison with SVR and wavelet neural network models were used by Adamowski et al. (2012) for estimation of SPI in the Awash River (Ethiopia). Choubin et al. (2016) used ANFIS, M5 model tree (M5Tree) and an MLP algorithm to present SPI forecasting estimates. Drought forecasting using SPI thus has an extensive history in the available literature (Bonaccorso et al., 2003; Deo et al., 2017; Guttman, 1999; Hayes et al., 1999; Jalalkamali et al., 2015; Moreira et al., 2008; Paulo and Pereira, 2007; Sönmez et al., 2005).

Moreover, several member bootstrap and ensemble techniques have been extensively used to reduce the uncertainty and to produce more reliable and consistent predictions (Cannon and Whitfield, 2002; Jeong and Kim, 2005). The bootstrap and ensemble techniques are based on a computational procedure that utilizes intensive resampling with replacement, in order to minimize uncertainty (Efron and Tibshirani, 1994). Abrahart (2003) developed a bootstrap technique for rainfall–runoff modelling to continuously sample the input space. Jeong and Kim (2005) designed an ensemble neural network using bootstrap technique to forecast monthly rainfall–runoff. Srivastav et al. (2007) proposed a bootstrap ANN model for uncertainty analysis to forecast river flow. Tiwari and Chatterjee (2010b) developed an ensemble technique for hourly flood forecasting capable of quantifying uncertainty and found the outputs to be more stable and accurate than comparative models. Tiwari and Chatterjee (2011) developed a wavelet–bootstrap–ANN model for daily discharge forecasting. In another study (Tiwari and Chatterjee, 2010a), they designed a wavelet–bootstrap based ANN model for accurate and reliable hourly flood forecasting.

Drought hazard has impacted Pakistan, the present study region quite significantly since the last few decades (Pakistan, 1950–2015) with severe socio-economic impacts. The duration of drought in 1998 reduced Pakistan's national agricultural production by 2.6% during the period 2000–2001 (Ahmad et al., 2004). Drought modelling studies in this developing nation has been very limited. The blockage of the Western Depression that carries rainclouds to the northern areas of Pakistan from the Mediterranean can catastrophically shift weather patterns making accurate long-term predictions difficult. El Niño, in particular, has a strong effect on the sub-tropical jet stream that manifests itself in the form of a potential drought event in summer (Zaidi, 2016). Khan and Gadiwala (2013) conducted a study of drought behaviour using the SPI at different timescales in the province of Sindh, Pakistan. Similarly, Xie et al. (2013) used a spatiotemporal variability analysis based on SPI data to forecast drought and recently, the study of Ali et al. (2017) forecasted drought based on Standardized Precipitation-Evapotranspiration Index (*Spe*I) with a multilayer perceptron-based artificial neural network model. Ahmed et al. (2016) utilized antecedent SPI for characterization of seasonal drought in the Balochistan Province of Pakistan. However, to the best of the authors' knowledge, there has been no study focussing on forecasting drought using SPI modelling in Pakistan.

In context of the gaps identified in current literature that relate to

Pakistan's socio-economic future, the focus of this paper, is to illuminate the utilization of the Ensemble-ANFIS approach and its comparison with M5 Tree and MPMR models for SPI forecasting in drought-rich regions of Pakistan where drought management is contingent upon accurate forecasting, especially over multiple horizons to support different kinds of socio-economic activities. The objectives are as follows. (1) To use 34-years of rainfall (1981–2015) and compute multi-scale SPI for 1, 6 and 12 months. (2) To develop the forecasting model with an Ensemble-ANFIS approach incorporating statistically significant lagged historical SPI data (i.e., incorporating antecedent drought behaviour) to reduce the uncertainty and forecast the future drought behaviour more reliable and consistent. (3) To benchmark the three different data-driven models: Ensemble-ANFIS vs. M5 Tree and MPMR models. (4) To evaluate the ensemble-ANFIS model by checking for the severity (S), duration (D) and peak intensity (I) of drought events determined from the forecasted and observed SPI. (5) To provide a predictive uncertainty assessment framework for multi-scalar SPI forecasting at diverse study sites.

2. Theoretical overviews

An overview of the objective forecasting model, Ensemble-ANFIS with the comparative counterpart models, M5Tree and MPMR, is presented in this section.

2.1. Ensemble based adaptive neuro fuzzy inference system (ensemble-ANFIS) model

ANFIS was introduced by Jang in 1993 as a division of the adaptive tool (i.e. the outputs being dependent on the parameters belonging to the input nodes). This is an improved ANN technique that is fundamentally identical to the fuzzy inference systems (FIS) model, yet utilizing the merits of both ANN and FIS designed on a common paradigm. During the training process, two types of learning algorithms are used to tune the parameters for optimal performance. The foresaid system utilizes two inputs to generate one output employing the fuzzy 'if-then' rules of Takagi-Sugeno-Kang's (TSK) (Hoffmann et al., 2007) fuzzy model which can be defined as:

$$\text{Rule(a): if } \alpha \text{ is } \Gamma_1 \text{ and } \beta \text{ is } \Omega_1, \text{ then } f_1 = p_1\alpha + q_1\beta + s_1 \tag{1}$$

$$\text{Rule(b): if } \alpha \text{ is } \Gamma_2 \text{ and } \beta \text{ is } \Omega_2, \text{ then } f_2 = p_2\alpha + q_2\beta + s_2 \tag{2}$$

where α and β represents the input of the ANFIS whereas Γand Ω are the fuzzy set with $f_j(j = 1,2)$ being the first order output polynomial of the TSK fuzzy inference system, while p_j, q_j and $s_j(j = 1,2)$ are the set of consequent parameters. Fig. 1(a) describes the ANFIS architecture.

The elementary construction of ANFIS can be seen as a 5 layer feedforward neural network. Each node j is an adaptive node in layer 1 with a suitable membership function related to the input to node$j$.

$$\Theta_{1,j} = \mu_{\Gamma_j}(\alpha),\ \Theta_{1,j} = \mu_{\Omega_{(j-2)}}(\beta),\ (j = 1,2) \tag{3}$$

In Eq. (3), α, β denote the input nodes whileΓ and Ωare the linguistic labels with $\mu(\alpha)$, $\mu(\beta)$representing the membership function (usually bell-shaped) that specifies the degree to which the given input satisfies the quantifiers Γ, Ω. The membership function is defined as:

$$\mu(\alpha) = \frac{1}{1 + \left(\frac{(\alpha - c_j)}{a_j}\right)^{b_j}} \tag{4}$$

where a_j, $b_j$and c_j are the parameters. The bell-shaped function adopts different forms of membership with the variation of these parameters. Each parameter in this layer is termed as antecedent. In layer 2, the node is a fixed circular node, labelled as Πwhere the input is multiplied with the node function to act as the output which is following.

$$\Theta_{2,j} = \omega_j = \mu_{\Gamma_j}(\alpha) \times \mu_{\Omega_j}(\beta) \tag{5}$$

here ω_j is the firing strength of a rule. Every node in layer 3 is also a fixed circular node labelled N with normalized firing strength as output which is basically the ratio of jth rule's firing strength to the sum of all rule's firing strength. Mathematically

$$\Theta_{3,j} = \overline{\omega_j} = \frac{\omega_j}{\omega_1 + \omega_2} \tag{6}$$

In layer 4, each node is turned to be an adaptive node marked by a square that is given by,

$$\Theta_{4,j} = \overline{\omega_j} f_j = \overline{\omega_j}(P_j\alpha + q_j\beta + s_j) \tag{7}$$

where $\overline{\omega_j}$ refers to the 3rd layer output with (p_j, q_j, s_j) being the parameters that are called consequent parameters. In layer 5, the consequent parameters are then expressed into a linear combination to compute the overall output of the fixed nodes. Consider,

$$\left.\begin{aligned}\Theta_{5,j} = f_{output} &= \sum_j \overline{\omega_j} f_j = \overline{\omega_1} f_1 + \overline{\omega_2} f_2 \\ &= \tfrac{\overline{\omega_1}}{\overline{\omega_1} + \overline{\omega_2}} f = \tfrac{\overline{\omega_2}}{\overline{\omega_1} + \overline{\omega_2}} f \\ &= (\overline{\omega_1}\alpha)p_1 + (\overline{\omega_1}\beta)q_1 + (\overline{\omega_1})s_1 \\ &+ (\overline{\omega_2}\alpha)p_2 + (\overline{\omega_2}\beta)q_2 + (\overline{\omega_2})s_2\end{aligned}\right\} \tag{8}$$

In the forward pass, the premise parameters (α_j, β_j, c_j) are stable while the least square estimation tries to identify the optimal consequent parameters. Then the backward pass starts with these parameters fixed to back propagate the ratio of the output node from the output end towards the input end. The gradient descent then updates these premise/antecedent parameters (Goyal et al., 2014; Jang et al., 1997; Karthika and Deka, 2015; Mayilvaganan and Naidu, 2011; Moosavi et al., 2013; Nayak et al., 2004; Pérez et al., 2012; Sehgal et al., 2014; Shirmohammadi et al., 2013). In this paper, several membership function were tested where the π-shaped curve membership function *(pimf)* was found to generate better results. The mathematical formulation of the π-shaped membership function is:

$$\mu(\alpha, a, b, c, d) = \left\{\begin{array}{cc} 0, & \alpha \le a \\ 2\left(\frac{\alpha - a}{b - a}\right)^2, & a \le \alpha \le \frac{a+b}{2} \\ 1 - 2\left(\frac{\alpha - b}{b - a}\right)^2, & \frac{a+b}{2} \le \alpha \le b \\ 1, & b \le \alpha \le c \\ 1 - 2\left(\frac{\alpha - c}{d - c}\right)^2, & c \le \alpha \le \frac{c+d}{2} \\ 2\left(\frac{\alpha - d}{d - c}\right)^2, & \frac{c+d}{2} \le \alpha \le d \\ 0, & \alpha \ge d \end{array}\right\} \tag{9}$$

To improve the versatility of a classical ANFIS model, we have designed 10-member ensemble frameworks to attain the most accurate forecast of SPI over multiple timescales. It is especially noted that several studies have used this 10-model ensemble approach, such as Burges et al. (2011) who developed 10 ensemble model using the lambda-gradient method, Deo et al. (2009) who developed a 10 member ensemble for modelling daily climate extremes including droughts with global climate models in eastern Australia, and Rajathi and Jayashree (2016) who designed 10 ensemble models to forecast soil moisture in India. Some other literature can also be seen in previous works (e.g., Bachman et al., 2014; Jamroz et al., 2016; Opitz and Maclin, 1999; Strauss et al., 2017; Taniar, 2009). Generally, an ensemble (Lei and Wan, 2012; Zhou et al., 2002) is a multiple component learner technique where several lots of training procedures are applied to perform the same objective task. In this study, the technique of random sampling of the training data was adopted to construct the ensemble forecasting systems (Zhou et al., 2002).

2.2. Minimax probability machine regression (MPMR) model

MPMR, a non-linear probabilistic machine regression model that

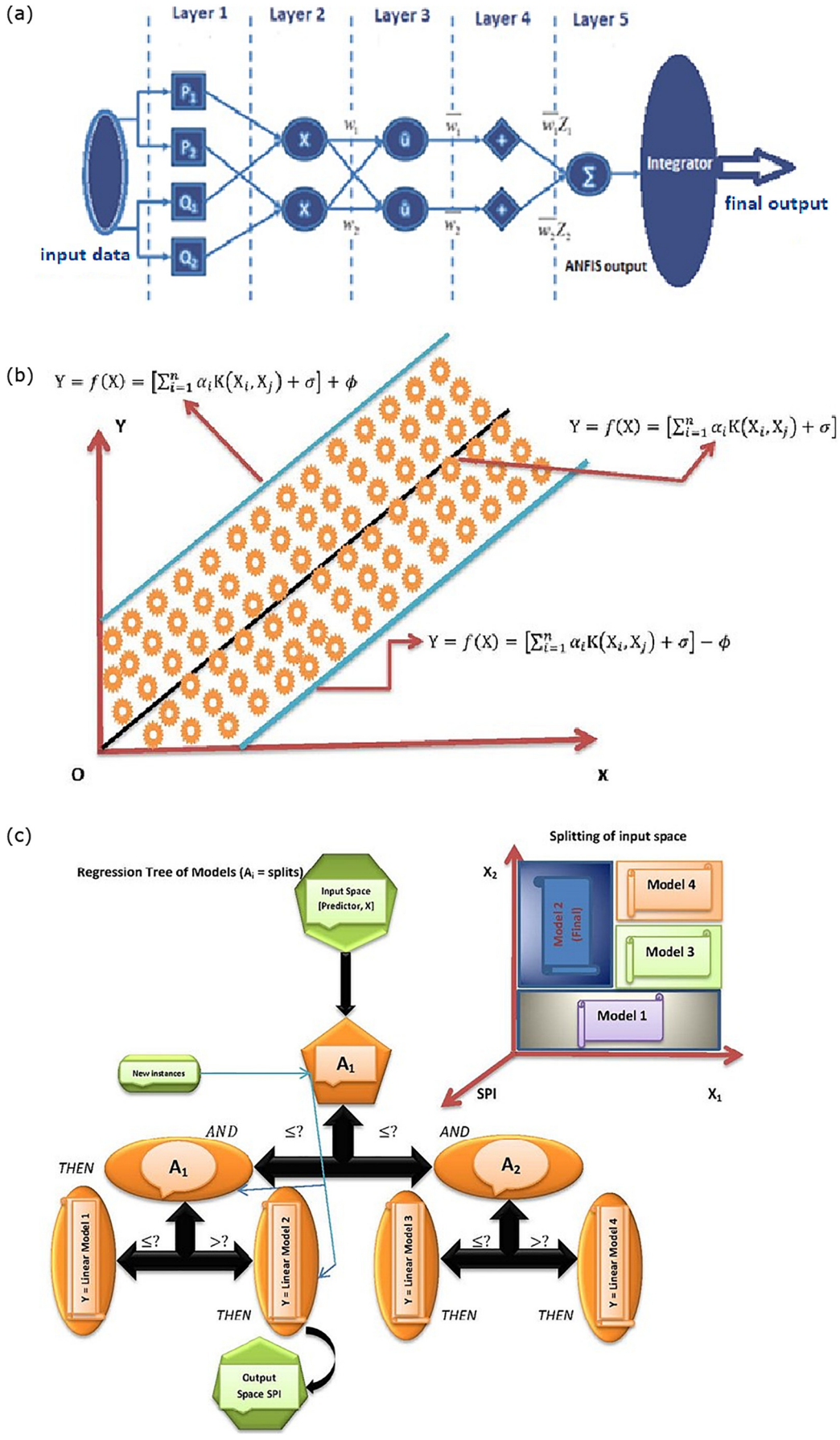


Fig. 1. Basic structure of: (a) Ensemble-ANFIS, (b) MPMR and (c) M5 Tree models.

has the capability of maximizing the least probability within the interval of true regression of the objective function, was also applied to benchmark the Ensemble-ANFIS model. Fig. 1(b) outlines the MPMR model.

MPMR is developed on linear discriminant and convex optimizations (Strohmann and Grudic, 2003) that makes MPMR an advanced and improved version of Support Vector Machines (Strohmann and Grudic, 2003). The data are analysed by shifting all of the regression data between $+\phi$ and $-\phi$ along the axis of the dependent variable. The area between them is called a regression surface where upper and lower bounds of probability identify the regions for misclassifying a point without making distributional assumptions by the model (Bertsimas and Sethuraman, 2000). The learning (d-dimensional inputs) generated from an unknown regression is a function of the form (Strohmann and Grudic, 2003):

$$f: \Re^{D} \rightarrow \Re$$

$$Y = f(x) + \sigma \tag{10}$$

where $x \in R^{D}$ is the input vector according to a bounded distributionΩ, $Y \in \Re$ is an output vector and variance $(\sigma) = \rho^2$ $(\rho) = \sigma^2 \in \Re$. The function $\hat{f}$ is an approximated function in MPMR for x_i generated from Ω. $\hat{Y}$ is given by

$$\hat{Y} = \hat{f}(x) \tag{11}$$

The estimated bounds calculated by the model is based on the minimum probability (Ψ) that $\hat{f}(x)$is within ϕ of Y (Strohmann and Grudic, 2003).

$$\Psi = \inf \ \Pr\{|\hat{Y} \ - Y| \leq \phi\} \tag{12}$$

The predictive ability of a true regression is then evaluated from Eq. (12) by a boundary based on the minimax probability to deduce Ψ directly between ϕ of the true function shown below:

$$\hat{Y} = \hat{f}(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x_j) + \sigma \tag{13}$$

In Eq. (13), $K_{i,\,j} = \theta(x_i, x_j)$ is a kernel function satisfying the mercer condition, vector x_i is from the learned data andα_i, σ are the output parameters.

2.3. M5 tree model

Ensemble-ANFIS is benchmarked with respect to the M5 Tree model. The original M5 Tree model was pioneered by Quinlan (1992) as a hierarchical model developed on binary decision structure. The linear regression at terminal (tree) nodes develop connections between inputs and output (Mitchell, 1997). For the construction of the decision tree, the input-output matrix is partitioned into subsets in two distinct phases (Rahimikhoob et al., 2013). The characterization of N-sample training matrices through input patterns in relation to a predictor, is related by a model that is developed by the M5 Tree algorithm (Bhattacharya and Solomatine, 2005). Based on matching attributes, the M5 Tree establishes a relationship between inputs and the SPI data used in drought forecasting. Fig.1(c) describes the M 5Tree structure.

The model utilizing an M5 Tree paradigm is developed based on the divide-and-conquer rule in which the associated N data points to a leaf or test criteria that branches into subgroups parallel to the test outcome. The process is recursive based on which N data points splits into subgroups trailing a principal that relies on the standard deviation and decreasing the model training error, ρ_R (Bhattacharya and Solomatine, 2005; Kisi, 2015):

$$\rho_R = \rho(\Omega) - \sum \left(\frac{\Omega_j}{\Omega} \rho(\Omega_j) \right) \tag{14}$$

here Ω represents the set of examples while $\Omega_j$is the subset of thejthoutcome.

In order to design an optimal M5 Tree model with a lowest ρ_R, M5 Tree picks them to optimize ρ_R by attaining the maximum split comprising the patterns and data attributes. The process of splitting stops either when the class value of all split cases reach a stationary node or only a few instances are left. As a result, the division rules that operate on input data may lead to a very large complex network structure that needs to be pruned back. To overcome the sudden discontinuities appearing from a smaller set of training data, a smoothing technique is needed to tackle what can happen between neighboring linear models at the leaves of the pruned tree (Bhattacharya and Solomatine, 2005; Kisi, 2015). This optimizes the tuned model to achieve a better accuracy. The purpose of the smoothing is to modify the linear equations in order to bring closer the equations corresponding to the forecasted output of the input variables (Quinlan, 1992).

2.4. Standardized precipitation index (SPI)

SPI is essentially a probability-based drought metric that provides a representation of abnormal wetness and dryness conditions. Prior to the design of a forecast model for the selected study regions in Pakistan, the short- and long-term SPI index was calculated following the notion of McKee et al. (1993). The computation of SPI involved the fitting of a Pearson Type III distribution to monthly rainfall (ppt) data.

Mathematically, the Pearson Type III distribution/gamma distribution function is given below:

$$g(ppt) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} (ppt)^{\alpha-1} e^{-x/\beta} \tag{15}$$

where α and β denotes the estimated parameters using the maximum likelihood solution:

$$\alpha = \frac{1}{4A}\left(1 + \sqrt{\frac{1 + 4A}{3}}\right) \tag{16}$$

$$\beta = \frac{\alpha}{\overline{P}} \tag{17}$$

and

$$A = \ln(\overline{ppt}) - (\sum \ln(ppt))/N \tag{18}$$

where N = the number of rainfall observation months. The cumulative probability can then be given by

$$G(ppt) = \int_0^P g(ppt)dppt = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} \int_0^P x^{\alpha-1} e^{-P/\beta} dppt \tag{19}$$

Suppose that $m = ppt/\beta$, that reduces Eq. (18) to an incomplete gamma function:

$$G(ppt) = \frac{1}{\Gamma(\alpha)} \int_0^m m^{\alpha-1} e^{-m} dm \tag{20}$$

As for $ppt = 0$, the gamma function is undefined, so the cumulative probability becomes:

$$H(ppt) = q + (1 - q)\, G(ppt) \tag{21}$$

where q is the probability of zero. The cumulative probability $H(ppt)$ can be transformed into the standard normal random variable with mean zero and variance of one. This yields the monthly value of SPI, viz:

$$SPI = \begin{cases} +\left(m - \frac{c_0 + c_1 m + c_2 m^2}{1 + d_1 m + d_2 m^2 + d_3 m^3}\right), & 0.5 < H(ppt) \leq 1.0 \\ -\left(m - \frac{c_0 + c_1 m + c_2 m^2}{1 + d_1 m + d_2 m^2 + d_3 m^3}\right), & 0 < H(ppt) \leq 0.5 \end{cases} \tag{22}$$

In Eq. (22), m is given by:

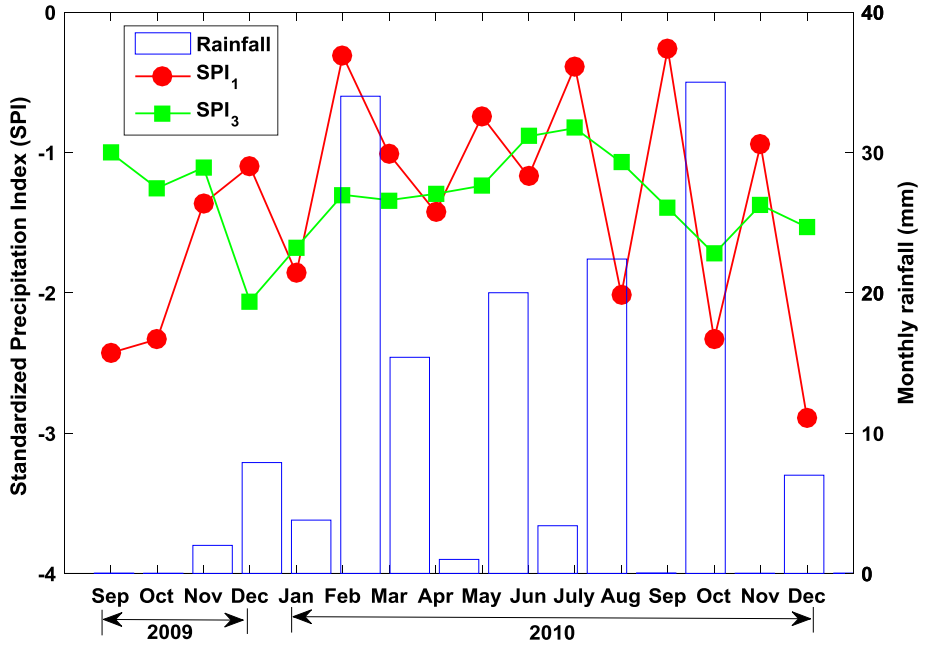


Fig. 2. The 1 and 3 month Standardized Precipitation Index (SPI) with rainfall data for drought periods (Sept 2009 to Dec 2010) at Islamabad station, Pakistan.

$$m = \begin{cases} \sqrt{\ln\left(\frac{1}{[H(ppt)]^2}\right)} & 0 < H(ppt) \leq 0.5 \\ \sqrt{\ln\left(\frac{1}{[1.0 - H(ppt)]^2}\right)} & 0.5 < H(ppt) \leq 1.0 \end{cases} \quad (23)$$

The values of the constants in Eq. (21) are as follows: $c_0 = 2.515517$, $c_1 = 0.802853$, $c_2 = 0.010328$, $d_1 = 1.432788$, $d_2 = 0.189269$ and $d_3 = 0.001308$ (McKee et al., 1993). These constants are used in the transformation of cumulative probability H(ppt) into the standard normal random variable SPI with mean zero and variance 1. Further, these constants are helpful in categorize the drought conditions. Moreover, the So, the standardized precipitation index represents an SPI-score or the number of standard deviations that an event deviates from the mean (Loukas and Vasiliades, 2004; McKee et al., 1993; Sujitha, 2017). Based on the calculated SPI, drought can be categorized as moderately dry = ($-1.5 < \text{SPI} \leq 1.0$), severely dry = ($-2.0 < \text{SPI} \leq -1.5$), and extremely dry = (SPI ≤ -2.0).

Fig. 2 illustrates the 1 and 3 month SPI time-series, depicting the progression of drought events from Sep 2009–Dec 2010 for the Islamabad station. The onset of drought using the running sum approach of Yevjevich (1967a, 1991) can be deduced as the particular month when the SPI value declined below 0 and the termination of drought when the SPI value first returned to positivity. In concurrence with this, the cumulative rainfall appears to be reduced significantly in this dry period. The drought duration is then the sum of all months with SPI < 0 and the drought's peak intensity occurs when the SPI value is at its minimum point.

3. Materials and methods

In this section the description of the acquired rainfall data for the study region in Pakistan, the development of data-intelligent models and the performance evaluation are presented.

3.1. Rainfall data

The rainfall data (1981–2015) were sourced from the Pakistan Meteorological Department, Pakistan (Department). The rainfall of TRACE (amount of rainfall < 0.1 mm) per month was replaced by the average of the respective time-averaged value from the climatological period. Prior to developing the drought model, the rainfall data were used to compute multi-scale short- and long term SPI from 1981 to 2015 for the study regions.

3.2. Study region

The three study regions for this work are: Dera Ismail Khan (denoted 'D. I. Khan'), Islamabad and Nawabshah. Fig. 3 shows a map of the study regions.

D. I. Khan (31.8424° N, 70.8952° E) is situated in Khyber Pakhtunkhwa (KPK) province. The climate consists of hot summers and mild winters. Water scarcity due to drought in D. I. Khan had badly affect wheat production in 2012 (Amir, 2012). The average annual rainfall is about 268.8 mm. Islamabad (33.7294° N, 73.0931° E) is the capital of Pakistan which has a humid subtropical climate with five seasons: winter, spring, summer, rainy monsoon and autumn. The average monsoon rainfall is about 790.8 mm. Islamabad received the heaviest 620 mm rainfall in just 10 h on 23 July 2002, which was the heaviest rainfall in the past 100 years. The annual average rainfall is 1142.1 mm. Nawabshah (26.2442° N, 68.4100° E) is a city of Sindh province, located 50 km from the River Indus. The climate is considered to be the hottest of all 3 locations with summer temperatures reaching as high as 53 °C. A record breaking severe heatwave hit Nawabshah in 2010 (Department). The average annual rainfall is 114.1 mm and the climate is also affected by monsoons in Nawabshah.

Table 1 describes the latitude, longitude, elevation, minimum, maximum, standard deviation, skewness and kurtosis of rainfall and SPI data of the selected study regions that are used to develop the forecasting models in this paper. A 35 year rainfall dataset from 1981 to 2015 was acquired from the Pakistan Meteorological Department (Department) to calculate SPI. Antecedent months from this dataset were used as significant SPI lags to develop the forecasting model where SPI data for 35 years were partitioned into 70% (training) and 30% (testing) periods.

To investigate the dependence of the input variable with itself at two points in time (that recognises the role of memory in forecasting SPI), partial autocorrelation function (PACF) was used to deduce out the autocorrelation that measured the dependency of one variable after removing the effect of the other variable(s). The autocorrelation can be obtained following Box et al. (2015) and Hamilton (1994).

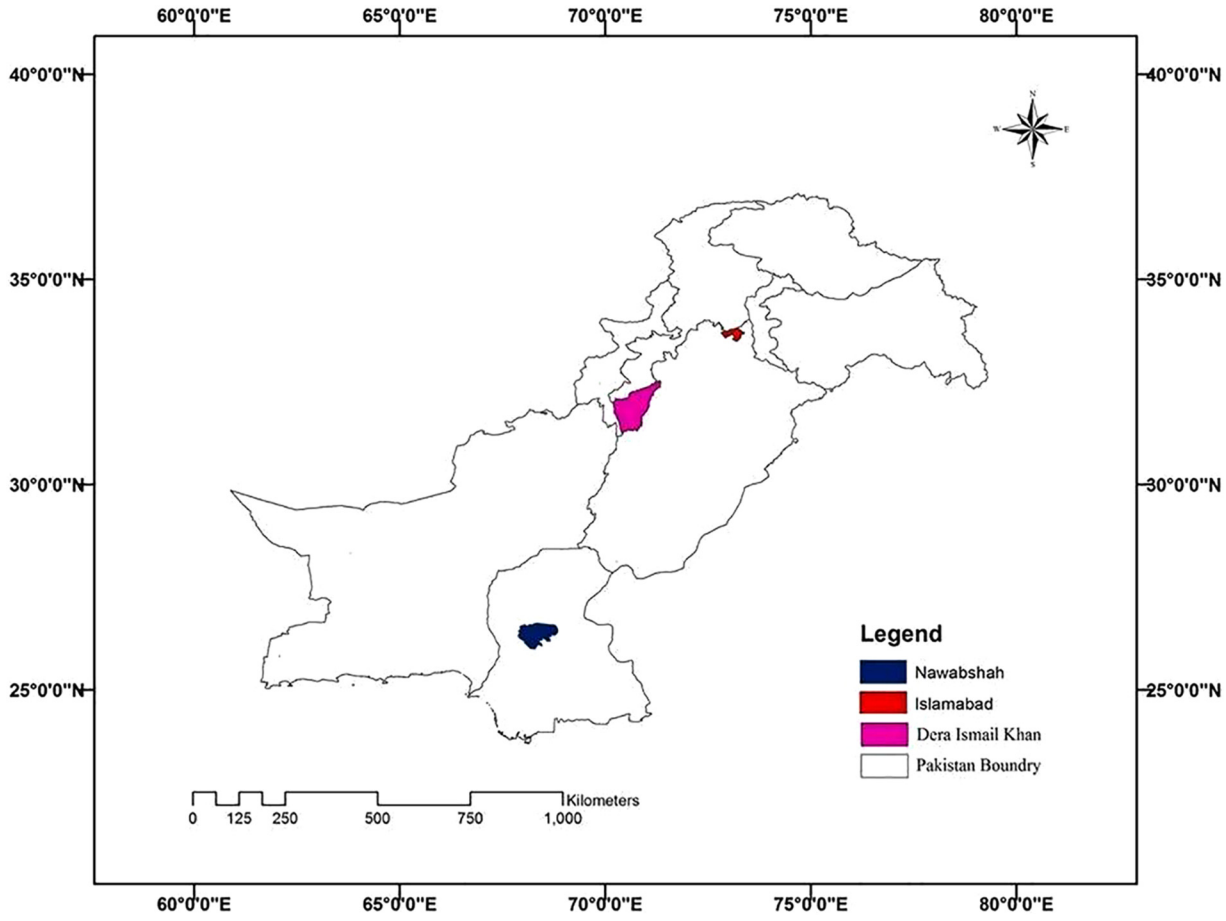


Fig. 3. Map of the study sites.

$$\rho_k = \frac{\frac{1}{n-k}\sum_{t=k+1}^{n}(x_t - \overline{x})(x_{t-k} - \overline{x})}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}(x_t - \overline{x})}\sqrt{\frac{1}{n-k}\sum_{t=k+1}^{n}(x_{t-k} - \overline{x})}} \quad (24)$$

where $k = 0, 1, 2, \cdots$ indicates the order of autocorrelation with observed seriesx_t for$t = 1, 2, 3, \cdots, n$ and $\overline{x}$ represents the mean SPI and the partial autocorrelation is defined as a regression series:

$$x_t = \Phi_{21}x_{t-1} + \Phi_{22}x_{t-1} + e_t \quad (25)$$

where x_t represents the original series minus the sample mean and the estimate of ϕ_{22} is expected to yield the value of the partial autocorrelation of the order 2. Extending the regression with k additional lags, the estimate of the last term is expected to give the partial autocorrelation of order$k$. A larger positive partial autocorrelation is better for extracting the features in developing an accurate forecasting model.

Fig. 4(a–c) displays ρ_k vs. significant time lags of the multi-scalar SPI for 3, 6 and 12 months presented at 90% to 95% confidence interval for statistically significant ρ_k of D. I. Khan, Islamabad and Nawabshah. At lag (t–1) with the shifting of input data by 1 unit timescale, the largest ρ_k attained was about 95% and this was accomplished for the partial autocorrelation of SPI_{12}; followed by about 90% for the SPI_6 and SPI_3 at

Table 1
Descriptive statistics of the study sites' geographic, hydrologic and drought characteristics.

Station	Geographic characteristics			Hydrological statistics (1981–2015): Rainfall (ppt)					
	Longitude	Latitude	Elevation (m)	Mean	Std.	Min	Max	Skewness	Kurtosis
D. I. Khan	70.91	31.83	175.00	29.46	38.59	0.30	376.00	3.30	18.49
Islamabad	73	33.74	604.00	107.21	130.70	0.01	743.30	2.03	4.49
Nawabshah	68.41	26.24	35.08	16.57	36.88	0.10	328.20	4.59	26.31

Drought statistics												
SPI_3					SPI_6				SPI_{12}			
	Mean	Std.	Min	Max	Mean	Std.	Min	Max	Mean	Std.	Min	Max
D. I. Khan	0.00	1.00	−2.10	3.13	1.00	0.00	−2.12	2.61	0.00	1.00	0.00	2.29
Islamabad	−0.01	1.00	−2.38	2.52	−0.01	1.00	−1.78	2.40	−0.01	1.00	−1.80	2.08
Nawabshah	−0.04	0.98	−1.88	2.86	−0.03	0.98	−1.88	2.65	−0.03	0.98	−2.04	1.94

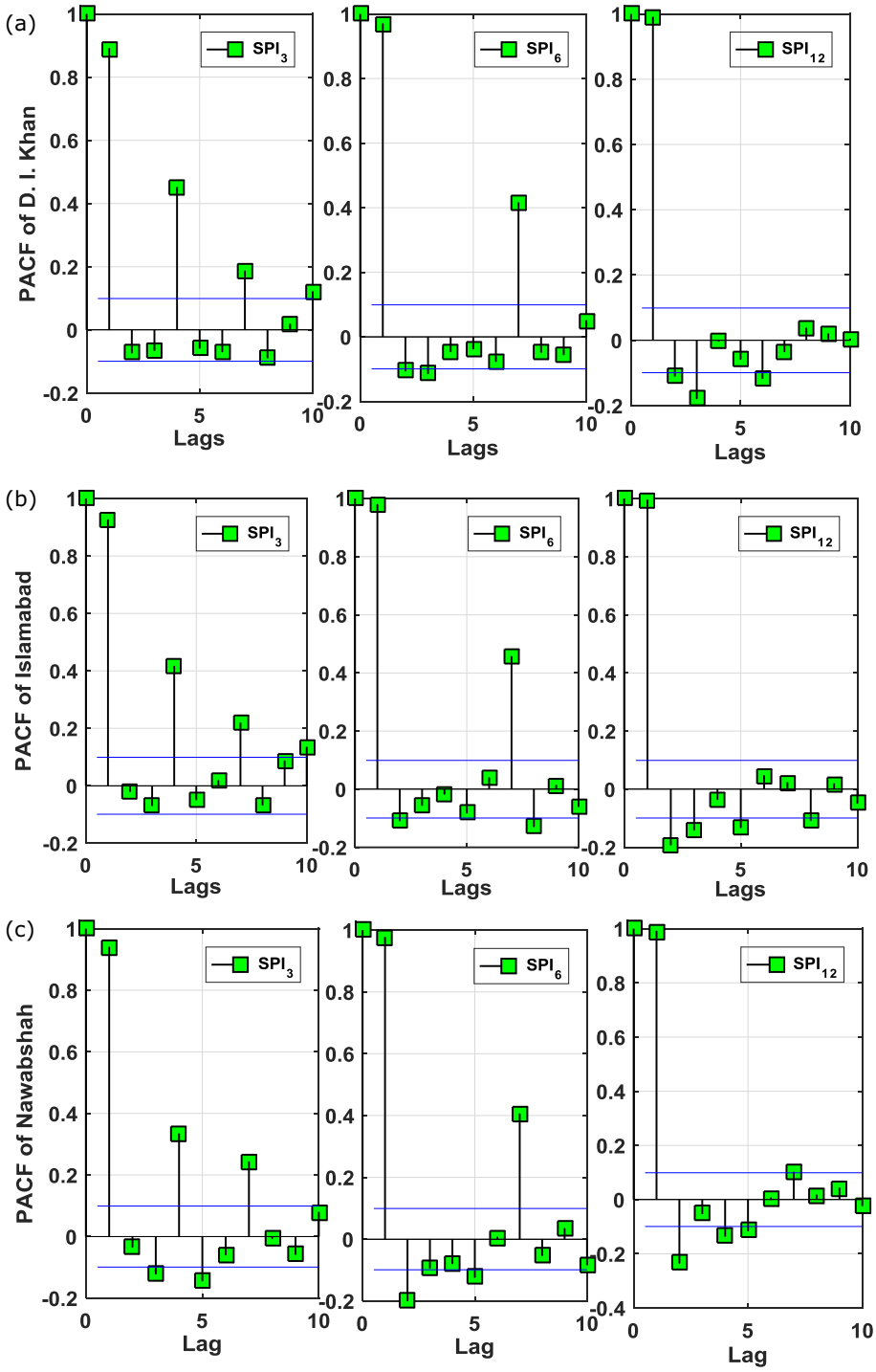


Fig. 4. Partial autocorrelation function (PACF) of historical SPI time-series based on 3, 6, and 12 month scale for stations: (a) D. I. Khan, (b) Islamabad, (c) Nawabshah. The blue line denotes the statistically significant boundary at the 95% confidence interval. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

about 85%. For Islamabad station, the partial autocorrelation was found to be high for the SPI_{12} and SPI_6 at about 95%, followed by the SPI_3 at about 90%. Similarly, the same pattern was attained for the station Nawabshah. This confirms the pivotal role of the predictor variables in the forecasted value of SPI. Moreover, the partial auto-correlation analysis aids in verifying that the chosen variables used to forecast SPI are appropriate, at least in the statistical sense, and they comprise of the inherent predictive features that could explain the multi-scalar evaluation of future SPI. Thus, the SPI at significant lag (t–1) for each tested station was used as the significant input to develop the new Ensemble-ANFIS and its counterpart models.

3.3. Development of drought model

Normalization of multi-scalar SPI input-target data was performed by statistical rules to overcome the numerical difficulties generated by the data attributes, patterns and fluctuations (Hsu et al., 2003). Mathematically, it is written as:

$$x_{SPI_{normalized}} = \frac{(x_{SPI} - x_{SPI_{\min}})}{(x_{SPI_{\max}} - x_{SPI_{\min}})} \quad (26)$$

where x_{SPI} describes any datum point of the SPI (input or output), $x_{SPI_{\min}}$ is the minimum value of the whole SPI dataset, $x_{SPI_{\max}}$ is the maximum value whereas $x_{SPI_{normalized}}$ denotes the normalized datum point.

In this paper multi-scalar SPI data were used to develop the forecasting Ensemble-ANFIS model, executed under the MATLAB environment with a Pentium 4, 2.93 GHz dual core Central Processing Unit. Prior to model design, PACF was used to find the significant lags for the purpose of obtaining the features and the optimization of the model. The inputs based on significant lags were selected randomly in Layer 1 to construct the ensembles based on the ANFIS model. The fuzzy membership functions were then allocated to this randomly selected data in Layer 2 which were then passed into Layer 3 for the assignment of the antecedent rules as per FIS framework. The weighted functions were used in Layer 4 to yield the normalization strengths to the antecedent fuzzy rule membership functions. The rule consequences were applied to them in Layer 5. These layers (2–5) were referred to as hidden layers, while the inference of the defuzzification process was carried out in Layer 6 (i.e., the output layer) to obtain the final ANFIS model output. The final output in Layer 7 was the mean of 10 different ANFIS model outputs (i.e., a 10-fold model development stage was adopted to reduce model stochasticity by a universally averaged set of output). The correlation coefficient '*r*', in combination with the mean squared error, *MSE* was applied to investigate the performance of the models in the training phase. The results generated by the Ensemble-ANFIS model have been summarised in Table 2.

r and *MSE* values attained in training of Ensemble-ANFIS model for SPI_3, SPI_6 and SPI_{12} forecasting at D. I. Khan were found to be: *r(0.888, 0.966, 0.988)* and *MSE(0.207, 0.065, 0.022)* respectively. Equivalent metrics for Islamabad are found to be: *r(0.923, 0.977, 0.994)* and *MSE (0.142, 0.043, 0.011)* and for Nawabshah are found to be: *r(0.936, 0.976, 0.989)* and *MSE(0.116, 0.044, 0.001)*. Overall, the training performance of the Ensemble-ANFIS model was considerably high for all of the study regions. It is thus envisaged that the Ensemble-ANFIS model testing performance, as seen later, will be relatively accurate for forecasting multi-scalar drought events at these sites. It is noteworthy, however, that the Ensemble-ANFIS model was able to attain a better training performance (in terms of both *r* and *MSE*) for the Islamabad station compared to D. I. Khan and Nawabshah but as a whole assessment, the accuracy was high at all 3 study sites (Table 2).

With M5 Tree algorithm, a tree-based forecast model was designed that operated on the 'divide-and-conquer' principle (Kisi, 2015; Rahimikhoob et al., 2013). For D. I. Khan station, *r* and *MSE* values attained in the training phase for forecasting multi-scalar SPI yielded: *r (0.882, 0.961, 0.989)* and *MSE(0.192, 0.062, 0.017)*; for Islamabad, the trained model yielded: *r(0.941, 0.979, 0.994)* and *MSE(0.099, 0.037, 0.009)* and for Nawabshah, the trained model yielded *r(0.947, 0.978, 0.992) and MSE(0.089, 0.037, 0.013)* for SPI_3, SPI_6, and SPI_{12} forecasts, respectively. This analysis showed that the trained M5 Tree model was able to forecast slightly better in terms of both a higher *r* and a lower *MSE* values for Islamabad compared to D. I. Khan and Nawabshah but as a whole assessment, the accuracy of M5 Tree was lower than the Ensemble-ANFIS model (Table 3).

Development of the MPMR model depended on a kernel function due to its reliance on the inner product of support vectors. In our case, a linear kernel equation was considered most appropriate. Moreover, the width and size of error threshold can also affect the performance of the MPMR model. The width of the error tube in this study was found to be 1.0 for the optimal model, whereas the kernel parameters, ker.p1 and ker.p2 were set to 2.0 and 23.0, respectively. At D. I. Khan station, the *r* and *MSE* values in the training period used to forecast multi-scalar SPI were: *r(0.859, 0.961, 0.986)* and *MSE(0.225, 0.062, 0.021)* while for Islamabad station, these were *r(0.926, 0.978, 0.994)* and *MSE(0.123, 0.038, 0.009)*. The correlation coefficient and mean squared error used to estimate SPI_3, SPI_6, and SPI_{12} at Nawabshah station were: *r(0.936, 0.977, 0.991),* and *MSE (0.106, 0.039, 0.013)*. In this respect, it is noticeable that the M5 Tree model forecasted slightly better in terms of both *r* and *MSE* for Islamabad compared to the other stations but as a whole, the accuracy of MPMR was high at all 3 stations. Table 3 illustrates these comparisons.

3.4. Model performance and their interpretation

The American Society for Civil Engineering (Yen, 1995) recommends two categories of model evaluation procedures that comprise of the statistical (or visual comparison of the observed and forecasted data) and the standardized performance metrics. The statistical metrics are used to investigate the differences between the minimum, maximum, mean, variance, standard deviation, skewness, and kurtosis factors while the standardized metrics are used for the validation of the predicted outcomes with respect to the observed data.

The mathematical formulations are as follows (Dawson et al., 2007; Deo et al., 2016c; Legates and McCabe, 1999; Willmott, 1981; Willmott, 1982; Willmott, 1984).

I. Mean square error (*MSE*) is expressed as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(SPI_{FOR,i} - SPI_{OBS,i})^2 \quad (27)$$

II. Correlation coefficient (r) is expressed as:

Table 2
Training performance of 10-member ('ensemble-ANFIS') model with correlation coefficient (*r*) and mean square error (*MSE*). Note: The inputs are based on statistically significant lagged data, SPI (*t* – 1).

D. I. Khan							Islamabad						Nawabshah					
SPI_3			SPI_6		SPI_{12}		SPI_3		SPI_6		SPI_{12}		SPI_3		SPI_6		SPI_{12}	
Ensemble No.	MSE	r	MSE	r	MSE	r	MSE	r	MSE	r	MSE	r	MSE	r	MSE	r	MSE	r
M_1	0.218	0.892	0.067	0.960	0.020	0.990	0.148	0.921	0.049	0.972	0.012	0.994	0.1088	0.934	0.038	0.978	0.022	0.988
M_2	0.201	0.887	0.061	0.969	0.024	0.987	0.145	0.918	0.041	0.980	0.010	0.995	0.1058	0.944	0.045	0.974	0.022	0.988
M_3	0.198	0.892	0.062	0.968	0.024	0.986	0.127	0.933	0.047	0.976	0.010	0.994	0.1166	0.937	0.046	0.973	0.017	0.990
M_4	0.213	0.885	0.066	0.968	0.025	0.987	0.128	0.929	0.045	0.978	0.010	0.994	0.1334	0.929	0.052	0.970	0.021	0.988
M_5	0.212	0.893	0.064	0.968	0.021	0.989	0.152	0.920	0.042	0.978	0.011	0.993	0.1182	0.936	0.036	0.982	0.022	0.988
M_6	0.238	0.867	0.070	0.966	0.018	0.990	0.155	0.918	0.045	0.978	0.012	0.993	0.1081	0.945	0.044	0.978	0.017	0.991
M_7	0.195	0.897	0.066	0.964	0.022	0.989	0.146	0.917	0.042	0.977	0.011	0.994	0.1200	0.934	0.048	0.973	0.018	0.990
M_8	0.199	0.891	0.062	0.966	0.019	0.990	0.142	0.917	0.039	0.979	0.012	0.993	0.1253	0.931	0.046	0.976	0.019	0.989
M_9	0.203	0.876	0.064	0.969	0.022	0.988	0.119	0.937	0.046	0.977	0.011	0.993	0.1338	0.926	0.041	0.978	0.017	0.990
M_{10}	0.190	0.902	0.064	0.965	0.025	0.986	0.157	0.921	0.038	0.978	0.012	0.993	0.0969	0.948	0.041	0.977	0.023	0.987
Average	0.207	0.888	0.065	0.966	0.022	0.988	0.142	0.923	0.043	0.977	0.011	0.994	0.1167	0.936	0.044	0.976	0.020	0.989

Table 3
Training performance of M5 Tree and MPMR models. Note: The inputs are based on statistically significant lagged data, SPI (t – 1).

Model	M5Tree						MPMR											
	SPI_3		SPI_6		SPI_{12}		Parameters					SPI_3		SPI_6		SPI_{12}	
Station	MSE	r	MSE	r	MSE	r	ϕ	scale	Kernel	Ker.p_1	Ker.p_2	MSE	r	MSE	r	MSE	r
D. I. Khan	0.192	0.882	0.062	0.961	0.015	0.989	1.0	none	linear	2	23	0.225	0.859	0.062	0.961	0.021	0.986
Islamabad	0.099	0.941	0.037	0.979	0.009	0.994	1.0	none	linear	2	23	0.123	0.926	0.038	0.978	0.009	0.994
Nawabshah	0.089	0.947	0.037	0.978	0.013	0.992	1.0	none	linear	2	23	0.106	0.936	0.039	0.977	0.013	0.991

$$r = \left(\frac{\sum_{i=1}^{N} (SPI_{OBS,i} - \overline{SPI}_{OBS,i})(SPI_{FOR,i} - \overline{SPI}_{FOR,i})}{\sqrt{\sum_{i=1}^{N} (SPI_{OBS,i} - \overline{SPI}_{OBS,i})^2} \sqrt{\sum_{i=1}^{N} (SPI_{FOR,i} - \overline{SPI}_{FOR,i})^2}} \right) \quad (28)$$

III. Willmott's index (WI) is expressed as:

$$WI = 1 - \left[\frac{\sum_{i=1}^{N} (SPI_{FOR,i} - SPI_{OBS,i})^2}{\sum_{i=1}^{N} (|SPI_{FOR,i} - \overline{SPI}_{OBS,i}| + |SPI_{OBS,i} - \overline{SPI}_{OBS,i}|)^2} \right], \quad 0 \le d \le 1 \quad (29)$$

IV. Nash-Sutcliffe coefficient (EV) is expressed as:

$$NS_E = 1 - \left[\frac{\sum_{i=1}^{N} (SPI_{OBS,i} - SPI_{FOR,i})^2}{\sum_{i=1}^{N} (\overline{SPI}_{OBS,i} - \overline{SPI}_{FOR,i})^2} \right] \quad (30)$$

V. Root mean square error (RMSE) is expressed as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (SPI_{FOR,i} - SPI_{OBS,i})^2} \quad (31)$$

VI. Mean absolute error (MAE) is expressed as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |(SPI_{FOR,i} - SPI_{OBS,i})| \quad (32)$$

VII. Legates-McCabe's (LM) is expressed as:

$$LM = 1 - \left[\frac{\sum_{i=1}^{N} |SPI_{FOR,i} - SPI_{OBS,i}|}{\sum_{i=1}^{N} |SPI_{OBS,i} - \overline{SPI}_{OBS,i}|} \right] \quad (33)$$

where SPI_{OBS} and SPI_{FOR} are the observed and forecasted i^{th} value of multi-scale SPI. $\overline{SPI}_{OBS}$ and $\overline{SPI}_{FOR}$ are the observed and forecasted mean of SPI in the (test) set and N is the number of tested data points.

Due to the standardization of the observed and forecasted means and variance, the robustness of r can be limited (Chai and Draxler, 2014). The "goodness-of-fit" relevant to high values are measured by *RMSE* while in contrast; *MAE* evaluates all deviations from observed data both in the same manner regardless of sign. The performance of a model can be reduced to partial peaks and higher magnitudes that can exhibit larger error and obtuse to small magnitudes (Dawson et al., 2007). Willmott's Index *(WI)* was introduced to counter this issue by considering the ratio of mean squared error instead of the differences (Mohammadi et al., 2015; Willmott, 1981; Willmott, 1982; Willmott, 1984; Willmott et al., 2012). Nash-Sutcliffe efficiency (NS_E) is another normalized metric that determines the relative magnitude of residual variance of forecasted data in comparison to the measured variance (Nash and Sutcliffe, 1970). Legates-McCabe's (*LM*) is a more advanced and powerful metric than both *WI* and NS_E which utilizes the adjustment of comparison in the evaluation of *WI* and NS_E. *LM* can be quite robust in evaluating the results by addressing the weaknesses of r and using *WI* and NS_E as baseline-adjusted indices together with an evaluation of *RMSE* and *MAE* (Legates and McCabe, 1999).

4. Results and discussion

In this section the results of multi-scalar SPI forecasts (i.e., 3-, 6- and 12-months) generated by the new ensemble-ANFIS model (with 10-fold validation procedure) are analysed. In order to construct the model, the importance of input variables (i.e., significant lags) was checked in terms of the forecast accuracy. Then, based on statistical criteria (Eqs. (27)–(33)), the Ensemble-ANFIS models were developed and compared with M5 Tree and MPMR models.

In Table 4(a–d) we show the accuracy of the 10 model ensembles of ANFIS in the testing phase constructed with significant lag (t–1) SPI as an input. Each table presents 10 ensemble-ANFIS with an average ensemble-ANFIS as a final (boldfaced) result. The magnitude of *RMSE, MAE, r, WI,* NS_E and *LM* between forecasted and observed multi-scalar SPIs are examined for each site with the respective input lag combinations to generate ten ensemble-ANFIS modelling scenarios (denoted as 'M_1–M_{10}') (Table 4) with average metrics representing the final ensemble-ANFIS model.

According to the results for Nawabshah station, the (average) ensemble-ANFIS model yielded a value of *RMSE (0.315), MAE (0.214), r (0.946), WI (0.927),* NS_E *(0.893), LM (0.721)* for SPI_3. For SPI_6, the Table shows *RMSE(0.223), MAE(0.141), r(0.974), WI(0.966),* NS_E*(0.948), LM(0.824)* and for SPI_{12}, it shows *RMSE(0.033), MAE (0.019), r(0.990), WI(0.987),* NS_E*(0.981), LM(0.902)* with significant lagged (t–1) inputs. It is noteworthy that models denoted as M_9 *(*SPI_3*),* M_4 *(*SPI_6*)* and M_8 *(*SPI_{12}*)* appear to be the best ensemble-ANFIS models for the Nawabshah study site (Table 4).

The magnitude of performance metrics for the study regions Islamabad and D. I. Khan can also be seen in Table 4 respectively. Moreover, for these sites, the models M_8 *(*SPI_3*),* M_7 *(*SPI_6*)* and M_{10} *(*SPI_{12}*)* seem to be the most accurate ensemble-ANFIS models for Islamabad (Table 4) whereas the models M_6 *(*SPI_3*),* M_1 *(*SPI_6*)* and M_{10} *(*SPI_{12}*)* are the most accurate ensemble-ANFIS models for D. I. Khan (Table 4). Overall, the statistical performance of the ensemble-ANFIS model for all three study sites are quite impressive, and those for Nawabshah station appears to be the best, followed by slightly lower performance for Islamabad and D. I. Khan stations, respectively.

Fig. 5(a–c) represents the uncertainties in multi-scalar SPI forecasts for the test period. Here, the ensemble-ANFIS model results attained from 10-fold simulation are shown in terms of the maximum and minimum absolute forecasting error (AFE) (i.e., indicated as error bars) and the average of all 10 model ensemble (i.e., shown in green) which is compared with the real (i.e., observed) SPI data for 3-, 6-, and 12-months. These error bars are practically useful, as they provide a clear confidence interval of worst and average simulations and the varying degrees of uncertainties at each of the test points for the ensemble member, and are more easily comparable using the length of the bars.

Table 4
Ensemble-ANFIS model evaluated in testing phase with Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Correlation Coefficient (*r*), Willmott Index (WI), Nash-Sutcliff (NS_E) and Legates & McCabes Index (LM) for 3, 6, and 12 month forecasts. The best model is boldfaced (blue).

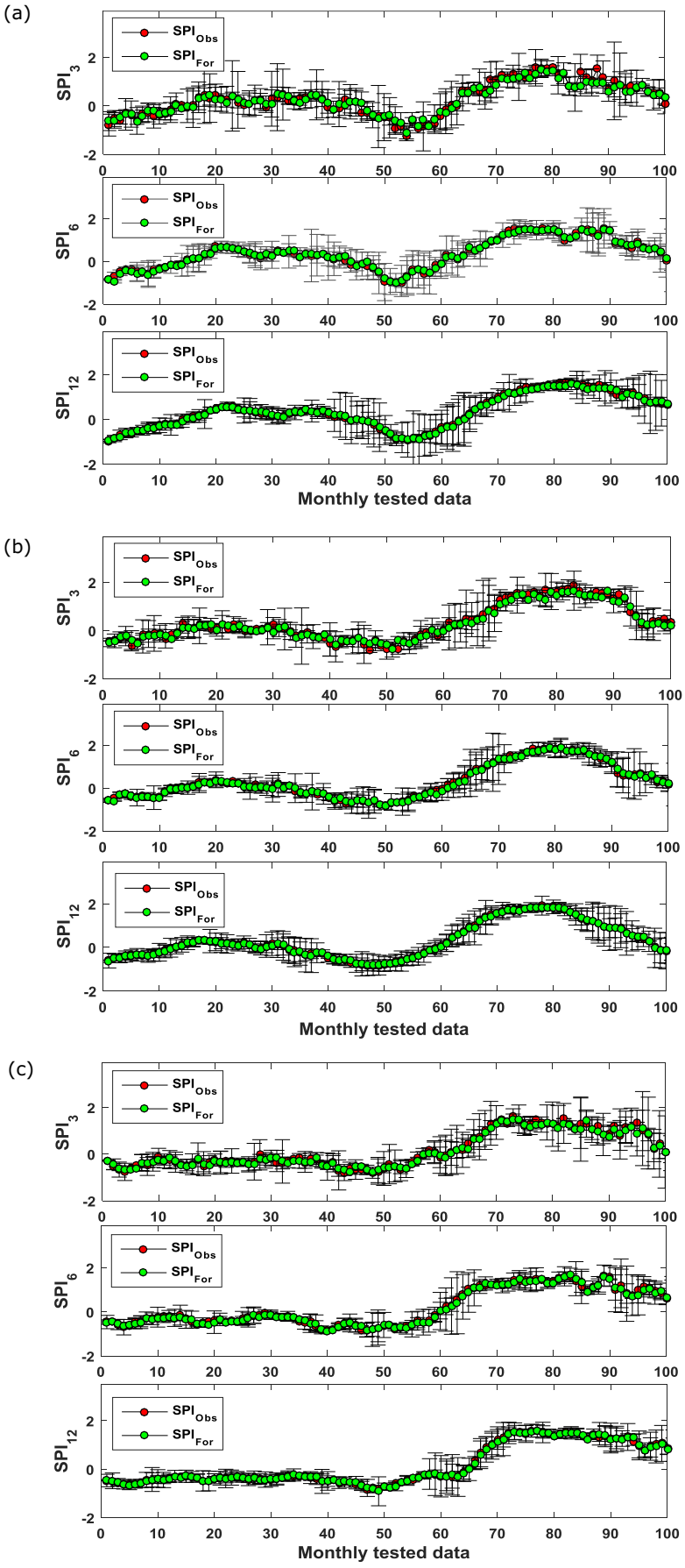
D. I. Khan																		
	SPI_3						SPI_6						SPI_{12}					
Ensemble No.	RMSE	MAE	r	WI	E_{NS}	LM	RMSE	MAE	r	WI	E_{NS}	LM	RMSE	MAE	r	WI	E_{NS}	LM
M_1	0.427	0.336	0.881	0.871	0.775	0.561	**0.241**	**0.175**	**0.977**	**0.9754**	**0.9534**	**0.816**	0.172	0.113	0.984	0.983	0.969	0.866
M_2	0.467	0.360	0.895	0.879	0.797	0.569	0.262	0.180	0.963	0.963	0.927	0.784	0.146	0.100	0.989	0.988	0.978	0.880
M_3	0.473	0.347	0.883	0.862	0.775	0.589	0.260	0.174	0.966	0.956	0.929	0.787	0.148	0.101	0.990	0.989	0.980	0.885
M_4	0.439	0.345	0.898	0.890	0.806	0.590	0.244	0.178	0.965	0.964	0.928	0.773	0.133	0.096	0.989	0.988	0.979	0.875
M_5	0.443	0.354	0.881	0.873	0.774	0.541	0.248	0.185	0.965	0.966	0.929	0.771	0.166	0.107	0.985	0.985	0.970	0.871
M_6	**0.377**	**0.294**	**0.929**	**0.919**	**0.863**	**0.656**	0.225	0.163	0.970	0.970	0.939	0.787	0.180	0.119	0.984	0.984	0.968	0.859
M_7	0.480	0.350	0.872	0.870	0.758	0.570	0.238	0.168	0.973	0.970	0.946	0.806	0.157	0.104	0.986	0.986	0.973	0.869
M_8	0.479	0.351	0.882	0.842	0.768	0.571	0.259	0.191	0.970	0.966	0.938	0.777	0.177	0.115	0.984	0.984	0.967	0.860
M_9	0.467	0.358	0.909	0.893	0.814	0.607	0.250	0.177	0.963	0.963	0.926	0.771	0.162	0.112	0.988	0.986	0.975	0.872
M_{10}	0.487	0.376	0.861	0.842	0.741	0.536	0.250	0.171	0.970	0.968	0.941	0.806	**0.138**	**0.097**	**0.991**	**0.990**	**0.982**	**0.889**
Average	0.454	0.347	0.889	0.874	0.787	0.579	0.248	0.176	0.968	0.966	0.935	0.788	0.158	0.107	0.987	0.986	0.974	0.873
Islamabad																		
M_1	0.372	0.285	0.929	0.900	0.861	0.640	0.184	0.146	0.985	0.982	0.970	0.835	0.103	0.075	0.993	0.992	0.987	0.897
M_2	0.382	0.297	0.933	0.911	0.870	0.657	0.223	0.170	0.971	0.963	0.942	0.766	0.123	0.090	0.991	0.989	0.982	0.873
M_3	0.428	0.324	0.904	0.857	0.816	0.584	0.195	0.143	0.980	0.972	0.960	0.815	0.122	0.091	0.992	0.990	0.985	0.889
M_4	0.424	0.315	0.914	0.874	0.831	0.613	0.207	0.151	0.975	0.970	0.951	0.798	0.126	0.093	0.992	0.990	0.984	0.888
M_5	0.360	0.277	0.929	0.913	0.863	0.632	0.219	0.164	0.976	0.970	0.951	0.793	0.115	0.084	0.994	0.992	0.988	0.902
M_6	0.356	0.279	0.931	0.916	0.866	0.651	0.207	0.157	0.974	0.968	0.948	0.780	0.105	0.082	0.994	0.993	0.989	0.902
M_7	0.381	0.287	0.934	0.909	0.869	0.667	**0.218**	**0.160**	**0.976**	**0.970**	**0.952**	**0.799**	0.118	0.086	0.991	0.988	0.983	0.881
M_8	**0.390**	**0.289**	**0.932**	**0.908**	**0.866**	**0.672**	0.232	0.171	0.974	0.968	0.948	0.792	0.104	0.075	0.995	0.993	0.989	0.910
M_9	0.445	0.335	0.897	0.858	0.802	0.583	0.196	0.146	0.977	0.971	0.955	0.795	0.115	0.083	0.994	0.992	0.988	0.904
M_{10}	0.349	0.262	0.930	0.899	0.860	0.631	0.238	0.175	0.975	0.968	0.950	0.799	**0.103**	**0.078**	**0.995**	**0.993**	**0.990**	**0.910**
Average	0.389	0.295	0.923	0.894	0.850	0.633	0.212	0.158	0.976	0.970	0.953	0.797	0.114	0.084	0.993	0.991	0.987	0.896
Nawabshah																		
M_1	0.360	0.240	0.948	0.921	0.886	0.733	0.250	0.163	0.970	0.961	0.941	0.810	0.121	0.078	0.991	0.988	0.983	0.897
M_2	0.347	0.234	0.933	0.913	0.869	0.698	0.220	0.137	0.979	0.971	0.957	0.842	0.121	0.072	0.992	0.990	0.985	0.911
M_3	0.315	0.217	0.946	0.935	0.894	0.720	0.213	0.133	0.979	0.973	0.958	0.842	0.159	0.096	0.987	0.984	0.975	0.888
M_4	0.263	0.182	0.961	0.949	0.924	0.761	**0.183**	**0.120**	**0.984**	**0.980**	**0.969**	**0.859**	0.128	0.079	0.991	0.988	0.983	0.901
M_5	0.309	0.210	0.946	0.923	0.896	0.708	0.255	0.164	0.961	0.948	0.923	0.780	0.123	0.073	0.991	0.989	0.983	0.903
M_6	0.341	0.229	0.926	0.905	0.856	0.664	0.224	0.139	0.970	0.960	0.939	0.808	0.158	0.090	0.985	0.980	0.970	0.882
M_7	0.305	0.209	0.953	0.932	0.905	0.732	0.201	0.128	0.980	0.975	0.960	0.844	0.152	0.083	0.987	0.982	0.974	0.893
M_8	0.284	0.182	0.956	0.941	0.915	0.763	0.212	0.134	0.974	0.964	0.948	0.819	0.139	0.083	0.989	0.986	0.978	0.890
M_9	**0.256**	**0.190**	**0.967**	**0.957**	**0.933**	**0.763**	0.238	0.149	0.970	0.960	0.940	0.808	0.158	0.088	0.989	0.987	0.977	0.903
M_{10}	0.374	0.250	0.923	0.899	0.852	0.670	0.236	0.144	0.972	0.966	0.944	0.824	**0.113**	**0.069**	**0.993**	**0.991**	**0.987**	**0.916**
Average	0.315	0.214	0.946	0.927	0.893	0.721	0.223	0.141	0.974	0.966	0.948	0.824	0.137	0.081	0.990	0.986	0.979	0.898

The red values represents the highest values of an ANFIS model out of 10 model.

Note that the maximum error indicated on this plot is the difference of the actual error between the highest or worst ensemble-ANFIS model member (i.e., as indicated by the performance metrics) and the average error in forecasted multi-scalar SPI. Likewise, the minimum error is the forecasting error between the average and the lowest or worst ensemble-ANFIS model member (i.e., as indicated by the performance metrics). For D. I. Khan, the error bars for the forecasted and observed SPI_{12} for each test point demonstrate a relatively small error value as compared to SPI_3 and SPI_6 forecasts (see Fig. 5(a)). This is then followed for Islamabad station (Fig. 5(b)) and the Nawabshah station (Fig. 5(d)) as the SPI_3 has larger errors which are reduced gradually in the SPI_6 and SPI_2 forecasts, respectively. Overall, the error bars of Nawabshah appear to be relatively small, which is then followed by Islamabad and then D. I. Khan station.

To compare directly the forecasted and observed multi-scalar SPI, Fig. 6(a-d) plots the 10-member averaged ensemble-ANFIS vs. M5 Tree and the MPMR model simulated absolute forecasted error for each tested month. Further, to analyse the model performance more closely, a scatterplot showing the goodness-of-fit and its correlation coefficient *r* is shown to depict the extent of agreement between forecasted and observed multi-scalar SPI. The ensemble-ANFIS model convincingly outperforms the M5 Tree and the MPMR model in all tested points. The absolute forecasted error is seen to exhibit a reasonably larger magnitude for M5 Tree and MPMR as compared to the ensemble-ANFIS model throughout the testing phase. Therefore, it is clear that the ensemble-ANFIS model has a better ability to simulate the multi-scalar SPI with good accuracy, as confirmed by the larger *r*-value.

In Table 5, the preciseness of the ensemble-ANFIS model in relation



(caption on next page)

Fig. 5. Ensemble-ANFIS model results attained from 10-fold simulations, analysed in terms of the maximum and minimum absolute forecasting error (AFE) (i.e., bars) and the average of 10 model ensemble forecasts (i.e., green) compared with observed SPI for 3, 6, and 12 month scale data for testing months. (a) D. I. Khan, (b) Islamabad, (c) Nawabshah. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

to the M5 Tree and MPMR models is evaluated where the results of SPI_3, SPI_6 and SPI_{12} for all three study sites in the testing phase are shown. For all sites considered, the accuracy of ensemble-ANFIS appears to have generally improved in the forecasts of SPI_6 and SPI_{12} data, as compared to those of SPI_3. This can be proved by the remarkable rise in the correlation coefficient computed between the observed and forecasted multi-scalar SPI and a corresponding decrease in the *RMSE* and *MAE* values. However, for the ensemble-ANFIS model, the improvement in forecasting accuracy declined with no further increase in value of *r*, or a reduction in the value of *RMSE* and *MAE* with an additional combination of significant lagged data after (*t*–1) (not shown here). This is plausibly due to the 80%–95% feature extraction evident from the lag (*t*–1) while the additional lagged historical SPI did not significantly contribute to an improvement in the final model. For the comparative analysis among the forecast models (i.e., ensemble-ANFIs vs. M5 Tree and MPMR) between the observed and the forecasted multi-scalar SPI data for all three study regions, we refer to Table 5. The ensemble-ANFIS model for the site Nawabshah appears to be the most accurate in terms of these assessment metrics computed between the observed and the forecasted multi-scalar SPI in contrast to those of MPMR and M5 Tree model.

Fig. 7(a-c) displays the frequency distribution of the ensemble-

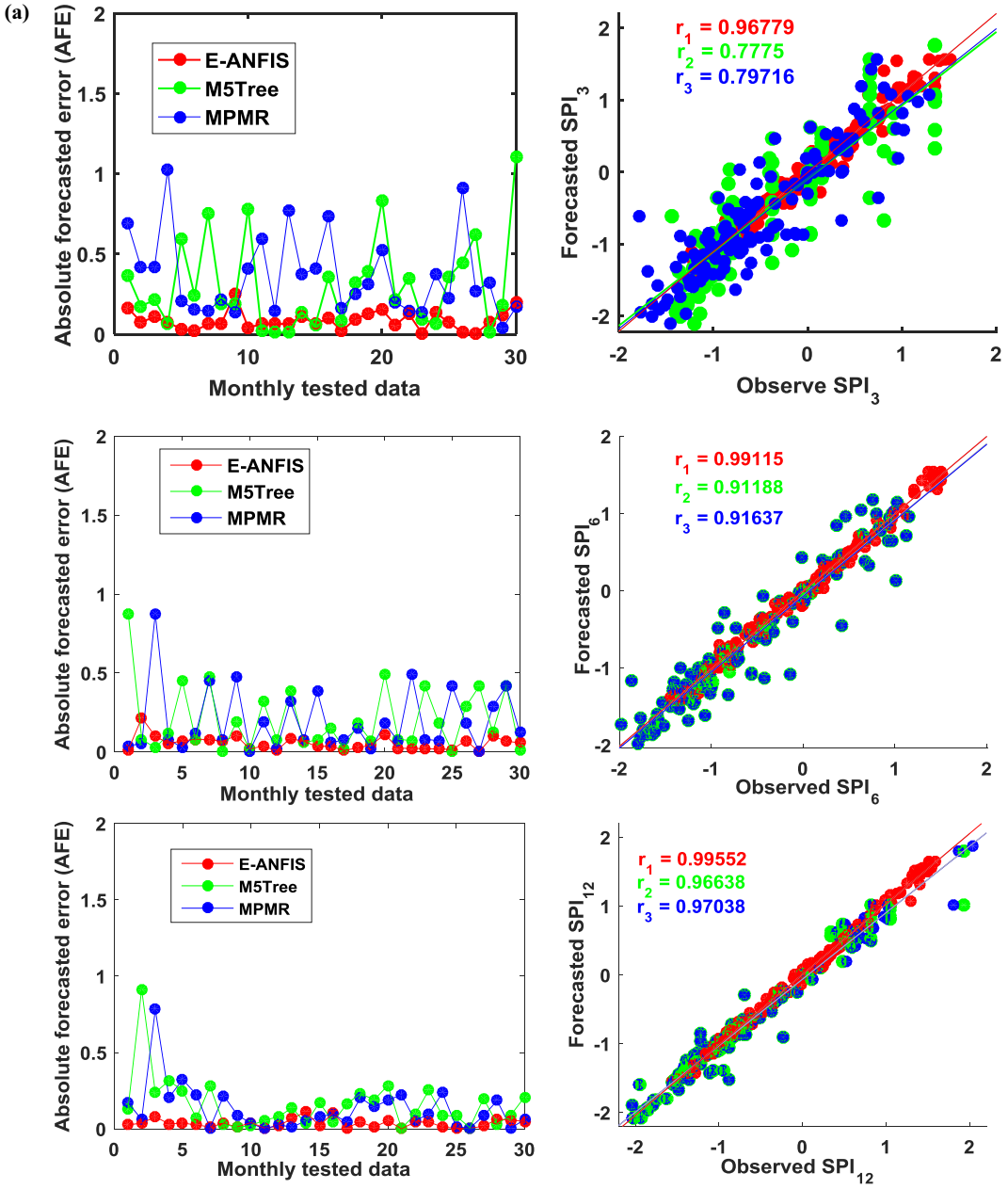


Fig. 6. (Left) Times-series of Ensemble-ANFIS vs. M5Tree and MPMR models-generated absolute forecasting error (AFE). (Right) Scatterplot of the forecasted and observed SPI in the testing phase based on 3, 6, and 12 month scales. (a) D. I. Khan, (b) Islamabad, (c) Nawabshah. For each scatterplot, the least square fitting line and its respective correlation coefficient is shown.

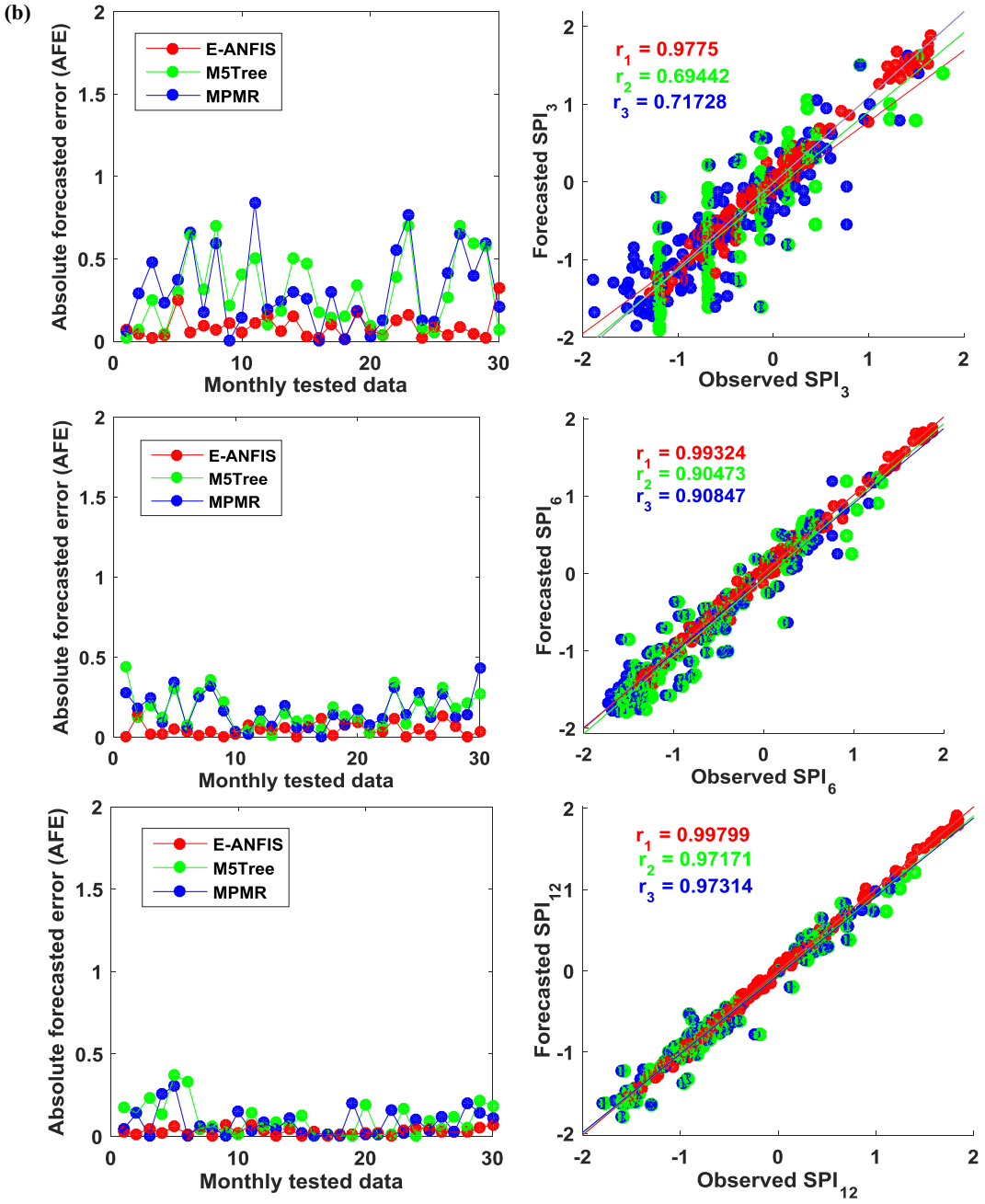


Fig. 6. (*continued*)

ANFIS vs. M5Tree and MPMR model's absolute forecasting error in the testing phase. Moreover, the percentage of each month in the testing period with an error level (± 0.12) has also been presented. The forecasting error frequency generated by the ensemble-ANFIS model in the forecasting of multi-scalar SPI, within the smallest error range (± 0.12)was observed to be the highest for D. I. Khan station (i.e., $SPI_3 \approx 56\%$, $SPI_6 \approx 79\%$, $SPI_{12} \approx 94\%$,), followed by Islamabad station ($SPI_3 \approx 52\%$, $SPI_6 \approx 89\%$, $SPI_{12} \approx 83\%$,) and Nawabshah station ($SPI_3 \approx 79\%$, $SPI_6 \approx 89\%$, $SPI_{12} \approx 86\%$,). This showed that the overall performance attained with the ensemble-ANFIS model was considerably better with a majority of the lower forecasting errors in the test phase (Fig. 7).

In Fig. 8(a–c), we illustrate a boxplot of the ensemble-ANFIS vs. M5Tree and MPMR model's forecasting error for multi-scale SPI of all the study sites. The outliers specified by + in every boxplot represent the extreme magnitudes of the forecasting error within the testing phase along with their upper quartile, median and lower quartile values. The distributed forecasting errors are justified by these boxplots showing a much lower spread was achieved by ensemble-ANFIS with a relative smaller magnitude of quartile statistics and median values in comparison with M5 Tree and the MPMR model. The net shift for the M5 Tree and MPMR models in the forecasting errors towards larger magnitudes are consistent with Fig. 7. Accordingly, the ensemble-ANFIS model remains the superior and highly optimized model for D. I. Khan, Islamabad and Nawabshah stations applied to forecast multi-scalar SPI data in terms of the illustrated clustered error distribution towards smaller magnitude.

Table 6 displays the computed values of duration (D), peak intensity (I), and severity (S) with their actual differences ($\Delta = SPI_{For} - SPI_{Obs}$) between D, I and S (10 sets of values plus a mean value) deduced from forecasted and observed datasets for all the three study site Dera Ismail Khan, Islamabad, and Nawabshah. The accuracy of the ensemble-ANFIS

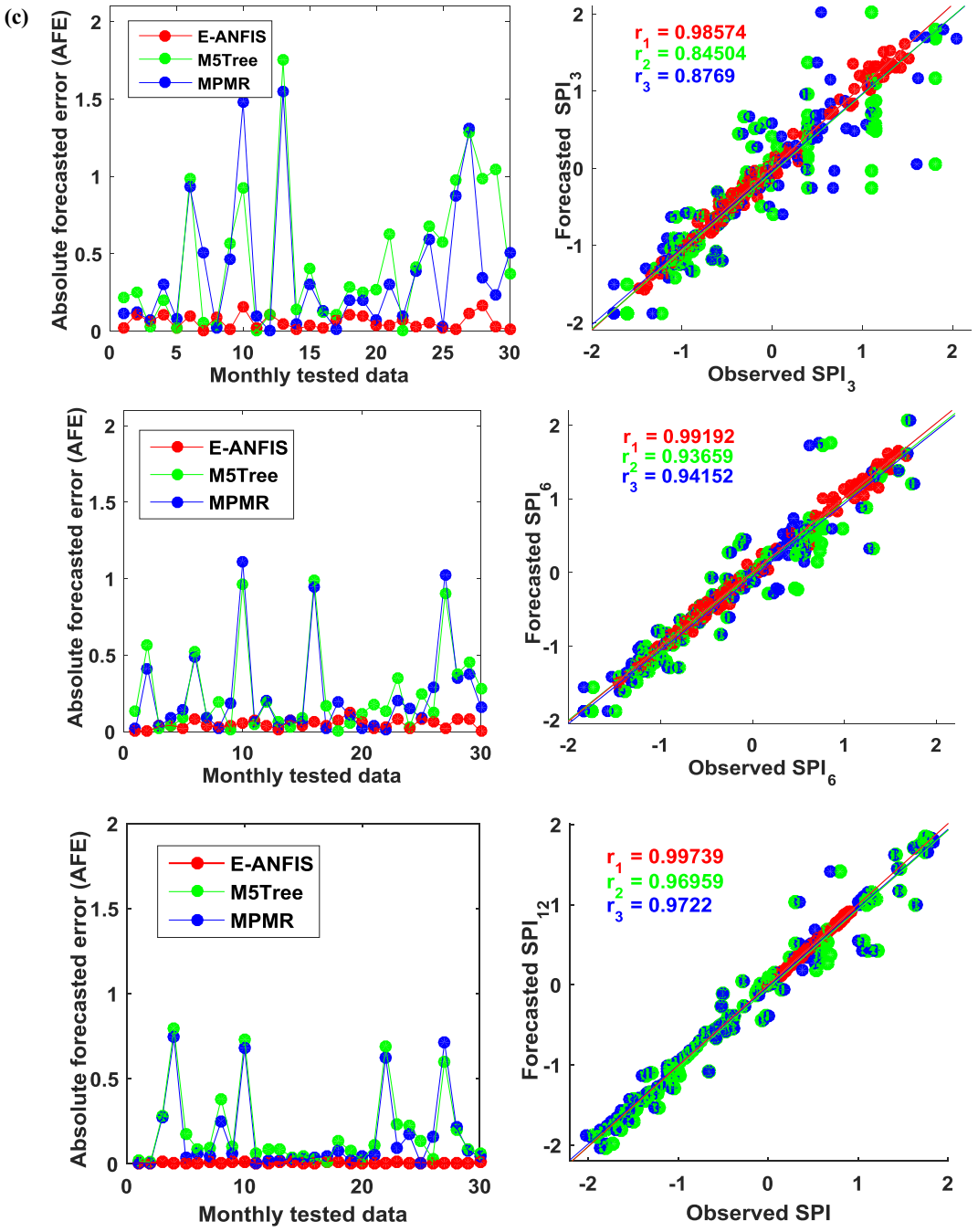


Fig. 6. (*continued*)

model in forecasting of drought properties was tested by quantifying the D, I and S using running-sum approaches (Kim et al., 2009; Yevjevich, 1967b). Based on the forecasted values of multi-scale SPI, a drought was seen to be occur in a month when the monthly value of the SPI was negative (i.e. rainfall conditions were lower rainfall than the normal period). The severity of the drought was then the accumulated value of the negative SPI and the duration as the sum of all months when this drought status was sustained. It shows the analysis of forecasted drought properties (in terms of the errors) of 10 model ensemble-ANFIS where the uncertainty in every model is shown. 10 model ANFIS can be used a decision tool for drought-risk assessment. When compared by S, there was a significant difference in observed and forecasted magnitudes of SPI_3 in comparison with SPI_6 and SPI_{12} for all stations (see. Table 6). However, the drought severity for Dera Ismail Khan and Islamabad was greater than Nawabshah in SPI_3, SPI_6 and SPI_{12} respectively. Similarly the drought intensity was higher for Islamabad and Dera Ismail Khan in regard to Nawabshah while the drought duration was found to be larger in Nawabshah for almost all 10 ensemble models as compared to other two stations.

The ability of the ensemble-ANFIS model has been extended to assess the predictive uncertainty in multi-scale SPI forecasting for all study regions in Table 7. In the forecasted 10 ensemble models the fluctuation (or variation) in the drought property is seen to occur in every simulation; hence this can be used to decide on a particular confidence interval of the simulated drought from 10 members, compared to the observed. The 10 ensemble models are first used to generate the range of simulations, and then we extract the properties of drought for each model. By doing so, a decision-support system can be developed to provide the stakeholders more strategic information in terms of the risk (i.e. how much can a simulation varies from the

Table 5
Comparison of 10-member average of the Ensemble-ANFIS vs. the M5 Tree and MPMR-based SPI forecasting performance. Note that the best model is boldfaced (blue).

D. I. Khan																		
	SPI_3						SPI_6						SPI_{12}					
Model	RMSE	MAE	r	WI	E^{NS}	LM	RMSE	MAE	r	WI	E^{NS}	LM	RMSE	MAE	r	WI	E^{NS}	LM
E-ANFIS	**0.454**	**0.347**	**0.889**	**0.874**	**0.787**	**0.579**	**0.248**	**0.176**	**0.968**	**0.966**	**0.935**	**0.788**	**0.158**	**0.107**	**0.987**	**0.986**	**0.974**	**0.873**
M5Tree	0.452	0.365	0.870	0.771	0.739	0.503	0.260	0.181	0.954	0.937	0.905	0.744	0.174	0.112	0.982	0.974	0.961	0.848
MPMR	0.393	0.306	0.891	0.822	0.779	0.557	0.258	0.179	0.956	0.940	0.910	0.753	0.166	0.106	0.984	0.977	0.966	0.859
Islamabad																		
E-ANFIS	**0.389**	**0.295**	**0.923**	**0.894**	**0.850**	**0.633**	**0.212**	**0.158**	**0.976**	**0.970**	**0.953**	**0.797**	**0.114**	**0.084**	**0.993**	**0.991**	**0.987**	**0.896**
M5Tree	0.465	0.356	0.831	0.700	0.668	0.458	0.246	0.184	0.950	0.927	0.896	0.709	0.139	0.096	0.985	0.979	0.968	0.855
MPMR	0.447	0.336	0.843	0.784	0.693	0.489	0.238	0.174	0.952	0.938	0.903	0.724	0.135	0.092	0.986	0.980	0.970	0.861
Nawabshah																		
E-ANFIS	**0.315**	**0.214**	**0.946**	**0.927**	**0.893**	**0.721**	**0.223**	**0.141**	**0.974**	**0.966**	**0.948**	**0.824**	**0.142**	**0.086**	**0.989**	**0.986**	**0.978**	**0.892**
M5Tree	0.409	0.273	0.916	0.883	0.836	0.655	0.264	0.179	0.967	0.954	0.933	0.784	0.191	0.116	0.984	0.980	0.967	0.871
MPMR	0.359	0.227	0.935	0.915	0.874	0.712	0.249	0.151	0.970	0.961	0.940	0.817	0.178	0.100	0.986	0.983	0.971	0.888

observed value). On the other hand, if a single simulation is used without the ensemble approach, which does not determine the range of uncertainty in drought properties which is a very important task for decision-making. Upper bound (UB) and lower bound (LB) of confidence intervals show the uncertainty involved in the drought duration predictions. It is clear from the Table 7 that the drought duration uncertainty values are slightly higher for SPI_3 as compared to SPI_6 and SPI_{12} in all stations. Geographically, the predictive uncertainty in monthly multi-scale SPI forecasting is higher for Nawabshah, followed by Dera Ismail Khan and Islamabad station.

5. Further discussion: limitations and future direction

To strategically address the foregoing challenges, researchers must design versatile predictive models for better forecasting of drought, particularly in nations that are sensitive to climate change (2016; Ahmad et al., 2004; Amir, 2012; Pakistan, 1950–2015; Zaidi, 2016). The 1998 drought in Pakistan was the worst in 50 years, and it was a primary factor responsible for poor economic growth (Haider, 2016). The Baluchistan province, especially the western and the central parts, often remain in the grip of drought almost all year round in Pakistan (Haider, 2016). Improved prediction of SPI is likely to assist such regions in preparing for expected changes in rainfall distribution. In earlier studies, data-driven models based on climatic parameters found the variability of models over large, sparsely distributed regions (Abbot and Marohasy, 2012; Abbot and Marohasy, 2014; Deo and Şahin, 2015a; Deo and Sahin, 2016; Deo et al., 2016b; Xie et al., 2013). The development of models to target specific localities will likely be necessary to improve drought prediction.

This study has designed for the first time an ensemble-ANFIS framework tested against the M5 Tree and MPMR model for multi-scalar SPI forecasting with significantly lagged inputs. It is worth mentioning that 10-member ensemble-ANFIS was evaluated to yield a large statistical correlation between observed and forecasted SPI based on the Legates-McCabe's Index (Table 4). The performance of the 10-member ensemble-ANFIS model was remarkable, compared to the M5 Tree and MPMR model in terms of the attained statistical accuracy. Therefore, the 10-member ensemble-ANFIS model was proven to be a valuable predictive tool for forecasting drought in the present work. Data-driven models are likely to become important tools in core decision-making in hydrology and water sciences that aim to address drought-related issues arising from imminent global warming scenarios and the increasing risk of water scarcity in the first world as well as developing nations.

In spite of the enormous accuracy, our study does have some limitations that could seed new, future research. ANFIS was employed via an ensemble technique, but in terms of the model's optimization, the hybridization of different data-intelligent 'add-in' algorithms could yield more promising results (Behmanesh et al., 2014; Liang et al., 2015; Vairappan et al., 2009; XingXing et al., 2008; Zhou et al., 2011). Some of the other advanced optimisation methods that enable more robust feature extraction applied could include: the Particle Swarm Optimization (PSO) (Chen and Yu, 2005), Quantum-Behaved Particle Swarm Optimization (Q-PSO) (Zhisheng, 2010), Genetic Algorithm (GA) (Reeves, 1995) and the Firefly Algorithm (FA) (Yang, 2010) technique that has been tested in climate applications as an improvement to the ANFIS model (Deo et al., 2018; Ghorbani et al., 2017; Raheli et al., 2017; Yaseen et al., 2017). In a following up work, extreme learning machine (Huang et al., 2006), artificial neural network (Abbot and Marohasy, 2014), support vector machine (Cortes and Vapnik, 1995) etc. can also be utilized to forecast drought events in Pakistan.

While the above-mentioned approaches are well-established, a new, more generalized framework for hybridizing the ANN with the intuitionistic fuzzy (Takeuti and Titani, 1984) and neutrosophic logic (Smarandache, 2001) that may also be achieved instead of using a fuzzy logic integrator. The later are known to handle uncertainty, indeterminacy, incompleteness and inconsistency in the predictor-target data. Since the standard statistical approaches tend to avoid the hurdle of model uncertainty that potentially leads to over-confident inferences and risky agricultural decisions, Bayesian Model Averaging (BMA) (Foresee and Hagan, 1997) is another data-intelligent tool to model uncertainties which can be used in ranking model performance. Multi-resolution analysis (e.g., empirical wavelet transform (Gilles, 2013), empirical mode composition (Huang et al., 1998), maximum overlap wavelet (Kormylo and Mendel, 1982) and singular value decomposition (Golub and Reinsch, 1970) could broaden the accuracy and scope of this study.

6. Conclusion

Machine learning-based simulation of drought behaviours utilizing the standardized rainfall deficits and surpluses can be used as a key task for developing effective drought mitigation strategies. Due to the severe impacts of drought hazard, SPI based drought forecasts are an acceptable tool utilized globally that can manage the associated future risk to agriculture, water management, demand, pricing and policy. In this

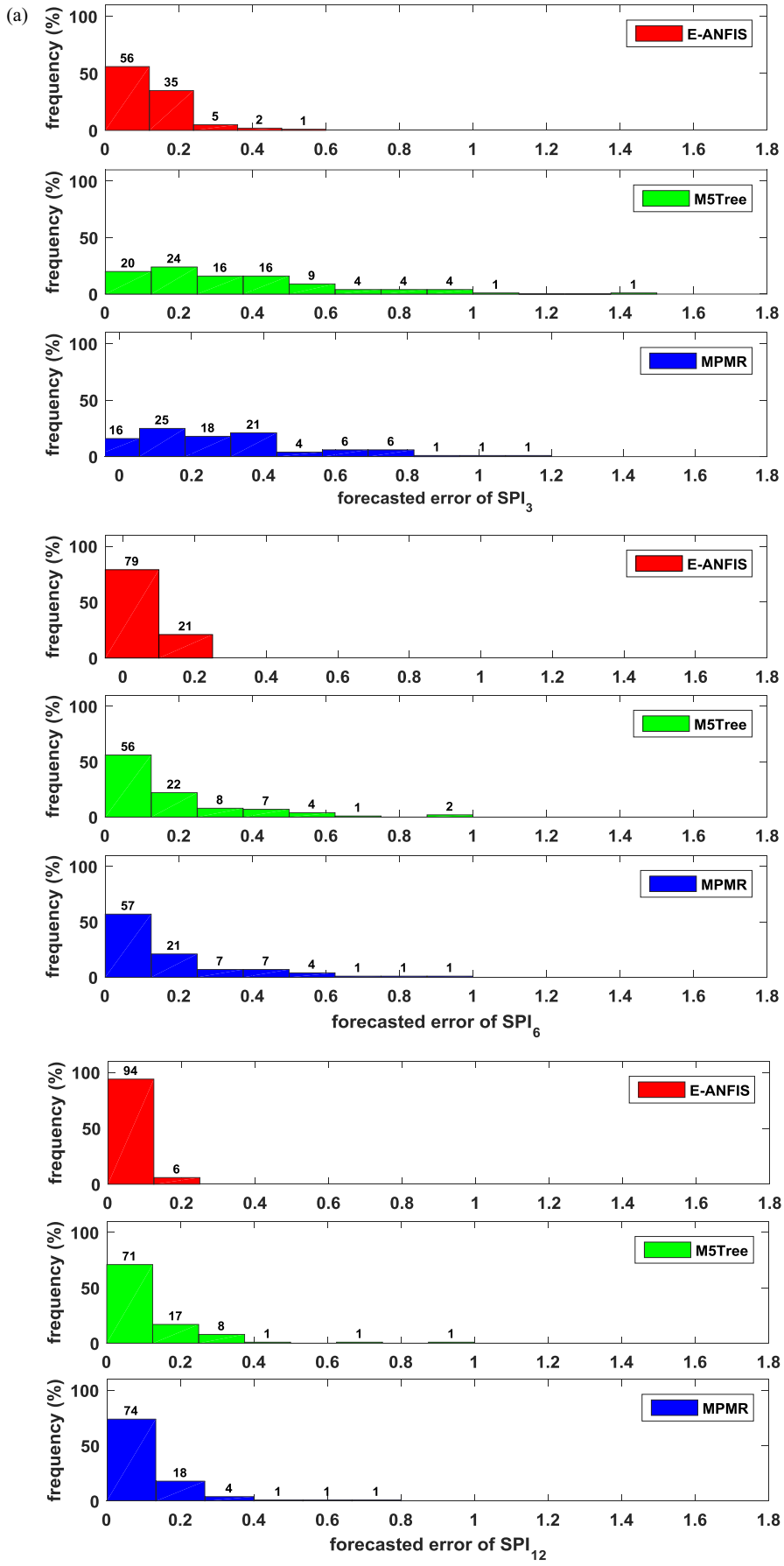


Fig. 7. Cumulative frequency of errors generated by the ensemble-ANFIS vs. M5Tree and MPMR model- based on the 3, 6 and 12 month SPI forecasts. (a) D. I. Khan, (b) Islamabad, (c) Nawabshah. Note that the percentage is shown in the respective error bracket.

(b)

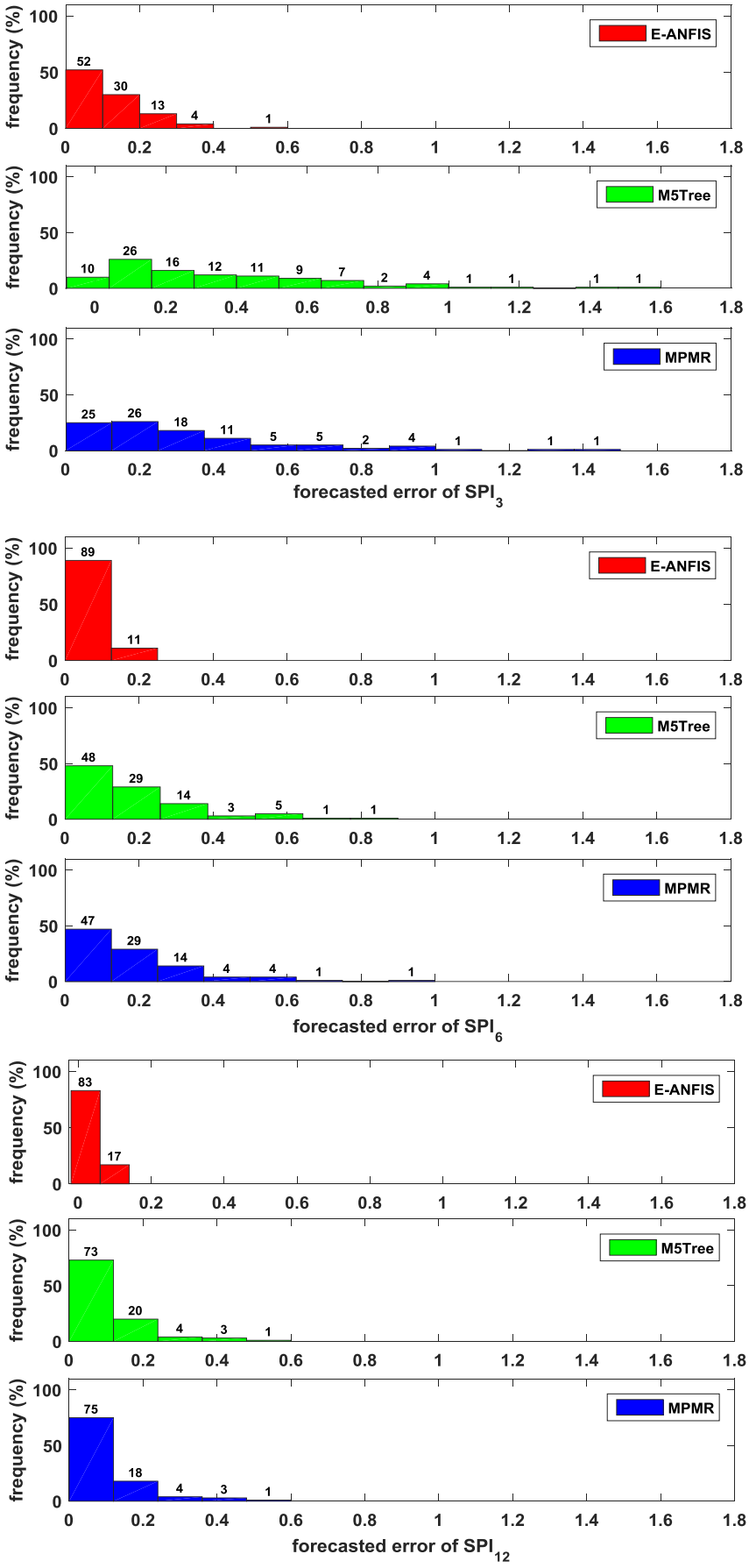


Fig. 7. (*continued*)

(c)

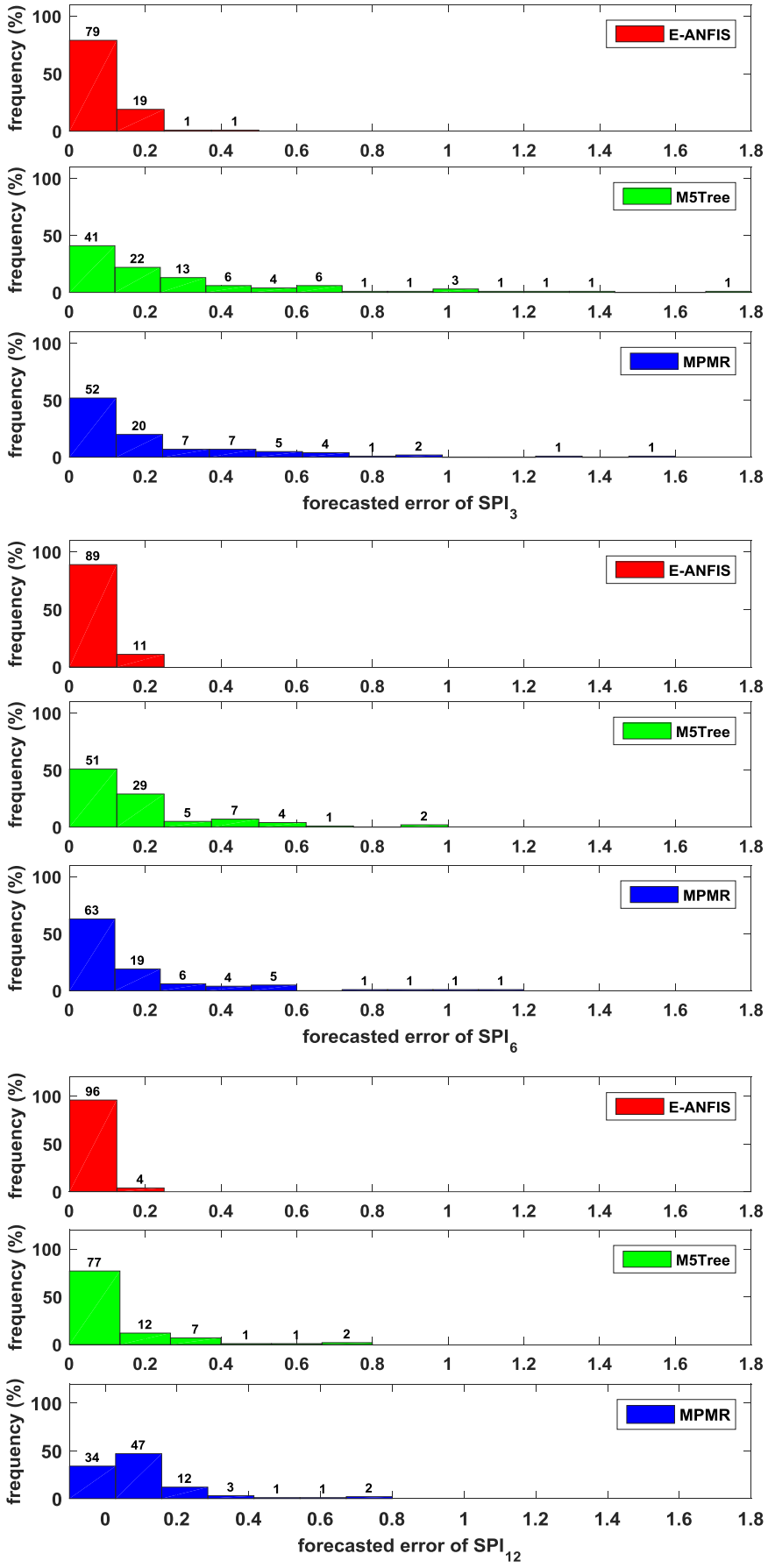


Fig. 7. (*continued*)

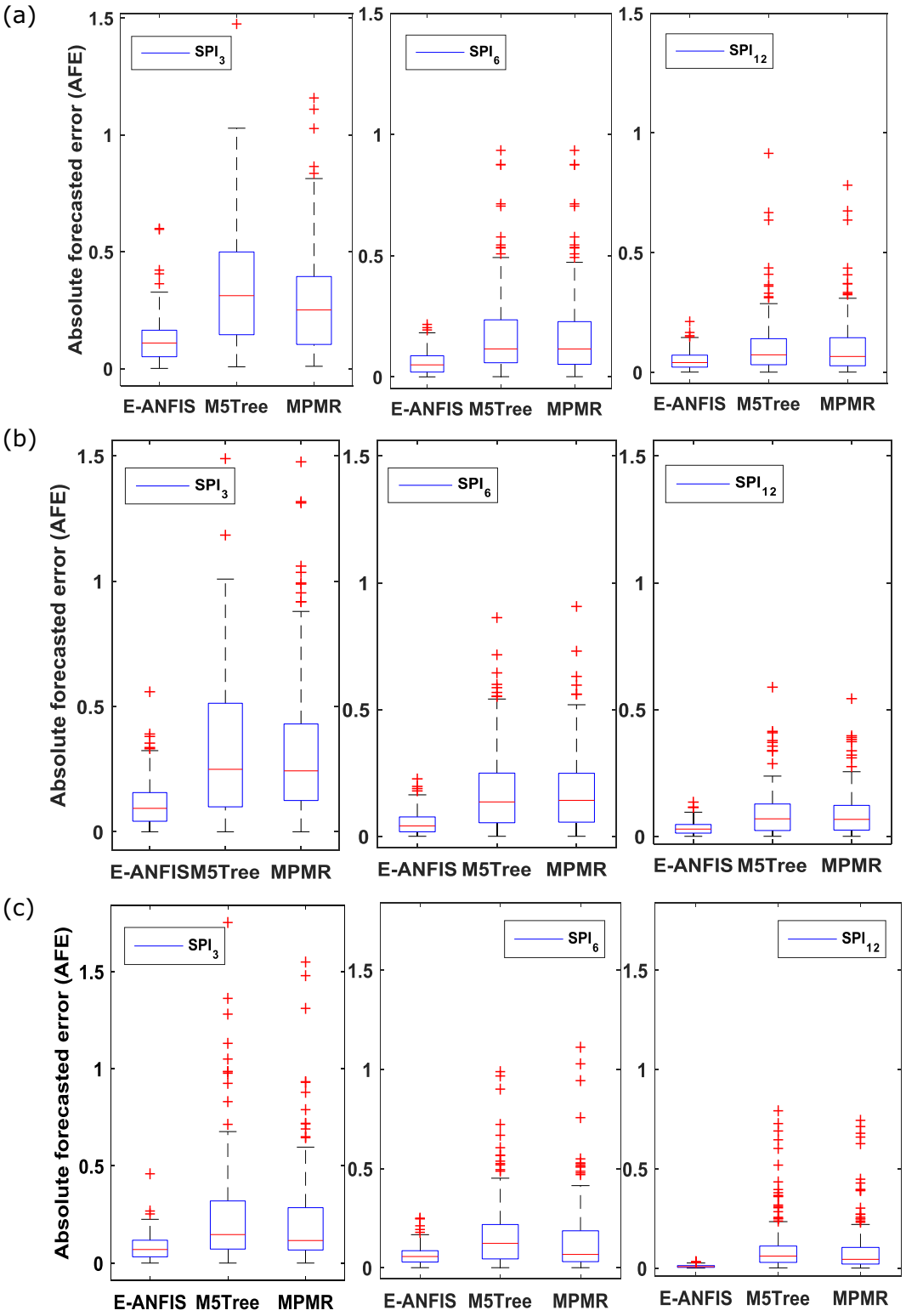


Fig. 8. Boxplot of the distribution of absolute forecasting error (AFE) generated by the ensemble-ANFIS vs. M5 Tree and the MPMR model 3, 6 and 12 month scale SPI forecasts. (a) D. I. Khan, (b) Islamabad, (c) Nawabshah.

study, the 10-member ensemble-ANFIS model was benchmarked with the M5 Tree and MPMR models for both short- and long-term SPI forecasting at three major locations in Pakistan where agriculture plays a significant role in the nation's food supply and economic growth. The models were developed using significant lags of SPI (i.e., historical behaviour of drought) in order to test the relative performance of the models for multi-scalar SPI forecasting. In terms of assessing the model accuracy, *MSE* and *r* (in the training phase), *RMSE, MAE, r, WI,* NS_E and *LM* (in the testing phase) were used between the forecasted and observed multi-scalar SPI for 3-, 6- and 12-month forecast horizons.

Our results revealed that using the SPI of the antecedent period to forecast in the consequent month plays a vital role in terms of forecasting accuracy of the data-intelligent models. Comparison of these models revealed that the 10-member ensemble-ANFIS, M5 Tree and

Table 6
The performance of the ensemble-ANIS model in all station for quantifying drought, as measured by the actual difference ($\Delta = SPI_{For} - SPI_{Obs}$) between the forecasted and the observed properties of drought over the testing period; Severity, S ≡ accumulated negative SPI after the onset of drought is detected (SPI < 0); Intensity, I ≡ minimum value of the SPI; Duration, D ≡ sum of the consecutive months in which drought status is sustained for all stations. Note that the average model is boldfaced (blue).

Dera Ismail Khan									
SPI_3									
	Drought Duration (D)			**Drought Intensity (I)**			**Drought Severity (S)**		
Models	SPI_{For}	SPI_{Obs}	Δ	SPI_{For}	SPI_{Obs}	Δ	SPI_{For}	SPI_{Obs}	Δ
M_1	89	86	3	-2.085	-1.912	-0.173	-54.769	-58.560	3.790
M_2	82	80	2	-2.046	-1.912	-0.134	-53.734	-60.234	6.500
M_3	79	83	4	-1.86	-2.385	0.525	-46.609	-55.689	9.080
M_4	85	79	6	-1.450	-2.385	0.935	-45.201	-51.378	6.176
M_5	80	74	6	-1.944	-2.061	0.116	-46.985	-47.729	0.743
M_6	70	71	1	-1.842	-2.385	0.543	-38.641	-44.897	6.256
M_7	73	75	2	-2.058	-1.912	-0.146	-45.874	-56.758	10.883
M_8	81	73	8	-1.591	-2.079	0.488	-47.693	-52.061	4.36
M_9	80	75	5	-1.685	-2.061	0.375	-45.402	-56.939	11.537
M_{10}	85	83	2	-1.402	-2.385	0.983	-50.499	-55.342	4.843
Average	**80.4**	**77.9**	**2.5**	**-1.796**	**-2.148**	**0.351**	**-47.541**	**-53.959**	**6.417**
SPI_6									
M_1	80	78	2	-1.693	-1.781	-0.088	-59.681	-59.763	-0.081
M_2	81	81	0	-1.722	-1.781	-0.059	-51.124	-51.303	-0.178
M_3	83	84	1	-1.713	-1.781	-0.068	-54.742	-57.933	-3.191
M_4	76	73	3	-1.695	-1.781	-0.086	-50.415	-50.051	0.364
M_5	80	80	0	-1.777	-1.732	0.044	-58.769	-57.394	1.374
M_6	76	76	0	-1.657	-1.778	-0.121	-46.511	-45.335	1.176
M_7	75	74	1	-1.584	-1.732	-0.148	-48.382	-48.199	0.182
M_8	74	76	2	-1.705	-1.752	-0.047	-52.211	-52.546	-0.334
M_9	88	86	2	-1.720	-1.669	0.051	-56.61	-55.859	0.750
M_{10}	79	79	0	-1.666	-1.752	-0.085	-57.465	-57.230	0.234
Average	**79.2**	**78.7**	**1.1**	**-1.693**	**-1.754**	**-0.060**	**-53.591**	**-53.561**	**0.029**
SPI_{12}									
M_1	79	78	1	-1.611	-1.635	-0.023	-51.934	-51.623	0.311
M_2	79	80	1	-1.778	-1.795	-0.017	-61.016	-62.480	-1.464
M_3	72	73	1	-1.586	-1.640	-0.053	-50.293	-51.169	-0.876
M_4	77	78	1	-1.772	-1.640	0.132	-60.806	-60.637	0.168
M_5	80	80	0	-1.783	-1.795	-0.011	-58.792	-59.360	-0.568
M_6	76	75	1	-1.783	-1.636	0.146	-59.847	-59.924	-0.077
M_7	79	81	2	-1.582	-1.795	-0.213	-49.001	-50.366	-1.365
M_8	78	79	1	-1.635	-1.640	-0.004	-61.792	-62.556	-0.764
M_9	77	77	0	-1.584	-1.795	-0.211	-55.376	-55.396	-0.019
M_{10}	80	79	1	-1.602	-1.795	-0.193	-55.967	-56.458	-0.491
Average	**77.7**	**78**	**0.9**	**-1.672**	**-1.717**	**-0.045**	**-56.482**	**-56.997**	**-0.514**

(*continued on next page*)

Table 6 (*continued*)

Islamabad									
SPI_3									
M_1	69	71	2	-1.443	-1.737	-0.294	-52.432	-55.772	-3.339
M_2	69	65	4	-1.806	-2.103	-0.296	-50.440	-56.048	-5.608
M_3	74	70	4	-1.902	-1.939	-0.037	-58.248	-66.626	-8.378
M_4	62	61	1	-1.763	-2.002	-0.239	-46.926	-52.938	-6.012
M_5	70	66	4	-1.447	-1.972	-0.525	-49.508	-50.404	-0.895
M_6	72	68	4	-1.844	-1.972	-0.128	-54.404	-58.782	-4.377
M_7	63	62	1	-1.987	-2.103	-0.116	-49.349	-50.633	-1.284
M_8	72	82	10	-1.837	-2.002	-0.164	-55.729	-67.2971	-11.567
M_9	61	63	2	-1.752	-2.103	-0.350	-43.57	-50.331	-6.761
M_{10}	65	71	6	-1.542	-1.767	-0.224	-48.371	-57.031	-8.6606
Average	**67.7**	**67.9**	**3.8**	**-1.732**	**-1.970**	**-0.237**	**-50.898**	**-56.586**	**-5.688**
SPI_6									
M_1	70	69	1	-2.096	-2.088	0.007	-65.068	-63.709	1.359
M_2	64	66	2	-1.925	-1.835	0.090	-54.697	-55.526	-0.828
M_3	70	73	3	-1.739	-1.866	-0.126	-50.533	-58.045	-7.512
M_4	63	65	2	-2.026	-1.867	0.159	-49.412	-50.229	-0.816
M_5	65	65	0	-1.927	-1.835	0.091	-58.561	-58.053	0.507
M_6	63	60	3	-1.976	-1.835	0.140	-47.414	-44.442	2.971
M_7	66	69	3	-2.012	-2.088	-0.075	-52.348	-54.835	-2.486
M_8	61	69	8	-1.995	-2.088	-0.093	-51.981	-54.761	-2.780
M_9	75	72	3	-2.018	-2.088	-0.069	-60.148	-58.795	1.357
M_{10}	68	71	3	-1.95	-1.972	-0.022	-57.773	-58.396	-0.622
Average	**66.5**	**67.9**	**2.8**	**-1.9667**	**-1.956**	**0.010**	**-54.793**	**-55.678**	**-0.885**
SPI_{12}									
M_1	64	65	1	-1.953	-2.073	-0.119	-52.803	-53.870	-1.067
M_2	71	72	1	-2.027	-2.076	-0.045	-60.739	-62.236	-1.496
M_3	69	71	2	-2.045	-2.076	-0.030	-62.476	-62.201	0.275
M_4	67	66	1	-2.015	-2.076	-0.061	-54.734	-55.656	-0.922
M_5	64	62	2	-2.039	-2.076	-0.037	-52.032	-52.277	-0.245
M_6	62	61	1	-2.054	-2.074	-0.020	-52.642	-52.089	0.552
M_7	65	65	0	-2.038	-2.074	-0.036	-50.357	-50.534	-0.177
M_8	64	65	1	-2.065	-2.076	-0.011	-60.285	-59.392	0.892
M_9	64	64	0	-1.936	-1.988	-0.051	-54.563	-57.813	-3.250
M_{10}	66	68	2	-2.013	-2.076	-0.063	-55.702	-56.251	-0.549
Average	**65.6**	**65.9**	**1.1**	**-2.018**	**-2.067**	**-0.048**	**-55.633**	**-56.232**	**-0.598**

(*continued on next page*)

Table 6 (*continued*)

Nawabshah									
SPI$_3$									
M$_1$	85	85	0	-1.770	-1.889	0.092	-51.842	-59.104	7.262
M$_2$	84	85	1	-1.797	-1.889	0.147	-56.499	-59.698	3.199
M$_3$	82	79	3	-1.741	-1.889	0.149	-49.420	-49.764	0.343
M$_4$	89	90	1	-1.739	-1.889	0.135	-55.285	-57.678	2.392
M$_5$	99	97	2	-1.754	-1.889	0.152	-63.464	-64.957	1.492
M$_6$	93	92	1	-1.736	-1.889	0.136	-57.802	-59.155	1.353
M$_7$	94	92	2	-1.753	-1.889	0.121	-62.330	-61.710	-0.62
M$_8$	94	95	1	-1.767	-1.889	0.178	-67.39	-68.681	1.290
M$_9$	86	81	5	-1.710	-1.889	0.148	-54.291	-54.92	0.636
M$_{10}$	89	84	5	-1.740	-1.889	0.148	-53.712	-55.350	1.637
Average	**89.5**	**88**	**2.1**	**-1.751**	**-1.889**	**0.141**	**-57.204**	**-59.102**	**1.8988**
SPI$_6$									
M$_1$	84	80	4	-1.799	-1.889	0.089	-53.780	-53.005	-0.775
M$_2$	87	88	1	-1.818	-1.889	0.070	-57.890	-58.713	0.822
M$_3$	84	82	2	-1.792	-1.889	0.096	-51.616	-53.163	1.546
M$_4$	85	79	6	-1.793	-1.889	0.096	-53.071	-52.169	-0.902
M$_5$	89	88	1	-1.809	-1.889	0.080	-56.092	-56.245	0.153
M$_6$	87	84	3	-1.814	-1.889	0.075	-52.619	-52.91	0.298
M$_7$	88	83	5	-1.836	-1.889	0.052	-59.959	-59.903	-0.056
M$_8$	91	88	3	-1.802	-1.889	0.087	-56.063	-56.733	0.670
M$_9$	87	82	5	-1.813	-1.889	0.076	-54.043	-53.180	-0.863
M$_{10}$	83	83	0	-1.831	-1.889	0.058	-50.874	-51.521	0.646
Average	**86.5**	**83.7**	**3**	**-1.811**	**-1.889**	**0.078**	**-54.601**	**-54.755**	**0.1540**
SPI$_{12}$									
M$_1$	94	94	0	-1.774	-1.986	0.211	-61.211	-60.999	-0.211
M$_2$	87	85	2	-1.836	-1.813	-0.022	-63.156	-62.129	-1.026
M$_3$	91	91	0	-1.941	-2.038	0.096	-57.427	-57.929	0.501
M$_4$	92	93	1	-1.945	-1.889	-0.056	-62.577	-63.483	0.906
M$_5$	92	93	1	-1.944	-1.986	0.0416	-58.909	-59.569	0.659
M$_6$	85	83	2	-1.559	-1.564	0.005	-48.092	-47.550	-0.542
M$_7$	91	89	2	-1.804	-1.763	-0.040	-51.893	-51.844	-0.048
M$_8$	91	92	1	-1.854	-2.038	0.183	-59.241	-59.520	0.278
M$_9$	79	70	9	-1.894	-1.986	0.091	-56.006	-56.876	0.870
M$_{10}$	89	89	0	-1.99	-1.885	-0.107	-58.414	-58.562	0.1481
Average	**89.1**	**87.9**	**1.8**	**-1.854**	**-1.895**	**0.040**	**-57.692**	**-57.846**	**0.1534**

MPMR models can all be successfully adopted for multi-scalar SPI forecasting, although there was a significant variation (in terms of model forecasting accuracy) (i.e., between the performance of the ensemble-ANFIS versus M5 Tree and MPMR models). The ensemble-ANFIS model achieved the highest accuracy for D. I. Khan, Islamabad and Nawabshah stations, which was confirmed in both the training (Table 2) and the testing phase of the model (Table 4). That is, in the case of forecasting moderate drought, severe drought and extreme drought, the ensemble-ANFIS performance was quite impressive for D. I. Khan, Islamabad and Nawabshah due to the model exhibiting the

Table 7
Ninety five percentage of confidence band using ensemble-ANFIS model for low, and upper forecasted multi-scale drought duration property for the stations (a) Dera Ismail Khan, (b) Islamabad, and (c) Nawabshah. Note than Per = Percentile, $\Delta = SPI_{For} - SPI_{Obs}$, and FDP (m ± e) = Forecasted Duration Property (mean ± error).

D. I. Khan						Islamabad				Nawabshah				
Confidence bound	Per	SPI_{FOR}	SPI_{Obs}	Δ	FDP(m ± e)	SPI_{FOR}	SPI_{Obs}	Δ	FDP(m ± e)	SPI_{FOR}	SPI_{Obs}	Δ	FDP(m ± e)	
Upper bound	P_{95}	87.2	84.6	2.6	80.4 ± 2.6	73.1	77.0	3.9	67.7 ± 3.9	96.7	96.1	0.6	89.5 ± 0.6	SPI_3
Lower bound	P_{25}	79.2	74.2	5.0	80.4 ± 5.0	63.5	63.5	0.0	67.7 ± 0.0	85.2	84.2	1.0	89.5 ± 1.0	SPI_3
Upper bound	P_{95}	85.7	85.1	0.6	79.2 ± 0.6	72.7	72.5	0.2	66.5 ± 0.2	90.1	88.0	2.1	86.5 ± 2.1	SPI_6
Lower bound	P_{25}	76.0	76.0	0.0	79.2 ± 0.0	63.2	65.2	2.0	66.5 ± 2.0	84.2	82.0	2.2	86.5 ± 2.2	SPI_6
Upper bound	P_{95}	80.0	80.5	0.5	79.2 ± 0.5	70.1	71.5	1.4	65.6 ± 1.4	93.1	93.5	0.4	89.1 ± 0.4	SPI_{12}
Lower bound	P_{25}	77.0	77.2	0.2	79.2 ± 0.2	64.0	64.2	0.2	65.6 ± 0.2	87.5	86.0	1.5	89.1 ± 1.5	SPI_{12}

lowest value of *RMSE* and *MAE* and highest magnitudes of *r, WI,* NS_E and *LM*.

This research has set the foundation for the potential of using more extensive predictor data products such as climate predictors (i.e., rainfall, temperature, humidity etc.), climate mode indices (SOI, PDO, EMI etc.) and satellite data with reanalysis and ground-based products for forecasting future drought in Pakistan and elsewhere. The technique applied is practically beneficial for multi-period drought analysis in respect to the forecasting variable. While the present study presents only a case study in Pakistan, the developed framework can be extended to any other location in the world. For example, to increase its practicality in future drought forecasting, the ensemble-ANFIS model could be applied to other agricultural and water resource reservoir zones where water scarcity threatens long term sustainability. Due to the aforementioned potential and abilities of the ensemble-ANFIS, it is possible to apply such a model for the forecasting of stream and river flow, wind forecasting, soil moisture forecasting and crop yield estimation in Pakistan. Another possible extension to the present study may employ several different types of climate data (temperature, humidity, sunshine, rainfall) with climate indices for drought forecasting in different regions as each of these variables are likely to impact drought severity and duration.

Acknowledgment

This study was supported by the University Of Southern Queensland Office Of Graduate Studies Postgraduate Research Scholarship (2017–2019) (USQPRS 2017). Data were acquired from the Pakistan Meteorological Department, Pakistan, which are duly acknowledged.

References

Abbot, J., Marohasy, J., 2012. Application of artificial neural networks to rainfall forecasting in Queensland, Australia. Adv. Atmos. Sci. 29 (4), 717–730.
Abbot, J., Marohasy, J., 2014. Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. Atmos. Res. 138, 166–178.
Abrahart, R.J., 2003. Neural network rainfall-runoff forecasting based on continuous resampling. J. Hydroinf. 5 (1), 51–61.
Adamowski, J., Fung Chan, H., Prasher, S.O., Ozga-Zielinski, B., Sliusarieva, A., 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resour. Res. 48 (1).
Ahmad, S., Hussain, Z., Qureshi, A.S., Majeed, R., Saleem, M., 2004. Drought Mitigation in Pakistan: Current Status and Options for Future Strategies, 85. (IWMI).
Ahmed, K., Shahid, S., bin Harun, S., Wang, X.-j., 2016. Characterization of seasonal droughts in Balochistan Province, Pakistan. Stoch. Env. Res. Risk A. 30 (2), 747–762.
Ali, Z., et al., 2017. Forecasting drought using multilayer perceptron artificial neural network model. Adv. Meteorol. 2017.
Almedeij, J., 2016. Long-term periodic drought modeling. Stoch. Env. Res. Risk A. 30 (3), 901–910.
Amir, I., 2012. Tough Times Ahead for Farmers in Southern KP. (DAWN).
Bachman, P., Alsharif, O., Precup, D., 2014. Learning with pseudo-ensembles. Adv. Neural Inf. Proces. Syst. 3365–3373.
Bates, B., Kundzewicz, Z.W., Wu, S., Palutikof, J., 2008. Climate Change and Water: Technical Paper vi. Intergovernmental Panel on Climate Change (IPCC).
Behmanesh, M., Mohammadi, M., Naeini, V.S., 2014. Chaotic time series prediction using improved ANFIS with imperialist competitive learning algorithm. Int. J. Soft Comput. Eng. 4 (4), 25–33.
Bertsimas, D., Sethuraman, J., 2000. Moment problems and semidefinite optimization. In: Handbook of Semidefinite Programming, pp. 469–509.
Bhattacharya, B., Solomatine, D.P., 2005. Neural networks and M5 model trees in modelling water level–discharge relationship. Neurocomputing 63, 381–396.
Bonaccorso, B., Bordi, I., Cancelliere, A., Rossi, G., Sutera, A., 2003. Spatial variability of drought: an analysis of the SPI in Sicily. Water Resour. Manag. 17 (4), 273–296.
Box, G.E., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time Series Analysis: Forecasting and Control. John Wiley & Sons.
Brown, S.C., Versace, V.L., Lester, R.E., Walter, M.T., 2015. Assessing the impact of drought and forestry on streamflows in south-eastern Australia using a physically based hydrological model. Environ. Earth Sci. 74 (7), 6047–6063.
Burges, C., Svore, K., Bennett, P., Pastusiak, A., Wu, Q., 2011. Learning to rank using an ensemble of lambda-gradient models. In: Proceedings of the Learning to Rank Challenge, pp. 25–35.
Byun, H.-R., Wilhite, D.A., 1999. Objective quantification of drought severity and duration. J. Climatol. 12 (9), 2747–2756.
Cai, W., Cowan, T., 2008. Dynamics of late autumn rainfall reduction over southeastern Australia. Geophys. Res. Lett. 35 (9).
Cancelliere, A., Bonaccorso, B., Mauro, G., 2006. A non-parametric approach for drought forecasting through the standardized precipitation index. Met. Stat. Mat. l Anal. Ser. Idrol. 1 (1), 1–8.
Cancelliere, A., Di Mauro, G., Bonaccorso, B., Rossi, G., 2007. Drought forecasting using the standardized precipitation index. Water Resour. Manag. 21 (5), 801–819.
Cannon, A.J., Whitfield, P.H., 2002. Downscaling recent streamflow conditions in British Columbia, Canada using ensemble neural network models. J. Hydrol. 259 (1), 136–151.
Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7 (3), 1247–1250.
Chen, G.-C., Yu, J.-S., 2005. Particle swarm optimization algorithm. In: Information and Control-Shenyang. 34(3). pp. 318.
Choubin, B., Malekian, A., Golshan, M., 2016. Application of several data-driven techniques to predict a standardized precipitation index. Atmosfera 29 (2), 121–128.
Climate change impacts on capital Islamabad. The Times of Islamabad.
Cortes, C., Vapnik, V., 1995. Support vector machine. Mach. Learn. 20 (3), 273–297.
Crimp, S., et al., 2015. Bayesian space–time model to analyse frost risk for agriculture in Southeast Australia. Int. J. Climatol. 35 (8), 2092–2108.
Dawson, C.W., Abrahart, R.J., See, L.M., 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ. Model. Softw. 22 (7), 1034–1052.
Dayal, K., Deo Ravinesh, C., Apan, A., 2016a. Application of Hybrid Artificial Neural Network Algorithms for the Prediction of Standardized Precipitation Index, IEEE TENCON 2016 — Technologies for Smart Nation. IEEE, Singapore.
Dayal, K., Deo Ravinesh, C., Apan, A., 2016b. Drought Modelling Based on Artificial Intelligence and Neural Network Algorithms: A Case Study in Queensland, Australia. In: Leal Filho, W. (Ed.), Climate Change Adaptation in Pacific Countries: Fostering Resilience and Improving the Quality of Life. Springer, Berlin.
Dayal, K., Deo, Ravinesh C., Apan, A., 2017. Investigating drought duration-severity-intensity characteristics using the Standardised Precipitation-Evapotranspiration Index: case studies in drought-prone southeast Queensland. J. Hydrol. Eng. 23 (1). http://dx.doi.org/10.1061/(ASCE)HE.1943-5584.0001593.
Deo, R.C., Şahin, M., 2015a. Application of the artificial neural network model for prediction of monthly standardized precipitation and evapotranspiration index using hydrometeorological parameters and climate indices in eastern Australia. Atmos. Res. 161–162, 65–81.
Deo, R.C., Şahin, M., 2015b. Application of the extreme learning machine algorithm for the prediction of monthly effective drought index in eastern Australia. Atmos. Res. 153, 512–525.
Deo, R.C., Şahin, M., 2016. An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland. Environ. Monit. Assess. http://dx.doi.org/10.1007/s10661-016-5094-9.
Deo, R.C., Sahin, M., 2016. An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland. Environ. Monit. Assess. 188 (2), 1.
Deo, R.C., et al., 2009. Impact of historical land cover change on daily indices of climate extremes including droughts in eastern Australia. Geophys. Res. Lett. 36 (8).

Deo, R.C., Byun, H.-R., Adamowski, J., Begum, K., 2015. Application of effective drought index for quantification of meteorological drought events: a case study in Australia. Theor. Appl. Climatol. http://dx.doi.org/10.1007/s00704-015-1706-5(http://link.springer.com/article/10.1007%2Fs00704-015-1706-5.
Deo, R.C., Byun, H.-R., Adamowski, J.F., Begum, K., 2016a. Application of effective drought index for quantification of meteorological drought events: a case study in Australia. Theor. Appl. Climatol. 1–21.
Deo, R.C., Tiwari, M.K., Adamowski, J., Quilty, J., 2017b. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. Stoch. Env. Res. Risk A. 31 (5), 1211–1240. http://dx.doi.org/10.1007/s00477-016-1265-z.
Deo, R.C., Wen, X., Qi, F., 2016c. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Appl. Energy 168, 568–593.
Deo, R.C., Kisi, O., Singh, V.P., 2017. Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. Atmos. Res. 184, 149–175.
Deo, R.C., et al., 2018. Multi-layer perceptron hybrid model integrated with the firefly optimizer algorithm for windspeed prediction of target site using a limited set of neighboring reference station data. Renew. Energy 116, 309–323.
Department, P.M., Pakistan Meteorological Department.
Dijk, A.I., et al., 2013. The millennium drought in Southeast Australia (2001–2009): natural and human causes and implications for water resources, ecosystems, economy, and society. Water Resour. Res. 49 (2), 1040–1057.
Efron, B., Tibshirani, R.J., 1994. An Introduction to the Bootstrap. CRC press.
Foresee, F.D., Hagan, M.T., 1997. Gauss-Newton approximation to Bayesian learning, Neural networks, 1997. In: International Conference on IEEE, pp. 1930–1935.
Ghorbani, M.A., Deo, R.C., Yaseen, Z.M.K., Mahasa, H., Mohammad, B., 2017. Pan evaporation prediction using a hybrid multilayer perceptron-firefly algorithm (MLP-FFA) model: case study in North Iran. Theor. Appl. Climatol. http://dx.doi.org/10.1007/s00704-017-2244-0.
Gilles, J., 2013. Empirical wavelet transform. IEEE Trans. Signal Process. 61 (16), 3999–4010.
Golub, G.H., Reinsch, C., 1970. Singular value decomposition and least squares solutions. Numer. Math. 14 (5), 403–420.
Goyal, M.K., Bharti, B., Quilty, J., Adamowski, J., Pandey, A., 2014. Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, fuzzy logic, and ANFIS. Expert Syst. Appl. 41 (11), 5267–5276.
Guttman, N.B., 1999. Accepting the standardized precipitation index: a calculation algorithm. JAWRA J. Am. Water Res. Assoc. 35 (2), 311–322.
Haider, N., 2016. Living with Disasters.
Hamilton, J.D., 1994. Time Series Analysis, 2. Princeton university press Princeton.
Hayes, M.J., Svoboda, M.D., Wilhite, D.A., Vanyarkho, O.V., 1999. Monitoring the 1996 drought using the standardized precipitation index. Bull. Am. Meteorol. Soc. 80 (3), 429–438.
Hayes, M., Svoboda, M., Wall, N., Widhalm, M., 2011. The Lincoln declaration on drought indices: universal meteorological drought index recommended. Bull. Am. Meteorol. Soc. 92 (4), 485–488.
Hoffmann, F., Schauten, D., Holemann, S., 2007. Incremental evolutionary design of TSK fuzzy controllers. IEEE Trans. Fuzzy Syst. 15 (4), 563–577.
Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A Practical Guide to Support Vector Classification.
Huang, N.E., et al., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. Lond. A Math. Phys. Eng. Sci. 903–995 (The Royal Society).
Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2006. Extreme learning machine: theory and applications. Neurocomputing 70 (1), 489–501.
IPCC, 2012. Summary for policymakers: A special report of working groups I and II of the intergovernmental panel on climate change. In: Field, C.B. (Ed.), Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. Cambridge University Press, Cambridge, UK.
Jalalkamali, A., Moradi, M., Moradi, N., 2015. Application of several artificial intelligence models and ARIMAX model for forecasting drought using the standardized precipitation index. Int. J. Environ. Sci. Technol. 12 (4), 1201–1210.
Jamroz, M., Kolinski, A., Kihara, D., 2016. Ensemble-based evaluation for protein structure models. Bioinformatics 32 (12), i314–i321.
Jang, J.-S., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst. Man Cybernet. 23 (3), 665–685.
Jang, J.-S.R., Sun, C.-T., Mizutani, E., 1997. Neuro-fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence.
Jeong, D.I., Kim, Y.O., 2005. Rainfall-runoff models using artificial neural networks for ensemble streamflow prediction. Hydrol. Process. 19 (19), 3819–3835.
Karthika, B., Deka, P.C., 2015. Prediction of air temperature by hybridized model (Wavelet-ANFIS) using wavelet decomposed data. Aqua. Proc. 4, 1155–1161.
Keyantash, J., Dracup, J.A., 2002. The quantification of drought: an evaluation of drought indices. Bull. Am. Meteorol. Soc. 83 (8), 1167–1180.
Khan, M., Gadiwala, M., 2013. A study of drought over Sindh (Pakistan) using standardized precipitation index (SPI) 1951 to 2010. Pakistan J. Meteorol. 9 (18).
Kim, D.-W., Byun, H.-R., Choi, K.-S., 2009. Evaluation, modification, and application of the effective drought index to 200-year drought climatology of Seoul, Korea. J. Hydrol. 378, 1): 1–12.
Kisi, O., 2015. Pan evaporation modeling using least square support vector machine, multivariate adaptive regression splines and M5 model tree. J. Hydrol. 528, 312–320.
Koehn, J., 2015. Managing people, water, food and fish in the Murray–Darling Basin, south-eastern Australia. Fish. Manag. Ecol. 22 (1), 25–32.
Kormylo, J., Mendel, J., 1982. Maximum likelihood detection and estimation of Bernoulli-Gaussian processes. IEEE Trans. Inf. Theory 28 (3), 482–488.
Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241.
Lei, K.S., Wan, F., 2012. Applying Ensemble Learning Techniques to ANFIS for Air Pollution Index Prediction in Macau, International Symposium on Neural Networks. Springer, pp. 509–516.
Liang, Z., Xie, B., Liao, S., Zhou, J., 2015. Concentration degree prediction of AWJ grinding effectiveness based on turbulence characteristics and the improved ANFIS. Int. J. Adv. Manuf. Technol. 80.
Loukas, A., Vasiliades, L., 2004. Probabilistic analysis of drought spatiotemporal characteristics inThessaly region, Greece. Nat. Hazards Earth Syst. Sci. 4 (5/6), 719–731.
Mayilvaganan, M., Naidu, K., 2011. Comparison of membership functions in adaptive-network-based fuzzy inference system (ANFIS) for the prediction of groundwater level of a watershed. J. Comput. Appl. Res. Dev. 1, 35–42.
McAlpine, C., et al., 2007. Modeling the impact of historical land cover change on Australia's regional climate. Geophys. Res. Lett. 34 (22).
McKee, T.B., Doesken, N.J., Kleist, J., 1993. The relationship of drought frequency and duration to time scales. In: Proceedings of the 8th Conference on Applied Climatology. American Meteorological Society Boston, MA, pp. 179–183.
Mishra, A.K., Singh, V.P., 2010. A review of drought concepts. J. Hydrol. 391 (1), 202–216.
Mishra, A.K., Singh, V.P., 2011. Drought modeling–a review. J. Hydrol. 403 (1), 157–175.
Mitchell, T.M., 1997. Machine Learning, Ser. Computer Science Series. McGraw-Hill Companies, Inc., Singapure.
Mohammadi, K., et al., 2015. A new hybrid support vector machine–wavelet transform approach for estimation of horizontal global solar radiation. Energy Convers. Manag. 92, 162–171.
Moosavi, V., Vafakhah, M., Shirmohammadi, B., Behnia, N., 2013. A wavelet-ANFIS hybrid model for groundwater level forecasting for different prediction periods. Water Resour. Manag. 27 (5), 1301–1321.
Moreira, E.E., Coelho, C.A., Paulo, A.A., Pereira, L.S., Mexia, J.T., 2008. SPI-based drought category prediction using loglinear models. J. Hydrol. 354 (1), 116–130.
Moreira, E., Martins, D., Pereira, L., 2015. Assessing drought cycles in SPI time series using a Fourier analysis. Nat. Hazards Earth Syst. Sci. 15 (3), 571–585.
Mpelasoka, F., Hennessy, K., Jones, R., Bates, B., 2008. Comparison of suitable drought indices for climate change impacts assessment over Australia towards resource management. Int. J. Climatol. 28 (10), 1283–1292.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. 10 (3), 282–290.
Nayak, P.C., Sudheer, K., Rangan, D., Ramasastri, K., 2004. A neuro-fuzzy computing technique for modeling hydrological time series. J. Hydrol. 291 (1), 52–66.
Opitz, D.W. and Maclin, R., 1999. Popular ensemble methods: An empirical study. J. Artif. Intell. Res. (JAIR), 11: 169–198.
Pakistan, D.I., 1950–2015. List of natural disasters in Pakistan.
Palmer, W.C., 1965. Meteorological Drought, 30. US Department of Commerce, Weather Bureau Washington, DC.
Palmer, W.C., 1968. Keeping Track of Crop Moisture Conditions, Nationwide: The New Crop Moisture Index.
Paulo, A.A., Pereira, L.S., 2007. Prediction of SPI drought class transitions using Markov chains. Water Resour. Manag. 21 (10), 1813.
Pérez, E.C., Algredo-Badillo, I. and Rodríguez, V.H.G., 2012. Performance analysis of ANFIS in short term wind speed prediction. (arXiv preprint arXiv:1212.2671).
Quinlan, J.R., 1992. Learning With Continuous Classes, 5th Australian Joint Conference on Artificial Intelligence, Singapore. pp. 343–348.
Qureshi, M., Ahmad, M., Whitten, S., Reeson, A., Kirby, M., 2016. Impact of climate variability including drought on the residual value of irrigation water across the Murray–Darling Basin, Australia. Water Econ. Policy 1550020.
Raheli, B., Aalami, M.T., El-Shafie, A., Ghorbani, M.A., Deo, R.C., 2017. Uncertainty assessment of the multilayer perceptron (MLP) neural network model with implementation of the novel hybrid MLP-FFA method for prediction of biochemical oxygen demand and dissolved oxygen: a case study of Langat River. Environ. Earth Sci. 76 (14), 503.
Rahimikhoob, A., Asadi, M., Mashal, M., 2013. A comparison between conventional and M5 model tree methods for converting pan evaporation to reference evapotranspiration for semi-arid region. Water Resour. Manag. 27 (14), 4815–4826.
Rahmat, S.N., Jayasuriya, N., Bhuiyan, M.A., 2016. Short-term droughts forecast using Markov chain model in Victoria, Australia. Theor. Appl. Climatol. 1–13.
Rajathi, N., Jayashree, L., 2016. Soil Moisture Forecasting Using Ensembles of Classifiers, Proceedings of First International Conference on Information and Communication Technology for Intelligent Systems: Volume 1. Springer, pp. 235–244.
Reeves, C.R., 1995. A genetic algorithm for flowshop sequencing. Comput. Oper. Res. 22 (1), 5–13.
Riebsame, W.E., Changnon Jr., S.A., Karl, T.R., 1991. Drought and natural resources management in the United States. In: Impacts and Implications of the 1987–89 Drought. Westview Press Inc..
Santos, C.A.G., Morais, B.S., Silva, G.B., 2009. Drought forecast using an artificial neural network for three hydrological zones in San Francisco River basin, Brazil. IAHS Publ. 333, 302.
Sehgal, V., Sahay, R.R., Chatterjee, C., 2014. Effect of utilization of discrete wavelet components on flood forecasting performance of wavelet based ANFIS models. Water Resour. Manag. 28 (6), 1733–1749.
Shirmohammadi, B., Moradi, H., Moosavi, V., Semiromi, M.T., Zeinali, A., 2013. Forecasting of meteorological drought using wavelet-ANFIS hybrid model for different time steps (case study: southeastern part of East Azerbaijan province, Iran). Nat. Hazards 69 (1), 389–402.
Smarandache, F., 2001. A Unifying Field in Logics: Neutrosophic Logic, Math/0101228.
Sönmez, F.K., KÖmÜscÜ, A.Ü., Erkan, A., Turgu, E., 2005. An analysis of spatial and

temporal dimension of drought vulnerability in Turkey using the standardized precipitation index. Nat. Hazards 35 (2), 243–264.
Srivastav, R., Sudheer, K., Chaubey, I., 2007. A simplified approach to quantifying predictive and parametric uncertainty in artificial neural network hydrologic models. Water Resour. Res. 43 (10).
Strauss, T., Hanselmann, M., Junginger, A. and Ulmer, H., 2017. Ensemble methods as a defense to adversarial perturbations against deep neural networks. (arXiv preprint arXiv:1709.03423).
Strohmann, T., Grudic, G.Z., 2003. A formulation for minimax probability machine regression. Adv. Neural Inf. Proces. Syst. 785–792.
Sujitha, E., 2017. Analysis of dry/wet conditions using the standardized precipitation index and its potential usefulness for drought/flood monitoring in the regions of Trichy. J. Pharmacog. Phytochem. 6 (4), 452–457.
Svoboda, M., Hayes, M., Wood, D., 2012. Standardized Precipitation Index User Guide. World Meteorological Organization Geneva, Switzerland.
Takeuti, G., Titani, S., 1984. Intuitionistic fuzzy logic and intuitionistic fuzzy set theory. J. Symb. Log. 49 (3), 851–866.
Taniar, D., 2009. Strategic Advancements in Utilizing Data Mining and Warehousing Technologies: New Concepts and Developments. IGI Global.
Timbal, B., Hendon, H., 2011. The role of tropical modes of variability in recent rainfall deficits across the Murray-Darling Basin. Water Resour. Res. 47 (12).
Tiwari, M.K., Chatterjee, C., 2010a. Development of an accurate and reliable hourly flood forecasting model using wavelet–bootstrap–ANN (WBANN) hybrid approach. J. Hydrol. 394 (3), 458–470.
Tiwari, M.K., Chatterjee, C., 2010b. Uncertainty assessment and ensemble flood forecasting using bootstrap based artificial neural networks (BANNs). J. Hydrol. 382 (1), 20–33.
Tiwari, M.K., Chatterjee, C., 2011. A new wavelet–bootstrap–ANN hybrid model for daily discharge forecasting. J. Hydroinf. 13 (3), 500–519.
Vairappan, C., Tamura, H., Gao, S., Tang, Z., 2009. Batch type local search-based adaptive neuro-fuzzy inference system (ANFIS) with self-feedbacks for time-series prediction. Neurocomputing 72 (7), 1870–1877.
Vicente-Serrano, S., 2016. Foreword: drought complexity and assessment under climate change conditions. Cuad. Investig. Geog. 42 (1), 7–11.
Wilhite, D.A., Hayes, M.J., 1998. Drought Planning in the United States: Status and Future Directions, The Arid Frontier. Springer, pp. 33–54.
Wilhite, D.A., Hayes, M.J., Knutson, C., Smith, K.H., 2000. Planning for Drought: Moving from Crisis to Risk management1. Wiley Online Library.
Williams, A., et al., 2015. Quantifying the response of cotton production in eastern Australia to climate change. Clim. Chang. 129 (1–2), 183–196.
Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2 (2), 184–194.
Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 63 (11), 1309–1313.
Willmott, C.J., 1984. On the Evaluation of Model Performance in Physical Geography, Spatial Statistics and Models. Springer, pp. 443–460.
Willmott, C.J., Robeson, S.M., Matsuura, K., 2012. A refined index of model performance. Int. J. Climatol. 32 (13), 2088–2094.
Wittwer, G., Adams, P.D., Horridge, M., Madden, J.R., 2002. Drought, regions and the Australian economy between 2001–02 and 2004–05. Aust. Bull. Lab. 28 (4), 231.
Xie, H., Ringler, C., Zhu, T., Waqas, A., 2013. Droughts in Pakistan: a spatiotemporal variability analysis using the Standardized Precipitation Index. Water Int. 38 (5), 620–631.
XingXing, W., XiLin, Z., HuiXiao, Y., 2008. Application of improved ANFIS in optimization of machining parameters. Chin. J. Mech. Eng. 44 (1), 199–204.
Yang, X.-S., 2010. Firefly algorithm, stochastic test functions and design optimisation. Int. J. Bio-Insp. Comput. 2 (2), 78–84.
Yaseen, Z.M., et al., 2018. Rainfall pattern forecasting using novel hybrid intelligent model based ANFIS-FFA. Water Resour. Manag. 32 (1), 105–122. http://dx.doi.org/10.1007/s11269-017-1797-0.
Yen, B.C., 1995. Discussion and closure: criteria for evaluation of watershed models. J. Irrig. Drain. Eng. 121 (1), 130–132.
Yevjevich, V., 1967a. An Objective Approach to Definitions and Investigations of Continental Hydrologic Droughts. Colorado State University, Fort Collins.
Yevjevich, V.M., 1967b. An objective approach to definitions and investigations of continental hydrologic droughts. (Hydrology papers (Colorado State University); no. 23).
Yevjevich, V., 1991. Tendencies in hydrology research and its applications for 21st century. Water Resour. Manag. 5 (1), 1–23.
Yuan, W.-P., Zhou, G.-S., 2004. Comparison between standardized precipitation index and Z-index in China. Acta Phytoecol. Sin. 4.
Zaidi, H.B., 2016. Expect more Extreme Weather this Year. (Dawn).
Zhisheng, Z., 2010. Quantum-behaved particle swarm optimization algorithm for economic load dispatch of power system. Expert Syst. Appl. 37 (2), 1800–1803.
Zhou, Z.-H., Wu, J., Tang, W., 2002. Ensembling neural networks: many could be better than all. Artif. Intell. 137 (1–2), 239–263.
Zhou, Q., et al., 2011. A new method to obtain load density based on improved ANFIS. Power Syst. Prot. Control 39 (1), 29–34.

Chapter 5

Multi-stage committee based extreme learning machine model incorporating the influence of climate parameters and seasonality on drought forecasting

Foreword

This chapter is an exact copy of the published article in *Computers and Electronics in Agriculture* journal (Vol. 152, Pages 149-165).

The ensemble-ANFIS model in Chapter 4 has the limitation of short term (i.e. monthly) drought forecasting. Therefore addressing this issue, a new soft-computing committee based drought modelling framework for short term forecasting was designed in this study. Committee of modelling technique is a model combination technique, which is uncommon in climatological studies. In this study, the ELM-based committee (Comm-ELM) was developed using multiple meteorological input predictors and explore for monthly drought forecasting in three regions of Pakistan.

For comparison purpose, the PSO-ANFIS-based committee (Comm-PSO-ANFIS) and MLR-based committee (Comm-MLR) models were also designed. The results shows the Comm-ELM outperformed in relation to Comm-PSO-ANFIS and Comm-MLR. The Comm-ELM is only capable of forecasting short term (i.e. monthly) drought. It is unable to forecast medium and long term drought forecasting.

Computers and Electronics in Agriculture 152 (2018) 149–165

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

Multi-stage committee based extreme learning machine model incorporating the influence of climate parameters and seasonality on drought forecasting

Mumtaz Ali, Ravinesh C. Deo*, Nathan J. Downs, Tek Maraseni

School of Agricultural, Computational and Environmental Sciences, Institute of Agriculture and Environment, University of Southern Queensland, Springfield, QLD 4300, Australia

ARTICLE INFO

Keywords:
Standardized precipitation index
Drought forecasting
Committee model
Extreme learning machine
Particle swarm optimization based adaptive neuro fuzzy inference system
Multi-linear regression

ABSTRACT

Drought forewarning is an important decisive task since drought is perceived a recurrent feature of climate variability and climate change leading to catastrophic consequences for agriculture, ecosystem sustainability, and food and water scarcity. This study designs and evaluates a soft-computing drought modelling framework in context of Pakistan, a drought-stricken nation, by means of a committee extreme learning machine (Comm-ELM) model in respect to a committee particle swarm optimization-adaptive neuro fuzzy inference system (Comm-PSO-ANFIS) and committee multiple linear regression (Comm-MLR) model applied to forecast monthly standardized precipitation index (SPI). The proposed Comm-ELM model incorporates historical monthly rainfall, temperature, humidity, Southern Oscillation Index (SOI) at monthly lag ($t-1$) and the respective month (*i.e.*, periodicity factor) as the explanatory variable for the drought's behaviour defined by SPI. The model accuracy is assessed by root mean squared error, mean absolute error, correlation coefficient, Willmott's index, Nash-Sutcliffe efficiency and Legates McCabe's index in the independent test dataset. With the incorporation of periodicity as an input factor, the performance of the Comm-ELM model for Islamabad, Multan and Dera Ismail Khan (D. I. Khan) as the test stations, was remarkably improved in respect to the Comm-PSO-ANFIS and Comm-MLR model. Other than the superiority of Comm-ELM over the alternative models tested for monthly SPI forecasting, we also highlight the importance of the periodicity cycle as a pertinent predictor variable in a drought forecasting model. The results ascertain that the model accuracy scales with geographic factors, due to the complexity of drought phenomenon and its relationship with the different inputs and data attributes that can affect the overall evolution of a drought event. The findings of this study has important implications for agricultural decision-making where future knowledge of drought can be used to develop climate risk mitigation strategies for better crop management.

1. Introduction

Drought is a socio-economic hazard posing severe threats to groundwater reservoirs, leading to a scarcity of water resources, crop failure and socio-economic challenges (Mpelasoka et al., 2008; Riebsame et al., 1991; Wilhite et al., 2000; Deo et al., 2009). A general scarcity of sufficient palatable water due to rapid growth of human populations and an expansion of agricultural, energy and industrial sectors is also a growing concern (IPCC, 2012; McAlpine et al., 2007). Climate change further escalates the potential severity and frequency of drought events (Deo et al., 2009; McAlpine et al., 2009). Large-scale climate mode indices including the Southern Oscillation Index (SOI) is significantly correlated with fluctuations in rainfall and onset of drought (Mishra and Singh, 2010; Morid et al., 2007; Nguyen-Huy et al., 2017; Özger et al., 2012). Hence drought forecasts that incorporate climate factors can assist hydrologists, agriculturalists and resource planners in strategic decisions to address socio-economic challenges posed by a drought, particularly in cases of prolonged events (Bates et al., 2008; Deo et al., 2016a; Mishra and Singh, 2010, 2011; Wilhite and Hayes, 1998).

Traditionally, drought indices are used to measure, monitor and forecast drought. Several different indices have been developed according to their appropriateness for a given geographic region (Mishra and Singh, 2010, 2011). The Palmer (1965) Drought Severity Index (PDSI) measures the overall dryness based on precipitation and temperature datasets. PDSI is particularly useful in pointing out long-term drought events and it is not appropriate for a region with generally high surface run-off (Mishra and Singh, 2010, 2011). To tackle the

* Corresponding author.
E-mail addresses: Mumtaz.Alli@usq.edu.au (M. Ali), ravinesh.deo@usq.edu.au (R.C. Deo).

https://doi.org/10.1016/j.compag.2018.07.013
Received 21 April 2018; Received in revised form 4 July 2018; Accepted 5 July 2018

complexities associated with PDSI, researchers have developed the Crop Moisture Index (CMI), particularly for assessing agricultural drought events (Palmer, 1968). CMI is based on the rank of the precipitation records and aims to compute both positive and negative precipitation anomalies. However, CMI in the short-term is insufficient to offset long-term issues. Byun and Wilhite (1999) developed the Effective Drought Index (EDI) based on daily precipitation data. EDI is a good index for operational monitoring of meteorological and agricultural drought, although the consideration of precipitation alone does not take into account other environmental parameters that also cause drought (*e.g.*, the impact of temperature). Several drought indices have been developed built on PDSI to take into account additional data on precipitation and crop moisture, however the standardized precipitation index (SPI) is used universally as a standard metric (Deo et al., 2017a).

The importance of modelling SPI is derived from the notion that: (1) SPI is able to assess the water shortage situations built on a statistical distribution of rainfall that can enable both short-term (i.e. monthly) and long-term (i.e., seasonal and annual) drought assessments (ranging from 1 up to 48 months). (2) SPI is presented as a normalized standard metric of rainfall surpluses/deficits in relation to a benchmark climatological period (Hayes et al., 1999; McKee et al., 1993; Yuan and Zhou, 2004), and therefore, it can enable a comparison of the drought behaviour in geographically and climatologically diverse regions. (3) SPI has been explored and validated for drought mitigation studies in diverse climatic regions (Almedeij, 2016; Choubin et al., 2016; Svoboda et al., 2012). (4) SPI has the potential to represent both short (1 and 3 months) and long-term (6–12 months) drought in a probabilistic fashion, largely on multiple timescales. In the case of agricultural drought, the SPI makes it possible to examine the soil moisture status with respect to precipitation anomalies on a comparatively short timescale using hydrological reservoirs to replicate long-term climate anomalies (Svoboda et al., 2012). Due to the advantageous features, the SPI is an ideal metric for the management of not only hydrological, but also for agricultural drought events (Deo et al., 2017a; Guttman, 1999). In this study, SPI based on 1 month is utilized for drought forecasting as there is no study on a monthly drought in the selected study regions, although the study of Ali et al. (2018a) designed a study using ensemble-ANFIS model for long term (3–12 months) drought forecasting and Ali et al. (2018b) developed a multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm for rainfall forecasting, both focussed in Pakistan.

In the existing literature, data-driven models have been used to model SPI for drought forecasting. For example, an SPI-based methodology was designed by Cancelliere et al. (2007) to forecast probabilistic drought alterations in Sicily, Italy. Jalalkamali et al. (2015) in Yazd, Iran conducted a study to forecast the SPI using several model variants, including a multilayer perceptron artificial neural network (MLP ANN), an adaptive neuro-fuzzy inference system (ANFIS), support vector machines (SVM), and an autoregressive integrated moving average (ARIMA) multivariate model. In another study, models based on ANFIS and ANN Wavelet tools were adopted by Shirmohammadi et al. (2013) to forecast SPI in Azerbaijan. An SPI-based forecasting study was also performed by Santos et al. (2009) using an ANN model for San Francisco, USA. Drought forecasts using SPI have an extensive history in the current literature (Adamowski et al., 2012; Bonaccorso et al., 2003; Cancelliere et al., 2006; Choubin et al., 2016; Deo et al., 2017b; Guttman, 1999; Hayes et al., 1999; Jalalkamali et al., 2015; Moreira et al., 2015; Moreira et al., 2008; Paulo and Pereira, 2007; Sönmez et al., 2005). However, SPI based drought forecasts are yet to be explored for agricultural regions in Pakistan where the influence of drought is a major impeding factor for crop productivity and farmers' livelihoods.

Drought hazard continues to severely affect the agriculturally dependent nation of Pakistan (Report, 1950-2015). The severe drought event of 1998 led to a significant reduction in Pakistan's national agricultural productivity by 2.6% over 2000–2001 (Ahmad et al., 2004). In spite of such pressing issues, drought models for local agricultural zones in Pakistan have been very limited: (1) Khan and Gadiwala (2013) aimed to investigate drought behaviour using SPI at multi timescales for the province of Sindh: (2) Xie et al. (2013) applied a spatiotemporal variability analysis based on the SPI data to also forecast drought behavior, and most recently, (3) the study of Ali et al. (2017) forecasted drought based on Standardized Precipitation-Evapotranspiration Index (SPEI) where a multilayer perceptron-based artificial neural network model was employed. (4) The study of Ahmed et al. (2016) has utilized antecedent SPI data for the characterization of future seasonal drought events in Balochistan, Pakistan and (5) Ali et al. (2018a) implemented an ensemble strategy based on the ANFIS model to forecast the SPI using the historical SPI to forecast future SPI. While it was not specifically on drought forecasting, a recent study of Ali et al. (2018b) has forecasted rainfall in Pakistan using a multi-stage online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm. However, there has been no study in Pakistan specifically on future drought models or drought indices utilising different climatological parameters.

Considering a lack of drought models for developing nations like Pakistan, the aim of this research is to develop and evaluate Comm-ELM (a data-intelligent model) with universal approximation capabilities. The specific objectives are:

(i) To develop a committee based extreme learning machine (Comm-ELM) model following its successful application elsewhere, *e.g.*, (Barzegar et al., 2018; Prasad et al., 2018) and evaluate its preciseness for forecasting future SPI incorporating the relevant hydro-meteorological dataset (*i.e.*, temperature, rainfall, humidity, Southern Oscillation Index) and the seasonality metric (i.e., periodicity) as predictors for the period of 1081–2015. Here, a multi-stage model strategy is adopted incorporating antecedent climate-based variables at $(t-1)$ and corresponding periodicity in the first stage, and the forecasted SPI.
(ii) To elucidate the importance of periodicity as a pertinent factor in drought forecast models and seasonal influences on drought progression using data monitoring for Pakistan, a nation vulnerable to significant agricultural and water management issues.
(iii) To compare the performances of Comm-ELM in respect to Comm-PSO-ANFIS and Comm-MLR models for SPI-forecasting.

2. Theoretical background

2.1. Standardized precipitation index (SPI)

The SPI drought forecasting metric, centred on normalized probabilities of dryness relative to a base climatology, provides a depiction of irregular wetness and dryness situation. Before developing the forecast model for the designated regions in Pakistan, the monthly SPI index was calculated using precipitation (PCN) data (Deo et al., 2017a; McKee et al., 1993). Pearson Type III distribution/gamma distribution function is given by the following mathematical expression:

$$g(PCN) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}(PCN)^{\alpha-1}e^{-x/\beta} \quad (1)$$

where α and β represents the estimated parameters using the maximum likelihood. The cumulative probability can be given by

$$G(PCN) = \int_0^P g(PCN)dPCN = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\int_0^P x^{\alpha-1}e^{-P/\beta}dPCN \quad (2)$$

Suppose that $m = PCN/\beta$, this reduces Eq. (2) to an incomplete gamma function:

$$G(PCN) = \frac{1}{\Gamma(\alpha)}\int_0^m m^{\alpha-1}e^{-m}dm \quad (3)$$

As for $PCN = 0$, the gamma function is undefined, so the cumulative probability becomes:

$$H(\mathrm{PCN}) = q + (1-q)G(PCN) \tag{4}$$

where q is the probability of zero. This yields the monthly value of SPI, viz:

$$SPI = \begin{cases} +\left(m - \frac{c_0 + c_1 m + c_2 m^2}{1 + d_1 m + d_2 m^2 + d_3 m^3}\right), & 0.5 < H(PCN) \leqslant 1.0 \\ -\left(m - \frac{c_0 + c_1 m + c_2 m^2}{1 + d_1 m + d_2 m^2 + d_3 m^3}\right), & 0 < H(PCN) \leqslant 0.5 \end{cases} \tag{5}$$

The magnitudes of the constants in Eq. (5) are as follows: $c_0 = 2.515517$, $c_1 = 0.802853$, $c_2 = 0.010328$, $d_1 = 1.432788$, $d_2 = 0.189269$ and $d_3 = 0.001308$ (McKee et al., 1993). These constants are used in the transformation of the cumulative probability H (PCN) into a standard normal random variable SPI with mean zero and variance 1. The standardized precipitation index represents an SPI-score (Loukas and Vasiliades, 2004; McKee et al., 1993; Sujitha, 2017). Based on the calculated SPI, drought can be categorized as moderately dry = ($-1.5 <$ SPI ≤ 1.0), severely dry = ($-2.0 < SPI \leq -1.5$), and extremely dry = (SPI ≤ -2.0).

Fig. 2 show the monthly SPI time-series, portraying the evolution of drought episodes from May 1991–Mar 1992 for Multan, Pakistan. The onset of drought using the running sum approach of Yevjevich (1967, 1991) can be deduced as the particular month when the SPI value declined below 0 and the termination of drought when the SPI value first returned to positivity. In accord with this, the cumulative rainfall appears to be reduced significantly in this dry period. The drought severity is then the sum of all months with SPI < 0 and the drought's peak intensity occurs when the SPI value is at its minimum point.

In the proposed Comm-ELM model, the SPI is used as an objective/observed variable in this study.

2.2. Extreme learning machine (ELM)

ELM is a state-of-the-art data intelligent model developed by Huang et al. (2006) used for the purpose of designing a Single Layer Feedforward Neural Network (SLFN). ELM is relatively faster, and computationally efficient compared to traditional algorithms such as back propagation (BP) or support vector machines (SVM) (Rajesh and Prakash, 2011). The standard Single Layer Feedforward Neural Network (SLFN) with M hidden nodes of N arbitrary inputs $(x_k, y_k) \in \Gamma^n \times \Gamma^n$ with an activation function $f(.)$ can be mathematically formulated as:

$$\sum_{i=1}^{M} \rho_i f(x_k, ;c_i, w_i) = y_k \tag{6}$$

where $k = 1, 2, \cdots, N$, $c_i \in \Gamma$ is the bias of ith node which is assigned randomly whereas $w_i \in \Gamma$ is a random input weight vector that connects the ith hidden node with the output node. The function $g(x_k, ;c_{i,w_i})$ denotes the output corresponding to ith hidden node with respect to input x_k. Therefore Eq. (1) reduces to the following form:

$$\mathrm{H}\beta = Y \tag{7}$$

$$\text{where}\mathrm{H} = \begin{bmatrix} f(x_1, ;c_1, w_1) & \cdots & f(x_1, ;c_M, w_M) \\ \vdots & \cdots & \vdots \\ f(x_N, ;c_1, w_1) & \cdots & f(x_N, ;c_M, w_M) \end{bmatrix}_{N \times M},$$

$\mathrm{H}\beta = (\beta_1{}^T \beta_2{}^T, \cdots, \beta_L{}^T)^T{}_{m \times M}$ and $Y = (t_1{}^T t_2{}^T, \cdots, t_L{}^T)^T{}_{m \times M}$. The least square solution of the linear systems provides the following output weight:

$$\beta = \mathrm{H}^{+} Y \tag{8}$$

where H^{+} represents the Moore–Penrose generalized inverse of$\mathrm{H}$. The SLFNs with random input weight selection effectively acquire distinct training examples with minimum chance of error (Huang, 2003; Tamura and Tateishi, 1997).

Fig. 1(a) illustrates the basic structure of the ELM model.

2.3. Particle swarm optimization based adaptive neuro fuzzy inference system (PSO-ANFIS)

The ANFIS model was developed by Jang (1993) as a division of the adaptive tool (*i.e.* the outputs being dependent on the parameters belonging to the input nodes). The ANFIS model utilizes two inputs to generate one output employing the fuzzy 'if-then' rules of the Takagi-Sugeno-Kang (TSK) (Hoffmann et al., 2007) fuzzy model which can be defined as;

Rule (a): if α is A_1 and β is B_1, then $f_1 = p_1\alpha + q_1\beta + s_1$ (9)

Rule (b): if α is A_2 and β is B_2, then $f_2 = p_2\alpha + q_2\beta + s_2$ (10)

where $\alpha$and β represent the input of the ANFIS model whereas Aand Bare the fuzzy set with $f_j\,(j = 1, 2)$ being the first order output polynomial of the TSK fuzzy inference system, while p_j, $q_j$and $s_j\,(j = 1, 2)$ are the set of consequent parameters. The elementary construction of ANFIS can be seen as a 5 layer feedforward neural network. Each node $j$is an adaptive node in layer 1 with a suitable membership function related to the input to node$j$.

$$\Phi_{1,j} = \mu_{\mathrm{A}_j}(\alpha),\ \Phi_{1,j} = \mu_{\mathrm{B}_{(j-2)}}(\beta),\ (j = 1, 2), \tag{11}$$

In Eq. (11), α, β denote the input nodes whileA and Bare the linguistic labels with $\mu(\alpha)$, $\mu(\beta)$ representing the membership function (usually bell-shaped) that specifies the degree to which the given input satisfies the quantifiersA, B. The membership function is defined as:

$$\mu(\alpha) = \frac{1}{1 + \left(\frac{(\alpha - c_j)}{a_j}\right)^{b_j}} \tag{12}$$

where a_j, $b_j$and $c_j$are the parameters. The bell-shaped function adopts different forms of membership with the variation of these parameters. These outputs are the fire strength rules given by,

$$\Phi_{2,j} = \omega_j = \mu_{\mathrm{A}_j}(\alpha) \times \mu_{\mathrm{B}_j}(\beta) \tag{13}$$

Here $\omega_j$is the firing strength of a rule. Every node in layer 3 is also a fixed circular node labelled N with normalized firing strength as output which is basically the ratio of jth rule's firing strength to the sum of all rule's firing strength. In layer 4, each node is turned to be an adaptive node marked by a square whereas in layer 5, the consequent parameters are then expressed into a linear combination/summation to compute the overall output of all the fixed nodes. In ANFIS, the epoch of hybrid learning consists of a forward and backward pass (Goyal et al., 2014; Jang et al., 1997; Karthika and Deka, 2015; Mayilvaganan and Naidu, 2011; Moosavi et al., 2013; Nayak et al., 2004; Pérez et al., 2012; Sehgal et al., 2014; Shirmohammadi et al., 2013).

In this paper, the Fuzzy c-mean clustering method (FCM) designed by Dunn (1973) and Bezdek et al. (1984) is utilized to determine the antecedent membership functions. FCM is a data clustering algorithm which allows one piece of data to belong to two or more clusters in such a way that each data point belongs to a cluster to a degree specified by a membership grade. The collection of n vectors X_i, $(i = 1, 2, 3, \cdots, n)$ is partitioned through FCM into C-fuzzy groups to determine a cluster centre in each group to minimize the cost function of dissimilarity measure.

To improve the versatility of ANFIS model using the FCM algorithm, we used the Particle Swarm Optimization (PSO) technique (Çavdar, 2016) to tune the ANFIS parameters. A particle changes its position each time by tuning its velocity. The velocity vector is then updated corresponding to the position of global best ($gbest$) and personal best ($pbest$) position of each particle that can be defined mathematically as:

$$v_i(t+1) = w \times v_i(t) + C_1 \times \aleph_1(pbest - p(t)) + C_2 \times \aleph_2(gbest - x_i(t)) \tag{14}$$

$$x_i(t+1) = x_i(t) + v_i(t+1) \tag{15}$$

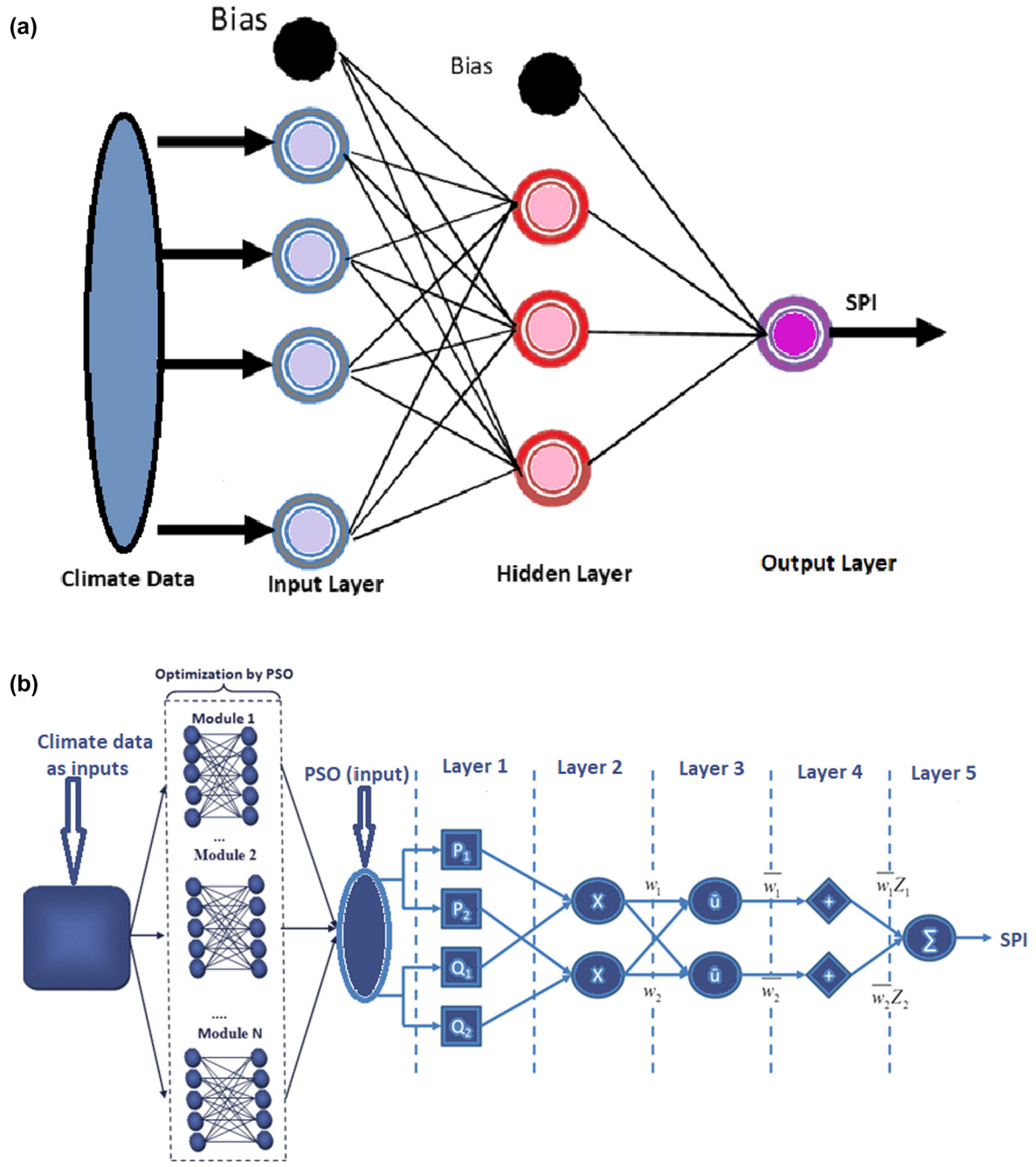


Fig. 1. Schematic view of the (a) ELM and (b) PSO-ANFIS models.

Here, $v_i(t)$ is called the agent velocity, $x_i(t)$ is the agent current position at iteration$t$; C_1, C_2 are cognitive and social accelerations; w is the weight of inertia; $\aleph_1$, $\aleph_2 \in (0, 1)$ are uniformly distributed random numbers whereas ($gbest$), and($pbest$) is the global best and personal best positions related within the swarm and ithparticle respectively. Fig. 1(b) describes the schematic view of PSO-ANFIS.

2.4. Multiple linear regression (MLR)

MLR is a generalized form of the simple regression model from single to multiple predictors where the objective is to deduce a model that can exhibit the maximum deviations in the predictor data to evaluate their corresponding regression coefficients (Deo and Sahin, 2017). MLR guarantees that the forecast model minimizes the variations that appear in data due to unexplained "noise". For n observations of k predictor variables, an MLR model adapts the following regression equation (Draper and Smith, 1981; Montgomery et al., 2012):

$$\Im = c + \alpha_1 \mathrm{M}_1 + \alpha_2 \mathrm{M}_2 + \ldots + \alpha_k \mathrm{M}_k \tag{16}$$

where $\Im(n + 1)$ is a matrix of forecasting SPI, $\mathrm{M}(n \times k)$ is a vector of input/predictor variable(s), $c$is the y-intercept and $\alpha$is the coefficient of multiple regression for each regressor variable(s) (Civelekoglu et al., 2007; Şahin et al., 2013).

It should be noted that the value of $\alpha$is approximated through least squares (*e.g.*, (Apaydın et al., 1994; Ozdamar, 2004)) for each predictor variable. The multiple linear equations are fitted to a model with a set of $\Im$ and Mmatrix in the training period. The fitted MLR model then uses the coefficients and the y-intercept to generate the forecasts of $\Im$ values as well as Mvalues in the testing phase.

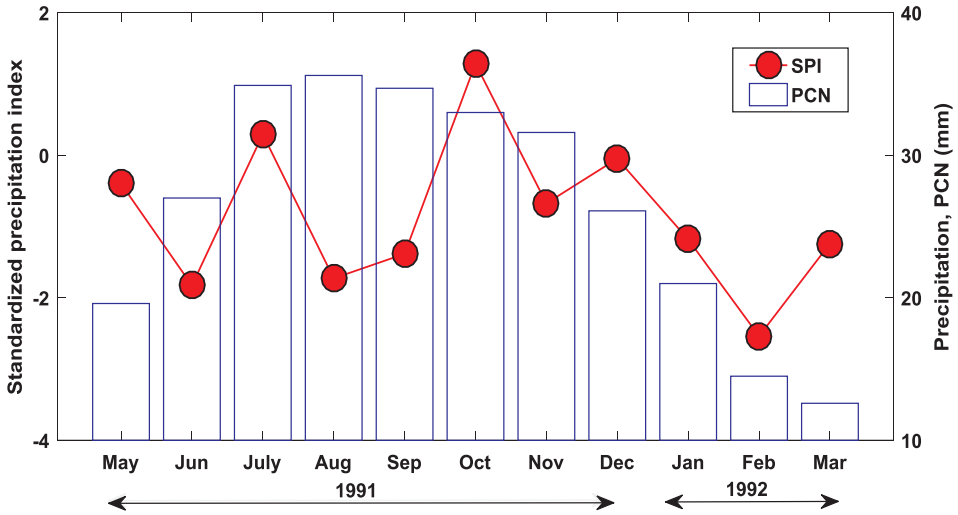


Fig. 2. Monthly standardized precipitation index showing drought characteristics and the correspondence with precipitation (PCN) data of the drought period (May 1991 to Mar 1992) for Multan Station. The unit of PCN is millimetre (mm).

3. Materials and method

3.1. Climate data

The climate data utilized in this study includes precipitation (PCN), mean monthly temperature (T) and mean monthly relative humidity (H). These parameters were obtained from the Pakistan Meteorological Department (PMD, 2016) for the period 1981–2015. Prior to developing the drought forecasting model, the monthly SPI index was computed using the rainfall data from 1981 to 2015.

As machine learning models rely on a prognostic features in historical data to forecast the future drought, the climate index named SOI (Nicholls, 2004) was used as an input with the climate data (PCN, T, H) of historical antecedent month at $(t-1)$ in the proposed drought forecasting models. This followed several earlier studies where drought and streamflow has been forecasted using synoptic-scale climate drivers (Deo et al., 2017a; Deo and Sahin, 2016). SOI computes the difference in surface air pressure between Tahiti and Darwin which is best at a monthly or longer duration. The positive values of SOI indicate La Niña (cold phase) while negative values represent El Niño (warm phase).

SOI has been utilized in this study because it has significant influence on the climate in Pakistan (Afzal et al., 2013) as compared to other climate index. Finally, the periodicity/seasonality (number of months) was incorporated in the proposed forecasting models to study their effect on the drought events and analyse their contribution to the model accuracy.

3.2. Study region

The regions for this study are: Islamabad, Dera Ismail Khan (denoted 'D. I. Khan') and Multan (Fig. 3). These stations are important for agriculture production as well as different geographical environments in which to test the model.

Islamabad is the capital city of Pakistan (population approximately 1.152 million) and has a humid subtropical climate with four seasons: winter (December-February), spring (March-May), summer (June-September) and autumn (October-November). The average monsoon rainfall is about 790.8 mm. Islamabad received the heaviest 620 mm rainfall in just 10 h on 23 July 2002, which was the heaviest rainfall in the past 100 years. The annual average rainfall is 1142.1 mm. D. I. Khan (a regional centre of population 1.627 million) is located in the province of Khyber Pakhtunkhwa (KPK). The climate entails hot summers and mild winters. In 2012, water scarcity due to drought in D. I. Khan had badly affected the agriculture sector (Amir, 2012). The average annual rainfall is about 268.8 mm while the average annual temperature is 24.5 °C. Multan (another regional centre, population million) is located in southern part of Punjab province. The climate is arid with hot summers and cold winters. The average temperature of Multan in summer is 42.3 °C while the record breaking highest temperature was 50.0 °C in May 2010 (Department, 2010). The average annual rainfall is 186.8 mm and the climate is also affected by monsoon seasons. The average annual temperature is 25.6 °C in Multan. Multan has also experienced the worst heat waves in Pakistan's history on three occasions (Department, 2010; Salman, 2006).

Table 1 describes the latitude, longitude, elevation, minimum, maximum, standard deviation, skewness and kurtosis of rainfall, temperature, and humidity and SPI data for the designated regions that has been utilized during the development of the forecasting models.

3.3. Development of data-driven committee based ELM model

Historical climate data (T, PCN, H), SOI index and periodicity (months) were used to develop the proposed Comm-ELM model in relation with Comm-PSO-ANFIS and Comm-MLR. It is important to note the proposed approach concurs with earlier research performed (Barzegar et al., 2018). However, this study adopt the ELM as the committee-based model which is more advanced and highly optimized data intelligent model (Huang et al., 2006; Rajesh and Prakash, 2011). The origional data with the historical antecedent month at $(t-1)$ as input predictors were employed to forecast the drought. The proposed Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR models were developed under MATLAB environment on a Pentium 4 2.93 GHz dual core Central Processing Unit. The development and validation of the proposed Comm-ELM model can be describe in the following steps:

Step 1: First the historical antecedent month at $(t-1)$ were calculated from the T, PCN, H, SOI and corresponding periodicity month at first stage.

Step 2: In the second stage, after incorporating historical temperature at $(t-1)$ as input predictor, the ELM model was applied to simulate the SPI. Next, the temperature and rainfall at lag $(t-1)$ were employed together to compute SPI. Further, the combination of humidity was incorporated with temperature and rainfall to forecast SPI index. Moreover, the SOI index was taken together with temperature, rainfall and humdity to estmiate SPI and finally, the respective number of month (periodicty) were combined with temperature, rainfall, humidity and SOI index to forecast the drought index at significant lag $(t-1)$. The regression ELM type for training was selected first with hidden neurons from 1 to 60. Moreover, all the activation functions

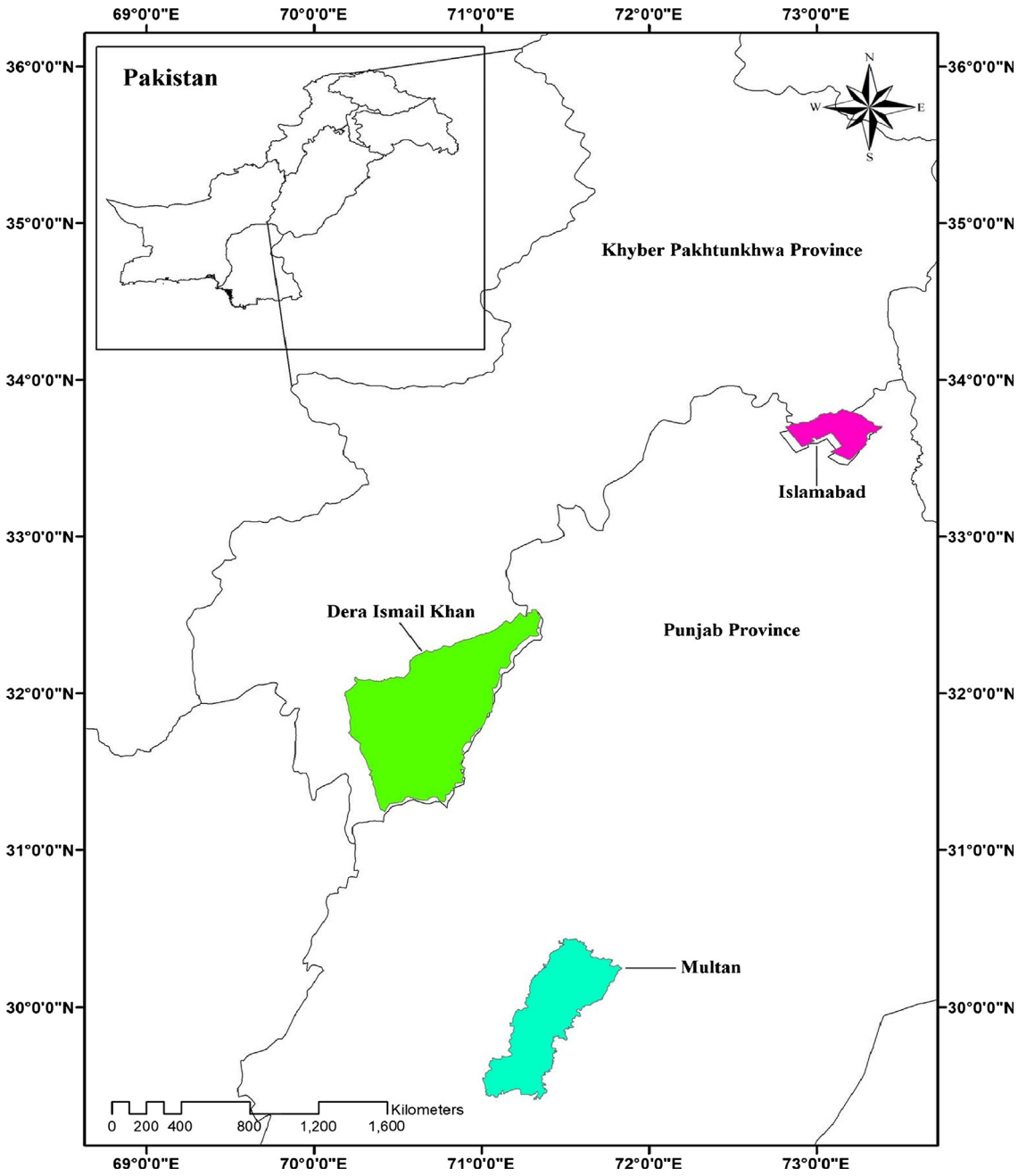


Fig. 3. Map of the study regions Islamabad, Multan and D. I. Khan in Pakistan.

(sigmoidal, sine, hardlim, triangular basis, radial basis) were tested to select the best (optimum) activation function for the optimum forecasting model.

Step 3: In this stage, the benckmark PSO-ANFIS and MLR models were also applied to forecast SPI index based on ELM, PSO-ANFIS and MLR.

Step 4: At this stage, the corresponding simulated SPI index on the basis of ELM, PSO-ANFIS and MLR models were incorporated together in the ELM model as a predictor varaibles to develop the committee bassed ELM (Comm-ELM) mode for the forecasting of final SPI index. Similarly, the Comm-PSO-ANFIS and Comm-MLR models were developed by using the same strategy adopted in the Comm-ELM model.

Step 5: The final forecasted SPI is obtained in stage 5. Fig. 4 summarizes these stages of the proposed Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR model.

Following (Deo et al., 2016c), in this paper, the data were partitioned into 70% and 30% for training and testing purposes. Normalization of input-target data was accomplished by statistical rules to overcome the numerical difficulties caused by the data features, patterns and fluctuations (Hsu et al., 2003). Mathematically, this is written as:

$$x_{SPI_{NORMALIZED}} = \frac{(x_{SPI} - x_{SPI_{MIN}})}{(x_{SPI_{MAX}} - x_{SPI_{MIN}})} \quad (17)$$

where x_{SPI} describes any datum point of input or output variable, $x_{SPI_{MIN}}$ is the minimum value of the whole dataset, $x_{SPI_{MAX}}$ is the maximum value, whereas $x_{SPI_{NORMALIZED}}$ denotes the normalized datum point.

In the development phase of the Comm-ELM model, a major task was to optimize effectively the network architecture between inputs and the objective variable for a robust predictive modelling system (Barzegar et al., 2018; Prasad et al., 2018). The correlation coefficient **'*r*'**, in combination with the mean squared error (***MSE***) was applied to

Table 1
Descriptive statistics of the study sites' geographic, drought and hydrologic characteristics over the study period (1981 to 2015).

Station	Geographic characteristics			Drought statistics: SPI					
	Longitude	Latitude	Elevation (m)	Mean	Std.	Min	Max	Skewness	Kurtosis
Islamabad	70.91°	31.83°	175.00	−0.0409	0.9781	−3.1445	2.9328	−0.0580	−0.1169
Multan	71.47°	30.19°	129.00	−0.0138	0.9888	−2.5380	4.4304	0.4934	1.7644
D. I. Khan	68.41°	26.24°	35.08	−0.0033	0.9830	−2.7480	3.50972	0.1228	0.1228

Hydrological statistics

Temperature (T)					Precipitation (PCN)				Humidity (H)			
	Mean (°C)	Std. (°C)	Min (°C)	Max (°C)	Mean (mm)	Std. (mm)	Min (mm)	Max (mm)	Mean (%)	Std. (%)	Min (%)	Max (%)
Islamabad	21.83	7.286	8.3	33.5	107.2	130.7	0.01	743.30	33.258	12.090	16	71
Multan	25.58	8.001	11	49.4	21.27	30.16	0.1	217.3	40.411	11.212	15	69
D. I. Khan	24.28	7.594	10.4	35.6	29.46	38.59	0.30	376.00	42.973	10.862	5	70

investigate the performance of the proposed Comm-ELM model in the training phase. The results generated by the Comm-ELM model have been summarised in Table 2.

The magnitudes of r and *MSE* attained in training of the Comm-ELM model for monthly SPI forecasting at Islamabad were seen to be: *(*r = 0.973, *MSE* = 0.115*)* with periodicity (M_5 model). Equivalent metrics for Multan were found to be: *(*r = 0.953, *MSE* = 0.213*)* model (M5). Finally, for D. I. Khan the metrics were: *(*r = 0.983, *MSE* = 0.077*)* model (M5). Overall, the training performance of the Comm-ELM model was considerably high for all of the study regions. It is thus envisaged that the Comm-ELM model testing performance, as seen later, will be relatively accurate for forecasting drought events at these sites.

The comparative forecasting model Comm-PSO-ANFIS was required to tune the parameters for fuzzy membership functions (Gaussian in this case) by the PSO technique to achieve the desired optimum model. After incorporating the forecasted SPI based on ELM, PSO-ANFIS and MLR again into the PSO-ANFIS in the form of inputs to develop the Comm-PSO-ANFIS model, the process of modifying the ANFIS parameters by PSO tuning is continued until the desired performance is achieved. The data was then passed in the ANFIS layer 1 to establish the ANFIS training process, within the training layer. The trained ANFIS was then passed through the testing layer for testing to attain the output (*i.e.*, the final forecasted SPI). For Islamabad, these metrics were seen to be: *(*r = 0.986, *MSE* = 0.091*)* with periodicity. Similarly for Multan, they are: *(*r = 0.923, *MSE* = 0.300*)* with periodicity and finally for D. I. Khan these were found to be: *(*r = 0.995, *MSE* = 0.031*)*.

Development of the Comm-MLR model depends on the inspection of the cause and effect association between forecasted SPI and predictor variables. The values of r and *MSE* for Islamabad were found to be: *(*r = 0.456, *MSE* = 0.882*)*. Equivalently, for Multan, they were: (*r* = 0882, *MSE* = 0.456) with periodicity and finally for D. I. Khan these were found to be: (*r* = 0.983, *MSE* = 0.099). Table 2 illustrates these comparisons.

3.4. Model performance measures and their interpretation

American Society for Civil Engineering (Yen, 1995) recommends two classification metrics for model assessment methodology that involve the factual (or visual correlation of the observed and forecasted information) and the institutionalized execution measurements. The mathematical formulations for these are as follows (Dawson et al., 2007; Deo et al., 2016c; Legates and McCabe, 1999; Willmott, 1981, 1982, 1984).

I. Mean square error (*MSE*) is expressed as:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(SPI_{FOR,i} - SPI_{OBS,i})^2 \quad (18)$$

II. Correlation coefficient (r) is expressed as:

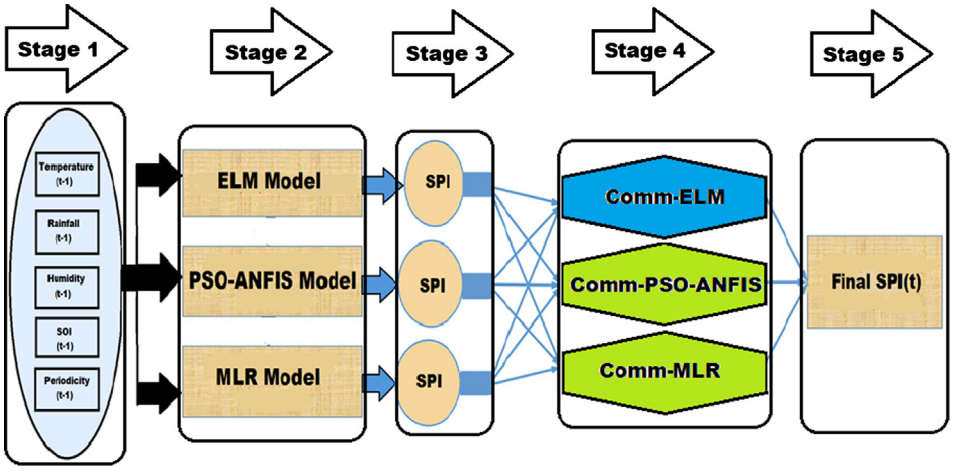


Fig. 4. Schematic of the proposed multi-stage, Comm-ELM mode (blue)l *vs.* Comm-PSO-ANFIS (green) and Comm-MLR (green) models. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2
Training performance of Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR models with correlation coefficient (*r*) and mean square error (*MSE*). [Definition of the model inputs as stated in Section 3.1 are: T = mean monthly temperature; PCN = Precipitation; H = mean relative humidity; SOI = Southern Oscillation Index and MP = monthly cycle or eriodicity].

Model	Input Combination	Comm-ELM											
		Islamabad				Multan				D. I. Khan			
		No. Hidden Neuron	Activation Function	Training period		No. Hidden Neuron	Activation Function	Training period		No. Hidden Neuron	Activation Function	Training period	
				MSE	r			MSE	r			MSE	r
M_1	T	60	Sine	0.188	0.479	60	Sigmoid	0.230	0.206	5	Sine	1.256	0.131
M_2	T + PCN	60	Sine	0.133	0.937	60	Sigmoid	0.087	0.957	5	Sine	0.982	0.599
M_3	T + PCN + H	60	Sine	0.200	0.947	60	Sigmoid	1.165	0.671	5	Sine	0.822	0.727
M_4	T + PCN + H + SOI	60	Sine	0.869	0.669	60	Sigmoid	0.569	0.637	5	Sine	0.620	0.711
M_5	T + PCN + H + SOI + MP	60	Sine	0.115	0.973	60	Triangular	0.213	0.953	5	Sine	0.077	0.983

Model	Input Combination	Comm-PSO-ANFIS									
		Design Parameters				Islamabad		Multan		D. I. Khan	
		ANFIS Type	Epoch	Clusters	Radii	Training period		Training period		Training period	
						MSE	r	MSE	r	MSE	r
M_1	T	TSK	100	auto	2	0.559	0.648	0.478	0.726	0.011	0.998
M_2	T+PCN	TSK	100	auto	2	0.013	0.998	0.024	0.994	1.211	0.787
M_3	T+PCN+H	TSK	100	auto	2	0.067	0.988	1.174	0.652	0.673	0.888
M_4	T+PCN+H+SOI	TSK	100	auto	2	1.014	0.813	0.709	0.808	0.782	0.869
M_5	T+PCN+H+SOI+MP	TSK	100	auto	2	0.091	0.986	0.300	0.923	0.031	0.995

Model	Input Combination	Comm-MLR																	
		Islamabad						Multan						D. I. Khan					
		c	α_1	α_2	α_3	Training period		c	α_1	α_2	α_3	Training period		c	α_1	α_2	α_3	Training period	
						MSE	r					MSE	r					MSE	r
M_1	T	0.260	0.834	0.079	−0.240	0.790	0.267	0.758	−0.348	−0.043	−0.200	0.962	0.369	0.038	0.004	−0.022	0.900	0.023	0.996
M_2	T+PCN	0.662	0.749	−0.019	−0.889	0.109	0.982	−0.664	0.777	0.072	1.252	0.103	0.974	−0.292	0.848	−0.103	0.655	1.933	0.608
M_3	T+PCN+H	0.974	0.593	−0.001	−1.091	0.316	0.949	0.183	0.135	0.040	0.456	1.594	0.475	−0.116	1.076	−0.123	0.336	1.019	0.817
M_4	T+PCN+H+SOI	0.769	−0.442	−0.166	0.120	1.997	0.611	0.157	0.304	−0.114	0.439	1.599	0.473	−0.176	1.157	−0.208	0.160	1.281	0.763
M_5	T+PCN+H+SOI+MP	−0.241	1.209	0.007	−0.020	0.282	0.958	−0.003	0.830	0.106	−0.097	0.456	0.882	−0.216	1.123	0.026	0.076	0.099	0.983

$$r = \left(\frac{\sum_{i=1}^{N} (SPI_{OBS,i} - \overline{SPI}_{OBS,i})(SPI_{FOR,i} - \overline{SPI}_{FOR,i})}{\sqrt{\sum_{i=1}^{N} (SPI_{OBS,i} - \overline{SPI}_{OBS,i})^2} \sqrt{\sum_{i=1}^{N} (SPI_{FOR,i} - \overline{SPI}_{FOR,i})^2}} \right), \; 0 \leqslant d \leqslant 1 \tag{19}$$

III. Willmott's Index (*WI*) is expressed as:

$$WI = 1 - \left[\frac{\sum_{i=1}^{N} (SPI_{FOR,i} - SPI_{OBS,i})^2}{\sum_{i=1}^{N} (|SPI_{FOR,i} - \overline{SPI}_{OBS,i}| + |SPI_{OBS,i} - \overline{SPI}_{OBS,i}|)^2} \right], \; 0 \leqslant d \leqslant 1 \tag{20}$$

IV. Nash-Sutcliffe coefficient (E_{NS}) is expressed as:

$$E_{NS} = 1 - \left[\frac{\sum_{i=1}^{N} (SPI_{OBS,i} - SPI_{FOR,i})^2}{\sum_{i=1}^{N} (\overline{SPI}_{OBS,i} - \overline{SPI}_{FOR,i})^2} \right], \; 0 \leqslant d \leqslant 1 \tag{21}$$

V. Root mean square error (*RMSE*) is expressed as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (SPI_{FOR,i} - SPI_{OBS,i})^2} \tag{22}$$

VI. Mean absolute error (*MAE*) is expressed as:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |(SPI_{FOR,i} - SPI_{OBS,i})| \tag{23}$$

VII. Legates-McCabe's (*LM*) is expressed as:

$$LM = 1 - \left[\frac{\sum_{i=1}^{N} |SPI_{FOR,i} - SPI_{OBS,i}|}{\sum_{i=1}^{N} |SPI_{OBS,i} - \overline{SPI}_{OBS,i}|} \right], \; 0 \leqslant d \leqslant 1 \tag{24}$$

where $SPI_{OBS,i}$ and $SPI_{FOR,i}$ are the observed and forecasted i^{th} value of multi-scale SPI. $\overline{SPI}_{OBS,i}$ and $\overline{SPI}_{FOR,i}$ are the observed and forecasted mean of SPI in the (test) set and N is the number of tested data points. The *MSE* indicates how close a regression line is to a set of points by calculating the distance from the points to the regression line and then squaring. The correlation coefficient (*r)* lies in [0, 1], and demonstrates the proportion of variance in observed data that can be explained by the data intelligent model (Dawson et al., 2007). Due to the standardization of the observed and forecasted means and variance, the robustness of *r* can be limited (Chai and Draxler, 2014). The 'goodness-of-fit' relevant to high values are measured by *RMSE*, while *MAE* evaluates all deviations from observed and forecasted data regardless of sign (Chai and Draxler, 2014). Moreover, *RMSE* is useful when model errors follow the normal distribution whereas *MAE* is better for uniform model error distribution (Chai and Draxler, 2014). The performance of a model can be reduced to partial peaks and higher magnitudes that can exhibit larger error and obtuse to small magnitudes (Dawson et al., 2007). Willmott's Index *(WI)* was introduced to counter this issue by considering the ratio of MSE instead of their differences (Mohammadi et al., 2015; Willmott, 1981, 1982, 1984; Willmott et al., 2012). Nash-Sutcliffe efficiency (E_{NS}) is another normalized metric that determines the relative magnitude of residual variance of forecasted data in comparison to the measured variance (Nash and Sutcliffe, 1970). Legates-McCabe's (*LM*) is a more advanced and powerful metric than both *WI* and E_{NS} which utilizes the adjustment of comparison in the evaluation of *WI* and E_{NS}. *LM* can be quite robust in evaluating the results by addressing the weaknesses of *r* and using *WI* and E_{NS} as baseline-adjusted indices together with an evaluation of *RMSE* and *MAE* (Legates and McCabe, 1999).

4. Results

A scatterplot diagram (*i.e.*, Fig. 5) was constructed using the forecasted and observed SPI data in the testing phase to compare the performance of the Comm-ELM *vs.* the Comm-PSO-ANFIS and Comm-MLR model with periodicity factors included for all three study sites. Further, to analyse the model performance more closely, a scatterplot showing the goodness-of-fit and its correlation coefficient (*r*) is shown to depict the extent of agreement between forecasted and observed monthly SPI. The Comm-ELM model convincingly outperforms the Comm-PSO-ANFIS and Comm-MLR model with periodicity in the testing months for all tested stations. Overall, the Comm-ELM model has a better ability to simulate with periodicity forecast monthly SPI, as confirmed by the larger *r*-value.

The empirical cumulative distribution function (ECDF) was plotted at each station for different forecasting abilities in Fig. 6(a–c). According to this figure, the Comm-ELM method was better than Comm-MLR and Comm-PSO-ANFIS for Islamabad, Multan and D. I. Khan stations to forecast SPI. Based on the percentage of errors in the error bracket (0 to ± 2) for all stations, Fig. 6(a–c) clearly confirms that the Comm-ELM method was the most responsive model in forecasting monthly SPI. Thus, the performance of Comm-ELM with periodicity is more accurate in drought forecasting for the selected regions of study.

In Table 3 we show the accuracy of Comm-ELM *vs.* Comm-PSO-ANFIS and the Comm-MLR model in the testing phase constructed by adding the historical temperature, rainfall, humidity, SOI and periodicity one by one as an input predictor to model current and future drought. Each table presents models M_1 to M_5 of Comm-ELM, Comm-PSO-ANFIS and Comm-MLR with periodicity. The model $\mathbf{M_5}$ with periodicity was considered the final (blue boldfaced) model.

By analysing the results of Islamabad, the Comm-ELM model performed better in terms of forecasting accuracy as compared to Comm-PSO-ANFIS and Comm-MLR models. By incorporating the periodicity with other predictor variables together, the Comm-ELM, Comm-PSO-ANFIS and Comm-MLR models yielded values of *RMSE* (0.307, 0. 0.674, 0.469) and *MAE* (0.237, 0.541, 0.369) while the magnitudes of standard normalized metrics are *r* (0.976, 0.946, 0.961) respectively. According to the results for Multan station, the Comm-ELM model returns the values of *RMSE* (0.441), *MAE* (0.296) and *r* (0.960). The Comm-PSO-ANFIS and Comm-MLR models generate the following results with periodicity, *RMSE* (Comm-PSO-ANFIS = 0.595, Comm-MLR = 0.391), *MAE* (Comm-PSO-ANFIS = 0.370, Comm-MLR = 0.319), *r* (Comm-PSO-ANFIS = 0.906, Comm-MLR = 0.967). The magnitude of performance metrics for the study region D. I. Khan for the Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR models yielded values of *RMSE* (0.372, 0.516, 0.415), *MAE* (0.279, 0.402, 0.324), *r* (0.973, 0.962, 0.972). The results of combination of other predictor variables are presented in Table 3. Overall, the Comm-ELM forecasts monthly SPI better for all study regions compared to Comm-PSO-ANFIS and Comm-MLR models.

Fig. 7 is a Taylor diagram, providing a more concrete and conclusive argument about the statistical summary of how well the forecasted SPI matched with the observed SPI in terms of their correlation. The similarity between forecasted and observed SPI is quantified in terms of their correlation and standard deviations. For Islamabad station, the correlation of the Comm-ELM model with observation was about 0.98, followed by Comm-MLR ≈ 0.96 and Comm-PSO-ANFIS ≈ 0.94. The Comm-ELM model was closer to the observed SPI as its correlation is about 0.96 as compared to Comm-PSO-ANFIS (0.94) and Comm-MLR (0.95) for Multan station. Similarly the Comm-ELM again appeared to be the best model for D. I. Khan station because its correlation lies within close neighbourhood of the observed SPI data. Overall, the correlation of the Comm-ELM model is closer to the observed SPI compared to Comm-PSO-ANFIS and Comm-MLR models for Islamabad, Multan and D. I. Khan stations.

Table 4 shows the accuracy of Comm-ELM *vs.* Comm-PSO-ANFIS and the Comm-MLR model in the testing period on the basis of LM in terms of forecasting accuracy. By incorporating the periodicity with other predictor variables together, the *LM* values for Islamabad are Comm-ELM (0.799), Comm-PSO-ANFIS (0.668) and Comm-MLR (0.748). The Comm-ELM, Comm-PSO-ANFIS and Comm-MLR models

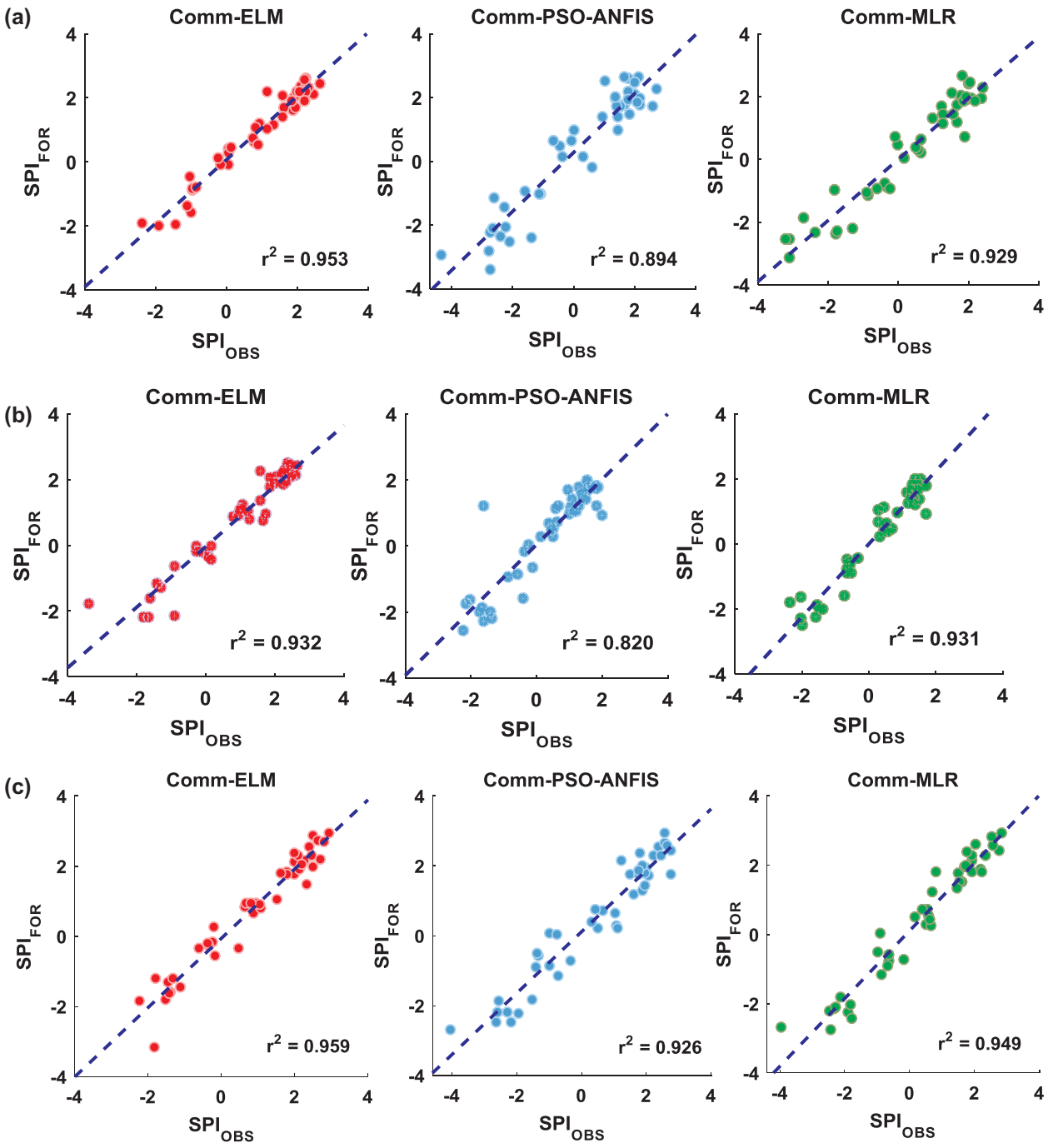


Fig. 5. Scatterplot of the forecasted (SPI_{FOR}) and observed (SPI_{OBS}) data in the testing phase using the multi-stage, Comm-ELM *vs*. Comm-PSO-ANFIS and Comm-MLR models with periodicity (i.e. month) as an input parameter with the coefficient of determination (r2) inserted in each panel for study zones (a) Islamabad, (b) Multan and (c) Dera Ismail Khan.

yielded values of *WI* (0.981, 0.952, 0.968) and E_{NS} (0.950, 0.868, 0.923) respectively. Again, the Comm-ELM model was superior for Islamabad compared to Comm-PSO-ANFIS and Comm-MLR models.

According to the results for Multan station, the Comm-ELM model returns the highest values of *WI* (0.963), E_{NS} (0.910), *and LM* (0.766). The Comm-PSO-ANFIS and Comm-MLR models generate the following results with periodicity, *WI* (Comm-PSO-ANFIS = 0.925, Comm-MLR = 0.965), E_{NS} (Comm-PSO-ANFIS = 0.821, Comm-MLR = 0.924) *and LM* (Comm-PSO-ANFIS = 0.692, Comm-MLR = 0.742). The performance of all three models is reasonably good but Comm-ELM generates better accuracy for this station.

The magnitude of performance metrics for the study region D. I. Khan for the Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR models yielded values of *WI* (0.975, 0.965, 0.977), E_{NS} (0.943, 0906, 0.944) *and LM* (0.791, 0.724, 0.786). The Comm-ELM again appeared to be the best model with both periodicity on the basis of *LM* followed by Comm-MLR and Comm-PSO-ANFIS.

In Fig. 8(a–c), we illustrate a boxplot of the Comm-ELM *vs*. Comm-PSO-ANFIS and Comm-MLR model's forecasting error for monthly SPI of all the study sites. The outliers specified by + in every boxplot represent the extreme magnitudes of the forecasting error within the testing months along with their upper quartile, median and lower quartile values. The distributed forecasting errors are justified by these boxplots showing a much lesser spread was achieved by Comm-ELM in

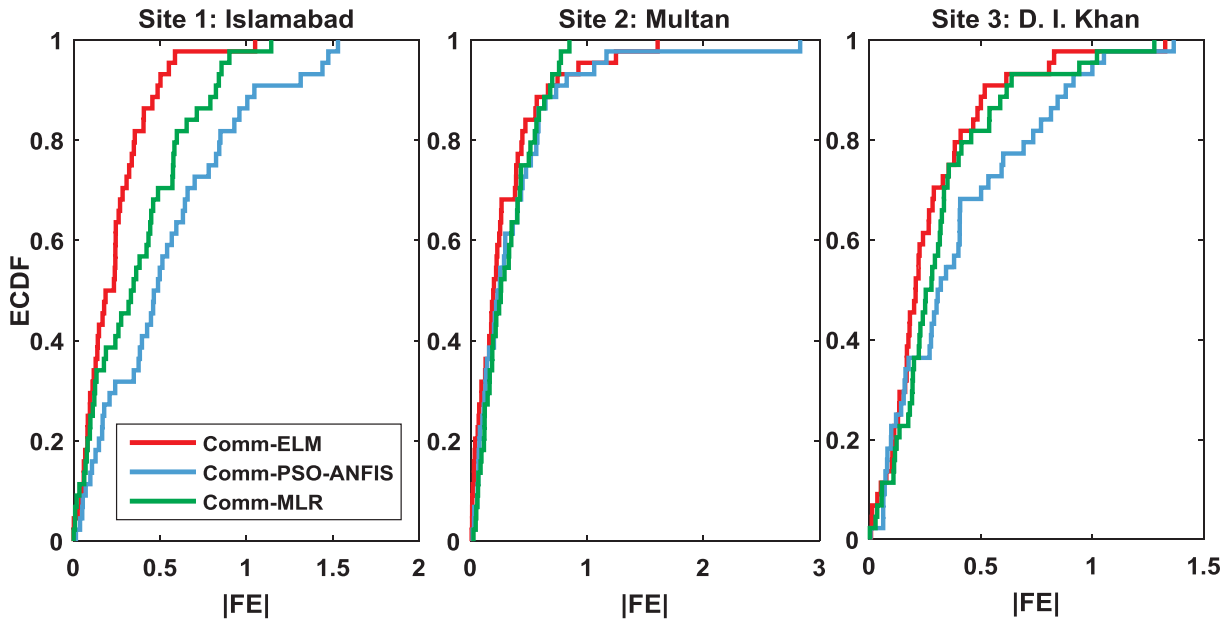


Fig. 6. Empirical cumulative distribution function (ECDF) of the forecast error, |FE| in the testing period using Comm-ELM *vs*. Comm-PSO-ANFIS and Comm-MLR models for the stations (a) Islamabad, (b) Multan and (c) Dera Ismail Khan.

Table 3
Influence of the model input combinations applied for forecasting of monthly SPI using the Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR models measured by root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (r). Note that the best model is boldfaced.

Model	Testing period	Islamabad								
		Comm-ELM			Comm-PSO-ANFIS			Comm-MLR		
	Input Combination	RMSE	MAE	r	RMSE	MAE	r	RMSE	MAE	r
M_1	T	1.276	0.630	-0.092	1.239	1.018	0.071	1.085	0.914	0.059
M_2	T + PCN	0.807	0.525	0.892	0.336	0.253	0.980	0.389	0.305	0.974
M_3	T + PCN + H	0.763	0.644	0.847	0.702	0.535	0.921	0.599	0.511	0.937
M_4	T + PCN + H + SOI	0.883	0.710	0.649	1.905	1.502	0.446	1.336	1.047	0.627
M_5	**T + PCN + H + SOI + MP**	**0.307**	**0.237**	**0.976**	0.674	0.541	0.946	0.469	0.369	0.961
		Multan								
M_1	T	0.218	0.137	0.266	0.936	0.663	-0.052	0.971	0.701	-0.157
M_2	T + PCN	0.329	0.240	0.942	0.973	0.374	0.768	0.902	0.326	0.832
M_3	T + PCN + H	1.050	0.792	0.715	1.235	1.017	0.506	1.287	1.109	0.446
M_4	T + PCN + H + SOI	0.958	0.744	0.638	1.829	1.408	0.178	1.262	1.065	0.538
M_5	**T + PCN + H + SOI + MP**	**0.441**	**0.296**	**0.960**	0.595	0.370	0.906	0.391	0.319	0.967
		D. I. Khan								
M_1	T	1.155	1.035	-0.095	0.234	0.189	0.990	0.234	0.185	0.993
M_2	T + PCN	0.919	0.698	0.649	1.512	1.188	0.574	1.692	1.416	0.396
M_3	T + PCN + H	0.732	0.500	0.842	1.181	0.927	0.723	1.232	1.054	0.726
M_4	T + PCN + H + SOI	1.013	0.612	0.471	1.389	1.123	0.665	1.777	1.341	0.376
M_5	**T + PCN + H + SOI + MP**	**0.372**	**0.279**	**0.973**	0.516	0.402	0.962	0.415	0.324	0.972

both periodicity and non-periodicity followed by Comm-MLR and Comm-PSO-ANFIS. Accordingly, the Comm-ELM model with both periodicity and non-periodicity remains the superior and highly optimized model compared to other two counterparts for all stations to forecast monthly SPI data.

Table 5 shows a geographical comparison of the proposed Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR models using relative root mean squared error (*RRMSE*) and relative mean absolute error (*RMAE*) for the different locations (Islamabad, Multan and D. I. Khan). Islamabad appears to be the most accurate station in forecasting Comm-ELM *(RRMSE* ≈32.14%*)* on the basis of *RRMSE*. On the other hand, D. I. Khan is the most accurate station by considering *RRMAE* (30.35%). In terms of site-averaged performance, the Comm-ELM model was found to yield the lowest relative percentage errors (*RRMSE, RMAE*).

Fig. 9 shows the magnitude of the average values of absolute forecasting errors |FE| accumulated over the monthly timescale for the proposed Comm-ELM *vs*. Comm-PSO-ANFIS and Comm-MLR models in the testing period. Although the forecasting skills generated by the Comm-ELM were slightly different than the Comm-PSO-ANFIS and Comm-MLR for all three stations. This performance was confirmed by the low magnitude of relative forecasted errors. For example, the relative error in January-February, May and July to December were significantly smaller for the Comm-ELM model as compared to Comm-PSO-ANFIS and Comm-MLR models in Islamabad station. The performance of Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR for Multan and D. I. Khan stations can be seen (Fig. 9, site 2 & 3). Overall, the Comm-ELM model provided better performance, including lower error statistics (Figs. 8 and 9) and higher correlation coefficients (Fig. 7) when it was coupled with periodicity.

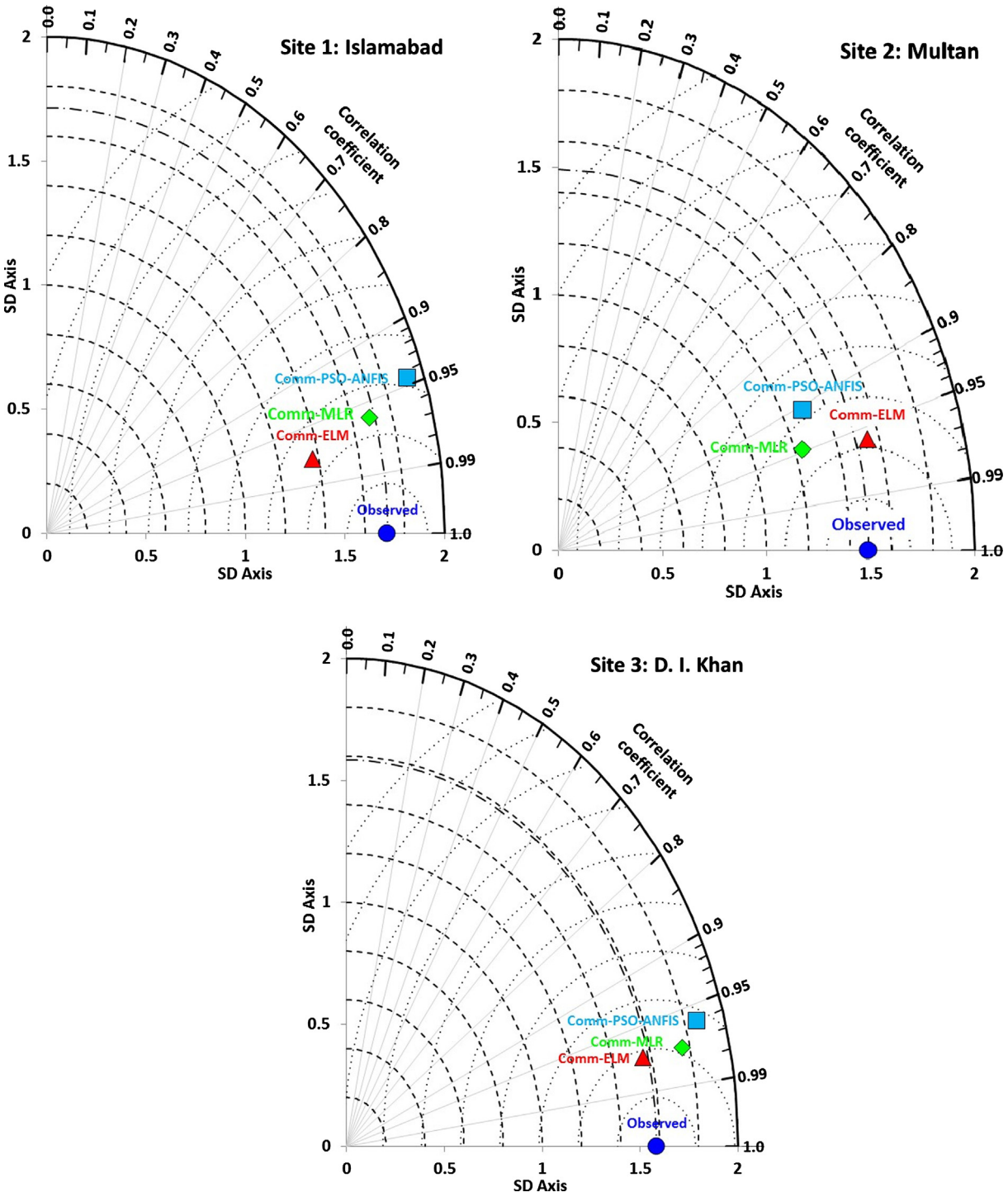


Fig. 7. Taylor diagram showing the correlation coefficient between observed and forecasted SPI and standard deviation of the proposed Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR models for the stations (a) Islamabad, (b) Multan and (c) Dera Ismail Khan.

5. Further discussion

Data-driven models are likely to become important decision-support tools in major decision-making for hydrology, agriculture and water science, aiming to address drought-related issues arising from the imminent global warming. The increasing risk of water scarcity remains a problem in first world nations as well as developing nations such as Pakistan. This work has developed the most accurate drought forecasting models to date for three locations in Pakistan utilizing Comm-ELM against the Comm-PSO-ANFIS and Comm-MLR models to forecast monthly SPI with consolidating environmental parameters as input predictors. The results reveal that Comm-ELM yields a substantial measurable accuracy between observed and forecasted SPI in light of the Legates-McCabe's Index (Table 4). The execution of the Comm-ELM showed a large value of LM, contrasted with the Comm-PSO-ANFIS and Comm-MLR models as far as the accomplished measurable precision. Along these lines, with periodicity (as also revealed in earlier studies elsewhere *e.g.*, (Deo et al., 2017a), the Comm-ELM proved to be an important prescient apparatus for forecasting drought in the present work.

Table 4
Evaluation of the Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR models using Willmott's index (WI), Nash-Sutcliffe (E_{NS}) and Legates-McCabe's (LM) agreement, for (a) Islamabad, (b) Multan; and (c) Dera Ismail Khan. Note that the best model is boldfaced.

Model	Testing period	Islamabad								
		Comm-ELM			Comm-PSO-ANFIS			Comm-MLR		
	Input Combination	WI	E_{NS}	LM	WI	E_{NS}	LM	WI	E_{NS}	LM
M_1	T	0.062	−1.904	−0.444	0.028	−0.288	−0.152	0.299	−0.157	−0.058
M_2	T + PCN	0.733	0.698	0.491	0.981	0.954	0.815	0.977	0.947	0.792
M_3	T + PCN + H	0.822	0.647	0.403	0.927	0.843	0.668	0.946	0.875	0.652
M_4	T + PCN + H + SOI	0.640	0.375	0.262	0.545	−0.146	0.039	0.469	0.378	0.286
M_5	**T + PCN + H + SOI + MP**	**0.981**	**0.950**	**0.799**	0.952	0.868	0.668	0.968	0.923	0.748
		Multan								
M_1	T	0.176	0.032	−0.135	0.139	−0.050	−0.027	−0.009	−0.287	−0.150
M_2	T + PCN	0.907	0.884	0.690	0.780	0.498	0.690	0.857	0.595	0.736
M_3	T + PCN + H	0.591	0.447	0.360	0.521	0.237	0.171	0.478	0.175	0.103
M_4	T + PCN + H + SOI	0.204	0.355	0.256	0.431	−0.681	−0.167	0.584	0.207	0.138
M_5	**T + PCN + H + SOI + MP**	**0.963**	**0.910**	**0.766**	0.925	0.821	0.692	0.965	0.924	0.742
		D. I. Khan								
M_1	T	0.215	−0.086	−0.073	0.991	0.979	0.866	0.993	0.982	0.878
M_2	T + PCN	0.340	0.415	0.239	0.629	0.196	0.184	0.461	0.065	0.065
M_3	T + PCN + H	0.799	0.694	0.560	0.751	0.510	0.363	0.726	0.504	0.305
M_4	T + PCN + H + SOI	0.272	0.126	0.294	0.701	0.321	0.228	0.444	−0.031	0.115
M_5	**T + PCN + H + SOI + MP**	**0.975**	**0.943**	**0.791**	0.965	0.906	0.724	0.977	0.944	0.786

In order to address existing difficulties, scientists must plan the future using adaptable decision-support models for accurate drought forecasting, especially in countries that are affected by recent environmental changes (2016; Ahmad et al., 2004; Amir, 2012; Pakistan, 1950-2015; Zaidi, 2016). The 1998 dry season in Pakistan was the most severe in the last 50 years, and it was an essential factor leading to poor economic development (Haider, 2016). The Baluchistan region, particularly the western and central parts were more often on drought throughout the year in Pakistan (Haider, 2016). Enhanced forecast of SPI is likely to help such locales in getting ready for expected changes in

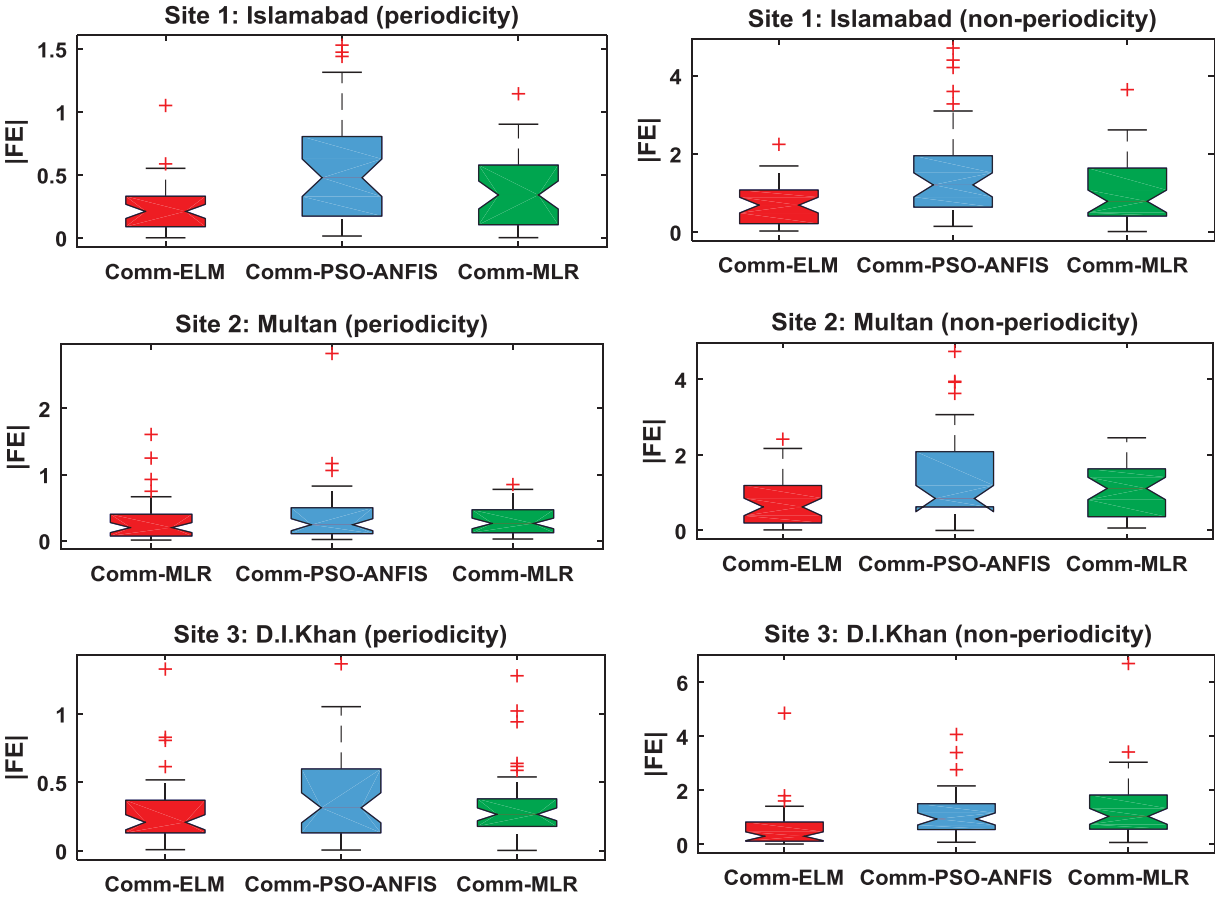


Fig. 8. Box-plots of forecasted error |FE| of Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR models forecasted monthly SPI index with periodicity (left) in comparison with non-periodicity (right) for Islamabad, Multan and Dera Ismail Khan.

Table 5
Geographic comparison of the study regions using Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR models using relative root mean squared error (*RRMSE*, %) and the relative mean absolute error (*RMAE*, %) computed within the test sites. Note that the best model is boldfaced.

Model	Testing period	Islamabad					
		Comm-ELM		Comm-PSO-ANFIS		Comm-MLR	
	Input Combination	RRMSE (%)	RMAE (%)	RRMSE (%)	RMAE (%)	RRMSE (%)	RMAE (%)
M_1	T	−237.68	2362.15	549.69	198.30	776.91	146.96
M_2	T + PCN	−46.27	36.61	83.15	74.11	102.68	26.79
M_3	T + PCN + H	−192.70	279.23	377.43	58.73	157.92	53.49
M_4	T + PCN + H + SOI	−366.82	144.04	353.35	152.33	352.42	96.65
M_5	**T + PCN + H + SOI + MP**	**32.14**	**33.16**	151.06	58.79	123.68	37.43
		Multan					
M_1	T	14.52	8.30	1256.02	131.06	−1254.30	224.63
M_2	T + PCN	−261.47	75.28	7221.28	62.48	314.58	29.81
M_3	T + PCN + H	−499.34	79.37	323.75	105.91	449.11	90.49
M_4	T + PCN + H + SOI	−266.33	75.51	477.80	192.61	440.22	87.58
M_5	**T + PCN + H + SOI + MP**	**58.04**	**72.60**	155.60	60.46	136.36	26.32
		D. I. Khan					
M_1	T	76.78	311.59	58.06	22.62	47.59	16.57
M_2	T + PCN	−110.74	55.79	285.83	259.95	344.62	139.79
M_3	T + PCN + H	−149.00	159.44	223.27	229.69	250.96	133.07
M_4	T + PCN + H + SOI	166.12	905.65	361.89	133.76	361.89	133.76
M_5	**T + PCN + H + SOI + MP**	**47.82**	**30.35**	84.47	80.94	84.48	80.94

precipitation. In prior studies, data-driven models based on climatic parameters found the variability of models over large, sparsely distributed regions (Abbot and Marohasy, 2012, 2014; Deo and Şahin, 2015; Deo and Sahin, 2016; Deo et al., 2016b; Xie et al., 2013). The development of models to target specific localities in this study is clearly beneficial and major contribution in improving drought prediction in Pakistan. The proposed Comm-ELM modelling approach concurs with earlier research performed by Barzegar et al. (2018); Prasad et al. (2018) in a sense where the authors used ELM, SVR, M5 Tree and MARS algorithms as a based models to design the committee based ANN model for groundwater contamination risk of multiple aquifers.

Regardless of the high precision found in predicting SPI using the models developed in this research, a few caveats require further discussion. In this paper, Comm-ELM was utilized through a periodicity procedure; however, the model advancement through hybridization of various data-intelligent models with ELM may improve the accuracy (Behmanesh et al., 2014; Liang et al., 2015; Vairappan et al., 2009; XingXing et al., 2008; Zhou et al., 2011). Some of the other advanced optimisation methods that enable more robust feature extraction could include: Particle Swarm Optimization (PSO) (Chen and Yu, 2005), Quantum-Behaved Particle Swarm Optimization (Q-PSO) (Zhisheng, 2010), Genetic Algorithm (GA) (Reeves, 1995) and Firefly Algorithms (FA) (Yang, 2010). These techniques have been tested in different climate applications and have been shown to be an improvement to the ELM model (Ghorbani et al., 2017; Raheli et al., 2017; Yaseen et al., 2017). These models could be implemented as hybrids with ELM in follow up work.

While the above-mentioned approaches are well-established, a new, more generalized framework for hybridizing the ANN with intuitionistic fuzzy (Takeuti and Titani, 1984) and neutrosophic logic (Smarandache, 2001) may be achieved by integrating with ELM. The latter is known to deal with vulnerability, indeterminacy, deficiency and irregularity in predictor target information since the standard statistical approaches tend to avoid the hurdle of model uncertainty. Bayesian Model Averaging (BMA) (Foresee and Hagan, 1997) is another data-intelligent tool to model uncertainties which can be used in ranking model performance. Multi-resolution analysis (*e.g.*, empirical wavelet transform (Gilles, 2013), empirical mode composition (Huang et al., 1998), maximum overlap wavelet (Kormylo and Mendel, 1982) and singular value decomposition (Golub and Reinsch, 1970) could broaden the accuracy and scope of this study. Moreover, the more recent techniques in rainfall and precipitation index forecasting tested for Pakistan (Ali et al., 2018a, 2018b) using multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm and ensemble ANFIS uncertainity model could also be adopted in SPI forecasting, particularly to check the ability of these data-driven techniques in context of the present problem.

6. Conclusion

Due to serious impacts of drought peril, the SPI based drought forecasting can be considered as a promising tool to deal with future hazards implicated on horticulture, water administration, water demand, pricing and policy. In this study, Comm-ELM model was benchmarked with two alternative techniques: the Comm-PSO-ANFIS and Comm-MLR based models where the monthly SPI forecasting model was constructed and evaluated for three agricultural regions in Pakistan where farming activity is a critical part of the country's economy. The developed models utilized the historical monthly atmospheric inputs (temperature, rainfall, humidity, SOI) at $(t-1)$ with respective periodicity as an improvement factor to forecast monthly SPI.

The results showed that utilizing the periodicity to forecast SPI led to an improvement in the exactness of the data-intelligent model. Comparison of the three models revealed that the Comm-ELM can be successfully applied for SPI forecasting, although there was a significant variation (in terms of the model forecasting accuracy) among the performance of these models with and without periodicity. That is, the Comm-ELM model attained the highest accuracy for Islamabad, Multan and D. I. Khan stations which was confirmed for both the training (Table 2) and the testing phase of the model (Tables 3–5).

While this study presented only a case study for Pakistan, the developed modelling framework can be extended to any location in the world where drought poses a catastrophic impact on national economy. For example, to increase the practicality of the Comm-ELM model for future drought forecasting, Comm-ELM model can be applied to several agricultural and water reservoir zones where water scarcity conditions are likely to threaten the long term sustainability of this resource. Due to the aforementioned potential and abilities of the Comm-ELM model, it is possible to also apply such a model to the forecasting of streamflow

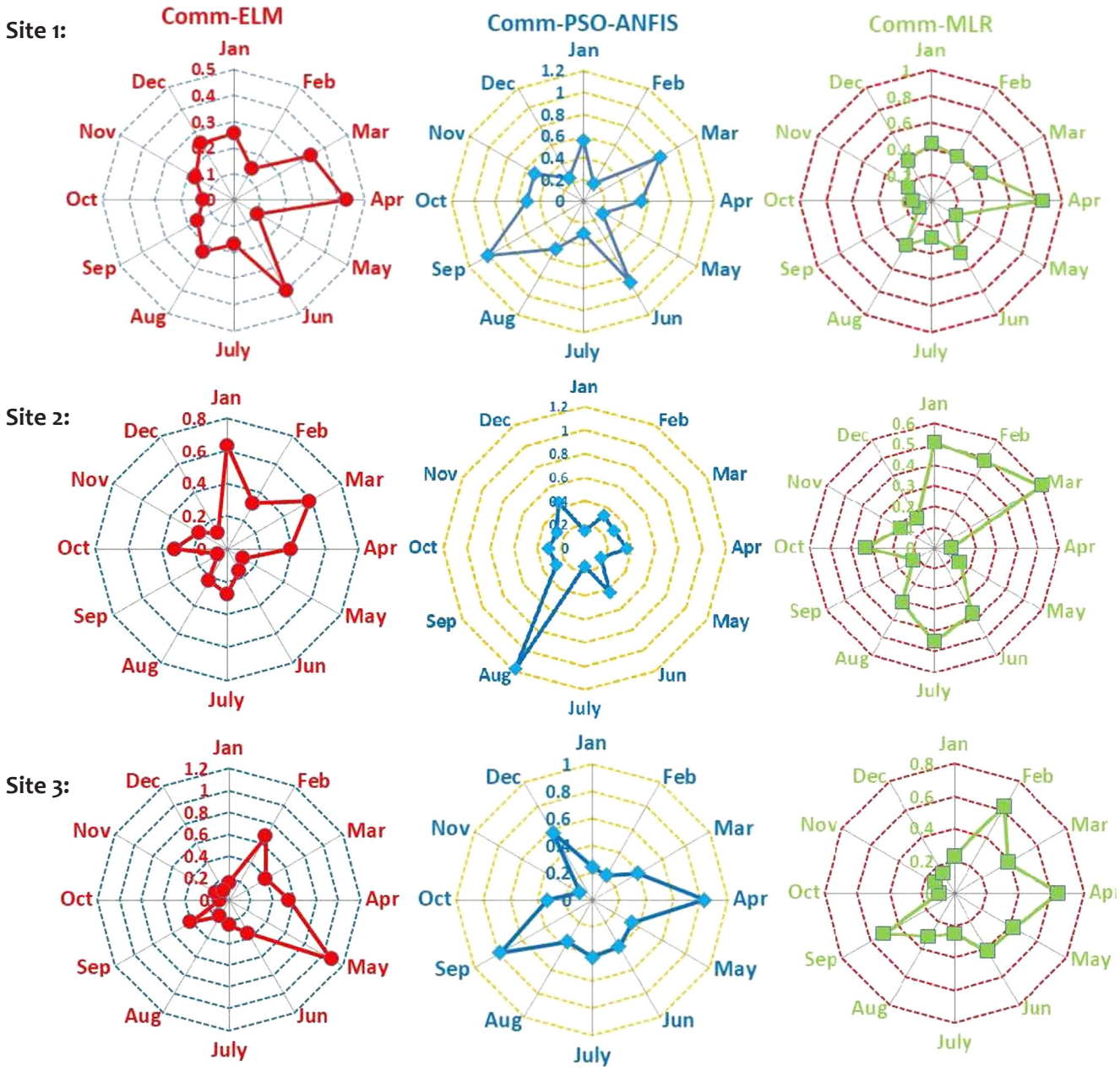


Fig. 9. Polar plots shows the monthly average values of |FE| generated from the Comm-ELM *vs.* Comm-PSO-ANFIS and Comm-MLR models in forecasting SPI for Site 1: Islamabad, Site 2: Multan and Site 3: D. I. Khan.

and river systems, including soil moisture and crop yield for agriculturally-dependent nations like Pakistan. Being a preliminary study in this drought-prone, agricultural region of Pakistan, this research has set a clear foundation for the potential of using more extensive predictor data products (e.g., satellite data with reanalysis and ground-based products) for forecasting future drought events in Pakistan and elsewhere, where the Comm-ELM model may be a preferable tool over the other comparative models.'

Acknowledgment

This study was supported by the University of Southern Queensland Office of Graduate Studies Postgraduate Research Scholarship (2017–2019). Data were acquired from Pakistan Meteorological Department, Pakistan, which are duly acknowledged. The authors express their sincere gratitude to all reviewers and the Editor for their constructive comments that has improved the final version of this paper.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.compag.2018.07.013.

References

2016. Climate change impacts on capital Islamabad. The Times of Islamabad.
Abbot, J., Marohasy, J., 2012. Application of artificial neural networks to rainfall forecasting in Queensland, Australia. Adv. Atmos. Sci. 29 (4), 717–730.
Abbot, J., Marohasy, J., 2014. Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. Atmos. Res. 138, 166–178.
Adamowski, J., Fung Chan, H., Prasher, S.O., Ozga-Zielinski, B., Sliusarieva, A., 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resources

Research. 48 (1).
Afzal, M., Haroon, M., Rana, A., Imran, A., 2013. Influence of North Atlantic oscillations and Southern oscillations on winter precipitation of Northern Pakistan. Pakistan J. Meteorol. 9 (18).
Ahmad, S., Hussain, Z., Qureshi, A.S., Majeed, R., Saleem, M., 2004. Drought Mitigation in Pakistan: Current Status and Options for Future Strategies. IWMI.
Ahmed, K., Shahid, S., bin Harun, S., Wang, X.-J., 2016. Characterization of seasonal droughts in Balochistan Province, Pakistan. Stochastic Environ. Res. Risk Assess. 30 (2), 747–762.
Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2018a. An ensemble-ANFIS based uncertainty assessment model for forecasting multi-scalar standardized precipitation index. Atmos. Res.
Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2018b. Multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm for rainfall forecasting. Atmos. Res. https://doi.org/10.1016/j.atmosres.2018.07.005.
Ali, Z., et al., 2017. Forecasting drought using multilayer perceptron artificial neural network model. Adv. Meteorol.
Almedeij, J., 2016. Long-term periodic drought modeling. Stochastic Environ. Res. Risk Assess. 30 (3), 901–910.
Amir, I., 2012. Tough times ahead for farmers in southern KP. DAWN.
Apaydın, A., Kutsal, A., Atakan, C., Ankara, 1994. The Statistics in practice. Hacettepe Pub.
Barzegar, R., Moghaddam, A.A., Deo, R., Fijani, E., Tziritis, E., 2018. Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms. Sci. Total Environ.. 621, 697–712.
Bates, B., Kundzewicz, Z.W., Wu, S., Palutikof, J., 2008. Climate change and water: Technical paper vi. Intergovernmental Panel on Climate Change (IPCC).
Behmanesh, M., Mohammadi, M., Naeini, V.S., 2014. Chaotic time series prediction using improved ANFIS with imperialist competitive learning algorithm. Int. J. Soft Comput. Eng. 4 (4), 25–33.
Bezdek, J.C., Ehrlich, R., Full, W., 1984. FCM: the fuzzy c-means clustering algorithm. Comput. Geosci.. 10 (2–3), 191–203.
Bonaccorso, B., Bordi, I., Cancelliere, A., Rossi, G., Sutera, A., 2003. Spatial variability of drought: an analysis of the SPI in Sicily. Water Resources Manage. 17 (4), 273–296.
Byun, H.-R., Wilhite, D.A., 1999. Objective quantification of drought severity and duration. J. Clim. 12 (9), 2747–2756.
Cancelliere, A., Bonaccorso, B., Mauro, G., 2006. A non-parametric approach for drought forecasting through the standardized precipitation index. Metodi statisticie matematici per l Analisi delle serie idrologiche. 1 (1), 1–8.
Cancelliere, A., Di Mauro, G., Bonaccorso, B., Rossi, G., 2007. Drought forecasting using the standardized precipitation index. Water Resour. Manage. 21 (5), 801–819.
Çavdar, T., 2016. PSO tuned ANFIS equalizer based on fuzzy C-means clustering algorithm. AEU-Int. J. Electron. Commun. 70 (6), 799–807.
Chai, T., Draxler, R.R., 2014. Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. Geoscientific Model Develop. 7 (3), 1247–1250.
Chen, G.-C., Yu, J.-S., 2005. Particle swarm optimization algorithm. Inform. Control-Shenyang-. 34 (3), 318.
Choubin, B., Malekian, A., Golshan, M., 2016. Application of several data-driven techniques to predict a standardized precipitation index. Atmósfera 29 (2), 121–128.
Civelekoglu, G., Yigit, N., Diamadopoulos, E., Kitis, M., 2007. Prediction of bromate formation using multi-linear regression and artificial neural networks. Ozone: Sci. Eng. 29 (5), 353–362.
Dawson, C.W., Abrahart, R.J., See, L.M., 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ. Modell. Software. 22 (7), 1034–1052.
Deo, R.C., Kisi, O., Singh, V.P., 2017a. Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. Atmos. Res. 184, 149–175.
Deo, R.C., Sahin, M., 2017. Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland. Renew. Sustain. Energy Rev. 72, 828–848.
Deo, R.C., Sahin, M., 2015. Application of the artificial neural network model for prediction of monthly standardized precipitation and evapotranspiration index using hydrometeorological parameters and climate indices in eastern Australia. Atmos. Res. 161-162, 65–81.
Deo, R.C., Sahin, M., 2016. An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland. Environ. Monitor. Assess. https://doi.org/10.1007/s10661-016-5094-9.
Deo, R.C., Byun, H.-R., Adamowski, J.F., Begum, K., 2016a. Application of effective drought index for quantification of meteorological drought events: a case study in Australia. Theor. Appl. Climatol. 1–21.
Deo, R.C., Sahin, M., 2016. An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland. Environ. Monitor. Assess. 188 (2), 1.
Deo, R.C., Tiwari, M.K., Adamowski, J., Quilty, J., 2016b. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. Stoch. Env. Res. Risk Assess. https://doi.org/10.1007/s00477-00016-01265-z.
Deo, R.C., Tiwari, M.K., Adamowski, J.F., Quilty, J.M., 2017b. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. Stochastic Environ. Res. Risk Assess. 31 (5), 1211–1240.
Deo, R.C., Wen, X., Qi, F., 2016c. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Appl. Energy. 168, 568–593.
Deo, R.C., Syktus, J., McAlpine, C., Lawrence, P., McGowan, H., Phinn, S.R., 2009. Impact of historical land cover change on daily indices of climate extremes including droughts in eastern Australia. Geophys. Res. Lett. 36 (8), 10.
Department, P.M., 2010. Dry weather predicted in the country during Friday/Monday.
Draper, N., Smith, H., 1981. Applied Regression Analysis. John Wiley, New York, pp. 709.
Dunn, J.C., 1973. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters.
Foresee, F.D., Hagan, M.T., 1997. Gauss-Newton approximation to Bayesian learning. Neural networks, 1997., international conference on, IEEE, 1930–1935.
Ghorbani, M.A., Deo, R.C., Yaseen, Z.M.K., Mahasa, H., Mohammad, B., 2017. Pan evaporation prediction using a hybrid multilayer perceptron-firefly algorithm (MLP-FFA) model: case study in North Iran. Theoret. Appl. Climatol. https://doi.org/10.1007/s00704-017-2244-0.
Gilles, J., 2013. Empirical wavelet transform. IEEE Trans. signal Process. 61 (16), 3999–4010.
Golub, G.H., Reinsch, C., 1970. Singular value decomposition and least squares solutions. Numerische mathematik. 14 (5), 403–420.
Goyal, M.K., Bharti, B., Quilty, J., Adamowski, J., Pandey, A., 2014. Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS. Expert Systems with Appl. 41 (11), 5267–5276.
Guttman, N.B., 1999. Accepting the standardized precipitation index: a calculation algorithm. JAWRA J. Am. Water Resources Assoc. 35 (2), 311–322.
Haider, N., 2016. Living with disasters.
Hayes, M.J., Svoboda, M.D., Wilhite, D.A., Vanyarkho, O.V., 1999. Monitoring the 1996 drought using the standardized precipitation index. Bull. Am. Meteorol. Soc.. 80 (3), 429–438.
Hoffmann, F., Schauten, D., Holemann, S., 2007. Incremental evolutionary design of TSK fuzzy controllers. IEEE Trans. Fuzzy Systems 15 (4), 563–577.
Hsu, C.-W., Chang, C.-C., Lin, C.-J., 2003. A practical guide to support vector classification.
Huang, G.-B., 2003. Learning capability and storage capacity of two-hidden-layer feedforward networks. IEEE Trans. Neural Networks. 14 (2), 274–281.
Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2006. Extreme learning machine: theory and applications. Neurocomputing 70 (1), 489–501.
Huang, N.E., et al., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. In: Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences. The Royal Society, pp. 903–995.
IPCC, 2012. Summary for Policymakers: A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change. in: Field, C.B. et al. (Eds.), In: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation Cambridge University Press, Cambridge, UK.
Jalalkamali, A., Moradi, M., Moradi, N., 2015. Application of several artificial intelligence models and ARIMAX model for forecasting drought using the Standardized Precipitation Index. Int. J. Environ. Sci. Technol. 12 (4), 1201–1210.
Jang, J.-S., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Systems, Man, Cybernet. 23 (3), 665–685.
Jang, J.-S.R., Sun, C.-T., Mizutani, E., 1997. Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence.
Karthika, B., Deka, P.C., 2015. Prediction of air temperature by hybridized model (Wavelet-ANFIS) using wavelet decomposed data. Aquatic Procedia. 4, 1155–1161.
Khan, M., Gadiwala, M., 2013. A Study of drought over Sindh (Pakistan) using standardized precipitation index (SPI) 1951 to 2010. Pakistan J. Meteorol. 9 (18).
Kormylo, J., Mendel, J., 1982. Maximum likelihood detection and estimation of Bernoulli-Gaussian processes. IEEE Trans. Inform. Theory 28 (3), 482–488.
Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241.
Liang, Z., Xie, B., Liao, S., Zhou, J., 2015. Concentration degree prediction of AWJ grinding effectiveness based on turbulence characteristics and the improved ANFIS. Int. J. Adv. Manuf. Technol. 80.
Loukas, A., Vasiliades, L., 2004. Probabilistic analysis of drought spatiotemporal characteristics inThessaly region, Greece. Natural Hazards Earth System Sci. 4 (5/6), 719–731.
Mayilvaganan, M., Naidu, K., 2011. Comparison of membership functions in adaptive-network-based fuzzy inference system (ANFIS) for the prediction of groundwater level of a watershed. J. Comput. Appl. Res. Dev. 1, 35–42.
McAlpine, C., et al., 2007. Modeling the impact of historical land cover change on Australia's regional climate. Geophys. Res. Lett. 34 (22).
McAlpine, C., Syktus, J., Ryan, J., Deo, R.C., McKeon, G., McGowan, H., Phinn, S., 2009. A continent under stress: interactions, feedbacks and risks associated with impact of modified land cover on Australia's climate. Global Change Biol. 15 (9), 2206–2223.
McKee, T.B., Doesken, N.J., Kleist, J., 1993. In: The relationship of drought frequency and duration to time scales. American Meteorological Society, Boston, MA, pp. 179–183.
Mishra, A.K., Singh, V.P., 2010. A review of drought concepts. J. Hydrol. 391 (1), 202–216.
Mishra, A.K., Singh, V.P., 2011. Drought modeling–A review. J. Hydrol. 403 (1), 157–175.
Mohammadi, K., et al., 2015. A new hybrid support vector machine–wavelet transform approach for estimation of horizontal global solar radiation. Energy Convers. Manage. 92, 162–171.
Montgomery, D.C., Peck, E.A., Vining, G.G., 2012. Introduction to linear regression analysis. John Wiley & Sons.
Moosavi, V., Vafakhah, M., Shirmohammadi, B., Behnia, N., 2013. A wavelet-ANFIS hybrid model for groundwater level forecasting for different prediction periods. Water Resour. Manage. 27 (5), 1301–1321.
Moreira, E., Martins, D., Pereira, L., 2015. Assessing drought cycles in SPI time series using a Fourier analysis. Nat. Hazards Earth Syst. Sci. 15 (3), 571–585.

Moreira, E.E., Coelho, C.A., Paulo, A.A., Pereira, L.S., Mexia, J.T., 2008. SPI-based drought category prediction using loglinear models. J. Hydrol. 354 (1), 116–130.
Morid, S., Smakhtin, V., Bagherzadeh, K., 2007. Drought forecasting using artificial neural networks and time series of drought indices. Int. J. Climatol. 27 (15), 2103–2111.
Mpelasoka, F., Hennessy, K., Jones, R., Bates, B., 2008. Comparison of suitable drought indices for climate change impacts assessment over Australia towards resource management. Int. J. Climatol. 28 (10), 1283–1292.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. J. Hydrol. 10 (3), 282–290.
Nayak, P.C., Sudheer, K., Rangan, D., Ramasastri, K., 2004. A neuro-fuzzy computing technique for modeling hydrological time series. J. Hydrol. 291 (1), 52–66.
Nguyen-Huy, T., Deo, R.C., An-Vo, D.-A., Mushtaq, S., Khan, S., 2017. Copula-statistical precipitation forecasting model in Australia's agro-ecological zones. Agric. Water Manag. 191, 153–172.
Nicholls, N., 2004. The changing nature of Australian droughts. Climatic Change 63 (3), 323–336.
Ozdamar, K., 2004. The Statistical Data Analysis with Software Packages. Kaan press, Eskisehir.
Özger, M., Mishra, A.K., Singh, V.P., 2012. Long lead time drought forecasting using a wavelet and fuzzy logic combination model: a case study in Texas. J. Hydrometeorol. 13 (1), 284–297.
Pakistan, D.i., 1950-2015. List of natural disasters in Pakistan.
Palmer, W.C., 1965. Meteorological drought. US Department of Commerce, Weather Bureau Washington, DC.
Palmer, W.C., 1968. Keeping track of crop moisture conditions, nationwide: the new crop moisture index.
Paulo, A.A., Pereira, L.S., 2007. Prediction of SPI drought class transitions using Markov chains. Water Resources Manage. 21 (10), 1813.
Pérez, E.C., Algredo-Badillo, I., Rodríguez, V.H.G., 2012. Performance analysis of ANFIS in short term wind speed prediction. arXiv preprint arXiv:1212.2671.
PMD, 2016. Pakistan Meteorological Department, Pakistan.
Prasad, R., Deo, R.C., Li, Y., Maraseni, T., 2018. Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydrometeorological predictors. Soil Tillage Res. 181, 63–81.
Raheli, B., Aalami, M.T., El-Shafie, A., Ghorbani, M.A., Deo, R.C., 2017. Uncertainty assessment of the multilayer perceptron (MLP) neural network model with implementation of the novel hybrid MLP-FFA method for prediction of biochemical oxygen demand and dissolved oxygen: a case study of Langat River. Environ. Earth Sci. 76 (14), 503.
Rajesh, R., Prakash, J.S., 2011. Extreme learning machines-a review and state-of-the-art. Int. J. Wisdom Based Comput. 1 (1), 35–49.
Reeves, C.R., 1995. A genetic algorithm for flowshop sequencing. Comput. Oper. Res. 22 (1), 5–13.
Report, 1950-2015. List of natral disaster.
Riebsame, W.E., Changnon Jr, S.A., Karl, T.R., 1991. Drought and natural resources management in the United States. Impacts and implications of the 1987–89 drought. Westview Press Inc.
Şahin, M., Kaya, Y., Uyar, M., 2013. Comparison of ANN and MLR models for estimating solar radiation in Turkey using NOAA/AVHRR data. Adv. Space Res. 51 (5), 891–904.
Salman, A., 2006. Blistering heatwave claims 33 lives.
Santos, C.A.G., Morais, B.S., Silva, G.B., 2009. Drought forecast using an artificial neural network for three hydrological zones in San Francisco River basin, Brazil. IAHS Publication 333, 302.
Sehgal, V., Sahay, R.R., Chatterjee, C., 2014. Effect of utilization of discrete wavelet components on flood forecasting performance of wavelet based ANFIS models. Water Resour. Manage. 28 (6), 1733–1749.
Shirmohammadi, B., Moradi, H., Moosavi, V., Semiromi, M.T., Zeinali, A., 2013. Forecasting of meteorological drought using Wavelet-ANFIS hybrid model for different time steps (case study: southeastern part of east Azerbaijan province, Iran). Nat. Hazards. 69 (1), 389–402.
Smarandache, F., 2001. A unifying field in logics: neutrosophic logic. math/0101228.
Sönmez, F.K., KÖmÜscÜ, A.Ü., Erkan, A., Turgu, E., 2005. An analysis of spatial and temporal dimension of drought vulnerability in Turkey using the standardized precipitation index. Nat. Hazards. 35 (2), 243–264.
Sujitha, E., 2017. Analysis of dry/wet conditions using the Standardized Precipitation Index and its potential usefulness for drought/flood monitoring in the regions of Trichy. J. Pharmacognosy Phytochem. 6 (4), 452–457.
Svoboda, M., Hayes, M., Wood, D., 2012. Standardized precipitation index user guide. World Meteorological Organization Geneva, Switzerland.
Takeuti, G., Titani, S., 1984. Intuitionistic fuzzy logic and intuitionistic fuzzy set theory. J. Symbol. Logic 49 (3), 851–866.
Tamura, S.i., Tateishi, M., 1997. Capabilities of a four-layered feedforward neural network: four layers versus three. IEEE Trans. Neural Networks 8 (2), 251–255.
Vairappan, C., Tamura, H., Gao, S., Tang, Z., 2009. Batch type local search-based adaptive neuro-fuzzy inference system (ANFIS) with self-feedbacks for time-series prediction. Neurocomputing 72 (7), 1870–1877.
Wilhite, D.A., Hayes, M.J., 1998. Drought planning in the United States: Status and future directions. In: The arid frontier. Springer, pp. 33–54.
Wilhite, D.A., Hayes, M.J., Knutson, C., Smith, K.H., 2000. Planning for drought: Moving from crisis to risk management1. Wiley Online Library.
Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2 (2), 184–194.
Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc.. 63 (11), 1309–1313.
Willmott, C.J., 1984. On the evaluation of model performance in physical geography. In: Spatial statistics and models. Springer, pp. 443–460.
Willmott, C.J., Robeson, S.M., Matsuura, K., 2012. A refined index of model performance. Int. J. Climatol. 32 (13), 2088–2094.
Xie, H., Ringler, C., Zhu, T., Waqas, A., 2013. Droughts in Pakistan: a spatiotemporal variability analysis using the Standardized Precipitation Index. Water Int. 38 (5), 620–631.
XingXing, W., XiLin, Z., HuiXiao, Y., 2008. Application of improved ANFIS in optimization of machining parameters. Chinese J. Mech. Eng. 44 (1), 199–204.
Yang, X.-S., 2010. Firefly algorithm, stochastic test functions and design optimisation. Int. J. Bio-Inspired Comput. 2 (2), 78–84.
Yaseen, Z.M., et al., 2017. Rainfall pattern forecasting using novel hybrid intelligent model based ANFIS-FFA. Water Resour. Manage. 32, 105–122.
Yen, B.C., 1995. Discussion and closure: criteria for evaluation of watershed models. J. Irrig. Drain. Eng. 121 (1), 130–132.
Yevjevich, V., 1967. An Objective Approach to Definitions and Investigations of Continental Hydrologic Droughts. Colorado State University Fort Collins.
Yevjevich, V., 1991. Tendencies in hydrology research and its applications for 21st century. Water Resour. Manage. 5 (1), 1–23.
Yuan, W.-P., Zhou, G.-S., 2004. Comparison between standardized precipitation index and Z-index in China. Acta Phytoecologica Sinica. 4.
Zaidi, H.B., 2016. Expect more extreme weather this year. Dawn.
Zhisheng, Z., 2010. Quantum-behaved particle swarm optimization algorithm for economic load dispatch of power system. Expert Syst. Appl. 37 (2), 1800–1803.
Zhou, Q., et al., 2011. A new method to obtain load density based on improved ANFIS. Power Syst. Protect. Control 39 (1), 29–34.

Chapter 6

Improving SPI-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridized with simulated annealing and kernel ridge regression algorithms

Foreword

This chapter is an exact copy of the published article in the *Journal of Hydrology* (Volume 576, Pages 164-184).

To address the limitations in ensemble-ANFIS (Chapter 4) and Comm-ELM (chapter 5), there is a desire to develop an artificial intelligence model that can forecast drought at a shorter, medium and longer forecast horizon. Therefore, new multivariate empirical mode decomposition (MEMD) integrated with simulated annealing (SA) and Kernel ridge regression (KRR) models to devise a hybrid MEMD-SA-KRR. The MEMD modeling approach allows the utilization of multiple predictor inputs to decompose into signals (i.e. IMFs). The SA algorithm selects best IMFs that are later used in KRR model to forecast 1-, 3-, 6- and 12-month drought at three diverse geographic location in Pakistan. The MEMD based models address the non-stationarity and non-linearity issues within the multiple predictor inputs.The MEMD and SA hybridization with RF models leads to MEMD-SA-RF. The forecasting accuracy of MEMD-SA-KRR is better over MEMD-SA-RF, Standalone KRR and Standalone RF models.

Journal of Hydrology 576 (2019) 164–184

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Improving SPI-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridized with simulated annealing and kernel ridge regression algorithms

Mumtaz Ali[1], Ravinesh C. Deo*, Tek Maraseni, Nathan J. Downs

School of Agricultural, Computational and Environmental Sciences, Centre for Applied Climate Sciences and Centre for Sustainable Agricultural Systems, Institute for Life Sciences and the Environment, University of Southern Queensland, Springfield, QLD 4300, Australia

ARTICLE INFO

This manuscript was handled by G. Syme, Editor-in-Chief, with the assistance of Jesús Mateo-Lázaro, Associate Editor

Keywords:
Hybrid drought
Forecast model
Multivariate empirical mode decomposition
Simulated annealing
Kernel ridge regression

ABSTRACT

New and improved drought models based on the World Meteorological Organization approved Standardized Precipitation Index, principally at multiple timescale horizons, are providing significant benefits to the hydrological community, by its widespread acceptance in the sub-field of water resources management, sustainable water use and precision agriculture. In this research paper, the existing challenges faced by a drought forecasting model trained at multiple time-scales are resolved where a new multi-phase, multivariate empirical mode decomposition model integrated with simulated annealing and Kernel ridge regression algorithms (*i.e.*, MEMD-SA-KRR) is designed to attain significantly accurate drought forecasts for 3 agricultural sites (*i.e.*, Faisalabad, Islamabad and Jhelum, located in Pakistan). Utilizing the multi-scalar Standardized Precipitation Index (SPI) time series as a target variable for characterization of drought, twelve multivariate datasets (derived from statistically significant lagged combinations of precipitation, temperature & humidity), that are enriched with eight synoptic-scale climate mode indices and periodicity, are utilized in designing a new drought model. The study constructs a hybrid MEMD-SA-KRR model, where firstly, the data are partitioned into their respective training and testing subsets after creating historically lagged SPI at timescale ($t-1$). Secondly, the MEMD algorithm is conditioned to demarcate multivariate climate indices from their training and testing sets, separately, into their decomposed intrinsic mode functions (IMFs) and residues. Thirdly, the SA method is employed to decide the most suitable IMFs. Finally, the KRR algorithm is applied to the selected IMFs to forecast multi-scaler SPI, at 1-, 3-, 6- and 12-monthly forecast horizon. The results are benchmarked with Random Forest, integrated with MEMD and SA to develop the MEMD-SA-RF equivalent model. The multi-phase MEMD-SA-KRR model is tested geographically in Pakistan, revealing that the MEMD-SA-KRR hybrid model generates reliable performance in forecasting multi-scaler SPI series, relative to the comparative models based on error analysis metrics. The hybrid drought model incorporating the most pertinent synoptic-scale climate drivers, as the model inputs has significant implications for hydrological applications and water resources management including its potential use in drought policy and drought recovery plans.

1. Introduction

Drought is characterized as a climatological menace that can occur in arid, semi-arid, or tropical rainforest zones (Keyantash and Dracup, 2002; Vicente-Serrano, 2016; Wilhite et al., 2000). Drought events can last from short to long period ranging from one month to four years as recent climate change significantly affects rainfall patterns (Vicente-Serrano, 2016). Drought severely disturbs water resources, agriculture crops, energy supply and industrial sectors, and it is a growing concern (Deo et al., 2009; IPCC, 2012; McAlpine et al., 2007; Yaseen et al., 2018b). Long term drought significantly poses challenges to groundwater reservoirs and also cause significant water scarcity (Cai and Cowan, 2008) and the related socio-economic costs (Dijk et al., 2013; Wittwer et al., 2002). Hydrologists and government and non-

* Corresponding author.
E-mail addresses: Mumtaz.Alli@usq.edu.au, mumtaz.ali@deakin.edu.au (M. Ali), ravinesh.deo@usq.edu.au (R.C. Deo).
[1] Deakin-SWU Joint Research Centre on Big Data, School of Information Technology, Deakin University, 221 Burwood Highway, Burwood 3125 Vic, Australia.

https://doi.org/10.1016/j.jhydrol.2019.06.032
Received 13 March 2019; Received in revised form 28 May 2019; Accepted 12 June 2019
Available online 14 June 2019
0022-1694/

government based policy makers planning to develop new strategies to manage drought risks for water management (Bates et al., 2008; Deo et al., 2017c; Mishra and Singh, 2011; Mouatadid et al., 2018).

Hydrological decision support systems are considerably as important standards in characterizing different watersheds and their scenarios (Chen et al., 2013; Gebremariam et al., 2014; Romagnoli et al., 2017; Sommerlot et al., 2016), but this model's accuracy can rest on the physics, initial conditions, and calibrations based on conceptual (*i.e.*, regional-scale) sub-models and the spatio-temporal aspects of the input and output variables (He et al., 2014; Sun et al., 2012). Recently, artificial intelligence models, that do not require any details about the physics of the watershed or any associated hydrological behaviors, are seen to have a good ability to forecast precipitation, run-off, streamflow and drought events (Ali et al., 2018b; Deo et al., 2017a; Deo and Şahin, 2016; Deo et al., 2017a; Deo et al., 2019; Mouatadid et al., 2018; Prasad et al., 2017; Yaseen et al., 2018a). Such models that can rely purely on how climate variables and related large-scale indices change over time can provide great insights into a drought forecasting and forewarning system.

Drought can be quantified by large-scale climate indices in different geographic zones, which is adopted to monitor drought intensity (Mishra and Singh, 2010; Mishra and Singh, 2011). Traditionally, the Palmer Drought Severity Index (PDSI) was developed to monitor drought (Palmer, 1965). The PDSI is beneficial in handling lengthy drought period but is not recommended for of high run-off region (Mishra and Singh, 2010; Mishra and Singh, 2011). Crop Moisture Index (CMI) is developed to handle the complexities in PDSI especially to quantify agronomic droughts (Palmer, 1968). CMI uses the rainfall record rankings to determine the positive and negative precipitation anomalies which are useful for the short-term offset. Byun and Wilhite (1999) designed the Effective Drought Index (EDI) for meteorological and agricultural drought events which only based on precipitation. The SPI is designed to overcome these hurdles and is a worldwide acceptable standard metric to monitor and investigate drought scenarios.

Modelling SPI, as advocated in this research study, is beneficial for future drought assessment, as follows: (1): This index can monitor water deficiency scenarios that based on statistical rainfall distribution ranging from monthly to seasonal and annual. (2): It is considered to be a globally acceptable, standardized metric dependent on normalized rainfall deficits (Hayes et al., 1999; McKee et al., 1993; Yuan and Zhou, 2004), that models the drought behavior in a climatological diverse regions (Almedeij, 2016; Choubin et al., 2016; Svoboda et al., 2012). (3) Handling multiple timescales of drought probabilistically, the SPI is an instrument to inspect soil moisture condition (Svoboda et al., 2012).

Artificial intelligence models have been adopted to develop drought forecasting strategies based on SPI utilizing several environmental parameters: (1) Santos et al. (2009) and Jalalkamali et al. (2015) designed a multilayer perceptron artificial neural network to forecast the SPI in US and Iran, respectively; and (2) Wavelet based adaptive neuro fuzzy inference system were developed to estimate SPI by Shirmohammadi et al. (2013) in Azerbaijan. For the details of SPI based drought forecasting, please refer to (*e.g.*, (Adamowski et al., 2012; Ali et al., 2018b; Bonaccorso et al., 2003; Cancelliere et al., 2006; Choubin et al., 2016; Deo et al., 2017c; Guttman, 1999; Hayes et al., 1999; Jalalkamali et al., 2015; Moreira et al., 2015; Moreira et al., 2008; Paulo and Pereira, 2007; Sönmez et al., 2005)).

In this research work, we develop an SPI based hybrid artificial intelligence model for a drought-prone region in Pakistan where historical drought events have severely hindered socio-economic and agricultural production (Report, 1950–2015). To address some of the challenges faced due to drought events, Khan and Gadiwala (2013) has analyzed drought patterns using SPI, while Xie et al. (2013) has forecasted SPI at several spatio-temporal scales in Pakistan. Recently, Ahmed et al. (2016) has characterized seasonal drought using SPI in Balochistan, Pakistan and Ali et al. (2018a) modelled drought events based on lagged data of SPI using adaptive neuro-fuzzy inference system based ensemble (ANFIS-ensemble) approach. Similarly, Ali et al. (2018b) has utilized some of the primary climate-based datasets to forecast the SPI series. However, studies on future drought models utilizing different synoptic-scale climate mode indices are still very limited, particularly in the agricultural region of Pakistan.

Owing to the variability in climate-based for a drought model, a suite of multi-resolution analytical tools can be useful to extract embedded features in a non-static time series signal that are related to a drought variable, and thus, they may help to improve an existing drought model. To resolve this challenge, the empirical mode decomposition (EMD) method formalized by Huang et al. (1998), can provide a useful alterative tool to improve existing drought models, as it is able to isolate the largely fluctuating signals into their respective smaller, and more clearly resolved frequency components to improve a drought model. Since its inception, the EMD method has gained attention due to its self-adaptability (Alvanitopoulos et al., 2014). The EMD is completely data dependent; thus making it greatly useful to extract relevant features without any loss of information. Further, the decomposed prominent features preserve the physical structure of the input temporally (Wu et al., 2011). Due to its capability to improve the forecasting accuracy of artificial intelligence models, the EMD algorithm has been integrated with artificial neural network (ANN) model, proving it to be a successful tool to forecast environmental variables such as solar radiation (Alvanitopoulos et al., 2014). For instance, Wang et al. (2018) trialed EMD and local mean decomposition integrated with least squares support vector machine to forecast solar radiation.

In spite of its recent applicability, a key issue with EMD and its variant algorithm(s) is that it can only applicable to decompose a univariate data (Colominas et al., 2014; Torres et al., 2011; Wu and Huang, 2009)), for example, only the significant antecedent lagged dataset of SPI can be used to forecast the future drought series. This is a major limitation of the EMD algorithm since the variation in drought events is immensely reliant on dynamically-driven climatological factors such as atmospheric circulation, so the incorporation of the relevant synoptic-scale climate indices (*e.g.*, SOI) that modulate drought events is extremely important. Therefore, these input variables need to be intelligently used into the artificial intelligence model, in addition to the antecedent drought index series. The recent work of Ali et al. (2018a), which appears to be the only study that has attempted to forecasted the medium and long period droughts, has used the significant lags of SPI to forecast the future SPI series. Further, Ali et al. (2018b) has employed climate dataset to forecast SPI only over a short term (one month) period.

Following earlier studies (Ali et al., 2018a; Ali et al., 2018b) on short, medium and long term drought forecasting, this study aims to operate large-scale climate indices and climate predictors to subsequently extract most, if not all, pertinent features where an MEMD (Rehman and Mandic, 2009a) based hybridized modelling approach is developed to forecast multi-scaler SPI index. The MEMD is an advance generalized form of EMD and CEEMD which demarcates multivariate inputs to performs accurate investigation of composite and nonlinear procedures (Rehman and Mandic, 2009a). Additionally the MEMD fixes the mode alignment problems arise in the joint analysis within a multi-dimensional data (Looney and Mandic, 2009). The applicability of MEMD can be seen clearly in forecasting of evapotranspiration (Adarsh et al., 2017), soil water (Hu and Si, 2013), crude oil price (He et al., 2016), solar radiation (Prasad et al., 2019) and iceberg drift (Andersson et al., 2017), yet this the present research is a pilot application of this novel technique in drought forecasting particularly for the agricultural region of Pakistan. Further, this paper follows earlier methodology (Quilty and Adamowski, 2018), aiming to partitioned data prior developing hybrid MEMD-SA-KRR model. This approach is used to circumvent the issues associated with other kinds of decomposition-based models that were raised in earlier studies (Quilty and Adamowski, 2018).

The novelty of this research study lies in the enhanced capability of

the MEMD algorithm to address the non-stationarity issues encountered in the model design process, by using concurrent transformation of model inputs into their decomposed components that are likely to improve an existing drought forecasting model (as revealed later in results section). The primary issues related to the selection of best IMFs (*i.e.*, patterns in drought model input series), which are not known *a priori*, are determined by the proposed approach, by an implementation of a robust feature selection process: Simulated Annealing (SA) algorithm. Hence, a careful integration of MEMD and SA algorithm with Kernel Ridge Regression (KRR) is made to generate a much improved hybrid forecast model, denoted as the MEMD-SA-KRR. The model is also benchmarked against MEMD-SA-RF and a standalone (i.e., KRR and RF) model to forecast multi-scaler SPI at 1-, 3-, 6 and 12-months tested at 3 drought-rich locations.

2. Theoretical framework

A summary of the artificial intelligence model based on KRR, MEMD and SA approaches with its comparative models is now presented.

2.1. Kernel ridge regression (KRR) model

KRR model is a machine learning model based on kernels and a ridge regression approach (Zhang et al., 2013), which is used to deal with over-fitting in the regression using regularization and the kernel technique to capture non-linear relationships (You et al., 2018). Mathematically, KRR can be formulated as;

$$\arg\min \frac{1}{q}\sum_{o=1}^{q} ||f_o - y_o||^2 + \lambda\ ||f||_H^2 \tag{1}$$

$$f_o = \sum_{p=1}^{q} \alpha_p \Phi(x_p, x_o) \tag{2}$$

where $||.||_H$ is the Hilbert normed space in Eq. (1) (Zhang et al., 2013). For a given $m - by - m$ kernel matrix, K is constructed by $K_{p,o} = \Phi(x_p, x_o)$ from selected input data where y is the input q-by-1 regressand vector, and α is the q-by-1 unknown solution vector. Equation (2) reduces to the following.

$$(K + \lambda qI) = y \tag{3}$$

$$\widetilde{y} = \sum_{o=1}^{q} \alpha_o \Phi(x_o, \widetilde{x}) \tag{4}$$

The KRR method in the model training stage is approximated α by solving Eq. (3) whereas this α is used in the testing phase to predict the regression of unknown sample $\widetilde{x}$ in Eq. (4). Further, the KRR algorithm searches for optimum α and λ from the parameter set. In KRR, linear, polynomial and Gaussian kernels are used to get the optimum accuracy (Alaoui and Mahoney, 2015; Vovk, 2013; Welling, 2013; You et al., 2018). Mathematically, linear, polynomial and Gaussian kernel are defined as:

$$\Phi(x_p, x_o) = x_p^T.\ x_o \tag{5}$$

$$\Phi(x_p, x_o) = (x_p^T.\ x_o + r)^d \tag{6}$$

$$\Phi(x_p, x_o) = \exp(-||x_p - x_o||^2/(2\sigma^2)) \tag{7}$$

where T represents the transpose and d is the dimension of the vector.

2.2. Multivariate empirical mode decomposition (MEMD) method

The MEMD technique is able to fix the issues of mode mixing which handle the drawback of exhaustiveness and time consuming. To understand the mathematical structure of MEMD, we need to study the EMD theory. The EMD is described as:

$$\Gamma(s) = \sum_{k=1}^{l} C_k(s) + R_l(s) \tag{8}$$

where $\Gamma(s)$, $C_k(s)$ and $R_l(s)$ are representing input data, the k^{th} IMF and remainder (residue) respectively. The MEMD formulated by Rehman and Mandic (2009b) uses multivariate data to decompose into multiple dimensions of IMFs to avoid the issues of mode mixing incorporating White Gaussian noise (Ur Rehman and Mandic, 2011). The mean $M(s)$is derived as following:

$$M(s) = \frac{1}{p}\sum_{q=1}^{p} e^{\theta_q}(s) \tag{9}$$

where $e^{\theta_q}(s)$ is referred to envelop curves with length of the vectors$s$.

$$R(s) = \Gamma(s) - M(s) \tag{10}$$

The term $R(s)$ is called a multi-dimensional IMF satisfying the stopping criterion. The MEMD algorithm has been utilized in analyzing signal processing (Huang et al., 2013; Mandic et al., 2013), and hydrology (Hu and Si, 2013; She et al., 2015).

The MEMD is a self-adaptive algorithm which makes no assumptions a priori about the composition of the signal (Hu and Si, 2013; She et al., 2015). The MEMD uses spline interpolation between maxima and minima to successively trace out IMFs where each IMF is a single periodic oscillator (Huang et al., 2013; Mandic et al., 2013). The IMFs cannot be predicted prior it is empirically observed from the signal. Since the IMFs can change over time, MEMD makes no assumptions about the stationarity of the signal (or the signal components) and is therefore better suited to non-linear signals when analyzing signals from complex systems (Huang et al., 2013; Mandic et al., 2013).

2.3. Simulated annealing (SA) model

The SA is a bio-inspired feature selection algorithm to find a suitable solution to an optimization technique (Elleithy and Fattah, 2012). The SA is an adaptive non-deterministic algorithm which has been extensively used as optimization technique such as traveling salesman (Peng et al., 1996), computer generated holograms (Taniguchi et al., 1997), power efficiency (Wilson, 1997) and heat exchangers (Athier et al., 1997). The basic algorithm of SA involves the following steps.

1. Create a randomly appropriate solution.
2. Determine the cost of solution using some cost function.
3. Create another random neighbouring solution.
4. Compute again cost of the above new solution.
5. If the new solution cost is less than the old solution cost, then move to the new solution otherwise go to the following step 6.
6. Follow again stages 3–5 until an optimum solution is determined.

2.4. Random forest (RF)

Bootstraping and bagging is basically ensemble learning techinques which creates classfiers and sums the final outcomes in terms of decision trees (Breiman, 1996; Schapire et al., 1998). The RF model is fundamentally a decision tree model which adopts randomly a bagging approach in the forecasting scenarios (Breiman, 1996). Each node is separated randomly by chosing preeminent possible predictors to improve accuracy that are robust to avoid overfitting (Breiman, 2001). The steps followed in the designing of RF can be given as:

1. Generate bootstrapping of n_{trees} by incorporating the predictors variables with n denotes the number of trees.
2. The randomly input predictors sample m_{try} is created to choose maximum predictors splitting by growing an unpruned regression tree.
3. Cumulate the aggregative predictions of n_{trees} to forecast multi-

scaler SPI.

The applicabilty of RF model can be seen in soil attribute prediction (Moore et al., 1993), hydrology (Moore et al., 1991), environmental management (Ascough Ii et al., 2008), drought forecasting (Chen et al., 2012), solar index estimation (Deo et al., 2017b), rainfall forecasting (Ali et al., 2018c) and most recently, forecasting soil moisture (Prasad et al., 2018).

For more comprehensive studies on RF model, readers are referred to (Breiman, 2001; Liaw and Wiener, 2002; Prasad et al., 2018; Robert et al., 1998; Segal, 2004).

2.5. Multi-scale Standardized precipitation index (SPI)

The SPI quantifies the wet and dry scenarios based on statistical probability theory. Prior to design the proposed multi-phase MEMD-SA-KRR model, the multi-scaler SPI index was computed by incorporating precipitation (PTCN) data (McKee et al., 1993) in the following Equation (8).

$$g(PTCN) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}(PTCN)^{\alpha-1}e^{-x/\beta} \tag{11}$$

where *g(PTCN)* indicates the probability density function, α and β are the parameters determined by the maximum likelihood solution whereas Γshows the gamma function. Further, the cumulative probability is defined as:

$$G(PTCN) = \int_0^P g(PTCN)dR_F = \frac{1}{\beta^{\alpha}\Gamma(\alpha)}\int_0^P x^{\alpha-1}e^{-P/\beta}d(PTCN) \tag{12}$$

By substituting $n = PTCN/\beta$ in Eq. (13):

$$G(PTCN) = \frac{1}{\Gamma(\alpha)}\int_0^n n^{\alpha-1}e^{-n}dn \tag{13}$$

The cumulative probability reduces to the following form when $PTCN = 0$:

$H(PTCN) = p + (1- p)\ G\ (PTCN)$, (14)with p represents the probability of zero which determines the SPI index, viz:

$$SPI = \begin{cases} +\left(n - \frac{\varepsilon_0+\varepsilon_1 n+\varepsilon_2 n^2}{1+\omega_1 n+\omega_2 n^2+\omega_3 n^3}\right), & 0.5 < H(PTCN) \le 1.0 \\ -\left(n - \frac{\varepsilon_0+\varepsilon_1 n+\varepsilon_2 n^2}{1+\omega_1 n+\omega_2 n^2+\omega_3 n^3}\right), & 0 < H(PTCN) \le 0.5 \end{cases} \tag{15}$$

where ε_0, ε_1, ε_2, ε_3, ω_1, $\omega_2$and ω_3 are arbitrary constants with magnitudes: $\varepsilon_0 = 2.515517$, $\varepsilon_1 = 0.802853$, $\varepsilon_3 = 0.010328$, $\omega_1 = 1.432788$, $\omega_2 = 0.189269$ and $\omega_3 = 0.001308$ (McKee et al., 1993). Drought is characterized into three categories as moderate = ($-1.5 <$ SPI ≤ 1.0), severe = ($-2.0 < SPI \le -1.5$), and extreme = (SPI ≤ -2.0).

3. Materials and method

3.1. Data

Twelve meteorological data series, precipitation (PTCN), temperature (T), humidity (H), southern oscillation index (SOI), sea surface temperatures (Nino3SST, Nino3.4SST, Nino4SST), pacific decadal oscillation (PDO), Indian ocean dipole (IOD), El-Nino southern oscillation Modoki index (EMI), southern annular mode (SAM) and periodicity at a monthly (t) lag of $t - 1$ are acquired from Pakistan Meteorological Department (PMD, 2016), National Climate Prediction Centre (Nicholls, 2004; SST, 2018), Joint Institute of the Study of the Atmosphere and Ocean (JISAO, 2018), Bureau of Meteorology, Australia (BMA, 2018), Japan Agency for Marine-Earth Science (JAMSTEC, 2018) and from the British Antarctic Survey (BAS, 2018). Any precipitation less than 0.1 mm was replaced with the corresponding averaged value for the climatological period. The multi-scaler SPI index was computed in R-programing software using the rainfall time series data from 1981 to 2015.

Despite inherent complexities associated with accurate forewarning of drought, synoptic-scale climate mode indices, that are strongly correlated with drought occurrence (Mishra and Singh, 2010; Morid et al., 2007; Özger et al., 2012), can provide a consensus on the overall behaviour of drought events,as these indices can have a significant influence on rainfall and streamflow patterns (Ali et al., 2018c; Andreoli and Kayano, 2005; Chiew et al., 1998; Deo and Şahin, 2016; McBride and Nicholls, 1983; McGregor et al., 2014; Nicholls, 1983; Prasad et al., 2017; Yaseen et al., 2018a; Yaseen et al., 2018b). For example, the well-known association of the Inter-decadal Pacific Oscillation over the Pacific Ocean is seen to influence the El-Nino Southern Oscillation (ENSO) phenomena that governs the intensity of drought events (Dai, 2013; Salinger et al., 2001). The negative period of SOI brings El-Nino episodes whereas the positive values of SOI launches La-Nina events (Adnan et al., 2017; Philander, 1983). The Northern Atlantic Oscillation (NAO) also has a significant influence from seasonal to inter-decadal variability on atmosphere and environmental variables (Dickson et al., 2000; Hurrell, 1995; Souriau and Yiou, 2001). The IOD across the eastern Indian Ocean carries hefty showers over east Africa while drought and forest fires across the Indonesian zone (Adnan et al., 2017; Ashok et al., 2001). Furthermore, Sea Surface Temperatures (Nino3SST, Nino3.4SST, Nino4SST) over the southeast Indian Ocean may also lead to heavy precipitation (Priya et al., 2015; Terray et al., 2007). The ENSO Modoki index (EMI) based on strong anomalous warming in the central tropical Pacific and cooling in the eastern and western tropical Pacific, potentially impacts the temperature and rainfall patterns around the globe due to ocean atmosphere and the unique tri-polar sea level pressure pattern (Ashok et al., 2007). Similarly, the SAM significantly influences the monsoon rainfall anomaly (Pal et al., 2017).

3.2. Study locations

The study locations utilized in this work are: Faisalabad, Islamabad and Jhelum displayed in Fig. 1. The geographical, climatological and drought statistics of these locations is described in Table 1. Further, it also presents the multi-scaler SPI index

Faisalabad, is categorized as desert with average yearly rainfall of 375 mm and temperature 24.2 °C, and it is located in the grasslands of northeast Punjab (Table 1) (Servey, 2016). Major crops growing in Faisalabad are wheat, rice, cotton, sugarcane, maize, different vegetables and fruits.

Islamabad is the capital city of Pakistan. It experiences a subtropical climate with four seasons: winter, spring, summer and autumn, with average monsoon and annual precipitations of 790.8 mm and 1142.1 mm, respectively. The heaviest precipitation in Islamabad was 620 mm/month recorded in July 2002.

Jhelum is situated in the Pothohar region of the Punjab province, Pakistan. Agriculture highly depends on precipitation in Punjab. The average annual rainfall is about 1000 mm in the rainy season of monsoon (Department, 2010; PMD, 2016).

3.3. Development of multi-phase MEMD-SA-KRR model

The multi-phase MEMD-SA-KRR model is developed in MATLAB R2016b programming environment (The Math Works Inc. USA). All the simulations were obtained operating Pentium 4, 2.93 GHz dual-core Central Processing Unit. The data are partitioned straightly 60% and 40% prior into training and testing subsets respectively following by (Quilty and Adamowski, 2018) as it is the most common approach for data partitioning (Cannas et al., 2006). The first 21 years from 1981 to 2001 data were used to train the MEMD-SA-KRR model while the remaining 14 years data (2002–20015) were utilized for testing. Moreover, the cross-validation or any data randomized approach cannot be adopted as time-series data by definition occur in a temporal order/sequence and this order or sequence must be preserved in order to keep

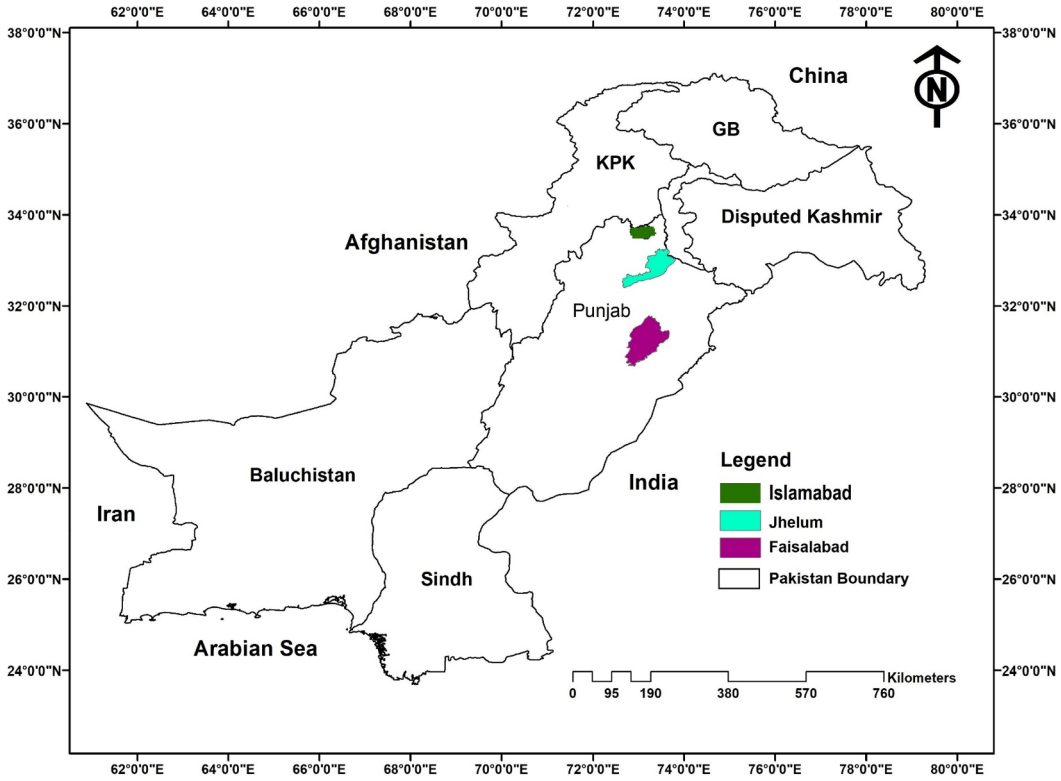


Fig. 1. Map of the selected study locations in Pakistan.

the structure of the series intact (Bergmeir and Benítez, 2012). Antecedent time time lagged inputs (*i.e.*, PTCN, T, H, SOI, Nino3SST, Nino3.4SST, Nino4SST, PDO, IOD, EMI, SAM, and periodicity) at $(t-1)$ are used to develop the hybrid model to enable SPI_1 (1-month), SPI_3(3-month), SPI_6 (6-month) and SPI_{12} (12-month) forecasts as elicudated in the following steps.

3.3.1. Phase 1: MEMD process

The MEMD method is applie to demarcates the input time-series variables into IMFs and residuals. The input variables were incoporated in (PTCN + T + H + SOI + Nino3SST + Nino3.4SST + Nino4SST + PDO + IOD + EMI + SAM + periodicity) for decompostion by the MEMD algorithm. Additionally, the predefined parameters include the ensemble number ($N = 500$) and the amplitude of the added white noise ($\varepsilon = 0.2$) (Ouyang et al., 2016; Ren et al., 2015; Wang et al., 2013; Wu and Huang, 2009). To acquire the same number of IMFs in training and testing period, the MEMD method is controlled using total projection, a stop vector (i.e. tolerance and threshold values) and stopping criterion (see Table 2). Total one hundred-twenty IMFs with residuals (Table 3) for multi-scaler SPI in Faisalabad sites is extracted with each input has (IMFs = 9, residual = 1) whereas for Islamabad, this number is ninety-six with each input has (IMFs = 7, residual = 1). For Jhelum, the MEMD algorithm disolved the input predictors into (IMFs = 8, residual = 1, No. of total IMFs = 108) in case of SPI_1, SPI_6, SPI_{12} while (IMFs = 7, residual = 1, No. of total IMFs = 96) for SPI_3.

The cross correlation is a linear method which determine the linear relationship between input predictor and target data. We avoided this method and used SA method instead to select from a pool of variables as the multiple climate input predictors have significant non-linearity. Moreover, we have 12 input predictors, and decomposed by MEMD into several sub-components, the resulting number of inputs is very large (i.e. approx. 120). This means if we use a manual method (e.g. cross correlation), there is a large amount of work but it uses linear method, and so, will not generate smart way of selecting the best inputs. Instead, we have used SA (which is a non-linear approach, so this method is more suitable than correlation analysis). We have identified this issue and revised the paper.

3.3.2. Phase 2: SA algorithm

The SA approach is adopted to choose the most appropriate IMFs for only the training period using a feature selection strategy for model development. Further, some parameters were defined prior that includes the number of maximum iterations (=20) and the parameter initial temperature (=10). The number of selected best IMFs (feature) is kept 25 which were defined prior to run the SA model. The selected IMFs for each training period are described in Table 3. For testing periods, those IMFs are used for model validation, based on selected IMFs of the training dataset.

3.3.3. Phase 3: Normalization process

The data are normalized between [0, 1] using Eq. (13) and due invertible nature of the normalization, the results will not be affected (Hsu et al., 2003). The normalization is carried out following Eq. (16) to handle large variation in the data (Hsu et al., 2003):

$$\Theta_{norm} = \frac{(\Theta - \Theta_{\min})}{(\Theta_{\max} - \Theta_{\min})} \tag{16}$$

In Eq. (16), Θ represents the input/output, $\Theta_{\min}$is the minimum value, $\Theta_{\max}$is the maximum value of the data and $\Theta_{norm}$is the corresponding normalized numeric value.

3.3.4. Phase 4: Kernal ridge regression (KRR) method

In the final phase of the modeling, the KRR model is then applied to forecast multi-scaler SPI series over the 1-, 3-, 6- and 12-monthly forecasting horizons to investigate its cability to predict drought events over multiple timescales. After incorporating the selected IMFs (for training perod) with lagged at $(t-1)$ in the KRR model, different types of kernals (*i.e.*, linear, polynomial and gaussian) are tried. It is remarkable that in this study, the gaussian and polynomial kernals are the appropriate kernals to acquire the optimum MEMD-SA-KRR and

Table 1
Summary of the basic statistics of the meteorological predictors, large scale climate mode indices and target variable of the study locations. Further, std. denotes standard deviation, skew represents skewness, and kurt indicates kurtosis.

Study Sites	Geographic characteristics			Climatological statistics (1981–2015)																				
				Precipitation PTCN (mm)						Temperature T (°C)						Humidity H (g/m3)								
	Longitude (°E)	Latitude (°N)	Elevation (m)	Mean	Max	Min	Std.	Skew	Kurt	Mean	Max	Min	Std.	Skew	Kurt	Mean	Max	Min	Std.	Skew	Kurt			
Faisalabad	73.13°	31.45°	184	35.4	435.3	0.1	49.3	3.0	14.0	26.9	37.2	9.5	7.1	−0.4	−1.2	38.6	80.0	16.0	10.4	0.5	0.5			
Islamabad	73.08°	33.73°	604	109.1	743.3	0.5	129.5	2.1	4.6	21.9	33.5	8.3	7.3	−0.2	−1.4	43.4	71.0	16.0	12.0	0.0	−0.7			
Jhelum	73.72°	32.74°	234	75.7	648.6	0.2	92.3	2.3	7.1	23.8	34.9	10.5	7.1	−0.3	−1.3	46.2	77.0	15.0	13.2	−0.3	−0.8			
Drought statistics																								
SPI_1							SPI_3						SPI_6						SPI_{12}					
	Mean	Max	Min	Std.	Skew	Kurt	Mean	Max	Min	Std.	Skew	Kurt	Mean	Max	Min	Std.	Skew	Kurt	Mean	Max	Min	Std.	Skew	Kurt
Faisalabad	0.1	2.4	−2.4	0.9	0.4	−0.2	0.1	2.3	−1.9	0.9	0.3	−0.6	0.0	2.1	−2.0	0.9	0.3	−0.8	0.0	2.0	−1.7	0.9	0.3	−0.9
Islamabad	0.0	2.8	−3.3	1.0	0.0	0.1	0.0	2.3	−2.7	1.0	−0.2	−0.2	0.0	2.0	−3.0	1.0	−0.4	−0.2	0.0	2.0	−2.6	1.0	−0.1	−0.4
Jhelum	0.1	2.4	−2.4	0.9	0.4	−0.2	0.1	2.3	−1.9	0.9	0.3	−0.6	0.0	2.1	−2.0	0.9	0.3	−0.8	0.0	2.0	−1.7	0.9	0.3	−0.9
Synoptic climate mode indices statistics (1981–2015)																								
Climate indices	Notation	Mean	Max	Min	Std.	Skew	Kurt	Climate indices	Notation	Mean	Max	Min	Std.	Skew	Kurt									
Southern Oscillation index	SOI	−1.5	27.1	–33.3	11.0	−0.1	−0.2	Pacific Decadal Oscillation	PDO	1.1	184.0	−2.3	11.7	14.4	210.3									
Sea Surface Temperature 1	Nino3SST	25.9	29.2	23.1	1.3	0.2	−0.6	Indian Ocean Dipole	IOD	0.2	1.5	−0.7	0.3	0.5	1.3									
Sea Surface Temperature 2	Nino3.4SST	27.1	29.2	24.5	1.0	−0.1	−0.4	ENSO Modoki index	EMI	−0.1	1.1	−1.5	0.5	−0.6	−0.4									
Sea Surface Temperature 3	Nino4SST	28.6	30.2	26.6	0.7	−0.6	−0.1	Southern Annular Mode	SAM	0.1	4.9	−3.3	1.1	0.0	0.8									
Periodicity	month	6.5	12.0	1.0	3.5	0.0	−1.2																	

Table 2
Design parameters involved to decomposed IMFs and residuals for training and testing period in each study site using multivariate empirical mode decomposition (MEMD) method.

Multi-scaler SPI	Training period				Testing period			
	No. of total projections	Stop vector tolerance values	threshold	No. of total sub-series in each input (IMFs & Res.)	No. of total projections	Stop vector tolerance values	threshold	No. of total sub-series of each input (IMFs & Res.)
Site 1: Faisalabad								
SPI_1	100	[0.05 0.05]	0.5	10	100	[0.05 0.05]	0.5	10
SPI_3	100	[0.05 0.05]	0.5	10	100	[0.05 0.05]	0.5	10
SPI_6	100	[0.05 0.05]	0.5	10	100	[0.05 0.05]	0.5	10
SPI_{12}	100	[0.05 0.05]	0.5	10	140	[0.05 0.05]	0.5	10
Site 2: Islamabad								
SPI_1	26	[0.05 0.05]	0.5	8	26	[0.05 0.05]	0.5	8
SPI_3	100	[0.05 0.05]	0.5	8	100	[0.05 0.05]	0.5	8
SPI_6	26	[0.05 0.05]	0.5	8	26	[0.05 0.05]	0.5	8
SPI_{12}	140	[0.05 0.05]	0.5	8	140	[0.05 0.05]	0.5	8
Site 3: Jhelum								
SPI_1	100	[0.05 0.05]	0.5	9	70	[0.05 0.05]	0.5	9
SPI_3	49	[0.05 0.05]	0.5	8	45	[0.05 0.05]	0.5	8
SPI_6	100	[0.05 0.05]	0.5	9	100	[0.05 0.05]	0.5	9
SPI_{12}	140	[0.05 0.05]	0.5	9	140	[0.05 0.05]	0.5	9

Table 3
Design parameters involving in the selected IMFs for training period for each study site using simulate annealing (SA) algorithm. The number of total IMFs is also given.

Multi-scaler SPI	Max. Iteration	Initial Temperature	Training period	
			No. of total IMFs	No. of selected IMFs
Site 1: Faisalabad				
SPI_1	20	10	120	25
SPI_3	20	10	120	25
SPI_6	20	10	120	25
SPI_{12}	20	10	120	25
Site 2: Islamabad				
SPI_1	20	10	96	25
SPI_3	20	10	96	25
SPI_6	20	10	96	25
SPI_{12}	20	10	96	25
Site 3: Jhelum				
SPI_1	20	10	108	25
SPI_3	20	10	96	25
SPI_6	20	10	108	25
SPI_{12}	20	10	108	25

standalone KRR model accuracy. For validating the performance of the hybrid MEMD-SA-KRR model, the same IMFs are picked in the testing period following the IMFs of the training period. For comparison purposes, the RF model is also hybridized with the MEMD-SA to design MEMD-SA-RF model. The number of trees (1000) and predictors (5) are defined prior to develop the MEMD-SA-RF model. Further, the standalone KRR and Standalone RF (number of trees (1000) and predictors (3)) models are also evaluated (Table 4).

Fig. 2 illustrates the schematic understanding of the multi-phase MEMD-SA-KRR hybrid model.

The *r* and *MSE* metrics were adopted to assess the MEMD-SA-KRR accuracy in training against MEMD-SA-RF, standalone KRR and RF models.

The values of *r* and *RMSE* generated by MEMD-SA-KRR model for multi-scaler SPI forecasting at Faisalabad are seen to be: SPI_1 ($r = 0.970$, $MSE = 0.045$), SPI_3 ($r = 0.993$, $MSE = 0.009$), SPI_6 ($r = 0.994$, $MSE = 0.007$) and SPI_{12} ($r = 0.997$, $MSE = 0.003$). These metrics for comparison models are MEMD-SA-RF ($r = 0.968$ (SPI_1), 0.985(SPI_3), 0.994 (SPI_6), 0.996 (SPI_{12}) , $MSE = 0.076$ (SPI_1), 0.024 (SPI_3), 0.007 (SPI_6), 0.003 (SPI_{12})), standalone KRR ($r = 0.829$ (SPI_1), 0.906 (SPI_3), 0.917 (SPI_6), 0.923 (SPI_{12}) , $MSE = 0.328$ (SPI_1), 0.184 (SPI_3), 0.150 (SPI_6), 0.140 (SPI_{12})) and standalone RF ($r = 0.965$ (SPI_1), 0.973 (SPI_3), 0.973 (SPI_6), 0.976 (SPI_{12}) , $MSE = 0.111$ (SPI_1), 0.052 (SPI_3), 0.047 (SPI_6), 0.044 (SPI_{12})).

Equivalent metrics of the MEMD-SA-KRR model for Islamabad (site 2) are found to be: (($r = 0.931$, $MSE = 0.144$) SPI_1, ($r = 0.974$, $MSE = 0.049$) SPI_3, ($r = 0.982$, $MSE = 0.031$) SPI_6, ($r = 0.993$, $MSE = 0.010$) SPI_{12}). The values of *r* and *MSE* generated by comparative models can be seen in Table 4. Similarly, the proposed multi-phase MEMD-SA-KRR model reasonably performs better for site 3 Jhelum as compared to other models. Consequently, it is evident that the multi-phase MEMD-SA-KRR model accuracy in the testing phase, as shown later, is relatively high for the multi-scaler SPI forecasts at all tested locations.

3.4. The performance assessing criterion

The newly designed hybrid MEMD-SA-KRR *vs*. MEMD-SA-RF, standalone KRR and RF models were assessed using several distinct evaluation criterions during multi-scaler SPI forecasting. The well-known statistical metrics based on earlier approaches (ASCE, 1993; ASCE, 2000; Yen, 1995) are employed in this work (Dawson et al., 2007; Deo et al., 2016; Legates and McCabe, 1999; Willmott, 1981; Willmott, 1982; Willmott, 1984).

I. Correlation coefficient (r) is defined as:

$$r = \left(\frac{\sum_{i=1}^{N} (SPI_{Obs,i} - \bar{SPI}_{Obs,i})(SPI_{For,i} - \bar{SPI}_{For,i})}{\sqrt{\sum_{i=1}^{N} (SPI_{Obs,i} - \bar{SPI}_{Obs,i})^2} \sqrt{\sum_{i=1}^{N} (SPI_{For,i} - \bar{SPI}_{For,i})^2}} \right) \quad (17)$$

II. Willmott's Index (EWI) is formulated as:

$$E_{WI} = 1 - \left[\frac{\sum_{i=1}^{N} (SPI_{FOR,i} - SPI_{OBS,i})^2}{\sum_{i=1}^{N} (|SPI_{FOR,i} - \bar{SPI}_{OBS,i}| + |SPI_{OBS,i} - \bar{SPI}_{OBS,i}|)^2} \right], \quad 0 \le E_{WI} \le 1 \quad (18)$$

III. Nash-Sutcliffe efficiency (ENS) value is described as:

$$E_{NS} = 1 - \left[\frac{\sum_{i=1}^{N} (SPI_{Obs,i} - SPI_{For,i})^2}{\sum_{i=1}^{N} (\bar{SPI}_{Obs,i} - \bar{SPI}_{For,i})^2} \right], \quad 0 \le E_{NS} \le 1 \quad (19)$$

IV. Root mean square error (RMSE) is mathematically derived as:

Table 4
Performance of training period using multivariate empirical mode decomposition model hybridized with simulated annealing and Kernel ridge regression (i.e. **MEMD-SA-KRR)** model *vs.* MEMD-SA-RF, Standalone KRR and Standalone RF models in terms of *r* and *MSE*. The choice of Kernel types were: polynomial, linear and Gaussian.

Multi-scaler SPI	MEMD-SA-KRR			MEMD-SA-RF				Standalone KRR			Standalone RF			
	Kernel type	Training period		No. tress	No. split predictor	Training period		Kernel type	Training period		No. tress	No. split predictor	Training period	
		MSE	*r*			*MSE*	*r*		*MSE*	*r*			*MSE*	***r***
					Site 1: Faisalabad									
$SP1_1$	Polynomial	0.045	0.970	1000	5	0.076	0.968	Gaussian	0.328	0.829	1000	3	0.111	0.965
SPI_3	Polynomial	0.009	0.993	1000	5	0.024	0.985	Gaussian	0.184	0.906	1000	3	0.052	0.973
SPI_6	Polynomial	0.007	0.994	1000	5	0.007	0.994	Gaussian	0.150	0.917	1000	3	0.047	0.973
SPI_{12}	Polynomial	0.003	0.997	1000	5	0.003	0.996	Gaussian	0.140	0.923	1000	3	0.044	0.976
					Site 2: Islamabad									
$SP1_1$	Polynomial	0.144	0.931	1000	5	0.152	0.977	Gaussian	0.496	0.812	1000	3	0.173	0.973
SPI_3	Polynomial	0.049	0.974	1000	5	0.089	0.981	Gaussian	0.372	0.850	1000	3	0.120	0.970
SPI_6	Polynomial	0.031	0.982	1000	5	0.036	0.986	Gaussian	0.291	0.894	1000	3	0.090	0.978
SPI_{12}	Polynomial	0.010	0.993	1000	5	0.011	0.994	Gaussian	0.299	0.885	1000	3	0.091	0.978
					Site 3: Jhelum									
$SP1_1$	Polynomial	0.127	0.910	1000	5	0.131	0.918	Gaussian	0.330	0.831	1000	3	0.116	0.969
SPI_3	Polynomial	0.017	0.987	1000	5	0.026	0.983	Gaussian	0.252	0.863	1000	3	0.078	0.970
SPI_6	Polynomial	0.010	0.992	1000	5	0.011	0.991	Gaussian	0.201	0.888	1000	3	0.059	0.972
SPI_{12}	Polynomial	0.002	0.998	1000	5	0.003	0.998	Gaussian	0.183	0.895	1000	3	0.054	0.975

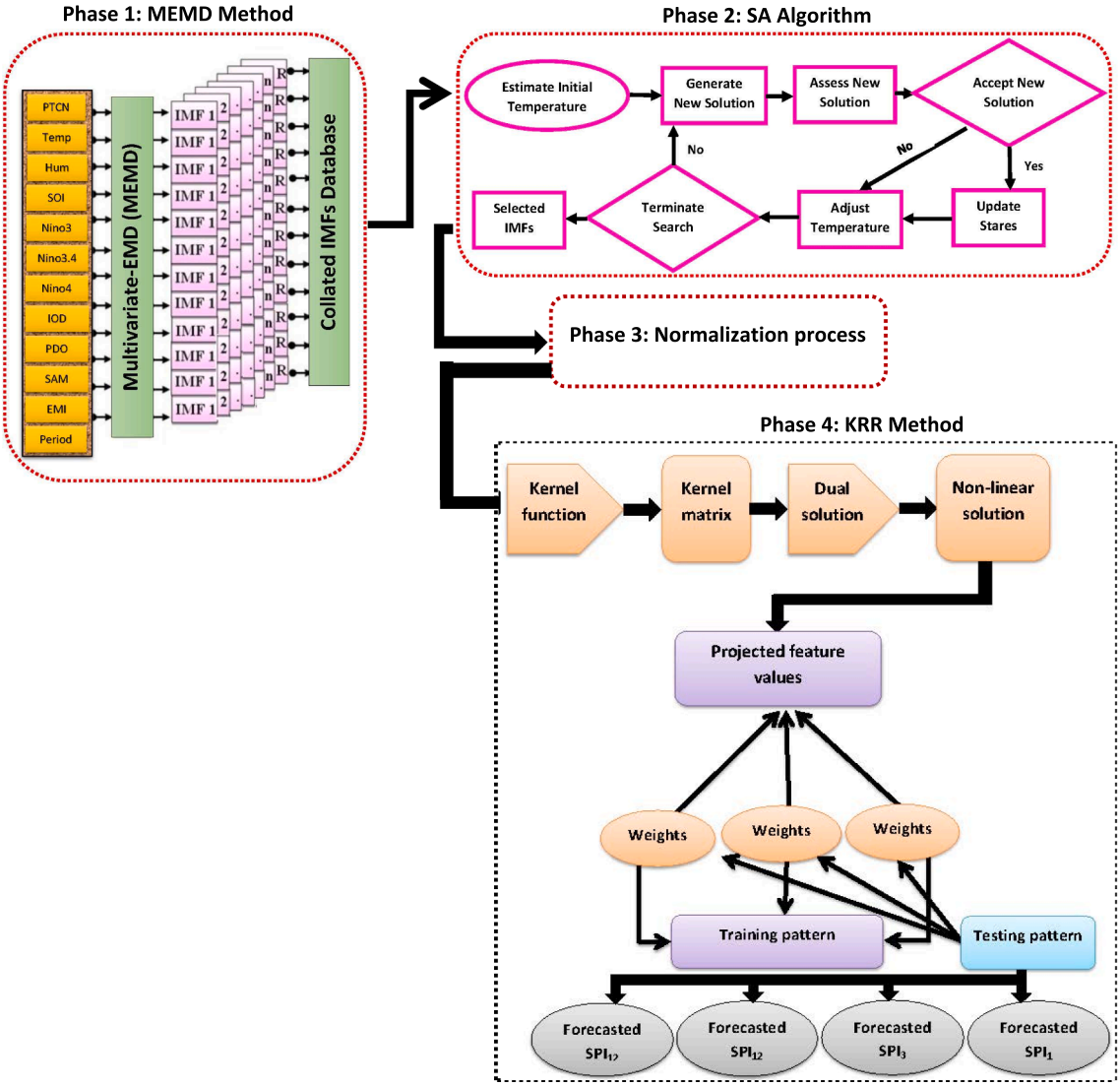


Fig. 2. Schematic structure of the proposed multi-phase (MEMD-SA-KRR) model integrating multivariate empirical mode decomposition (MEMD) at phase 1 and simulated annealing (SA) at phase 2 with kernel ridge regression (KRR) model at phase 4.

Table 5
Multi-scale analysis of testing period using the **MEMD-SA-KRR** *vs.* MEMD-SA-RF, Standalone KRR, and Standalone RF models measured by *RMSE*, *MAE* and *r*. The optimum model is blue bold faced.

Multi-scaler SPI	$SP1_1$			SPI_3			SPI_6			SPI_{12}		
Models	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*
Site 1: Faisalabad												
Standalone RF	0.332	0.254	0.959	0.263	0.199	0.976	0.273	0.208	0.976	0.292	0.224	0.974
Standalone KRR	0.580	0.440	0.819	0.512	0.392	0.901	0.540	0.435	0.899	0.573	0.481	0.896
MEMD-SA-RF	0.252	0.188	0.971	0.149	0.118	0.993	0.092	0.068	0.997	0.055	0.041	0.999
MEMD-SA-KRR	**0.180**	**0.133**	**0.980**	**0.078**	**0.059**	**0.997**	**0.066**	**0.050**	**0.998**	**0.034**	**0.025**	**0.999**
Site 2: Islamabad												
Standalone RF	0.445	0.353	0.978	0.404	0.315	0.972	0.393	0.303	0.976	0.415	0.338	0.977
Standalone KRR	0.733	0.584	0.841	0.702	0.556	0.847	0.712	0.565	0.874	0.771	0.617	0.882
MEMD-SA-RF	0.367	0.295	0.975	0.291	0.232	0.982	0.192	0.142	0.990	0.133	0.095	0.994
MEMD-SA-KRR	**0.314**	**0.246**	**0.953**	**0.119**	**0.087**	**0.994**	**0.146**	**0.106**	**0.992**	**0.111**	**0.077**	**0.995**
Site 3: Jhelum												
Standalone RF	0.357	0.282	0.963	0.362	0.297	0.964	0.364	0.303	0.966	0.382	0.321	0.964
Standalone KRR	0.606	0.473	0.800	0.636	0.515	0.836	0.642	0.541	0.854	0.668	0.588	0.858
MEMD-SA-RF	0.309	0.229	0.944	0.142	0.116	0.991	0.122	0.080	0.994	0.052	0.039	0.999
MEMD-SA-KRR	**0.313**	**0.215**	**0.938**	**0.098**	**0.073**	**0.995**	**0.092**	**0.067**	**0.996**	**0.047**	**0.034**	**0.999**

Table 6
Multi-scale analysis in testing period of **MEMD-SA-KRR** *vs.* MEMD-SA-RF, Standalone KRR and Standalone RF models using E_{WI}, E_{NS} and E_{LM} . Note that the best model is boldfaced (blue).

Multi-scaler	$SP1_1$			SPI_3			SPI_6			SPI_{12}		
Models	E_{WI}	E_{NS}	E_{LM}	E_{WI}	E_{NS}	E_{LM}	E_{WI}	E_{NS}	E_{LM}	E_{WI}	E_{NS}	E_{LM}
Site 1: Faisalabad												
Standalone RF	0.885	0.861	0.653	0.953	0.933	0.771	0.954	0.933	0.782	0.952	0.928	0.782
Standalone KRR	0.634	0.576	0.399	0.798	0.745	0.550	0.794	0.740	0.544	0.782	0.725	0.531
MEMD-SA-RF	0.938	0.920	0.743	0.986	0.978	0.864	0.995	0.992	0.929	0.998	0.997	0.960
MEMD-SA-KRR	**0.971**	**0.959**	**0.818**	**0.996**	**0.994**	**0.932**	**0.998**	**0.996**	**0.948**	**0.999**	**0.999**	**0.976**
Site 2: Islamabad												
Standalone RF	0.850	0.798	0.559	0.898	0.851	0.624	0.926	0.875	0.666	0.913	0.867	0.629
Standalone KRR	0.545	0.452	0.270	0.646	0.549	0.337	0.712	0.590	0.379	0.649	0.542	0.325
MEMD-SA-RF	0.905	0.862	0.630	0.953	0.922	0.723	0.985	0.970	0.843	0.993	0.986	0.896
MEMD-SA-KRR	**0.940**	**0.900**	**0.693**	**0.993**	**0.987**	**0.896**	**0.992**	**0.983**	**0.884**	**0.995**	**0.991**	**0.915**
Site 3: Jhelum												
Standalone RF	0.861	0.839	0.615	0.903	0.873	0.658	0.912	0.882	0.682	0.911	0.877	0.687
Standalone KRR	0.596	0.536	0.354	0.676	0.606	0.408	0.695	0.632	0.433	0.688	0.626	0.427
MEMD-SA-RF	0.907	0.879	0.687	0.988	0.980	0.867	0.992	0.987	0.916	0.999	0.998	0.962
MEMD-SA-KRR	**0.911**	**0.877**	**0.707**	**0.994**	**0.991**	**0.916**	**0.995**	**0.992**	**0.930**	**0.999**	**0.998**	**0.967**

Table 7
Geographic evaluation of the **MEMD-SA-KRR** *vs.* MEMD-SA-RF, Standalone KRR and Standalone RF models using relative percentage error (*RPE*, %). Note that the best model is boldfaced (blue).

Multi-scaler SPI	$SP1_1$	SPI_3	SPI_6	SPI_{12}
	RPE (%)	RPE (%)	RPE (%)	RPE (%)
Site 1: Faisalabad				
Standalone RF	60.13	51.52	26.65	23.07
Standalone KRR	84.99	79.86	47.07	48.10
MEMD-SA-RF	49.15	31.28	8.79	4.29
MEMD-SA-KRR	**34.86**	**18.75**	**6.51**	**2.91**
Site 2: Islamabad				
Standalone RF	71.71	130.57	48.89	64.66
Standalone KRR	121.88	282.69	153.95	97.39
MEMD-SA-RF	104.26	342.50	75.89	30.58
MEMD-SA-KRR	**126.31**	**81.35**	**84.67**	**29.22**
Site 3: Jhelum				
Standalone RF	54.90	70.59	34.29	32.56
Standalone KRR	87.31	102.30	57.21	58.94
MEMD-SA-RF	64.39	38.51	9.56	4.06
MEMD-SA-KRR	**63.48**	**26.31**	**8.63**	**3.66**

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(SPI_{For,i} - SPI_{Obs,i})^2} \tag{20}$$

V. Mean absolute error (MAE) is expressed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|(SPI_{For,i} - SPI_{Obs,i})| \tag{21}$$

VI. Legates and McCabe's (ELM) is expressed as:

$$E_{LM} = 1 - \left[\frac{\sum_{i=1}^{N}|SPI_{For,i} - SPI_{Obs,i}|}{\sum_{i=1}^{N}|SPI_{Obs,i} - \bar{SPI}_{Obs,i}|}\right], 0 \le E_{LM} \le 1 \tag{22}$$

VII. Relative percentage error (RPE; %), is stated as

$$RPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{(SPI_{For,i} - SPI_{Obs,i})}{SPI_{Obs,i}}\right| \times 100 \tag{23}$$

In Eq. (23), $SPI_{Obs,i}$ and $SPI_{For,i}$ shows the observed and forecasted i^{th} magnitudes of the SPI index, $\bar{SPI}_{obs,i}$ and $\bar{SPI}_{For,i}$ are the observed and

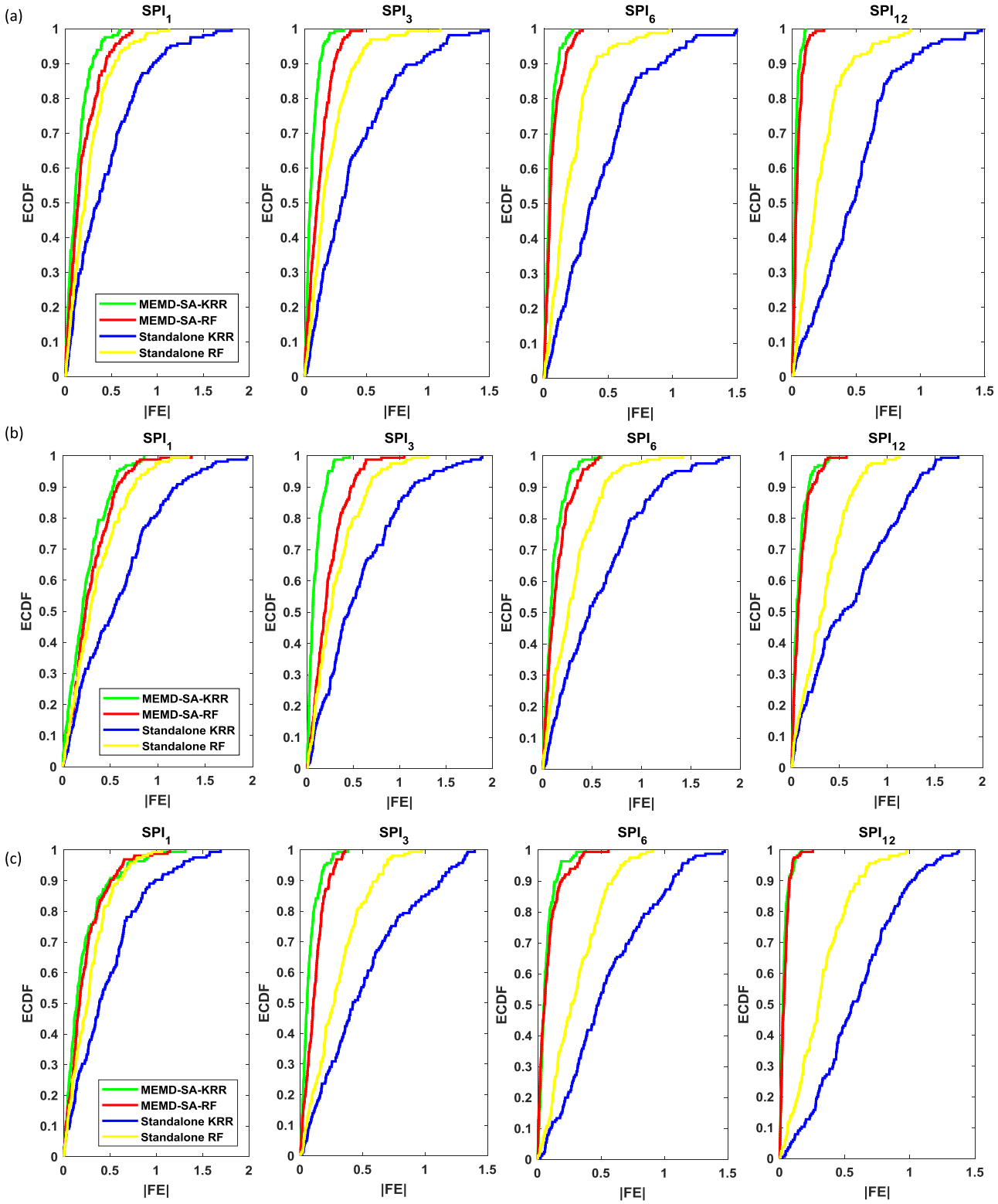


Fig. 3. Empirical cumulative distribution function (ECDF) of forecasted error |FE| 1-month (SP_1), 3-month (SPI_3), 6-month (SPI_6) and 12-month (SPI_{12}) generated by the proposed multi-phase MEMD-SA-KRR *vs.* MEMD-SA-RF, Standalone KRR and Standalone RF models for (a): Faisalabad, (b): Islamabad, and (c): Jhelum.

forecasted average of SPI and N is the total number of tested data points.

4. Results

Based on the above developed artificial intelligence models, drought forecasting has been performed for both short-term (1 ~ 3 months) and long term (6 ~ 12 months) horizons by modeling SPI as a drought indicator and utilizing synoptic-scale climate mode indices and relevant climate datasets. The MEMD algorithm applied in combination with SA and KRR methods in this study, aimed to design a multi-phase drought forecast model, denoted as MEMD-SA-KRR. The MEMD-SA-KRR model

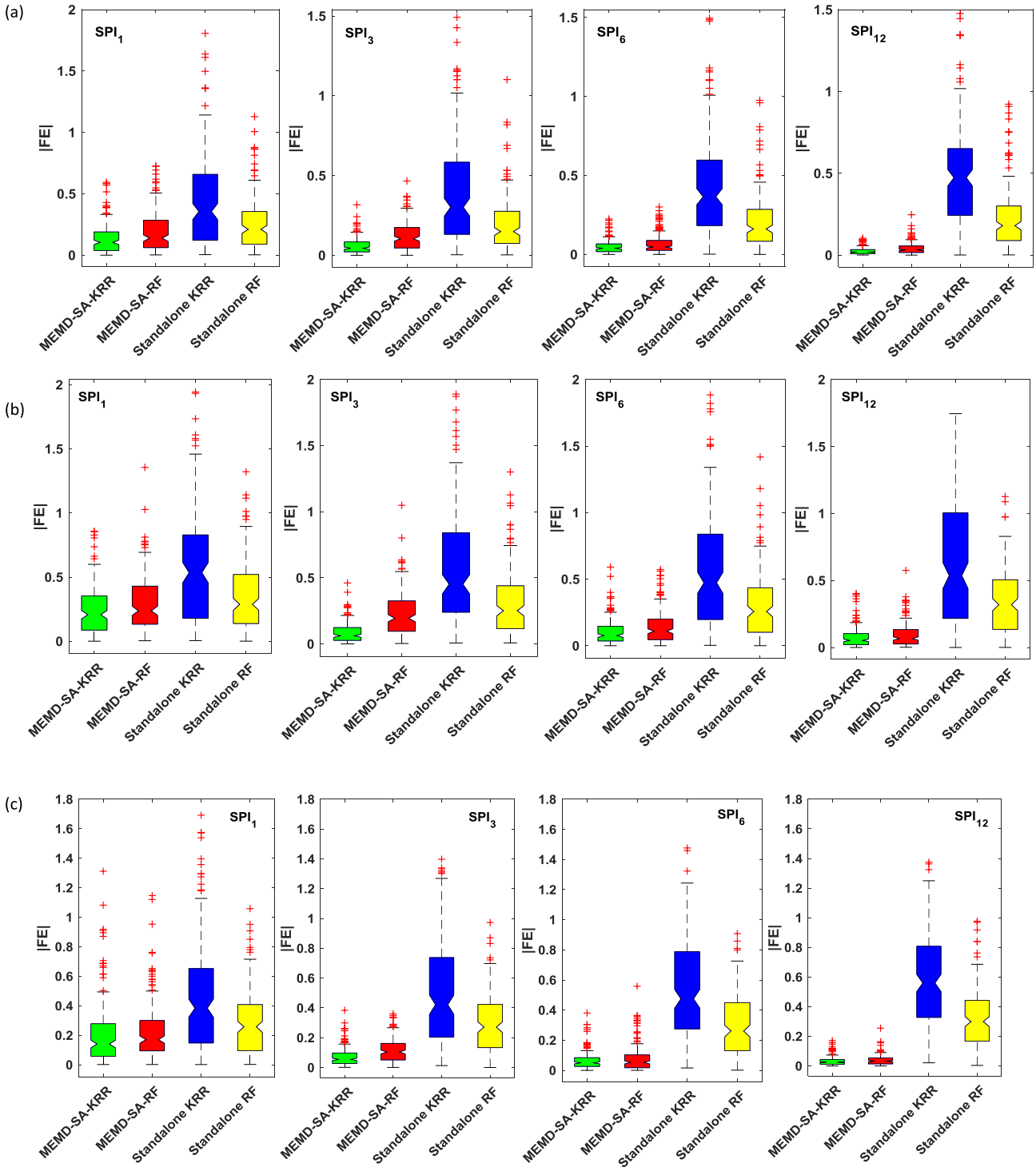


Fig. 4. Box-plots of forecasted error |FE| of 1-month (SP_1), 3-month (SPI_3), 6-month (SPI_6) and 12-month (SPI_{12}) in testing period generated by multi-phase MEMD-SA-RF *vs.* MEMD-SA-RF, Standalone KRR and Standalone RF models for (a): Faisalabad, (b): Islamabad, and (c): Jhelum.

is seen to be a well-established model in terms with ability to extract features from multivariate data comprised of meteorological variables and climate indices; hence, selecting the best features out of extracted oscillatory modes to forecast the multi-scaler SPI series. The performance is assessed with the help of some well-known statistical measures, visual and graphical plots with error distributions strategies in testing period.

4.1. Assessment of MEMD-SA-KRR model using statistical measures

In this paper, the well-designed multi-phase forecasting model MEMD-SA-KRR *vs.* MEMD-SA-RF, Standalone KRR and Standalone RF models is numerically evaluated using several acceptable performance metrics.

The performance of MEMD-SA-KRR model is appraised with MEMD-SA-RF, standalone KRR and standalone RF models utilizing *r*, *RMSE* and *MAE* metrics in Table 5. The hybrid MEMD-SA-KRR model developed

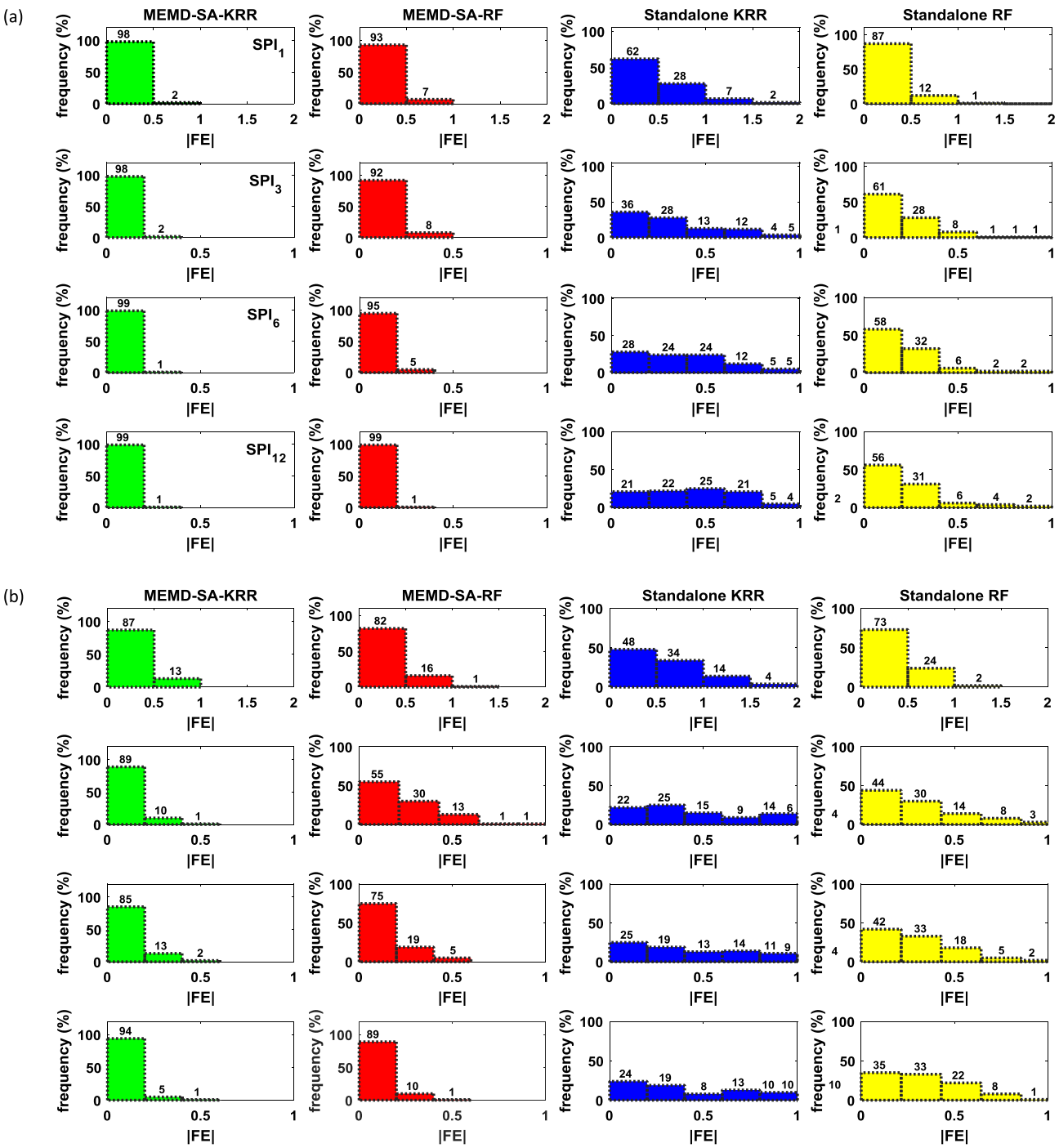


Fig. 5. Cumulative frequency of 1-month (SP_1), 3-month (SPI_3), 6-month (SPI_6) and 12-month (SPI_{12}) generated by MEMD-SA-KRR *vs.* MEMD-SA-RF, standalone KRR and standalone RF models of |FE| error for (a): Faisalabad, (b): Islamabad, and (c): Jhelum.

for Faisalabad attain the highest magnitudes of *r* and lowest *RMSE* and *MAE* values in SPI_1 ($r \approx 0.980$, $RMSE \approx 0.180$, $MAE \approx 0.133$) in comparison with MEMD-SA-RF ($r \approx 0.971$, $RMSE \approx 0.252$, $MAE \approx 0.188$), standalone KRR ($r \approx 0.819$, $RMSE \approx 0.580$, $MAE \approx 0.440$), and the standalone RF ($r \approx 0.959$, $RMSE \approx 0.332$, $MAE \approx 0.254$) model. Analogously, the multi-phase MEMD-SA-KRR model attains highest accuracy to forecast SPI_3, SPI_6 and SPI_{12} in response to the comparison models. Likewise, the performance of MEMD-SA-KRR model is significantly better for Islamabad and Jhelum (Table 5) to forecast multi-scaler SPI. This confirms that MEMD-SA-KRR model can be adopted as a well-established data-driven technique to forecast drought against MEMD-SA-RF, standalone KRR and standalone RF models.

Table 6 uses multi-scale E_{WI}, E_{NS} and E_{LM} criterion to analyse the performance of MEMD-SA-KRR model *vs.* MEMD-SA-RF, standalone KRR, and standalone RF models. The score of these metrics generated by MEMD-SA-KRR model for Faisalabad in SPI_1 are ($E_{WI} \approx 0.971$, $E_{NS} \approx 0.959$ and $E_{LM} \approx 0.818$), followed by MEMD-SA-RF ($E_{WI} \approx 0.938$, $E_{NS} \approx 0.920$ and $E_{LM} \approx 0.743$), standalone RF ($E_{WI} \approx 0.885$, $E_{NS} \approx 0.861$ and $E_{LM} \approx 0.653$) and standalone KRR ($E_{WI} \approx 0.634$, $E_{NS} \approx 0.576$, $E_{LM} \approx 0.399$) models. The multi-phase MEMD-SA-KRR models also outperform the counterpart models in forecasting medium-scale (*i.e.*, SPI_3) and long term drought scenarios (*i.e.*, SPI_6, SPI_{12}).

For site 2: Islamabad, again the proposed multi-phase MEMD-SA-KRR model appears to be the best model, SPI_1 ($E_{WI} \approx 0.940$, $E_{NS} \approx$

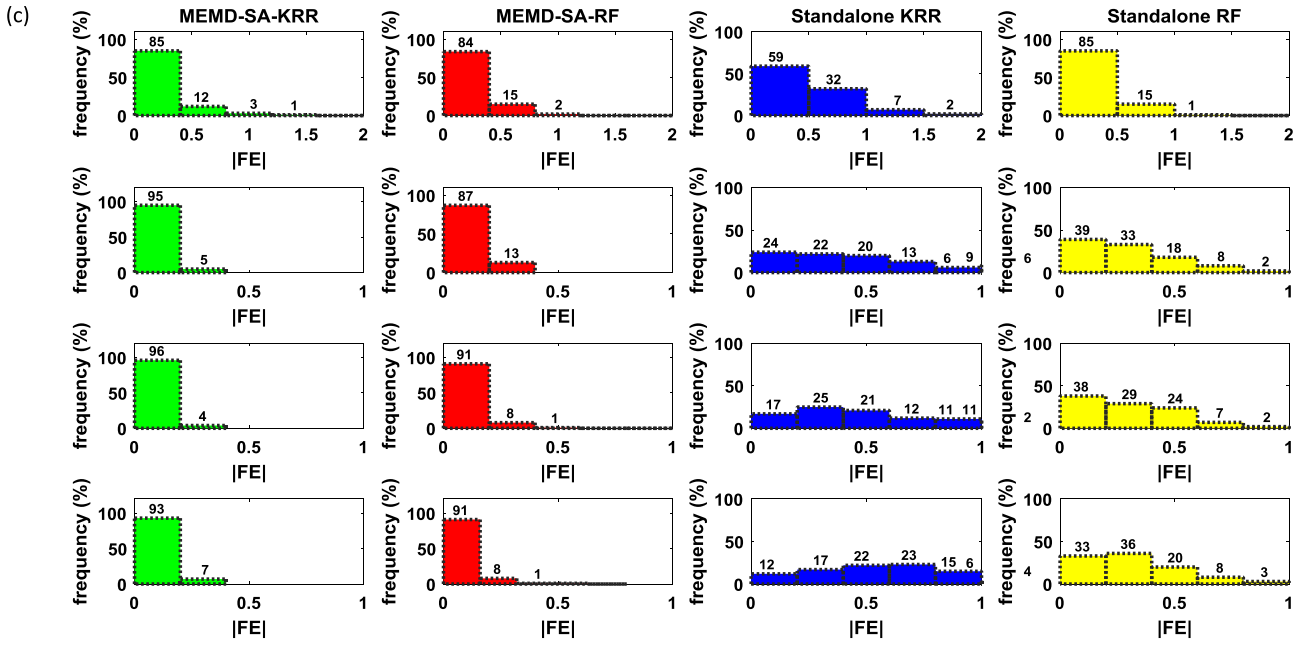


Fig. 5. (*continued*)

0.900 and E_{LM} ≈ 0.693), SPI_3 (E_{WI} ≈ 0.993, E_{NS}≈ 0.987 and E_{LM} ≈ 0.896), SPI_6 (E_{WI} ≈ 0.992, E_{NS}≈0.983 and E_{LM} ≈ 0.884) and SPI_{12} (E_{WI} ≈ 0.995, E_{NS}≈ 0.991 and E_{LM}≈ 0.915). The results of comparative models (Table 6) support the dominance of the multi-phase MEMD-SA-KRR over the comparative approaches.

The MEMD-SA-KRR hybrid model is also assessed for forecasting short, medium and long term drought for site 3: Jhelum. The acquired metrics, evident from the present results, are: for SPI_1 (E_{WI} ≈ 0.911, E_{NS} ≈ 0.877 and E_{LM} ≈ 0.707), for SPI_3 (E_{WI} ≈ 0.994, E_{NS}≈ 0.991 and E_{LM} ≈ 0.916), for SPI_6 (E_{WI} ≈ 0.995, E_{NS}≈ 0.992 and E_{LM} ≈ 0.930) and for SPI_{12} (E_{WI} ≈ 0.999, E_{NS}≈ 0.998 and E_{LM} ≈ 0.967). The results of the comparative models can be tabulated in Table 6 that also ascertain the superiority of MEMD-SA-KRR model for the selected study locations.

Table 7 demonstrates the magnitudes of *RPE* in terms of percentage produced in selected study regions. Based on the achieved precision, Site 1: Faisalabad generates the maximum accurateness in forecasting multi-scaler SPI index where MEMD-SA-KRR *(RPE* ≈ 2.91%*)* for SPI_{12}, *(RPE* ≈ 6.51%*)* for SPI_6, *(RPE* ≈ 18.75%*)* for SPI_3 and *(RPE* ≈ 34.86%*)* for SPI_1. Jhelum is the second most responsive site in forecasting multiple drought indexes using a multi-phase MEMD-SA-KRR model followed by Islamabad respectively. The proposed multi-phase MEMD-SA-KRR model was seen to exhibit lower percentage errors (*RPE, %*) in all study locations for long term drought forecasting within the range of 10% recommended threshold for an excellent model classification.

4.2. Assessment of MEMD-SA-KRR using visual and error distributions plots

The empirical cumulative distribution function (ECDF, Fig. 3) analyses different forecasting abilities plots for each study sites. The hybrid MEMD-SA-KRR model was seen reasonably well against MEMD-SA-RF, Standalone KRR and Standalone RF models. The generated error (0 to ± 1.5) for Faisalabad and Jhelum while (0 to ± 2) for Islamabad in forecasting multi-scaler SPI, Fig. 3 clearly proves that MEMD-SA-KRR model was the most precise and responsive model.

Fig. 4 compares the MEMD-SA-KRR *vs.* the MEMD-SA-RF, Standalone KRR and Standalone RF models in the form of a boxplot. The + sign demonstrates the larger forecasting error |FE| as an outliers of the testing period. The distributed error |FE| is confirmed with a much smaller quartile and was attained by MEMD-SA-KRR method in each study location to forecast multi-scaler SPI followed by MEMD-SA-RF, the Standalone RF and Standalone KRR. By analyzing Fig. 4, the preciseness of the hybrid MEMD-SA-KRR method appeared to be healthier than the comparative counterparts.

To present accurate accounts on the forecasting ability of the prescribed data-driven models, Fig. 5 demonstrates a detailed interpretation by plotting the frequency distribution of datasets using MEMD-SA-KRR method's forecasting error |FE|, together with the relevant comparison models. The acquired |FE| error of the MEMD-SA-KRR approach was within the smallest error range (±2) to forecast SPI_1, (±1)for SPI_3, SPI_6 and SPI_{12} for Faisalabad, Islamabad and Jhelum. Fig. 5 depicts the overall performance of the well-designed MEMD-SA-KRR method was better as compared to other models.

Fig. 6 describe a more tangible and conclusive information by plotting Taylor diagram in terms of a statistical demonstration of how accurate the forecasted and observed multi-scaler SPI coincides using their correlation coefficient (*r*). For site 1: Faisalabad, the *r* of the MEMD-SA-KRR model with actual was around 0.98 (SPI_1), trailed by MEMD-SA-RF≈0.97 (SPI_1), Standalone RF≈0.95 (SPI_1) and Standalone KRR≈0.80 (SPI_1) respectively. Again, the multi-phase MEMD-SA-KRR approach was found nearer to the actual SPI_6 with *r* (MEMD-SA-KRR≈0.999, MEMD-SA-RF≈0.99, Standalone RF≈0.96, and Standalone KRR≈0.90). Similarly, the multi-phase MEMD-SA-KRR was nearer in case of SPI_6 and SPI_{12} in relation to other comparative models. This argument is also established in site 2 and 3 where the proposed multi-phase MEMD-SA-KRR is closely matched with observed SPI_1, SPI_3, SPI_6 and SPI_{12} as compared to the counterpart models.

Fig. 7 displays a scatterplot with goodness-of-fit and r^2 between forecasted and observed multi-scaler SPI index. The proposed multi-phase MEMD-SA-KRR model is clearly better than comparative methods in terms of r^2 (MEMD-SA-KRR ≈0.960 (SPI_1), 0.994 (SPI_3), 0.996(SPI_6), 0.999(SPI_{12}), MEMD-SA-RF≈0.945 (SPI_1), 0.985 (SPI_3), 0.993(SPI_6), 0.997(SPI_{12}), Standalone KRR≈0.695 (SPI_1), 0.811(SPI_3), 0.808(SPI_6), 0.804(SPI_{12}), Standalone RF≈0.923(SPI_1), 0.952(SPI_3), 0.952(SPI_6), 0.948 (SPI_{12})) for Faisalabad. The proposed multi-phase MEMD-SA-KRR model for other sites Islamabad and Jhelum is reasonably good compared to counterpart models (Fig. 7). On the basis of attaining the larger r^2-value, the well-established MEMD-SA-KRR approach shows higher accuracy against other comparison models.

As it is always the case, long-term drought indicators are much

smoother than short-term. For example, if a drought index is analyzed over 6 or 12 monthly period, the changes in the behavior of drought has a lesser fluctuation. Hence, there are no large fluctuation/spikes in the long term SPI time-series, and this is the reason why we have a better forecasting result for these horizons. That is, the 6- and 12 month based SPI index become more stationary. Due to this reason, the proposed model accuracy is high in long term SPI index forecasting. We have checked the entire results and it appears there is no issue with our current results.

5. Discussion: Limitations and future remarks

In this research work, the appropriateness of the hybrid MEMD-SA-KRR model (compared against the MEMD-SA-RF, standalone KRR and standalone RF models) for multiple scale droughts forecasting has been investigated. In comparison with the other models, the hybrid MEMD-SA-RF outperformed the alternatives for all study locations, thus enlightening that the MEMD-SA-KRR method was well-organized and effective in extracting features from climatological variables in a tangible way. The performance of MEMD-SA-KRR has revealed that SA algorithm was beneficial in choosing the relevant features to assist the KRR in better emulating the future multi-scaler SPI.

In addition to the overall superiority of the hybrid MEMD-SA-KRR over the comparative counterpart models, the results also confirmed the appropriateness of Simulated Annealing (SA) in sorting out relevant feature with the assessment criterions for MEMD-SA-KRR method (*i.e.*, Tables 4–6) were remarkably improved than the hybrid MEMD-SA-RF and standalone counterpart models. Since the artificial intelligence models exclusively depend on past data that may significantly affect the 'learning' and forecasting process, the outcomes here establish that an appropriate feature collection should be performed carefully before to implement data-driven models. This is in accordance with the strategies

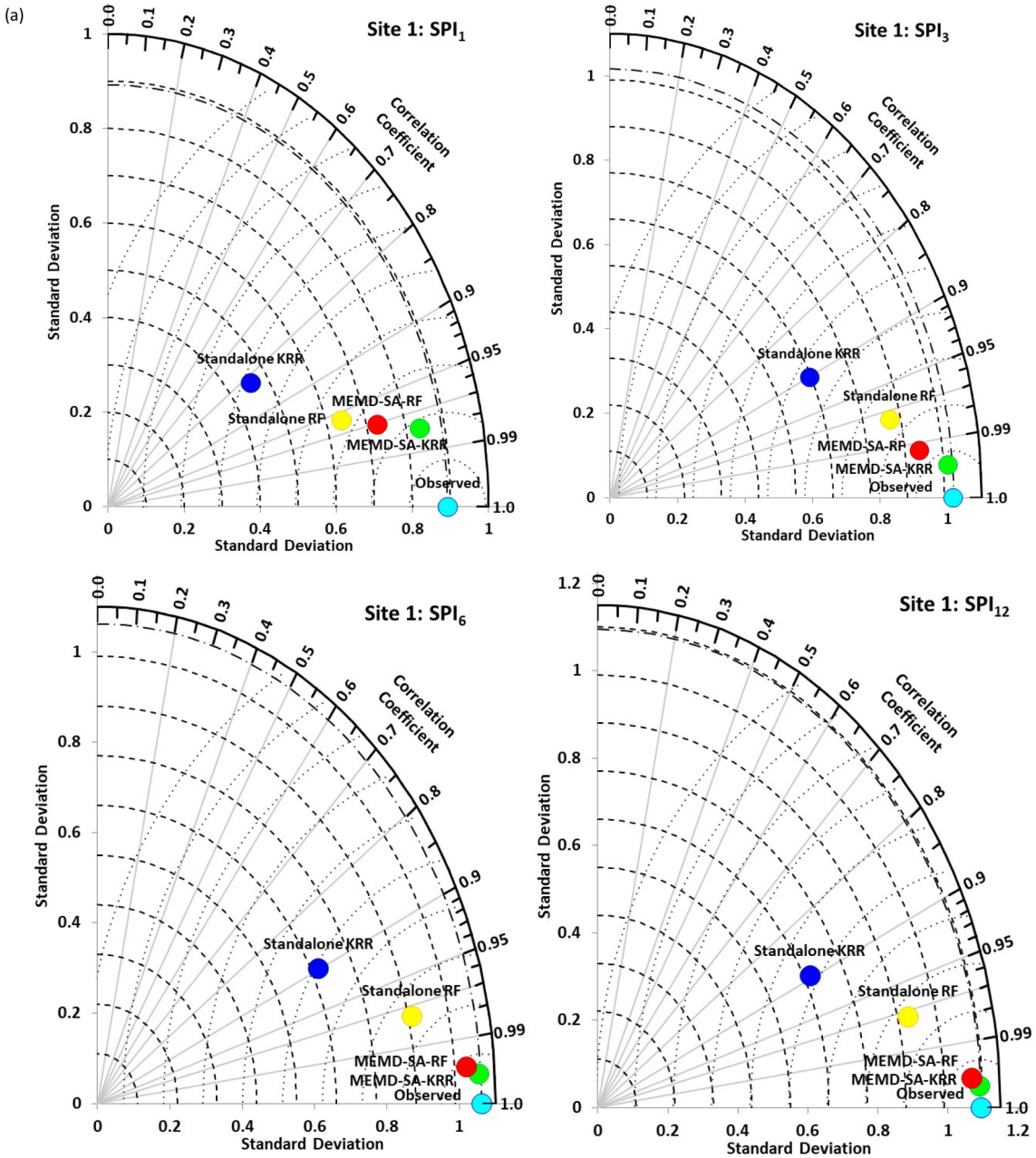


Fig. 6. Taylor plots depicting the predictive skill of MEMD-SA-RF *vs.* MEMD-SA-RF, Standalone KRR and Standalone RF models for (a): Faisalabad, (b): Islamabad, and (c): Jhelum in the testing period of 1-month (SP_1), 3-month (SPI_3), 6-month (SPI_6) and 12-month (SPI_{12}).

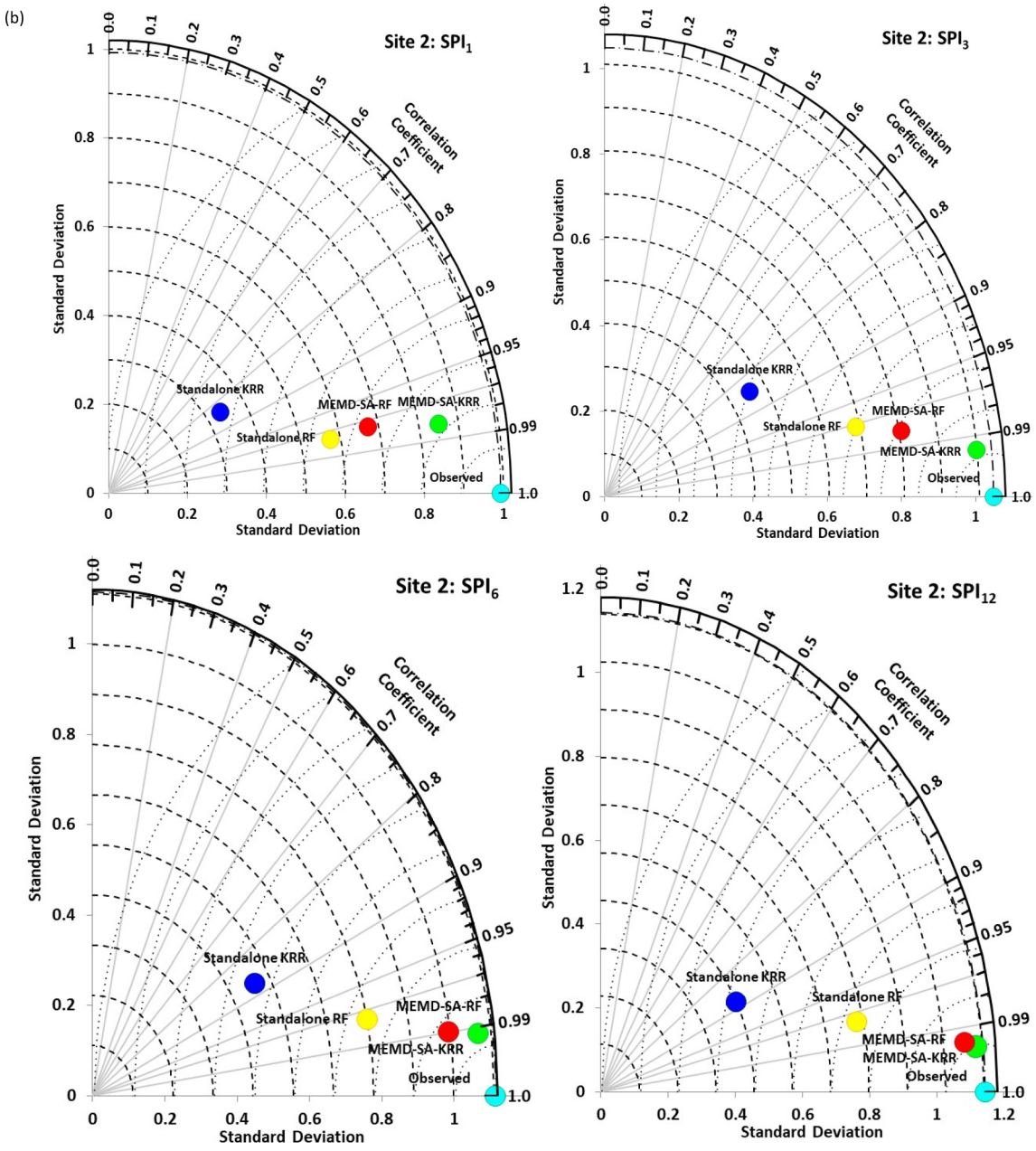


Fig. 6. (*continued*)

followed in (Badr and Fahmy, 2004; Cordón García et al., 2002; Mullen et al., 2009; Singh et al., 2012; Sweetlin et al., 2017). Other key perception is that less number of predictors input needs minimum output associations that results a parsimonious and computationally good KRR model. Another key outcome is the unique input IMFs combinations are compulsory in periodically determining future multi-scaler SPI (Table 2).

In addition to hybridizing the KRR and RF models with the robust SA approach, an extra improvement in model preciseness was accomplished via integration of MEMD model to demarcate the inputs that results into hybrid MEMD-SA-KRR (and MEMD-SA-RF) models. The MEMD is successfully classified and segregate the relevant features inside the climatological inputs to establish a more consistent physical foundation for a particular artificial intelligence method. The usefulness of empirical mode decomposition (EMD), ensemble empirical mode decomposition (EEMD), complete ensemble empirical mode decomposition (CEEMDAN), and improved complete ensemble empirical mode decomposition (ICEEMDAN) has been revealed in numerous studies (Alvanitopoulos et al., 2014; Hong et al., 2013; Wang et al., 2018; Wu et al., 2011), yet in the current research MEMD algorithm is employed for concurrent data pre-processing of numerous climatological predictors. The MEMD is able to identify concurrently the signal's main frequency to capture the respective features.

It should be noted that the feasibility of the MEMD algorithm (Huang et al., 2013; Mandic et al., 2013) for multi-scaler drought forecasting is a major advancement, performed in this research study, to improve the predicting ability of the standalone KRR model. The performance confirmed that the MEMD-SA-KRR can provide better forecasts of multi-scale SPI for the selected study locations in contrast to the corresponding KRR and RF and hybrid MEMD-SA-RF model. It was undoubtedly apparent that better understandings of the physical procedure were given to the hybrid model, mainly by the MEMD method, further enabling the artificial intelligence model to effectively capture the information in the meteorological variables and large-scale climate

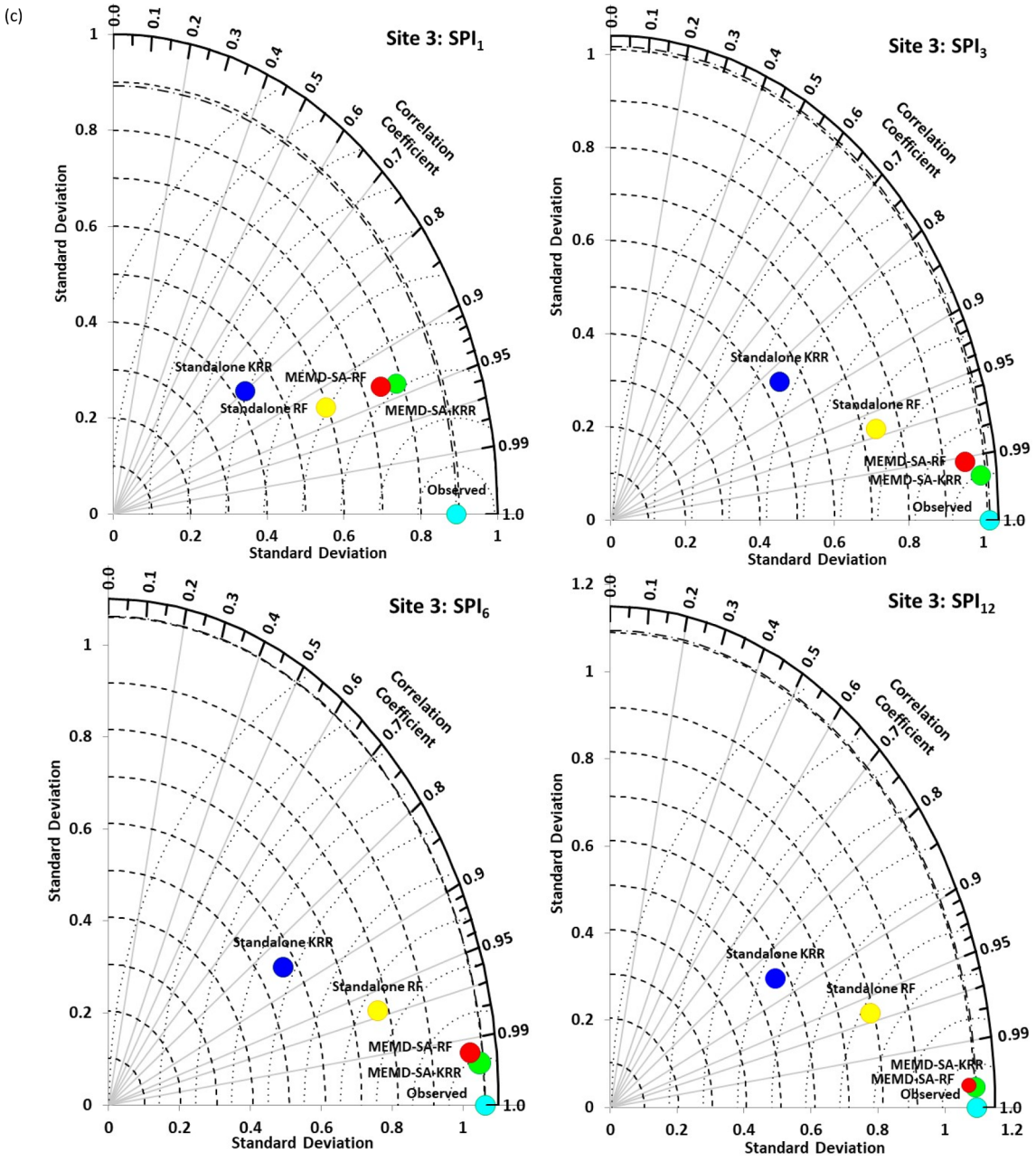


Fig. 6. (*continued*)

indices in relation to modelling the SPI series. One plausible reason for better performance of MEMD-SA-KRR compared to the standalone models, is perhaps attributable to the more effective translation of information on deterministic processes in meteorological variables and climate model indices, resolved into various frequency levels, consequently, resulting in lower model errors and improved performance for multiscale SPI forecasting.

In addition, a primal advantage of the MEMD as utilized in this study is its self-adaptive nature (Huang et al., 2013; Mandic et al., 2013) which completely based on the input predictors and involves minor human effort while decomposing the inputs. Furthermore, the MEMD accurately performs data-driven-based time–frequency analysis of the complex multivariate predictors, while considering the nonlinear behaviours by means of a multi-channelled dynamical process (Rehman and Mandic, 2009a). Another significant merit of the proposed MEMD algorithm is that it is able to overcome the mode alignment issues, which is remained unresolved in the EMD and CEEMD (Looney and Mandic, 2009). Further, MEMD has the ability of decompose multivariate input data while EMD and CEEMD are applicable to decompose a univariate data. Therefore, the hybrid MEMD-SA-KRR approach has the potential for drought management systems. With historically simulated multi-scaler SPI, this improved forecasting tool (such as MEMD-SA-KRR) can amicably be used in an environmental modelling system that can better forecast future drought events and utilize quickly to those requirements reducing the downtimes with growing efficiency.

6. Conclusions

In this research paper, a significant contribution towards drought modelling was made by developing a reliable drought forecast model incorporating most suitable input features derived from a suite of IMFs utilizing the Simulated Annealing (SA) approach. Here, the decomposed datasets were based on the antecedent values of the meteorological variables (*i.e.*, precipitation, temperature, humidity) including the most

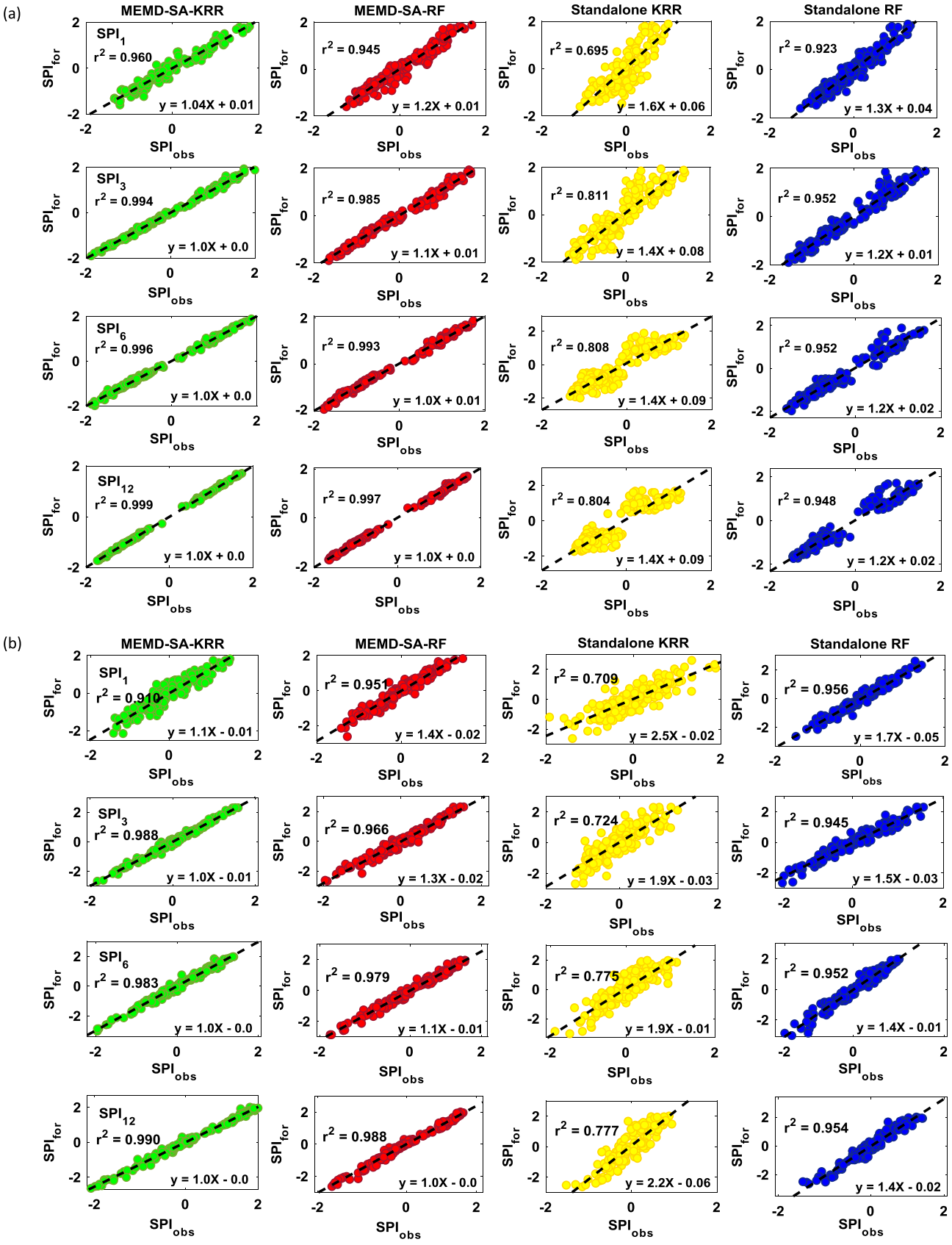


Fig. 7. The 1-month, 3-month, 6-month and 12-month observed *vs.* forecasted SPI in the testing period generated by all four models for (a): Faisalabad, (b): Islamabad, and (c): Jhelum. Least square regression line with the coefficient of determination (r^2) is shown.

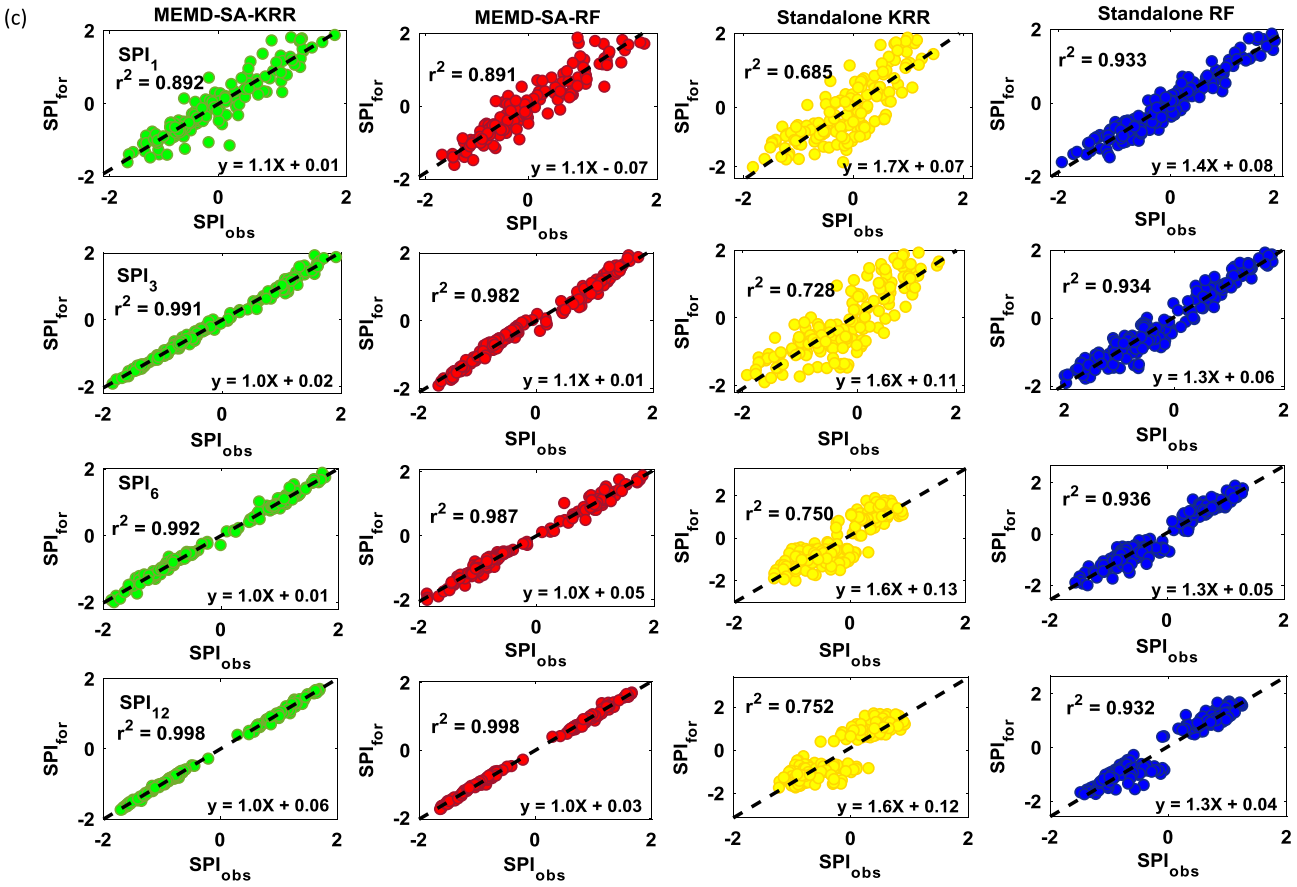


Fig. 7. (*continued*)

relevant climate mode indices (*i.e.*, SOI, Nino3SST, Nino3.4SST, Nino4SST, PDO, IOD, EMI, SAM) and the periodicity factor (*i.e.*, the monthly cycle) in the model's training procedure to forecast the multi-scaler SPI series for 3 agricultural regions in Pakistan: *i.e.*, Faisalabad, Islamabad, and Jhelum.

Long-term monthly datasets over the period 1981–2015 were decomposed for the candidate study sites in their corresponding IMFs and residual factors using multivariate empirical mode decomposition algorithm in which SA was employed to screen the most suitable IMFs in respect to the actual SPI index. The selected IMFs were incorporated into the Kernel Ridge Regression (KRR) algorithm to design the multi-phase hybrid MEMD-SA-KRR model, and the subsequent model's performance was compared against the hybrid MEMD-SA-RF, KRR, and RF model. Several evaluation criterions (*r, RMSE, MAE,* E_{WI}, E_{NS}, E_{LM} & *RPE*) were utilized to assess the multi-phase MEMD-SA-KRR model in modelling drought index at multiple time-scales of 1-, 3-, 6- and 12 month period.

Based on the relatively small forecast errors according to *RMSE* and *MAE*, and the high-performance metrics utilizing correlations, *r,* E_{WI}, E_{NS} and E_{LM}, the accuracy of the multi-phase hybrid MEMD-SA-KRR model was demonstrated to be a highly potent tool in comparison to its counterpart models (*e.g.*, Tables 4 and 5). The *RPE* were found to register values of approximately 34.86%, 18.75%, 6.51%, 2.91% for the hybrid MEMD-SA-KRR model compared with 49.15%, 31.28%, 8.79%, 4.29% for the hybrid MEMD-SA-RF model, that contrasted a value of 60.13%, 51.52%, 26.65%, 23.07% for standalone RF and 84.99%, 79.86%, 47.07%, 48.10% for standalone KRR model for the Faisalabad site for forecasting SPI_1, SPI_3, SPI_6 and SPI_{12} respectively. A remarkable amount of geographic variability in drought model performance was also evident for the multi-phase MEMD-SA-KRR hybrid model using *RPE*, where a primal performance was attained for Faisalabad (based on *RPE*) as compared to Islamabad and Jhelum.

This research study was of the first kind in Pakistan, particularly in introducing the newly designed MEMD approach to resolve meteorological input datasets into their relevant decomposed input signals to select the best candidate features based on SA and to forecast multi-scaler SPI employing KRR model. Expanding the scope of the proposed methodology, future studies can adopt the multi-phase hybrid MEMD-SA-KRR model in other areas such as the prediction of rainfall, streamflow, flood events, solar radiation and energy demand, to enable policymakers in the management of the climate change and energy crises issues.

Finally, drought forecast model attained in this research study can also amicably enable Governments and other relevant stakeholders in water resources and crop management, including decisions on infrastructural areas (*e.g.*, dam or irrigation operation) and other hydro-physical applications in the current phase of a changing climate where models can be used to make informed decisions.

CRediT authorship contribution statement

Mumtaz Ali: Writing - original draft, Conceptualization, Methodology, Software. **Ravinesh C. Deo:** Visualization, Conceptualization, Writing - review & editing, Investigation, Supervision. **Tek Maraseni:** Writing - review & editing. **Nathan J. Downs:** Writing - review & editing.

Acknowledgement

This paper has utilized wheat yield acquired from Bureau of Statistics, Government of Pakistan that is duly acknowledged. We acknowledge that this research project was sponsored by University of

Southern Queensland's Postgraduate Research Scholarship (2017–2019) awarded to the first author, managed by USQ Graduate Research School.

References

Adamowski, J., Fung Chan, H., Prasher, S.O., Ozga-Zielinski, B., Sliusarieva, A., 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resour. Res. 48 (1).

Adarsh, S., Sanah, S., Murshida, K.K., Nooramol, P., 2017. Scale dependent prediction of reference evapotranspiration based on Multi-Variate Empirical mode decomposition. Ain Shams Eng. J.

Adnan, M., Rehman, N., Khan, A.A., Mir, K.A., Khan, M.A., 2017. Influence of natural forcing phenomena on precipitation of Pakistan. Pakistan J. Meteorol. 12 (24).

Ahmed, K., Shahid, S., bin Harun, S., Wang, X.-J., 2016. Characterization of seasonal droughts in Balochistan Province, Pakistan. Stoch. Env. Res. Risk Assess. 30 (2), 747–762.

Alaoui, A., Mahoney, M.W., 2015. Fast randomized kernel ridge regression with statistical guarantees. Adv. Neural Information Processing Syst. 775–783.

Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2018a. An ensemble-ANFIS based uncertainty assessment model for forecasting multi-scalar standardized precipitation index. Atmos. Res. 207, 155–180.

Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2018b. Multi-stage committee based extreme learning machine model incorporating the influence of climate parameters and seasonality on drought forecasting. Comput. Electron. Agric. 152, 149–165.

Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2018c. Multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm for rainfall forecasting. Atmos. Res. 213, 450–464.

Almedeij, J., 2016. Long-term periodic drought modeling. Stoch. Env. Res. Risk Assess. 30 (3), 901–910.

Alvanitopoulos, P.-F., Andreadis, I., Georgoulas, N., Zervakis, M. and Nikolaidis, N., 2014. Solar Radiation Time-Series Prediction Based on Empirical Mode Decomposition and Artificial Neural Networks. In: 10th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI). IFIP Advances in Information and Communication Technology, AICT-436. Springer, Rhodes, Greece, pp. 447-455.

Andersson, L.E., Aftab, M.F., Scibilia, F., Imsland, L., 2017. Forecasting using multivariate empirical mode decomposition-Applied to iceberg drift forecast, IEEE Conference on Control Technology and Applications (CCTA), Kohala Coast. Hawai'i, USA.

Andreoli, R.V., Kayano, M.T., 2005. ENSO-related rainfall anomalies in South America and associated circulation features during warm and cold Pacific decadal oscillation regimes. Int. J. Climatol. 25 (15), 2017–2030.

ASCE, 1993. Criteria for Evaluation of Watershed Models. J. Irrig. Drain. Eng. 119 (3), 429–442.

ASCE, 2000. Artificial neural networks in hydrology. II: Hydrologic applications. J. Hydrologic Eng.' 5 (2), 124–137.

Ascough Ii, J., Maier, H., Ravalico, J., Strudley, M., 2008. Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. Ecol. Model. 219 (3–4), 383–399.

Ashok, K., Behera, S.K., Rao, S.A., Weng, H., Yamagata, T., 2007. El Niño Modoki and its possible teleconnection. J. Geophys. Res. Oceans 112 (C11).

Ashok, K., Guan, Z., Yamagata, T., 2001. Impact of the Indian Ocean dipole on the relationship between the Indian monsoon rainfall and ENSO. Geophys. Res. Lett. 28 (23), 4499–4502.

Athier, G., Floquet, P., Pibouleau, L., Domenech, S., 1997. Synthesis of heat-exchanger network by simulated annealing and NLP procedures. AIChE J. 43 (11), 3007–3020.

Badr, A., Fahmy, A., 2004. A proof of convergence for ant algorithms. Inf. Sci. 160 (1–4), 267–279.

BAS, 2018. British Antarctic Survey.

Bates, B., Kundzewicz, Z.W., Wu, S., Palutikof, J., 2008. Climate change and water: Technical paper vi. Intergovernmental Panel on Climate Change (IPCC).

Bergmeir, C., Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Inf. Sci. 191, 192–213.

BMA, 2018. Bureau of Meteorology, Australia.

Bonaccorso, B., Bordi, I., Cancelliere, A., Rossi, G., Sutera, A., 2003. Spatial variability of drought: an analysis of the SPI in Sicily. Water Resour. Manage. 17 (4), 273–296.

Breiman, L., 1996. Bagging predictors. Machine learning 24 (2), 123–140.

Breiman, L., 2001. Random forests. Machine learning 45 (1), 5–32.

Byun, H.-R., Wilhite, D.A., 1999. Objective quantification of drought severity and duration. J. Climatol. 12 (9), 2747–2756.

Cai, W., Cowan, T., 2008. Dynamics of late autumn rainfall reduction over southeastern Australia. Geophys. Res. Lett. 35 (9).

Cancelliere, A., Bonaccorso, B., Mauro, G., 2006. A non-parametric approach for drought forecasting through the standardized precipitation index. Metodi statisticie matematici per l Analisi delle serie idrologiche 1 (1), 1–8.

Cannas, B., Fanni, A., See, L., Sias, G., 2006. Data preprocessing for river flow forecasting using neural networks: wavelet transforms and data partitioning. Phys. Chem. Earth, Parts A/B/C 31 (18), 1164–1171.

Chen, H., Yang, D., Hong, Y., Gourley, J.J., Zhang, Y., 2013. Hydrological data assimilation with the Ensemble Square-Root-Filter: Use of streamflow observations to update model states for real-time flash flood forecasting. Adv. Water Resour. 59, 209–220.

Chen, J., Li, M. and Wang, W., 2012. Statistical uncertainty estimation using random forests and its application to drought forecast. Mathematical Problems in Engineering, 2012.

Chiew, F.H., Piechota, T.C., Dracup, J.A., McMahon, T.A., 1998. El Nino/Southern Oscillation and Australian rainfall, streamflow and drought: links and potential for forecasting. J. Hydrol. 204 (1), 138–149.

Choubin, B., Malekian, A., Golshan, M., 2016. Application of several data-driven techniques to predict a standardized precipitation index. Atmósfera 29 (2), 121–128.

Colominas, M.A., Schlotthauer, G., Torres, M.E., 2014. Improved complete ensemble EMD: a suitable tool for biomedical signal processing. Biomed. Signal Process. Control 14, 19–29.

Cordón García, O., Herrera Triguero, F. and Stützle, T., 2002. A review on the ant colony optimization metaheuristic: Basis, models and new trends. Mathware & Soft Computing. 2002 Vol. 9 Núm. 2 [-3].

Dai, A., 2013. The influence of the inter-decadal Pacific oscillation on US precipitation during 1923–2010. Clim. Dyn. 41 (3–4), 633–646.

Dawson, C.W., Abrahart, R.J., See, L.M., 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ. Modell. Software 22 (7), 1034–1052.

Deo, R.C., Kisi, O., Singh, V.P., 2017a. Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. Atmos. Res. 184, 149–175.

Deo, R., C and Şahin, M.,, 2016. An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland. Environ. Monit. Assess. https://doi.org/10.1007/s10661-016-5094-9.

Deo, R.C., Downs, N., Parisi, A., Adamowski, J., Quilty, J., 2017b. Very short-term reactive forecasting of the solar ultraviolet index using an extreme learning machine integrated with the solar zenith angle. Environ. Res. 155, 141–166.

Deo, R.C., et al., 2009. Impact of historical land cover change on daily indices of climate extremes including droughts in eastern Australia. Geophys. Res. Lett. 36 (8).

Deo, R.C., Tiwari, M.K., Adamowski, J.F., Quilty, J.M., 2017c. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. Stoch. Env. Res. Risk Assess. 31 (5), 1211–1240.

Deo, R.C., Wen, X., Qi, F., 2016. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Appl. Energy 168, 568–593.

Deo, R.C. et al., 2019. Two-phase extreme learning machines integrated with complete ensemble empirical mode decomposition with adaptive noise for multi-scale runoff prediction. J. Hydrol., In Press (03-Jan-2019).

Department, P.M., 2010. Dry weather predicted in the country during Friday/Monday.

Dickson, R., et al., 2000. The Arctic ocean response to the North Atlantic oscillation. J. Clim. 13 (15), 2671–2696.

Dijk, A.I., et al., 2013. The Millennium Drought in southeast Australia (2001–2009): natural and human causes and implications for water resources, ecosystems, economy, and society. Water Resour. Res. 49 (2), 1040–1057.

Elleithy, K.M., Fattah, E., 2012. A Simulated Annealing Algorithm for Register Allocation. College of Engineering & Islamic Architecture, Umm al-Qura University.

Gebremariam, S.Y., et al., 2014. A comprehensive approach to evaluating watershed models for predicting river flow regimes critical to downstream ecosystem services. Environ. Modell. Software 61, 121–134.

Guttman, N.B., 1999. Accepting the standardized precipitation index: a calculation algorithm. JAWRA J. Am. Water Resour. Assoc. 35 (2), 311–322.

Hayes, M.J., Svoboda, M.D., Wilhite, D.A., Vanyarkho, O.V., 1999. Monitoring the 1996 drought using the standardized precipitation index. Bull. Am. Meteorol. Soc. 80 (3), 429–438.

He, K., Zha, R., Wu, J., Lai, K., 2016. Multivariate EMD-Based Modeling and Forecasting of Crude Oil Price. Sustainability 8 (4), 387.

He, Z., Wen, X., Liu, H., Du, J., 2014. A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. J. Hydrol. 509, 379–386.

Hong, Y.-Y., Yu, T.-H., Liu, C.-Y., 2013. Hour-Ahead Wind Speed and Power Forecasting Using Empirical Mode Decomposition. Energies 6 (12), 6137–6152.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J., 2003. A practical guide to support vector classification.

Hu, W., Si, B.C., 2013. Soil water prediction based on its scale-specific control using multivariate empirical mode decomposition. Geoderma 193, 180–188.

Huang, J.-R., et al., 2013. Application of multivariate empirical mode decomposition and sample entropy in EEG signals via artificial neural networks for interpreting depth of anesthesia. Entropy 15 (9), 3325–3339.

Huang, N.E., et al., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proc. R. Soc. A 454, 903–995.

Hurrell, J.W., 1995. Decadal trends in the North Atlantic Oscillation: regional temperatures and precipitation. Science 269 (5224), 676–679.

IPCC, 2012. Summary for Policymakers: A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change. In: Field, C.B. (Ed.), Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation. Cambridge University Press, Cambridge, UK.

Jalalkamali, A., Moradi, M., Moradi, N., 2015. Application of several artificial intelligence models and ARIMAX model for forecasting drought using the Standardized Precipitation Index. Int. J. Environ. Sci. Technol. 12 (4), 1201–1210.

JAMSTEC, 2018. Japan Agency for Marine-Earth Science.

JISAO, 2018. Joint Institute of the Study of the Atmosphere and Ocean.

Keyantash, J., Dracup, J.A., 2002. The quantification of drought: an evaluation of drought indices. Bull. Am. Meteorol. Soc. 83 (8), 1167–1180.

Khan, M., Gadiwala, M., 2013. A Study of Drought over Sindh (Pakistan) Using Standardized Precipitation Index (SPI) 1951 to 2010. Pakistan J. Meteorol. 9 (18).

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in

hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241.
Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R news 2 (3), 18–22.
Looney, D., Mandic, D.P., 2009. Multiscale Image Fusion Using Complex Extensions of EMD. IEEE Trans. Signal Process. 57 (4), 1626–1630.
Mandic, D.P., ur Rehman, N., Wu, Z., Huang, N.E., 2013. Empirical mode decomposition-based time-frequency analysis of multivariate signals: the power of adaptive data analysis. IEEE Signal Process Mag. 30 (6), 74–86.
McAlpine, C., et al., 2007. Modeling the impact of historical land cover change on Australia's regional climate. Geophys. Res. Lett. 34 (22).
McBride, J.L., Nicholls, N., 1983. Seasonal relationships between Australian rainfall and the Southern Oscillation. Mon. Weather Rev. 111 (10), 1998–2004.
McGregor, S., et al., 2014. Recent Walker circulation strengthening and Pacific cooling amplified by Atlantic warming. Nat. Clim. Change 4 (10), 888.
McKee, T.B., Doesken, N.J., Kleist, J., 1993. In: The relationship of drought frequency and duration to time scales. American Meteorological Society, Boston, MA, pp. 179–183.
Mishra, A.K., Singh, V.P., 2010. A review of drought concepts. J. Hydrol. 391 (1–2), 202–216.
Mishra, A.K., Singh, V.P., 2011. Drought modeling–A review. J. Hydrol. 403 (1), 157–175.
Moore, I.D., Gessler, P., Nielsen, G., Peterson, G., 1993. Soil attribute prediction using terrain analysis. Soil Sci. Soc. Am. J. 57 (2), 443–452.
Moore, I.D., Grayson, R., Ladson, A., 1991. Digital terrain modelling: a review of hydrological, geomorphological, and biological applications. Hydrol. Process. 5 (1), 3–30.
Moreira, E., Martins, D., Pereira, L., 2015. Assessing drought cycles in SPI time series using a Fourier analysis. Nat. Hazards Earth Syst. Sci. 15 (3), 571–585.
Moreira, E.E., Coelho, C.A., Paulo, A.A., Pereira, L.S., Mexia, J.T., 2008. SPI-based drought category prediction using loglinear models. J. Hydrol. 354 (1), 116–130.
Morid, S., Smakhtin, V., Bagherzadeh, K., 2007. Drought forecasting using artificial neural networks and time series of drought indices. Int. J. Climatol. 27 (15), 2103–2111.
Mouatadid, S., Raj, N., Deo, R.C., Adamowski, J.F., 2018. Input selection and data-driven model performance optimization for predicting Standardized Precipitation and Evaporation Index in a drought-prone region. Atmos. Res.
Mullen, R.J., Monekosso, D., Barman, S., Remagnino, P., 2009. A review of ant algorithms. Expert Syst. Appl. 36 (6), 9608–9617.
Nicholls, N., 1983. Predicting Indian monsoon rainfall from sea-surface temperature in the Indonesia–north Australia area. Nature 306 (5943), 576.
Nicholls, N., 2004. The changing nature of Australian droughts. Clim. Change 63 (3), 323–336.
Ouyang, Q., et al., 2016. Monthly Rainfall Forecasting Using EEMD-SVR Based on Phase-Space Reconstruction. Water Resour. Manage. 30 (7), 2311–2325.
Özger, M., Mishra, A.K., Singh, V.P., 2012. Long lead time drought forecasting using a wavelet and fuzzy logic combination model: a case study in Texas. J. Hydrometeorol. 13 (1), 284–297.
Pal, J., Chaudhuri, S., Roychowdhury, A., Basu, D., 2017. An investigation of the influence of the southern annular mode on I ndian summer monsoon rainfall. Meteorol. Appl. 24 (2), 172–179.
Palmer, W.C., 1965. Meteorological drought, 30. US Department of Commerce, Weather Bureau Washington, DC.
Palmer, W.C., 1968. Keeping track of crop moisture conditions, nationwide: the new crop moisture index.
Paulo, A.A., Pereira, L.S., 2007. Prediction of SPI drought class transitions using Markov chains. Water Resour. Manage. 21 (10), 1813.
Peng, M., Gupta, N.K., Armitage, A.F., 1996. An investigation into the improvement of local minima of the Hopfield network. Neural networks 9 (7), 1241–1253.
Philander, S.G.H., 1983. El Nino southern oscillation phenomena. Nature 302 (5906), 295.
PMD, 2016. Pakistan Meteorological Department, Pakistan.
Prasad, R., Ali, M., Kwan, P., Khan, H., 2019. Designing a multi-stage multivariate empirical mode decomposition coupled with ant colony optimization and random forest model to forecast monthly solar radiation. Appl. Energy 236, 778–792.
Prasad, R., Deo, R.C., Li, Y., Maraseni, T., 2017. Input selection and performance optimization of ANN-based streamflow forecasts in a drought-prone Murray Darling Basin using IIS and MODWT algorithm. Atmos. Res. 197, 42–63.
Prasad, R., Deo, R.C., Li, Y., Maraseni, T., 2018. Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. Geoderma 330, 136–161.
Priya, P., Mujumdar, M., Sabin, T., Terray, P., Krishnan, R., 2015. Impacts of Indo-Pacific sea surface temperature anomalies on the summer monsoon circulation and heavy precipitation over northwest India-Pakistan region during 2010. J. Clim. 28 (9), 3714–3730.
Quilty, J., Adamowski, J., 2018. Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models for real-world applications with best practices and a new forecasting framework. J. Hydrol.
Rehman, N., Mandic, D.P., 2009. Multivariate empirical mode decomposition. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 466 (2117), 1291–1302.
Rehman, N., Mandic, D.P., 2009b. Multivariate empirical mode decomposition, Proceedings of The Royal Society of London A: Mathematical, Physical and Engineering Sciences. The Royal Society, pp. rspa20090502.
Ren, Y., Suganthan, P.N., Srikanth, N., 2015. A comparative study of empirical mode decomposition-based short-term wind speed forecasting methods. IEEE Trans. Sustainable Energy 6 (1), 236–244.
Report, 1950-2015. List of natral disaster.
Robert, B., Yoav, F., Peter, B., Sun, L.W., 1998. Boosting the margin: a new explanation for the effectiveness of voting methods. The annals of statistics 26 (5), 1651–1686.
Romagnoli, M., et al., 2017. Assessment of the SWAT model to simulate a watershed with limited available data in the Pampas region, Argentina. Sci. Total Environ. 596, 437–450.
Salinger, M., Renwick, J., Mullan, A., 2001. Interdecadal Pacific oscillation and south Pacific climate. Int. J. Climatol. 21 (14), 1705–1721.
Santos, C.A.G., Morais, B.S., Silva, G.B., 2009. Drought forecast using an artificial neural network for three hydrological zones in San Francisco River basin, Brazil. IAHS publication 333, 302.
Schapire, R.E., Freund, Y., Bartlett, P., Lee, W.S., 1998. Boosting the margin: a new explanation for the effectiveness of voting methods. Ann. Stat. 1651–1686.
Segal, M.R., 2004. Machine learning benchmarks and random forest regression.
Servey, B., 2016. Asian Urban Information of Kobe.
She, D., Zheng, J., Shao, M.a., Timm, L.C. and Xia, Y.,, 2015. Multivariate empirical mode decomposition derived multi-scale spatial relationships between saturated hydraulic conductivity and basic soil properties. CLEAN–Soil, Air. Water 43 (6), 910–918.
Shirmohammadi, B., Moradi, H., Moosavi, V., Semiromi, M.T., Zeinali, A., 2013. Forecasting of meteorological drought using Wavelet-ANFIS hybrid model for different time steps (case study: southeastern part of east Azerbaijan province, Iran). Nat. Hazards 69 (1), 389–402.
Singh, G., Kumar, N., Verma, A.K., 2012. Ant colony algorithms in MANETs: a review. J. Network Comput. Appl. 35 (6), 1964–1972.
Sommerlot, A.R., Wagena, M.B., Fuka, D.R., Easton, Z.M., 2016. Coupling the short-term global forecast system weather data with a variable source area hydrologic model. Environ. Modell. Software 86, 68–80.
Sönmez, F.K., KÖmÜscÜ, A.U., Erkan, A., Turgu, E., 2005. An analysis of spatial and temporal dimension of drought vulnerability in Turkey using the standardized precipitation index. Nat. Hazards 35 (2), 243–264.
Souriau, A., Yiou, P., 2001. Grape harvest dates for checking NAO paleoreconstructions. Geophys. Res. Lett. 28 (20), 3895–3898.
SST, 2018. National Climate Prediction Centre.
Sun, S., et al., 2012. Past and future changes of streamflow in Poyang Lake Basin. Southeastern China.
Svoboda, M., Hayes, M., Wood, D., 2012. Standardized precipitation index user guide. World Meteorological Organization Geneva, Switzerland.
Sweetlin, J.D., Nehemiah, H.K., Kannan, A., 2017. Feature selection using ant colony optimization with tandem-run recruitment to diagnose bronchitis from CT scan images. Comput. Methods Programs Biomed. 145, 115–125.
Taniguchi, M., Kurokawa, K., Itoh, K., Matsuoka, K., Ichioka, Y., 1997. Sidelobeless multiple-object discriminant filters recorded as discrete-type computer-generated holograms. Appl. Opt. 36 (35), 9138–9145.
Terray, P., Chauvin, F., Douville, H., 2007. Impact of southeast Indian Ocean sea surface temperature anomalies on monsoon-ENSO-dipole variability in a coupled ocean–atmosphere model. Clim. Dyn. 28 (6), 553–580.
Torres, M.E., Colominas, M.A., Schlotthauer, G., Flandrin, P., 2011. A complete ensemble empirical mode decomposition with adaptive noise. In: 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4144–4147.
Ur Rehman, N., Mandic, D.P., 2011. Filter bank property of multivariate empirical mode decomposition. IEEE Trans. Signal Process. 59 (5), 2421–2426.
Vicente-Serrano, S., 2016. Foreword: Drought complexity and assessment under climate change conditions. Cuadernos de Investigación Geográfica 42 (1), 7–11.
Vovk, V., 2013. Kernel ridge regression. Empirical inference. Springer 105–116.
Wang, W.-C., Xu, D.-M., Chau, K.-W., Chen, S., 2013. Improved annual rainfall-runoff forecasting using PSO–SVM model based on EEMD. J. Hydroinf. 15 (4), 1377–1390.
Wang, Z., Tian, C., Zhu, Q., Huang, M., 2018. Hourly Solar Radiation Forecasting Using a Volterra-Least Squares Support Vector Machine Model Combined with Signal Decomposition. Energies 11 (1), 68.
Welling, M., 2013. Kernel ridge regression. Max Welling's Classnotes in Machine Learning 1–3.
Wilhite, D.A., Hayes, M.J., Knutson, C., Smith, K.H., 2000. Planning for drought: Moving from crisis to risk management1. Wiley Online Library.
Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2 (2), 184–194.
Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 63 (11), 1309–1313.
Willmott, C.J., 1984. On the evaluation of model performance in physical geography. Spatial statistics and models. Springer 443–460.
Wilson, J.D., 1997. A simulated annealing algorithm for optimizing RF power efficiency in coupled-cavity traveling-wave tubes. IEEE Trans. Electron Devices 44 (12), 2295–2299.
Wittwer, G., Adams, P.D., Horridge, M., Madden, J.R., 2002. Drought, regions and the Australian economy between 2001–02 and 2004–05. Australian Bulletin of Labour 28 (4), 231.
Wu, Z., Huang, N.E., 2009. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv. Adaptive Data Analysis 1 (1), 1–41.
Wu, Z., Huang, N.E., Wallace, J.M., Smoliak, B.V., Chen, X., 2011. On the time-varying trend in global-mean surface temperature. Clim. Dyn. 37 (3–4), 759–773.
Xie, H., Ringler, C., Zhu, T., Waqas, A., 2013. Droughts in Pakistan: a spatiotemporal variability analysis using the Standardized Precipitation Index. Water Int. 38 (5), 620–631.
Yaseen, Z.M., et al., 2018a. Application of the Hybrid Artificial Neural Network Coupled with Rolling Mechanism and Grey Model Algorithms for Streamflow Forecasting Over

Multiple Time Horizons. Water. Resour. Manag. 1–17.
Yaseen, Z.M., Sulaiman, S.O., Deo, R.C., Chau, K.-W., 2018b. An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction. J. Hydrol.
Yen, B.C., 1995. Discussion and Closure: Criteria for Evaluation of Watershed Models. J. Irrig. Drain. Eng. 121 (1), 130–132.
You, Y., Demmel, J., Hsieh, C.-J. and Vuduc, R., 2018. Accurate, Fast and Scalable Kernel Ridge Regression on Parallel and Distributed Systems. arXiv preprint arXiv:1805.00569.
Yuan, W.-P., Zhou, G.-S., 2004. Comparison between standardized precipitation index and Z-index in China. Acta Phytoecologica Sinica 4.
Zhang, Y., Duchi, J., Wainwright, M., 2013. Divide and conquer kernel ridge regression. Conference on Learning Theory 592–617.

Chapter 7

Cotton yield prediction with Markov Chain Monte Carlo-based simulation model integrated with genetic programing algorithm: A new hybrid copula driven approach

Foreword

This chapter is an exact copy of the published article in *Agricultural and Forest Meteorology* journal (Vol. 263, Pages 428-448).

The agricultural crop yields have been severely impacted under the recent climate change and global warming. Thus, a cotton yield prediction with Markov Chain Monte Carlo-based simulation model integrated with genetic programing algorithm using multiple meteorological data of rainfall, temperature and humidity. Several different types of GP-MCMC-copula models were developed, each with the well-known copula families (i.e., Gaussian, student t, and Clayton, Gumble Frank and Fischer-Hinzmann functions) to screen and utilize an optimal cotton yield forecast model for the present study region.

The results of hybrid GP-MCMC-copula models were compared against different MCMC based copula models as well as GP model. The performances of GP-MCMC-copula models are better than comparative counterparts. The developed GP-MCMC-Copula modelling framework can be adopted by farmers, agricultural system modelling experts and agricultural policy-makers in strategic decision-making processes.

Agricultural and Forest Meteorology 263 (2018) 428–448

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

Cotton yield prediction with Markov Chain Monte Carlo-based simulation model integrated with genetic programing algorithm: A new hybrid copula-driven approach

Mumtaz Ali*, Ravinesh C. Deo*, Nathan J. Downs, Tek Maraseni

School of Agricultural, Computational and Environmental Sciences, Institute of Life Sciences and the Environment, University of Southern Queensland, Springfield, QLD, 4300, Australia

ARTICLE INFO

Keywords:
Crop yield prediction
Cotton yield
Climate data
Genetic programming
Markov Chain Monte Carlo based copula model

ABSTRACT

Reliable data-driven models designed to accurately estimate cotton yield, an important agricultural commodity, can be adopted by farmers, agricultural system modelling experts and agricultural policy-makers in strategic decision-making processes. In this paper a hybrid genetic programing model integrated with the Markov Chain Monte Carlo (MCMC) based Copula technique is developed to incorporate climate-based inputs as the predictors of cotton yield, for selected study regions: Faisalabad (31.4504 °N, 73.1350 °E), Multan (30.1984 °N, 71.4687 °E) and Nawabshah (26.2442 °N, 68.4100 °E), as important cotton growing hubs in the developing nation of Pakistan. Several different types of GP-MCMC-copula models were developed, each with the well-known copula families (*i.e.*, Gaussian, student t, Clayton, Gumble Frank and Fischer-Hinzmann functions) to screen and utilize an optimal cotton yield forecast model for the present study region. The results of the GP-MCMC based hybrid copula model were evaluated with a standalone GP and the MCMC based copula model in accordance with statistical analysis of the predicted yield based on correlation coefficient (r), Willmott's index (WI), Nash-Sutcliffe coefficient (NS_E), root mean squared error (RMSE) and mean absolute error (MAE) in the independent test phase. Further performance preciseness was evaluated by the Akiake Information Criterion (AIC), the Bayesian Information Criterion (BIC) and the Maximum Likelihood (Max_L) for the GP-MCMC based copula as well as the MCMC based copula model. GP-MCMC-Clayton copula model generated the most accurate result for the Multan station. For the optimal GP-MCMC-Clayton copula model, the acquired model evaluation metrics for Multan were: (LM ≈ 0.952; RRMSE ≈ 2.107%; RRMAE ≈ 1.771%) followed by the MCMC based Gaussian copula model (LM ≈ 0.895; RRMSE ≈ 4.541%; RRMAE ≈ 0.3.214%) and the standalone GP model (LM ≈ 0.132; RRMSE ≈ 23.638%; RRMAE ≈ 22.652%), indicating the superiority of the GP-MCMC-Clayton copula model in respect to the other benchmark models. The performance of GP-MCMC based copula model was also found to be superior in the case of Faisalabad and Nawabshah station as confirmed by AIC, BIC, Max_L metrics, including a larger value of the Legates-McCabe's (LM) index, utilized in conjunction with the relative percentage RRMSE and the relative mean absolute error (RMAE). Accordingly, it is averred that the developed GP-MCMC copula model can be considered as a pertinent data-intelligent tool used for accurate prediction of cotton yield, utilizing the readily available climate datasets in agricultural regions and is of relevance to agricultural yield simulation and sectoral decision-making.

1. Introduction

Timely information on the crop yield is important for agriculture-dependent nations (e.g., Pakistan), as this can generate crucial ideas for agricultural policy making, and forward planners and agricultural markets. Agriculture in Pakistan is known to contribute to about 21% of the county's GDP (Sarwar, 2014), which include cotton as an important cash crop. This is because cotton is an integral commodity for the economic development of Pakistan as the nation is highly dependent on the cotton industry and its related textile sector due to which the cotton crop has been given a principal status in the country. Cotton crop is grown from May-August as an industrial crop in 15% of the nation's available land area producing 15 million bales during 2014-15 (Reporter, 2015). Pakistan is placed at fourth position among cotton

* Corresponding authors.
E-mail addresses: Mumtaz.Ali@usq.edu.au (M. Ali), ravinesh.deo@usq.edu.au (R.C. Deo).

https://doi.org/10.1016/j.agrformet.2018.09.002
Received 21 January 2018; Received in revised form 18 July 2018; Accepted 1 September 2018
Available online 25 September 2018
0168-1923/

growers, third largest exporter and fourth largest consumer (Banuri, 1998). In 2013, about 1.6 million farmers (out of a total of 5 million in all sectors) engaged in cotton farming, growing more than 3 million hectares (Banuri, 1998; Reporter, 2015).

Data-intelligent models, utilizing past data can offer an accurate solutions to the problems related to the projection of future trends in agriculture, crop yield, rainfall and drought that affects agricultural productivity (Ali et al., 2018a, b; Bauer, 1975; Nguyen-Huy et al., 2017, 2018). Machine learning models, which are highly non-linear models, utilize data that has input features valued for the prediction of crop yield. In the work of Kern et al. (2018), multiple linear regression models were constructed to simulate the yield of the four major crop types in Hungary using environmental and remote sensing information. Moreover, Bokusheva et al. (2016) developed copula models for crop yields on VH indices and Craparo et al. (2015) built an ARIMA model to forecast the decline of coffee yield in Tanzania. Debnath et al. (2013) predicted area and cotton yield in India using an ARIMA model. Blanc et al. (2008) utilized a multiple regression model of the main climatic determinants of rain fed cotton yield in West Africa. Yang et al. (2014) assessed cotton yield and water demand under climate change and future adaptation measures using APSIM-OzCot model. Chen et al. (2011) studied the impact of climate change on cotton production and water consumption using COSIM model in China. Hearn (1994) design a simulation model named OZCOT for cotton crop management in Australia. Papageorgiou et al., (2011) predicted cotton yield using fuzzy cognitive maps in 2011, Greece. Jin and Xu (2012) conducted a study on the estimation of cotton yield using Carnegie Ames Stanford Approach model in China. The aforementioned models were developed to study the climate change impacts on cotton yield prediction.

In summary, existing literature shows that there are few studies in Pakistan that have developed methods for the prediction of cotton yield, despite its relevance as a world leader in cotton production. Ali et al. (2015) used a forecasting ARIMA model for the production of sugarcane and cotton crops of Pakistan from 2013–2030. Hina Ali et al. (2013) also analyzed production forecasting of cotton in Pakistan. Ahmad et al. (2017) developed an ARIMA model to forecast area, production and yield of major crops in Pakistan in 2017. Raza and Ahmad (2015) studied the impact of climate change on cotton productivity in Punjab and Sindh, Pakistan using fixed effect models. Ayaz et al. (2015) studied weather effect on cotton crop in Sindh, Pakistan. Carpio and Ramirez (2002) used yield and acreage models to forecast cotton yield in India, Pakistan and Australia. Ahmad (1975) designed a time series prediction for the supply response of cotton in Punjab, Pakistan in 1975.

All the previous studies indicate that the prediction of cotton yields have been based primarily on the effect of climate change with the adoption of ARIMA model only. In addition to that, all these studies have been conducted for a large area, either for a whole province, or national region, but not for a small locality. Moreover, there is a limitation of applying advanced data-intelligent algorithms for more accurate prediction models at a micro scale which can provide help for decision-making in precision agriculture and farming systems which may be the way future farming trends are analyzed. To address these mentioned issues, there is an apparent need for data intelligent models to predict cotton yield more accurately and at a much finer scale than attempted previously. In this study, for the first time, a hybrid genetic programing integrated with a Markov Chain Monte Carlo (GP-MCMC) based copula model has been developed for the prediction of cotton yield in Faisalabad, Multan and Nawabshah in Pakistan. The novelty of this study is to utilize as yet untested accurate GP-MCMC based copula models for the prediction of cotton yield in Pakistan.

To advance the application of copula models, especially in agriculture where they have been relatively scarcely applied the present study aims to address four primary objectives. (1) To apply GP and MCMC based copula, MCMC based copula models and a standalone GP model to determine which is of these models is the most accurate data-intelligent tool for predicting cotton yield in the developing nation of Pakistan. (2) To model influence of climate dataset (*i.e.*, temperature, rainfall and humidity) to predict effectively the cotton yield in the proposed districts of Punjab and Sindh, the primary agricultural hubs in Pakistan. (3) To develop and optimize the copula-based models by tuning the GP and the MCMC techniques as well as to evaluate their performances in comparison with MCMC based copula and standalone GP model. (4) To validate the predictive ability of each model with respect to cotton yield in Pakistan, making a major contribution to the use of data-driven models for agricultural yield estimation.

2. Theoretical framework

In this section an overview of the proposed predictive GP-MCMC based copula models with its comparative counterparts, MCMC based copula models and GP are presented.

2.1. Genetic programming (GP) model

Genetic programming is a heuristic evolutionary algorithm which has the potential to offer solutions of any form without the user specifying the problem (Deo and Samui, 2017; Koza, 1992). Evolutionary principles are utilized to acquire the persistent patterns in the structure of data without requiring prior knowledge. According to McPhee et al. (2008), an organized domain-independent method is used to genetically breed a population in genetic programming for getting computers to solve the problem that is starting from a high-level statement of what needs to be done. Fig. 1(a) demonstrates the basic structure of a GP model. More specifically, a population is transformed iteratively to produce successive new generations of programs by using similar genetic operations that occur naturally. These genetic operations are divided into five components: crossover (sexual recombination), mutation; reproduction; gene duplication; and gene deletion. Huang et al. (2006) showed that a GP model has the skill of self-parameterizing to extract the features bypassing the user, tuning the model, and due to this capability resembles to some extent the Extreme Learning Machine model.

In a GP model, the input data goes through a number of routes where (1) analyzation of attributes occurs; (2) selection of the best fitness functions are made for the purpose of minimizing the mean-squared error; (3) generation of functional and terminal sets; and (4) parameterization of genetic operations (Sreekanth and Datta, 2011). The GP model is optimized by the emulation of an evolutionary process to an adequate agreement between the response and input variable. A functional node performs the arithmetic operations (þ, −, ×, ÷), Boolean logic functions (AND, OR, NOT), conditionals (IF, THEN, ELSE), or any other functions (SIN, EXP) that may be used. A random tree structure is developed using these functions (Deo and Samui, 2017; Mehr et al., 2013). GP is developed in this paper by (1) randomly creating the initial population (i.e., computer program); (2) performing the execution of the program with the best fitness values; (3) based on reproduction, mutation, and crossover, generation of a new population of computer programs; (4) comparison and evaluation of fitness; and (5) finally the selection of the best program by the evolutionary process (Mehr et al., 2013). For this purpose, a randomly equated population is being created and best fitness is determined where “parents” are chosen individually and the “off-springs” are developed from the parents through the process of reproduction, mutation, and crossover (Deo and Samui, 2017).

2.2. Markov Chain Monte Carlo (MCMC) based copula models

In this study we hybridize the MCMC-copula models used previously (*e.g.*, (Ali et al., 2018b)) with the GP algorithm. Basically, a copula model, which has recently found important applications in the agricultural sector, is a powerful mathematical tool that has the ability to connect two or more time-independent variables (Nelsen, 2003; Nguyen-Huy et al., 2017, 2018). A copula function is basically a

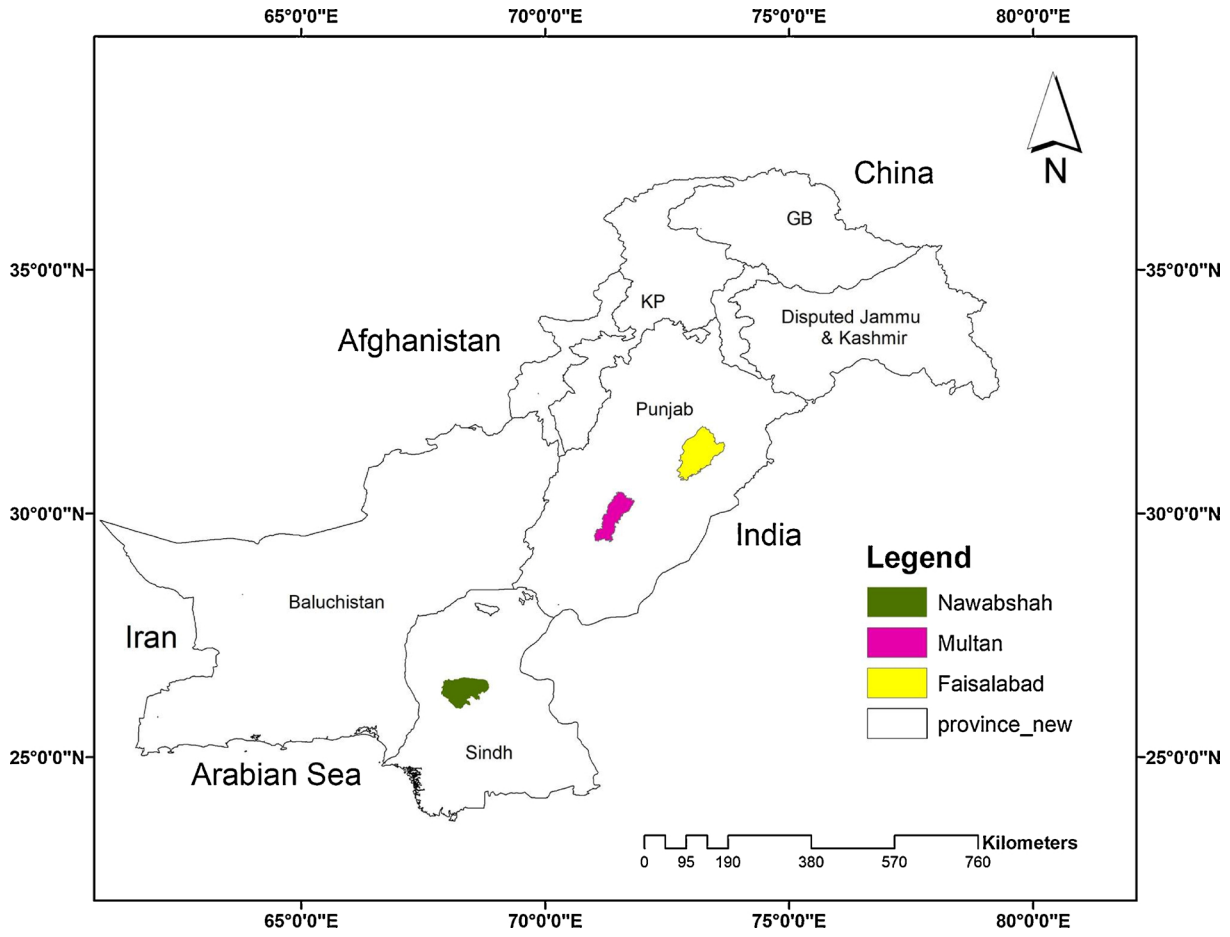


Fig. 1. Map of the selected study region in Pakistan.

mathematical function that is defined from $I^2(F, G)$ to $I(H)$ such that $[F(x), G(y), H(x, y)]$ is a point in I^3 with $I \in [0, 1]$ and X, Y are continuous random variables with distribution functions $F(x) = P(X \leq x)$ and $G(y) = P(Y \leq y)$, and $H(x, y) = P(X \leq x, Y \leq y)$ is a function that describes their joint distribution.

In this paper we utilize those types of copulas that are available widely in the literature such as Gaussian, student t, Clayton, Fischer-Hinzmann, Frank and Gumble copulas that are shown in the following Equations where,

I the Gaussian copula (Li et al., 2013) is expressed as:

$$\int_{-\infty}^{\phi^{-1}(a)} \int_{-\infty}^{\phi^{-1}(b)} \frac{1}{2\pi\sqrt{1-\theta^2}} \exp\left(\frac{2\theta xy - x^2 - y^2}{2(1-\theta^2)}\right) dxdy\,,\ \theta \in [-1, 1] \tag{1}$$

II a student t-copula (Li et al., 2013) can be formulated as:

$$\int_{-\infty}^{t_{\theta_2}^{-1}(a)} \int_{-\infty}^{t_{\theta_2}^{-1}(b)} \frac{\Gamma\left({}^{(\theta_2+2)}/_{2}\right)}{\Gamma\left({}^{\theta_2}/_{2}\right)\pi\theta_2\sqrt{1-\theta_1^2}} \left(1 + \frac{x^2 - 2\theta_1 xy + y^2}{\theta_2}\right)^{{}^{(\theta_2+2)}/_{2}} dxdy,\ \theta_1 \in [-1, 1], \theta_2 \in (0, \infty) \tag{2}$$

III a Clayton copula (Clayton, 1978) can be written as:

$$\max(a^{-\theta} + b^{-\theta} - 1, 0)^{-{}^{1}/_{\theta}},\ \theta_2 \in [-1, \infty)\backslash 0 \tag{3}$$

IV a Frank copula (Li et al., 2013) has the following mathematical formulation:

$$-\frac{1}{\theta}\ln\left(1 + \frac{(\exp(-\theta a) - 1)(\exp(-\theta b) - 1)}{\exp(-\theta) - 1}\right),\ \theta \in \mathbb{R}\backslash 0 \tag{4}$$

V a Gumble copula (Li et al., 2013) can be expressed as:

$$\exp\left(-(((-\ln(a))^{\theta}) + (-\ln(b))^{\theta})^{\frac{1}{\theta}}\right),\ \theta \in [1, \infty) \tag{5}$$

VI a Fischer-Hinzmann copula (Fischer and Hinzmann, 2006) can be given as:

$$[\theta_1(\min(a, b))^{\theta_2} + (1 - \theta_1)(ab)^{\theta_2}]^{\frac{1}{\theta_2}},\ \theta_1 \in [0, 1], \theta_2 \in \mathbb{R} \tag{6}$$

In all types of copula-based models an unknown process κ links observation $\widetilde{Y}$ to parameters θ^* in the modelling inference analysis (Mojtaba Sadegh and AghaKou, 2017) and these can be given through the following equation.

$$\widetilde{Y} = \kappa(\theta^*) + \xi \tag{7}$$

Where ξ indicates a vector of measurement errors. The vector $e = \widetilde{Y} - Y$ is called the error residual and $e = \{e_1, e_2, \cdots, e_n\}$ where n is the number observations that include the effects of model structural errors (Mojtaba Sadegh and AghaKou, 2017). Bayesian analysis is going to be carried for model inference and uncertainty quantification purposes because Bayesian analysis quantifies uncertainty with a probability distribution (Mojtaba Sadegh and AghaKou, 2017).

Bayes' law attributes all modelling uncertainties to the parameters and estimates the posterior distribution of model parameters by the

following equations (Mojtaba Sadegh and AghaKou, 2017):

$$p(\theta|\widetilde{Y}) = \frac{p(\theta)p(\widetilde{Y}|\theta)}{p(\widetilde{Y})} \quad (8)$$

Where $p(\theta)$ and $p(\theta|\widetilde{Y})$ defines prior and posterior distribution of parameters, respectively. Further, $p(\widetilde{Y}|\theta) \simeq L(\theta|\widetilde{Y})$ denotes the likelihood given as,

$$L(\theta|\widetilde{Y}) = \frac{n}{2}\ln\frac{\sum_{i=1}^{n}[\widetilde{y}_i - y_i(\theta)]^2}{n}. \quad (9)$$

To solve Eq. (9) analytically and numerically, a Markov Chain Monte Carlo (MCMC) simulation technique will be adopted to sample from the posterior distribution. For more details, readers are referred to (Ali et al., 2018b). Further literature on the Markov Chain Monte Carlo (MCMC) algorithms can be found elsewhere (*e.g.*, (Andrieu and Thoms, 2008; Duan et al., 1993; Gelman and Rubin, 1992; Gilks et al., 1994; Haario et al., 1999, 2001; Roberts and Rosenthal, 2009; Roberts and Sahu, 1997; Storn and Price, 1995, 1997; Ter Braak, 2006; ter Braak and Vrugt, 2008)).

3. Materials and method

In this Section, the description of acquired climate and cotton yield data, study regions, design of predictive models and performance criteria have been provided.

3.1. Climate and cotton yield data

The climate data for three regions (Multan, Nawabshah and Faisalabad) in Pakistan that includes rainfall (R), mean monthly temperature (T) and mean monthly relative humidity (H) were obtained from the Pakistan Meteorological Department, Pakistan for the year 1981 to 2013 (PMD, 2016). The meteorological data has recently been used in earlier studies (*e.g.*, (Ali et al., 2018a, b)). These three locations are the major producers of cotton yield in Pakistan (Districts, 2008; Service, 2012, 2014). The missing values of monthly rainfall were substituted by average of the respective time-averaged value from the climatological period because the rainfall for those months is not available. The cotton yield data was sourced from the Federal Bureau of Statistics (Economic wing), Islamabad, Pakistan and Agriculture Marketing Information Service, Directorate of Agriculture (Economics & Marketing) Punjab, Lahore Pakistan (Districts, 2008; Service, 2012, 2014). The source area and production data of cotton estimates were supplied by the provincial Crop Reporting Services and compiled by the Economic Wing of the devolved Ministry of Food and Agriculture and later by the Federal Bureau of Statistics. Cotton yield in the year 2009 was not available in the acquired dataset. To overcome this situation, the average value of all the cotton yield data from 1981 to 2013 was substituted for the missing 2009 data.

3.2. Study region

The selected regions for this study are: Multan, Faisalabad districts in Punjab and Nawabshah district of Sindh that can be seen on the map in Fig. 2.

Multan (30.1984 °N, 71.4687 °E) is located in southern part of Punjab province. The climate is arid with hot summers that bear some of the most extreme temperatures in the country as well as cold winters. The average temperature of Multan in summer is 42.3 °C while the record breaking highest temperature was 50.0 °C in May 2010 (Department, 2010; PMD, 2016). The average annual rainfall is 186.8 mm. Cotton and Mango are the major economic crops in Multan.

Faisalabad (31.4504 °N, 73.1350 °E) is situated in the rolling flat plains of northeast Punjab. The climate of Faisalabad is classified as dessert (Servey, 2016). Faisalabad is also a major producer of cotton. The average annual rainfall is approximately 375 mm (14.8 in. and the average temperature in Faisalabad is 24.2 °C (Servey, 2016).

Nawabshah (26.2442 °N, 68.4100 °E) is a city of Sindh province, located 50 km from the left bank of the River Indus. The climate is the hottest of all study sites with summer temperatures reaching as high as 53 °C A record breaking severe heatwave hit Nawabshah in 2010 (Department, 2010). The average annual rainfall in Nawabshah is 114.1 mm, making it also the driest of the selected study sites. Nawabshah is national hub of cotton production.

Table 1 describes the geographic characteristics as well as agricultural crop yield and hydrological statistics of the designated regions that have been utilized during the development of the forecasting models presented in this research. The acquired cotton yield data (Districts, 2008; Service, 2012, 2014) for the selected study regions were in hectors (ha) and tones. First, the yield data were transformed into kilograms (kg) from tones and then divided by hectors to get the data in standardized units of kilogram per hector ($kgha^{-1}$). The cotton yield data of Multan and Faisalabad have negative skewness while the temperatures of Multan and Nawabshah have negative skewness. The data were log-transformed between 0 and 1 to avoid the differences in skewness. The logarithmic transformation for the data is invertible, and hence will not affect the fitting results (Cong and Brady, 2012; Kim and Ahn, 2009). Following the approach of (Cong and Brady, 2012; Kim and Ahn, 2009; Mojtaba Sadegh and AghaKou, 2017), the data were not divided into training and testing.

3.3. Development of the proposed GP-MCMC based copula models

The proposed GP-MCMC based copula model (Ali et al., 2018b) was developed under the MATLAB environment on a Pentium 4 2.93 GHz dual core Central Processing Unit. In the proposed hybrid Genetic Programming integrated with Markov Chain Monte Carlo (GP-MCMC) based Copula model, the climate data (temperature, rainfall, humidity) is incorporated as the input predictor in the GP model where these input predictors are analyzed to assign a suitable fitness function. Further, functional and terminal sets are constructed to parametrize the genetic operation for predictor variables (Sreekanth and Datta, 2011). An evolutionary process is established between the inputs and cotton yield during optimization. A random initial population is created to obtain the best fitness based on reproduction, mutation, and crossover to evaluate and compare the fitness to calculate the forecasted cotton yield. The GP model is optimized by the emulation of an evolutionary process to get the optimum predicted cotton yield.

The GP based forecasting yield is then used as predictor in MCMC based copula model for final prediction. After selecting the desired copula, the next step is to adopt either local optimization or MCMC (which used global optimization). Here only the MCMC algorithm is used which starts from multiple runs in a single execution. This approach is capable of finding a global optimum approximation, estimates the posterior distribution of parameters and searches multiple regions of attraction. The MCMC simulation is employed within a Bayesian framework to estimate copula parameters. Further, the MCMC algorithm guarantees to find an estimate of the global optimum and characterizes the underlying uncertainty. Further, the MCMC approach is used to explore a wide range of copulas and evaluate them relative to the fitting uncertainties. Fig. 2 shows the flow chart of the proposed hybrid GP-MCMC based copula models.

Fig. 3(a–c) shows the modelling process in terms of marginal distribution based on Kendall's tau of average climate data (temperature, rainfall, humidity), and GP based simulated C_{pred} and observed C_{obs} for each station whereas Fig. 4(a–c) represents the closeness between the empirical cumulative probability distribution (ECP) of the data. The ECP also supports the marginal distributions in Fig. 3(a–c).

Further, Fig. 5(a–c) establishes the joint distribution of average climate data (temperature, rainfall, humidity), GP based simulated C_{pred} and observed C_{obs} for each station. The data were log-transformed

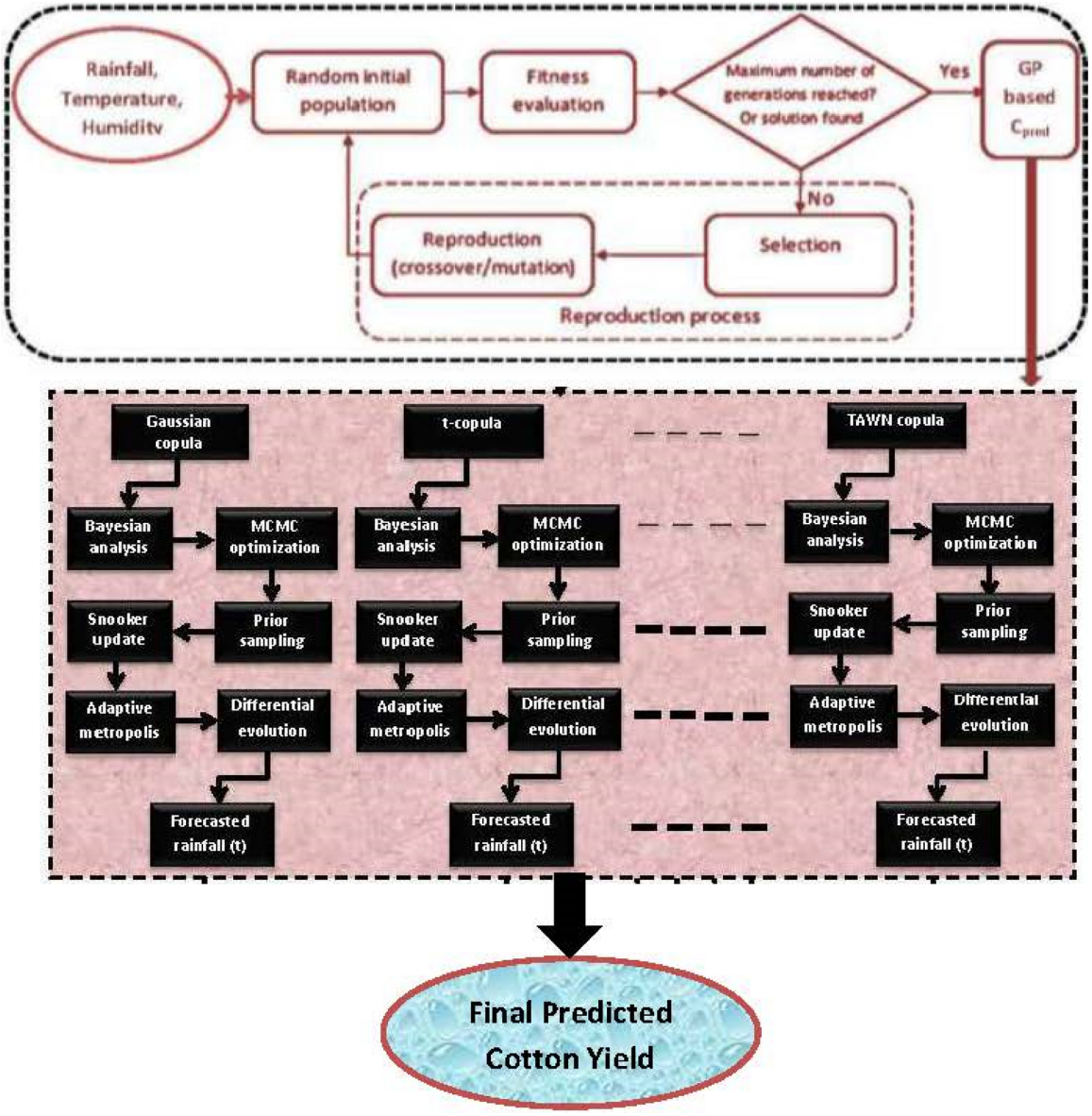


Fig. 2. A flow chart of the hybrid genetic programming algorithm integrated with a Markov Chain Monte Carlo based copula model.

Table 1
Descriptive statistics of the study sites' geographic, crop and hydrologic characteristics.

Station	Geographic characteristics			Agriculture crop statistics: Cotton yield (kg/ha)					
	Longitude	Latitude	Elevation (m)	Mean	Std.	Min	Max	Skewness	Kurtosis
Multan	71.47°	30.19°	129.00	813.09	213.32	271.00	1184.00	-0.67	0.03
Nawabshah	68.41°	26.24°	35.08	840.45	377.62	375.00	2209.00	1.72	4.30
Faisalabad	73.08°	31.42°	184.00	2008.45	903.99	475.00	3872.00	-0.13	-0.49

Hydrological statistics (1981–2013)

	Temperature (°C)				Rainfall (mm)				Humidity (%)			
	Mean	Std.	Min	Max	Mean	Std.	Min	Max	Mean	Std.	Min	Max
Multan	25.58	8.00	11.00	49.4	21.27	30.16	0.10	217.3	40.41	11.21	15	69.00
Nawabshah	31.22	3.35	18.62	33.36	28.18	28.99	0.06	124.86	43.22	4.99	35.00	53.20
Faisalabad	33.43	1.39	29.60	33.06	56.94	24.92	17.82	128.02	45.21	3.62	37.80	52.60

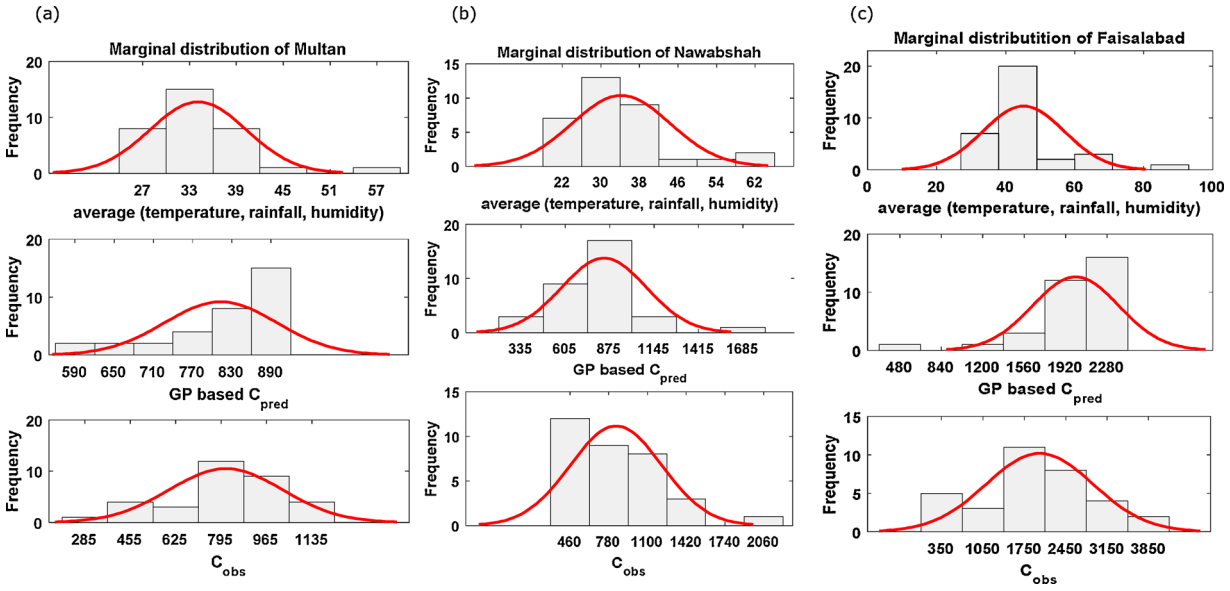


Fig. 3. Marginal distribution of average climate (temperature, rainfall, humidity), GP based predicted cotton yield (Cpred) and observed cotton yield (Cobs) with Kendall's tau (red) for (a) Multan, (c) Nawabshah and (c) Faisalabad station. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

between 0 and 1 to avoid the differences in skewness. The logarithmic transformation for the data is invertible, and hence will not affect the fitting results (Cong and Brady, 2012; Kim and Ahn, 2009).

3.4. Model performance criteria

To evaluate the performance of GP-MCMC based copula, MCMC based copula and GP models; we used different types of assessment tools for model evaluation that comprises statistical and standardized metrics. The mathematical formulations of all these assessment metrics are given as follows:

I The Likelihood value (Max_L) (Thyer et al., 2009) is calculated as :

$$\max_L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\tilde{\sigma}^2 - \frac{1}{2}\tilde{\sigma}^{-2}\sum_{i=1}^{n}[C_{obs,i} - C_{pred,i}]^2. \quad (10)$$

II The Akaike Information Criterion (AIC) (Akaike, 1974) is given by:

$$AIC = 2D + n.\ln - \left(\frac{\sum_{i=1}^{n}[C_{obs,i} - C_{pred,i}]^2}{n}\right) - 2CS. \quad (11)$$

III The Bayesian Information Criterion (BIC) (Schwarz, 1978) is given by:

$$BIC = D.\ln + n.\ln - \left(\frac{\sum_{i=1}^{n}[C_{obs,i} - C_{pred,i}]^2}{n}\right) - 2CS. \quad (12)$$

IV Confidence of Interval (CI) (Gardner and Altman, 1986) can be calculated as:

$$CI = \overline{C_{obs}} \pm z^*\frac{\sigma}{\sqrt{N}}. \quad (13)$$

V Correlation coefficient (r) (Dawson et al., 2007) is expressed as:

$$r = \left(\frac{\sum_{i=1}^{N}(C_{obs,i} - \overline{C}_{obs,i})(C_{pred,i} - \overline{C}_{pred,i})}{\sqrt{\sum_{i=1}^{N}(C_{obs,i} - \overline{C}_{obs,i})^2}\sqrt{\sum_{i=1}^{N}(C_{pred,i} - \overline{C}_{pred,i})^2}}\right). \quad (14)$$

VI Willmott's Index (WI) Willmott (1981) is expressed as:

$$WI = 1 - \left[\frac{\sum_{i=1}^{N}(C_{pred,i} - C_{pred,i})^2}{\sum_{i=1}^{N}(|C_{pred,i} - \overline{C}_{pred,i}| + |C_{pred,i} - \overline{C}_{pred,i}|)^2}\right],\ 0 \le d \le 1. \quad (15)$$

VII Nash-Sutcliffe coefficient (NS_E) Nash and Sutcliffe (1970) is expressed as:

$$NS_E = 1 - \left[\frac{\sum_{i=1}^{N}(C_{obs,i} - C_{pred,i})^2}{\sum_{i=1}^{N}(C_{pred,i} - \overline{C}_{pred,i})^2}\right]. \quad (16)$$

VIII Root mean square error (RMSE) (Dawson et al., 2007) is expressed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_{pred,i} - C_{obs,i})^2}. \quad (17)$$

IX Mean absolute error (MAE) (Dawson et al., 2007) is expressed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|(C_{pred,i} - C_{obs,i})|. \quad (18)$$

X Legates-McCabe's Index (LM) (Legates and McCabe, 1999) is expressed as:

$$LM = 1 - \left[\frac{\sum_{i=1}^{N}|C_{pred,i} - C_{obs,i}|}{\sum_{i=1}^{N}|C_{obs,i} - \overline{C}_{obs,i}|}\right]. \quad (19)$$

XI Relative root mean square error (RRMSE,%) (Legates and McCabe, 1999) is expressed as:

$$RRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(C_{pred,i} - C_{obs,i})^2}}{\frac{1}{N}\sum_{i=1}^{N}(C_{obs,i})} \times 100. \quad (20)$$

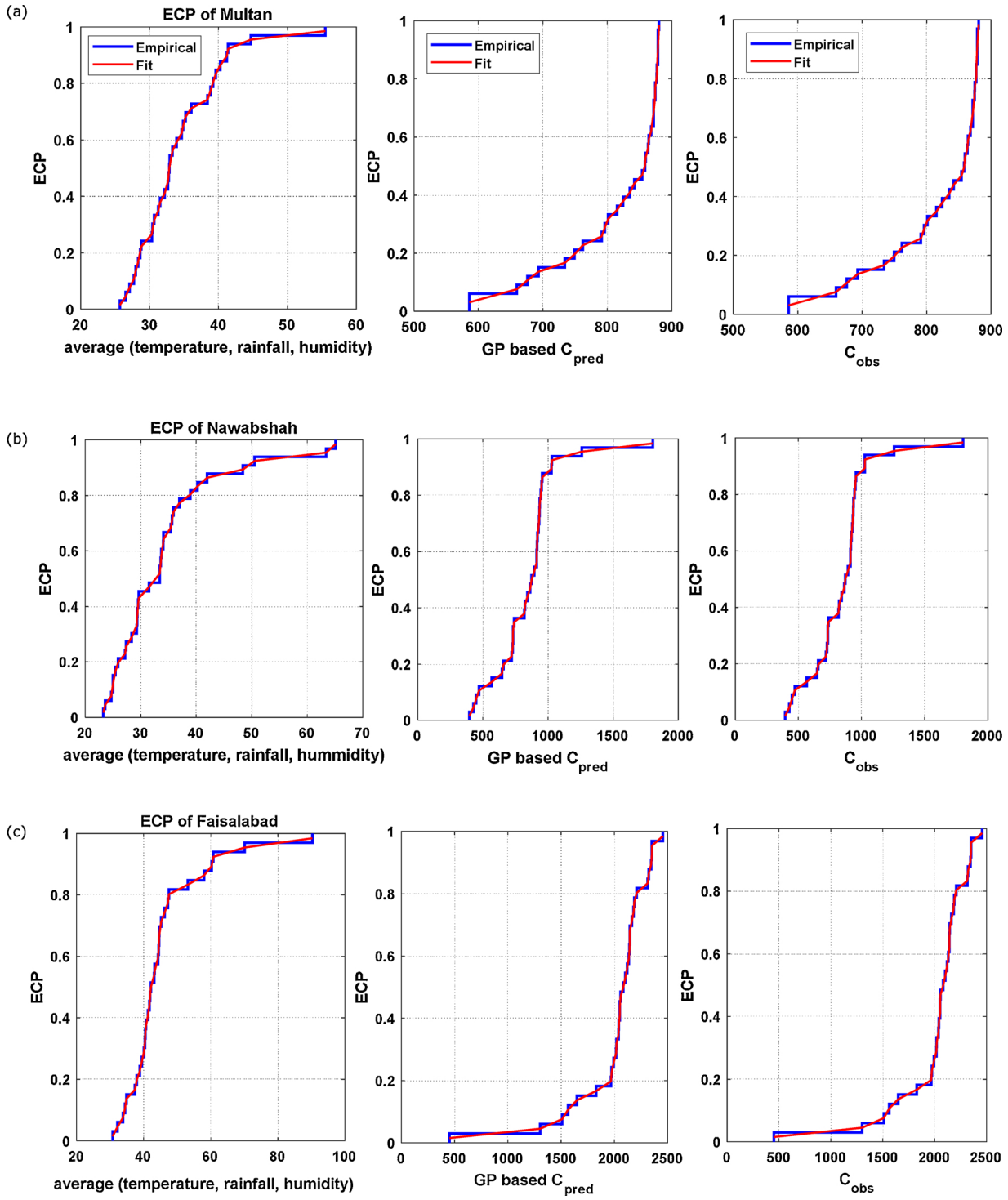


Fig. 4. Empirical cumulative probability distribution (ECP) of average (temperature, rainfall, humidity), GP based predicted cotton yield (Cpred) and observed cotton yield (Cobs) with Kendall' tau (fit) for (a) Multan station, (b) Nawabshah station and (c) Faisalabad station.

XII Relative mean absolute percentage error (RMAE; %) (Legates and McCabe, 1999) is expressed as

$$RMAE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{(C_{pred,i} - C_{obs,i})}{C_{obs,i}}\right| \times 100. \quad (21)$$

where C_{obs} and C_{pred} are the observed and simulated ith value of cotton yield C_Y, $\overline{C}_{obs}$ and $\overline{C}_{pred}$ are the observed and simulated mean of C_Y, z* represents the appropriate z*-value from the standard normal distribution and N is the number of tested data points.

The maximum likelihood (Max_L) minimizes the residuals between model simulations and observations. AIC, in contrast to the ad hoc likelihood value, takes into account both complexity of the model and minimization of error residuals and provides a more robust measure of quality of model predictions. A lower AIC value associates with a better model fit. Similar to AIC, a lower BIC value is associated with a better model fit. The correlation coefficient (r) lies in [0, 1], and demonstrates the proportion of variance in observed yield explained by the data

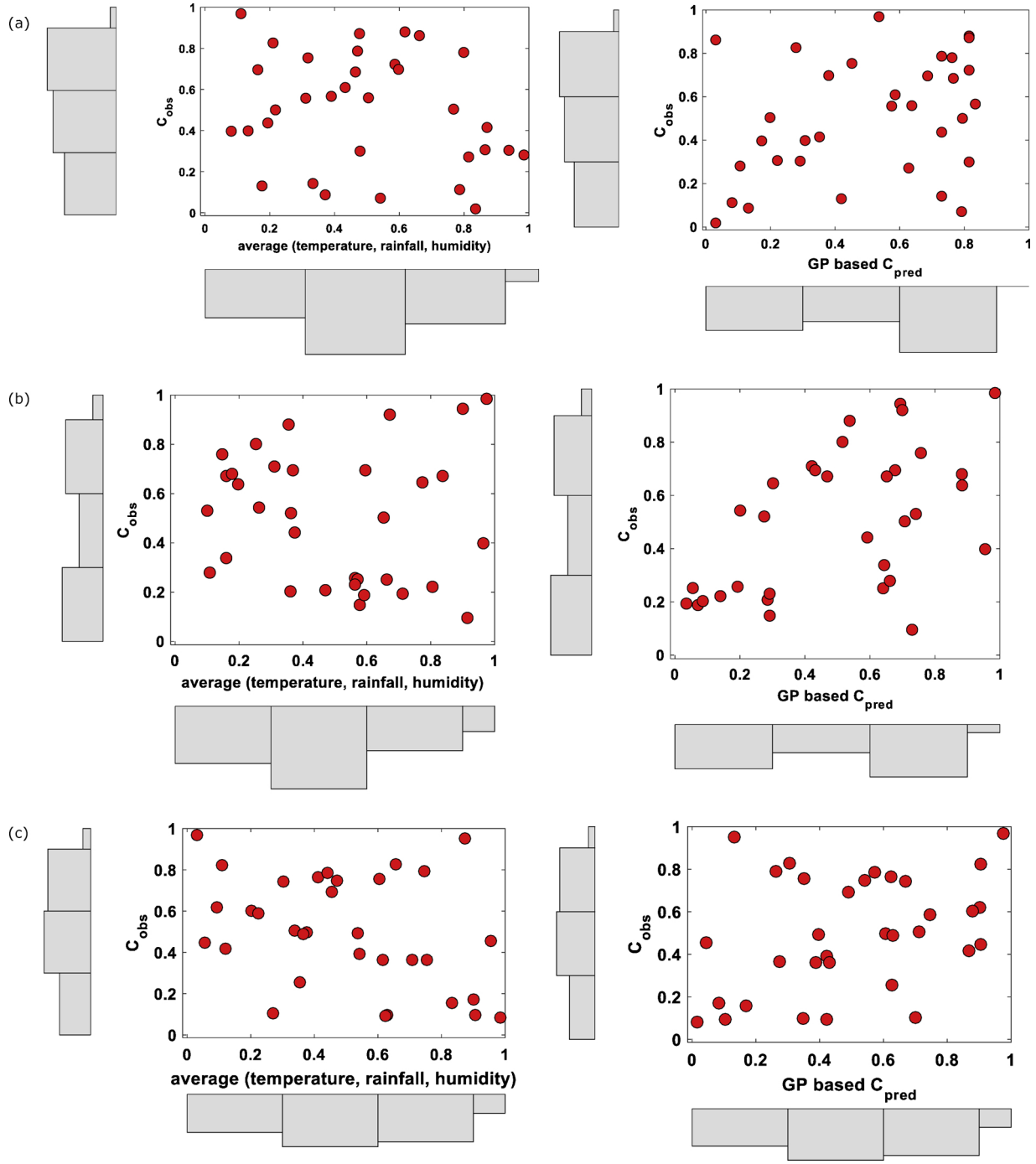


Fig. 5. Joint and marginal distribution of uniformly distributed average (temperature, rainfall, humidity), GP based predicted cotton yield (Cpred) and observed cotton yield (Cobs) for (a) Multan station, (b) Nawabshah station and (c) Faisalabad station.

intelligent model (Dawson et al., 2007). Further, these metrics (AIC, BIC, Max_L and CI) considers the number of the model parameters which is crucial important modelling (parsimony rules). Due to standardization of the observed and predicted means and variance, the robustness of r is limited. The goodness-of-fit relevant to high values are measured by RMSE while in contrast; MAE evaluates all deviations from observed data both in the same manner regardless of sign. The performance can reach to partial peaks and higher magnitudes that can exhibit larger error (Dawson et al., 2007). Willmott Index (WI) was introduced to counter this issue by considering the ratio of mean squared error instead of the differences (Willmott, 1981, 1982; Willmott, 1984; Willmott et al., 2012). Nash-Sutcliffe efficiency (NS_E) is another normalized statistical assessment metric used to determine the relative magnitude of residual variance of predicted data in comparison to the measured variance of observed data (Nash and Sutcliffe, 1970). Legates-McCabe's (LM) is a more advanced and powerful statistical assessment metric than both WI and NS_E which utilizes the adjustment of comparison in the evaluation of WI and EV. LM was found to be the best in evaluating the results by ignoring r and use WI and EV as baseline-adjusted indices together with an evaluation of RMSE and MAE (Legates and McCabe, 1999).

Due to the geographic differences among the regions of this study, the relative mean squared error (RRMSE) and relative mean absolute error (RMAE) are also calculated (Ali et al., 2018a, b; Dawson et al.,

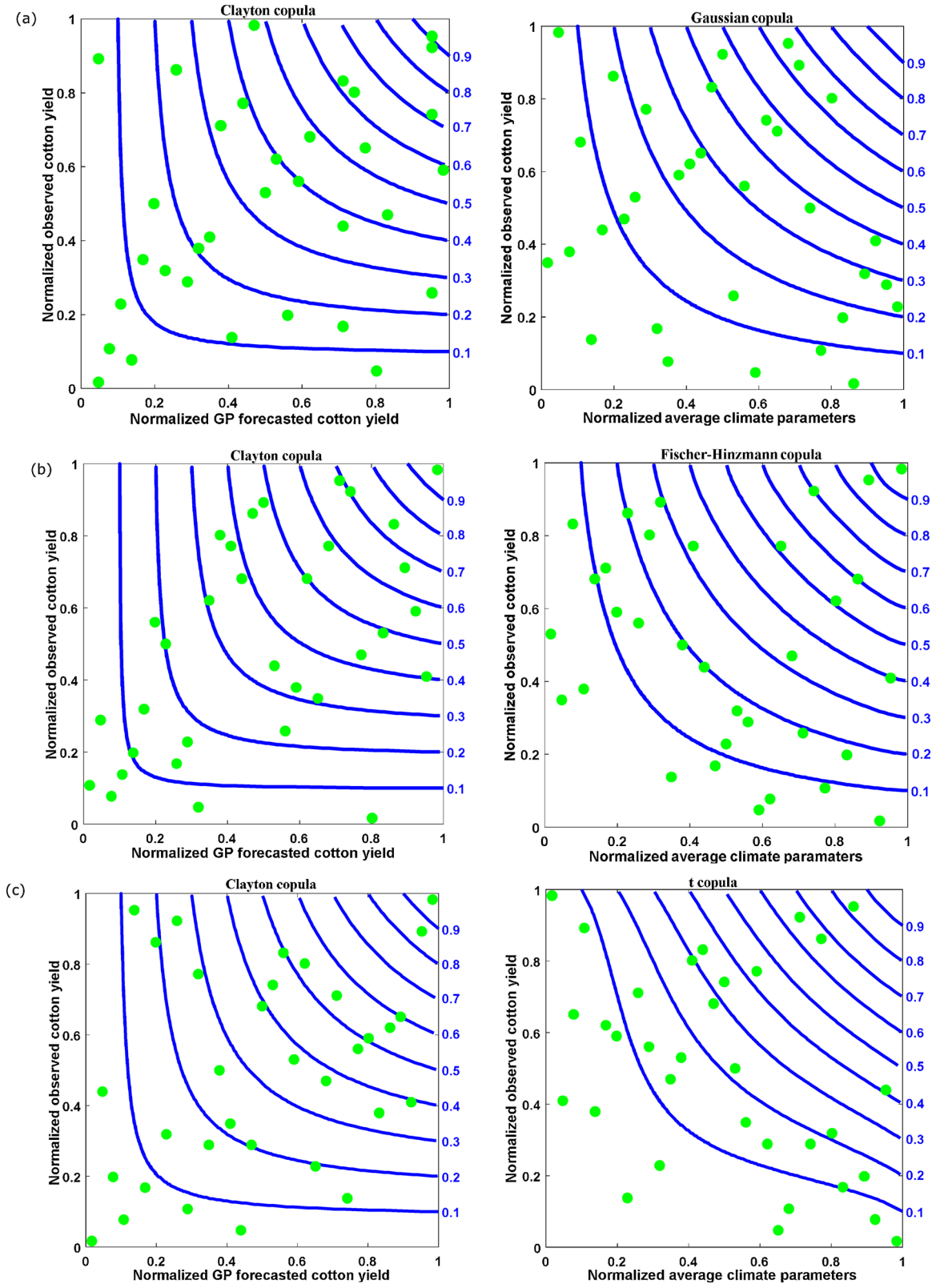


Fig. 6. Dependence structure of the GP-MCMC-copula based prediction cotton yield, MCMC based copula with average temperature, rainfall, humidity climate parameters versus the observed cotton yield for (a) Multan, (b) Nawabshah and (c) Faisalabad.
Note: Both the GP based predicted cotton yield, climate parameters (x-axis) and the cotton yield (y-axis) are presented in the probability space. Blue lines present the copula isolines and green circles show the observed normalized cotton yield.

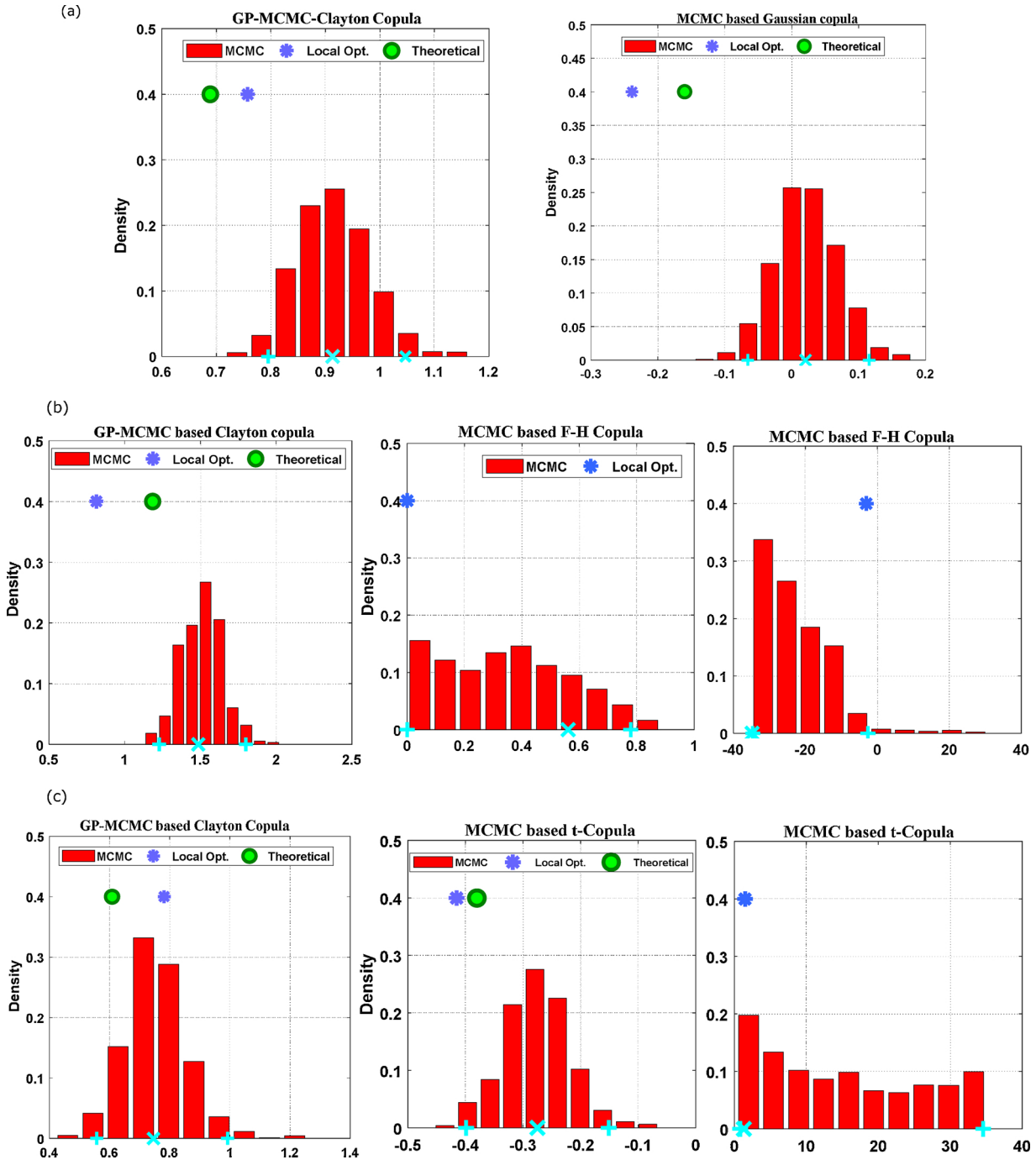


Fig. 7. Posterior distribution of the GP copulas derived by MCMC simulation within a Bayesian framework for the stations (a) Multan, (b) Nawabshah and (c) Faisalabad. Blue asterisks show the copula parameter derived by the local optimization while the green circles show the parameters derived by theoretical optimization. The red bins are the MCMC-derived parameters and the aqua (×) crosses show the maximum likelihood parameter of the MCMC. (For interpretation of the references to colour in the text, the reader is referred to the web version of this article.)

2007; Mohammadi et al., 2015) to evaluate and compare the model performance over geographically diverse study sites. If the (RRMSE, RMAE) < 10%, the performance of the developed model is considered outstanding while good if 10% < (RRMSE, RMAE) < 20%, fair if 20% < (RRMSE, RMAE) < 30% and poor if the (RRMSE, RMAE) > 30% (Ertekin and Yaldiz, 2000).

4. Results and discussion

The results of the GP-MCMC based copula model have been compared against MCMC based copula models and a standalone GP model based on the evaluation criterion described above (Eqs. (10)–(21)).

Fig. 6(a–c) demonstrates the joint dependence structure between GP based forecasted cotton yield and observed cotton yield anomalies using MCMC-copula models for the 33-year seasonal dataset. The asymmetric and skewed dependence structure of the observed data with the probability isolines derived through GP-MCMC based Clayton copula models in all stations can be seen to be more skewed than the other developed models. This pattern, however, has not been followed by the MCMC based Gaussian copula (in case of Multan), the Fischer-Hinzmann copula (in case of Nawabshah) or the t-copula (in case of Faisalabad). The uncertainty ranges of the GP-MCMC based Clayton copulas

Table 2
Local and Markov-Chain Monte Carlo (MCMC) copula parameters with 95% confidence interval (CI) for genetic programming (*i.e.*, GP-MCMC) based copula model and Markov-Chain Monte Carlo (MCMC) based copula model for predicting cotton yield. The parameters (para) with respective the CI is boldfaced (blue) and all models inputs are: T = temperature, R = rainfall, H = humidity.

Multan

Model		Copula	Local para 1	Local para 2	MCMC para 1	MCMC para 2	95% CI Local-MCMC para 1	95% CI Local-MCMC para 2
GP-MCMC Model	**M_1**	Gaussian	0.345		0.476		[0.41 0.53]	
	M_2	student t	0.405	2.017	0.469	2.478	[0.41 0.53]	[1.61 34.34]
	M_3	**Clayton**	**0.757**		**0.913**		**[0.79 1.04]**	
	M_4	Frank	2.374		2.922		[2.46 3.45]	
	M_5	Gumble	1.275		1.450		[1.35 1.58]	
	M_6	Fischer-Hinzmann	0.239	3.074	0.241	3.033	[0.13 0.33]	[1.78 5.39]
MCMC Model	**M_1**	Gaussian	-0.23		0.021		[-0.06 0.11]	
	M_2	student t	-0.25	13019593.4	0.021	33.291	[-0.08 0.11]	[3.44 34.25]
	M_3	Clayton	0		0.052		[0.00 0.19]	
	M_4	Frank	1.25		0.922		[0.27 1.69]	
	M_5	Gumble	1		1.003		[1.00 1.07]	
	M_6	Fischer-Hinzmann	0	-15.60	0.014	1.738	[0.00 0.41]	[-34.40 2.20]

Nawabshah

GP-MCMC	M_1	Gaussian	0.505		0.625		[0.54 0.68]	
	M_2	student t	0.528	7.648	0.624	33.383	[0.56 0.69]	[4.01 34.42]
	M_3	**Clayton**	**0.813**		**1.486**		**[1.22 1.79]**	
	M_4	Frank	3.408		4.406		[3.72 5.41]	
	M_5	Gumble	1.482		1.766		[1.57 2.08]	
	M_6	Fischer-Hinzmann	0.513	1.857	0.517	1.803	[0.14 0.66]	[0.31 10.93]
MCMC Model	M_1	Gaussian	-0.041		-0.153		[-0.29 -0.03]	
	M_2	student t	-0.143	3.458	-0.152	16.473	[-0.26 -0.03]	[2.56 34.09]
	M_3	Clayton	0		0		[0.00 0.08]	
	M_4	Frank	-0.841		-0.958		[-1.65 -0.28]	
	M_5	Gumble	1.084		1.000		[1.00 1.06]	
	M_6	Fischer-Hinzmann	0	-3.071	0.562	-34.796	[0.00 0.77]	[-34.61 -2.62]

Faisalabad

GP-MCMC Model	**M_1**	Gaussian	0.423		0.403		[0.31 0.47]	
	M_2	student t	0.358	1.559	0.398	1.993	[0.31 0.48]	[1.14 33.98]
	M_3	**Clayton**	**0.782**		**0.746**		**[0.55 0.99]**	
	M_4	Frank	2.365		2.374		[1.77 3.01]	
	M_5	Gumble	1.382		1.338		[1.23 1.48]	
	M_6	Fischer-Hinzmann	0.235	2.567	0.233	2.595	[0.01 0.34]	[1.07 12.56]
MCMC Model	**M_1**	Gaussian	-0.432		-0.282		[-0.39 -0.17]	
	M_2	student t	-0.415	1.504	-0.275	1.348	[-0.39 -0.15]	[0.84 34.57]
	M_3	Clayton	0		0		[0.00 0.09]	
	M_4	Frank	-2.557		-1.624		[-2.28 -0.94]	
	M_5	Gumble	1		1		[1.00 1.06]	
	M_6	Fischer-Hinzmann	0	12.814	0	3.347	[0.00 0.59]	[-34.47 4.81]

are tightly constrained for GP based on simulated and observed cotton yield (all stations) as compared to the uncertainty for wider ranges of MCMC based Gaussian, Fischer-Hinzmann and t-copula models. This means that the GP based predicted cotton yield has a stronger relationship (close or lies on to the fitted probability lines) which shows the GP-MCMC copula models are more accurate as compared to the MCMC based copula models. Comparing contours in Fig. 6, it is clear that the return period of the GP based yield is more close to the observed cotton yield.

Fig. 7(a–c) plots the posterior parameter distributions (red bins) of Clayton, Gaussian, Fischer-Hinzmann and student-t copula between GP based forecasted and actual cotton yield as well as average climate

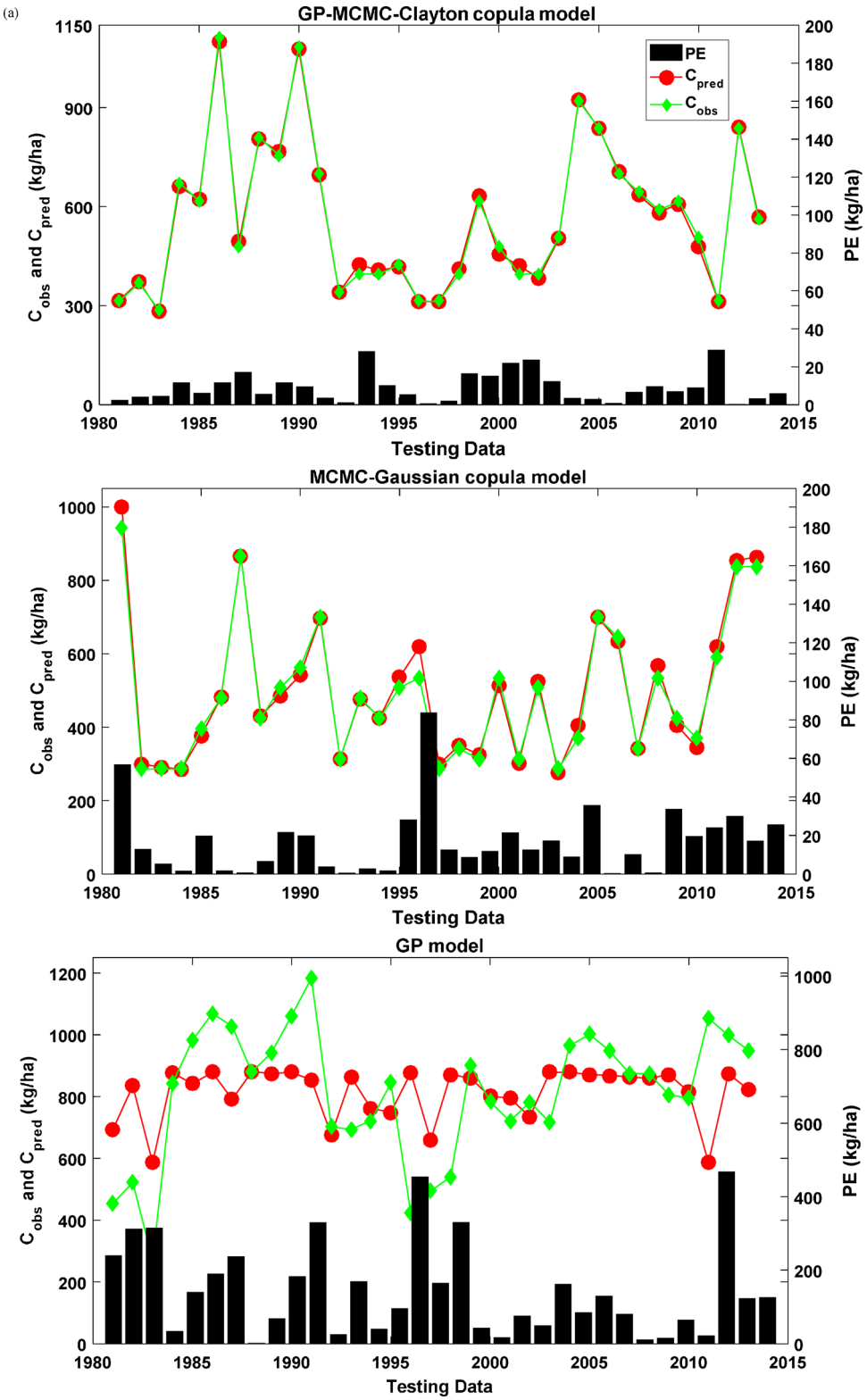


Fig. 8. Predicted (C_{pred}) and observed (C_{obs}) cotton yield generated by GP-MCMC copulas against MCMC copulas and GP models in the seasons of the testing period for the stations (a) Multan, (b) Nawabshah and (c) Faisalabad. Note that the bars shows the absolute prediction errors, PE = $|C_{obs} - C_{pred}|$.

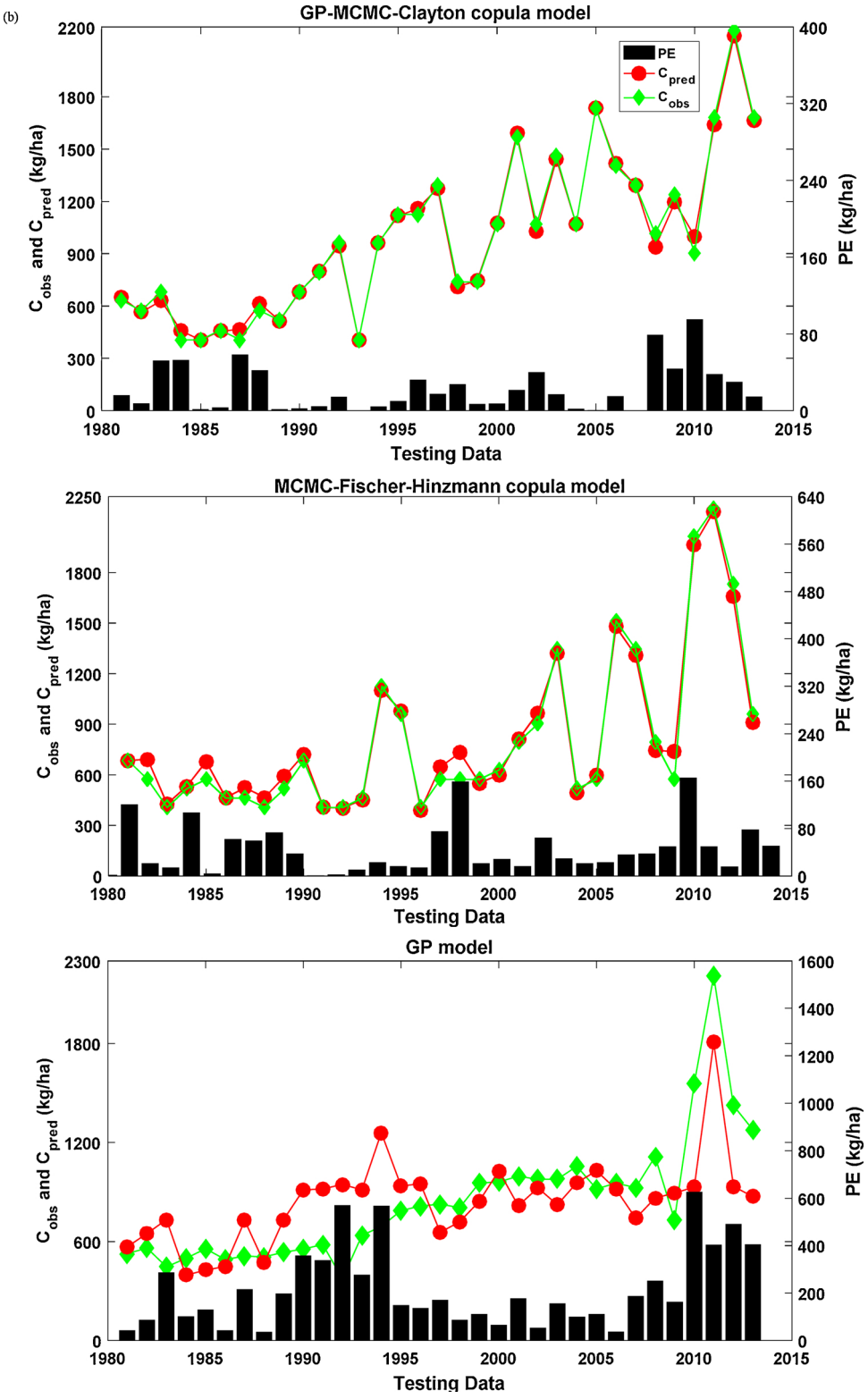


Fig. 8. (*continued*)

parameters against actual cotton yield. The copula parameters derived by MCMC algorithm using the global (local) optimization technique (blue asterisks) coincide with maximum likelihood parameters (aqua cross) of MCMC-derived posterior distributions (red bins). The inferred parameter of Gaussian copula (for Multan), Clayton copula (Nawabshah) from the global (local) optimization algorithm diverge significantly from their counterparts from the MCMC simulation. The parameters of theoretical optimization (green circles) with the maximum likelihood value of the MCMC algorithm are also coincided. The MCMC-derived posterior distributions (red bins) for the Clayton copula between GP based forecasted and observed cotton yield are nicely constrained for all stations, and their modes coincide with the parameter value inferred by the local optimization (blue asterisks) as compared to their counterpart copulas.

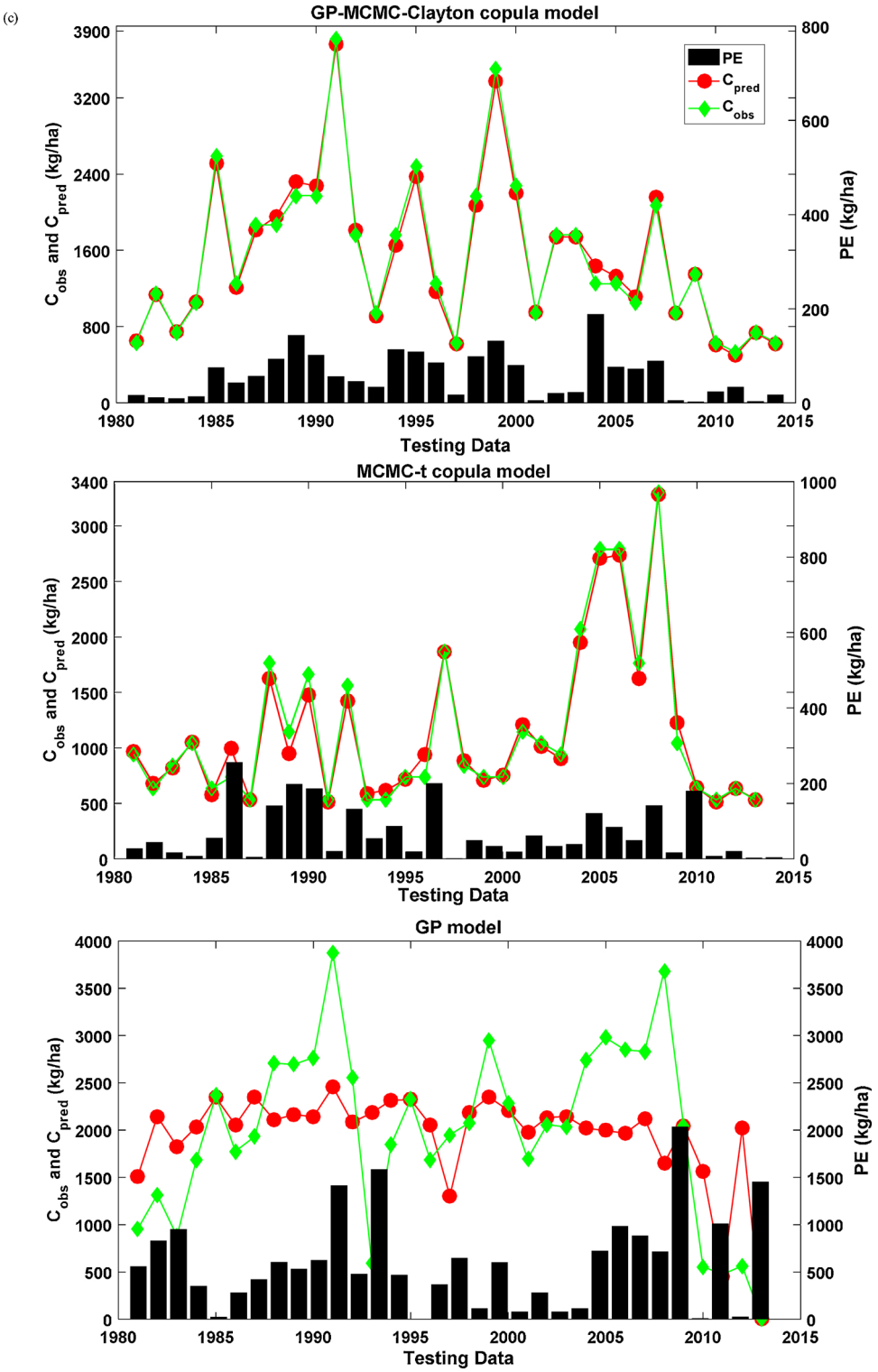


Fig. 8. (*continued*)

Table 2(a–c) shows the values of Local and MCMC parameters with 95% confidence interval (CI) of GP-MCMC and MCMC based (Clayton, Gaussian, t, Gumble, Frank and Fischer-Hinzmann) copulas to predict cotton yield for the study regions. The results of GP-MCMC based copulas are better than MCMC based copula models. The Clayton copula (Model M_3) appears to be best on the basis of the GP-MCMC optimization technique for all stations as compared to MCMC based copulas. In Table 2(a), the values of Local and MCMC parameters with 95% confidence of interval for Multan in terms of best copulas (on the basis of LM) are Local(0.757(GP-MCMC-Clayton), -0.23(MCMC-Gaussian)), MCMC(0.913 (GP-MCMC-Clayton), 0.021(MCMC-Gaussian)) and CI ([0.79 1.04] (GP-MCMC-Clayton), [−0.06 0.11](MCMC-Gaussian)). For Nawabshah Table 2(b), these values are GP-MCMC-Clayton copula (Local ≈ 0.813, MCMC ≈ 1.486, CI = [1.22 1.79]), MCMC-t-copula (Local ≈ 0, -3.071, MCMC ≈ -0.562, -34.796, CI = [0.00 0.77], [-34.61 -2.62]) and for Faisalabad Table 2(c), GP-MCMC-Clayton copula (Local ≈ 0.782, MCMC ≈ 0.746, CI = [0.55 0.99]), MCMC-t-copula (Local ≈ -0.415, 1.504, MCMC ≈ –0.275, 1.348, CI = [-0.39 -0.15], [0.84

Table 3
Testing performance of the genetic programming-based Markov Chain Copula Model (GP-MCMC) *vs*. the MCMC based copula and a standalone GP model in terms of the Root Mean Squared Error (*RMSE*, *kg/ha*), Mean Absolute Error (*MAE, kg/ha*), Correlation Coefficient (*r*), Nash-Sutcliff (NS_E), Willmott Index (*WI*), including Akiake information criterion (*AIC*), Bayesian information criterion (*BIC*) and Maximum-likelihood (Max_L) for the prediction of cotton yield. The optimal model is boldfaced (blue).

(a) Multan									
GP-MCMC									
Models	**Copula**	***RMSE (kg/ha)***	***MAE (kg/ha)***	***r***	NS_E	***WI***	***AIC***	***BIC***	Max_L
M_1	Gaussian	16.513	12.923	0.998	0.994	0.996	-262.877	-261.380	132.438
M_2	student t	15.527	11.958	0.998	0.995	0.997	-263.912	-260.919	133.956
M_3	Clayton	11.763	8.910	0.999	0.997	0.998	-284.194	-282.697	143.097
M_4	Frank	16.693	13.212	0.998	0.994	0.996	-261.169	-259.673	131.585
M_5	Gumble	20.190	15.333	0.997	0.992	0.995	-248.609	-247.112	125.304
M_6	Fischer-Hinzmann	14.822	11.296	0.998	0.995	0.997	-267.945	-264.952	135.973
MCMC									
M_1	Gaussian	21.713	14.807	0.993	0.984	0.990	-245.128	-243.631	123.564
M_2	student t	21.282	15.153	0.993	0.985	0.990	-243.115	-240.122	123.557
M_3	Clayton	21.092	14.893	0.993	0.985	0.990	-245.707	-244.211	123.854
M_4	Frank	21.246	15.024	0.993	0.985	0.990	-228.084	-226.588	115.042
M_5	Gumble	21.336	15.325	0.993	0.985	0.990	-244.932	-243.436	123.466
M_6	Fischer-Hinzmann	30.954	21.638	0.992	0.967	0.979	-243.014	-240.021	123.507
				GP					
M_1	GP	192.197	145.568	0.404	0.163	0.371			
(b) Nawabshah									
GP-MCMC									
M_1	Gaussian	38.464	30.050	0.997	0.993	0.996	-253.051	-251.554	127.525
M_2	student t	38.495	30.105	0.997	0.993	0.996	-251.010	-248.017	127.505
M_3	Clayton	33.468	23.559	0.997	0.994	0.997	-262.244	-260.748	132.122
M_4	Frank	40.707	31.573	0.996	0.992	0.995	-249.346	-247.850	125.673
M_5	Gumble	46.188	36.736	0.996	0.989	0.994	-240.992	-239.496	121.496
M_6	Fischer-Hinzmann	46.469	36.358	0.995	0.989	0.994	-238.588	-235.595	121.294
MCMC									
M_1	Gaussian	56.176	45.014	0.995	0.986	0.986	-227.072	-225.575	114.536
M_2	student t	56.152	45.348	0.995	0.986	0.986	-225.097	-222.104	114.548
M_3	Clayton	62.592	47.478	0.993	0.982	0.981	-220.922	-219.426	111.461
M_4	Frank	54.770	43.978	0.995	0.986	0.986	-228.734	-227.237	115.367
M_5	Gumble	62.598	47.484	0.993	0.982	0.981	-220.924	-219.428	111.462
M_6	Fischer-Hinzmann	60.329	43.563	0.993	0.983	0.982	-220.373	-217.380	112.187
GP									
M_1	GP	271.105	215.464	0.684	0.468	0.501			
(c) Faisalabad									
GP-MCMC									
M_1	Gaussian	85.214	74.271	0.995	0.989	0.993	-242.002	-240.506	122.001
M_2	student t	80.882	65.284	0.995	0.990	0.993	-242.674	-239.681	123.337
M_3	Clayton	72.785	55.591	0.996	0.992	0.995	-250.641	-249.144	126.320
M_4	Frank	86.423	73.128	0.994	0.988	0.993	-240.294	-238.797	121.147
M_5	Gumble	92.812	76.710	0.994	0.987	0.992	-235.613	-234.117	118.807
M_6	Fischer-Hinzmann	82.493	66.715	0.995	0.989	0.993	-241.351	-238.358	122.676
MCMC									
M_1	Gaussian	107.180	78.697	0.991	0.978	0.979	-228.301	-226.805	115.151
M_2	student t	98.548	69.865	0.991	0.981	0.982	-229.631	-226.638	116.815
M_3	Clayton	136.578	96.874	0.988	0.964	0.959	-209.114	-207.617	105.557
M_4	Frank	102.782	77.585	0.991	0.980	0.980	-228.859	-227.363	115.430
M_5	Gumble	138.627	99.800	0.988	0.963	0.958	-209.105	-207.608	105.552
M_6	Fischer-Hinzmann	138.620	99.800	0.988	0.963	0.958	-207.123	-204.130	105.561
GP									
M_1	GP	809.737	627.899	0.415	0.173	0.378			

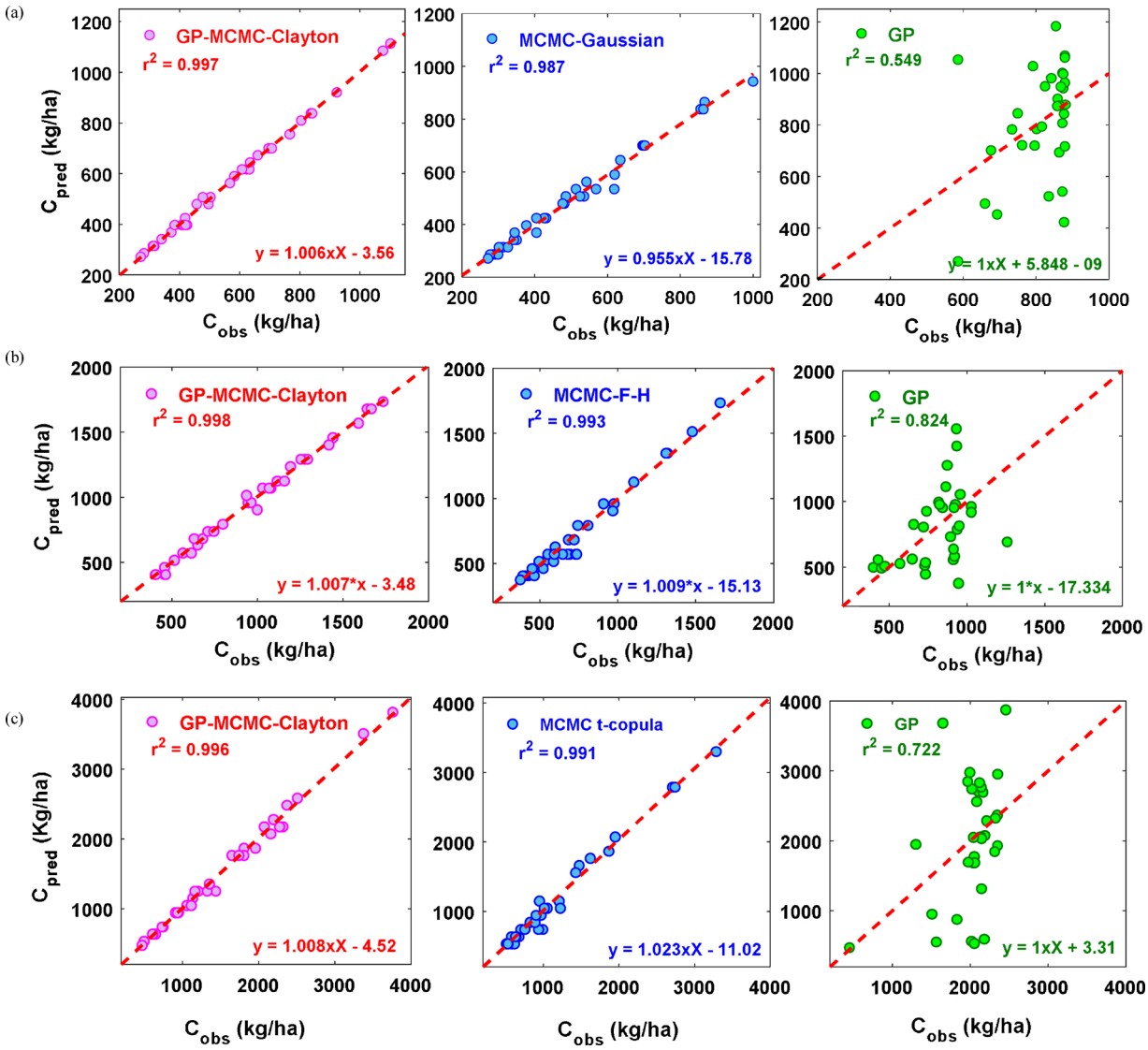


Fig. 9. Scatterplot of predicted (C_{pred}) and observed (C_{obs}) cotton yield using the GP-MCMC-copula, MCMC-copula and standalone GP models, with the coefficient of determination (r^2) inserted in each panel of study zones. (a) Multan, (b) Nawabshah and (c) Faisalabad.

34.57]). Thus, the hybrid GP-MCMC based Clayton copula performed well in all study stations compared to the MCMC based copula and standalone GP models.

Fig. 8(a–c) plots a comparison of the predicted and observed cotton yield with prediction error (PE) as bar graphs, generated by the GP-MCMC (Clayton) copula models together with the MCMC based (Gaussian, Fischer-Hinzmann, t) copulas and standalone GP model. It is to be noted that the best copula models only plotted for Multan, Nawabshah and Faisalabad. There is compelling evidence that the GP-MCMC-Clayton copula models perform very accurately for all stations in comparison with the MCMC copulas (Gaussian, Fischer-Hinzmann, t) and standalone GP model. On a season-by-season basis analysis, the GP-MCMC-Clayton copula models are seen to exhibit a reasonable accuracy by predicting the yield that is almost the same as the observed cotton yield from 1981 to 2013 followed by MCMC based copula models. This can also be confirmed by the smaller bar graphs that revealed the smaller the bar graph, the more accurate the model. The standalone GP model on the other hand, predicted a large amount of prediction errors in all stations from 1981 to 2013 while exhibit a significant difference between the predicted and the observed cotton yield.

In Table 3(a–c), the preciseness of the proposed GP-MCMC based copula model is evaluated in relation to MCMC based copula model and standalone GP model in terms of RMSE, MAE, r, WI, EV, AIC, BIC and Max_L. The GP-MCMC-Clayton copula applied to Multan station attained the highest values of RMSE ≈ 11.763 kg/ha, MAE ≈ 8.910 kg/ha, r ≈ 0.999, NS_E ≈ 0.997, WI ≈ 0.998, AIC ≈ -284.194, BIC ≈ -282.697, Max_L ≈ 143.097, as compared to the MCMC-Gaussian copula where these metrics are RMSE ≈ 21.713 kg/ha, MAE ≈ 14.807 kg/ha, r ≈ 0.993, NS_E ≈ 0.984, WI ≈ 0.990, AIC ≈ -245.128, BIC ≈ -243.631, Max_L ≈ 123.564, followed by standalone GP model where RMSE ≈ 192.197 kg/ha, MAE ≈ 145.568 kg/ha, r ≈ 0.404, NS_E ≈ 0.163, WI ≈ 0.371. For Nawabshah, the best copula appeared to be Clayton on the basis of the GP-MCMC algorithm that obtained RMSE ≈ 33.468 kg/ha, MAE ≈ 23.559 kg/ha, r ≈ 0.997, NS_E ≈ 0.994, WI ≈ 0.997, AIC ≈ -262.244, BIC ≈ -260.748, Max_L ≈ 132.122, against MCMC based Fischer-Hinzmann copula RMSE ≈ 60.329 kg/ha, MAE ≈ 43.563 kg/ha, r ≈ 0.993, NS_E ≈ 0.983, WI ≈ 0.982, AIC ≈ -220.373, BIC ≈ -217.380, Max_L ≈ 112.187, and GP model RMSE ≈ 271.105 kg/ha, MAE ≈ 215.464 kg/ha, r ≈ 0.684, NS_E ≈ 0.468, WI ≈ 0.501. Finally in Faisalabad station, the GP-MCMC based Clayton copula again shows high performance accuracy as RMSE ≈ 72.785 kg/ha, MAE ≈ 55.591 kg/ha, r ≈ 0.996, NS_E ≈ 0.992, WI ≈ 0.995, AIC ≈ 250.641, BIC ≈ -249.144, Max_L ≈ 126.320, and are reasonably better than the MCMC based t-copula model having RMSE ≈ 98.548 kg/ha, MAE ≈ 69.865 kg/ha, r ≈ 0.991,

Table 4
Evaluation of the GP-MCMC *vs.* the MCMC and a standalone GP model using the relative root mean squared error (*RRMSE*), relative mean absolute error (*RMAE*) and Legates & McCabe's Index (*LM*). The optimal model is boldfaced (blue).

Mode	Copula	*LM*	*RRMSE*	*RMAE (%)*
(a) Multan				
GP-MCMC				
M_1	Gaussian	0.929	2.912	2.659
M_2	student t	0.935	2.781	2.352
M_3	Clayton	0.952	2.107	1.771
M_4	Frank	0.929	2.989	2.789
M_5	Gumble	0.917	3.616	3.151
M_6	Fischer-Hinzmann	0.938	2.614	2.281
MCMC				
M_1	Gaussian	0.895	4.541	3.214
M_2	student t	0.892	4.451	3.257
M_3	Clayton	0.894	4.411	3.220
M_4	Frank	0.893	4.443	3.228
M_5	Gumble	0.891	4.462	3.276
M_6	Fischer-Hinzmann	0.846	6.474	4.231
GP				
M_1	GP	0.132	23.638	22.652
(b) Nawabshah				
GP-MCMC				
M_1	Gaussian	0.918	3.789	3.587
M_2	student t	0.918	3.792	3.588
M_3	Clayton	0.936	3.297	2.873
M_4	Frank	0.914	4.010	3.841
M_5	Gumble	0.900	4.550	4.329
M_6	Fischer-Hinzmann	0.901	4.578	4.214
MCMC				
M_1	Gaussian	0.875	7.072	5.988
M_2	student t	0.874	7.069	6.075
M_3	Clayton	0.870	7.756	6.862
M_4	Frank	0.878	6.895	5.778
M_5	Gumble	0.870	7.756	6.863
M_6	Fischer-Hinzmann	0.879	7.595	6.692
GP				
M_1	GP	0.215	32.257	29.204
(c) Faisalabad				
GP-MCMC				
M_1	Gaussian	0.887	5.513	5.788
M_2	student t	0.901	5.233	4.769
M_3	Clayton	0.917	4.807	3.548
M_4	Frank	0.889	5.591	5.767
M_5	Gumble	0.883	6.005	6.297
M_6	Fischer-Hinzmann	0.899	5.337	4.657
MCMC				
M_1	Gaussian	0.859	9.312	6.906
M_2	student t	0.875	8.562	6.947
M_3	Clayton	0.825	12.075	11.218
M_4	Frank	0.861	8.930	7.227
M_5	Gumble	0.821	12.045	11.557
M_6	Fischer-Hinzmann	0.821	12.044	11.557
GP				
M_1	GP	0.113	40.316	51.765

$NS_E \approx 0.981$, $WI \approx 0.982$, $AIC \approx$ -229.631, $BIC \approx$ -226.638, $Max_L \approx 116.815$ and standalone GP as $RMSE \approx 809.737$ kg/ha, $MAE \approx 627.899$ kg/ha, $r \approx 0.415$, $NS_E \approx 0.173$, $WI \approx$ 0.378.

Further, to analyze the model performance more closely, Fig. 9(a–c) displays a scatterplot showing the goodness-of-fit and its correlation coefficient r is shown to depict the extent of agreement between predicted and observed cotton yield. The GP-MCMC based Clayton copula models for all stations convincingly outperform the MCMC based Gaussian (Multan), Fischer-Hinzmann (Nawabshah), t (Faisalabad) copulas and standalone GP model in all seasons from 1981 to 2013. Therefore, it is clear that the GP-MCMC based Clayton copula model has a better ability to simulate the cotton yield with good accuracy, as confirmed by the larger r-value.

Table 4(a–c) demonstrates the comparison of the GP-MCMC based copula models vs. the MCMC based copula models and standalone GP model using relative root mean squared error (RRMSE), relative mean absolute error (RMAE) and Legates & McCabe's Index (LM) for the stations Multan, Nawabshah, and Faisalabad. The best GP-MCMC based Clayton copula for Multan attained the values of $LM \approx 0.952$, $RRMSE \approx 2.107\%$ and $RMAE \approx 1.771\%$ followed by MCMC based Gaussian copula $LM \approx 0.895$, $RRMSE \approx 4.541\%$, $RMAE \approx 3.214\%$ and GP model $LM \approx 0.132$, $RRMSE \approx 23.638\%$ and $RMAE \approx 22.652\%$ (See, Table 4(a)). Similarly for Nawabshah, the GP-MCMC based Clayton copula appeared to be the best giving values $LM \approx 0.936$, $RRMSE \approx 3.297\%$ and $RMAE \approx 2.873\%$ as compared to the MCMC based Fischer-Hinzmann copula $LM \approx 0.879$, $RRMSE \approx 7.595\%$ and $RMAE \approx 6.692\%$, GP model $LM \approx 0.215$, $RRMSE \approx 32.257\%$ and $RMAE \approx 29.204\%$ (See, Table 4(a)). In Faisalabad, the best GP-MCMC-Clayton copula obtained $LM \approx 0.917$, $RRMSE \approx 4.807\%$ and $RMAE \approx 3.548\%$ benchmarked with MCMC-t-copula $LM \approx 0.875$, $RRMSE \approx 8.562\%$ and $RMAE \approx 6.947\%$ and the standalone GP model $LM \approx 0.113$, $RRMSE \approx 40.316\%$ and $RMAE \approx 51.765\%$. Overall, the hybrid GP-MCMC based Clayton copula was better than the MCMC based copula (Gaussian, t, Clayton, Gumble, Frank, Fischer-Hinzmann) and GP models.

In Fig. 10(a–c), we illustrate a boxplot of the GP-MCMC based copula vs. MCMC based copula and standalone GP model's prediction error for the seasonal cotton yield of all the study sites. The outliers specified by + in every boxplot represent the extreme magnitudes of the prediction error within the testing seasons along with their upper quartile, median and lower quartile values. The distributed prediction error is justified by these boxplots showing a much lesser spread was achieved by GP-MCMC-copula models compared with MCMC-copula models giving a relatively smaller magnitude of quartile statistics and median values followed by GP models. Accordingly, the GP-MCMC based Clayton copula models remain as highly optimized superior models followed by MCMC based (Gaussian, Fischer-Hinzmann and t) copula models and standalone GP models giving good results for all stations applied to predict seasonal cotton yield in terms of the illustrated clustered error distribution tending towards smaller magnitude. The red circle of each boxplot shows the mean prediction error while the dashed lines (brown) connected the mean value of each boxplot.

Table 5 shows a geographical comparison of the GP-MCMC based best copula model, MCMC based best copula model and standalone GP predictive performance using relative root mean squared error (RRMSE), relative mean absolute error (RMAE) and Legates & McCabe's Index (LM) for the different locations (Multan, Nawabshah, and Faisalabad). In terms of site-averaged performance, the GP-MCMC based copula model was found to yield the highest Legates-McCabe's agreement and lowest relative percentage errors (RRMSE, RMAE). Multan appears to be the most accurate station in predicted cotton yield ($LM \approx 0.952$, $RRMSE \approx 2.107\%$, $RMAE \approx 1.771\%$) followed by Nawabshah ($LM \approx 0.936$, $RRMSE \approx 3.297\%$, $RMAE \approx 2.873\%$) and Faisalabad ($LM \approx 0.917$, $RRMSE \approx 4.807\%$, $RMAE \approx 3.548\%$) respectively. But overall, the prediction produced by all three models was low in terms of relative error being within the recommended 10% threshold for an

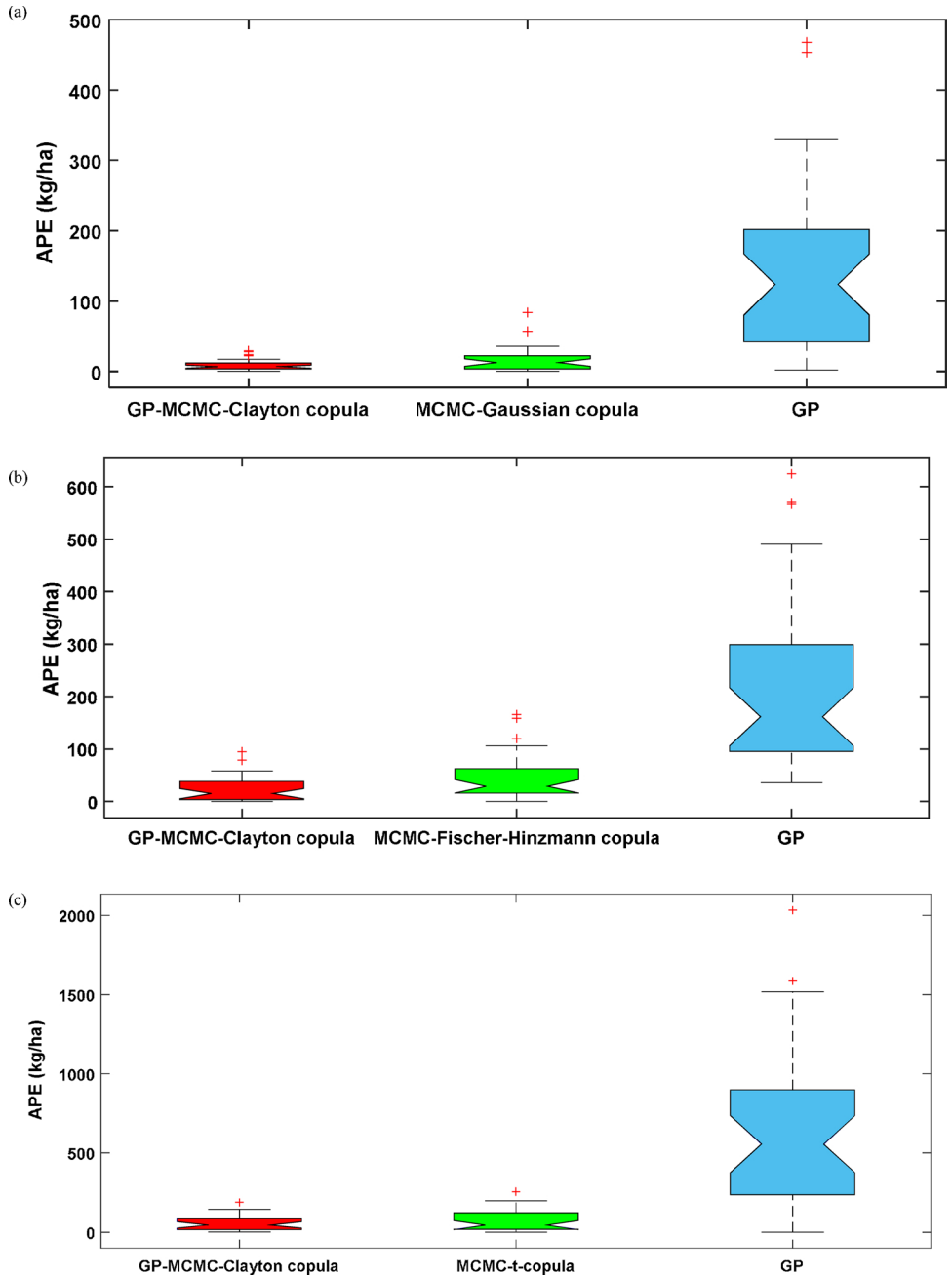


Fig. 10. Box-plots of absolute prediction error (APE, kg/ha) of GP-MCMC based copula models *vs.* MCMC based copula models and the standalone GP model in predicted cotton yield for the stations for (a) Multan (b) Nawabshah, and (c) Faisalabad.

excellent model Ertekin and Yaldiz (2000).

To address the accurate estimation for an increased demand of cotton in Pakistan, there is a desire to develop strong predictive models for cotton crop production. This paper is dedicated to modelling cotton yield within the agriculture regions in Pakistan using climate and cotton yield data from 1981 to 2013. Moreover, the proposed GP-MCMC-copula models were significantly accurate in utilizing climate and cotton yield data with there being less than 3.0% and 2.0% uncertainty based on RRMSE and RMAE respectively with a reasonably large statistical correlation of Legates-McCabe's ($LM \approx 0.952$) for Multan station and similarly for the other two stations (Table 5). Due to the aforementioned qualities of the GP-MCMC copula models, it is possible to apply the proposed model for the prediction of other crop yields in Pakistan as no such studies have been conducted in the past. Where such data is available, the developed techniques could also be applied and tested to alternative micro-scale farming communities globally.

Despite the mentioned merits of the GP-MCMC based Clayton copula models, there are some shortcomings which can be addressed in follow-up work. The cotton yield data of neighbouring stations may be used to predict the cotton yield in a central station which shares a common boundary to the predictor stations. Further, the impact of fertilizer may also be coupled with meteorological data to explore in later iterations of the developed copula models because this will also likely have a strong effect on the production of crop yields. Drought is a factor and a likely important predictor variable with meteorological parameters that will influence the prediction of accurate cotton yield. Irrigation statistics (water supply) can also be utilized in this regard to improve the estimation of crop yield in follow-up work. Solar radiation has a great effect on crop production which may also be focused on in

Table 5
Geographical comparison showing the performance skills of the GP-MCMC *vs*. the MCMC based best copula models and a standalone GP model using the relative root mean squared error (*RRMSE*), relative mean absolute error (*RMAE*) and the Legates & McCabe's Index (*LM*). Note that the best model is boldfaced (blue).

Model	*LM*	*RRMSE (%)*	*RMAE (%)*
Multan			
GP-MCMC Clayton copula	0.952	2.107	1.771
MCMC Gaussian copula	0.895	4.541	3.214
GP	0.132	23.638	22.652
Nawabshah			
GP-MCMC Clayton copula	0.936	3.297	2.873
MCMC Fischer-Hinzmann copula	0.879	7.595	6.692
GP	0.215	32.257	29.204
Faisalabad			
GP-MCMC Clayton copula	0.917	4.807	3.548
MCMC student t-copula	0.875	8.562	6.947
GP	0.113	40.316	51.765

follow up work. The use of more data (variables) however, means more scarce resources are needed for their collection/collation. The inclusion of additional data may ultimately impact upon the efficiency of the developed models and as such these should be tested individually to find the optimum number of inputs for each local environment.

In terms of model optimization for crop prediction, it is believed that the hybridization of different models can generate better estimation than standalone models. Therefore, the MCMC based copula models could be optimized with ensemble methods (Dietterich, 2002; Lei and Wan, 2012; Yun et al., 2008) to achieve more accurate results. The ANFIS algorithms developed by Jang et al. (1997) which is very powerful can be another optimization method to be considered in this regard. Moreover, the more advanced models such as the Ensemble method (Dietterich, 2002), Particle Swarm Optimization (Chen and Yu, 2005; Zhisheng, 2010), Genetic algorithms (Davis, 1991), chaos theory (Briggs and Peat, 1989) etc. coupled with copula (Nelsen, 2003) may generate good results. The regression analysis (Draper and Smith, 2014) can be coupled with copula in a way to optimize the prediction ability of possible hybrid copula model. A more generalized framework of hybridizing copula techniques with generalized mixed models (Draper and Smith, 2014) to develop MCMC-copula mixed models for crop yield prediction is planned in upcoming work. Autoregressive fractionally integrated moving average (Ling and Li, 1997) based copula (ARFIMA-copula) and least square support vector machine (Yuan et al., 2017) based copula (LSSVM-copula) models can be used to predict cotton yield. Support vector machine designed by Cortes and Vapnik (1995), Extreme learning machine studied by Huang et al. (2006) etc. may also be very good options for the prediction of cotton yield and other crops in Pakistan. As the standard statistical approaches avoid the hurdle of model uncertainty that leads to over-fitting and riskier decisions, Bayesian model averaging BMA techniques (Hoeting et al., 1998) have the ability to model uncertainty for accurate predictions. Therefore, BMA techniques provide yet another option to be used to model uncertainty in crop yield that is produced due to several factors such as missing climate data, extreme weather conditions and the likely influence of climate change.

Multi-resolution tools like frequency resolution can also be applied in this area to broaden the scope of this study. In this regard, wavelet transformation (maximum overlap discrete wavelet) (Holschneider, 1988; Khalighi et al., 2011), empirical mode decomposition (EMD) (Rilling et al., 2003), and singular value decomposition (SVD) (De Lathauwer et al., 1994) can be utilized for prediction purposes. This study advocates the possibility of using reanalysis and satellite climate data to predict cotton yield as well as to extend the scope to other crop yields and global locations.

The proposed GP-MCMC based copula models can be applied to other agricultural crop yield prediction scenarios that will assist agricultural policy makers in Pakistan in the optimal management of crop estimation. Moreover, accurate future wheat yield prediction can warn the government and impacted stakeholders prior to significant food security. Further, as processed based modelling requires extensive number of resource for implementation. For example, modelling the effects of systematic nutrient transfers requires additional attention and very expansive (Snow et al., 2014). Thus developing countries like Pakistan can't afford, so therefore, the proposed GP-MCMC based copula models could be used as a convenient option.

5. Conclusion

This paper has developed a suite of GP-MCMC based copula models using climate data (temperature, rainfall, humidity) as predictor variables and cotton yield data as an objective variable to predict cotton yield for different geographical sites in Pakistan. To attain an accurate GP-MCMC-copula model, the MCMC algorithm adopted a global optimization technique to find the best copula parameters. Evidently, the performance of the GP-MCMC based copula was found to be much better than the MCMC based copula and the standalone GP models, as evident by low relative forecasting errors and high performance metrics.

By assessing the performance of the GP-MCMC-Clayton copula in relation to the MCMC-Gaussian copula and the GP model for Multan using the most advanced normalized metrics of Legates-McCabe's Index, the GP-MCMC-Clayton copula was again found to have the highest agreement. The obtained LM agreement values between the predicted and observed cotton yield for Multan station were LM ≈ 0.952 (GP-MCMC-Clayton copula), 0.895 (MCMC-Gaussian copula) and 0.132 (GP) respectively whereas the relative percentage errors RRMSE and RMAE were only 2.107%, 1.771% (GP-MCMC-Clayton copula) compared with 4.541%, 3.214% (MCMC-Gaussian copula) and 23.638%, 22.652% (GP). The GP-MCMC based copula models also appeared to be the best in Nawabshah and Faisalabad stations.

In summary, to improve the prediction accuracy, the GP-MCMC based copula models can be optimized and tuned with other advanced techniques including ensemble methods. Generalized mixed models that develop MCMC-copula mixed models for crop yield prediction are planned in upcoming work. Autoregressive fractionally integrated moving average based copulas (ARFIMA-copula) and least square support vector machine based copula (LSSVM-copula) models have future potential. This study can be extended to other locations where cotton yield data is available to provide an accurate estimation of cotton yield on a small scale, affecting individual stations. Thus, we propose that suitably optimized GP-MCMC based copula models may be used to predict cotton yield in the future where the prediction of such a commodity will likely become even more important due to increasing demand and for economic growth in terms of exporting to international markets.

Acknowledgements

This research utilized cotton yield data acquired from the Pakistan Bureau of Statistics, Government of Pakistan: Islamabad, Pakistan and climate data were acquired from Pakistan Meteorological Department, Pakistan, that are duly acknowledged. This study was supported by the University of Southern Queensland's Office of Graduate Studies Postgraduate Research Scholarship (2017–2019). We thank all reviewers and the journal Editor for their useful comments that have improved the clarity of the final manuscript.

References

Ahmad, B., 1975. Supply response of cotton in Punjab: a time series analysis. Pak. Cottons 29, 19–32.

Ahmad, D., Chani, M., Humayon, A., 2017. Major crops forecasting area, production and yield evidence from agriculture sector of Pakistan. Sarhad J. Agric. 33 (3), 385–396.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19 (6), 716–723.

Ali, S., Badar, N., Fatima, H., 2015. Forecasting production and yield of sugarcane and cotton crops of Pakistan for 2013-2030. Sarhad J. Agric. 31 (1), 1–9.

Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2018a. An ensemble-ANFIS based uncertainty assessment model for forecasting multi-scalar standardized precipitation index. Atmos. Res. 207, 155–180.

Ali, M., Deo, R.C., Downs, N.J., Maraseni, T., 2018b. Multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm for rainfall forecasting. Atmos. Res. 213, 450–464.

Andrieu, C., Thoms, J., 2008. A tutorial on adaptive MCMC. Stat. Comput. 18 (4), 343–373.

Ayaz, M., Khalid, K., Malik, K.M., 2015. An Analysis of Weather and Cotton Crop Development in Lower Sindh (Tandojam), vol. 10. pp. 2007–2012 Retrieved on January (2015).

Banuri, T., 1998. Pakistan: Environmental Impact of Cotton Production and Trade. International Institute for Sustainable Development.

Bauer, M.E., 1975. The role of remote sensing in determining the distribution and yield of crops. Adv. Agron. 27, 271–304.

Blanc, E., Quirion, P., Strobl, E., 2008. The climatic determinants of cotton yields: evidence from a plot in West Africa. Agric. For. Meteorol. 148 (6-7), 1093–1100.

Bokusheva, R., Kogan, F., Vitkovskaya, I., Conradt, S., Batyrbayeva, M., 2016. Satellite-based vegetation health indices as a criteria for insuring against drought-related yield losses. Agric. For. Meteorol. 220, 200–206.

Briggs, J., Peat, F.D., 1989. Turbulent Mirror: an Illustrated Guide to Chaos Theory and the Science of Wholeness. HarperCollins Publishers.

Carpio, C.E., Ramirez, O.A., 2002. Forecasting foreign Cotton production: the case of India, Pakistan and Australia. Paper presented and published. The Proceedings of The 2002 Beltwide Cotton Conference.

Chen, G.-c., Yu, J.-s., 2005. Particle swarm optimization algorithm. Inf. Control Shenyang 34 (3), 318.

Chen, C., Pan, X., Zhang, L., Pang, Y., 2011. Impact of climate change on cotton production and water consumption in Shiyang River Basin. Trans. Chin. Soc. Agric. Eng. 27 (1), 57–65.

Clayton, D.G., 1978. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. Biometrika 65 (1), 141–151.

Cong, R.-G., Brady, M., 2012. The interdependence between rainfall and temperature: copula analyses. Sci. World J 2012.

Cortes, C., Vapnik, V., 1995. Support vector machine. Mach. Learn. 20 (3), 273–297.

Craparo, A., Van Asten, P., Läderach, P., Jassogne, L., Grab, S., 2015. Coffea arabica yields decline in Tanzania due to climate change: Global implications. Agric. For. Meteorol. 207, 1–10.

Davis, L., 1991. Handbook of Genetic Algorithms.

Dawson, C.W., Abrahart, R.J., See, L.M., 2007. HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts. Environ. Model. Softw. 22 (7), 1034–1052.

De Lathauwer, L., De Moor, B., Vandewalle, J., by Higher-Order, B.S.S., 1994. Singular Value Decomposition, Proc. EUSIPCO-94. Edinburgh, Scotland, UK. pp. 175–178.

Debnath, M., Bera, K., Mishra, P., 2013. Forecasting area, production and yield of cotton in India using ARIMA model. Res. Rev. J. Space Sci. Technol. 2 (1), 16–20.

Deo, R.C., Samui, P., 2017. Forecasting evaporative loss by least-square support-vector regression and evaluation with genetic programming, Gaussian process, and minimax probability machine regression: case study of Brisbane City. J. Hydrol. Eng., 05017003.

Department, P.M., 2010. Dry Weather Predicted in the Country During Friday/Monday.

Dietterich, T.G., 2002. Ensemble learning. The Handbook of Brain Theory and Neural Networks, vol. 2. pp. 110–125.

Districts, C.a.a.p.b, 2008. Crops Area and Production by Districts 1981-2008. Food and Cash crops, Federal Bureau of Statistics (Economic wing), Islamabad, Pakistan.

Draper, N.R., Smith, H., 2014. Applied Regression Analysis. John Wiley & Sons.

Duan, Q., Gupta, V.K., Sorooshian, S., 1993. Shuffled complex evolution approach for effective and efficient global minimization. J. Opt. Theory Appl. 76 (3), 501–521.

Ertekin, C., Yaldiz, O., 2000. Comparison of some existing models for estimating global solar radiation for Antalya (Turkey). Energy Convers. Manag. 41 (4), 311–330.

Fischer, M.J., Hinzmann, G., 2006. A New Class of Copulas With Tail Dependence and a Generalized Tail Dependence Estimator. Diskussionspapiere//Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Statistik und Ökonometrie.

Gardner, M.J., Altman, D.G., 1986. Confidence intervals rather than P values: estimation rather than hypothesis testing. Br. Med. J. Clin. Res. Ed. (Clin. Res. Ed.) 292 (6522), 746–750.

Gelman, A., Rubin, D.B., 1992. Inference from iterative simulation using multiple sequences. Stat. Sci. 457–472.

Gilks, W.R., Roberts, G.O., George, E.I., 1994. Adaptive direction sampling. Statistician 179–189.

Haario, H., Saksman, E., Tamminen, J., 1999. Adaptive proposal distribution for random walk Metropolis algorithm. Comput. Stat. 14 (3), 375–396.

Haario, H., Saksman, E., Tamminen, J., 2001. An adaptive metropolis algorithm. Bernoulli 7 (2), 223–242.

Hearn, A.B., 1994. OZCOT: a simulation model for cotton crop management. Agric. Syst. 44 (3), 42.

Hina Ali, H.A., Faridi, Zahir, Ali, Hira, 2013. Production and forecasting trends of cotton in Pakistan: an analytical view. J. Basic Appl. Sci. Res. 3 (12), 5.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1998. Bayesian model averaging. Proceedings of the AAAI Workshop on Integrating Multiple Learned Models. pp. 77–83.

Holschneider, M., 1988. On the wavelet transformation of fractal objects. J. Stat. Phys. 50 (5), 963–993.

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2006. Extreme learning machine: theory and applications. Neurocomputing 70 (1), 489–501.

Jang, J.-S.R., Sun, C.-T., Mizutani, E., 1997. Neuro-fuzzy and Soft Computing: a Computational Approach to Learning and Machine Intelligence.

Jin, X., Xu, X., 2012. Estimation of cotton yield based on net primary production model in xinjiang, China, agro-geoinformatics (agro-geoinformatics). 2012 First International Conference on. IEEE. pp. 1–4.

Kern, A., et al., 2018. Statistical modelling of crop yield in Central Europe using climate data and remote sensing vegetation indices. Agric. For. Meteorol. 260, 300–320.

Khalighi, S., Sousa, T., Oliveira, D., Pires, G., Nunes, U., 2011. Efficient feature selection for sleep staging based on maximal overlap discrete wavelet transform and SVM, engineering in medicine and biology society, EMBC. 2011 Annual International Conference of the IEEE. IEEE. pp. 3306–3309.

Kim, T.-W., Ahn, H., 2009. Spatial rainfall model using a pattern classifier for estimating missing daily rainfall data. Stoch. Environ. Res. Risk Assess. 23 (3), 367–376.

Koza, J.R., 1992. Genetic Programming: on the Programming of Computers by Means of Natural Selection. MIT press, pp. 1.

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35 (1), 233–241.

Lei, K.S., Wan, F., 2012. Applying ensemble learning techniques to ANFIS for air pollution index prediction in Macau. International Symposium on Neural Networks 509–516.

Li, C., Singh, V.P., Mishra, A.K., 2013. A bivariate mixed distribution with a heavy-tailed component and its application to single-site daily rainfall simulation. Water Resour. Res. 49 (2), 767–789.

Ling, S., Li, W., 1997. On fractionally integrated autoregressive moving-average time series models with conditional heteroscedasticity. J. Am. Stat. Assoc. 92 (439), 1184–1194.

McPhee, N.F., Poli, R., Langdon, W.B., 2008. Field Guide to Genetic Programming.

Mehr, A.D., Kahya, E., Olyaie, E., 2013. Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. J. Hydrol. (Amst) 505, 240–249.

Mohammadi, K., et al., 2015. A new hybrid support vector machine–wavelet transform approach for estimation of horizontal global solar radiation. Energy Convers. Manag. 92, 162–171.

Mojtaba Sadegh, E.R., AghaKou, Amir, 2017. Multivariate Copula Analysis Toolbox (MvCAT): Describing dependence and underlying uncertainty using a Bayesian framework. Water Resour. Res. 53 (6), 17.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—a discussion of principles. J. Hydrol. (Amst) 10 (3), 282–290.

Nelsen, R.B., 2003. Properties and applications of copulas: a brief survey. Dhaene, J., Kolev, N., Morettin, P.A. (Eds.), Proceedings of the First Brazilian Conference on Statistical Modeling in Insurance and Finance 10–28.

Nguyen-Huy, T., Deo, R.C., An-Vo, D.-A., Mushtaq, S., Khan, S., 2017. Copula-statistical precipitation forecasting model in Australia's agro-ecological zones. Agric. Water Manag. 191, 153–172.

Nguyen-Huy, T., Deo, R.C., Mushtaq, S., An-Vo, D.-A., Khan, S., 2018. Modeling the joint influence of multiple synoptic-scale, climate mode indices on Australian wheat yield using a vine copula-based approach. Eur. J. Agron. 98, 65–81.

Papageorgiou, E.I., Markinos, A.T., Gemtos, T.A., 2011. Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application. Appl. Softw. Comput. 11 (4), 3643–3657.

PMD, 2016. Pakistan Meteorological Department, Pakistan.Pakistan Meteorological Department, Pakistan.

Raza, A., Ahmad, M., 2015. Analysing the Impact of Climate Change on Cotton Productivity in Punjab and Sindh, Pakistan.Analysing the Impact of Climate Change on Cotton Productivity in Punjab and Sindh, Pakistan.

Reporter, T.Ns.S., 2015. Cotton Production Reaches 14.838 Million Bales, the Dawn.

Rilling, G., Flandrin, P., Goncalves, P., 2003. On empirical mode decomposition and its algorithms. IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing. IEEER 8–11.

Roberts, G.O., Rosenthal, J.S., 2009. Examples of adaptive MCMC. J. Comput. Graph. Stat. 18 (2), 349–367.

Roberts, G.O., Sahu, S.K., 1997. Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. J. R. Stat. Soc. Ser. B Stat. Methodol. 59 (2), 291–317.

Sarwar, U., 2014. Agriculture in Pakistan – an Overview.Agriculture in Pakistan – an Overview.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.

Servey, B., 2016. Asian Urban Information of Kobe.

Service, A.M.I., 2012. DISTRICT-WISE AREA OF WHEAT CROP. Directorate of Agriculture (Economics & Marketing) Punjab, Lahore.DISTRICT-WISE AREA OF WHEAT CROP. Directorate of Agriculture (Economics Marketing) Punjab, Lahore.

Service, A.M.I., 2014. DISTRICT-WISE AREA OF WHEAT CROP. Directorate of Agriculture (Economics & Marketing) Punjab, Lahore Pakistan.DISTRICT-WISE AREA OF WHEAT CROP. Directorate of Agriculture (Economics Marketing) Punjab, Lahore Pakistan.

Snow, V.O., et al., 2014. The challenges–and some solutions–to process-based modelling

of grazed agricultural systems. Environ. Model. Softw. 62, 420–436.
Sreekanth, J., Datta, B., 2011. Coupled simulation-optimization model for coastal aquifer management using genetic programming-based ensemble surrogate models and multiple-realization optimization. Water Resour. Res. 47 (4).
Storn, R., Price, K., 1995. Differential Evolution–a Simple and Efficient Adaptive Scheme for Global Optimization Over Continuous Spaces: Technical Report TR-95-012. International Computer Science, Berkeley, California.
Storn, R., Price, K., 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. J. Glob. Optim. 11 (4), 341–359.
ter Braak, C.J., Vrugt, J.A., 2008. Differential evolution Markov chain with snooker updater and fewer chains. Stat. Comput. 18 (4), 435–446.
Ter Braak, C.J., 2006. A Markov Chain Monte Carlo version of the genetic algorithm Differential Evolution: easy Bayesian computing for real parameter spaces. Stat. Comput. 16 (3), 239–249.
Thyer, M., et al., 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: a case study using Bayesian total error analysis. Water Resour. Res. 45 (12).
Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2 (2), 184–194.
Willmott, C.J., 1982. Some comments on the evaluation of model performance. Bull. Am. Meteorol. Soc. 63 (11), 1309–1313.
Willmott, C.J., 1984. On the Evaluation of Model Performance in Physical Geography, Spatial Statistics and Models. Springer, pp. 443–460.
Willmott, C.J., Robeson, S.M., Matsuura, K., 2012. A refined index of model performance. Int. J. Climatol. 32 (13), 2088–2094.
Yang, Y., Yang, Y., Han, S., Macadam, I., Li Liu, D., 2014. Prediction of cotton yield and water demand under climate change and future adaptation measures. Agric. Water Manag. 144, 42–53.
Yuan, X., Tan, Q., Lei, X., Yuan, Y., Wu, X., 2017. Wind power prediction using hybrid autoregressive fractionally integrated moving average and least square support vector machine. Energy 129, 122–137.
Yun, Z., et al., 2008. RBF neural network and ANFIS-based short-term load forecasting approach in real-time price environment. IEEE Trans. Power Syst. 23 (3), 853–858.
Zhisheng, Z., 2010. Quantum-behaved particle swarm optimization algorithm for economic load dispatch of power system. Expert Syst. Appl. 37 (2), 1800–1803.

Chapter 8

Two-phase ant colony optimization algorithm integrated with online sequential extreme learning machine to predict wheat yield

Foreword

This chapter is an exact copy of the submitted (under 2nd review) manuscript to the *Journal of IEEE Access*.

This chapter is based on the development of a universal wheat yield data intelligent model utilizing wheat yield data at district level. To develop a two-phase hybrid ACO-OSELM model using feature based input selection ant colony optimization (ACO) algorithm and OSELM model to predict wheat yield. The ACO algorithm is conditioned to search for the suitable, statistically relevant data sites for the model's training, and the corresponding testing sites by virtue of a feature selection strategy utilizing a total of 27 agricultural counties' datasets in the agro-ecological zones in Punjab province in Pakistan.

The ACO-OSELM model is compared against ACO-ELM and ACO-RF models, showing improved performance in response to the comparison models. The developed model can be explored as a decision-support tenet for crop yield estimation in regions where a statistically significant relationship with historical agricultural crop is well-established.

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Two-phase ant colony optimization algorithm integrated with online sequential extreme learning machine to predict wheat yield

Mumtaz Ali[1], Ravinesh C Deo*[2], Tek Maraseni[3], Nathan J. Downs[4]
School of Agricultural, Computational and Environmental Sciences
Centre for Applied Climate Sciences and Centre for Sustainable Agricultural Systems
Institute for Life Sciences and the Environment
University of Southern Queensland
Springfield, QLD 4300, AUSTRALIA

Corresponding author: Ravinesh C Deo[2] (e-mail: ravinesh.deo@usq.edu.au).

This research project has been sponsored by The University of Southern Queensland's Postgraduate Research Scholarship (2017–2019) awarded to the first author, managed by the Office of Research and Graduate Studies Division

ABSTRACT Reliable artificial intelligence models designed to predict wheat yield over relatively large and spatially spread agricultural fields can be adopted as important decision support tools to develop strategic farming management practices. In this paper, an optimally trained two-phase machine-learning model is designed to predict the wheat yield (Wpred), utilizing 27 agricultural districts data in agro-ecological zones in the Punjab province of Pakistan, a developing nation that relies on agricultural productivity for the survival of its citizens. The universally-trained model, denoted as the two-phase ACO-OSELM is designed to utilize the online sequential extreme learning machine (OSELM) model coupled with the ant colony optimization (ACO) algorithm incorporating statistically significant annual yield lagged at (t – 1) as the model's predictor to generate the future yield at 6 tested field sites. In the first phase, the ACO algorithm is conditioned to search for the suitable, statistically relevant data sites for the model's training, and the corresponding testing sites by virtue of a feature selection strategy. An annual wheat yield time series input data are constructed utilizing data from each selected training sites and applied against 6 test site cases to evaluate the hybrid ACO-OSELM model. The partial autocorrelation function is adopted to deduce statistically significantly lagged data, and OSELM is applied to generate wheat yield. The two-phase hybrid ACO-OSELM model is tested within the 6 agricultural sites of Punjab province, and the results are benchmarked with extreme learning machine (ELM) and random forest (RF) integrated with ACO to design hybrid ACO-ELM and hybrid ACO-RF models, respectively. Testing performance of hybridized models, according to robust metrics, was satisfactory; however, the two-phase hybrid ACO-OSELM model was proven to be a reliable wheat yield prediction tool with high performance in the present context, and the method can be replicated over some of the other regions globally where it may assist in better farming management applications.

INDEX TERMS Agricultural precision; wheat yield model; ant colony optimization; OSELM

1.0: INTRODUCTION

Culminating knowledge about best approaches to farming, with strategic crop management systems, and learning from the best practices in neighborhood cropping zones, are considered as a useful approach for agronomists to help formulate timely information on crop yield to foster benefits to agriculture-reliant nations [1, 2]. In Pakistan, wheat, which is the subject of this paper, is cultivated in winter season largely in the agricultural lands of the province of Punjab, the largest producing jurisdiction [3]. Wheat accounts for 2.6% of Pakistan's GDP and 12.5% to the GDP of the agronomy sector [4]. According to United Nations Food and Agriculture Organization, Pakistan was placed in the eighth position as a global wheat producer from 2007 to 2009 [5]. The modelling of wheat production with intelligent learning systems that also incorporate historical practices, and relevant knowledge from past

yields to predict the future yield can provide new strategic frameworks for improving current and future agricultural productions, and support future food security issues in both developing and first world nations.

Considering the vitality of wheat as a daily food grain supporting millions of human lives, accurately predicting the annual yield can assist governments and agricultural-climate policy experts in making decisions on their national imports and exports, maintaining sufficient reservations of wheat as a proxy for national food security and setting the relevant prices for agriculture markets. In 2005, the actual yield in Pakistan was relatively low compared to the predicted yield, and as such, poor estimations have moderated the market price and prompted the government to export the grains from the international market [6, 7].

In the past, Pakistan has faced significant crises of wheat supply, particularly in the period of 2012-2013, which occurred due to the failure of the province of Punjab, the present study region, to meet its target production value. A plausible reason for this deficit was attributed to the poor agricultural planning and inaccurate estimations to satisfy the national grain needs [8]. A report published in the Express Tribune [9] indicates that, similar to the past experience, Pakistan is likely to further face wheat shortages into the future. Due to such uncertainties that directly have a detrimental impact on income and food security for the already staggering economies of developing Pakistan, the government and policymakers require improved forecast models to facilitate them to estimate the potential reductions and associated food security risks due to a shortage of wheat yield. This justifies the pivotal role of data-intelligent models that encapsulate historical patterns in yield with a provision accounting for the surplus and shortfalls, to be embraced for the prediction of yield not only in Pakistan but also in other agricultural nations that may suffer from similar potential risks.

Data-intelligent models have great adaptability for crop planning due to their user friendly implementation, competitive performance and the evolution of data analytic techniques that employ feature detection and subsequent implementation in predictive models [10]. Data-intelligent algorithms are also very attractive tools for policymakers that can enable them to utilize systematic ways for estimating future yield [11]. Nasser and Mehmoud developed accurate photovoltaic power forecasting models using deep LSTM-RNN [12]. Che et al. applied recurrent neural networks for multivariate time series with missing values [13]. There are several examples of data intelligent algorithms in agronomy. The study of Dempewolf et al. [14] aimed to predict wheat yield in Punjab using the vegetation index and measured crop statistics whereas Hamid [15] investigated the wheat economy and its likely future prospects while Muhammad [16] studied historical background of the wheat improvements in Balochistan region. Specifically, Iqbal *et al*.[17] applied an autoregressive moving average (ARIMA) model to project future wheat belt areas and productions up to the year 2022 in Pakistan. Also, Saeed *et al*. [18] predicted wheat in Pakistan using an ARIMA model while Sher and Ahmad [19] developed a study on the prediction of wheat through a Cobb-Douglas function and an ARIMA model for each input data and the respective province. However, these studies have applied simplistic regression models (*e.g.*, ARIMA) that is often discredited due to their assumptions of linearity in the relationships between wheat yield and its predictor variables.

Other than univariate statistical models, there have also been some studies based on rainfall, temperature, fertilizer and other related variables. One such example is the research of Azhar *et al*. [20, 21], that developed a model for the prediction of wheat yield with rainfall-based inputs for the month of November to January in the Punjab Province. Sabir and Tahir [22] developed a model based on exponential smoothing for wheat yield in 2011-12. In spite of being used quite profusely, these types of models embraced general ideas, without incorporating expert information to capitalize on the attributes in historical data, and were largely restricted to linear methods without optimally extracted features. Notwithstanding this, the advent of data-intelligent models at an astonishing rate in the current era can be useful for decision-makers to develop automatic expert systems containing optimal rules based on historical knowledge in farming strategies and crop management [23-25].

Although the literature on application of data-intelligent algorithms for wheat (and other crop) prediction is relatively sparse, some studies show that such a contemporary approach can be an effective way to model future yield. Pantazi et al. [26] predict what yield using self-organizing map and counter-propagation artificial neural networks (CP-ANN), XY-fused Networks (XY-Fs) and Supervised Kohonen Networks (SKNs) in UK. Kumar et al. [27] introduced a crop selection method (CSM) based on advance machine learning technique to improve yield rate. Sanchez et al. [28] applied MLR, M5tree, SVR, MLP and KNN methods for massive crop yield prediction. While Balakrishnan and Muthkumarasamy [29] developed ensemble based AdaSVM and AdaNaive crop yield prediction models in India. Rahman *et al*. [30] developed a machine learning model for rice prediction in Bangladesh, while a neural network integrated model was used [31] for rice crop yield monitoring. Monisha *et al.* [32] applied ANN to predict corn and soybean yield in Malaysia. In addition, these studies have been conducted for a large area, either for a province, or a national region, but not for a small locality (better in terms of accuracy; applicability etc.) such as the site used in this study. Crop prediction could be a difficult task as many variables are interrelated so the yield can be affected by human decisions or activities (*e.g.*, irrigated water, land, fertilization and crop rotation) and uncontrollable, natural factors *(e.g.*, weather) [2].

Therefore, crop planners could potentially generate weather scenario and feed yield forecast models. Despite this, no study has utilized wheat yield of several locations for training purposes to predict the yield of other sites. This sort of strategic modelling can assist in decision-making about the efficiency of production and developing precise agricultural practices.

Techniques utilizing data from several other study sites for training purposes to predict the objective sited data is practically useful since it can enable the modellers to extract similar features and patterns prevalent at the predictor site to be analyzed to estimate the objective site data. This approach can enable agricultural experts to develop farming protocols by comparing site-specific yield and make appropriate deductions in respect to the presence of favorable (or unfavorable) environmental or soil fertility conditions and also implement better management decisions necessary to reach optimal yield. Mehdizadeh [33] developed several data intelligent models (ANN, ANFIS, SVM and MARS) using temperature data from several stations for training to forecast in the objective station in Iran. Deo and Şahin [34] forecasted long-term solar radiation using an ANN model at two sites to develop a model to forecast solar radiation at another site. Considering the needs for accurate future wheat yield prediction, the modelling of crop yield using several sites' yield data for model development can provide a comparative framework for different farmers in identifying more cost-effective and productive agricultural management practices.

In this paper, for the first time, a two-phase hybrid OSELM model integrated with ACO algorithm is developed, denoted as the "ACO-OSELM model". For the purpose of comparison, the standalone extreme learning machine (ELM) and random forest (RF) models are also developed as ELM and RF are considered to be good benchmark models. The two-phase hybrid ACO-OSELM model is tested for wheat yield prediction in agricultural sites: Rahimyar Khan, Dera Ghazi Khan (denoted as D. G. Khan), Kasur, Sialkot, Rawalpindi, and Jhang located in Punjab province, Pakistan where several sites (26 sites) in each case were used to develop the model. The selected study sites are spread throughout the whole Punjab province and are the major wheat producer (see, Figure 1(c)).

To test the applicability of the proposed two-phase hybrid ACO-OSELM model, this study aims to fulfil four objectives: (1) To develop a bio-inspired ACO algorithm to select the best possible sites located in Punjab province, Pakistan for training purposes using feature selection strategy; (2) To incorporate the significant lag at *(t-1)* of the selected training sites in the OSELM model to develop a two-phase hybrid ACO-OSELM hybrid prediction tool; (3) To incorporate the significant antecedent lag of wheat yield effectively to predict the current and future wheat yield; and (4) To validate the predictive ability of the proposed two-phase hybrid ACO-OSELM model for wheat yield prediction universally in whole Punjab province, Pakistan. The novelty of this study is therefore, to design and apply the newly proposed two-phase hybrid ACO-OSELM wheat yield prediction model in Pakistan where out of the 27 sites, 26 are used for training and the remainder 27th site is used for testing in 6 different combinations of the target yield site.

2.0: THEORETICAL FRAMEWORK

Basic theory about the construction of two-phase hybrid ACO-OSELM model for wheat yield prediction is presented in this section.

2.1: Ant colony optimization (ACO) algorithm

ACO applied in this study for the selection of training sites to construct wheat yield time series, is a feature selection algorithm based on swarm optimization technique introduced by Dorigo and Di Caro [35]. ACO has been widely used in different applications [36-40], as a bio-inspired algorithm follow the behavior of ant colonies. In this paper, the ACO algorithm is adopted to locate the minimum possible distance between wheat yield (W) of the training sites, and the testing sites, a feature that can be used to select the respective training site for yield prediction at the testing site. A constant amount of pheromone, a parameter of the ACO algorithm, is assigned to input sites, to classify them against the test site at the beginning of the search. The ant uses the pheromone trial to calculate the probability of the selecting site for training against the testing site where the pheromone values change by traversing the training sites, and consequently, the probability is increased for the new ants to select the best training site. In mathematical notations, the probability $\rho_{js}(t)$ of selecting the shortest distance between the target and the branch is given by:

$$\rho_{js}(\tau)=\frac{\left(\tau_s+\Theta_{js}(\tau)\right)^{\beta}}{\left(\tau_s+\Theta_{js}(\tau)\right)^{\beta}+\left(\tau_s+\Theta_{jl}(\tau)\right)^{\beta}} \quad (1)$$

where $j\in\{1,2\}$ is called a decision point with s and l denote the short and long branch at an instant τ of the total amount of pheromone $\Theta_{js}(\tau)$. The value of $\beta=2$ was computed by Deneubourg, et al. [41]. The probability of the longest path is also computed in this way where $\rho_{js}(\tau)+\rho_{jl}(\tau)=1$. The trail update on the two branches is described as follows:

$$\Theta_{js}(\tau)=\Theta_{js}(\tau-1)+\rho_{js}(\tau-1)m_j(\tau-1)+\rho_{ks}(\tau-1)m_k(\tau-1) \quad (2)$$

$$\Theta_{jl}(\tau)=\Theta_{jl}(\tau-1)+\rho_{jl}(\tau-1)m_j(\tau-1)+\rho_{kl}(\tau-r)m_k(\tau-r) \quad (3)$$

where $j,k \in (1,2,)$ and $m_j(\tau)$ denotes the number of ants on the node j at the time τ which is given by:

$$m_j(\tau) = \rho_{ks}(\tau-1)m_k(\tau-1) + \rho_{kl}(\tau-r)m_k(\tau-r) \quad (4)$$

The ACO algorithm has successfully been applied to traveling salesman problems [42], group shop scheduling [43], vehicle routing [44] and telecommunication networks [45].

The pseudo code of ACO algorithm is following:
N = food source, t_{ij} = initial pheromone, p = pheromone deposited.

Initialize pheromone t_{ij};
Repeat for all ants *i*: construct solution (*i*);
For all ants *i*: global pheromone update (*i*);
For all ants' edges: evaporate pheromone;
($t_{i\text{-}j}$:= (1-p).$t_{i\text{-}j}$)

Construct solution (i):
Initialize ant;
While not yet solution:
Expand the solution by one edge probabilistically according to the pheromone;
($t_{pi\text{-}j}$) / ($\text{sum}_{\text{pi-j}}.t_{pi\text{-}j}$,;)

Global pheromones update (i):
For all edges in the solution;
Increase the pheromone according to the quality;

In this paper, the novelty of ACO algorithm is hybridization with OSELM model to develop a wheat yield prediction tool (ACO-OSELM). The ACO algorithm searches for relevant training sites for the development of ACO-OSELM model. The ACO-ELM model is tested in the agricultural rich districts in Punjab, Pakistan.

2.2: Extreme Learning Machine (ELM)

ELM is an advanced data intelligent model designed by Huang et al. [46] which used a Single Layer Feedforward Neural Network (SLFN). ELM is relatively faster, and thus more computationally efficient than existing data driven models [47]. The ELM can be mathematically formulated as:

$$\sum_{i=1}^{M} \rho_i f(W_k,;c_i,w_i) = W_{pred} \quad (5)$$

In Eq. (5) k ranges from 1 to M with $c_i \in \Gamma$ is the bias of ith node assigned randomly whereas $w_i \in \Gamma$ is a random input weight vector. The function $g(W_k,;c_i,w_i)$ denotes the predicted wheat yield W_{pred} corresponding to the ith hidden node with respect to wheat yield W_k at a lag (*t-1*). Therefore Eq. (5) reduces to the following form:

$$H\beta = Y \quad (6)$$

Where

$$H = \begin{bmatrix} f(W_1,;c_1,,w_1) & \cdots & f(W_1,;c_M,,w_M) \\ \vdots & \cdots & \vdots \\ f(W_N,;c_1,,w_1) & \cdots & f(W_N,;c_M,,w_M) \end{bmatrix}_{N\times M},$$

$$\beta = \left(\beta_1^T \beta_2^T, \cdots, \beta_L^T\right)^T_{m\times M},$$

and $Y = \left(t_1^T t_2^T, \cdots, t_L^T\right)^T_{m\times M}$.

The linear system with solution provides the following output weight:

$$\beta = H^{+}Y \quad (7)$$

Where H^{+} is the inverse of H. The SLFNs with random input weight selection effectively acquire training with least chance of error [48, 49].

2.3: Online Sequential Extreme Learning Machine (OSELM)

The standalone ELM uses all *N*-samples of data for training purposes but data may be used chunk-by-chunk in real world problems because the learning process is time consuming in ELM which requires new training data each time the model is run [50]. As a variant of the standalone ELM model, the OSELM operates in two learning stages *i.e.*, initialization and a sequential learning stage. In OSELM, the matrix H in the initialization stage is packed which is later used in the learning stage. In the initialization stage, the random weights and biases are assigned to the small chunk of initial wheat yield (W) training data to compute the hidden layer output matrix. The sequential learning phase is then commenced either on a one-by-one or lump-by-lump basis and the used data is not allowed to be used again. For more details on OS-ELM, the readers are referred to (*e.g.*, [50-53]).

2.4: Random Forest (RF)

It is noteworthy that the ensemble learning strategies such as bootstraping and bagging generates classifiers and aggregates the results in the form of decision trees [54, 55]. Therefore, the random forest (RF) model is basically a baggging approach with an additional layer of randomness in the prediction process [55]. Every node is split with randomly chosen best subsets of predictors that perform very well and are robust against overfitting [56]. The strategy of RF can be presented as:

Step 1: Construct n_{trees} of bootstrapping from the wheat yield W at lag (*t-1*) where *n* is the number of trees.

Step 2: Grow an unpruned regression tree where the wheat yield (W) as a predictors sample m_{try} is randomly to select optimum split among the predictors.

Step 3: Aggregate the predictions of n_{trees} to predict the wheat yield (W_{pred}).
Detailed analysis on RF can be seen in [56-60].

3.0: DATA AND METHODS

3.1: Study Region and Wheat Yield Data

Wheat yield data has been sourced from Federal Bureau of Statistics (Economic Wing) in Pakistan and the Agriculture Marketing Information Services, Directorate of Agriculture (Economics & Marketing) [61-63]. The study sites included are the agriculture-intensive sites, situated in Punjab, Pakistan. Agricultural sectors in Punjab province play a vital role in the economy with contributions ranging from 56.1% to 61.5% [64]. Further, extensive irrigation systems make this region a rich agricultural locality. Considering the region as a major agricultural belt, the development of data-intelligent models for wheat yield prediction is an interesting research endeavor. To construct the time series wheat yield dataset, the site (district level) productions of wheat were acquired.

Figure 1 illustrates the study sites of wheat farming. Figure 1(a) shows the provinces in Pakistan whereas Figure 1(b) is the map of all sites in Punjab province (current study region). Figure 1(c) is representing total of 6 maps which represents the testing site (yellow colour), training sites (red colour), and the sites where wheat yield data is not available (green colour) and the sites which are not selected by ACO algorithm (blue colour). A total of 27 sites were considered with data from 1981-2013. To obtain the wheat yield time series, out of 27 sites, 26 sites were used for the selection of the best sites for training to develop the model in relation to the remaining (1) testing site. Each time, 26 sites were used to pick the best sites for training subsets against the 6 testing sites. Table 1 presents basic statistics (latitude, longitude, elevation, maximum, minimum, standard deviation, skewness and kurtosis) of the present study sites.

3.2: Model Performance Evaluation

To evaluate the performance of the proposed two-phase hybrid ACO-OSELM *vs*. ACO-ELM and the ACO-RF models applied for wheat yield prediction, statistical and standardized metrics [65] were used. The mathematical formulations of these assessment metrics are given as follows [66-71].

I. Correlation coefficient (r) is expressed as:

$$r = \left(\frac{\sum_{i=1}^{N}\left(W_{obs,i} - \bar{W}_{obs,i}\right)\left(W_{pred,i} - \bar{W}_{pred,i}\right)}{\sqrt{\sum_{i=1}^{N}\left(W_{obs,i} - \bar{W}_{obs,i}\right)^2}\sqrt{\sum_{i=1}^{N}\left(W_{pred,i} - \bar{W}_{pred,i}\right)^2}} \right) \quad (8)$$

II. Willmott's Index (WI) is expressed as:

$$WI = 1 - \left[\frac{\sum_{i=1}^{N}\left(W_{pred,i} - W_{obs,i}\right)^2}{\sum_{i=1}^{N}\left(\left|W_{pred,i} - \bar{W}_{obs,i}\right| + \left|W_{obs,i} - \bar{W}_{obs,i}\right|\right)^2} \right], 0 \le WI \le 1 \quad (9)$$

III. Nash-Sutcliffe coefficient (NS_E) is expressed as:

$$NS_E = 1 - \left[\frac{\sum_{i=1}^{N}\left(W_{obs,i} - W_{pred,i}\right)^2}{\sum_{i=1}^{N}\left(W_{obs,i} - \bar{W}_{pred,i}\right)^2} \right], 0 \le NS_E \le 1 \quad (10)$$

IV. Root mean square error ($RMSE$, $kgha^{-1}$) is expressed as:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(W_{pred,i} - W_{obs,i}\right)^2} \quad (11)$$

V. Mean absolute error (MAE, $kgha^{-1}$) is expressed as:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\left(W_{pred,i} - W_{obs,i}\right)\right| \quad (12)$$

VI. Legates and McCabe's (LM) is expressed as:

$$LM = 1 - \left[\frac{\sum_{i=1}^{N}\left|W_{pred,i} - W_{obs,i}\right|}{\sum_{i=1}^{N}\left|W_{obs,i} - \bar{W}_{obs,i}\right|} \right], 0 \le LM \le 1 \quad (13)$$

VII. Relative root mean square error ($RRMSE$,%) is expressed as:

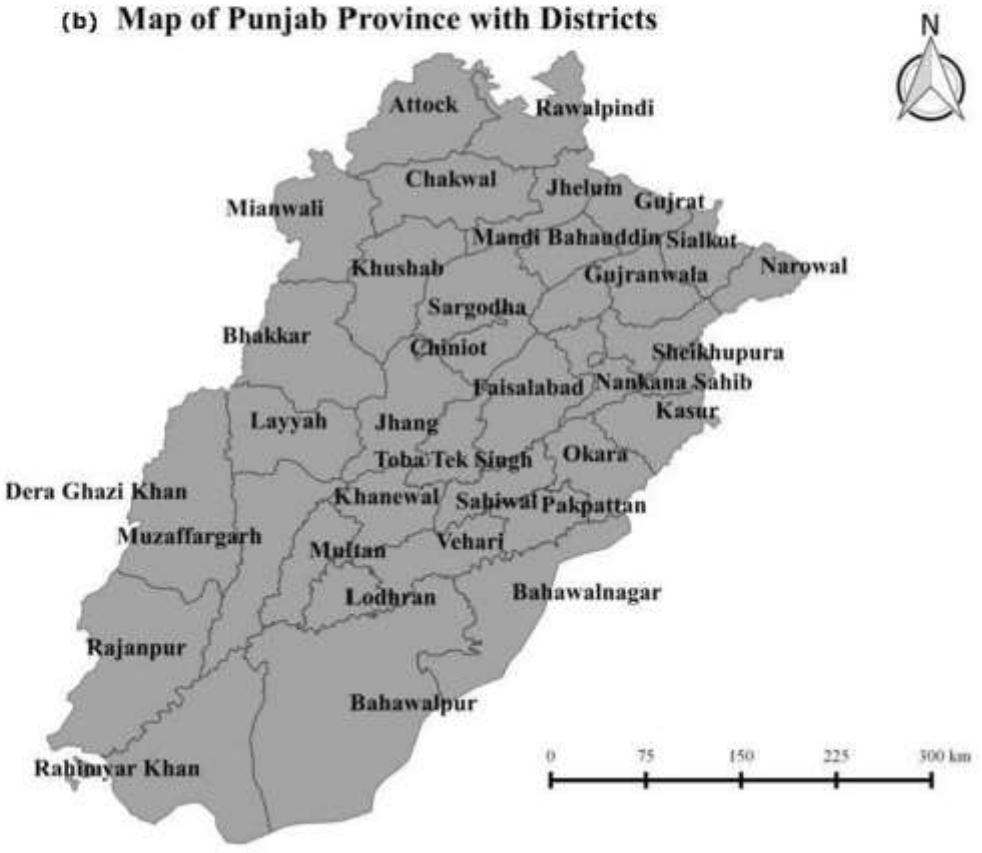
$$RRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(W_{pred,i} - W_{obs,i}\right)^2}}{\frac{1}{N}\sum_{i=1}^{N}\left(W_{obs,i}\right)} \times 100 \quad (14)$$

VIII. Relative mean absolute percentage error ($RMAE$; %), is expressed as

$$RMAE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\left(W_{pred,i} - W_{obs,i}\right)}{W_{obs,i}}\right| \times 100 \quad (15)$$

where $W_{obs,i}$ and $W_{pred,i}$ are the observed and predicted i^{th} value of the wheat yield W, $\bar{W}_{obs,i}$ and $\bar{W}_{pred,i}$ are the observed and predicted average of W and N is the total number of tested data points.


(a) Map of Pakistan
N
Legend
Punjab Province
Pakistan Map
Gilgit Baltistan
Khyber Pakhtunkhwa
Azad Kashmir
Disputed Territory
Fata
Punjab
Balochistan
Sindh
0
250
500
750
1000 km


(b) Map of Punjab Province with Districts
N
Attock
Rawalpindi
Chakwal
Jhelum
Gujrat
Mianwali
Mandi Bahauddin
Sialkot
Khushab
Gujranwala
Narowal
Sargodha
Bhakkar
Chiniot
Sheikhupura
Faisalabad
Nankana Sahib
Kasur
Layyah
Jhang
Toba Tek Singh
Okara
Dera Ghazi Khan
Khanewal
Sahiwal
Pakpattan
Muzaffargarh
Multan
Vehari
Lodhran
Bahawalnagar
Rajanpur
Bahawalpur
Rahimyar Khan
0
75
150
225
300 km

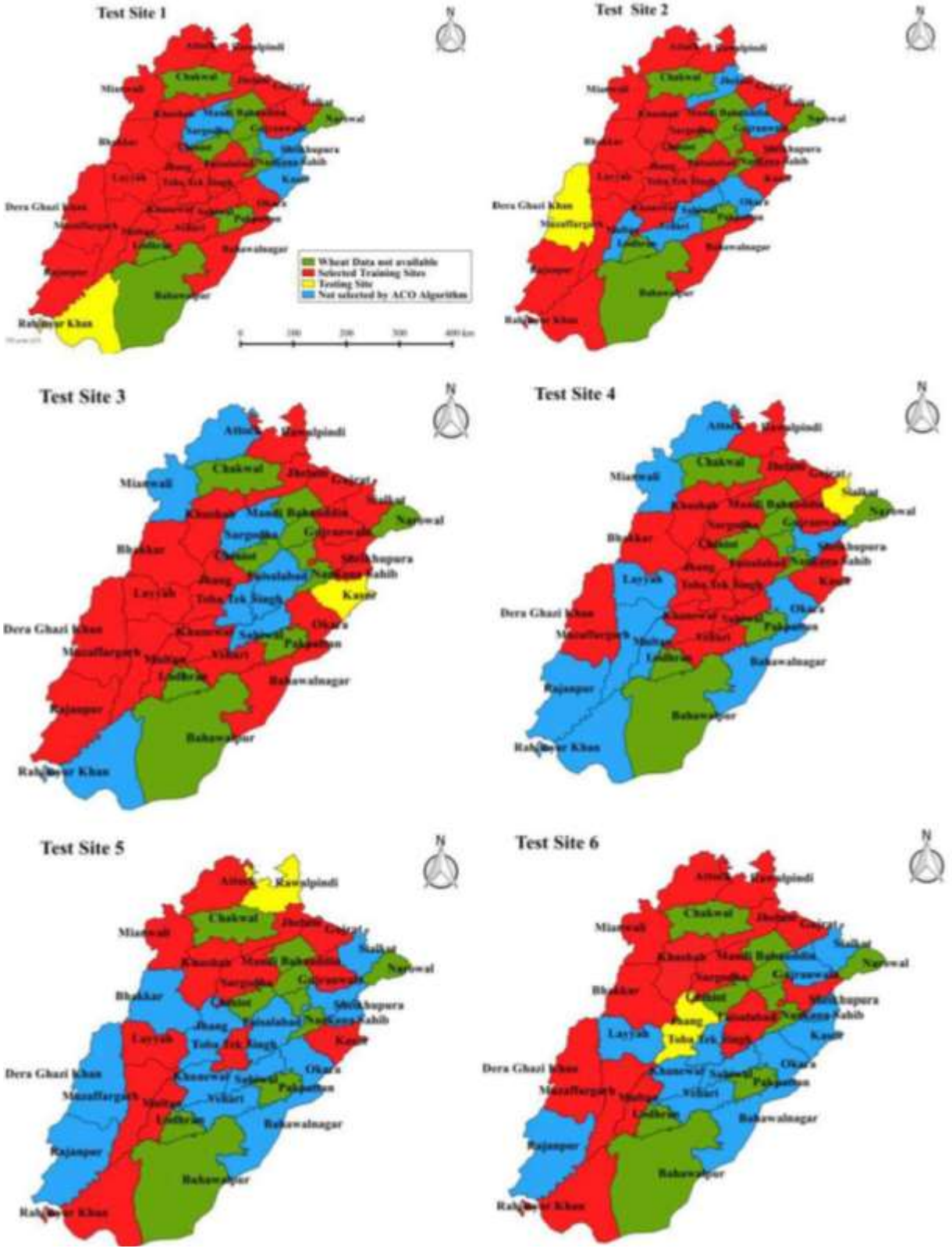


FIGURE 1: Map of the study region. (a) Provinces of Pakistan. (b) Districts of Punjab where the present study was undertaken. (c) Selected training sites in red and the corresponding test site in yellow. Note that the sites shown in green have 'no available wheat yield data' and those in blue were not selected by the Ant Colony Optimisation algorithm.

TABLE 1: Geographic properties and wheat yield statistics of the study sites for Punjab, Pakistan.

station	Geographic characteristics			Wheat yield statistics (kgha^{-1})					
	Latitude (N)	Longitude (E)	Elevation (m)	Mean	Std.	Min.	Max.	Skew.	Kurt.
Shekhupura	31.71°	73.98°	207	2297.46	519.54	1087.89	3062.99	-0.29	-0.65
Okara	30.81°	73.45°	105	2901.89	528.93	1888.08	3588.10	-0.26	-1.16
Sahiwal	30.66°	71.11°	152	2751.87	411.29	1955.17	3792.40	0.12	-0.10
Vehari	30.04°	72.34°	140	2568.94	456.80	1846.00	3462.03	0.09	-0.87
Multan	30.15°	71.52°	122	2307.97	377.77	1713.03	2952.97	0.16	-1.17
Muzaffar Garh	30.07°	71.18°	122	2151.16	470.69	1157.47	2890.72	-0.21	-0.99
Dera Ghazi Khan	30.03°	70.38°	129	2234.18	457.67	1050.86	2914.07	-0.66	0.32
Bakar	45.30°	14.53°	82	2049.52	601.05	1222.01	3440.61	0.79	0.18
Layyah	30.96°	70.94°	143	2025.44	485.42	1160.59	2900.23	0.21	-0.88
Khushab	32.32°	71.90°	195	1455.08	345.98	822.00	2103.04	0.09	-0.88
Sargodha	32.08°	72.66°	189	2253.13	308.08	1637.40	2742.39	0.02	-1.06
Faisalabad	31.45°	73.13°	184	2487.44	511.38	1565.83	3257.58	-0.07	-1.35
Toba Tek Singh	30.97°	72.48°	149	2757.25	715.26	1814.86	3310.60	2.20	5.95
Gujrat	32.57°	74.07°	233	1632.04	267.08	996.51	1985.37	-0.77	-0.03
Rawalpindi	33.56°	73.01°	508	1374.83	372.15	624.01	1993.06	-0.26	-0.68
Jhelum	32.74°	73.72°	234	1418.49	365.23	752.09	2110.03	-0.01	-0.64
Mianwali	32.58°	71.53°	210	1618.80	332.99	1037.09	2510.39	0.63	0.28
Lahore	31.52°	74.35°	217	2485.52	431.22	1385.93	3209.78	-0.97	1.08
Khanewal	30.28°	71.93°	128	2613.95	436.54	1850.94	3600.46	0.24	-0.67
Rajanpur	29.10°	70.32°	97	2138.37	512.46	1009.88	3012.48	-0.70	-0.10
Bahawal Nagar	30.00°	73.24°	163	2298.80	544.52	1401.50	3773.33	0.57	0.08
Attock	33.76°	72.36°	358	1271.96	326.45	685.73	2029.07	0.28	-0.36
Gujranwala	32.15°	74.18°	229	2437.28	604.31	1055.69	3484.13	-0.29	-0.56
Jhang	31.30°	72.32°	158	2353.19	431.16	1637.83	3089.02	0.03	-1.06
Kasur	31.11°	74.44°	218	2495.00	407.78	1688.95	3099.40	-0.07	-0.93
Rahimyar Khan	28.42°	70.29°	80	2308.74	522.80	1312.27	3369.46	0.29	-0.49
Sialkot	32.49°	74.52°	256	2051.86	612.39	598.89	3018.60	-0.44	-0.47

3.3: Design of two-phase hybrid ACO-OSELM Model

The two-phase hybrid ACO-OSELM model was developed using MATLAB R2016, (The Math Works Inc. USA) with Pentium 4 2.93 GHz dual core Central Processing Unit.

Historical wheat yield time series data were used to develop the proposed two-phase universal ACO-OSELM model. The original wheat yield data with statistically significant lagged values at (t – 1) as the input predictor was employed in the first phase of model development. The development of the

two-phase hybrid ACO-OSELM model involved the following phases:

Phase 1: The ACO algorithm is used to determine the best sites for model development in the training period using a feature selection strategy. Further, some predefined parameters were defined at this phase. The number of ants in this phase were 10 with 20 iterations where the initial pheromone is 1 were used. For each site, the numbers of selected sites (features) were defined prior to running the model. For Rahimyar Khan, the number of these selected sites (feature) is 22, D. G. Khan (20), Kasur (19), Sialkot (17), Rawalpindi (12) and Jhang (14). The proposed two-phase hybrid ACO-OSELM model was trained on a longer time series (Site Rahimyar Khan, 726 data points) and shorter time-series (396 data points) for the Rawalpindi site to assess the accuracy for the universal performance so it can applied anywhere in Pakistan in future. Moreover, pheromone exponential weight and heuristic exponential weight are also considered to be 1 here. The selected training sites with their correlation r against testing sites are described in Table 2 whereas Figure 2 plots the RMSE errors of the ACO algorithm between the cost and objective function during the selection of best sites.

After the selections of training sites against testing sites using the ACO algorithm, their correlation *r* (of selected training sites) against testing sites were calculated to confirm the linear relationship among them. For study site Rahimyar Khan, the training site Khanewal has the highest value of $r \approx 0.855$, followed by Bahawal Nagar ($r \approx 0.854$). Similarly, Muzaffar Garh ($r \approx 0.881$) and Rajanpur ($r \approx 0.861$) have the largest values of correlation with site D. G. Khan. For study site Kasur, Gujranwala and Shekhupura attained the highest values of ($r \approx 0.950$, 0.947). For other sites Sialkot, Rawalpindi and Jhang, the readers are referred to Table 2. On the other hand, Site Kasur has the smallest *RMSE* followed by Jhang site between the cost and objective function during feature selection (Figure 2).

Table 3 presents the number of datum points for training and testing purposes in each site with ratio of selected sites against testing sites, skewness, kurtosis, standard Deviation and mean of training and testing data. . The data were normalized between 0 and 1 to avoid the differences in skewness in training and testing period. The normalization for the data is invertible, and hence will not affect the results [72]. Figure 3 presents the time series of the tested study sites constructed from the selected features using the ACO algorithm.

Phase 2: The statistically significant lags of historical wheat yield were calculated from the constructed historical wheat yield time series data using the partial autocorrelation function (PACF) in Figure 4.

After incorporating the significant lag at ($t - 1$) as the input predictor in the OSELM model, different activation functions (sigmoid, sine, hardlim, radial basis) were tested to determine the best activation function. The optimal radial base and sigmoid function (rbf and sig) were found with different numbers of hidden neurons ranging from 7-35 and block size set to 100 in the development of ACO-OSELM. The second significant lag (t-2) was also utized in the proposed two-phase hybrid ACO-OSELM model to check whether it increase the model purformance. But upon utlizing the lag (t-2), it reduces the accuracy of proposed two-phase hybrid ACO-OSELM model, so it was not considered in this paper. For comparison purposes, extreme learning machine (ELM) and random forest (RF) models were also evaluated (Figure 5).

The normalization of the constructed wheat yield time series was accomplished following Eq. (16) to overcome data fluctuation caused by the features, patterns [72] using:

$$W_{norm} = (W - W_{min}) / (W_{max} - W_{min}) \quad (16)$$

In Equation (12), W indicates input/output of the wheat yield data, W_{min} is the smallest value, W_{max} is the largest value of wheat yield in the dataset and W_{norm} is the desired normalized value. To assess the training performance of the proposed two-phase hybrid ACO-OSELM, the correlation coefficient '*r*', and the root mean squared error, *RMSE* was used with ACO-ELM and ACO-RF models (Table 4).

The magnitudes of r and RMSE attained in the training period of the proposed two-phase hybrid ACO-OSELM model for wheat yield prediction at Rahimyar Khan and D. G. Khan were seen to be: (r = 0.812, 0.790, RMSE = 374.82, 381.57 kgha-1). Equivalent metrics for Kasur and Sialkot were found to be: (r = 0.804, 0.798, RMSE = 370.49, 386.18 kgha-1) and finally for Rawalpindi and Jhang were: (r = 0.832, 0.799, RMSE = 356.80, 353.55 kgha-1). For comparison, the ACO-ELM and ACO-RF models were also studied. The magnitudes of these assessment metrics for other sites can be seen in Table 4. The training accuracy of the proposed two-phase hybrid ACO-OSELM model was high. Therefore, it is foreseen that the ACO-OSELM model accuracy in the testing phase, as shown later, is relatively high for wheat yield prediction at these tested sites.

TABLE 2: Selected training stations using Ant Colony Optimization (ACO) algorithm with the correlation coefficient (*r*) for each training station against the testing station.

Test Sites	ACO Selected Training Sites	*Correlation (r)*	Test Sites	ACO Selected Training Sites	*Correlation (r)*	Test Sites	ACO Selected Training Sites	*Correlation (r)*
Site 1- Rahimyar Khan	Khanewal	0.855	**Site 2- D. G. Khan**	Rahimyar	0.781	**Site 3 - Kasur**	Sialkot	0.926
	Faisalabad	0.723		Attock	0.310		Gujranwala	0.950
	Bahawal	0.854		Sialkot	0.779		Jhelum	0.592
	Multan	0.432		Sargodha	0.810		Layyah	0.817
	Gujranwala	0.817		Rajanpur	0.861		Rajanpur	0.664
	D. G. Khan	0.708		Jhang	0.764		Bakkar	0.837
	Khushab	0.638		Layyah	0.816		Bahawal	0.911
	Okara	0.644		Khushab	0.720		Vehari	0.942
	Vehari	0.659		Mianwali	0.643		Jhang	0.930
	Toba Tek	0.556		Lahore	0.644		Lahore	0.406
	Rawalpindi	0.595		Toba Tek	0.685		khanewal	0.805
	Sialkot	0.782		Shekhupura	0.791		Muzaffar	0.858
	Sahiwal	0.602		Kasur	0.780		Okara	0.890
	Layyah	0.568		Bahawal	0.727		D. G. Khan	0.780
	Muzaffar	0.778		Faisalabad	0.812		Multan	0.905
	Attock	0.628		Khanewal	0.704		Shekhupura	0.947
	Jhang	0.848		Bakkar	0.828		Gujrat	0.649
	Bakkar	0.841		Rawalpindi	0.333		Rawalpindi	0.443
	Rajanpur	0.360		Gujrat	0.509		Khushab	0.819
	Mianwali	0.341		Muzaffar	0.881			
	Gujrat	0.480						
	Jhelum	0.185						
Site 4 – Sialkot	Faisalabad	0.420	**Site 5 - Rawalpindi**	Muzaffar	0.560	**Site 6 - Jhang**	Rawalpi	0.537
	D. G. Khan	0.425		Toba Tek	0.801		D. G.	0.764
	Sargodha	0.468		Mianwali	0.758		Multan	0.901
	Gujrat	0.183		Attock	0.876		Gujrat	0.722
	Khushab	0.382		Gujranwala	0.873		Jhelum	0.657
	Jhang	0.522		Multan	0.589		Khusha	0.873
	Vehari	0.479		Jhelum	0.944		Rahimy	0.919
	Lahore	0.310		Sargodha	0.945		Attock	0.571
	Sahiwal	0.346		Khushab	0.907		Sargodh	0.906
	Gujranwala	0.530		Gujrat	0.781		Bakkar	0.842
	Jhelum	0.649		Kasur	0.898		Muzaffa	0.855
	khanewal	0.489		Rahimyar	0.899		Mianwa	0.485
	Bakkar	0.370					Faisalab	0.818
	Rawalpindi	0.181					Shekhu	0.939
	Toba Tek	0.571						
	Khanewal	0.782						
	Kasur	0.463						

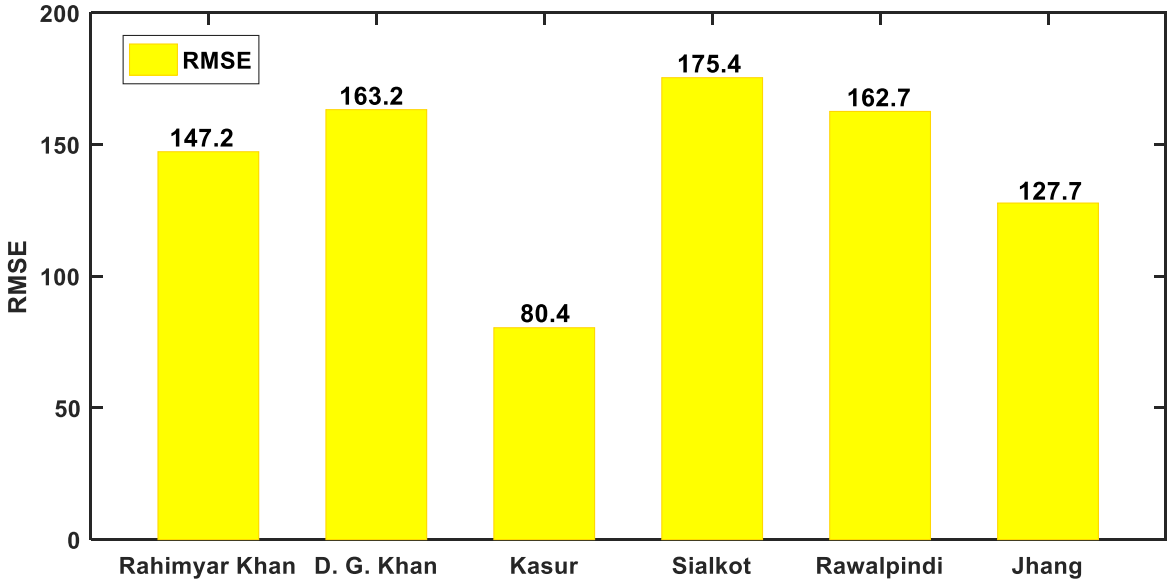


FIGURE 2: Bar graphs of the root mean squared error (RMSE) encountered by the Ant Colony Optimisation algorithm in the selection of training study sites for each testing study site: Site 1: Rahimyar Khan, Site 2: D. G. Khan, Site 3: Kasur, Site 4: Sialkot, Site 5: Rawalpindi, and Site 6: Jhang.

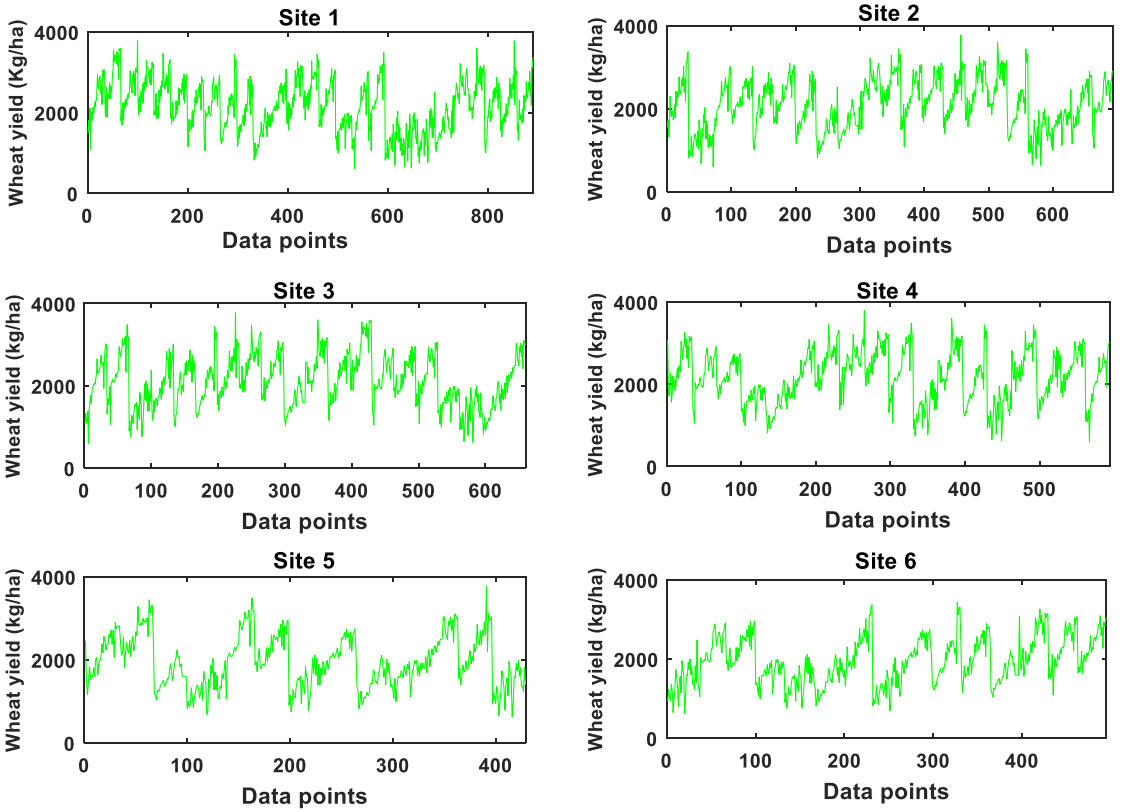


FIGURE 3: Time series of the annual wheat yield data for the training stations selected by the Ant Colony Optimisation algorithm for each testing study site. Site 1: Rahimyar Khan, Site 2: D. G. Khan, Site 3: Kasur, Site 4: Sialkot, Site 5: Rawalpindi, and Site 6: Jhang.

TABLE 3: Training data points (in terms of selected training sites) and testing data point for each testing site using ACO algorithm with skewness and kurtosis of training and testing data.

Testing Sites	No. selected sites	No. of data points in each station	No. of training data	No. of testing data	Skewness		Kurtosis		Standard Deviation		Mean	
					Training	Testing	Training	Testing	Training	Testing	Training	Testing
Rahimyar Khan	22	33	22x33 = 726	33	0.026	0.290	-0.660	-0.491	640.03	522.80	2126.59	2308.74
D. G. Khan	20	33	20x33 = 660	33	-0.023	-0.661	-0.639	0.322	604.39	457.67	2100.48	2234.18
Kasur	19	33	19x33 = 627	33	0.134	-0.072	-0.532	-0.934	623.22	407.78	2101.54	2495.00
Sialkot	17	33	17x33 = 561	33	-0.124	-0.440	-0.647	-0.472	620.67	612.39	2193.14	2051.86
Rawalpindi	12	33	12x33 = 396	33	0.638	-0.264	1.593	-0.680	646.02	372.15	2008.08	1374.83
Jhang	14	33	14x33 = 462	33	0.233	0.033	-0.537	-1.061	588.81	431.16	1918.63	2353.19

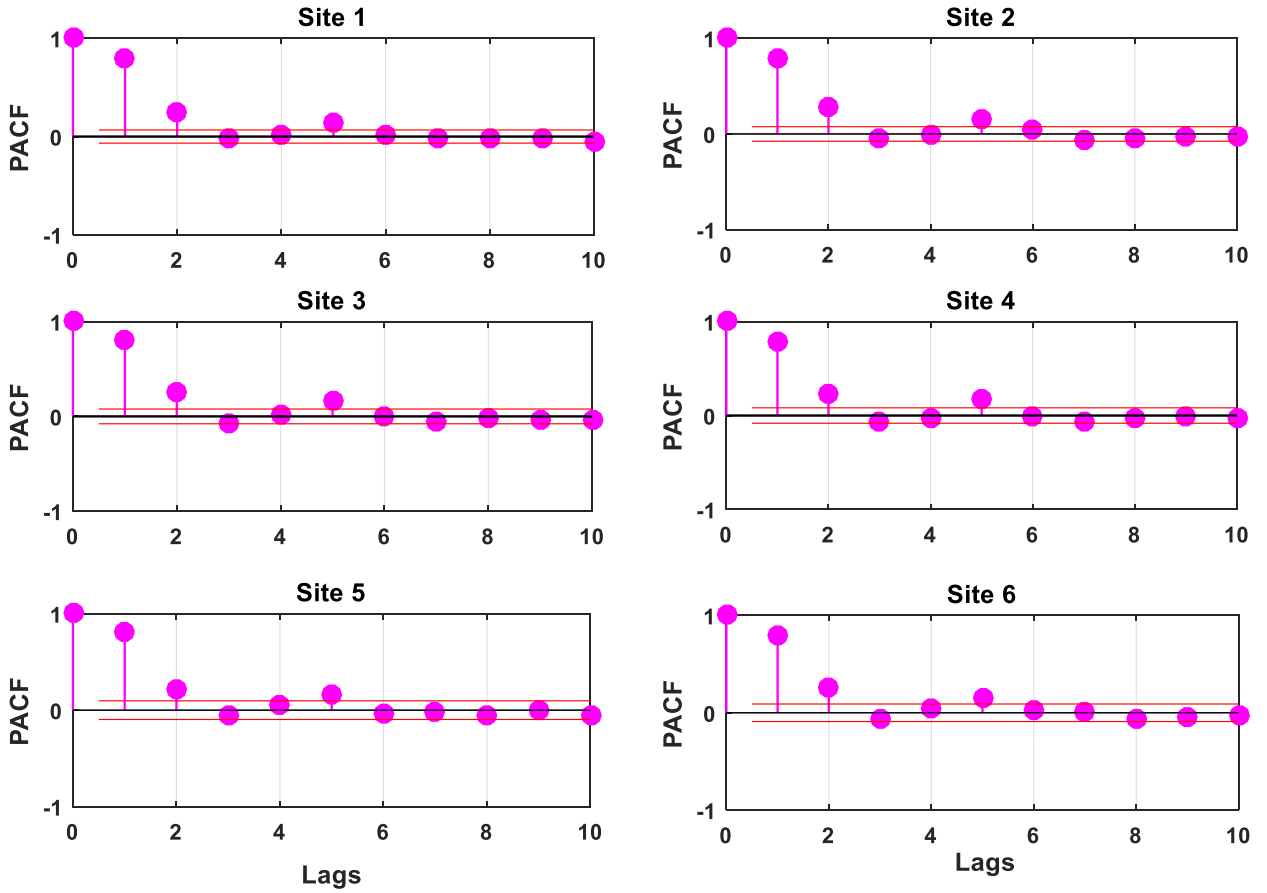


FIGURE 4: Partial autocorrelation function correlation coefficient (PACF) of the historical annual wheat yield time series for each testing study site: Site 1: Rahimyar Khan, Site 2: D. G. Khan, Site 3: Kasur, Site 4: Sialkot, Site 5: Rawalpindi, and Site 6: Jhang.

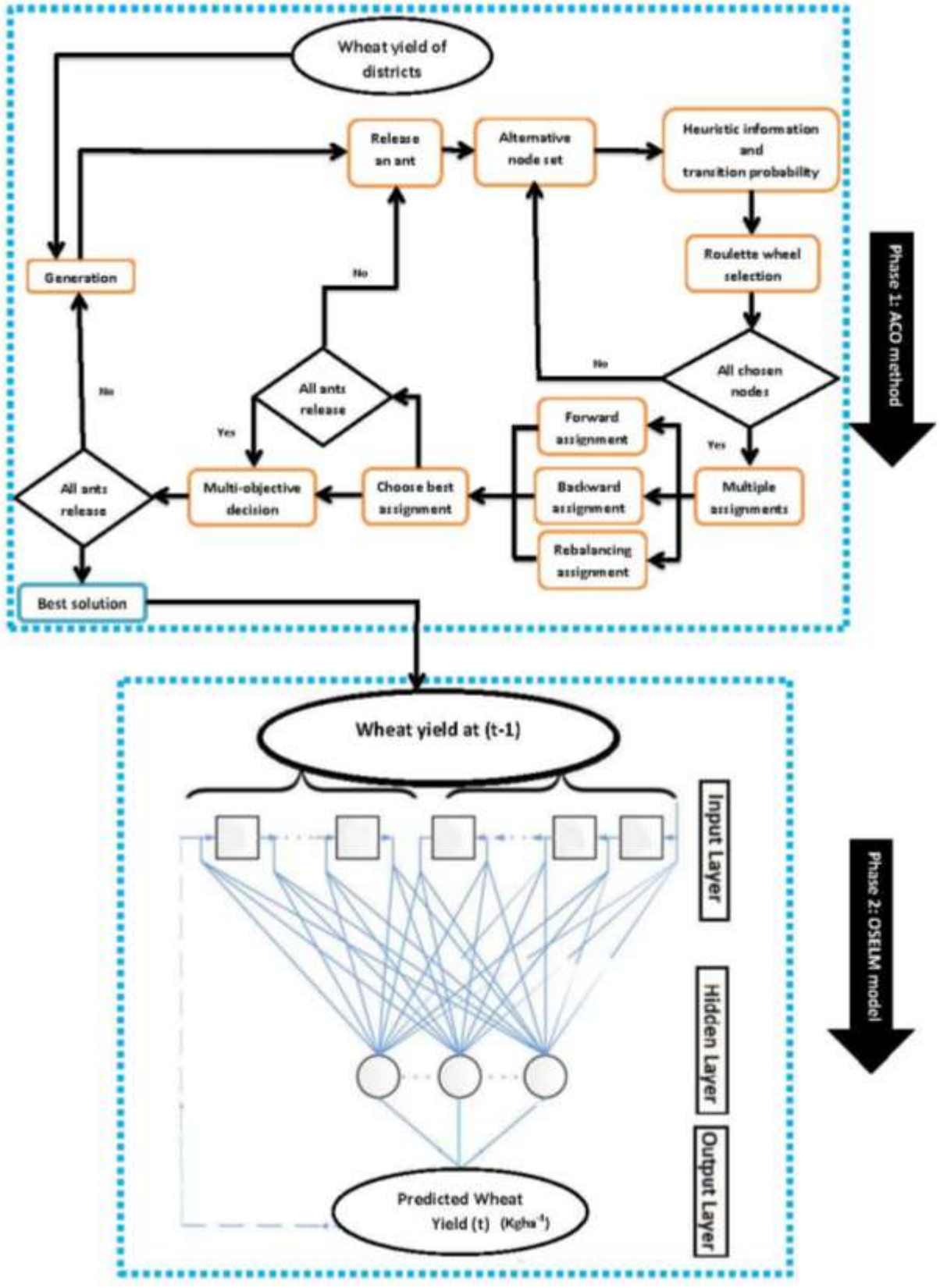


FIGURE 5: Flow chart of the proposed hybrid two-phase Ant Colony Optimization algorithm integrated with Online Sequential Extreme Learning Machine (OSELM) model.

TABLE 4: Training performance of two-phase hybrid **ACO-OSELM** *vs.* ACO-ELM and ACO-RF models with correlation coefficient (*r*) and root mean squared error (*RMSE, kgha*$^{-1}$).

	Lags	ACO-OSELM					ACO-ELM				ACO-RF			
		No. Hidden Neuron	Activation Function	No. Blocks	Training period		No. Hidden Neuron	Activation Function	Training period		No. tress	No. predictor split	Training period	
					RMSE (kg/ha)	*r*			*RMSE (kg/ha)*	*r*			*RMSE (kg/ha)*	*r*
Rahimyar Khan	W_{t-1}	35	rbf	100	374.82	0.812	15	sig	375.99	0.810	10000	2	205.88	0.949
D. G. Khan	W_{t-1}	11	rbf	100	381.57	0.790	15	sig	382.11	0.790	10000	2	212.60	0.942
Kasur	W_{t-1}	7	rbf	100	370.49	0.804	17	rbf	366.78	0.808	10000	2	201.27	0.948
Sialkot	W_{t-1}	15	rbf	100	386.18	0.798	9	rbf	386.57	0.797	10000	2	215.79	0.944
Rawalpindi	W_{t-1}	35	rbf	100	356.80	0.832	17	rbf	357.93	0.831	10000	2	213.64	0.946
Jhang	W_{t-1}	10	sig	100	353.55	0.799	15	sin	352.46	0.799	10000	2	214.46	0.933

4.0: RESULTS

The proposed two-phase hybrid ACO-OSELM is appraised in comparison with ACO-ELM and ACO-RF models, using statistical metrics, diagnostic plots and error distributions (Eq. (8) - (15)) between the predicted and observed yield.

Figure 6 displays a scatterplot with goodness-of-fit and correlation coefficient *r* between predicted and observed wheat yield. The proposed two-phase hybrid ACO-OSELM model is clearly better than ACO-ELM and ACO-RF in terms of r^2 (ACO-OSELM≈0.995, ACO-ELM≈0.996, ACO-RF≈0.862) for Kasur.

Again, the proposed two-phase hybrid ACO-OSELM model is more accurate for Sialkot, r^2 (ACO-OSELM≈0.974, ACO-ELM≈0.936, ACO-RF≈0.892), and Rawalpindi sites in terms of the achieved r^2 (ACO-OSELM≈0.945, ACO-ELM≈0.924, ACO-RF≈0.814). The proposed two-phase hybrid ACO-OSELM model for other sites Rahimyar Khan, D. G. Khan and Jhang is reasonably good compared to ACO-ELM and ACO-RF models (Figure 6). On the basis of attaining the larger r^2-value, the proposed two-phase hybrid ACO-OSELM model shows better accuracy against the comparison models for all the study regions, confirmed by attaining the larger r^2-value.

Figure 7 compares boxplots of the proposed two-phase hybrid ACO-OSELM model with ACO-ELM and ACO-RF models for each site. The + denotes the outliers of the extreme prediction error |PE| of the testing data together with their upper quartile, median and lower quartile. The distributed |PE| is confirmed with a much smaller quartile was acquired by the proposed two-phase hybrid ACO-OSELM model for Rahimyar Khan and D. G. Khan followed by the ACO-ELM and ACO-RF models. The proposed two-phase hybrid ACO-OSELM model again achieved a good accuracy in terms of |PE| for Rawalpindi and Jhang sites in relation to the counterpart models. Similarly, the proposed two-phase hybrid ACO-OSELM model performed well for Sialkot and Kasur sites in predicting wheat yield followed by the ACO-ELM and ACO-RF models. By observing Figure 7, the accuracy of the proposed two-phase hybrid ACO-OSELM model for all sites appeared to be better than the comparative models.

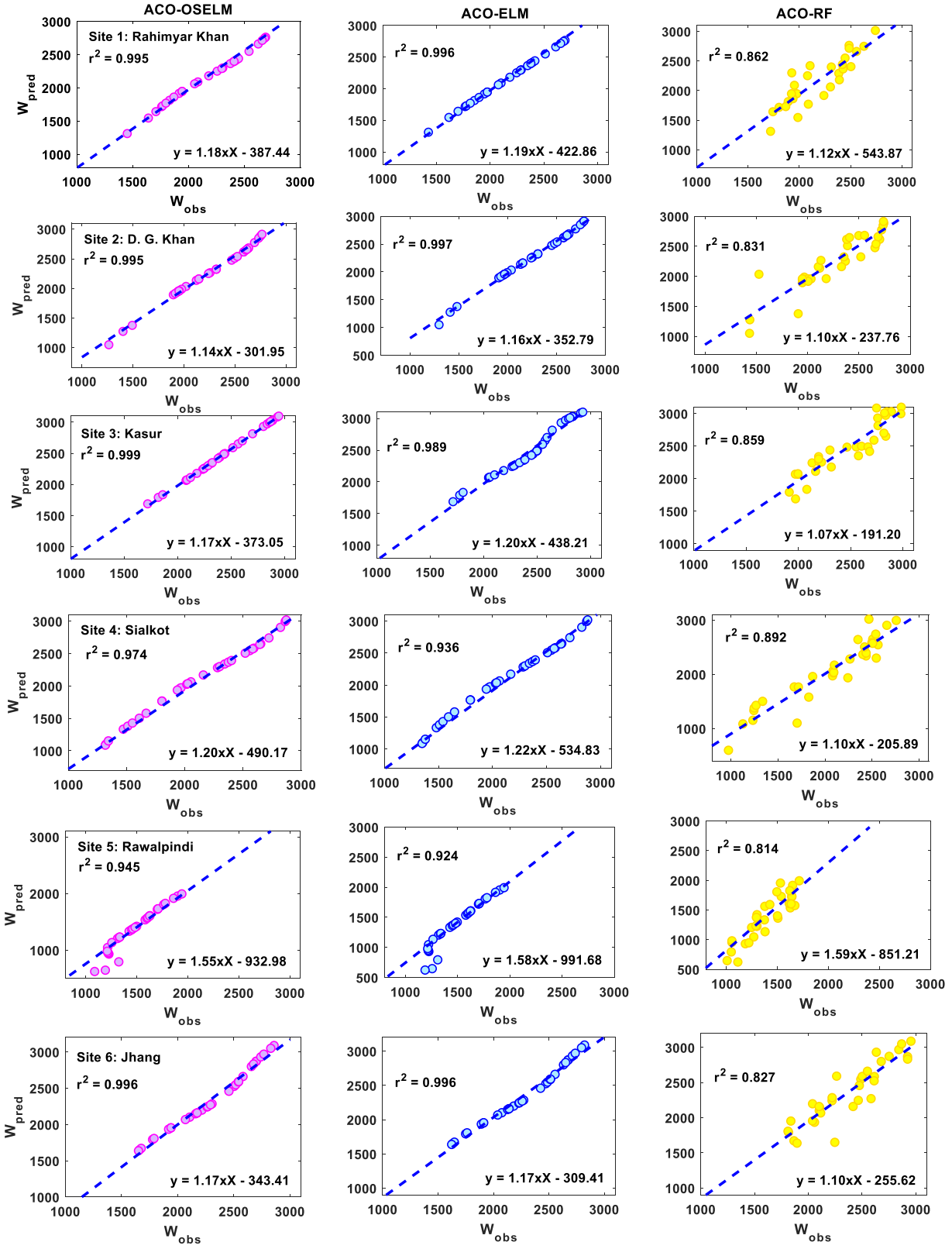


FIGURE 6: Scatterplots of the predicted (Wpred) and observed wheat yield (Wobs) (kgha-1) in the testing phase of the ACO-OSELM vs. ACO-ELM and ACO-RF models including the coefficient of determination (r2) and a linear fit inserted in each panel for the tested study zones.

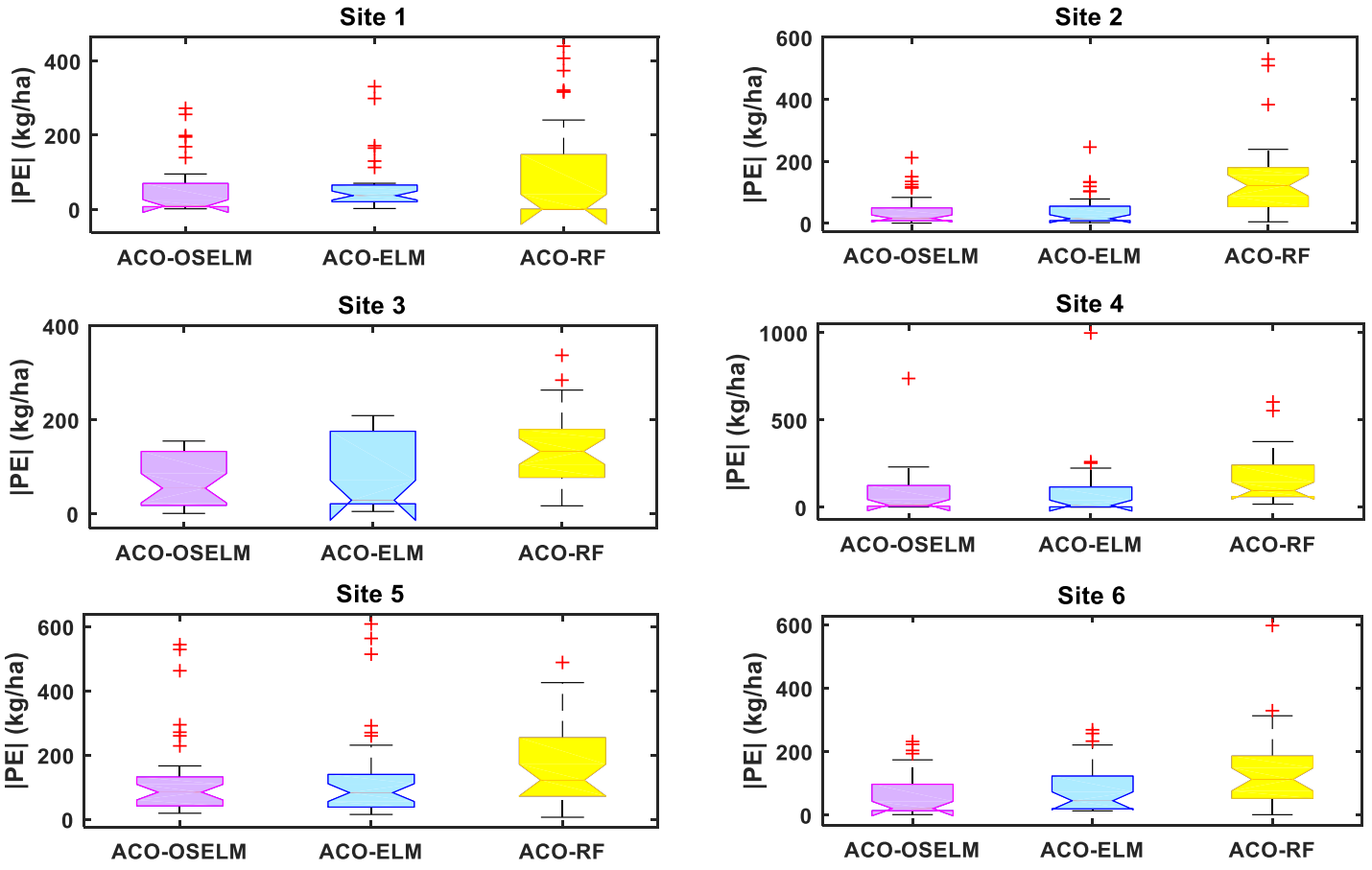


FIGURE 7: Box-plots of the prediction error |PE| (kgha-1) of ACO-OSELM vs. ACO-ELM and ACO-RF models between the predicted and observed wheat yield for Site 1: Rahimyar Khan, Site 2: D. G. Khan, Site 3: Kasur, Site 4: Sialkot, Site 5: Rawalpindi, and Site 6: Jhang.

In Table 5, the preciseness of the proposed two-phase hybrid ACO-OSELM is evaluated in relation to the ACO-ELM and ACO-RF models where the results for each tested site are shown on the basis of *r*, *RMSE* and *MAE*. The proposed two-phase hybrid ACO-OSELM model applied at Kasur site attained the highest correlation coefficient and smallest *RMSE* and *MAE* (r≈0.999, *RMSE*≈85.42kgha^{-1}, *MAE*≈66.54 kgha^{-1}) as compared to the ACO-ELM (r≈0.987, *RMSE*≈111.59 kgha^{-1}, *MAE*≈78.15 kgha^{-1}) and the ACO-RF (r≈0.926, *RMSE*≈154.36 kgha^{-1}, *MAE*≈135.24 kgha^{-1}) model. Moreover, for Sialkot site, these metrics were ACO-OSELM (r≈0.984, *RMSE*≈155.86 kgha^{-1}, *MAE*≈76.95 kgha^{-1}), followed ACO-ELM (r≈0.967, *RMSE*≈197.10 kgha^{-1}, *MAE*≈83.21 kgha^{-1}) and ACO-RF (r≈0.942, *RMSE*≈209.89 kgha^{-1}, *MAE*≈155.35 kgha^{-1}). Similarly, the performance of the proposed two-phase hybrid ACO-OSELM model is better for Site Rawalpindi, Jhang, Rahimyar Khan and D. G. Khan in terms of achieving largest magnitudes of *r* and smallest magnitudes of *RMSE* and *MAE*. This is a clear indication that the proposed two-phase hybrid ACO-OSELM model can be considered to be a better data-intelligent tool for wheat yield prediction as compared to the ACO-ELM and ACO-RF models. The empirical cumulative distribution function (ECDF, Figure 8) at each site depicts the different prediction skills. The proposed two-phase hybrid ACO-OSELM method was reasonably better and superior to both the ACO-ELM and ACO-RF models. Based on the error (0 to ±400 kgha^{-1}) for the Rahimyar Khan, D. G. Khan and Kasur sites, (0 to ±600 kgha^{-1}) for Rawalpindi and Jhang site while (0 to ±1000 kgha^{-1}) for Sialkot site, Figure 9 clearly proves that the proposed two-phase hybrid ACO-OSELM method was the most accurate model in predicting wheat yield.

Table 6 presents the preciseness of the proposed two-phase hybrid ACO-OSELM model in comparison with the ACO-ELM and ACO-RF models, evaluated for all sites in terms of *WI*, NS_E and *LM*. The proposed two-phase hybrid ACO-OSELM model in Rahimyar Khan attained almost similar values of *WI*≈0.980, NS_E≈0.966 and *LM*≈0.865 with ACO-ELM (*WI*≈0.978, NS_E≈0.963 and *LM*≈0.848) better than ACO-RF (*WI*≈0.876, NS_E≈0.830 and *LM*≈0.579) models.

TABLE 5: Testing performance of ACO-OSELM vs. ACO-ELM and ACO-RF models measured by root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (r).

		ACO-OSELM			ACO-ELM			ACO-RF		
Sites	**Input Lags**	***RMSE*** ***(kgha^{-1})***	***MAE*** ***(kgha^{-1})***	***r***	***RMSE*** ***(kgha^{-1})***	***MAE*** ***(kgha^{-1})***	***r***	***RMSE*** ***(kgha^{-1})***	***MAE*** ***(kgha^{-1})***	***r***
Rahimyar Khan	W_{t-1}	94.96	56.16	0.996	99.06	63.26	0.997	212.32	175.25	0.929
D. G. Khan	W_{t-1}	67.12	42.37	0.997	68.98	44.35	0.998	189.44	141.92	0.912
Kasur	W_{t-1}	85.42	66.54	0.999	111.59	78.15	0.987	154.36	135.24	0.926
Sialkot	W_{t-1}	155.86	76.95	0.984	197.10	83.21	0.967	209.89	155.35	0.942
Rawalpindi	W_{t-1}	191.89	129.86	0.967	204.59	134.39	0.955	203.53	165.47	0.898
Jhang	W_{t-1}	96.24	61.41	0.992	114.33	80.81	0.992	181.10	134.10	0.909

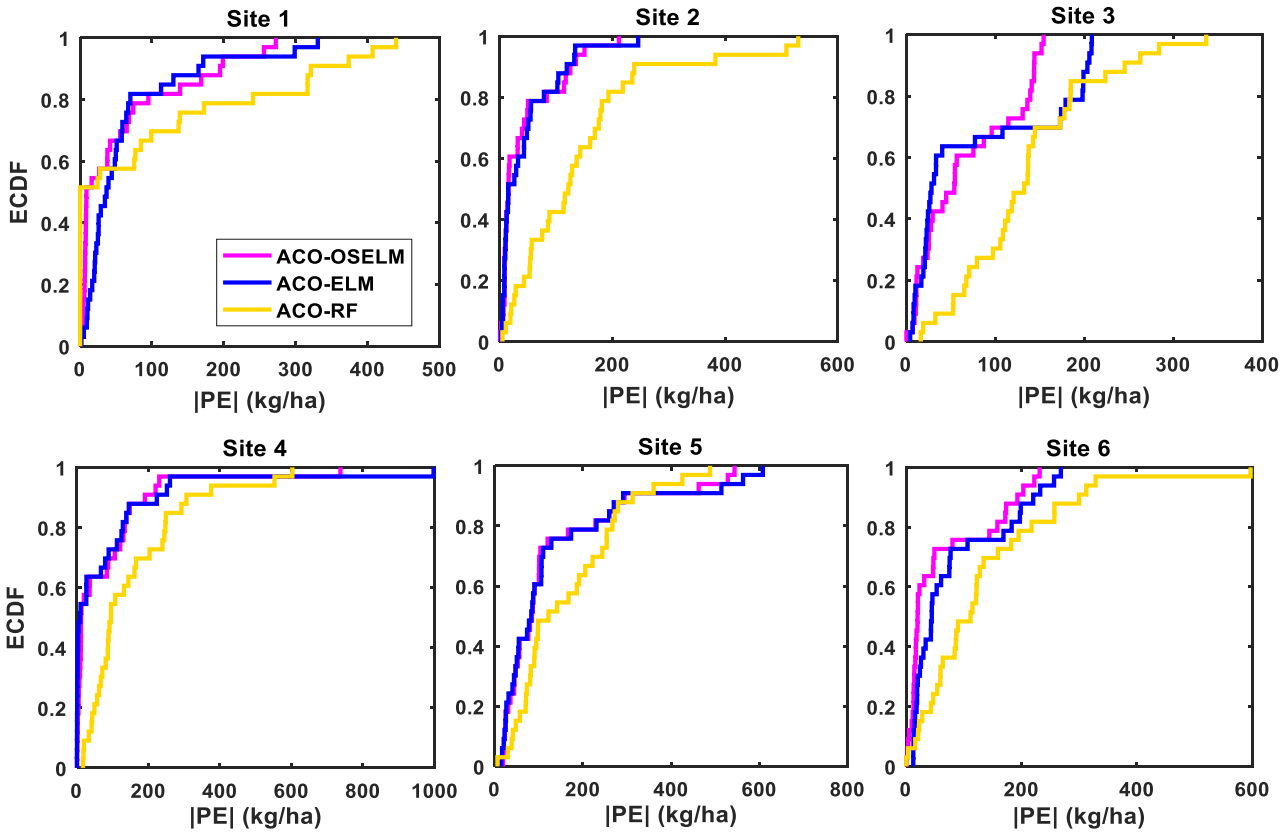


FIGURE 8: Empirical cumulative distribution function (ECDF) of the prediction error, |PE| (kgha-1) for the testing stations using ACO-OSELM vs. ACO-ELM and ACO-RF models.

For D. G. Khan and Kasur sites, the proposed two-phase hybrid ACO-OSELM appeared to have almost similar results (*WI*≈0.989, 0.977, NS_E≈0.978, 0.955, and *LM*≈0.884, 0.805), with ACO-ELM (*WI*≈0.988, 0.962, NS_E≈0.977, 0.923 and *LM*≈0.879, 0.766) but better than ACO-RF (*WI*≈0.903, 0.920, NS_E≈0.823, 0.852 and *LM*≈0.612, 0.595). For other sites Sialkot, Rawalpindi and Jhang, the proposed two-phase hybrid ACO-OSELM model appeared to be the best (see; Table 6) as compared to the counterpart models. The proposed two-phase hybrid ACO-OSELM model shows high accuracy in comparison with the other two counterparts.

Table 7 demonstrates the magnitudes of relative root mean squared error (*RRMSE*) and relative mean absolute error (*RMAE*) for the different locations (Rahimyar Khan, D. G. Khan, Kasur, Sialkot, Rawalpindi and Jhang). D. G. Khan appears to be the most accurate site in predicting wheat yield ACO-OSELM *(RRMSE* ≈3.00%*)* and *(RMAE* ≈2.25%*)* on the basis of *RRMSE* and *RMAE*. The ACO-OSELM model was seen to generate the lowest relative percentage errors (*RRMSE*, *RMAE*) for all tested sites except for Rawalpindi site. But overall, the predicted errors generated by the proposed two-phase hybrid ACO-OSELM model were low in terms of their relative error values, but more importantly, they were within the recommended range of 10% threshold except for Rawalpindi site for an excellent model classification [73].

Figure 9 presents the |PE| yield in each year from 1981-2013 of the proposed two-phase hybrid ACO-OSELM *vs.* ACO-ELM and ACO-RF models at the testing sites. The prediction errors generated by the proposed two-phase hybrid ACO-OSELM were very low compared to the ACO-ELM and ACO-RF models for all sites. This was justified by the minimum values of relative prediction errors. The |PE| errors were significantly smaller in each year for the proposed two-phase hybrid ACO-OSELM model as compared to ACO-ELM and ACO-RF models in Rahimyar Khan, D. G. Khan, Kasur, Sialkot, Rawalpindi and Jhang sites. Overall, the proposed two-phase hybrid ACO-OSELM model generated better significant accuracy with smaller error statistics (Figure 7 and 10) and higher *WI* (Figure 8).

5.0: DISCUSSION: LIMITATION AND FUTURE WORK

Develop strategies which address food scarcity issues, decision-making on national imports and exports and setting the prices in agriculture markets, accurate crop yield prediction can play an important role in policy-making, particularly in agricultural-based nations such as Pakistan. This study has aimed for the first time, to design a two-phase hybrid ACO-OSELM model using significant lag at $(t-1)$ to predict future wheat yield. The approach is practically useful for crop management in terms of using the wheat yield data from several nearby sites in developing better agricultural practices with efficient crop precision technologies. For example, the methodology can be used in the remote areas where meteorological and other agriculture data is not available due to limited resources. The research framework in this study can be applied to any other study site where wheat yield data are available to provide an accurate prediction.

The proposed two-phase hybrid ACO-OSELM model with its counterpart models (ACO-ELM and ACO-RF) was successfully appraised to generate smaller relative percentage errors in terms of *RRMSE* and *RMAE* respectively with a reasonably large statistical correlation of Legates-McCabe's between predicted and observed yield for D. G. Khan and similarly for other tested sites (see Table 6 & 7). The performance was high, according to the achieved relative percentage errors which were less than 10%. Thus, the proposed two-phase hybrid ACO-OSELM model can be used to predict wheat yield where the prediction of a crop commodity will likely become even more important due to an increase of population and for economic growth in terms of exporting to international markets.

The proposed two-phase hybrid OSELM model can be applied on those areas where only wheat yield data is available which could be of interest to the government's national policy-making and agricultural engineers to help minimize crop estimation uncertainties [7, 74]. Due to the aforementioned qualities of the proposed two-phase hybrid ACO-OSELM model, it is possible to apply such a model for the prediction of other crops such as Rice, Maize, Cotton, Sugarcane, Oilseeds and the other coarse grains and pulses that could also be utilized to generate similar predictions in a follow-up study. This may be of great interest to government policy makers and agricultural engineers to avoid the possibility of inaccurate estimation and predictions in the future [7, 74].

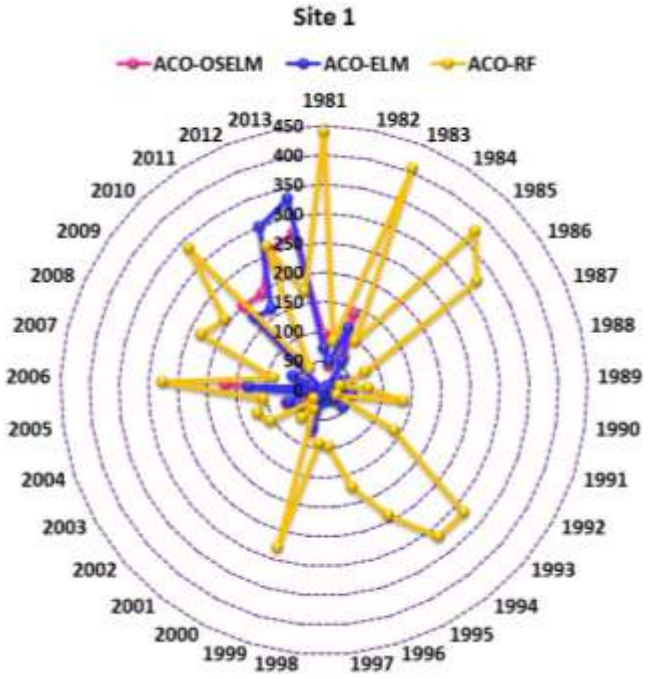
It is noted that the current study utilized the historical wheat yield data to predict the future yield and therefore, carries some limitations. To enhance the scope of this study, meteorological data such as land-surface and air temperatures, rainfall, soil moisture, wind, solar radiation etc. could also be used to predict the crop yield as these parameters greatly impact the crop production amounts. Such predictor variables (which whose data could be remotely sensed through satellites or atmospheric simulation models) (*e.g.*, [11, 14, 31, 75, 76]) are likely to be greatly valuable for modelling crop yield in remotely located agricultural areas. The incorporation of fertilizer (and the relevant soil properties) could also be coupled with meteorological data to explore their use in the proposed two-phase hybrid ACO-OSELM model. Irrigation statistics (*e.g.*, water supply) could also be utilized to improve crop yield and photosynthetically active radiation that governs crop production could be the focus of an independent research study. Further, as processed based modelling are.

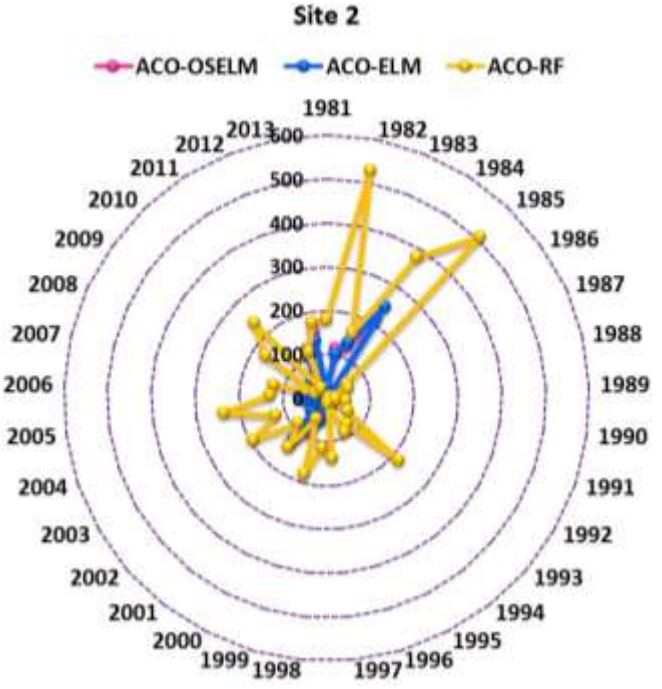
Table 6: The performance of ACO-OSELM vs. ACO-ELM and ACO-RF models using Willmott's index (WI), Nash-Sutcliffe (NSE) and Legates-McCabe's (LM) agreement, for Site 1: Rahimyar Khan, Site 2: D. G. Khan, Site 3: Kasur, Site 4: Sialkot, Site 5: Rawalpindi and Site 6: Jhang. Note that the best model is boldfaced (blue).

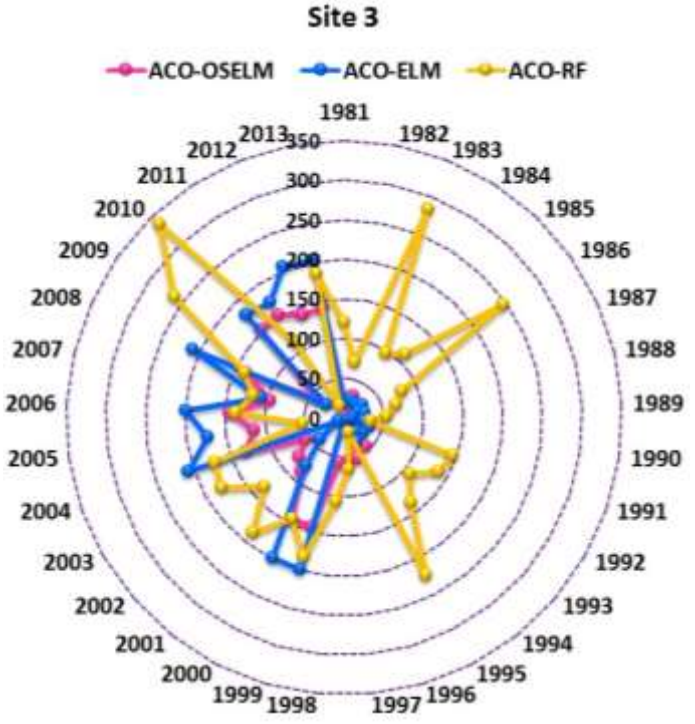
		ACO-OSELM			ACO-ELM			ACO-RF		
Sites	**Input Lags**	***WI***	***NS_E***	***LM***	***WI***	***NS_E***	***LM***	***WI***	***NS_E***	***LM***
Rahimyar Khan	W_{t-1}	**0.980**	**0.966**	**0.865**	0.978	0.963	0.848	0.876	0.830	0.579
D. G. Khan	W_{t-1}	**0.989**	**0.978**	**0.884**	0.988	0.977	0.879	0.903	0.823	0.612
Kasur	W_{t-1}	**0.977**	**0.955**	**0.805**	0.962	0.923	0.766	0.920	0.852	0.595
Sialkot	W_{t-1}	**0.960**	**0.933**	**0.845**	0.931	0.89	0.833	0.941	0.879	0.687
Rawalpindi	W_{t-1}	**0.712**	**0.726**	**0.570**	0.647	0.688	0.555	0.772	0.692	0.453
Jhang	W_{t-1}	**0.974**	**0.949**	**0.833**	0.966	0.927	0.781	0.883	0.818	0.636

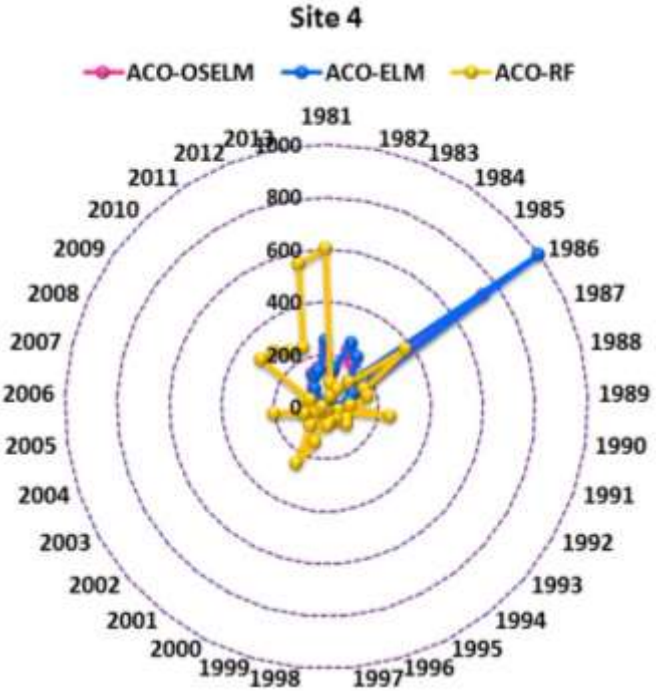
Table 7: Geographic comparison of the accuracy of the ACO-OSELM vs. ACO-ELM and ACO-RF models in terms of relative root mean squared error (RRMSE, %) and the relative mean absolute error (RMAE, %) computed within the test sites. . Note that the best model is boldfaced (blue).

Site and Model Data		**ACO-OSELM**		**ACO-ELM**		**ACO-RF**	
Sites	**Input Lags**	***RRMSE (%)***	***RMAE (%)***	***RRMSE (%)***	***RMAE (%)***	***RRMSE (%)***	***RMAE (%)***
Rahimyar Khan	W_{t-1}	4.11	2.27	4.29	2.57	9.20	8.20
D. G. Khan	W_{t-1}	**3.00**	**2.25**	3.09	2.39	8.48	7.33
Kasur	W_{t-1}	3.42	2.42	4.47	2.79	6.19	5.59
Sialkot	W_{t-1}	7.60	7.40	9.61	8.74	10.23	9.52
Rawalpindi	W_{t-1}	13.96	13.95	14.88	14.70	14.80	14.42
Jhang	W_{t-1}	4.09	2.26	4.86	3.06	7.70	6.09


Site 1
ACO-OSELM
ACO-ELM
ACO-RF
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
450
400
350
300
250
200
150
100
50
0


Site 2
ACO-OSELM
ACO-ELM
ACO-RF
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
500
400
300
200
100
0


Site 3
ACO-OSELM
ACO-ELM
ACO-RF
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
350
300
250
200
150
100
50
0


Site 4
ACO-OSELM
ACO-ELM
ACO-RF
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
1000
800
600
400
200
0

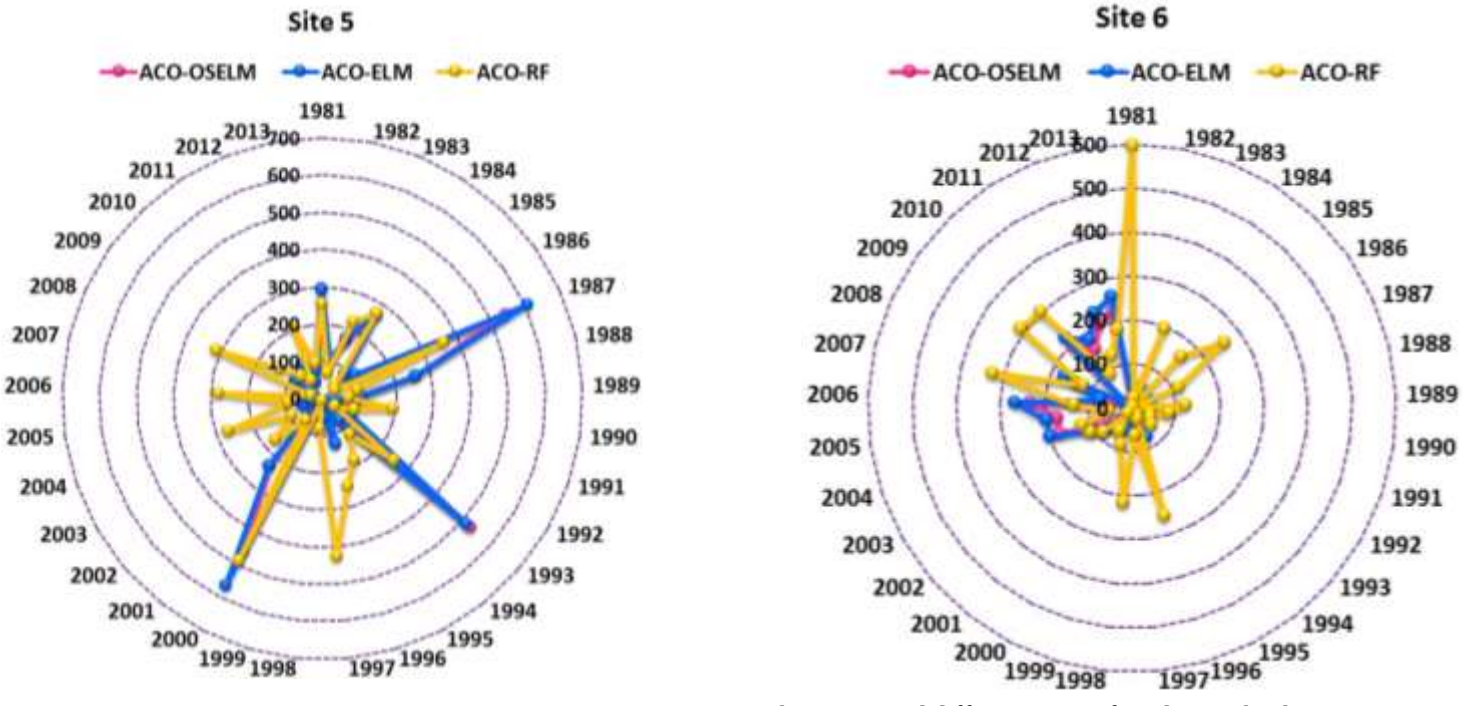


Figure 9: Polar plots showing the prediction error |PE| in each year generated from the ACO-OSELM vs. ACO-ELM and ACO-RF models in predicting wheat yield for Site 1: Rahimyar Khan, Site 2: D. G. Khan, Site 3: Kasur, Site 4: Sialkot, Site 5: Rawalpindi and Site 6: Jhang.

resource demanding and developing countries like Pakistan can't afford, so therefore, the proposed two-phase hybrid ACO-OSELM model could be used as a convenient option.

The two-phase hybrid ACO-OSELM model could be optimized by an ensemble modelling method (with confidence intervals of the yield simulations) to possibly achieves more accurate results. An ensemble of model simulations could lead to a more strategic decision-making model, with the provision of how accurate a simulated yield could be, in terms of the uncertainties between several forecasted data. Other advanced optimization methods applied could include: Quantum-Behaved PSO and the Firefly Algorithm to select training sites which have been tested to hybridize with the OSELM model (*e.g.*, [77-85]). Further, empirical wavelet transform [86], empirical mode composition [87] and singular value decomposition [88] could be another venue of this work. Future work could apply copula functions [89] as statistical tools where joint behaviour of multivariate data (*e.g.*, yield and corresponding predictors) can be modelled for any tested sites. Moreover, long-short term memory (LSTM) and recurrent neural network (RNN) can be used to further broaden the scope of this study in the follow up work [12, 13]. Further, we have used wheat yield data from different sites to train the OSELM, ELM and RF models and predicted the wheat yield in another testing site. Therefore, it's clear that the data isn't complex and convoluted as the input data has the same nature with target. Moreover, we complex and complicated nature. That's why ELM and RF generated close results with OSELM (slightly better than ELM and RF). Thus, it is most likely that the performance of LSTM will be almost similar to ACO-OSELM.

6.0: CONCLUSION

This paper designed a two-phase hybrid ACO-OSELM model to predict wheat yield on district level. In the first phase, the ACO algorithm selects the best district/site of for training using the wheat yield from 26 sites. The selected sites were used to construct a time series and PACF were utilized to determine the statistically significant lag at *(t-1)*. The PACF lagged were used to train the OSEL model and validated in testing site to predict wheat yield to develop the proposed two-phase hybrid ACO-OSELM model in order to achieve a high level of accuracy. Further, several types of evaluation criterion were adopted to judge the accuracy of the proposed two-phase hybrid ACO-OSELM model.

The proposed two-phase hybrid ACO-OSELM model was compared with ACO-ELM and ACO-RF. As evident by low relative prediction errors and high performance metrics, the performance of the proposed two-phase hybrid ACO-OSELM model is much better than that of counterpart models. The prediction errors for the best site

D. G. Khan and for the proposed two-phase hybrid ACO-OSELM were *RMSE*≈67.12 kgha^{-1} whereas the *MAE* was 42.37 kg/ha. In terms of normalized performance metrics, the values of these metrics for the D. G. Khan site of proposed two-phase hybrid ACO-OSELM model were *r*≈0.997, *WI*≈0.989 and NS_E≈0.978 (See; Table 5 & 6).

To assess the performance of the proposed two-phase hybrid ACO-OSELM in relation to ACO-ELM and the ACO-RF models using Legates-McCabe's. The obtained *LM* agreement values between the predicted and observed wheat yield for the D. G. Khan study site were *LM*≈0.884 (ACO-OSELM), 0.879 (ACO-ELM) and 0.612 (ACO-RF) respectively whereas the relative percentage errors *RRMSE* and *RMAE* were only 3.00%, 2.25% (ACO-OSELM) compared with 3.09%, 2.39% (ACO-ELM) and 8.48%, 7.33% (ACO-RF). A reasonable degree of geographic variability was evident on the basis of relative percentage errors *RRMSE* and *RMAE* with the optimal performance acquired for D. G. Khan site as compared to other sites being the accuracy of the proposed two-phase hybrid ACO-OSELM model.

This study provides a baseline in terms of using wheat yield data from several sites, being potentially utilized to predict wheat yield and other climatological parameters more accurately in future studies. The two-phase hybrid ACO-OSELM model may be applied to other agricultural crop yield prediction scenarios that will assist agricultural policy makers in Pakistan in the optimal management of crop estimation. Moreover, accurate wheat yield prediction can be used to alert the government and impacted stakeholders prior to significant food security.

ACKNOWLEDGMENT

This paper has utilized wheat yield acquired from Bureau of Statistics, Government of Pakistan that is duly acknowledged. We duly acknowledge that this research project has been sponsored by The University of Southern Queensland's Postgraduate Research Scholarship (2017–2019) awarded to the first author, managed by the Office of Research and Graduate Studies Division.

REFERENCES

[1] A. Raorane and R. Kulkarni, "Data Mining: An effective tool for yield estimation in the agricultural sector," *International Journal of Emerging Trends of Technology in Computer Science*, vol. 1, pp. 75-79, 2012.

[2] A. Gonzalez-Sanchez, J. Frausto-Solis, and W. Ojeda-Bustamante, "Attribute selection impact on linear and nonlinear regression models for crop yield prediction," *The Scientific World Journal*, vol. 2014, 2014.

[3] K. A. Khan S. B., Hafeezullah., Rehman, A. , " Agricultural Statistics of Pakistan " Pakistan Bureau of Statistics, Government of Pakistan, Islamabad, Pakistan2012.

[4] E. Survey, "Pakistan Economic Survey 2011–2012," Ministry of Finance of the Government of Pakistan2012.

[5] FAO, "Review of the sector and grain storage issue," 2013.

[6] P. Dorosh and A. Salam, "Wheat markets and price stabilisation in Pakistan: An analysis of policy options," *The Pakistan Development Review*, pp. 71-87, 2008.

[7] M. S. Niaz, "Wheat Policy—A Success or Failure," DAWN Newspaper2014.

[8] A. Bokhari, "Wheat crisis in the making," Dawn newspaper, Pakistan2013.

[9] S. A. Sajjad, "Story of Pakistan's elite wheat, Published in The Express Tribune," THE EXPRESS TRIBUNE Pakistan2017.

[10] A. Irmak, J. Jones, W. Batchelor, S. Irmak, K. Boote, and J. Paz, "Artificial neural network model as a data analysis tool in precision farming," *Transactions of the ASABE*, vol. 49, pp. 2027-2037, 2006.

[11] M. E. Bauer, "The role of remote sensing in determining the distribution and yield of crops," *Advances in Agronomy*, vol. 27, pp. 271-304, 1975.

[12] M. Abdel-Nasser and K. Mahmoud, "Accurate photovoltaic power forecasting models using deep LSTM-RNN," *Neural Computing and Applications*, pp. 1-14, 2017.

[13] Z. Che, S. Purushotham, K. Cho, D. Sontag, and Y. Liu, "Recurrent neural networks for multivariate time series with missing values," *Scientific reports*, vol. 8, p. 6085, 2018.

[14] J. Dempewolf, B. Adusei, I. Becker-Reshef, M. Hansen, P. Potapov, A. Khan, *et al.*, "Wheat yield forecasting for Punjab Province from vegetation index time series and historic crop statistics," *Remote Sensing*, vol. 6, pp. 9653-9675, 2014.

[15] N. P. Hamid, Alberto T. V., and Suzanne G.,, "The wheat economy of Pakistan setting and prospects.," International Food Policy Research Institute, Ministry of Food and Agriculture, Government of Pakistan, Islamabad, Pakistan1987.

[16] K. Muhammad, "Description of the historical background of wheat improvement in baluchistan, Pakistan," *Agriculture Research Institute (Sariab, Quetta, Baluchistan, Pakistan)*, 1989.

[17] N. Iqbal, K. Bakhsh, A. Maqbool, and A. S. Ahmad, "Use of the ARIMA model for forecasting wheat area and production in Pakistan," *Journal of Agriculture and Social Sciences*, vol. 1, pp. 120-122, 2005.

[18] N. Saeed, A. Saeed, M. Zakria, and T. M. Bajwa, "Forecasting of wheat production in Pakistan using ARIMA models," *International Journal of Agricultural Biology*, vol. 2, pp. 352-353, 2000.

[19] F. Sher and E. Ahmad, "Forecasting wheat production in Pakistan," 2008.

[20] B. Azhar, M. G. Chaudhry, and M. Shafique, "A model for forecasting wheat production in the Punjab," *The Pakistan Development Review*, vol. 12, pp. 407-415, 1973.

[21] B. Azhar, M. G. Chaudhry, and M. Shafique, "A Forecast of Wheat Production in the Punjab for 1973-74," *The Pakistan Development Review*, vol. 13, pp. 106-112, 1974.

[22] H. Sabir and S. Tahir, "Supply and demand projection of wheat in Punjab for the year 2011-2012," *Interdis. J. Contemp. Res. Bus*, vol. 3, pp. 800-808, 2012.

[23] R. J. McQueen, S. R. Garner, C. G. Nevill-Manning, and I. H. Witten, "Applying machine learning to agricultural data," *Computers and electronics in agriculture*, vol. 12, pp. 275-293, 1995.

[24] S. Dimitriadis and C. Goumopoulos, "Applying machine learning to extract new knowledge in precision agriculture applications," in *Informatics, 2008. PCI'08. Panhellenic Conference on*, 2008, pp. 100-104.

[25] E. J. Coopersmith, B. S. Minsker, C. E. Wenzel, and B. J. Gilmore, "Machine learning assessments of soil drying for agricultural planning," *Computers and electronics in agriculture*, vol. 104, pp. 93-104, 2014.

[26] X. E. Pantazi, D. Moshou, T. Alexandridis, R. Whetton, and A. M. Mouazen, "Wheat yield prediction using machine learning and advanced sensing techniques," *Computers and Electronics in Agriculture*, vol. 121, pp. 57-65, 2016.

[27] R. Kumar, M. Singh, P. Kumar, and J. Singh, "Crop Selection Method to maximize crop yield rate using machine learning technique," in *2015 international conference on smart*

technologies and management for computing, communication, controls, energy and materials (ICSTM), 2015, pp. 138-145.
[28] A. González Sánchez, J. Frausto Solís, and W. Ojeda Bustamante, "Predictive ability of machine learning methods for massive crop yield prediction," 2014.
[29] N. Balakrishnan and G. Muthukumarasamy, "Crop production-ensemble machine learning model for prediction," *International Journal of Computer Science and Software Engineering*, vol. 5, p. 148, 2016.
[30] M. M. Rahman, N. Haq, and R. M. Rahman, "Machine Learning Facilitated Rice Prediction in Bangladesh," in *Information and Computer Technology (GOCICT), 2014 Annual Global Online Conference on*, 2014, pp. 1-4.
[31] C. Chen and H. McNairn, "A neural network integrated approach for rice crop monitoring," *International Journal of Remote Sensing*, vol. 27, pp. 1367-1393, 2006.
[32] M. Kaul, R. L. Hill, and C. Walthall, "Artificial neural networks for corn and soybean yield prediction," *Agricultural Systems*, vol. 85, pp. 1-18, 2005.
[33] S. Mehdizadeh, "Assessing the potential of data-driven models for estimation of long-term monthly temperatures," *Computers and Electronics in Agriculture*, vol. 144, pp. 114-125, 2018.
[34] R. C. Deo and M. Şahin, "Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland," *Renewable and Sustainable Energy Reviews*, vol. 72, pp. 828-848, 2017.
[35] M. Dorigo and G. Di Caro, "Ant colony optimization: a new meta-heuristic," in *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, 1999, pp. 1470-1477.
[36] A. Badr and A. Fahmy, "A proof of convergence for ant algorithms," *Information Sciences*, vol. 160, pp. 267-279, 2004.
[37] R. J. Mullen, D. Monekosso, S. Barman, and P. Remagnino, "A review of ant algorithms," *Expert systems with Applications*, vol. 36, pp. 9608-9617, 2009.
[38] J. D. Sweetlin, H. K. Nehemiah, and A. Kannan, "Feature selection using ant colony optimization with tandem-run recruitment to diagnose bronchitis from CT scan images," *Computer methods and programs in biomedicine*, vol. 145, pp. 115-125, 2017.
[39] O. Cordón García, F. Herrera Triguero, and T. Stützle, "A review on the ant colony optimization metaheuristic: Basis, models and new trends," *Mathware & soft computing. 2002 Vol. 9 Núm. 2 [-3]*, 2002.
[40] G. Singh, N. Kumar, and A. K. Verma, "Ant colony algorithms in MANETs: A review," *Journal of Network and Computer Applications*, vol. 35, pp. 1964-1972, 2012.
[41] J.-L. Deneubourg, S. Aron, S. Goss, and J. M. Pasteels, "The self-organizing exploratory pattern of the argentine ant," *Journal of insect behavior*, vol. 3, pp. 159-168, 1990.
[42] L. Bianchi, L. M. Gambardella, and M. Dorigo, "An ant colony optimization approach to the probabilistic traveling salesman problem," in *International Conference on Parallel Problem Solving from Nature*, 2002, pp. 883-892.
[43] C. Blum, "ACO applied to group shop scheduling: A case study on intensification and diversification," in *International Workshop on Ant Algorithms*, 2002, pp. 14-27.
[44] O. Bräysy, "A reactive variable neighborhood search for the vehicle-routing problem with time windows," *INFORMS Journal on Computing*, vol. 15, pp. 347-368, 2003.
[45] R. Cordone and F. Maffioli, "Coloured Ant System and local search to design local telecommunication networks," in *Workshops on Applications of Evolutionary Computation*, 2001, pp. 60-69.
[46] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: theory and applications," *Neurocomputing*, vol. 70, pp. 489-501, 2006.
[47] R. Rajesh and J. S. Prakash, "Extreme learning machines-a review and state-of-the-art," *International journal of wisdom based computing*, vol. 1, pp. 35-49, 2011.
[48] S. i. Tamura and M. Tateishi, "Capabilities of a four-layered feedforward neural network: four layers versus three," *IEEE Transactions on Neural Networks*, vol. 8, pp. 251-255, 1997.
[49] G.-B. Huang, "Learning capability and storage capacity of two-hidden-layer feedforward networks," *IEEE Transactions on Neural Networks*, vol. 14, pp. 274-281, 2003.
[50] M. Ali, R. C. Deo, N. J. Downs, and T. Maraseni, "Multi-stage hybridized online sequential extreme learning machine integrated with Markov Chain Monte Carlo copula-Bat algorithm for rainfall forecasting," *Atmospheric Research*, vol. https://doi.org/10.1016/j.atmosres.2018.07.005, 2018.
[51] N.-Y. Liang, G.-B. Huang, P. Saratchandran, and N. Sundararajan, "A fast and accurate online sequential learning algorithm for feedforward networks," *IEEE Transactions on neural networks*, vol. 17, pp. 1411-1423, 2006.
[52] Y. Lan, Y. C. Soh, and G.-B. Huang, "Ensemble of online sequential extreme learning machine," *Neurocomputing*, vol. 72, pp. 3391-3395, 2009.
[53] B. Yadav, S. Ch, S. Mathur, and J. Adamowski, "Discharge forecasting using an online sequential extreme learning machine (OS-ELM) model: a case study in Neckar River, Germany," *Measurement*, vol. 92, pp. 433-445, 2016.
[54] R. E. Schapire, Y. Freund, P. Bartlett, and W. S. Lee, "Boosting the margin: A new explanation for the effectiveness of voting methods," *Annals of statistics*, pp. 1651-1686, 1998.
[55] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, pp. 123-140, 1996.
[56] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.
[57] B. Robert, F. Yoav, B. Peter, and L. W. Sun, "Boosting the margin: a new explanation for the effectiveness of voting methods," *The annals of statistics*, vol. 26, pp. 1651-1686, 1998.
[58] M. R. Segal, "Machine learning benchmarks and random forest regression," 2004.
[59] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news*, vol. 2, pp. 18-22, 2002.
[60] R. Prasad, R. C. Deo, Y. Li, and T. Maraseni, "Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition," *Geoderma*, vol. 330, pp. 136-161, 2018.
[61] A. M. I. Service. (2012). *DISTRICT-WISE AREA OF WHEAT CROP*. Available: http://www.amis.pk/Agristatistics/DistrictWise/2010-2012/Wheat.html
[62] A. M. I. Service. (2014). *DISTRICT-WISE AREA OF WHEAT CROP*. Available: http://www.amis.pk/Agristatistics/DistrictWise/2012-2014/Wheat.html
[63] C. a. a. p. b. Districts, "Crops area and production by Districts 1981-2008, Food and Cash crops," Federal Bureau of Statistics (Economic wing), Islamabad, Pakistan.2008.
[64] P. Punjab. (2015). *Population*. Available: https://en.wikipedia.org/wiki/Punjab,_Pakistan
[65] B. C. Yen, "Discussion and Closure: Criteria for Evaluation of Watershed Models," *Journal of Irrigation and Drainage Engineering*, vol. 121, pp. 130-132, 1995.
[66] C. W. Dawson, R. J. Abrahart, and L. M. See, "HydroTest: a web-based toolbox of evaluation metrics for the standardised assessment of hydrological forecasts," *Environmental Modelling & Software*, vol. 22, pp. 1034-1052, 2007.
[67] R. C. Deo, X. Wen, and F. Qi, "A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset," *Applied Energy*, vol. 168, pp. 568-593, 2016.
[68] D. R. Legates and G. J. McCabe, "Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation," *Water resources research*, vol. 35, pp. 233-241, 1999.
[69] C. J. Willmott, "Some comments on the evaluation of model performance," *Bulletin of the American Meteorological Society*, vol. 63, pp. 1309-1313, 1982.

[70] C. J. Willmott, "On the validation of models," *Physical geography*, vol. 2, pp. 184-194, 1981.
[71] C. J. Willmott, "On the evaluation of model performance in physical geography," in *Spatial statistics and models*, ed: Springer, 1984, pp. 443-460.
[72] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," 2003.
[73] F. D. Foresee and M. T. Hagan, "Gauss-Newton approximation to Bayesian learning," in *Neural networks, 1997., international conference on*, 1997, pp. 1930-1935.
[74] I. U. H. Akhtar, "Pakistan Needs a New Crop Forecasting System," 2014.
[75] D. Stathakis, I. Savin, and T. Nègre, "Neuro-fuzzy modeling for crop yield prediction," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 34, pp. p1-4, 2006.
[76] P. Kumar, D. K. Gupta, V. N. Mishra, and R. Prasad, "Comparison of support vector machine, artificial neural network, and spectral angle mapper algorithms for crop classification using LISS IV data," *International Journal of Remote Sensing*, vol. 36, pp. 1604-1617, 2015.
[77] N.-D. Hoang, A.-D. Pham, and M.-T. Cao, "A novel time series prediction approach based on a hybridization of least squares support vector regression and swarm intelligence," *Applied Computational Intelligence and Soft Computing*, vol. 2014, p. 15, 2014.
[78] N. Kayarvizhy, S. Kanmani, and R. Uthariaraj, "ANN models optimized using swarm intelligence algorithms," *WSEAS Transactions on Computers*, vol. 13, pp. 501-519, 2014.
[79] D. Kumar, R. K. Prasad, and S. Mathur, "Optimal design of an in-situ bioremediation system using support vector machine and particle swarm optimization," *Journal of contaminant hydrology*, vol. 151, pp. 105-116, 2013.
[80] S. K. Pal, C. Rai, and A. P. Singh, "Comparative study of firefly algorithm and particle swarm optimization for noisy non-linear optimization problems," *International Journal of Intelligent Systems and Applications*, vol. 4, p. 50, 2012.
[81] A. Sedki and D. Ouazar, "Hybrid Particle Swarm and Neural Network Approach for Streamflow Forecasting," *Mathematical Modelling of Natural Phenomena*, vol. 5, pp. 132-138, 2010.
[82] R. Taormina and K.-W. Chau, "Data-driven input variable selection for rainfall-runoff modeling using binary-coded particle swarm optimization and Extreme Learning Machines," *Journal of Hydrology*, 2015.
[83] R. Taormina, K.-W. Chau, and B. Sivakumar, "Neural network river forecasting through baseflow separation and binary-coded swarm optimization," *Journal of Hydrology*, vol. 529, pp. 1788-1797, 2015.
[84] B. Raheli, M. T. Aalami, A. El-Shafie, M. A. Ghorbani, and R. C. Deo, "Uncertainty assessment of the multilayer perceptron (MLP) neural network model with implementation of the novel hybrid MLP-FFA method for prediction of biochemical oxygen demand and dissolved oxygen: a case study of Langat River," *Environmental Earth Sciences*, vol. 76, p. 503, 2017.
[85] M. A. Ghorbani, C. Deo Ravinesh, Y. Zaher Mundher, K. Mahsa H, and M. Babak, "Pan Evaporation Prediction Using a Hybrid Multilayer Perceptron-Firefly Algorithm (MLP-FFA) Model: Case study in North Iran," *Theoretical and Applied Climatology*, vol. In Press., 2017.
[86] J. Gilles, "Empirical wavelet transform," *IEEE transactions on signal processing*, vol. 61, pp. 3999-4010, 2013.
[87] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, *et al.*, "The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 1998, pp. 903-995.
[88] K. Chau and C. Wu, "A hybrid model coupled with singular spectrum analysis for daily rainfall prediction," *Journal of Hydroinformatics*, vol. 12, pp. 458-473, 2010.
[89] T. Nguyen-Huy, R. C. Deo, D.-A. An-Vo, S. Mushtaq, and S. Khan, "Copula-statistical precipitation forecasting model in Australia's agro-ecological zones," *Agricultural Water Management*, vol. 191, pp. 153-172, 2017.

AUTHOR BIOGRAPHIES

MUMTAZ ALI received MSc and MPhil degrees from Quaid-i-Azam University Islamabad, Pakistan. He is currently pursuing the PhD degree within School of Agricultural, Computational and Environmental Sciences, University of Southern Queensland, Australia within the *Advanced Data Analytics*: Modelling & Simulation Research Group. His current research interests include statistical and machine learning, atmospheric and environmental modelling, agricultural modelling, recommender systems, decision support system, clustering algorithm, medical image segmentation and public health.

DR RAVINESH C DEO: Senior *M. IEEE* | *Aff. M. ASCE* | *M. Inst. P* | received PhD from The University of Adelaide. He worked as a Research Academic, Postdoctoral Fellow, Associate Lecturer, Lecturer and Senior Lecturer at The University of Queensland, University of the South Pacific, Griffith University, University of Sydney and University of Southern Queensland. He is affiliated as International Senior Scientist at the Chinese Academy of Sciences institute, and is a leading scientist in artificial intelligence, informatics, expert systems and heuristic neural networks. Dr Deo was awarded prestigious international awards to recognize outstanding achievements including: Advance Queensland United States Smithsonian Fellowship, Australia-China Young Scientist Exchange Program, Japan Society for Promotion of Science Fellowship, Chinese Academy of Science Presidential International Fellowship and Australian Endeavour Fellowship. Dr Deo is Associate Editor of *ASCE Journal of Hydrologic Engineering*, Editorial Member *Hydrology Research* and Guest Editor of Special Issues for *'Energies'*, and has received Pblication Awards, Head of Department Award & Teaching Excellence Award. He has professional membership of scientific bodies: IEEE, Australian Academy of Science affiliated Japan Society for Promotion of Science Alumni Association (Elected Executive), Elected Member Institute of Physics, Australian Mathematical Society and Australasian Association for Engineering Education, Australia Global Alumni, American Geophysical Union and Affiliate Member of American Society for Civil Engineers. He leads *Advanced Data Analytics*: Modelling & Simulation Research Group, and has supervised 9 Higher Degree by Research thesis with 4 PhD & Research by Masters, 5 Coursework Masters Dissertation and published more than 120 journal, 25 conference papers, 2 Edited Books and 6 book chapters.

ASSOCIATE PROFESSOR TEK MARASENI is a Natural Resources Management and Environmental Economics Expert, University of Southern Queensland, Australia. He has over 25 years work experience in the areas of forestry, carbon trading, governance, water, and greenhouse gas (GHG) emissions accounting/modelling research in different countries including Nepal, Thailand, Vietnam, Laos, PNG, China, Myanmar, Bhutan, India, Fiji and Australia. He completed a double BSc (Science and Forestry) in Nepal, and gained his MSc in NRM from Asian Institute of Technology, Thailand, His PhD (2006) on estimating optimal harvesting age of plantings and modelling carbon sequestration potential of different land use systems at the University of Southern Queensland (USQ), Australia. He has worked with the Ministry of Forests and Soil Conservation in Nepal (1992-2003), and is currently working as an Associate Professor with the USQ and a Visiting International Scientist with the Chinese Academy of Sciences, China. He has produced over 130 publications including four books in the last ten years. His publications have had an impact at several levels, from local and regional agricultural and forestry communities to global research networks. His research work has been recognized through national and international awards/fellowships including: (1) Research Excellence Award from USQ 2014; (2) "Senior International Scientist" award from the Chinese Academy of Sciences 2013; and (3) "Climate Change Professional Fellows" award from the US State Department 2011.

DR NATHAN DOWNS (BEng, BEd, Mphil, PhD) is a Senior Lecturer in Mathematics. Since commencing a postgraduate research career in 2000 and attaining a PhD in 2009 Dr Downs has worked in research and teaching positions at Griffith University, the University of Southern Queensland and James Cook University, publishing in the areas of photobiology and mathematical modeling. Current research interests include atmospheric studies, and assessing solar radiation exposure risks in different population groups for a variety of outdoor environments.

Chapter 9

Conclusion

9.1: Synthesis

This thesis has advanced the science of hydrological and agricultural crop prediction by constructing accurate and highly precision hybrid probabilistic and artificial intelligence models using computational intelligence techniques. These have been focused on rainfall and drought forecasting studies, as well as crop yield predictions within the three important agricultural-reliant provinces of Pakistan. Utilizing the new modelling approaches, rainfall was forecasted at a monthly forecast horizon whereas the crop yield was predicted at yearly (seasonal) forecast horizons. For drought forecasts, the models were evaluated at 12-, 6-, 3-month to 1-month forecast horizons to capture short and long-term drought prediction abilities of the developed models and also to realize near real-time forecasting especially for most agricultural activities that range from 1 to 12 monthly periods. In improving the hydrological forecasting tasks using rainfall, drought and crop yield objectives, the hybrid probabilistic and artificial intelligence models were designed with new methodological approaches. Given the complexity of drought phenomenon, a diverse range of data intelligence algorithms that were utilized to construct a set of hybrid models including the Markov Chain Monte Carlo based copula (MCMC-copula), online sequential extreme learning machine (OSELM), extreme learning machine (ELM), random forest (RF), Kernel ridge regression (KRR), ensemble based adaptive neuro fuzzy inference system (ensemble-ANFIS), mini-max probability machine regression (MPMR), M5 Tree, particle swarm optimization based ANFIS (PSO-ANFIS), multiple linear regression (MLR) and genetic programming (GP).

The following three important issues were addressed in this research:

1. The problem of full dependence structure between climatic inputs and crop yield can be captured through probabilities.
2. The problem of selection of non-redundant predictor inputs from sets of multivariate input in hydrological and crop yield forecasting and
3. The non-stationarity and non-linearity issue.

The Markov Chain Monte Carlo based copula modelling resolved the first issue. Input selection algorithms resolved the second issue of feature optimization, while the latter (third) issue was resolved by time-scale multi-resolution representation of the respective hydrological input time series.

In the first objective (Chapter 3), the MCMC based copula models were used to overcome the inter-dependency (i.e. probabilistic behavior) between the inputs and target. Then the Bat algorithm based on feature selection strategy screened the salient and most suitable MCMC-copula models. The selected MCMC-copula models were utilized in online sequential extreme learning machine (OSELM), extreme learning machine (ELM) and random forest (RF) models whilst designing the high precision rainfall forecasting model at monthly forecast horizons. Hybridization led to the formation of MCMC-Cop-Bat-OS-ELM model that outperformed the comparative MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF models.

Further, an ensemble modelling approach which enables the uncertainty between multi-models to be rationalized more efficiently, leading to a reduction in forecast error caused by stochasticity in drought behaviours, was developed and explored in Chapter 4 (Objective 2). Thus, an ensemble-ANFIS based uncertainty assessment modeling approach was developed and

explored for medium and long term (3-, 6-, 12-months) drought forecasting. The partial-auto correlation function (PACF) was utilized to determine the significant input lags for the development of the 10 member ensemble-ANFIS model. Applying 10-member simulations, the ensemble-ANFIS model was validated for its ability to forecast severity (S), duration (D) and intensity (I) of drought. The results are benchmarked with the M5 Model Tree (M5 tree) and Minimax Probability Machine Regression (MPMR) models. The ensemble-ANFIS model was found to have better performances in emulating the drought index (i.e. SPI) compared to the M5 tree and MPMR models.

Additionally, in Objective 2 (Chapter 5), a committee of modelling approach based on ELM (Comm-ELM) was designed and explored to forecast short term (i.e. monthly) drought using multiple meteorological inputs. Committee of modeling is a model combination technique, which is uncommon in climatological studies. In this study, the ELM-based committee was investigated and bench marked as PSO-ANFIS-based committee and MLR-based committee models. The Comm-ELM model was found to outperform the Comm-PSO-ANFIS and Comm-MLR model for short term drought forecasting.

Moreover, two self-adaptive techniques that do not require any basis function or pre-defined mother wavelet were utilized in Chapter 6 to further address the non-stationarity and non-linearity issues (Objective 2). The short, medium and long term drought forecasting was achieved by designing and employing a novel multivariate empirical mode decomposition (MEMD) approach. This technique was developed to permit the utilization of multiple predictor inputs in MEMD-based modelling approaches. A total of twelve predictor inputs were fed into MEMD to get signals (i.e. IMFs) followed by Simulated Annealing (SA) algorithm to select the best IMFs for KRR to develop the hybrid MEMD-SA-KRR model. The MEMD-SA-KRR

proved to be better in forecasting 1-, 2-, 6- and 12 month drought in comparison to the RF counterpart (MEMD-SA-RF) and the standalone KRR and RF models.

Devising new strategies to predict agricultural crop yield is important for agricultural and policy makers. Therefore in chapter 7 (Objective 3), cotton yield prediction with the Markov Chain Monte Carlo-based simulation model was integrated with the genetic programing algorithm using multiple meteorological data of rainfall, temperature and humidity. Several different types of GP-MCMC-copula models were developed, each with the well-known copula families (i.e., Gaussian, student t, and Clayton, Gumble Frank and Fischer-Hinzmann functions) to screen and utilize an optimal cotton yield forecast model for the present study region.

Last but not least, universal two-phase hybrid ACO-OSELM, ACO-ELM and ACO-RF models using feature based input selection and colony optimization (ACO) algorithm to predict wheat yield was constructed in Chapter 8 (Objective 3). The ACO algorithm is conditioned to search for the suitable, statistically relevant data sites for the model's training, and the corresponding testing sites by virtue of a feature selection strategy utilizing a total of 27 agricultural counties' datasets in the agro-ecological zones in Punjab province in Pakistan. The developed model can be explored as a decision-support tenet for crop yield estimation in regions where a statistically significant relationship with a historical agricultural crop is well-established.

The outcomes of this PhD thesis clearly showed improved performances of hybrid probabilistic and machine learning models developed with respect to standalone (i.e., individual) counterpart models. A further illustration of the research outcomes are as follows:

1. **Markov Chain Monte Carlo based Copula modelling**

   The MCMC based copula models served as an important tool to determine the underlying dependencies in forecasting rainfall and cotton yield prediction. Copulas are sets of powerful mathematical tools that have the ability to connect two or more time-independent variables deployed in Bayesian analysis that quantifies uncertainty with a probability distribution whereas the Markov Chain Monte Carlo (MCMC) simulation technique was adopted as an optimization technique to determine the copulas' parameters. Bayes' law attributes all modeling uncertainties to the parameters and estimates the posterior distribution of model parameters. Several different types of copulas were developed, each with the well-known copula families (i.e., Gaussian, student t, and Clayton, Gumble Frank and Fischer-Hinzmann functions).

2. **Feature selection strategies**

   a. The Bat algorithm played an important role in determination and ranking of inputs (i.e. best MCMC-copula models) for OSELM model enhanced performance in forecasting monthly rainfall. The key important feature of Bat algorithm is that the algorithm selects and determines the MCMC based copula models using the echolocation behavior of micro bats with velocity, frequency and location in the integration space. Further, the bats are then moved to update the rules in the search space to improve the best selecting model using random walks. Next, the best model is evaluated by conditional archiving to update the current model.

   b. The Simulate Annealing (SA) algorithm was a bio-inspired feature selection optimization technique used to determine suitable inputs for drought forecasting. Due to its non-deterministic adaptive nature, the SA algorithm starts random

selection of inputs to determine solution of cost function. Soon another random search for neighbouring inputs is started and if the cost value of this newly selected input is less than the previous one, then the search for other neighbouring inputs starts again until the optimum criteria is met.

c. The ant colony optimization (ACO) algorithm was another bi-inspired feature selection method based on swarm optimization. This algorithm was used to locate the minimum possible distance between the inputs (training sites) following the behavior of ant colonies. A parameter named "pheromone" is assigned to input sites to classify them against the test site at the beginning of the search. The ant uses the pheromone to calculate the probability of the selecting site for training against the testing site where the pheromone values change by traversing the training sites, and consequently, the probability is increased for the new ants to select the best training site.

Feature selection or input determination has been a critical process in the development of artificial intelligence models. This research has shown that appropriate feature selection was necessary in order to develop parsimonious and best performing predictive models applied in the agricultural and the drought sector.

3. Multivariate empirical mode decomposition (MEMD) method

The MEMD is a self-adaptive multi-resolution method; hence the number of IMFs and residual components (i.e., resolved frequencies) are contingent upon the embedded features within the data sets. The MEMD method improved the model performances with respect to the standalone models. The MEMD-SA-KRR outperformed the alternative

models at three sites applied for drought forecast and the performance was improved in medium and long-term drought forecasts.

The main advantage of the self-adaptive MRA tool, MEMD, integrated with SA and KRR was that the hybrid MEMD-SA-KRR model requires only trivial human interventions. This has provided the prospect of being embedded into advanced forecasting apps for portable devices such as tablets and mobile phones and to provide hydrological forecasts at local farm levels.

9.2: Novel contributions of the study

This PhD thesis has made novel contributions to science, particularly in the development of probabilistic and machine learning predictive models for hydrological forecasting and crop yield prediction. In addition to the development of hybridized probabilistic and artificial intelligence models, further novel methodological improvements are as follows:

1. **Hybrid probabilistic and machine learning models**

A major contribution of this study was the development of a new probabilistic based machine learning modelling approach. Generally, model combinations are lacking in hydrological, environmental and agricultural crop applications. The MCMC based copula models were integrated with Bat and OSELM, ELM and RF to construct MCMC-Cop-Bat-OS-ELM, MCMC-Cop-Bat-ELM and MCMC-Cop-Bat-RF for rainfall forecasting and genetic programming, based on the MCMC copula (i.e. GP-MCMC-Cop) for cotton yield prediction. This approach is first study undertaken in Pakistan. MCMC based copula models overcome the inter-dependency within the data while the Bat algorithm, OSELM, ELM and RF (in the case of rainfall forecasting) and GP (for cotton yield prediction) extract relevant features to improve forecasting accuracy.

2. **Ensemble based adaptive neuro-fuzzy inference system (ensemble-ANFIS)**

A novel ensemble-ANFIS based uncertainty assessment modelling approach for drought forecasting was another key novel contribution. The developed 10-member ensemble-ANFIS model was validated for its ability to forecast severity (S), duration (D) and intensity (I) of drought. This enabled uncertainty between multi-models to be rationalized more efficiently, leading to a reduction in forecast error caused by stochasticity in drought behaviours.

3. **Committee-based modelling approach**

A major contribution of this PhD research thesis is the design of a new committee-based modeling approach. Generally, model combinations are lacking in drought and other environmental applications. The ELM-based committee of models (Comm-ELM) was able to achieve a better performance as compared to the particle swarm optimization ANFIS based committee of models (Comm-PSO-ANFIS) and MLR-based committee of models (Comm-MLR). The Comm-ELM was able to further optimize and stabilize the forecasts, since ELM created suitable weights, rather than simply averaging out the sizes, in forecasting drought.

4. **Multivariate empirical mode decomposition**

Another novel contribution was the development of multivariate empirical mode decomposition (MEMD) modeling approach in this study. This approach is important since MEMD is able to be used as a multi-variable forecasting tool; previously, EMD, EEMD, CEEMD, ICEEMDAN etc. were only used as single-variable forecasting tools. The forecasting performance increased with the integration of the MEMD hybridization approach to forecast 1-, 3-, 6- and 12-month drought forecasting.

5. Feature selection based strategy using district level wheat yield

The development of a two-phase hybrid (AC-OSELM) model using feature based input selection ant colony optimization (ACO) algorithm and OSELM was undertaken to predict wheat yield. The ACO algorithm is conditioned to search for suitable, statistically relevant data sites and the corresponding testing sites for the model's training, by virtue of a feature selection strategy utilizing a total of 27 agricultural counties' datasets in the agro-ecological zones in Punjab province in Pakistan. The developed ACO-OSELM model can be explored as a decision-support tenet for crop yield estimation in regions where a statistically significant relationship with historical agricultural crop is well-established. Further, the modeling approach can be adopted in a region where meteorological data is unavailable.

6. Further contributions

- With the notion of real-time forecasting, the forecast horizon has been shortened gradually from yearly to seasonal (3-months) and monthly with evaluation of respective models at a shorter forecast horizon.
- An important finding is that the predictive performances of probabilistic and artificial intelligence models for hydrological and agricultural crops are highly data-sensitive and site dependent due to geographical influences.
- Important contributions have been the application of various machine learning models such as M5 Tree, RF, KRR, ANFIS, OSELM, ELM, MLR and GP for rainfall and drought forecasting, as well as crop yield prediction in the Pakistan agricultural hub.

The innovative approaches being explored showed promising outcomes and could provide the scientific tenets for integrated on-farm decision-support systems for hydrological and precision agricultural purposes.

9.3: Limitations of the current study and recommendations for future research

Despite the significant contributions to a PhD focused on research, this study had some limitations; suggestions for future research are as follows:

- The short length of crop yield data tends to affect the parameter estimation of the model and the evaluation of the forecast performance. To overcome this practical issue, a longer actual time series or simulated crop yield data from dynamic models would be preferred.
- Studies with improved complete empirical ensemble mode decomposition with adaptive noise (ICEEMDAN), empirical wavelet transform (EWT), variational mode decomposition (VMD) and advanced maximum overlap discrete wavelet transformation (MODWT) could also provide greater insight into the performance of these predictive models.
- Integration of add-on optimizer algorithms (e.g., firefly optimizer algorithm (FFA), or Quantum-Behaved Particle Swarm Optimization (Q-PSO)) could also be applied in these hydrological models.
- Incorporation of satellite-based data such as those from Giovanni and reanalysis or MODIS could also improve model performance.
- Since the standard statistical approaches tend to avoid the hurdle of model uncertainty that potentially leads to over-confident inferences and risky agricultural decisions,

Bayesian Model Averaging (BMA) is another data-intelligent tool to model uncertainties and can be used in ranking model performance.

- Alternative feature selection algorithms, iterative input selection (IIS), Neighborhood Component Analysis (NCA) based feature selection algorithm for regression (fsrnca), modified minimum redundancy maximum relevance (mMRMR) algorithm or joint mutual information maximization feature selection (JMIM) can be further explored.
- Dimensionality reduction techniques such as singular valued decomposition (SVD), linear discriminant analysis (LDA), principal component analysis (PCA) can be explored for inputs.

In conclusion, this PhD thesis has made novel contributions towards the practical problem of hydrological forecasting and agricultural crop prediction using hybridized probabilistic and machine learning techniques. The easy-to-implement, hybridized probabilistic and machine learning data-intelligent forecasting models used in this study have high computational efficiency and low latency. This could revolutionize hydrological and agricultural crop modelling and forecasting, concurrently serving as an important instrument for water resource management and agricultural management applications.

References

Aamir, E. and Hassan, I., 2018. Trend analysis in precipitation at individual and regional levels in Baluchistan, Pakistan, IOP Conference Series: Materials Science and Engineering. IOP Publishing, pp. 012042.

Adamowski, J. and Chan, H.F., 2011. A wavelet neural network conjunction model for groundwater level forecasting. Journal of Hydrology, 407(1-4): 28-40.

Adamowski, J., Fung Chan, H., Prasher, S.O., Ozga-Zielinski, B. and Sliusarieva, A., 2012. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. Water Resources Research, 48(1).

Alvanitopoulos, P.-F., Andreadis, I., Georgoulas, N., Zervakis, M. and Nikolaidis, N., 2014. Solar Radiation Time-Series Prediction Based on Empirical Mode Decomposition and Artificial Neural Networks, 10th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI). IFIP Advances in Information and Communication Technology, AICT-436. Springer, Rhodes, Greece, pp. 447-455.

Ayaz, M., Khalid, K. and Malik, K.M., 2015. An Analysis of Weather and Cotton Crop Development in Lower Sindh (Tandojam). Retrieved on January, 10(2015): 2007-2012.

Baker, L. and Ellison, D., 2008. The wisdom of crowds—ensembles and modules in environmental modelling. Geoderma, 147(1-2): 1-7.

Banuri, T., 1998. Pakistan: Environmental Impact of Cotton Production and Trade, International Institute for Sustainable Development.

Barredo, J.I., 2007. Major flood disasters in Europe: 1950–2005. Natural Hazards, 42(1): 125-148.

Barzegar, R., Moghaddam, A.A., Deo, R., Fijani, E. and Tziritis, E., 2018. Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms. Science of the Total Environment, 621: 697-712.

BAS, 2018. British Antarctic Survey

Bhalme, H.N. and Mooley, D.A., 1980. Large-scale droughts/floods and monsoon circulation. Monthly Weather Review, 108(8): 1197-1211.

BMA, 2018. Bureau of Meteorology, Australia

Bokhari, A., 2013. Wheat crisis in the making, Dawn newspaper, Pakistan.

Breiman, L., 1996. Bagging predictors. Machine learning, 24(2): 123-140.

Cai, W. and Cowan, T., 2008. Dynamics of late autumn rainfall reduction over southeastern Australia. Geophysical Research Letters, 35(9).

Çavdar, T., 2016. PSO tuned ANFIS equalizer based on fuzzy C-means clustering algorithm. AEU-International Journal of Electronics and Communications, 70(6): 799-807.

Deo, R.C., Byun, H.-R., Adamowski, J. and Begum, K., 2015. Application of effective drought index for quantification of meteorological drought events: a case study in Australia. Theoretical and Applied Climatology, DOI 10.1007/s00704-015-1706-5(http://link.springer.com/article/10.1007%2Fs00704-015-1706-5).

Deo, R.C. et al., 2009. Impact of historical land cover change on daily indices of climate extremes including droughts in eastern Australia. Geophysical Research Letters, 36(8).

Dijk, A.I. et al., 2013. The Millennium Drought in southeast Australia (2001–2009): Natural and human causes and implications for water resources, ecosystems, economy, and society. Water Resources Research, 49(2): 1040-1057.

Districts, C.a.a.p.b., 2008. Crops area and production by Districts 1981-2008, Food and Cash crops, Federal Bureau of Statistics (Economic wing), Islamabad, Pakistan.

Dorosh, P. and Salam, A., 2008. Wheat markets and price stabilisation in Pakistan: An analysis of policy options. The Pakistan Development Review: 71-87.

Draper, N. and Smith, H., 1981. Applied regression analysis, 709 pp. John Wiley, New York.

FAO, 2013. Review of the sector and grain storage issue.

Fischer, M.J. and Hinzmann, G., 2006. A new class of copulas with tail dependence and a generalized tail dependence estimator, Diskussionspapiere//Friedrich-Alexander-Universität Erlangen-Nürnberg, Lehrstuhl für Statistik und Ökonometrie.

Hejazi, M.I. and Cai, X., 2009. Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm. Advances in water resources, 32(4): 582-593.

Hicks, M.J. and Burton, M.L., 2010. Preliminary Damage Estimates for Pakistani Flood Events, 2010. Center for Business and Economic Research, Ball State University.

Hina Ali, H.A., Zahir Faridi, Hira Ali, 2013. Production and Forecasting Trends of Cotton in Pakistan: An Analytical View. Journal of Basic and Applied Scientific Research, 3(12): 5.

Hsu, C.-W., Chang, C.-C. and Lin, C.-J., 2003. A practical guide to support vector classification.

Huang, G.-B., Zhu, Q.-Y. and Siew, C.-K., 2006. Extreme learning machine: theory and applications. Neurocomputing, 70(1): 489-501.

IPCC, 2012. Summary for Policymakers: A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change. In: C.B. Field et al. (Editors), In: Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation Cambridge University Press, Cambridge, UK.

JAMSTEC, 2018. Japan Agency for Marine-Earth Science

Jang, J.-S., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE transactions on systems, man, and cybernetics, 23(3): 665-685.

JISAO, 2018. Joint Institute of the Study of the Atmosphere and Ocean.

Koza, J.R., 1992. Genetic programming: on the programming of computers by means of natural selection, 1. MIT press.

Kundzewicz, Z.W., Radziejewski, M. and Pinskwar, I., 2006. Precipitation extremes in the changing climate of Europe. Climate Research, 31(1): 51-58.

Langridge, R., Christian-Smith, J. and Lohse, K., 2006. Access and resilience: analyzing the construction of social resilience to the threat of water scarcity. Ecology and Society, 11(2).

Looney, D. and Mandic, D.P., 2009. Multiscale Image Fusion Using Complex Extensions of EMD. IEEE Transactions on Signal Processing, 57(4): 1626-1630.

Maier, H.R., Jain, A., Dandy, G.C. and Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environmental modelling & software, 25(8): 891-909.

Mansoor, H., 2010. Pakistan evacuates thousands in flooded south.

McAlpine, C. et al., 2007. Modeling the impact of historical land cover change on Australia's regional climate. Geophysical Research Letters, 34(22).

McKee, T.B., Doesken, N.J. and Kleist, J., 1993. The relationship of drought frequency and duration to time scales, Proceedings of the 8th Conference on Applied Climatology. American Meteorological Society Boston, MA, pp. 179-183.

Mitchell, T.M., 1997. Machine learning, ser. Computer Science Series. Singapure: McGraw-Hill Companies, Inc.

Montgomery, D.C., Peck, E.A. and Vining, G.G., 2012. Introduction to linear regression analysis, 821. John Wiley & Sons.

Mpelasoka, F., Hennessy, K., Jones, R. and Bates, B., 2008. Comparison of suitable drought indices for climate change impacts assessment over Australia towards resource management. Int. J. Climatol., 28(10): 1283-1292.

Nelsen, R.B., 2003. Properties and applications of copulas: A brief survey, Proceedings of the First Brazilian Conference on Statistical Modeling in Insurance and Finance,(Dhaene, J., Kolev, N., Morettin, PA (Eds.)), University Press USP: Sao Paulo, pp. 10-28.

News, D., 2010. Floods to hit economic growth: Finance Ministry, Dawn News.

Niaz, M.S., 2014. Wheat Policy—A Success or Failure, DAWN Newspaper.

Nicholls, N., 2004. The changing nature of Australian droughts. Climatic change, 63(3): 323-336.

NOAA, 2017. Billion-Dollar Weather and Climate Disasters: Table of Events, National Center for Environmental Information.

NWFP, 2010. Imperial Gazetteer of India,. Digital South Asia Library.

Odusola, A. and Abidoye, B., 2015. Effects of Temperature and Rainfall Shocks on Economic Growth in Africa.

Pachauri, R.K. et al., 2014. Climate change 2014: synthesis report. Contribution of Working Groups I, II and III to the fifth assessment report of the Intergovernmental Panel on Climate Change. IPCC.

PBS, 2017. Population of Punjab province, Pakistan Bureau of Statistics

PMD, 2016. Pakistan Meteorological Department, Pakistan.

Quinlan, J.R., 1992. Learning with continuous classes, 5th Australian joint conference on artificial intelligence. Singapore, pp. 343-348.

Rajesh, R. and Prakash, J.S., 2011. Extreme learning machines-a review and state-of-the-art. International journal of wisdom based computing, 1(1): 35-49.

Rehman, N. and Mandic, D.P., 2009. Multivariate empirical mode decomposition. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 466(2117): 1291-1302.

Report, 2013. Cold weather in upper areas & dry weather observed in almost all parts of the country.

Reporter, T.N.s.S., 2015. Cotton production reaches 14.838 million bales, The Dawn.

Riebsame, W.E., Changnon Jr, S.A. and Karl, T.R., 1991. Drought and natural resources management in the United States. Impacts and implications of the 1987-89 drought. Westview Press Inc.

Rostami, A., Hatampour, A., Amiri, M., Ghiasi-Freez, J. and Heidari, M., 2014. Developing a Committee Machine Model for Predicting Reservoir Porosity From Image Analysis of Thin Sections, 20th Formation Evaluation Symposium of Japan. Society of Petrophysicists and Well-Log Analysts.

Sajjad, S.A., 2017. Story of Pakistan's elite wheat, Published in The Express Tribune, THE EXPRESS TRIBUNE Pakistan.

Sarwar, U., 2014. Agriculture in Pakistan – An Overview.

Service, A.M.I., 2012. DISTRICT-WISE AREA OF WHEAT CROP. Directorate of Agriculture (Economics & Marketing) Punjab, Lahore.

Service, A.M.I., 2014. DISTRICT-WISE AREA OF WHEAT CROP. Directorate of Agriculture (Economics & Marketing) Punjab, Lahore Pakistan.

SST, 2018. National Climate Prediction Centre.

strategy, P., 2018. Accelerating Economic Growth and Improving Social Outcomes.

Strohmann, T. and Grudic, G.Z., 2003. A formulation for minimax probability machine regression, Advances in Neural Information Processing Systems, pp. 785-792.

Survey, E., 2012. Pakistan Economic Survey 2011–2012, Ministry of Finance of the Government of Pakistan.

Tarakzai, S., 2010. Pakistan battles economic pain of floods, Jakarta Globe.

Vörösmarty, C.J. et al., 2010. Global threats to human water security and river biodiversity. Nature, 467(7315): 555-561.

Wilhite, D.A., Hayes, M.J., Knutson, C. and Smith, K.H., 2000. Planning for drought: Moving from crisis to risk management1. Wiley Online Library.

Wittwer, G., Adams, P.D., Horridge, M. and Madden, J.R., 2002. Drought, regions and the Australian economy between 2001-02 and 2004-05. Australian Bulletin of Labour, 28(4): 231.

WMO, 2017. Rainfall extremes cause widespread socio-economic impacts, World Meteorological Organization.

Wu, Z., Huang, N.E., Wallace, J.M., Smoliak, B.V. and Chen, X., 2011. On the time-varying trend in global-mean surface temperature. Climate Dynamics, 37(3-4): 759-773.

Yaseen, Z.M., Sulaiman, S.O., Deo, R.C. and Chau, K.-W., 2018. An enhanced extreme learning machine model for river flow forecasting: state-of-the-art, practical applications in water resource engineering area and future research direction. J. Hydrol.

You, Y., Demmel, J., Hsieh, C.-J. and Vuduc, R., 2018. Accurate, Fast and Scalable Kernel Ridge Regression on Parallel and Distributed Systems. arXiv preprint arXiv:1805.00569.

Zhang, Y., Duchi, J. and Wainwright, M., 2013. Divide and conquer kernel ridge regression, Conference on Learning Theory, pp. 592-617.