UNIVERSITY OF SOUTHERN QUEENSLAND
AUSTRALIA

STREAMFLOW AND SOIL MOISTURE FORECASTING WITH HYBRID DATA INTELLIGENT MACHINE LEARNING APPROACHES: CASE STUDIES IN THE AUSTRALIAN MURRAY-DARLING BASIN

A thesis submitted by

Ramendra Prasad

Dip Ed (Mathematics and Science)

BSc GCEd (Mathematics and Physics)

MSc (Physics)

For the award of

Doctor of Philosophy

2018

Abstract

For a drought-prone agricultural nation such as Australia, hydro-meteorological imbalances and increasing demand for water resources are immensely constraining terrestrial water reservoirs and regional-scale agricultural productivity. Two important components of the terrestrial water reservoir *i.e.*, streamflow water level (SWL) and soil moisture (*SM*), are imperative both for agricultural and hydrological applications. Forecasted SWL and *SM* can enable prudent and sustainable decision-making for agriculture and water resources management. To feasibly emulate SWL and *SM*, machine learning data-intelligent models are a promising tool in today's rapidly advancing data science era. Yet, the naturally chaotic characteristics of hydro-meteorological variables that can exhibit non-linearity and non-stationarity behaviors within the model dataset, is a key challenge for non-tuned machine learning models. Another important issue that could confound model accuracy or applicability is the selection of relevant features to emulate SWL and *SM* since the use of too fewer inputs can lead to insufficient information to construct an accurate model while the use of an excessive number and redundant model inputs could obscure the performance of the simulation algorithm.

This research thesis focusses on the development of hybridized data-intelligent models in forecasting SWL and *SM* in the upper layer (surface to 0.2 m) and the lower layer (0.2–1.5 m depth) within the agricultural region of the Murray-Darling Basin, Australia. The SWL quantifies the availability of surface water resources, while, the upper layer *SM* (or the surface *SM*) is important for surface run-off, evaporation, and energy exchange at the Earth-Atmospheric interface. The lower layer (or the root zone) *SM* is essential for groundwater recharge purposes, plant uptake and transpiration. This research study is constructed upon four primary objectives designed for the forecasting of SWL and *SM* with subsequent robust evaluations by means of statistical metrics, in tandem with the diagnostic plots of observed and modeled datasets.

The first objective establishes the importance of feature selection (or optimization) in the forecasting of monthly SWL at three study sites within the Murray-Darling Basin. Artificial neural network (ANN) model optimized with iterative input selection (IIS) algorithm named IIS-ANN is developed whereby the

IIS algorithm achieves feature optimization. The IIS-ANN model outperforms the standalone models and a further hybridization is performed by integrating a non-decimated and advanced maximum overlap discrete wavelet transformation (MODWT) technique. The IIS selected inputs are transformed into wavelet sub-series via MODWT to unveil the embedded features leading to IIS-W-ANN model. The IIS-W-ANN outperforms the comparative IIS-W-M5 Model Tree, IIS-based and standalone models.

In the second objective, improved self-adaptive multi-resolution analysis (MRA) techniques, ensemble empirical mode decomposition (EEMD) and complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) are utilized to address the non-stationarity issues in forecasting monthly upper and lower layer soil moisture at seven sites. The *SM* time-series are decomposed using EEMD/CEEMDAN into respective intrinsic mode functions (IMFs) and residual components. Then the partial-auto correlation function based significant lags are utilized as inputs to the extreme learning machine (ELM) and random forest (RF) models. The hybrid EEMD-ELM yielded better results in comparison to the CEEMDAN-ELM, EEMD-RF, CEEMDAN-RF and the classical ELM and RF models.

Since *SM* is contingent upon many influential meteorological, hydrological and atmospheric parameters, for the third objective sixty predictor inputs are collated in forecasting upper and lower layer soil moisture at four sites. An ANN-based ensemble committee of models (ANN-CoM) is developed integrating a two-phase feature optimization via Neighborhood Component Analysis based feature selection algorithm for regression (*fsrnca*) and a basic ELM. The ANN-CoM shows better predictive performance in comparison to the standalone second order Volterra, M5 Model Tree, RF, and ELM models.

In the fourth objective, a new multivariate sequential EEMD based modelling is developed. The establishment of multivariate sequential EEMD is an advancement of the classical single input EEMD approach, achieving a further methodological improvement. This multivariate approach is developed to allow for the utilization of multiple inputs in forecasting *SM*. The multivariate sequential EEMD optimized with cross-correlation function and Boruta feature selection algorithm is integrated with the ELM model in emulating weekly *SM* at four sites. The resulting hybrid multivariate sequential EEMD-Boruta-ELM attained a better performance in comparison with the multivariate adaptive regression splines (MARS) counterpart (EEMD-Boruta-MARS) and standalone ELM and MARS models.

The research study ascertains the applicability of feature selection algorithms integrated with appropriate MRA for improved hydrological forecasting. Forecasting at shorter and near-real-time horizons (*i.e.*, weekly) would help reinforce scientific tenets in designing knowledge-based systems for precision agriculture and climate change adaptation policy formulations.

Certification of Thesis

This Thesis is the work of **Ramendra Prasad** except where otherwise acknowledged, with the majority of the authorship of the papers presented as a Thesis by Publication undertaken by the Student. The work is original and has not previously been submitted for any other award, except where acknowledged.

Principal Supervisor: **Dr Ravinesh C Deo**

Associate Supervisor: **Professor Yan Li**

Associate Supervisor: **Associate Professor Tek Maraseni**

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

Journal Publications and Statement of Author Contributions

The following provides details of the agreed share of contributions of the doctoral candidate and the co-authors in the publications presented in this research thesis:

Article 1: Chapter 3

Ramendra Prasad, Ravinesh C. Deo, Yan Li, Tek Maraseni, (2017). "Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray-Darling Basin region using IIS and MODWT algorithm" *Atmospheric Research*, vol. 197 (2017), pp. 42–63. (**Q1; Impact Factor: 3.85 and SNIP: 1.447; 89th percentile**).

The percentage contributions for this paper are RP 70%, RCD 20%, YL 5% and TM 5%.

Author	Tasks Performed
Ramendra Prasad (Candidate)	Establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript.
Ravinesh C. Deo (Principal Supervisor)	Supervised and assisted in scientific methodological development with important technical inputs, editing, and co-authorship of the manuscript.
Yan Li (Associate Supervisor)	Editing and proofreading of the manuscript.
Tek Maraseni (Associate Supervisor)	Proofreading of the manuscript.

Article 2: Chapter 4

Ramendra Prasad, Ravinesh C. Deo, Yan Li, Tek Maraseni, (2018) "Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition" *Geoderma,* vol. 330 (2018), pp. 136–161. (**Q1; Impact Factor: 4.21 and SNIP: 1.689; 93rd percentile**).

The percentage contributions for this paper are RP 70%, RCD 20%, YL 5% and TM 5%.

Author	Tasks Performed
Ramendra Prasad (Candidate)	Establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript.
Ravinesh C. Deo (Principal Supervisor)	Supervised and assisted in scientific methodological development, editing and co-authorship of the manuscript.
Yan Li (Associate Supervisor)	Proofreading and co-authorship of the manuscript.
Tek Maraseni (Associate Supervisor)	Proofreading and co-authorship of the manuscript.

Article 3: Chapter 5

Ramendra Prasad, Ravinesh C. Deo, Yan Li, Tek Maraseni, (2018) "Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors" *Soil & Tillage Research,* vol. 181 (2018), pp. 63–81. (**Q1; Impact Factor: 4.31 and SNIP: 1.946; 98th percentile**).

The percentage contributions for this paper are RP 70%, RCD 20%, YL 5% and TM 5%.

Author	Tasks Performed
Ramendra Prasad (Candidate)	Establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript.
Ravinesh C. Deo (Principal Supervisor)	Supervised and assisted in scientific methodological development, editing and co-authorship of the manuscript.
Yan Li (Associate Supervisor)	Editing and proofreading of the manuscript.

Tek Maraseni (Associate Supervisor)	Proofreading of the manuscript.

Article 4: Chapter 6 (Submitted manuscript)

Ramendra Prasad, Ravinesh C. Deo, Yan Li, Tek Maraseni, (2018). Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach.

Under review: *Catena*, **92nd percentile**.

The percentage contributions for this paper are RP 85%, RCD 10%, YL, and TM together 5%.

Author	Tasks Performed
Ramendra Prasad (Candidate)	Establishment of methodology, data analysis, preparation of tables and figures, compilation and writing of the manuscript.
Ravinesh C. Deo (Principal Supervisor)	Supervised and assisted in scientific methodological development, editing and co-authorship of the manuscript.
Yan Li (Associate Supervisor)	Proofreading and co-authorship of the manuscript.
Tek Maraseni (Associate Supervisor)	Proofreading and co-authorship of the manuscript.

Acknowledgements

Firstly, I would like to express my sincere gratitude and appreciation towards my Principal research supervisor, Dr. Ravinesh Deo. He continuously provided the much-needed support in the formulation of the project idea and kept me motivated throughout the project. His constant guidance and monitoring ensured the completion of this thesis.

In addition, I would also like to thank my associate supervisors, Professor Yan Li and Associate Professor Tek Maraseni in providing valuable advice and editorial support.

I would like to convey my appreciation to the whole supervisory team for contributing towards the preparation of high-quality journal articles. My appreciation also goes to the members of the "***Advanced Data Analytics*: Environmental Modelling & Simulation Research Group"** for any direct and indirect contributions.

I am also grateful to University of Southern Queensland (USQ) Office of Research and Graduate Studies for awarding the International Fees Research Scholarship (USQ-IFRS) and International Stipend Research Scholarship (USQ-ISRS) to pursue 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 New South Wales Department of Primary Industries-Office of Water, Scientific Information for Land Owners (SILO), the Australian Water Availability Project (AWAP) and Interim ERA European center for medium-range weather forecasting reanalysis (ECMWF reanalysis).

Next, I would like to pay my heartily gratitude to my parents Mr. Mahendra Prasad and Mrs. Sudha Rama Sharma for their blessings. I would also like to thank my siblings, Mr. Umendra Prasad, Mr. Parmendra Prasad, Ms. Keshni Devi, Ms. Riteshni Devi and Ms. Niteshni Devi for their encouragement. Also would like to acknowledge all my friends for their support and invaluable discussions throughout this project.

I would also like to express my gratitude towards my wife, Mrs. Sarita Prasad for her understanding and unconditional affection.

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3 (Article 1 – Published, *Atmospheric Research*, vol. 197 (2017), pp. 42–63)

Supplementary analysis and discussions

Chapter 4 (Article 2 – Published, *Geoderma,* vol. 330 (2018), pp. 136–161)

Supplementary analysis and discussions

Chapter 5 (Article 3 – Published, *Soil & Tillage Research,* vol. 181 (2018), pp. 63–81)

Supplementary analysis and discussions

Chapter 6 (Article 4 – Submitted, under review: *Catena*)

List of Tables

Chapter 2

Chapter 3 (Article 1 – Published, *Atmospheric Research*, vol. 197 (2017), pp. 42–63)

Supplementary analysis and discussions

Chapter 4 (Article 2 – Published, *Geoderma,* vol. 330 (2018), pp. 136–161)

Supplementary analysis and discussions

Chapter 5 (Article 3 – Published, *Soil & Tillage Research,* vol. 181 (2018), pp. 63–81)

Supplementary analysis and discussions

Chapter 6 (Article 4 – Submitted, under review: *Catena*)

Appendix

List of Acronyms

ANN	Artificial Neural Network
AWAP	Australian Water Availability Project
BF	Basis Functions
BOM	Bureau of Meteorology-Australia
CART	Classification And Regression Tree
CCF	Cross-Correlation Function
CEEMDAN	Complete Ensemble Empirical Mode Decomposition with Adaptive Noise
CSIRO	Commonwealth Scientific and Industrial Research Organization
DWT	Discrete Wavelet Transformation
E_{NS}	Nash–Sutcliffe Efficiency
ECMWF	European Center for Medium-range Weather Forecasting
EEMD	Ensemble Empirical Mode Decomposition
ELM	Extreme Learning Machine
EMD	Empirical Mode Decomposition
EMI	El Nino Modoki Index
EWT	Empirical Wavelet Transformation
FE	Forecasting Error
fsrnca	Feature Selection for Regression based on Neighborhood Component Analysis
GCV	Generalized Cross Validation
IIS	Iterative Input Selection
IMF	Intrinsic Mode Function
IOD	Indian Ocean Dipole
IPCC	International Panel for Climate Change
L	Legates-McCabe's Index
LQ	Lower Quartile
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MARS	Multivariate Adaptive Regression Splines

MDA	Mean Decrease Accuracy
MDB	Murray-Darling Basin
MLP	Multi-Layer Perceptron
MLR	Multi-Linear Regression
MODWT	Maximum Overlap Discrete Wavelet Transformation
MRA	Multi-Resolution Analysis
MSE	Mean Squared Error
MZSA	Maximum Z-score among Shadow Attributes
NCA	Neighbourhood Component Analysis
NSOR	Non-Stationary Oscillation Resampling
NSW	New South Wales
OLS	Orthogonal Least Squares
OOB	Out-of-Bag
PACF	Partial Auto-Correlation Function
PCA	Principal Component Analysis
PCN	Precipitation
PDO	Pacific Decadal Oscillation
PSO	Particle Swarm Optimization
QLD	Queensland
r	Pearson's Correlation coefficient
RF	Random Forest
RMSE	Root-Mean-Square-Error
RRMSE	Relative Root-Mean-Square Error
RVFL	Random Vector Functional-links
SD	Standard Deviation
SDR	Standard Deviation Reduction
SILO	Scientific Information for Land Owners
SLFN	Single Layer Feed-forward Neural network
SM	Soil Moisture
SM_{UL}	Upper Layer Soil Moisture
SM_{LL}	Lower Layer Soil Moisture
SSA	Singular Spectrum Analysis
SST	Sea Surface Temperatures

SVD	Singular Value Decomposition
SVM	Support Vector Machine
SWL	Streamflow Water Level
SWL_{obs}	observed Streamflow Water Level
SWL_{pred}	predicted Streamflow Water Level
UQ	Upper Quartile
VDM	Variational Mode Decomposition
WI	Willmott's Index of Agreement
WT	Wavelet Transform

Hybrid Models Notations

Chapter 3 (Published Article 1)

IIS-ANN Artificial neural network (ANN) optimized with iterative input selection algorithm.

IIS-M5 Tree M5 Model Tree optimized with iterative input selection algorithm.

IIS-W-ANN ANN optimized with iterative input selection algorithm and wavelet (MODWT) transformation.

IIS-W-M5 Tree M5 Model Tree optimized with iterative input selection algorithm and wavelet transformation (MODWT).

Chapter 4 (Published Article 2)

CEEMDAN-ELM Extreme learning machine (ELM) integrated with complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN).

CEEMDAN-RF Random forest (RF) integrated with CEEMDAN multi-resolution utility.

EEMD-ELM Extreme learning machine (ELM) integrated with ensemble empirical mode decomposition (EEMD).

EEMD-RF Random forest (RF) integrated with EEMD.

Chapter 5 (Published Article 3)

ANN-CoM Ensemble committee of model developed with ANN basis, where second-order Volterra, M5 Model Tree, random forest, and ELM served as the expert models of the ensemble.

Chapter 6 (Submitted Article 4)

EEMD-Boruta-ELM ELM optimized with multivariate sequential ensemble empirical mode decomposition (EEMD) and Boruta feature selection.

EEMD-Boruta-MARS multivariate adaptive regression splines (MARS) optimized with multivariate sequential EEMD and Boruta.

Chapter 1: Introduction

1.1 Background

The demand for water resources is ever-increasing with population growth, increased agricultural and industrial activities, the expansion of water-related sports and recreation. In addition, the changing weather patterns and climate due to anthropogenic factors further affect the distribution and accessibility of this valuable and limited resource. This growing demand and intermittent supply require farsighted and effective water resource management stratagems to avoid any probable catastrophes.

Principally, the terrestrial water reservoirs instead of the direct precipitation control the functioning of agricultural, hydrological, ecological and interrelated socio-economic systems (Loon and Laaha, 2015). In particular, two integral components of the terrestrial water reservoirs *i.e.*, streamflow water level (SWL) and soil moisture (*SM*) are imperative for water resources management and agriculture. The SWL is the accumulation of the surface runoff from a catchment or basin that serves as a storage and water-source for surface water usages. While the *SM* controls the interactions between the land and the atmosphere (Brocca *et al.*, 2010, Brocca *et al.*, 2008) and serves as an important driver of soil water retention, infiltration, evapotranspiration, groundwater recharge, and geophysical processes.

Prolonged precipitation deficits with a series of dry spell epochs cause meteorological drought events. This when aggravates into inadequate availability of surface and subsurface water resources causes hydrological drought and further leads to a decline in soil moisture (*SM*) causing crop failure (*i.e.*, agricultural drought). An escalation of such extreme event leads to a widespread socioeconomic drought (White *et al.*, 1999, Mishra and Singh, 2010, Wilhite and Glantz, 1985). Such events not only moderates the river and terrestrial ecology (IPCC, 2014) but also severely impacts many sectors of the society culminating into human health issues. Yet, prior to intensifying into a severe drought event, the terrestrial water storage (including SWL and *SM*) endures the immediate impacts of hydro-meteorological anomalies and anthropogenic changes.

The streamflow data have been widely used for hydrologic drought studies, while agricultural drought is largely contingent upon *SM* levels. Hence, the forecasted SWL and *SM* information are important aspects in properly managing this limited resource. Forecasted SWL and *SM* levels could assist in drought preparedness and design of early warning systems as well as gain an insight into the future availability of water resources. Precise and reliable future information on SWL levels would assist in constructing of prudent and timely procedures and techniques for optimal distribution and utilization of water for purposes like domestic, industrial, agricultural, hydro-electricity generation and recreation. In addition, the advanced or projected knowledge of the other important variable, *i.e.*, *SM* levels, at micro-scale would allow farmers and farm managers to make proactive sustainable decisions for efficient irrigation scheduling, grazing scheduling, 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 soil moisture and empowers precision agriculture.

Recent advances in computational capacity have allowed for application of the machine learning based predictive models in many areas. The predictive or data-driven models extract pertinent predictive features from historical data sets. Since forecasting is an important aspect of 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 applications as well. Therefore, new and advanced predictive models hybridized with feature optimization and multi-resolution analysis approaches are being explored in this study for SWL and *SM* forecasting within Australia's agricultural hub, the Murray-Darling Basin.

1.2 Statement of the problem

Australia, the driest inhabited continent on Earth with harsh environmental conditions (Ummenhofer *et al.*, 2009), is no exception to the extreme hydrological events. An increase in frequency and intensity of events such as longer-lasting and hotter drought and catastrophic floods has been noted since the 1950s (IPCC, 2014; Deo *et al.*, 2015). Hydrological anomalies comprising of frequent and long-lasting drought events are a common feature of Murray-Darling Basin (MDB) cluster (Deo *et al.*, 2009) where the present study is focussed. Historical droughts in the MDB region, Australia recorded were; the Federation (1895-1902), World War II (1937–1945), and the Millennium drought event (1997–2009) (Timbal *et al.*, 2015; Deo *et al.*, 2016; Ummenhofer *et al.*, 2009; CSIRO, 2012). Particularly, the most recent event (Millennium drought) resulted in the lowest volume of streamflow since 1783 (CSIRO, 2012; van Dijk *et al.*, 2013). With that, under the high emission scenario, the projected seasonal soil moisture changes for the year 2090 shows a significant decrease predominantly in winter and spring seasons while the annual-mean decrease of up to 10% for the MDB region has been estimated. (Timbal *et al.*, 2015). The scarcity of water resources in MDB, Australia continues to elevate and is exacerbated by the changing climate, rainfall variability and land-use changes (McAlpine *et al.*, 2009), which makes the management of water resources more difficult (Timbal *et al.*, 2015; Humphrey *et al.*, 2016). 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 hydrological variables such as streamflow water level (SWL) and soil moisture level (*SM*).

Robust predictive models with better accuracies could serve as suitable alternatives for forecasting SWL and *SM*. However, the foremost and critical issues of selecting the non-redundant (and most important) input data remains 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, streamflow water level and soil moisture 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 accuracy of data-driven models (Adamowski and Chan, 2011, Adamowski *et al.*, 2012). Multi-resolution analysis (MRA) which can perform a careful assessment of the input data in terms of the predictive features it may contain, can be applied to ameliorate this challenge in model development. Most generally, discrete wavelet transformation (DWT) has been used for this purpose. However, DWT can have two major disadvantages, i) the issue of decimation effect whereby half the wavelet coefficients are only used in subsequent transformation causing loss of information and ii) their dependence on the point of the commencement of wavelet transformation on input data (Rathinasamy *et al.*, 2014). Instead, a more refined and non-decimated version known as the maximum overlap discrete wavelet transformation (MODWT) algorithm, can overcome these challenges (Renaud *et al.*, 2002). MODWT has not been explored in forecasting of SWL, thus has been trialed in this study. In addition, we chose two alternative and improved white-noise-assisted data analysis methods called the ensemble empirical mode decomposition (EEMD) developed by Wu and Huang (2009) and the complete empirical ensemble mode decomposition with adaptive noise (CEEMDAN) (Torres *et al.*, 2011). These are the newer and improved versions of the original empirical mode decomposition (EMD) developed by Huang *et al.* (1998). The advantage of both EEMD and CEEMDAN methods is that, in overcoming the non-stationarity and non-linearity problem via decomposition of the original time series, they do not require the prescribing of frequency bands or imposing of any particular basis function, making EEMD and CEEMDAN completely self-adaptive.

Model combinations are also very uncommon in hydrological applications and have been overlooked in environmental applications (Baker and Ellison, 2008). In this research 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 (Hatampour, 2013, Barzegar *et al.*, 2017). In this study, a novel artificial neural network (ANN) based ensemble committee of models is developed and evaluated. After employing the individual expert models, ANN is used to further optimize and stabilize the forecasts.

Despite EEMD being a self-adaptive and advanced MRA utility, studies pertaining to forecasting only utilized single input variable. The practice is to use the lagged time series of the objective variable to forecast the future data (Beltran-Castro *et al.*, 2013, Jiao *et al.*, 2016, Ouyang *et al.*, 2016, Bai *et al.*, 2015, Basha *et al.*, 2015, Seo and Kim, 2016). Yet the environmental and hydrological variables are driven by many influencing parameters that may have been left out. To include many predictor inputs, a multivariate approach to EEMD is developed and evaluated that has not been undertaken previously.

Overall, this thesis intends to address issues of appropriate input selection, non-linearity and non-stationarity of the input data in forecasting the streamflow water level and soil moisture data within Murray-Darling Basin, Australia. In addition, a novel ensemble forecasting using committee based modelling is also explored with the multivariate sequential EEMD approach.

1.3 Objectives

The key aim of this research was to develop a set of high-precision hybrid data-intelligent model for hydrological purposes (streamflow water level and soil moisture level forecasts) within the Murray-Darling Basin in the state of New South Wales (NSW), Australia. Future knowledge of streamflow water level and soil moisture is important for water resource managers and farm managers alike, in strategic decision-making. The models are developed to forecast across medium forecast horizons (monthly) and converging to the short-term horizon (weekly). To achieve the key aim, the objectives of the study are:

1) To develop hybridized ANN and M5 Tree models using non-decimated wavelet multi-resolution utility, MODWT and iterative input selection (IIS) optimizer algorithms for forecasting streamflow water level at monthly forecast horizon. The preciseness of the hybrid models were validated with respect to their standalone counterparts. The article has been published in *Atmospheric Research* journal (Vol. 197, Pages 42-63).
2) Utilize two self-adaptive multi-resolution tools (EEMD and CEEMDAN) hybridized extreme learning machine (ELM) and random forest models to forecast monthly upper and lower layer soil moisture. The EEMD and CEEMDAN addressed the non-linearity within the stochastic hydrological inputs without the need for any predefined basis functions. The performance of EEMD and CEEMDAN based models were compared against each other

and with standalone models. This has been published in the journal *Geoderma* (Vol. 330, Pages 136-161).

3) Develop and explore a new committee of modelling approach for monthly upper and lower layer soil moisture forecasting. Committee of modelling is a model combination technique, which is uncommon in hydrological studies. In this study, the ANN-based committee was investigated and optimized with Neighborhood Component Analysis based feature selection algorithm for regression, *fsrnca* feature selection algorithm. This has been published in *Soil & Tillage Research* journal (Vol. 181, Pages 63-81).
4) Devise a new multivariate approach to EEMD modelling to allow for utilization of multiple predictor inputs. This new multivariate sequential EEMD modelling technique has been developed and evaluated to forecast near-real-time (weekly) soil moisture values with ELM and multivariate adaptive regression splines (MARS) models. Feature optimization was carried out with cross-correlation function (*CCF*) and random forest-based Boruta wrapper feature selection algorithm. The manuscript is under review in the journal *Catena*.

1.4 Thesis layout

The schematic illustrating an overview of the research is shown in Figure 1.1. It clearly outlines the linkages between the factors influencing the terrestrial water storage, hence the need for reliable and precise forecasting tool using the available resources. The thesis is organized into seven chapters as follows:

Chapter 1 presents the introductory background and the statement of problem pertaining to the research and presents the objectives of this study.

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 viewpoints 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* (DOI: 10.1016/j.atmosres.2017.06.014). It is devoted to the establishment of multi-resolution analysis, MODWT based ANN modelling approach for hydrological forecasting *i.e.*, SWL. It outlines the issues with traditional approaches, model development and outcomes with respect to comparative tree-based model (M5 Tree model). Chapter 3 addresses the first research objective of this study.

Chapter 4 This chapter is presented as a published article in the journal, *Geoderma* (DOI: https://doi.org/10.1016/j.geoderma.2018.05.035). This chapter describes the application of advanced MRA utilities, EEMD and CEEMDAN in ensemble modelling using fast and efficient ELM modelling approach for hydrological forecasting. Chapter 4 is in response to the second research objective of this study whereby monthly upper and lower layer *SM* is forecasted using newly developed EEMD-ELM and CEEMDAN-ELM models. It outlines the model development and the outcomes benchmarked against comparative random forest models (EEMD-RF and CEEMDAN-RF).

Chapter 5 This chapter is presented as a published journal article in the journal, *Soil & Tillage Research* (DOI: 10.1016/j.still.2018.03.021) and describes the application of an alternative ensemble committee of modelling approach for hydrological forecasting. The monthly upper and lower layer *SM* is forecasted using this innovative committee of models based on ANN in combination with four standalone expert models; 2nd order Volterra, M5 Model Tree, Random Forest, and ELM. It outlines the model development and performances of the ensemble committee with respect to standalone models. Chapter 5 captures the third research objective of this study.

Chapter 6 This chapter is presented as a submitted manuscript in the *Catena* journal. It presents the development of a novel multivariate sequential EEMD forecasting technique for hydrological forecasting. The upper layer *SM* is forecasted using multivariate sequential EEMD at a weekly forecast horizon with Boruta feature selection. It outlines the development of the novel multivariate sequential EEMD-Boruta-ELM model and its performances with respect to a comparative MARS model (EEMD-Boruta-MARS) and the standalone ELM and MARS models. Chapter 6 addresses the fourth research objective of this study.

Chapter 7 presents the synthesis of the study with concluding remarks, limitations, and recommendations for future works.

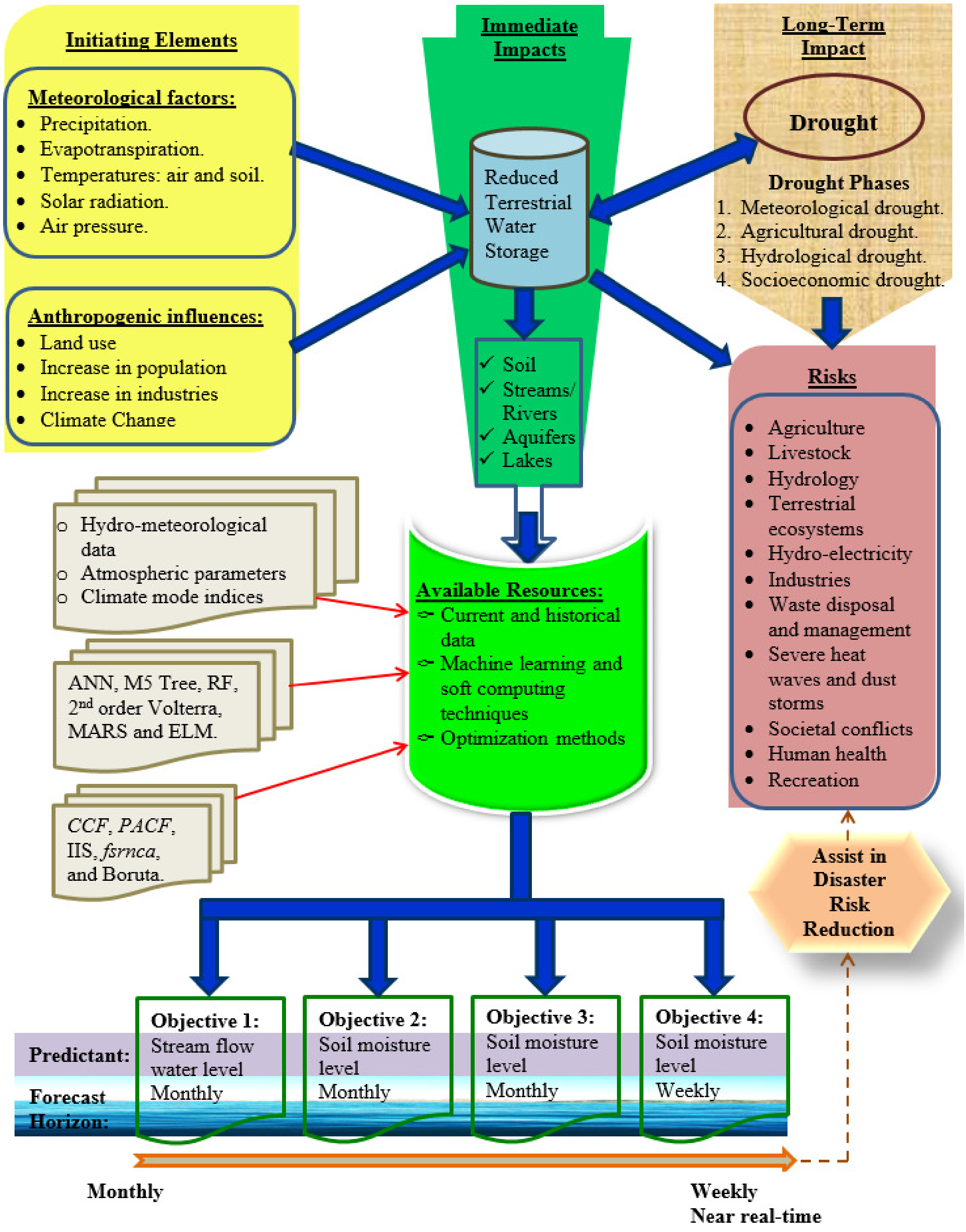


Figure 1.1 Schematic of the research thesis.

Chapter 2: Data and methodology

This chapter provides an overview of the location of the study sites in developing the hybrid data-intelligent hydrological forecasting 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-data driven model development.

2.1 Study area

In developing the hybrid hydrological forecasting models, the region of study is the Murray-Darling Basin (MDB) with an area of 1,042,730 km² (14% of mainland Australia) (The Murray–Darling Basin Authority, 2010; Bureau of Meteorology, 2018). The MDB is Australia's most important hydrological basin and is regarded as the agricultural hub of Australia. It encompasses 67% as agricultural land (Bureau of Meteorology, 2018; Australian Bureau of Statistics, 2010) where diverse agricultural activities account for 2% of the total economic output of Australia (Australian Bureau of Statistics, 2014) and contribute to 1/3 of Australia's food supply (Welsh *et al.*, 2013). Additionally, agriculture is the most important industry for rural and interior dwellers (Campbell and Scarlett, 2014). The study is specifically focussed in the state of New South Wales (NSW), located on the east coast of Australia as illustrated in Figure 2.1a (Shaded in green). The state of NSW accounted for ~23% of Australia's agricultural production by value in the financial year 2015-16 (Australian Bureau of Statistics, 2017). The key agricultural export commodities in NSW over the last 5 years were beef, vegetables, and fruit (NSW-Department of Industry, 2017). NSW accounted for ~ $1/4$ of Australia's wine exports by volume and 38% of Australian total sheep and lamb flock size in the last financial year (2015-2016) (Australian Bureau of Statistics, 2017) asserting that NSW is one of the most significant agricultural states in Australia.

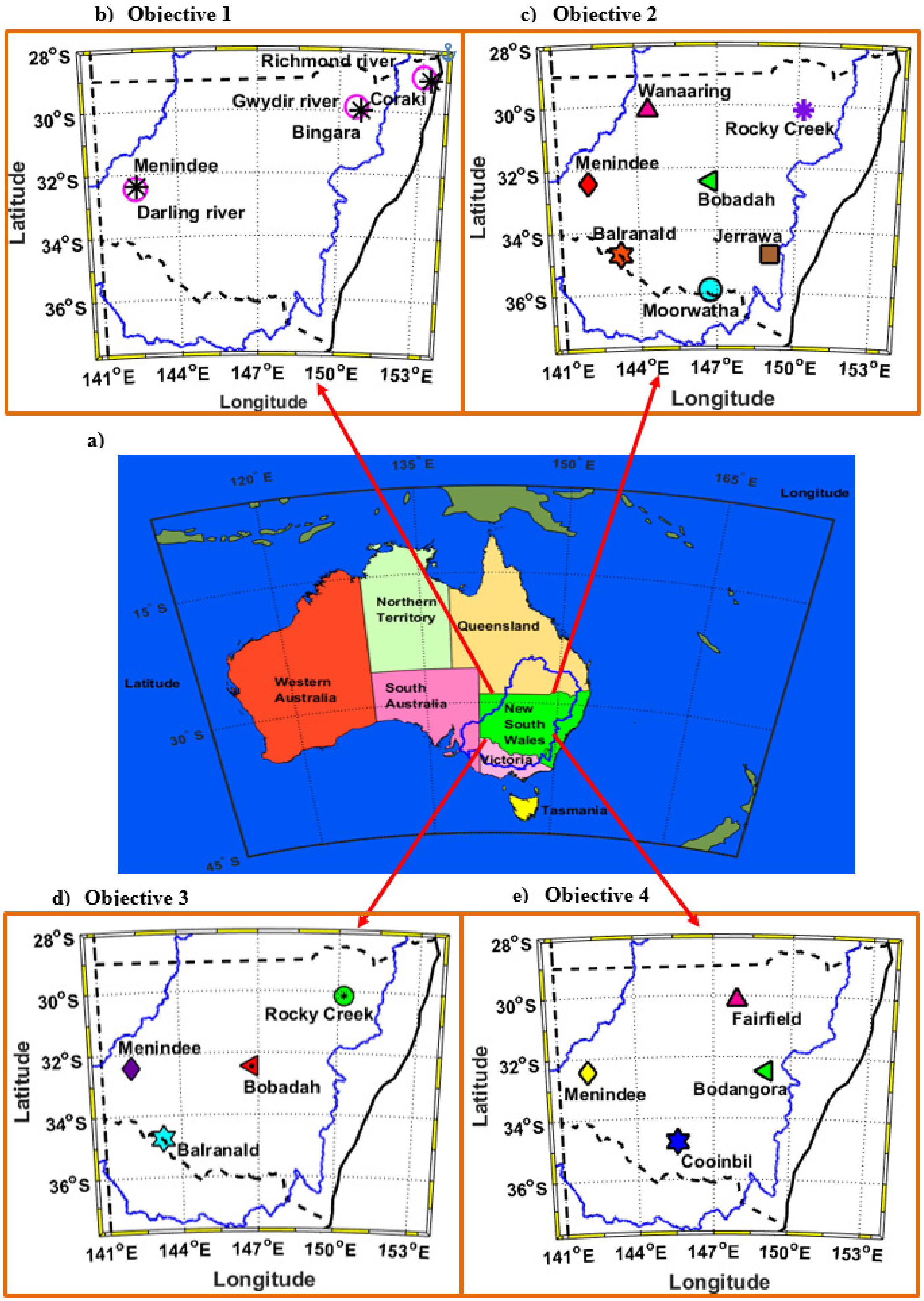


Figure 2.1 Map of the study region (a) the location of MDB and the state of NSW, with close-ups (b - e) illustrating the selected sites for objectives 1 to 4 respectively.

2.2 Data description

A variety of data sources was utilized in developing high precision data-intelligent hydrological models. 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 study.

Objective	**Data used**	**Source**	**Study period**	**Forecast horizon**	**Specific study area**
1 *(Chapter 3)*	**Predictors:** Meteorological variables	*Scientific Information for Land Owners (SILO)* (Jeffrey *et al.*, 2001; Tozer *et al.*, 2012; Beesley *et al.*, 2009)	January 1977 to May 2016	Monthly	Figure 2.1b
	Target: Streamflow water level	*NSW Department of Primary Industries* (NSW Department of Primary Industries-Office of Water, 2016)			
2 *(Chapter 4)*	**Predictors:** Antecedent upper and lower layer soil moisture	*Australian Water Availability Project (AWAP)* (Raupach *et al.*, 2012, 2009)	January 1990–December 2016	Monthly	Figure 2.1c
	Target: Upper and lower layer soil moisture				
3 *(Chapter 5)*	**Predictors:** Hydro-meteorological variables	*Australian Water Availability Project (AWAP)* (Raupach *et al.*, 2012, 2009)	January 1990–December 2016	Monthly	Figure 2.1d
	Atmospheric Parameters	*Interim ERA European center for medium-range weather forecasting reanalysis (ECMWF reanalysis)* (Dee *et al.*, 2011)			
	Synoptic scale climate mode indices	*Refer to Table 2.3*			
	Target: Upper and lower layer soil moisture	*Australian Water Availability Project (AWAP)* (Raupach *et al.*, 2012, 2009)			

4 *(Chapter 6)*	**Predictors:** Hydro-meteorological variables **Target:** Upper layer soil moisture	*Australian Water Availability Project (AWAP)* (Raupach *et al.*, 2012, 2009)	January 2007 to December 2016	Weekly	Figure 2.1e

Particularly, in the construction of the streamflow water level hybrid forecasting model (Objective 1), three hydrological sites (Richmond River, Gwydir River, and Darling River) located within the state of NSW were selected. Since the hydrological stations do not simultaneously observe the meteorological parameters or have only recently started the monitoring of rainfall, the most appropriate and nearest meteorological stations corresponding to the respective hydrological stations were selected from a list of weather stations in NSW from the Bureau of Meteorology (BOM), Australia data portal: http://www.bom.gov.au/climate/cdo/about/sitedata.shtml. Using GPS coordinates, the nearest direct distances between corresponding stations were computed. Figure 2.1b illustrates the corresponding hydrological and meteorological stations.

In constructing of monthly soil moisture forecasting models the relative soil moisture data for upper (SM_{UL}) and the lower layer (SM_{LL}) were sourced from Australian Water Availability Project (AWAP) (Raupach *et al.*, 2012, 2009). The upper layer soil moisture is up to a depth of 0.20 m from the surface and the lower layer is from 0.20 – 1.50 m depth and are characterized as surface *SM* and root-zone *SM*, respectively (Seneviratne *et al.*, 2010). The usual convention is to express *SM* as a dimensionless ratio of two masses or two volumes or as a ratio of a mass per unit volume (Petropoulos, 2014). However, the relative *SM* values derived by AWAP are in the range of 0 to 1 computed with respect to the base climatological reference period that is from the year 1961 to 1990 (Raupach *et al.*, 2009). Accordingly, the study period has been after 1990 for soil moisture forecasting studies.

For Objectives 2 and 3 the study period was January 1990–December 2016 at monthly *SM* forecasting horizon. While for Objective 4, the forecasting horizon was weekly and the period of study was January 2007 to December 2016 since AWAP commenced the weekly data generation in January 2007 (Raupach *et al.*, 2012, Raupach *et al.*, 2009). The final cut-off date was 01 January 2017 to account for final week overlap. The description of data from these sources is outlined as follows:

2.2.1 Streamflow water level (SWL) - NSW Department of Primary Industries

The monthly SWL data (in meters) were obtained from the NSW Department of Primary Industries (DPI) Office of Water data portal (http://realtimedata.water.nsw.gov.au/water.stm). The Department of Primary Industry – Office of Water is the organization responsible for monitoring of surface and groundwater resources and the development of water policy in NSW. It monitors the daily staff gauge readings, gas purge pressure and float well water level recording systems, electronic pressure sensors and telemetered digital logging systems at specific sites in river basins. The catchment area, mean annual rainfall and population statistics of the three selected river basins are shown in Table 2.2. It must be noted that Richmond river basin is a smaller one with high mean rainfall, while Darling river basin is large with relatively low mean annual rainfall.

Table 2.2 Catchment details of the river basins.

	Catchment area (km^2)	Mean annual rainfall (mm)	Population
Richmond river basin	6940	1525	115000
Gwydir river basin	25930	530	26000
Darling river basin	115880	300	30000

Source: (NSW Department of Primary Industries (DPI) Office of Water, 2018)

All data prepared by the DPI-Office of Water are quality coded during processing. Generally, the data were from direct gauging while some were adjusted during processing due to anomalies.

2.2.2 Meteorological data - Scientific Information for Land Owners

The meteorological data for neighboring hydrological sites were acquired from Scientific Information for Land Owners (SILO) database: https://www.longpaddock.qld.gov.au/silo/ppd/index.php developed by Queensland Department of Environment and Resource Management (Jeffrey *et al.*, 2001; Tozer *et al.*, 2012; Beesley *et al.*, 2009). The predictor inputs comprised of monthly precipitation (*PCN*; mm), maximum temperature (T_{max}; °C), minimum temperature (T_{min}; °C), evaporation (*Evap*; mm), solar radiation (*Rn*; MJ/m^2) and vapour pressure (*VP*; hPa). These SILO-based meteorological data were derived from Australian Bureau of Meteorology's observations and the missing values had been interpolated via statistical techniques (Beesley *et al.*, 2009, Zajaczkowski *et al.*, 2013, Tozer *et al.*, 2012). The SILO data generation can briefly be stated in three steps (Beesley *et al.*, 2009):

a) The monthly rainfall climatology parameters are interpolated via a thin plate smoothing spline.
b) Then ordinary kriging, which is a geo-statistical spatial estimation technique requiring error variance minimization, is applied to the normalized monthly data as above;
c) Finally, using the relative temporal distribution generated by ordinary kriging, the monthly values are disaggregated to the daily values within each month.

2.2.3 Soil moisture and hydro-meteorological data - Australian Water Availability Project (AWAP)

The AWAP-based soil moisture and hydro-meteorological data are derived at 0.05° × 0.05° spatial resolution from a "*WaterDyn*" physical model (Raupach *et al.*, 2012, Raupach *et al.*, 2009). The *WaterDyn* physical model was developed by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in collaboration with the Australian Bureau of Meteorology (BOM) as part of the Australian Water Availability Project (AWAP). This model simulates the soil's hydrological conditions via terrestrial water balance equation across Australian continent by incorporating meteorological forcing data (*i.e.*, solar radiation, precipitation, and minimum and maximum daily air temperature) coupled with continental parameter maps (*e.g.*, albedo, soil characteristics, seasonality of vegetation greenness) and generates magnitudes of soil moisture and several other hydrological parameters. The water balance equation calculates soil moisture as the sum of the water fluxes across the boundaries of the upper and the lower layer. For the upper layer soil moisture, the influx is precipitation while the out-fluxes include transpiration from this layer, soil evaporation, surface runoff, and leaching from upper to lower layer. While the influx into the lower layer is the leaching from upper to the lower layer and the out-fluxes are from deep drainage and transpiration from the lower layer.

Quality controlled daily meteorological fields for *WaterDyn* model are generated by BOM from its network of rain gauges (*i.e.*, up to approximately 7500 gauges, both open and closed) and weather stations while solar irradiance data is obtained using geostationary satellites (Raupach *et al.*, 2009, AWAP, 2016). The AWAP gridded data are created as follows (Beesley *et al.*, 2009, Tozer *et al.*, 2012, Jones *et al.*, 2009):

i. The observed data is decomposed into a monthly averages and the associated anomaly. Due to their weak associations with topography anomalies are used.
ii. Then, using the Barnes successive correction technique the anomalies are interpolated. While the three-dimensional smoothing splines are utilized to interpolate the monthly climatological averages.
iii. Finally, the gridded data sets are produced by multiplying the monthly climate average grids with the monthly anomaly grids.

An additional variance term is also added to allow for observational or measurement errors.

Limitation: The main limitation of the *WaterDyn* physical model is that it is accustomed to nowcasting, *i.e.*, it determines the soil moisture values at the instance when the meteorological input data are channeled. Essentially the *WaterDyn* model is hindcasting since the meteorological data are recorded beforehand and then the soil moisture values are quantified. For example, the soil moisture level for the present month can only be determined at the end of the month after the observation and the accumulation of all the essential meteorological parameters are complete. For key decision making, however, it is imperative to have advanced knowledge or forecasted information of soil moisture, particularly at the local scale (*e.g.*, at the farm level). The current restrictions of the *WaterDyn* physical model do not allow for forecasting, instead, the data-driven models could offer feasible local alternatives.

2.2.4 Atmospheric parameters - Interim ERA European center for medium-range weather forecasting (ECMWF) reanalysis

To achieve Objective 3, a total of 60 possible predictor inputs were collated of which 38 input series were atmospheric parameters acquired from ERA-Interim. A complete list of atmospheric parameters used is provided in Chapter 5. The ERA-Interim is a newer global atmospheric reanalysis model that commenced in 1989 and generates a large variety of global gridded data. It was developed by the European center for medium-range weather forecasting (ECMWF). The ERA-Interim produces 3-hourly surface parameters, (describing weather, ocean-wave and land-surface conditions), 6-hourly upper-air parameters covering the troposphere and stratosphere and vertical integrals of atmospheric fluxes (Dee *et al.*, 2011). The monthly averages for many of the parameters are also generated that were used in monthly forecasting study. Advancements in ERA-Interim in respect to its predecessor versions include assimilation of 12-hourly 4D-Var of the upper-air atmospheric state with a spectral resolution of the outer loop as T255 (~79 km), and two successive inner loops at T95 (~210 km) and T159 (~125 km) resolutions. In addition, the ERA-interim has the inclusion of automated bias correction scheme in satellite radiance observations. The introduction of wavelet-like weighting functions for background-error covariance and the utilization of rain-affected radiances rather than derived rain rates for rainfall assimilation are further enhancements (Dee *et al.*, 2011). Yet, uncertainties and biases in ERA-Interim are very difficult to quantify and a more robust approach with the inclusion of traditional observations are preferred. Hence, in this study, the ERA-interim reanalysis datasets were used in tandem with the observed hydro-meteorological data.

2.2.5 Synoptic scale climate indices – various sources

The monthly synoptic scale climate indices were sourced from various authentic and reliable databases as shown in Table 2.3 below. 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 the SSTs. As a result, the most recent version of SST, Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4) has been adopted in this study. The ERSST.v4 utilizes the up-to-date in-situ datasets with the precise ship and buoy bias adjustments throughout the entire analysis period that substantially improves its applicability (Huang *et al.*, 2015).

Table 2.3 Sources of synoptic scale climate indices.

Synoptic scale climate indices	***Acronym***	**Source**	**Website (URL)**
Sea Surface Temperature (SST) of NINO 1+2 region	NINO 1+2	Sea Surface Temperature (SST): Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4) - Climate Prediction Centre-NOAA (Huang *et al.*, 2015; Liu *et al.*, 2015; Henley *et al.*, 2015)	http://www.cpc.ncep.noaa.gov/data/indices/ersst4.nino.mth.81-10.ascii
SST of NINO3 region	NINO3		
SST of NINO4 region	NINO4		
SST of NINO3.4 region	NINO3.4		
Tripole Index for the Interdecadal Pacific Oscillation	TPI (IPO)		
Dipole Mode Index (Previously known as IOD) (Saji *et al.*, 1999; Abram *et al.*, 2008)	DMI	Japan Agency for Marine-Earth Science and Technology (JAMSTEC)	http://www.jamstec.go.jp/frcgc/research/d1/iod/DATA/emi.monthly.txt
El Nino Modoki Index (Taschetto and England, 2009; Ashok *et al.*, 2007)	EMI		
Pacific Decadal Oscillation (Newman *et al.*, 2016)	PDO	Joint Institute of the Study of Atmosphere and Ocean (JISAO)	http://research.jisao.washington.edu/pdo/PDO.latest
Southern Oscillation Index (Abawi *et al.*, 2000)	SOI	Bureau of Meteorology – Australia	ftp://ftp.bom.gov.au/anon/home/ncc/www/sco/soi/soiplaintext.html
Southern Annular Mode Index (Visbeck, 2009; Ho *et al.*, 2012)	SAM	Natural Environment Research Council (NERC)	http://www.nerc-bas.ac.uk/public/icd/gjma/newsam.1957.2007.txt

It must be noted that for Objective 2 (Chapter 4) monthly soil moisture was forecasted using only historic upper and lower *SM* time series, while for Objective 3 (in Chapter 5) a total of sixty potential predictor inputs were collated and then a robust feature selection was employed to determine the determine the best inputs. The initial predicament was a collation of the same set of input hydro-meteorological variables for all seven sites as in Chapter 4 since the notion was to have the same study sites. Yet, these historic data were not consistently available for all seven sites. While accounting for the limitations on the consistent availability of data, a total of sixty input variables were pooled for four study sites as presented in Chapter 5 (Figure 2.1 c-d).

2.3 General methodology

Prior to model development, data quality checking phase is necessary. A calendar averaging technique was applied to replace all missing data during this phase. The hydro-meteorological 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.

Therefore, all predictor inputs and the target were normalized to the range of zero and one using the following relation:

$$d_{norm} = \frac{d - d_{min}}{d_{max} - d_{min}} \tag{1}$$

where d_{norm} is the normalized value of the data d which are either inputs or the target values, d_{max} is the maximum value of d, and d_{min} is the minimum magnitude of d.

In order to get a robust modelling and forecasting approaches in emulating streamflow water level and soil moisture, a variety of forecasting models are considered for an evaluation of their preciseness. The models range from the well-known artificial neural network (ANN), M5 Model Tree, 2nd order Volterra, random forest (RF), and multivariate adaptive regression splines (MARS), to the more computationally efficient extreme learning machine (ELM) algorithms are adopted.

The broad classification of the modelling technique into different categories is illustrated in Figure 2.2.

The second order Volterra is a mathematical rule-based algorithm developed on the basis of Taylor series (Maheswaran and Khosa, 2012, 2015). 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, 2001; Liaw and Wiener, 2002). The MARS model is a linear regression model with basis functions for each spline (Friedman, 1991). In addition, the ANN and ELM models are neuronal algorithms (Deo and Şahin, 2015; Yang *et al.*, 2017; Huang, 2015). The main difference is that ANN model is a multiple layer perceptron type while the ELM model is a single layer feed-forward network (SLFN) type algorithm. ELM has added advantages including being fast and computationally efficient with a better universal approximation or generalization capability in many forecasting problems (Huang, 2015; Huang, *et al.*, 2015; Tang *et al.*, 2016; Mouatadid and Adamowski, 2016).

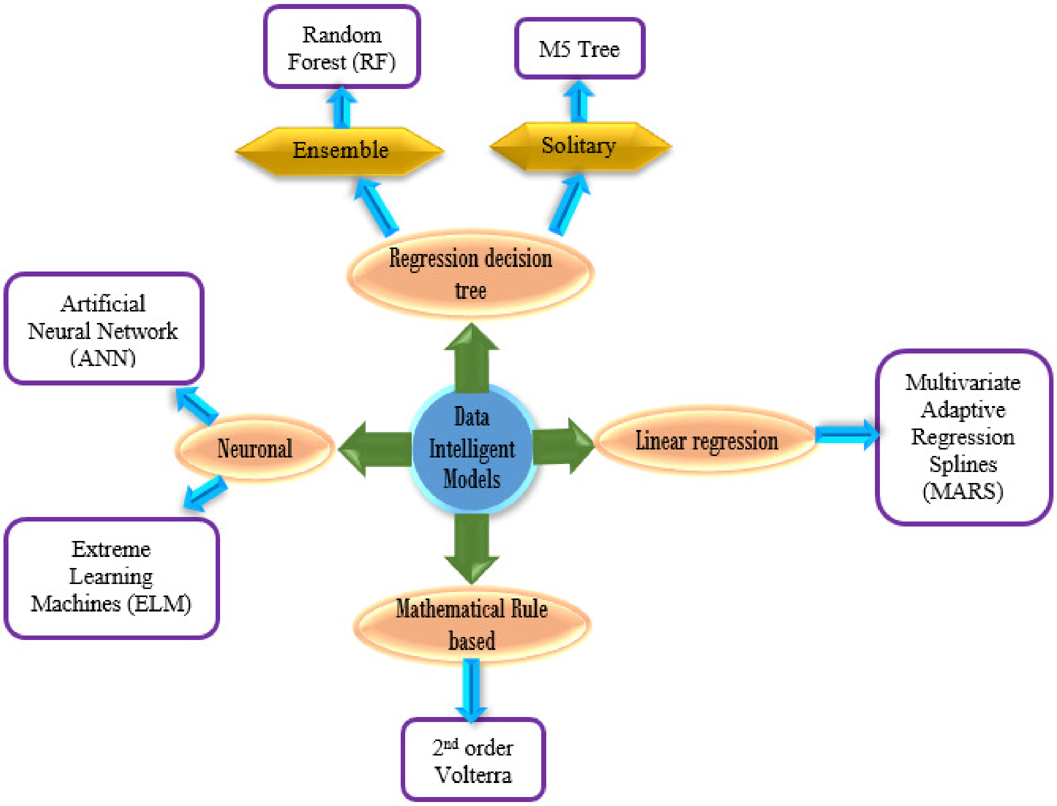


Figure 2.2 Categories of data intelligent models used in this research thesis.

For monthly SWL forecasting (Chapter 3), the ANN and M5 Model Tree models were utilized. In Chapter 4 (monthly *SM* forecasting with antecedent *SM* as inputs), ELM and random forest models were utilized. In Chapter 5 (monthly *SM* forecasting with multiple inputs), 2nd order Volterra, M5 Model Tree, random forest (RF) and ELM models were utilized with ANN based committee of models being developed. While in Chapter 6 (weekly *SM* forecasting with multiple inputs) the ELM and MARS models were adopted.

Data pre-processing via proper multi-resolution analysis tool is necessary for models to handle the non-stationarity features within the inputs (Adamowski *et al.*, 2012, Adamowski and Chan, 2011, Wang *et al.*, 2017, Deo *et al.*, 2017). Hence, hybridized models with advanced non-decimated wavelet multi-resolution utility (MODWT) (Chapter 3), two self-adaptive multi-resolution tools including ensemble empirical mode decomposition (EEMD) (Wu and Huang, 2009) and complete empirical ensemble mode decomposition with adaptive noise (CEEMDAN) (Torres *et al.*, 2011) are adopted (Chapter 4). In addition, new approaches are developed and explored including a committee of models (Chapter 5) and a multivariate sequential EEMD modelling approach (Chapter 6).

Appropriate input selection is imperative not only for input dimension reduction but also to improve the model performances. The optimization by means of feature selection approaches has its own advantages and disadvantages and therefore a number of algorithms were explored including the extra trees based iterative input selection (IIS) (Chapter 3), the partial-auto correlation function (*PACF*) (Chapter 4), Neighborhood Component Analysis (NCA) based feature selection algorithm for regression (*fsrnca*) (Chapter 5), and cross-correlation function (*CCF*) together with random forest driven Boruta wrapper-based algorithm (Chapter 6). In addition to the standalone approaches, the specific hybrid models developed in this study include:

1. IIS-W-ANN, IIS-W-M5 Tree, IIS-ANN and IIS-M5 Tree for monthly SWL forecasting. IIS was utilized for feature optimization with MODWT for addressing non-stationarity.
2. EEMD-ELM, EEMD-RF, CEEMDAN-ELM, CEEMDAN-RF for monthly SM_{UL} and SM_{LL} forecasting. *PACF* was utilized for determination of significant lags while EEMD and CEEMDAN were the MRA utilities addressing the non-stationarity in the input data.

3. ANN-CoM for monthly SM_{UL} and SM_{LL} forecasting with ELM, M5 Model Tree, RF and 2nd order Volterra as the underlying expert models. For feature optimization, two-phase feature selection approach via *fsrnca* followed by basic ELM were utilized.
4. Multivariate sequential EEMD based models to address non-stationarity within multiple predictor inputs in EEMD transformation was developed. Feature optimization was achieved using the *CCF* and Boruta input selection leading to hybridized multivariate sequential EEMD-Boruta-ELM and multivariate sequential EEMD-Boruta-MARS models.

For model evaluations, a diverse range of statistical metrics were used including the Pearson's correlation coefficient (*r*), root-mean-square-error (*RMSE*), mean absolute error (*MAE*), Willmott's Index (*WI*), Nash–Sutcliffe Efficiency (E_{NS}), and the Legates-McCabe's index (*L*). 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 percentage error (*MAPE*)) 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: Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray-Darling Basin region using IIS and MODWT algorithm

Foreword

This chapter is an exact copy of the published article in *Atmospheric Research* journal (Vol. 197, Pages 42-63).

It describes the hybridization of the widely used neuronal-based artificial neural network and a Cartesian and regression tree based M5 model tree for streamflow water level forecasting. Commonly, the discrete wavelet transformation (DWT) is applied in hydro-meteorological forecasting, however, the main drawback is the loss of information or the decimation effect. Hence, the classical models are hybridized with the advanced and non-decimated wavelet transformation, *i.e.*, maximum overlap discrete wavelet transformation (MODWT). The MODWT transformation is able to extract relevant information in time-frequency domain without any loss of information. In addition, an extra trees based iterative input selection (IIS) algorithm is utilized to further optimize the model performances.

The newly developed hybrid models IIS-W-ANN is evaluated against the comparative IIS-W-M5 Tree, the IIS-based models (IIS-ANN and IIS-M5 Tree) and the standalone ANN and M5 model tree in forecasting of monthly streamflow water level at three hydrological sites within drought-prone Murray-Darling Basin (MDB), Australia using meteorological data as predictor inputs.

Atmospheric Research 197 (2017) 42–63

Contents lists available at ScienceDirect

Atmospheric Research

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

Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm

Ramendra Prasad, Ravinesh C. Deo*, Yan Li, Tek Maraseni

School of Agricultural, Computational, and Environmental Sciences, Institute of Agriculture and Environment (IAg & E), University of Southern Queensland, Springfield, QLD 4300, Australia

ARTICLE INFO

Keywords:
Wavelet-hybrid model
Streamflow forecasting
Drought in Murray-Darling Basin
Iterative input selection algorithm, MODWT

ABSTRACT

Forecasting streamflow is vital for strategically planning, utilizing and redistributing water resources. In this paper, a wavelet-hybrid artificial neural network (ANN) model integrated with iterative input selection (IIS) algorithm (IIS-W-ANN) is evaluated for its statistical preciseness in forecasting monthly streamflow, and it is then benchmarked against M5 Tree model. To develop hybrid IIS-W-ANN model, a global predictor matrix is constructed for three local hydrological sites (Richmond, Gwydir, and Darling River) in Australia's agricultural (Murray-Darling) Basin. Model inputs comprised of statistically significant lagged combination of streamflow water level, are supplemented by meteorological data (i.e., precipitation, maximum and minimum temperature, mean solar radiation, vapor pressure and evaporation) as the potential model inputs. To establish robust forecasting models, iterative input selection (IIS) algorithm is applied to screen the best data from the predictor matrix and is integrated with the non-decimated maximum overlap discrete wavelet transform (MODWT) applied on the IIS-selected variables. This resolved the frequencies contained in predictor data while constructing a wavelet-hybrid (i.e., IIS-W-ANN and IIS-W-M5 Tree) model. Forecasting ability of IIS-W-ANN is evaluated via correlation coefficient (*r*), Willmott's Index (*WI*), Nash–Sutcliffe Efficiency (E_{NS}), root-mean-square-error (*RMSE*), and mean absolute error (*MAE*), including the percentage *RMSE* and *MAE*. While ANN models are seen to outperform M5 Tree executed for all hydrological sites, the IIS variable selector was efficient in determining the appropriate predictors, as stipulated by the better performance of the IIS coupled (ANN and M5 Tree) models relative to the models without IIS. When IIS-coupled models are integrated with MODWT, the wavelet-hybrid IIS-W-ANN and IIS-W-M5 Tree are seen to attain significantly accurate performance relative to their standalone counterparts. Importantly, IIS-W-ANN model accuracy outweighs IIS-ANN, as evidenced by a larger *r* and *WI* (by 7.5% and 3.8%, respectively) and a lower *RMSE* (by 21.3%). In comparison to the IIS-W-M5 Tree model, IIS-W-ANN model yielded larger values of $WI = 0.936–0.979$ and $E_{NS} = 0.770–0.920$. Correspondingly, the errors (*RMSE* and *MAE*) ranged from 0.162–0.487 m and 0.139–0.390 m, respectively, with relative errors, $RRMSE = (15.65–21.00)$ % and $MAPE = (14.79–20.78)$ %. Distinct geographic signature is evident where the most and least accurately forecasted streamflow data is attained for the Gwydir and Darling River, respectively. Conclusively, this study advocates the efficacy of iterative input selection, allowing the proper screening of model predictors, and subsequently, its integration with MODWT resulting in enhanced performance of the models applied in streamflow forecasting.

1. Introduction

Since the 1950s, increased frequency and intensity of extreme hydrological events (including longer-lasting and hotter drought and catastrophic floods) have been experienced in many parts of the World (Deo et al., 2015b; IPCC, 2014). This calamity has been coupled with significant hydrological imbalance over local, regional and continental scales that has impacted severely the agriculture, energy, recreation, domestic and industrial water supply and health sector. To develop robust disaster management strategies, including adaptive and mitigation measures for climate extreme, forecasting models for hydrological variables (e.g., streamflow water level) are extremely crucial. Forecasting models are beneficial for decision-makers and resource managers to construct a broader understanding of the future possibility of

* Corresponding author.
E-mail addresses: ramendra.prasad@usq.edu.au (R. Prasad), ravinesh.deo@usq.edu.au (R.C. Deo).

http://dx.doi.org/10.1016/j.atmosres.2017.06.014
Received 28 March 2017; Received in revised form 6 May 2017; Accepted 13 June 2017
Available online 20 June 2017

natural disasters (e.g. drought). Optimized forecasts are thus very important for facilitating prudent and strategic decisions by stakeholders in socio-economic sectors (Mehr et al., 2014; Mishra and Singh, 2011; Ni et al., 2010).

Numerical quantification of future streamflow dynamics can be undertaken by two categories of forecasting models: physically-based or dynamical models (e.g., Global Circulation Model) and statistically-based, or data driven models. Physically-based models are governed by the laws of physics where mathematical equations are applied to analyze the associations of conservation of mass, energy and momentum laws to the atmospheric and oceanic dynamics incorporated with relevant forcing (i.e., initial conditions) between several hydro-meteorological properties (CSIRO and Bureau of Meteorology, 2015). For instance, the Australian Bureau of Meteorology (BOM) utilizes Predictive Ocean Atmosphere Model for Australia (POAMA) based on a coupled ocean-atmosphere climate model (Cottrill et al., 2012; Zhao and Hendon, 2009) taking into account large-scale synoptic features like the progression of high and low pressure systems, large-scale oceanic currents and overturning in weather. It should be noted that, despite capturing the dynamics of physical processes at a broad range of spatio-temporal scales, physically-based models are generally reliant on very good quality and accurate input data that must be able to provide the radiative or other atmospheric forcing to execute the forecasting process. In forecasting streamflow, physical models would require data and mathematical relationships of additional determinants such as soil texture, watershed, and river network. Notwithstanding these, sophisticated programs are needed for implementation of differential equations, and consequently, they require rigorous optimization schemes compared to statistical or data-driven models (Abbot and Marohasy, 2012, 2014; Jain and Srinivasulu, 2004; Sehgal et al., 2014). On the other hand, due to the relative simplicity in their design and overall usage, data-driven models are able to be run by computational algorithms where the only requirement is historical data. Such models are gaining a lot of research attention in the hydrological simulation area (Dayal et al., 2016b; Deo and Şahin, 2015a; Deo et al., 2016a; Deo et al., 2017b; Deo and Sahin, 2016; Deo and Şahin, 2015b; Deo et al., 2016c; Humphrey et al., 2016; Rathinasamy et al., 2013; Taormina and Chau, 2015a; Yaseen et al., 2016a; Yaseen et al., 2016b). This approach has been popular in streamflow forecasting, particularly tailored to local hydrological forecasting, and so, it is attractive for decision-making in agriculture, crop management, irrigation, water pricing, allocation, and policy formulation.

Since data-driven models are viable tools for streamflow forecasting (Deo and Şahin, 2015a; Humphrey et al., 2016; Mehr et al., 2014; Onyari and Ilunga, 2013; Yaseen et al., 2016b), in this paper we apply an artificial neural network (ANN) and M5 Model Tree algorithm. Mimicking the neural structure of the brain, ANNs create a good approximation of unseen data through functional relationships between the past and future data. Due to the stochastic nature of the streamflow data, the use of the ANN model for prediction purposes is ideal as this model has the ability to capture complex and nonlinear relationship between predictors and the predictand (ASCE Task Committee on Application of ANN in Hydrology, 2000a; Xiong and O'Connor, 2002; Yilmaz et al., 2011). ANN models offer many advantages including: (1) the modelling is non-parametric so the predictor data used does not have to follow a Gaussian distribution; (2) the predictor data used may possess irregular seasonal variations but these can be analysed through ANN's non-linear modelling ability; (3) the ANN model is a nonlinear model and so, it performs well even when limited predictor data are available; (4) the ANN model is very robust and is able to deal with outliers and noisy input variables (Jain et al., 1999). Consequently, ANNs have been utilized in diverse hydrological catchments; for example, in Bangladesh (Liong et al., 2000; Liong and Sivapragasam, 2002); United Kingdom (Cameron et al., 2002); Vietnam (Phien and Kha, 2003); Italy (Alvisi et al., 2006; Campolo et al., 2003), Brazil (Pereira Filho and dos Santos, 2006) and Australia (Dayal et al., 2016a, b; Deo and Sahin, 2016; Deo et al., 2016c). However, an evaluation of ANN model against an alternative, M5 Model Tree (Deo et al., 2017a; Deo et al., 2017b) is also an interesting research endeavor since the latter (non-neural network) model is a hierarchical modular algorithm integrating classification and regression approaches. M5 Model Tree is built on assumptions that the dependency between predictors and predictand is approximated on smaller sub-domains encompassing a feature extraction platform (Solomatine and Xue, 2004). In the area of streamflow forecasting, M5 Model Tree model was used in India (Bhattacharya and Solomatine, 2003; Bhattacharya and Solomatine, 2005; Londhe and Dixit, 2012); Italy (Solomatine and Dulal, 2003; Solomatine and Siek, 2004); China (Solomatine and Xue, 2004); Nepal (Solomatine and Siek, 2004); Turkey (Sattari et al., 2013) and Africa (Onyari and Ilunga, 2013). However, to the best of the authors' knowledge, a comparison of ANN with M5 Tree in the present drought-prone region has not been undertaken.

Despite the ANN model's widespread use, pertinent issues of variable selection have not been addressed adequately with several works demonstrating the need to tackle this issue in model construction step (Abbot and Marohasy, 2012, 2014; Galelli and Castelletti, 2013b; Galelli et al., 2014; Hejazi and Cai, 2009; López et al., 2005; Quilty et al., 2016; Taormina and Chau, 2015b). To model a predictand from many exploratory variables, a high degree of uncertainty exists as to the best choice of the predictor candidates (George, 2000). Undoubtedly, irrelevant (or reductant) inputs are likely to worsen the model's underlying complexity (Hejazi and Cai, 2009; Maier et al., 2010) and trigger poor performance including ambiguity in model comparison (Maier and Dandy, 2000; Maier et al., 2010). By contrast, a set of carefully selected predictors is likely to ease the model's training process, increase the physical interpretability and provide a better understanding of the dynamics of the system that is modelled (Bowden et al., 2005). In spite of this, many studies that utilized ANNs and other data-driven models (e.g., (Alvisi et al., 2006; Chau, 2006, 2007; Deo and Şahin, 2016; Liong et al., 2000; Liong and Sivapragasam, 2002; Phien and Kha, 2003)) did not incorporate input selection algorithms with their primary model.

In literature, parsimonious and interpretable models have been constructed via input selection methods, which include, but are not limited to, cross-correlation and partial autocorrelations, '*hydrological expertise*' (Campolo et al., 2003; Deo et al., 2017a; Deo et al., 2017b; Deo and Sahin, 2017; Deo and Şahin, 2016; Deo et al., 2016c), average mutual information (Bhattacharya and Solomatine, 2005), bootstrap-ranked mutual information (Quilty et al., 2016) and minimum Redundancy Maximum Relevancy (Onyari and Ilunga, 2013). A recent study of Galelli and Castelletti (2013a) proposed an iterative input selection (IIS) as an alternative method, emerging as a novel ancillary data screening tool. Importantly, the IIS procedure can enable modelers to determine predictors from a global pool (via a tree-based algorithm), by acting as a safeguard for model's robustness against data redundancy and consequently the poor performance. In fact, a study on the evaluation of IIS procedure with partial mutual information, partial correlation, and Genetic Algorithm-ANN hailed the IIS as a better tool (Galelli et al., 2014). The results showed that IIS handled the persistence of collinearity amongst inputs and the potentially non-Gaussian, non-linear and interdependency factors. The veracity of IIS was demonstrated by Galelli and Castelletti (2013a) for a streamflow forecasting study in Ticino River (Switzerland). However, application of IIS in streamflow forecasting in the present study region is yet to be undertaken.

In spite of the skills of data-driven models to forecast a predictand by an analysis of non-linear patterns, the presence of non-stationarities in input data are likely to degrade a model's preciseness (Adamowski and Chan, 2011; Adamowski et al., 2012). This is because streamflow and the interrelated hydro-meteorological inputs are likely to exhibit complex temporal behavior with non-stationarity features (i.e., trends, seasonality, periodicity or jumps). To satisfy a search for a robust model, discrete wavelet transformation (DWT) has been adopted as a multi-resolution data pre-processing tool (Deo et al., 2016c). DWT itself exhibits challenges by virtue of the decimation effect induced in the wavelet coefficients. It generates half the wavelet coefficients of the detailed signal at the current level, while the other half of the smooth version are recursively processed by high pass and low pass filters, primarily at coarser resolution (Rathinasamy et al., 2014). The challenge, of course, is that the number of wavelet coefficients is halved with each shift of the analyzing mother wavelet. This issue is potentially addressed by *à trous* wavelet filter (Shensa, 1992). The alternative tool used in this study, non-decimated DWT (i.e., MODWT) involves non-decimation, that is able to retain the down sampled values at the various decomposition level (Cornish et al., 2005; Dghais and Ismail, 2013; Percival et al., 2011; Percival and Walden, 2000) while resolving the frequencies in the predictor (Steinbuch and Molengraft, 2005). Subsequently, the issue of decimation is addressed by MODWT where wavelet components at different timescales are of the same length (Maheswaran and Khosa, 2013; Rathinasamy et al., 2013; Rathinasamy et al., 2014). In spite of the use of MODWT in some studies (Maheswaran and Khosa, 2012; Rathinasamy et al., 2013), this method is yet to be tested for streamflow forecasting.

In this paper we develop an ANN model coupled with both, the iterative input selection (as a variable screening tool) integrated with non-decimated discrete wavelet transform (MODWT) algorithm (as a model's performance enhancement tool) in the drought-prone, Australia's agricultural (Murray-Darling) Basin where drought fosters economic repercussions (Helman, 2009). Previous studies have tested an ANN model for streamflow forecasting, such as those in Tallebudgera catchment (Gold Coast) (Fazel et al., 2014) and eastern Queensland (Deo and Sahin, 2016). Other applications were those in Ellen Brook River (Western Australia) where an ANN and wavelet-ANN was implemented (Badrzadeh et al., 2016), and in South Australia where an integrated hydrological (GR4J conceptual R-R model) and ANN was trialed (Humphrey et al., 2016).

The goal of our paper is to evaluate the utility of ANN and M5 Model Tree for streamflow forecasting in Australia's drought-prone (Murray-Darling) Basin. The novelty is to integrate iterative input selection (IIS) procedure with the non-decimated maximum overlap discrete wavelet transform (MODWT) algorithm using ANN as a primary model. Thus the main objectives of this paper are: (1) to utilize the IIS scheme, and enable a screening of the best set of predictors in order to attain a parsimonious, high-performance model, (2) to apply MODWT so as to construct a hybrid IIS-W-ANN model and compare its performance with a standalone ANN model, and (3) to evaluate the hybrid IIS-W-ANN model in respect to the IIS-W-M5 Model Tree.

2. Theory of machine learning algorithm

2.1. Artificial neural network

Since McCulloch and Pitts (1943) pioneered the idea of neural networks, utilization of ANN models in forecasting stochastic variables (e.g., streamflow, evaporation, drought, energy) has elevated (Deo and Sahin, 2016; Deo and Sahin, 2017; Deo and Şahin, 2015a, b; Deo et al., 2016c; Jain et al., 1999; Moustris et al., 2011). Generally, multilayer perceptron type ANN model comprises of three or more neuronal layers (Fig. 1a). Data is introduced through the input layer, while the output layer generating the forecasts corresponding to the features within the input(s), and one or more intermediate or hidden layers act as a platform for feature extraction. Nodes in the preceding and following layers are interconnected by weights such that the receiving node sums the weights from the preceding layer, add a bias and drives the result through a transfer function generating an output (Deo and Şahin, 2015a).

A typical ANN algorithm is written as (Deo and Şahin, 2015a, 2016; Kim and Valdés, 2003):

$$y(x) = F\left(\sum_{i=1}^{L} w_i\ (t).\ x_i(t) + b\right) \tag{1}$$

with $x_i(t)$ = predictor variable(s) in discrete time space t selected using input selection algorithm, $y\ (x)$ = forecasted streamflow in test set, L = hidden neurons determined iteratively, $w_j\ (t)$ = weight that connects the i^{th} neuron in the input layer, b = neuronal bias and $F(.)$ is the hidden transfer function.

ANN is a black-box and does not identify the training algorithm explicitly without a model identification process. Hence modelers trial several algorithms to attain an optimal model (Deo and Şahin, 2015a). Two MATLAB-based algorithms, used in this paper, are based on the quasi-Newton method (*trainlm* and *trainbfg*) (Huang, 1970). Both Levenberg-Marquardt (LM) (*trainlm*) and Broyden-Fletcher-Goldfarb-Shanno (BFGS) (*trainbfg*) (Dennis and Schnabel, 1996; Marquardt, 1963) minimize the mean square error (HariKumar et al., 2009). While LM is one of the fastest methods, BFGS uses Newton's method based on a hill-climbing optimization approach, seeking a stationary point (twice continuously differentiable) function. This algorithm has good performance for non-smooth optimizations (Avriel, 2003) where the Hessian matrix is not evaluated directly but instead it is approximated by rank-one updates specified by the gradient evaluation.

ANN models must be set appropriate architectures via hidden transfer and output layer functions. Normally, logarithmic sigmoid, tangent sigmoid and linear functions are applied (Deo and Sahin, 2016; Deo and Sahin, 2017; Deo and Şahin, 2015a; Vogl et al., 1988):

$$\begin{aligned}&\text{Tangent Sigmoid (tansig)} \Rightarrow F(x) = \frac{2}{1+\exp(-2x)} - 1\\&\text{Log Sigmoid (logsig)} \Rightarrow F(x) = \frac{1}{1+\exp(-x)}\\&\text{Positive Linear (purelin)} \Rightarrow F(x) = x\end{aligned} \tag{2}$$

During the learning phase, ANN is able to construct nonlinear relationships between inputs and output where the weights and biases minimize the objective function to yield the smallest mean squared error (*MSE*):

$$MSE = \frac{1}{N}\sum\nolimits_{p=1}^{N} (T_p - O_p)^2 \tag{3}$$

Note that T_p and O_p are, respectively, the targeted and output values for the p^{th} data point and N = total number of datum points.

As a safety measure, an early stopping criterion of the validation phase must be applied where *MSE* is monitored at each training and validation iteration and the training process is normally stopped with weights and biases derived before reaching the convergence. That can help prevent overfitting of the data (Rahimikhoob, 2014). Finally, the performance evaluation on the new (or unseen) data set is undertaken in the testing phase.

2.2. M5 model tree

In this paper, we evaluated ANN model's ability to forecast streamflow in respect to an M5 Model Tree. M5 Model Tree, pioneered by Quinlan (1992), is a hierarchical model with linear regression functions at the leaves to deal with continuous-class learning problems. Training is facilitated in two distinct phases: growing that commences with one node and recursively splits the input/output data into subsets/subspaces and a local specialized linear regression model is built within each subspace.

Fig. 1(b) plots the hierarchical display and also illustrates the corresponding splitting of the learning space. Training data (*T*) are split at nodes using ‘*divide-and-conquer*’ lemma where datum points are associated with a leaf or a test/split criterion to a split data set with a goal to minimize the intra-subset variation in the output variable's values down at each branch. The attribute that maximizes the expected error reduction/standard deviation reduction (*SDR*) is selected for splitting at the node. *SDR* is calculated as follows (Bhattacharya and Solomatine,

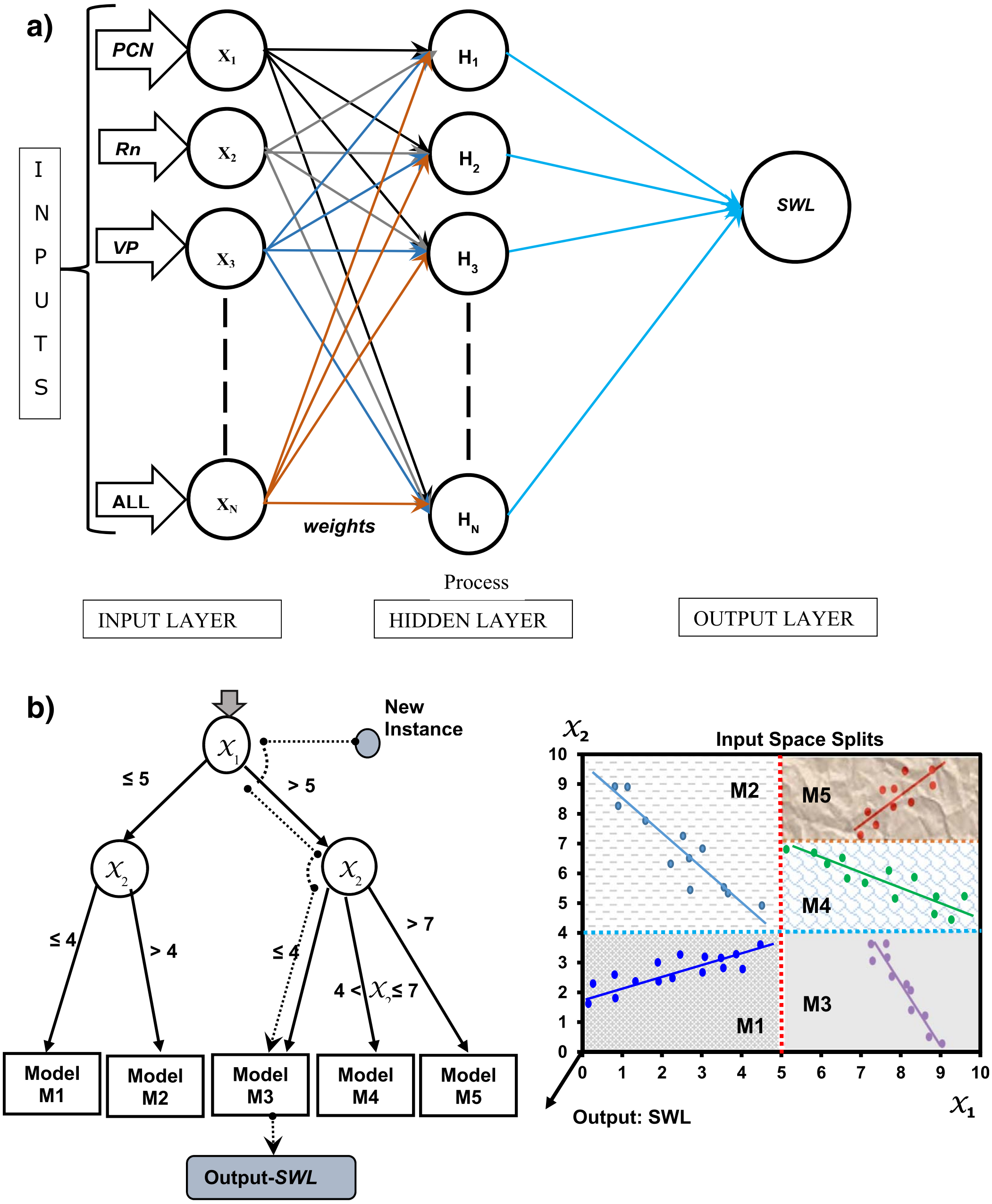


Fig. 1. Illustration of forecasting modelling frameworks. (a) Multilayer feedforward Neural Network Architecture (Details of input variables are provided in Table 1 with hidden neurons were successively increased from 1 to 40). b) The architecture of M5 Tree Model. (Hierarchical display and the corresponding input space splits with linear regression models for each subspace).

2005; Solomatine and Xue, 2004; Witten et al., 2011):

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times \mathrm{sd}(\mathrm{T_i}) \tag{4}$$

Here, T_1, T_2 … are the data sets that result from splitting the node, and sd (T) is the standard deviation of the class value. In the model optimization phase, the splitting process will cease when the class value of the instance reaching the node varies slightly which in this study, was 5%, or only a few splitting cases remained.

During the pruning phase, as long as the expected estimated error decreases, M5 Model Tree is pruned to prevent overfitting of data. Smoothing is then carried out to eliminate the sharp discontinuities resulting from a combination of multiple piece-wise linear regression functions, which in turn improves the forecasting accuracy (Wang and Witten, 1997; Witten et al., 2011).

2.3. Maximum overlap discrete wavelet transform

The premise of integrating an ANN with discrete wavelet transform is to generate a set of detailed frequencies in the predictor data (i.e., localized, transient, abrupt or stochastic phenomena) at various scales. Although discrete wavelet transform (DWT) is popular in wavelet-conjunction models, it has drawbacks such as the decimation effect that potentially forces loss of information in model training process and a consequent induction of bias in the forecasts (Rathinasamy et al., 2013). Ambiguity may also arise for the exact point of commencement of wavelet decomposition where the analyzing wavelet function is to be applied (Percival and Walden, 2000).

In this paper, MODWT (rather than conventional DWT), as proposed by Percival and Walden (2000), is adopted to address the decimation effect, allowing the same number of wavelets and scaling coefficients as

Table 1
Geographical and climate statistics of hydrological and meteorological stations utilized in this study.

Site	Name	Type	Location			Direct distance between Hyd. & Met. station (km)	Primary variables	Acronym	Annual climatic statistics (1977–2016)				
			Longitude (°E)	Latitude (°S)	Elevation (m)				Min.	Max.	Mean	Skewness	Kurtosis
1	Richmond River (Casino)	Hyd.	153.06	28.86	30.60	26.44	Streamflow water level	*SWL* (m)	0.72	1.18	0.95	0.05	1.43
	Coraki (Union st)	Met.	153.29	28.99	6.00		Precipitation	*P* (mm)	42.56	169.59	106.91	− 0.14	1.69
							Max. temp	*Tmax* (°C)	20.04	29.13	25.12	− 0.35	1.70
							Min. temp	*Tmin*°C	8.39	19.73	14.50	− 0.16	1.57
							Evaporation	*Evap* (mm)	69.38	184.82	129.83	− 0.15	1.66
							Solar radiation	*Rn* (MJm^{-2})	11.35	22.44	17.73	− 0.34	1.63
							Vapour pressure	*VP* (hPa)	11.58	23.31	17.49	− 0.01	1.55
2	Gwydir River (Pinegrove)	Hyd.	150.63	29.89	324.02	18.22		*SWL*	0.81	1.56	1.05	0.86	2.64
	Bingara (Keera)	Met.	150.78	29.99	333.00			*P*	35.37	93.68	58.84	0.57	1.82
								Tmax	16.89	32.23	25.16	− 0.21	1.59
								Tmin	1.77	17.30	9.68	− 0.05	1.50
								Evap	59.06	236.11	148.60	− 0.07	1.58
								Rn	11.01	24.97	18.92	− 0.33	1.61
								VP	8.62	18.23	13.26	0.10	1.60
3	Darling River (Menindee)	Hyd.	142.38	32.44	51.14	5.80		*SWL*	1.77	2.44	2.11	− 0.3	1.93
	Menindee Post Office	Met.	142.42	32.39	61.00			*P*	18.12	36.44	22.47	2.11	6.93
								Tmax	17.04	35.02	26.27	− 0.08	1.57
								Tmin	4.66	19.37	11.94	0.05	1.55
								Evap	60.53	349.82	194.30	0.10	1.58
								Rn	9.95	27.32	19.39	− 0.19	1.55
								VP	9.26	16.25	12.40	0.28	1.85

the observations at every level of the transform. In MODWT, the commencement point is not likely to influence the outcome of the decomposed data and the MODWT method is non-orthonormal, redundant and can be applied to all sample sizes (Cornish et al., 2005; Dghais and Ismail, 2013; Percival et al., 2011; Percival and Walden, 2000).

Basically, MODWT has the ability to yield wavelet coefficients with high and low pass filters applied to the input data. In matrix notation form, the transformation of an input time series vector denoted as $\widetilde{\mathbf{X}}$ having N number of samples is given by (Percival and Walden, 2000):

$$\widetilde{\mathbf{M}} = \mathrm{T}\widetilde{\mathbf{X}} \tag{5}$$

where T represents transformation matrix of dimension $(J_0 + 1)N \times N$, $\widetilde{\mathbf{M}}$ is a vector containing the MODWT wavelet and scaling coefficients with $(J_0 + 1)N$ dimension, where J_0 is the decomposition level of wavelet coefficients (Percival and Guttorp, 1994). Multi-resolution analysis or wavelet-based additive decomposition is obtained via a synthesis equation written as (Percival et al., 2011):

$$\widetilde{\mathbf{X}} = \sum\nolimits_{j=1}^{J_0} \widetilde{\boldsymbol{D}}_{\boldsymbol{j}} + \widetilde{\mathbf{S}}_{J_0} \quad \text{for } J_0 \geq 1 \tag{6}$$

where $\widetilde{\boldsymbol{D}}_{\boldsymbol{j}}$ and $\widetilde{\mathbf{S}}_{J_0}$are N-dimensional vectors having j^{th} level detail and $J_0{}^{\text{th}}$ approximation, respectively. For more details on the MODWT algorithm, readers can refer to the early works of Percival and Walden (2000) and also other recent papers (Maheswaran and Khosa, 2013; Rathinasamy et al., 2013; Rathinasamy et al., 2014) for usage in hydrology.

2.4. Iterative Input Selection (IIS)

IIS algorithm, adopted as input selection tool, executes in three phases (Galelli and Castelletti, 2013a). During the input ranking (IR) phase, the most significant inputs are selected in a forward selection method, however, the contribution of each input in explaining the output could be concealed due to the presence of several possibly redundant variables. The second phase is then applied to the predictor data, by adopting a Single Input Single Output (SISO) approach in order to overcome this potential issue where each of the first p-ranked variables is assessed independently through the identification of respective p-SISO models. The best-performing ones are selected (set p'). The third process in the IIS procedure is a Multiple Input Single Output (MISO) phase, which is applied to minimize the overfitting of the data where the prescribed screening model aims to assess the effectiveness of each input matrix in forecasting the output data. This successively adds the most significant ones from p', based on the coefficient of determination (R^2). To improve feature selections, the IIS algorithm executes a k-fold cross-validation and searches for optimal features when the R^2 value of the MISO model starts to either decrease or exhibit no significant improvement, leading to a termination of the algorithm. In this process, the IR and the model building algorithms are based on extremely randomized trees, as proposed by Geurts et al. (2006). In this paper, the MATLAB script for the IIS was used to screen the best predictors were from http://ivs4em.deib.polimi.it/. For more details on the IIS method, readers can consult the work of Galelli and Castelletti (2013a).

3. Materials and methodology

3.1. Study area and hydrological data

The present study area is located in New South Wales (NSW) within Australia's primary agricultural hub (i.e., Murray-Darling Basin). To construct the forecasting models, monthly streamflow water level (SWL) data were obtained from NSW Department of Primary Industries Data Portal (http://realtimedata.water.nsw.gov.au/water.stm), for hydrological sites located at Richmond, Gwydir and Darling River. To construct a large set of predictor matrix, supplementary predictor data for neighboring meteorological sites were also acquired from the Scientific Information for Land Owners (SILO) Portal developed by Queensland Department of Environment and Resource Management (Jeffrey et al., 2001). This data comprised of monthly precipitation, maximum and minimum temperature, evaporation, mean solar radiation and vapor pressure for the period January 1977 to May 2016. Table 1 shows a summary of sites and Fig. 2 plots a geographic map.

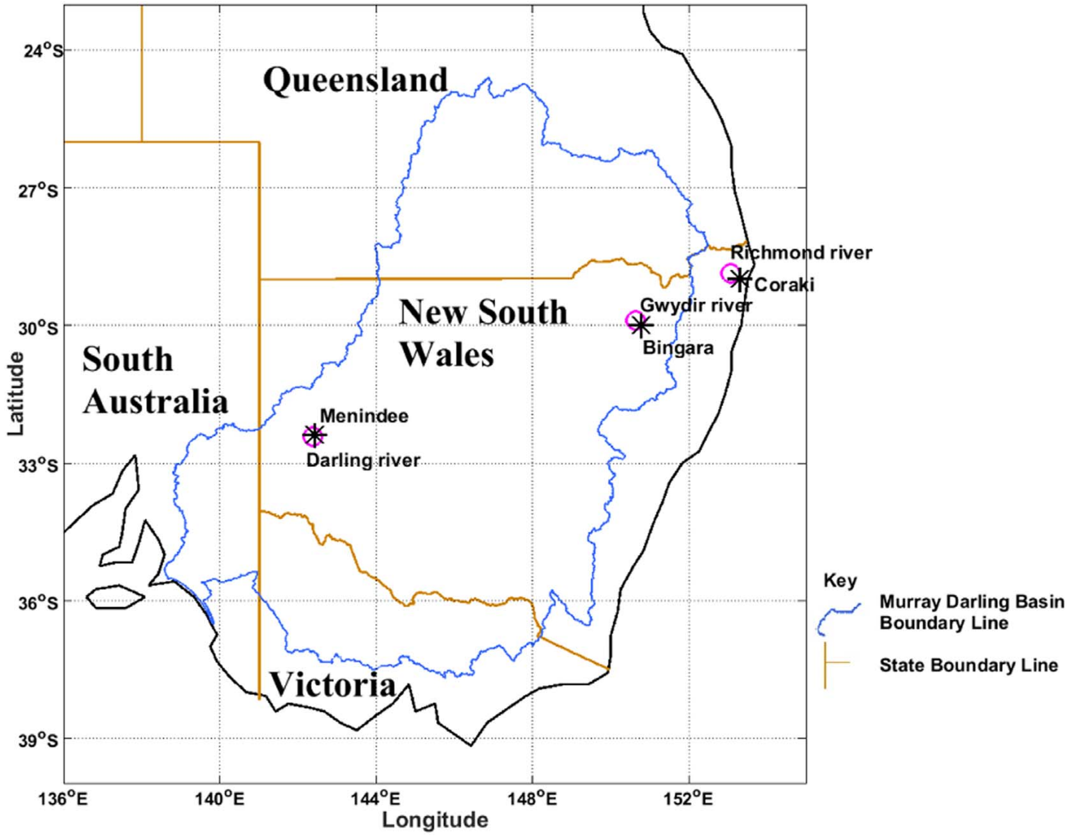


Fig. 2. Map of the study region showing the tested stations and their geographical location.

The shortest distance between meteorological and hydrological stations was 5.80 km while the furthest distance was 26.44 km. The site selection criteria was based on the agricultural surface water consumption and hectares (ha) of land used for agriculture within the vicinity of these selected sites as per the statistical data from Australian Bureau of Statistics. In the vicinity of Darling River station (Site 3), on an average, about 18,700–38,600 ha of land area is used for agricultural purposes while agricultural surface water consumption is 13–52 Giga Litres (GL). At Gwydir River station (Site 2), on an average 600–3700 ha of land area is being used for agricultural holding and consumes on average 13–52 GL of surface water for agricultural purposes. (Australian Bureau of Statistics, 2008). The Site 1, Richmond River (outside MDB) has been selected for a comparison of results.

During quality checking procedure, all missing data were replaced using respective calendar averaged values deduced from the entire period of study. As per Table 2, the amount of missing streamflow water level data were small (i.e., 1.27%, 1.90% and 1.27% for Richmond River, Gwydir River, and Darling River, respectively). SILO-based meteorological data were constructed from observational records from Australian Bureau of Meteorology where missing values had been interpolated via statistical techniques (Beesley et al., 2009; Tozer et al., 2012; Zajaczkowski et al., 2013).

Fig. 3 plots the mean climatological pattern of the objective variable (i.e., streamflow water level) and the respective predictor (maximum and minimum temperature, precipitation, evaporation, mean solar radiation and vapor pressure). Minimum temperature ($T_{\min}$), maximum temperature ($T_{\max}$) and mean solar radiation (Rn) plots show vivid minima during June and July (winter) and are seen to reach a maximum in December and January (summer period). Accordingly, the evaporation (*Evap*) and vapor pressure (*VP*) occupied the smallest magnitude in the winter season. The precipitation (*PCN*) recorded the lowest value from July to September, however, streamflow water level exhibited a very irregular pattern with no clear dependence on any of these meteorological variables. Out of the three hydrological sites, the lowest values were recorded as: Gwydir River station-lowest $T_{\max}$ (16.89 °C); $T_{\min}$ (1.77 °C); *VP* (8.62 h Pa); *Evap* (59.06 mm), Darling River station recorded the lowest *Rn* (9.95 MJ m^{-2}); while Richmond River station recorded the lowest value of streamflow water level (0.72 m).

While Darling River station recorded highest $T_{\max}$ (35.02 °C); Rn (27.32 MJ m^{-2}); *Evap* (349.82 mm) and interestingly the highest streamflow water level (2.44 m), Richmond River station recorded highest $T_{\min}$ (19.73 °C); *VP* (23.31 h Pa); and *PCN* (169.59 mm). Analysis showed that except for Darling River [where the skewness of *PCN* was 2.11 mm], all other variables had a skewness value much closer to zero to conform to near-normal distributions. Kurtosis factor (a

Table 2
Data partitioning for model development.

Period of study	Datum points	Data features after lags	Partitioning			Station name	Percentage of missing *SWL* data
			Training	Validation	Testing		
Jan 1977–May 2016	473	473–5 = 468	70%, 328	15%, 70	15%, 70	Richmond Gwydir River Darling River	1.27 1.90 1.27

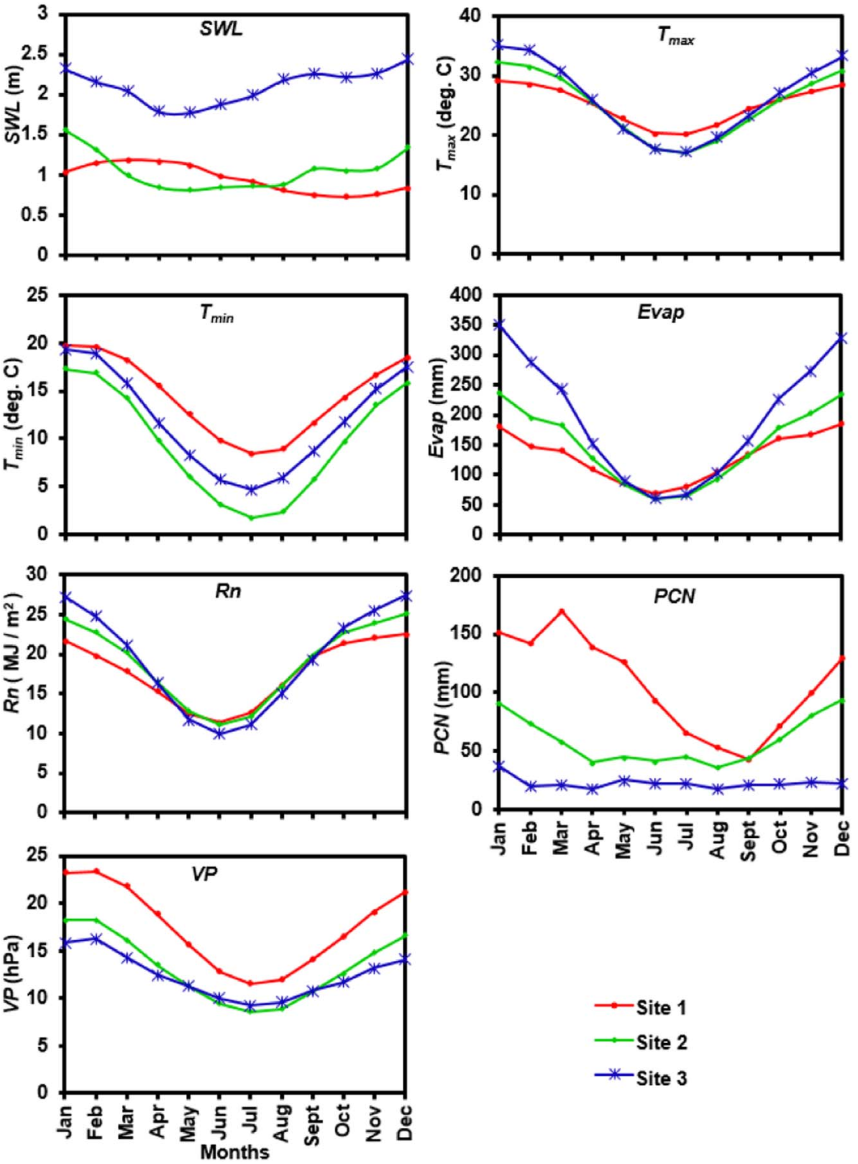


Fig. 3. Monthly climatological patterns (January 1977–May 2016) of the objective variable, stream water level (*SWL*) and the respective predictor variables (maximum temperatures, T_{max}; minimum temperatures, T_{min}; precipitation, *PCN*; evaporation, *Evap*; solar radiation, *Rn*; and vapor pressure, *VP*).

measure of whether the data were heavy or light-tailed) for Darling River precipitation registered a value of 6.93 mm for *PCN*, conforming to the leptokurtic distribution. This meant that the precipitation data had more outliers (with a heavy tail). By contrast, the other variables registered a kurtosis factor of < 3 (platykurtic), meaning the distribution seems to exhibit fewer and less extreme outliers (with light tail) than the normal (Table 1). It is clear that the three hydrological sites exhibit distinct climates, and therefore, offer a good comparison of the IIS-W-ANN and the respective counterpart forecasting models.

3.2. Construction of global predictor matrix and input selection process

Prior to developing the forecasting model, a global set of predictors related to streamflow, (i.e., hydrological and meteorological data) over a 40-year period, were constructed. It is imperative to note that any data-driven model must identify the patterns and trends in the predictors and predictand, where two primary approaches are capitalized. First, patterns in streamflow itself, which is partitioned in the training and testing sets are utilized for predictive modelling, since streamflow tends to display a high degree of serial correlation in time-space (i.e., persistence) arising from the groundwater storage and recharge that can act to amplify or dampen the effect of rainfall-runoff process; and hence, help in forecasting the future streamflow. If this is so, it can provide the streamflow data a memory of several (lagged) months representing the catchment hydrology and can also act as a driver for streamflow prediction (Chiew et al., 1998). Second, concurrent or time-lagged cross correlations between meteorological (i.e., rainfall, temperature, humidity, etc.) (McBride and Nicholls, 1983) and streamflow water level is used in a purely statistical-correlation sense (Chiew et al., 1998).

Fig. 4 illustrates a schematic view of different modelling stages. Following earlier work, both aforementioned approaches were adopted via a dual stage process: (i) MATLAB-based partial autocorrelation function (*PACF*) deduced at monthly lags for historical *SWL* data (Deo et al., 2016b; Yaseen et al., 2016b), (ii) MATLAB-based cross-correlation function (*CCF*) utilized with *SWL* vs. meteorological predictor variable (i.e., $T_{\max}$; T_{min}; *PCN*; *Rn*; *Evap*; *VP*) (Deo et al., 2017a; Deo and Sahin, 2017; Deo and Şahin, 2016; Deo et al., 2016c). At the respective lags, any variable with statistically significant relationship (i.e., at 95% confidence interval) with the predictand (streamflow) was screened as an input matrix, generating a global pool of statistically significant variables for streamflow water level forecasting. This yielded a global pool of 23, 16 and 20 model predictors for Richmond, Gwydir and Darling River, respectively (Table 3).

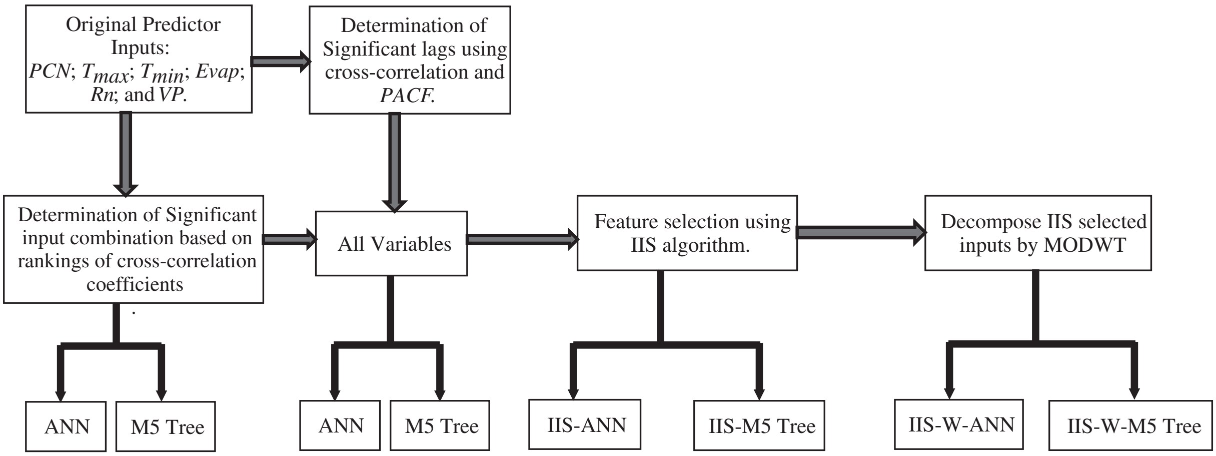

Original Predictor Inputs: PCN; Tmax; Tmin; Evap; Rn; and VP.
Determination of Significant lags using cross-correlation and PACF.
Determination of Significant input combination based on rankings of cross-correlation coefficients
All Variables
Feature selection using IIS algorithm.
Decompose IIS selected inputs by MODWT
ANN
M5 Tree
ANN
M5 Tree
IIS-ANN
IIS-M5 Tree
IIS-W-ANN
IIS-W-M5 Tree

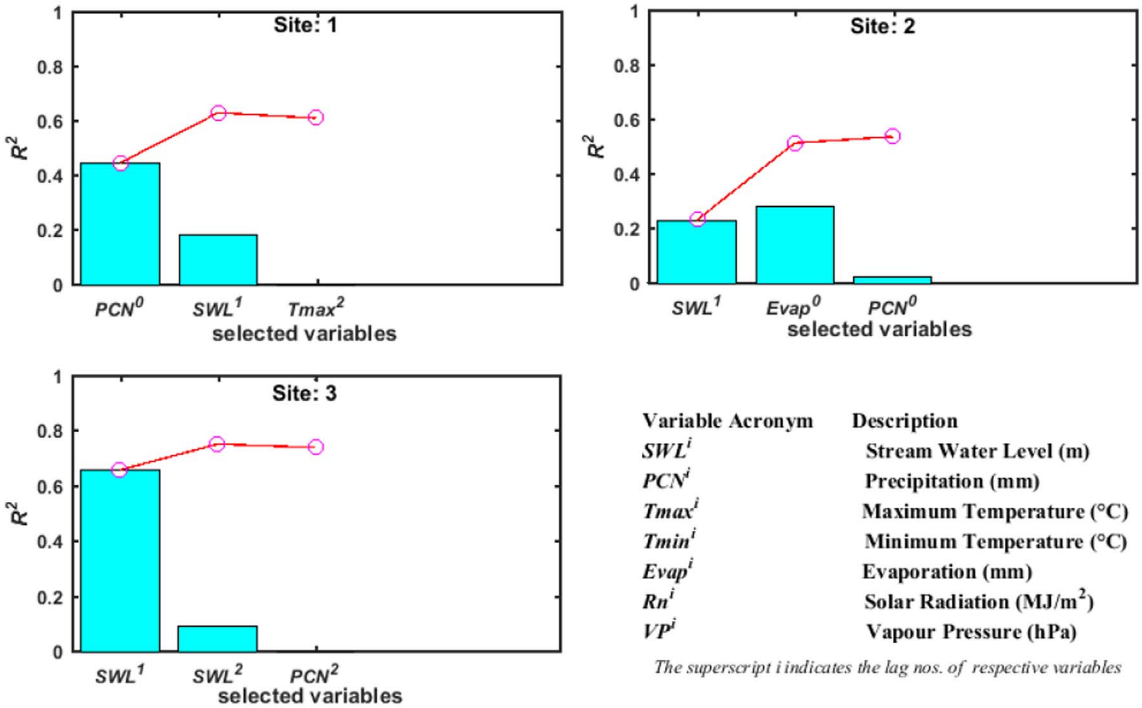
Fig. 4. A schematic view of the model development process.

To determine correct inputs with optimal features, a matrix of global predictors was screened via a tree-based iterative input selection (IIS) algorithm (Section 2.4) (Galelli and Castelletti, 2013b; Galelli et al., 2014) (Fig. 4). IIS utilized an underlying regression as an ensemble of Extra-Trees for randomizing attributes and cut-point choice while splitting a tree node (Galelli and Castelletti, 2013b; Geurts et al., 2006). This method is built on totally randomized trees whose structures are independent of the output values of the learning sample. Fig. 5 plots the cumulated performance, R^2 of Extra-Tree model within the IIS procedure and the contribution ΔR^2 of each screened variable evaluated as the variation of R^2 at each iteration. For Richmond River, cumulated performance increased monotonically with the number of the selected variables, up to the second variable. When the selection of an additional variable had no further significant increase in model performance and the algorithm tolerance, 'ε' was reached, the algorithm was terminated.

In congruence of this, it is construed that a significant proportion of streamflow processes can be described by means of two variables, which are the driver of streamflow water level at Richmond River: (i) the precipitation in the same month (lag = 0) (denoted as PCN^0) and (ii) previous month's streamflow water level (SWL^1) (lag = 1). *Note that*

Table 3
ANN and M5 Tree model structures with respective predictor variables.

Input combination	ANN						M5 Tree
	Number of neurons			Training algorithm	Hidden transfer function	Output transfer function	Number of rules
	Input layer	Hidden layer	Output layer				
SITE: 1 Richmond River							
PCN	1	3	1	trainbfg	logsig	purelin	14
PCN + *Rn*	2	24	1	trainlm	logsig	tansig	33
PCN + *Rn* + *Evap*	3	10	1	trainbfg	logsig	tansig	26
PCN + *Rn* + *Evap* + *VP*	4	24	1	trainbfg	tansig	tansig	36
PCN + *Rn* + *Evap* + *VP* + *Tmin*	5	16	1	trainbfg	tansig	purelin	39
PCN + *Rn* + *Evap* + *VP* + *Tmin* + *Tmax*	6	19	1	trainbfg	logsig	tansig	41
ALL 23 variables (*including significant lags*)	23	40	1	trainbfg	logsig	purelin	53
IIS selected variables: *PCN* (lag 0) + *SWL* (lag 1).	2	5	1	trainbfg	tansig	purelin	30
IIS-wavelet (*db3*-level 4 & *db4*-level 4)	10	4	1	trainlm	logsig	purelin	52
SITE: 2 Gwydir River							
Evap	1	13	1	trainbfg	logsig	tansig	15
Evap + *Rn*	2	33	1	trainbfg	logsig	purelin	31
Evap + *Rn* + *Tmax*	3	4	1	trainlm	logsig	purelin	37
Evap + *Rn* + *Tmax* + *Tmin*	4	4	1	trainlm	tansig	tansig	48
Evap + *Rn* + *Tmax* + *Tmin* + *VP*	5	23	1	trainbfg	tansig	logsig	39
Evap + *Rn* + *Tmax* + *Tmin* + *VP* + *PCN*	6	10	1	trainbfg	tansig	logsig	47
ALL 16 variables (*including significant lags*)	16	26	1	trainbfg	logsig	tansig	57
IIS selected variables - *SWL* (lag 1) + *Evap* (lag 0) + *PCN* (lag 0)	3	35	1	trainbfg	logsig	logsig	32
IIS-wavelet (*db3*-level 4 & *db3*-level 3)	10	11	1	trainbfg	logsig	tansig	57
SITE: 3 Darling River							
PCN	1	1	1	trainlm	tansig	tansig	11
PCN + *Rn*	2	26	1	trainbfg	tansig	purelin	21
PCN + *Rn* + *Tmin*	3	4	1	trainbfg	tansig	tansig	29
PCN + *Rn* + *Tmin* + *VP*	4	5	1	trainbfg	logsig	logsig	39
PCN + *Rn* + *Tmin* + *VP* + *Evap*	5	16	1	trainbfg	tansig	purelin	40
PCN + *Rn* + *Tmin* + *VP* + *Evap* + *Tmax*	6	18	1	trainlm	logsig	logsig	37
ALL 20 variables (*including significant lags*)	20	40	1	trainbfg	logsig	purelin	45
IIS selected variables: *SWL* (lag 1) + *SWL* (lag 2)	2	31	1	trainbfg	logsig	logsig	26
IIS-wavelet (*db4*-level 4 & *db4*-level 4)	10	2	1	trainlm	tansig	tansig	45



Variable Acronym	Description
SWLⁱ	Stream Water Level (m)
PCNⁱ	Precipitation (mm)
Tmaxⁱ	Maximum Temperature (°C)
Tminⁱ	Minimum Temperature (°C)
Evapⁱ	Evaporation (mm)
Rnⁱ	Solar Radiation (MJ/m^2)
VPⁱ	Vapour Pressure (hPa)

The superscript i indicates the lag nos. of respective variables

Fig. 5. Variable selection performed by Iterative Input selection (IIS) algorithm. Bars show the contribution to the coefficient of determination (ΔR^2) of each selected variable and the continuous (red) line denotes the cumulative performance coefficient of determination (R^2) of the underlying algorithm in IIS scheme.

the superscripts '0' and '1' denote the respective lag applied to construct the input data. Subsequently, three inputs were selected for Gwydir River with (i) SWL^1, (ii) evaporation in the same month ($EVAP^0$) and the PCN^0 data, while for Darling River, two significant variables were selected with (i) SWL^1 and (ii) SWL^2. It is interesting to note that one month lagged streamflow data was the common input for all hydrological sites while the maximum precipitation and evaporation data were important for Gwydir River. For Darling River site, the IIS algorithm did not identify precipitation and/or evaporation as a potential model input whereas 2-monthly lagged data was screened in this instance as an important predictor variable.

3.3. Model development and maximum overlap discrete wavelet transform

MATLAB software running over Intel *i*7, 3.40 GHz processor was utilized for the development of the ANN and the comparative M5 Tree model. Table 3 lists the sequential order of inputs applied in our model. This order was established via cross-correlation analysis performed with the predictand (i.e., streamflow water level) where model's improvement was monitored by a successive addition of variables. An input set for 'all predictors' consisted of all original time series and statistically significant lagged variables for each study site, whereas the final row of inputs (listed in Table 3) for each study site was determined from the IIS selected variables. A combination of the inputs was used in standalone non-IIS (i.e., ANN, M5 Model Tree) and the standalone IIS-integrated models (i.e., IIS-ANN, IIS-M5 Tree). Before the training process, all inputs were normalized to conform in the range of [0, 1] (Deo and Sahin, 2016; Deo and Şahin, 2015a; Deo et al., 2016c). Data partitioning was kept consistent (as indicated in Table 2). As there is no a set rule for data division, researchers have used different training, validation and testing set (Deo et al., 2016c). In this work, the subsets had training (70%), validation (15%) and testing (15%) data, while the target data were the time-series of the observed streamflow water level.

Table 3 shows the parameters of the ANN architecture with the corresponding inputs. In the case of ANN, the determination of an ideal network architecture (i.e., optimal neurons in the hidden layer), is important. It is noteworthy that a small architecture can lack sufficient degrees of freedom to correctly learn the predictor data, whereas an unnecessarily large architecture may not converge in a reasonable modelling time, and it may also overfit and memorize rather than generalize the data (Karunanithi et al., 1994). To avoid such issues, in the hidden layer a series of hidden neurons (h_n) starting at $h_n = 1$ to 40 in an incremental step of 1 were trialed and the model architecture that performed the best in terms of the lowest mean square error (*MSE*) criterion (Eq. (3)) was selected. To attain an accurate ANN model, various combinations of hidden transfer and output functions (via Eq. (2)) interchanged with the training algorithms were trialed one by one (Table 3). This resulted in a total of 480 ANN models executable with unique hidden neuronal architectures and predictor variables. The testing data (i.e., data unseen in the training phase) were utilized to assess the generalization capability of the optimal ANN network.

Improvement in ANN (including M5 Tree) model was facilitated by an integrating multi-resolution analysis that utilized maximum overlap discrete wavelet transform (MODWT). This generated wavelet hybrid models (i.e., IIS-W-ANN and IIS-W-M5 Tree) where inputs had already been selected using the IIS screening process (Table 3; Section 2.4, 3.2). However, prior to the integration of ANN and M5 Tree models with MODWT, the input data were partitioned to create independent matrices for training, validation, and testing. This prevented the inclusion of future data that were not truly available at a particular time step to be used in the model development, and the consequent unintentional induction of bias in the forecasted streamflow water level (Deo et al., 2016b; Kim and Valdes, 2003). MODWT was carried out independent of testing sets with coefficients combined in matrix used as an input.

In spite of the merits of the MODWT, it acts as a multi-resolution (frequency) identification tool for predictors, a challenge faced in model development phase was to deduce the best mother wavelet for the MODWT, as no explicit rule currently exists (Rathinasamy et al., 2013). That is, it was unclear from the current literature on the best wavelet function, as this aspect of model development is likely to depend on a particular problem of interest. To resolve this ambiguity, we adopted Daubechies mother wavelet which is normally used in hydrology, with three different forms: *db2* (2-vanishing moment), *db3* (3-vanishing moment) and *db4* (4-vanishing moment) (Deo et al., 2016c;

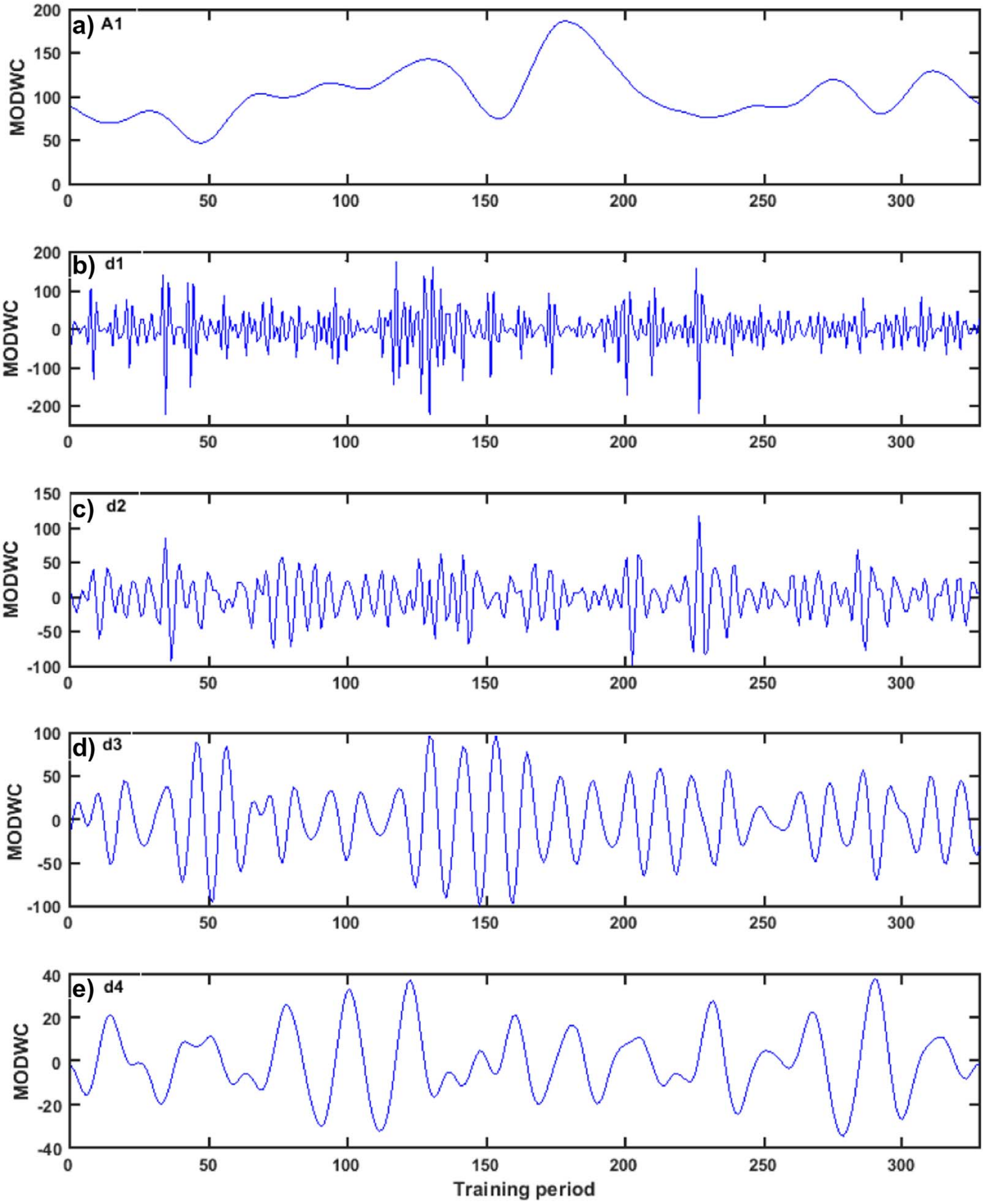


Fig. 6. Plot of maximum overlap discrete wavelet coefficients (MODWC) in the training period for monthly precipitation (*PCN*) lag 0 at Site 1: Richmond River.

Tiwari and Adamowski, 2013; Tiwari and Chatterjee, 2011). It is imperative to note that Daubechies wavelet was selected following literature, that an irregular mother wavelet was suited for hydrological prediction (Cannas et al., 2006; Mehr et al., 2014; Nourani et al., 2011).

The required minimum wavelet decomposition level was determined according to (Adamowski and Chan, 2011; Nourani et al., 2011):

$$L = \text{int}[\log(\text{N})] \tag{7}$$

Here, N (= 468) is the number of datum points and L = minimum level of decomposition (integer) ($L \approx 3$). As L is not known a priori, in this paper we tested decomposition levels 3 and 4, resulting in (4 and 5) wavelet decomposed outputs as per IIS-selected variable (i.e., one level for approximation, A and the other levels for detailed wavelet coefficients, d. It is also important to mention that *db2* did not yield satisfactory results and was thus ignored from all final analysis). Fig. 6a–d illustrate the MODWT coefficients generated at 4 levels of decomposition ($d1$, $d2$, $d3$, and $d4$) and one level of approximation ($A4$) for precipitation (*PCN*) input data in the case of Richmond River.

In order to evaluate the preciseness of the ANN model, the wavelet hybrid and standalone M5 Tree model were built using the same sequential addition of inputs as with ANN (with all and IIS-selected variables). For M5 Tree, software package at http://www.cs.rtu.lv/jekabsons/, developed by Jekabsons (2010) was utilized. During the training process, an initial model tree was erected and this was pruned later and validated using 10-fold cross-validation procedure as with earlier works (Deo et al., 2017a; Deo et al., 2017b; Kisi, 2015). The input/output data space was split into subspaces with localized rules built for each subspace and optimal number of rules for each optimal model was deduced, as shown in Table 3.

3.4. Model evaluation criteria

In this study, a statistical evaluation of the performance was conducted by means of statistical indicators: correlation coefficient (r), Willmott's Index (*WI*), Nash–Sutcliffe Efficiency (E_{NS}), Root-Mean-Square-Error (*RMSE*) and Mean Absolute Error (*MAE*) (Legates and McCabe, 1999a; Nash and Sutcliffe, 1970; Shamseldin, 1997; Willmott, 1981; Willmott, 1984). The percentage error measures, root mean square error (*RRMSE*) and mean absolute percentage error (*MAPE*) were also utilized to assess errors at different sites. As there is no universal and no single metric for assessing hydrological models (Chai and Draxler, 2014a; Krause et al., 2005a; Legates and McCabe, 1999b), this study is based on several performance metrics with mathematical equations as follows:

i. Correlation coefficient (r):

$$r = \frac{\sum_{i=1}^{N} (SWL_{OBS,i} - \overline{SWL}_{OBS})(SWL_{FOR,i} - \overline{SWL}_{FOR})}{\sqrt{\sum_{i=1}^{N} (SWL_{OBS,i} - \overline{SWL}_{OBS})^2}\sqrt{\sum_{i=1}^{N} (SWL_{FOR,i} - \overline{SWL}_{FOR})^2}} \tag{8}$$

ii. Willmott's Index (WI):

$$WI = 1 - \left[\frac{\sum_{i=1}^{N} (SWL_{FOR,i} - SWL_{OBS,i})^2}{\sum_{i=1}^{N} (|SWL_{FOR,i} - \overline{SWL}_{OBS}| + |SWL_{OBS,i} - \overline{SWL}_{OBS}|)^2}\right], \quad 0 \le WI \le 1 \tag{9}$$

iii. Nash–Sutcliffe Efficiency (E_{NS}):

$$E_{NS} = 1 - \left[\frac{\sum_{i=1}^{N} (SWL_{OBS,i} - SWL_{FOR,i})^2}{\sum_{i=1}^{N} (SWL_{OBS,i} - \overline{SWL}_{OBS})^2}\right], (-\infty \quad < E_{NS} < 1) \tag{10}$$

iv. Root mean square error ($RMSE$):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (SWL_{FOR,i} - SWL_{OBS,i})^2} \tag{11}$$

v. Mean absolute error (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |(SWL_{FOR,i} - SWL_{OBS,i})| \tag{12}$$

vi. Relative root mean square error ($RRMSE$, %):

$$RRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} (SWL_{FOR,i} - SWL_{OBS,i})^2}}{\frac{1}{N}\sum_{i=1}^{N} (SWL_{OBS,i})} \times 100 \tag{13}$$

vii. Mean absolute percentage error ($MAPE$; %):

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} \left|\frac{(SWL_{FOR,i} - SWL_{OBS,i})}{SWL_{OBS,i}}\right| \times 100 \tag{14}$$

In Eqs. (8–14), SWL_{OBS} is observed streamflow water level and SWL_{FOR} is the forecasted streamflow water level, i represents the occurrence time and N is the number of datum points within the testing period.

Inherent merits and weaknesses do not permit a single metric to independently evaluate the models. Therefore, it is prudent to use a combination of metrics (Chai and Draxler, 2014b). Different combinations of performance indicators, such as the correlation coefficient (r) and *MAE* (Legates and McCabe, 1999a); Nash–Sutcliffe Efficiency (E_{NS}) and *RMSE* (Humphrey et al., 2016; Mehr et al., 2014; Sajikumara and Thandaveswarab, 1999) are chosen for model evaluation. With correlation coefficient (r) the advantage is that it provides the information on the degree as well as the direction of the linear association between the observed and forecasted streamflow values, without which the information of this association is unclear. Additionally, r is parametric and insensitive to additive and proportional differences between simulated and observed homologous elements (Hora and Campos, 2015). Yet, the drawback is that r is oversensitive to extreme values (outliers) (Legates and McCabe, 1999a; Willmott, 1981). The Nash–Sutcliffe Efficiency (E_{NS}) is widely used criteria for evaluating the hydrological models and is considered to be a skill score computed as the comparative ability of a model with regards to a baseline model, which in this case is the mean of the observed streamflow water level values (Gupta et al., 2009). However, E_{NS} overestimates the larger values and the lower values are neglected (Legates and McCabe, 1999a). Willmott's index (*WI*) has also been used which is meritorious in comparison to r and E_{NS}, as in *WI* computation the differences between the observed and forecasted values are not squared (Legates and McCabe, 1999a) which overcomes the insensitivity issues. However, at times high values ($WI \ge 0.650$) are plausible even for poor model fits (Krause et al., 2005b).

As far as error measurements are concerned, both the error measures of *RMSE* and *MAE* are based on aggregation of residuals of observed and forecasted streamflow water level values (Nourani et al., 2011). The dissimilarity is that, in *RMSE* computation, the aggregation of residuals is squared, while in *MAE* it is not. Hence, *RMSE* is able to measure the goodness of fit relevant to high flows, while the *MAE* indicates the goodness of fit at moderate flow values (Galelli and Castelletti, 2013a) as *MAE* equally evaluates all deviations from the observed values (Deo et al., 2016b). A weakness, however, is that these are expressed in their absolute units, and thus should not be solely used to compare model performance at geographically diverse sites (Hora and Campos, 2015) (e.g., Fig. 2). Subsequently, we utilized relative errors; root mean square error (*RRMSE*) and mean absolute percentage error (*MAPE*) to describe the model's behavior over the range of statistically different hydrological flows, making it possible to compare the models evaluated for geographically (and climatically) diverse sites where *MAE* and *RMSE* alone do not make sense (Deo et al., 2017a; Deo et al., 2016d).

4. Results and general discussion

In this section the appraisal of wavelet-hybrid ANN integrated with iterative input selection algorithm (IIS-W-ANN) is undertaken for hydrological sites in drought-prone, Murray-Darling Basin. IIS-W-ANN is evaluated with respect to a standalone ANN (with & without IIS-screened predictors) including an equivalent M5 Tree model. To establish whether the IIS-W-ANN was a parsimonious model accomplishing a desired level of accuracy, an iterative modelling process was applied to optimize the input combinations, training algorithm and hidden transfer functions where the lowest mean square error for an optimal model was sought. In the test period, statistical metrics, as described in Eq. (8–14) are used to justify the results in the following section.

Table 4 evaluates both ANN and M5 Tree models integrated with maximum overlap discrete wavelet transform and iterative input selection algorithm. The performances attained by sequential addition of predictors used in model construction are shown for each study site. Interestingly, an incremental improvement in model accuracy was attained by the successive addition of variables where all available variables for each site seemed to demonstrate a better performance compared to single or incorrectly screened multiple variables. However, the ANN models executed with IIS-selected variables demonstrated a dramatic improvement where *WI* and E_{NS} were higher, and *RMSE/MAE* were lower than the models without IIS. In light of this, it is averred that the improved forecasting of streamflow requires appropriate input combinations, where IIS scheme was seen to enhance the performance of models with proper screening of variables.

In terms of numerical quantification of models, for the case of

Table 4
Evaluation of ANN and M5 Tree models, integrated with maximum overlap discrete wavelet transform and iterative input selection (IIS) algorithm. *r* = correlation coefficient; *WI* = Willmott's Index; E_{NS} = Nash–Sutcliffe Efficiency, *RMSE/MAE* = root mean square/mean absolute error. The incremental improvement attained by addition of successive variables, all variables, IIS-selected variables and subsequent wavelet models are shown.

Input combination	ANN					M5 model tree				
	r	*WI*	E_{NS}	*RMSE* (m)	*MAE* (m)	*r*	*WI*	E_{NS}	*RMSE* (m)	*MAE* (m)
SITE: 1 Richmond river & Coraki										
PCN	0.652	0.723	0.320	0.393	0.264	0.603	0.693	0.182	0.430	0.327
PCN + *Rn*	0.736	0.809	0.478	0.344	0.238	0.643	0.710	0.323	0.392	0.255
PCN + *Rn* + *Evap*	0.698	0.747	0.395	0.370	0.242	0.629	0.675	0.271	0.406	0.262
PCN + *Rn* + *Evap* + *VP*	0.777	0.785	0.450	0.353	0.238	0.687	0.705	0.339	0.387	0.252
PCN + *Rn* + *Evap* + *VP* + *Tmin*	0.804	0.790	0.482	0.343	0.232	0.685	0.703	0.329	0.390	0.256
PCN + *Rn* + *Evap* + *VP* + *Tmin* + *Tmax*	0.803	0.807	0.517	0.331	0.221	0.695	0.711	0.332	0.389	0.258
ALL 23 variables (*including significant lags*)	0.844	0.867	0.624	0.292	0.192	0.799	0.838	0.557	0.317	0.190
IIS selected: *PCN* (lag 0) + *SWL* (lag 1).	0.842	0.898	0.687	0.267	0.171	0.816	0.849	0.595	0.303	0.186
IIS-wavelet (*db3*-level 4/*db4*-level 4)	0.910	0.936	0.770	0.229	0.177	0.822	0.892	0.578	0.309	0.225
SITE: 2 Gwydir river & Bingara										
Evap	0.766	0.822	0.558	0.243	0.200	0.730	0.779	0.501	0.258	0.209
Evap + *Rn*	0.788	0.839	0.585	0.236	0.188	0.726	0.777	0.495	0.260	0.204
Evap + *Rn* + *Tmax*	0.758	0.839	0.564	0.242	0.179	0.705	0.790	0.481	0.263	0.197
Evap + *Rn* + *Tmax* + *Tmin*	0.793	0.836	0.599	0.231	0.185	0.686	0.789	0.458	0.269	0.200
Evap + *Rn* + *Tmax* + *Tmin* + *VP*	0.802	0.857	0.616	0.227	0.179	0.678	0.785	0.449	0.272	0.201
Evap + *Rn* + *Tmax* + *Tmin* + *VP* + *PCN*	0.830	0.859	0.645	0.218	0.169	0.698	0.799	0.477	0.264	0.196
ALL 16 variables (*including significant lags*)	0.835	0.904	0.697	0.201	0.153	0.750	0.856	0.560	0.243	0.182
IIS selected - *SWL* (lag 1) + *Evap* (lag 0) + *PCN* (lag 0)	0.824	0.900	0.679	0.207	0.154	0.780	0.870	0.609	0.229	0.175
IIS-wavelet (*db3*-level 4/*db3*-level 3)	0.908	0.944	0.803	0.162	0.139	0.882	0.929	0.769	0.176	0.147
SITE: 3 Darling river & Menindee post office										
PCN	0.508	0.162	0.056	1.678	1.165	− 0.170	0.081	− 0.069	1.786	1.274
PCN + *Rn*	0.440	0.505	0.189	1.555	1.125	− 0.100	0.144	− 0.080	1.795	1.264
PCN + *Rn* + *Tmin*	0.298	0.239	0.064	1.672	1.157	− 0.070	0.144	− 0.067	1.785	1.273
PCN + *Rn* + *Tmin* + *VP*	0.368	0.302	0.102	1.637	1.135	− 0.150	0.148	− 0.123	1.830	1.268
PCN + *Rn* + *Tmin* + *VP* + *Evap*	0.612	0.685	0.350	1.393	0.971	0.043	0.182	− 0.046	1.767	1.207
PCN + *Rn* + *Tmin* + *VP* + *Evap* + *Tmax*	0.686	0.827	0.329	1.415	1.063	0.205	0.262	0.028	1.703	1.149
ALL 20 variables (*including significant lags*)	0.871	0.926	0.759	0.849	0.573	0.861	0.925	0.738	0.885	0.524
IIS selected: *SWL* (lag 1) + *SWL* (lag 2)	0.921	0.958	0.847	0.676	0.373	0.876	0.936	0.765	0.837	0.487
IIS-wavelet (*db4*-level 4/*db4*-level 4)	0.960	0.979	0.920	0.487	0.390	0.922	0.952	0.843	0.684	0.501

Richmond River, the IIS-ANN model registered *WI* = 0.898, E_{NS} = 0.687 and *RMSE* = 0.267 m and *MAE* = 0.171 m. The IIS-M5 tree, however, also performed likewise, with *WI* = 0.849, E_{NS} = 0.595 and *RMSE* = 0.303 m and *MAE* = 0.186 m. A similar trend was evident for the model applied at Darling River site. Only for the case of Gwydir River, the inputs with 'all variables' for the ANN model performed slightly better with *WI* = 0.904, E_{NS} = 0.697, *RMSE* = 0.201 m and *MAE* = 0.153 m, while IIS selected models performance was second best and both IIS-ANN and IIS-M5 Tree had a similar performance. Interestingly, a similar trend was consistently demonstrated by the remainder of the forecasting error indicators, despite their advantages and drawbacks as discussed in the previous section (Section 3.4). Ideally, *RMSE* and *MAE* values must be as small as possible to reflect the lowest (or ideally 0) deviation of predictions from the observations, similarly, for perfect model fit, *r*, *WI*, and E_{NS} should be equal to unity.

When the performance of standalone ANN and M5 Tree models were compared for each hydrological site, it was evident that although the M5 Tree model yielded good performance, the ANN model outperformed the M5 Tree for all hydrological study sites. Although the exact cause of this is not yet known, this result could be due to the ANN model being a purely non-linear model, while the M5 Tree model being a hierarchical tree-based linear centered model. At all hydrological sites, the antecedent (one) month's streamflow has been selected as the preferred predictor variable deduced from the global pool. However, it was interesting to note that the precipitation data were not the sole predictor.

At the Gwydir River site, the results showed that precipitation data were not selected as the desired predictor variable, but rather the evaporation data seemed a better model predictor. Generally, the ability of all forecasting models in each input combination instance to provide an accurate estimation of streamflow was contingent upon proper input selection, and this was reflected very well with the IIS algorithm implemented into ANN and its comparative M5 Tree model. Based on model metrics, it can be confirmed that the IIS algorithm provided an ideal combination of inputs with the lowest errors (*RMSE/MAE*) and high performances (*r*, *WI*, E_{NS}) deduced within the testing period.

In this paper, we examined whether the forecasting models attained better accuracy when the non-decimated, maximum overlap discrete wavelet transform (MODWT) algorithm (Section 3.3) was implemented (i.e., leading to a set of hybrid ANN and M5 Tree models). In accordance with Table 4, MODWT-based decomposition of IIS selected variables led to a dramatic improvement of both IIS-W-ANN and IIS-W-M5 Tree compared to standalone IIS-ANN and IIS-M5 Tree model. When comparing the IIS-W-ANN against IIS-ANN, results showed that the value of *WI* had increased for all hydrological sites. For instance, at Richmond River site, the magnitude of *WI* increased from 0.898 to 0.936, at Gwydir River site, it increased from 0.900 to 0.944 and at Darling River, it increased from 0.958 to 0.979. Likewise, the value of E_{NS} also increased with MODWT algorithm was integrated with the model (i.e., Richmond River from 0.687 to 0.770; Gwydir River from 0.679 to 0.803; Darling River from 0.847 to 0.920). In the retrospect, the correlation coefficient, *r*, was relatively larger for the wavelet-hybrid model.

Other than *WI* and E_{NS} (normalized metrics) that justified a better utility of IIS-W-ANN and IIS-W-M5 Tree, there was a significant reduction in *RMSE* and *MAE* for all hydrological study sites. It is imperative to note that for hydrological evaluations, forecast models with E_{NS} value > 0.900 is considered to be 'very satisfactory', those between 0.800 and 0.900 are 'fairly good', and those below 0.800 are 'unsatisfactory' (Shamseldin, 1997). Therefore, the present wavelet-coupled model precision appears to be 'fairly good' when applied for

Table 5
Relative errors of wavelet-hybrid models integrated with iterative input selection (IIS) and subsequent variable combination using *RRMSE* and *MAPE*. Optimal models with smallest forecasted error (%) are shown in **boldface**.

Input combination	ANN		M5 Tree	
	RRMSE	*MAPE*	*RRMSE*	*MAPE*
SITE: 1 Richmond River				
PCN	36.02	21.23	39.50	33.83
PCN + *Rn*	31.55	18.92	35.95	20.19
PCN + *Rn* + *Evap*	33.98	18.92	37.29	20.11
PCN + *Rn* + *Evap* + *VP*	32.40	19.09	35.50	19.26
PCN + *Rn* + *Evap* + *VP* + *Tmin*	31.45	18.44	35.78	19.60
PCN + *Rn* + *Evap* + *VP* + *Tmin* + *Tmax*	30.37	17.31	35.71	19.95
ALL 23 variables (*including significant lags*)	26.77	15.11	29.06	13.60
IIS selected: *PCN* (lag 0) + *SWL* (lag 1).	24.46	**13.58**	**27.81**	**13.51**
IIS-wavelet (db3-level 4/db4-level 4)	**20.97**	17.00	28.37	18.71
SITE: 2 Gwydir River				
Evap	23.41	20.62	24.89	20.86
Evap + *Rn*	22.70	19.46	25.04	20.33
Evap + *Rn* + *Tmax*	23.27	18.02	25.37	19.88
Evap + *Rn* + *Tmax* + *Tmin*	22.30	18.29	25.93	20.22
Evap + *Rn* + *Tmax* + *Tmin* + *VP*	21.83	18.77	26.15	20.17
Evap + *Rn* + *Tmax* + *Tmin* + *VP* + *PCN*	20.98	17.16	25.48	19.57
ALL 16 variables (*including significant lags*)	19.38	14.84	23.38	16.87
IIS selected - *SWL* (lag 1) + *Evap* (lag 0) + *PCN* (lag 0)	19.96	14.84	22.02	16.81
IIS-Wavelet (*db3*-level 4/*db3*-level 3)	**15.65**	**14.79**	**16.92**	**15.40**
SITE: 3 Darling River				
PCN	72.30	51.94	76.93	57.52
PCN + *Rn*	67.01	52.96	77.34	56.18
PCN + *Rn* + *Tmin*	72.01	51.21	76.89	57.19
PCN + *Rn* + *Tmin* + *VP*	70.52	50.74	78.85	55.32
PCN + *Rn* + *Tmin* + *VP* + *Evap*	60.01	42.23	76.11	52.27
PCN + *Rn* + *Tmin* + *VP* + *Evap* + *Tmax*	60.97	56.85	73.38	48.71
ALL 20 variables (*including significant lags*)	36.56	26.87	38.13	21.22
IIS selected: *SWL* (lag 1) + *SWL* (lag 2)	29.13	**16.78**	36.05	**19.63**
IIS-wavelet (*db4*-level 4/*db4*-level 4)	**21.00**	20.78	**29.46**	24.06

streamflow water level forecasting at Gwydir River and while the models are 'very satisfactory' for the case of Darling River.

Considering the aforesaid, it is noteworthy that the results revealed a geographic signature in model accuracy such that the statistical performances for all hydrological sites were disparate in terms of the range of metrics attained. That is to say, the IIS-W-ANN model for the case of Darling River exhibited the largest correlation, while the Gwydir River recorded the lowest forecasted errors in terms of the metrics for all hydrological stations. Darling River, however, recorded the largest *WI* (≈ 0.979), E_{NS} (≈ 0.920) and r (≈ 0.960) followed by smaller values for Gwydir River and Richmond River. The lowest *RMSE* value was 0.162 m and the lowest *MAE* was 0.139 m was recorded at Gwydir River with the best performing model (IIS-W-ANN) followed by Richmond and then Darling Rivers. This result ascertains that the wavelet-hybrid ANN (and the comparative M5 Tree) model performance is not universally similar when the study sites with different hydrological conditions are considered (Table 1; Fig. 2).

To address the limitations of *RMSE/MAE* (that it promotes a robust evaluation of models applied at geographically diverse hydrological sites), Table 5 lists an alternative metric, the relative error values of wavelet-hybrid ANN (and M5 Tree) models integrated with iterative input selection (IIS). Here, we also show the subsequent variable combinations where the percentage of *RMSE/MAE* values are listed (i.e., *RRMSE* and *MAPE*). Evidently, the relative performance (Table 5) revealed that the IIS-W-ANN model exhibited the lowest value of *RRMSE* for all three hydrological sites which were apparently lower than those of the IIS-W-M5 Tree model. More precisely, the *RRMSE* values for each hydrological site in the combination [IIS-W-ANN: IIS-W-M5 Tree] are as follows: Richmond River: [20.97%: 28.37%]; Gwydir River: [15.65%: 16.92%]; Darling River: [21.00%: 29.46%]. Accordingly, the *RRMSE* values showed that the IIS-W-ANN performed the best for the case of Gwydir River, followed by Richmond and Darling Rivers, respectively, to concur with the architectures with input–hidden–output layer combinations of optimal models as 10-4-1 for Richmond River, 10-11-1 for Gwydir River, and 10-2-1 for Darling River (Table 3).

In Fig. 7 (a–b), a visual evaluation of the forecasted streamflow relative to the observed data has been performed with scatterplots prepared in the testing period. In each panel, a coefficient of determination (R^2) is used to examine the goodness-of-fit of the forecasting model developed for all three candidate stations. Note that here, we are interested in evaluating several models, including IIS-ANN, IIS-M5 Tree, IIS-W-ANN and IIS-W-M5 Tree. Evidently, the optimal model (i.e., IIS-W-ANN) is seen to register a large R^2 value in comparison with IIS-W-M5 Tree for all hydrological sites. Specifically, the results reflected the case of Richmond River ($R^2 \approx 0.827$), Gwydir River (≈ 0.825) and Darling River (≈ 0.921), which concurred with the results in Table 4.

Notably, the standalone IIS-ANN model (with IIS selected variables) outperformed the standalone IIS-M5 Tree, with R^2 values as follows: Richmond River ($R^2 \approx 0.709$), Gwydir River (≈ 0.679) and Darling River (≈ 0.848). The gradient (m) of linear fit, which is an alternative model performance metric, for the case of IIS-W-ANN model was found to be close to unity (i.e., 0.827 for Richmond River, 0.825 for Gwydir River, 0.921 for Darling River). On the other hand, the y-intercept, which should ideally be zero for the case of IIS-W-ANN, was 0.064 (Richmond River), 0.235 (Gwydir River) and 0.253 (Darling River) whereas for the case of IIS-W-M5 Tree models, y-intercept was 0.065, 0.293 and 0.535, respectively. In congruence with results presented in Table 4, for the IIS and non-wavelet models, the y-intercept deviated significantly from the ideal value of 0; indicating the superiority of the forecasting models where input selection with the IIS algorithm and multi-resolution analysis (with MODWT) was implemented.

Further evaluation of IIS-W-ANN relative to IIS-W-M5 Tree model and the respective standalone (ANN & M5 Tree) model is undertaken with a time-series plot of data in the test period (Figs. 8 a–b). The time-series plot (Fig. 8a) for hybrid models provides a definitive evidence that the IIS-W-ANN attained a better accuracy for all hydrological stations, such that the standalone model appeared to under-predict the streamflow water level data, while the inclusion of wavelet decomposed predictors acted to improve the forecasted result. Closer examination also showed that the low streamflow values were better forecasted by the standalone model and the higher streamflow values were better forecasted with the wavelet hybrid models.

Interestingly, a plateau was also observed for the case of Darling River for two major high flow events as depicted by the IIS-W-ANN model, while the IIS-W-M5 Tree model severely under-predicted this event. However, the standalone IIS-ANN and IIS-M5 Tree model provided much better forecasts for these anomalous hydrological events within the test period. Overall, in congruence with key statistical metrics (i.e., Table 4), there was a very good visual agreement between the observed and forecasted streamflow data within the test period, especially those produced by the IIS-W-ANN model relative to the other model counterparts for all hydrological sites.

So far, the analysis has provided compelling evidence of the superiority of input selection and multi-resolution wavelet decomposition in terms of the accuracy of the prescribed streamflow models. In Fig. 9 (a–b), the influence of input selection procedure (IIS) and MODWT is further checked for the ANN and M5 Tree models where the data for all three sites are pooled together. In this case, the percentage difference in the key model performance metrics (i.e., r, *WI* & *RRMSE*) are shown where the IIS scheme was applied with different predictor variables, while the percentage difference is also shown when the MODWT algorithm was integrated with the models having only the IIS-selected variable.

It is clear that the integration of the ANN model with the IIS scheme has produced 1.4% and 2.2% increase in r and *WI* values and about

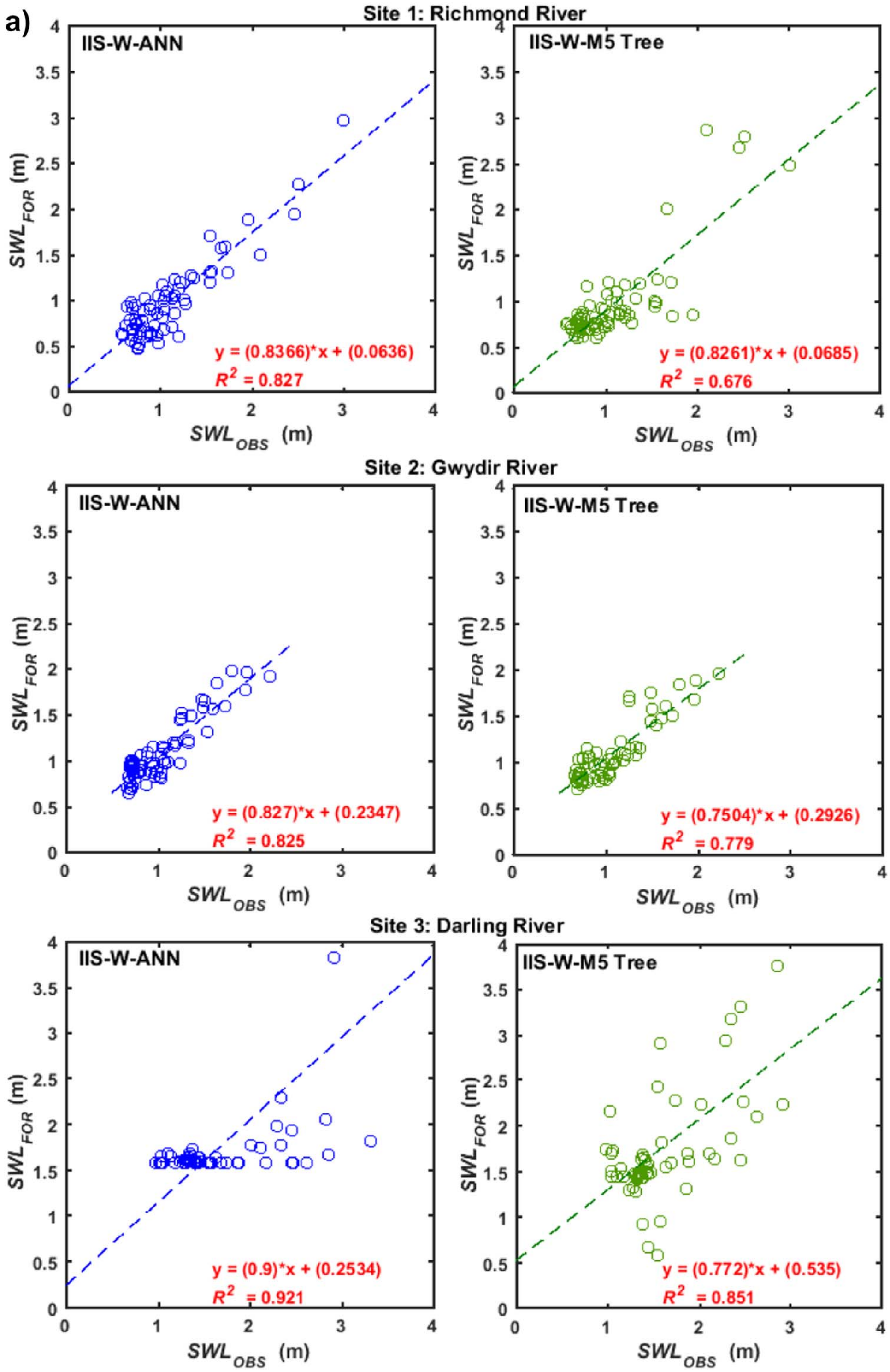


Fig. 7. Scatter plots of observed (SWL_{OBS}) and forecasted (SWL_{FOR}) stream water level for all the stations from: a) IIS-W-ANN and IIS-W-M5 Tree models, b) IIS-ANN and IIS-M5 Tree models (No wavelet applied). (Note: The dashed line in blue and green is the least-squares fit line to the respective scatter plots).

8.7% reduction in *RMSE* value (Fig. 9a). In terms of the effect of wavelet transform on models implemented with IIS-selected variables, there was a larger increase in *r* and *WI* by about 7.5% and 3.8%, respectively, whereas the *RMSE* has decreased by 21.3% when the IIS-W-ANN model was tested (Fig. 9b). A similar trend was also found for the M5 Tree model, albeit with a slightly lesser increase in *r* and a smaller reduction in *RRMSE*.

Additionally, the model preciseness was assessed using box plots illustrating the spread, with respect to quartiles, of the observed and forecasted streamflow water level values produced by the IIS and wavelet integrated model (Fig. 10a) and later with the IIS coupled model (without wavelets) (Fig. 10b). Being non-parametric, boxplots do not make assumptions about the underlying statistical distributions, allowing a better understanding of the degree of spread and skewness of a data set, while the whiskers indicate the variability outside of the lower (25th percentile) and upper (75th percentile) quartiles. For Richmond River, the spread of the IIS-W-ANN and the observed streamflow water level (SWL_{OBS}) had almost similar spread showing that the IIS-W-ANN has better forecasting ability for this station (Fig. 10a) while the IIS-W-M5 Tree displayed an unacceptably smaller spread. Likewise, the distribution of outlier data points registered by the IIS-W-ANN model compared well with the observed values.

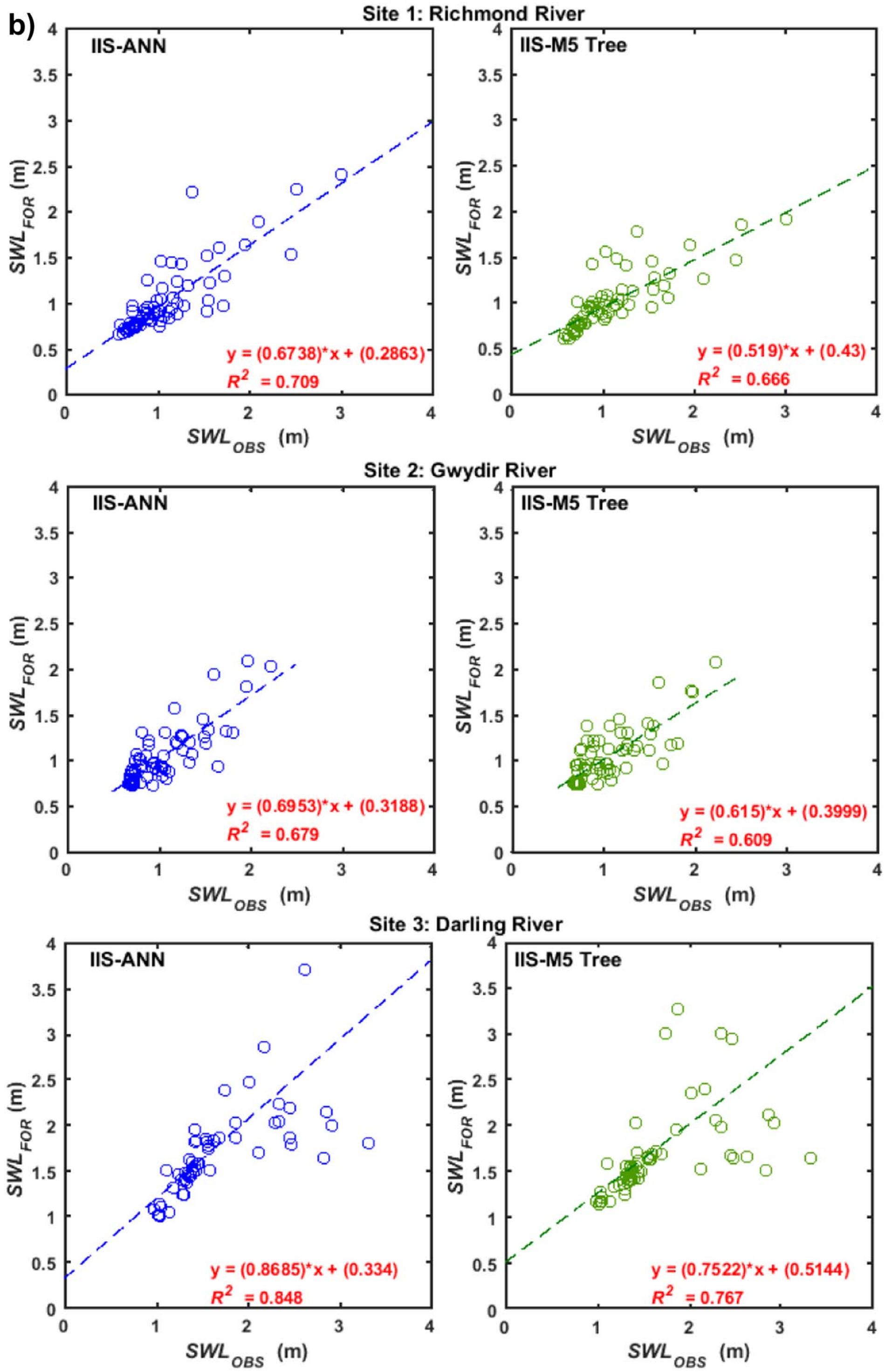


Fig. 7. (*continued*)

In the case of Gwydir River notably, the IIS-W-ANN registered a greater spread of the forecasts, as the streamflow water level were more skewed towards the higher end, which was consistent with the observed streamflow water level values and the medians of the two data sets were relatively the same. In the case of Darling River, where high streamflow levels were the main feature within the tested hydrological period, the IIS-W-ANN produced a better distribution of forecasts about median while surprisingly, the IIS-W-M5 Tree showed much-scattered spread. On the contrary, the IIS optimized ANN and M5 Tree models produced similar forecasting distributions at all the three sites (Fig. 10b), however, these distributions were not alike the spread of observed values. Thus, based on the forecast distributions produced by these boxplots, IIS-W-ANN had better predictive performance and is also affirmed by the assessment metrics (Table 4).

In this study, although a superior type of wavelet transform (i.e., non-decimated, maximum overlap discrete wavelet transformation, MODWT) was applied, the use of correct MODWT mother wavelet was still a challenging task, since different wavelets are expected to have different impacts on the frequency extraction process (Rathinasamy et al., 2013; Rathinasamy et al., 2014). Table 6 shows that the different Daubechies wavelets with different vanishing moments and decomposition levels that attained optimal results for our three hydrological sites. It is interesting to note that the IIS-W-ANN model generated an optimum performance where the Daubechies *db3* wavelet (3-vanishing

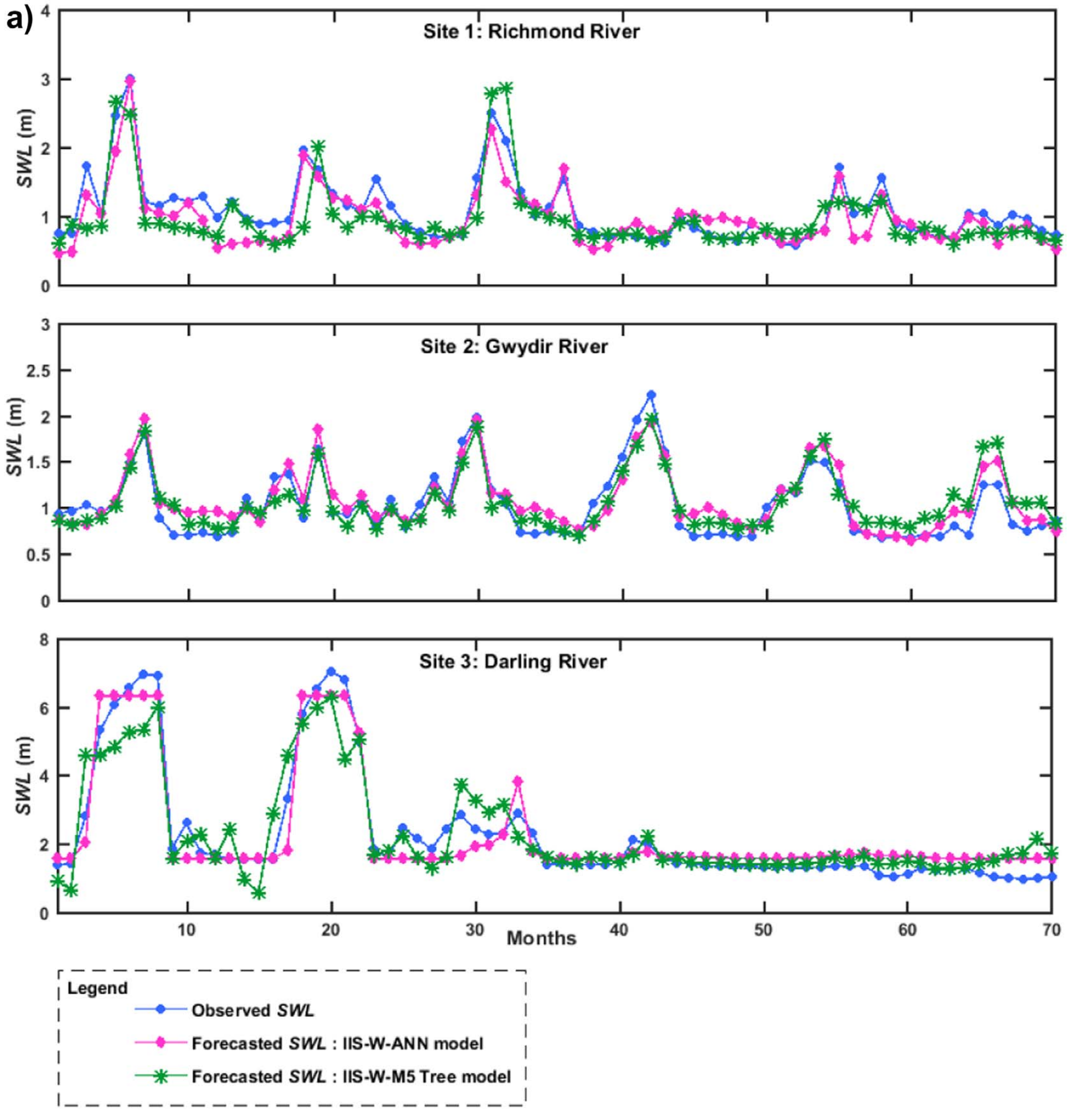


Fig. 8. Observed and forecasted streamflow water level (*SWL*) in the testing period, from: a) the optimal objective model, IIS-W-ANN, and tree based comparative counterpart, IIS-W-M5 Tree model, b) IIS-ANN and IIS-M5 Tree models (without wavelets).

moments) and 4 levels of decomposition for the case of Richmond River and Gwydir River yielded the highest *WI* values of about 0.936 and 0.944, respectively. The other performance measures followed alike trends with largest values of *r* and E_{NS} and minimum values of *RMSE* and *MAE*. However, for Darling River site, IIS-W-ANN was seen to produce the best results with *db4* wavelet, having 4 levels of decomposition.

In fact, performances with different mother wavelet and decomposition levels was significantly disparate. Similarly, although having lower performance level than the IIS-W-ANN model, the IIS-W-M5 Tree model also registered varying results with different combinations of the vanishing moment and decomposition levels of *db* mother wavelets. These results were in accordance with previous studies (Rathinasamy et al., 2013; Rathinasamy et al., 2014) that noted no universal mother wavelet can generate the most accurate forecast. It is construed that the decomposition level and choice of the analyzing wavelet remains an open problem of interest and should form the subject of an independent follow-up study.

5. Further discussion

In this paper, the preciseness of the iterative input selection (IIS) optimized wavelet coupled ANN model (benchmarked with M5 Tree model) was investigated for streamflow forecasting in a drought-prone region. ANN was a multilayer feed-forward perceptron model with one 'hidden' layer consisting of between 1 and 40 prescribed hidden nodes were utilized to examine the input data features in order to optimize the model architecture. Analysis of results showed that ANN outperformed the M5 Tree model for all tested hydrological sites, revealing that the model was efficient in extraction of features within hydro-meteorological inputs in a physically meaningful way to forecast the streamflow data.

The ANN model being a pattern recognizing algorithm is able to extract vital information from hydro-meteorological variables using non-linearly connected elements (i.e., neurons) and subsequently, the model simulates the stochastic and complex hydrological system to generate forecasts. The benefit is that the ANN model is simple to develop and does not require in-depth knowledge of the internal physical structure of the data. A merit of ANN is its ability to generalize any linear or nonlinear system without being constrained to a specific form. ANN has the capability to simultaneously weave through large volumes of data, providing prospects for parallel implementation. Owing to the distributed processing within the network, ANN is able to model input data with embedded noise and measurement errors without severe loss

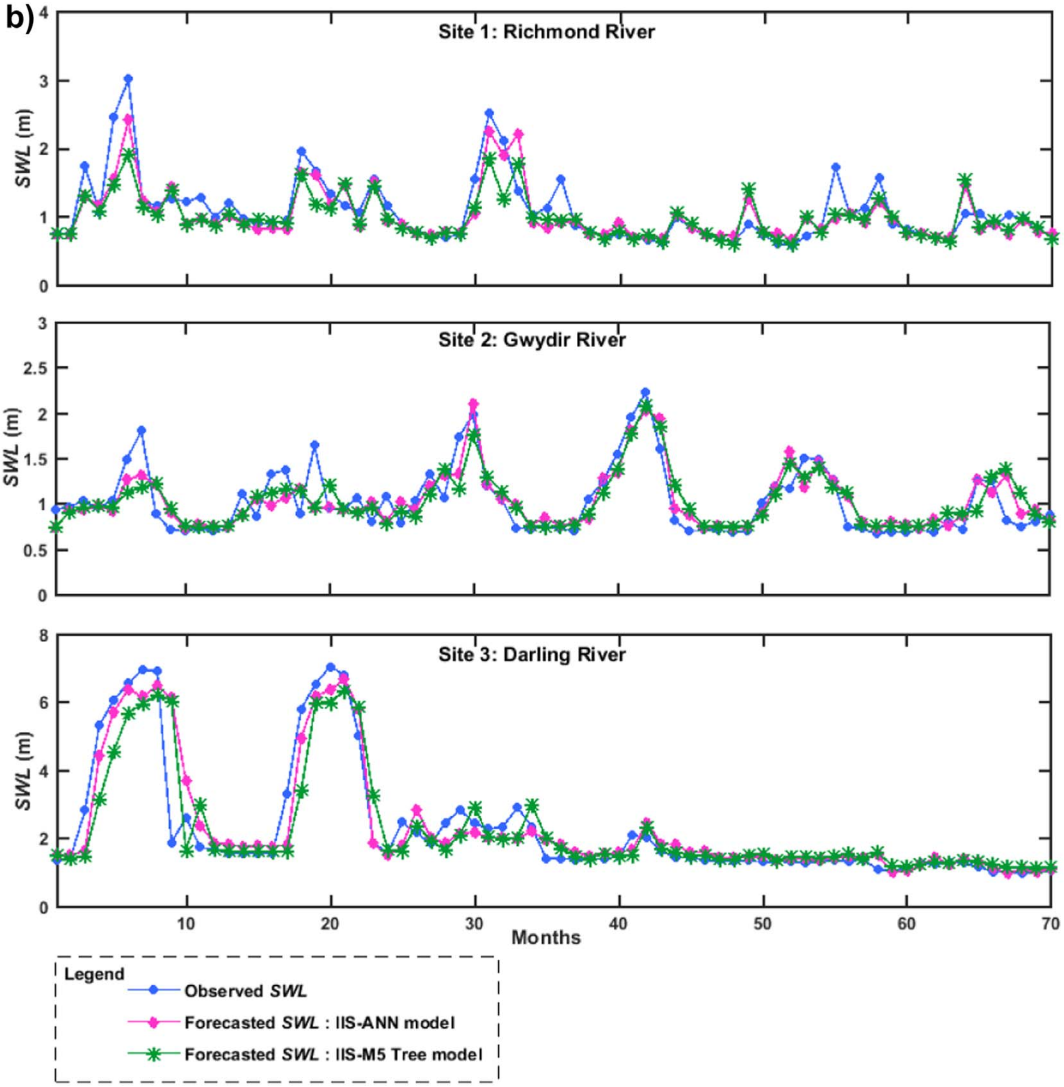


Fig. 8. (*continued*)

of accuracy (ASCE Task Committee on Application of ANN in Hydrology, 2000a; Xiong and O'Connor, 2002; Yilmaz et al., 2011).

The robustness is evident as ANN does not assume any probability distribution like normality or equal dispersion and covariance matrix requirements (Moghaddamnia et al., 2009). With that during the generalization, the transfer functions or activation functions may limit model's sensitivity (Demirel et al., 2015) based on the objective function, allowing an ANN model to react to a certain range of inputs and this feature enables it to generate lower error values (Table 4). ANN provides liberty to select the number of hidden layers and the

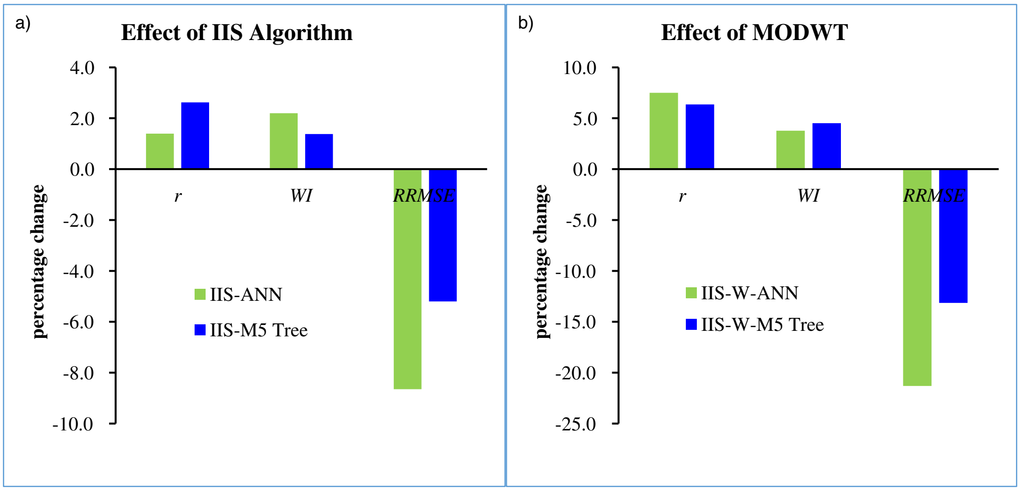


Fig. 9. Effect of (a) input selection procedure (IIS algorithm) and (b) multi-resolution analysis (MODWT) on the performance of ANN and M5 Tree models. Note: The IIS algorithm was applied to ALL variables while MODWT was applied to IIS-selected variables. (NB: r and WI are unitless, while RRMSE values are in percentage (%)).

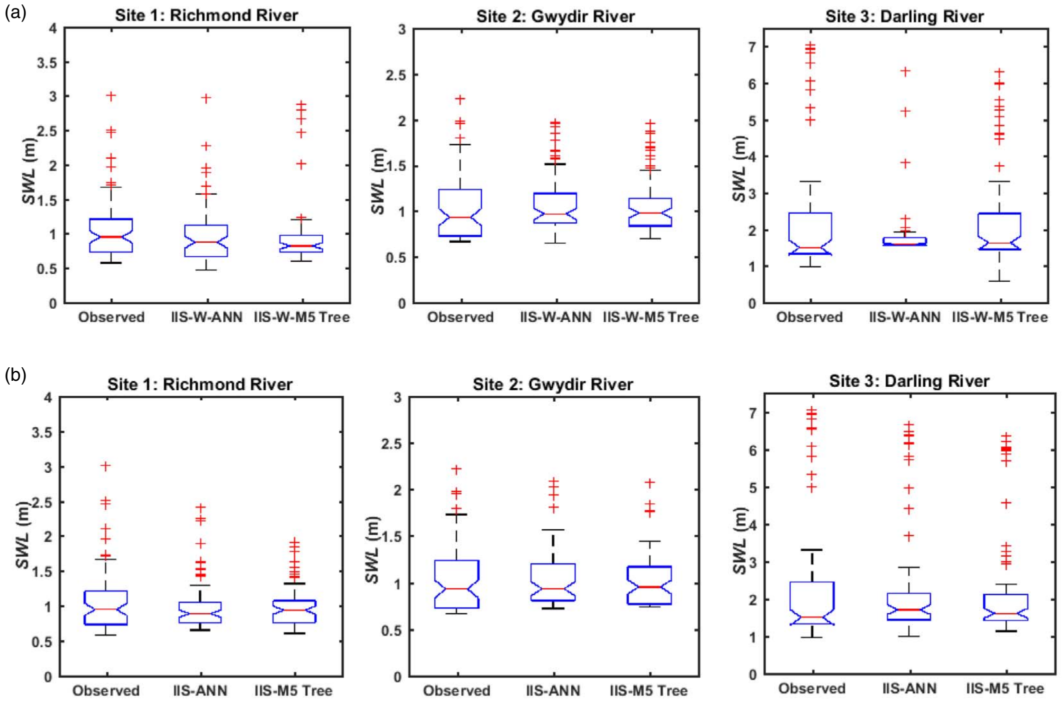


Fig. 10. Box-plots of observed compared with forecasted streamflow water level, *SWL*: (a) models integrated with IIS and MODWT, (b) models with IIS (without wavelet).

Table 6
SWL forecasting performance of different wavelets and decomposition levels at the selected sites. Mother wavelets and decomposition levels with best performances has been **bold** faced.

Wavelet type	Decomposition level	ANN					M5 Tree				
		r	*WI*	E_{NS}	*RMSE* (m)	*MAE* (m)	*r*	*WI*	E_{NS}	*RMSE* (m)	*MAE* (m)
SITE: 1 Richmond River											
db3	3	0.878	0.912	0.702	0.260	0.201	0.805	0.867	0.587	0.406	0.225
db3	4	**0.910**	**0.936**	**0.770**	**0.229**	**0.177**	0.789	0.849	0.521	0.329	0.226
db4	3	0.770	0.872	0.509	0.333	0.242	0.800	0.879	0.562	0.315	0.232
db4	4	0.763	0.851	0.528	0.327	0.246	**0.822**	**0.892**	**0.578**	**0.309**	**0.225**
SITE: 2 Gwydir River											
db3	3	0.898	0.943	0.793	0.166	0.138	**0.882**	**0.929**	**0.769**	**0.176**	**0.147**
db3	4	**0.908**	**0.944**	**0.803**	**0.162**	**0.139**	0.881	0.922	0.750	0.183	0.155
db4	3	0.809	0.879	0.650	0.216	0.175	0.739	0.824	0.537	0.249	0.195
db4	4	0.823	0.883	0.645	0.218	0.178	0.726	0.820	0.511	0.256	0.191
SITE: 3 Darling River											
db3	3	0.958	0.978	0.916	0.502	0.391	0.795	0.766	0.511	1.207	0.738
db3	4	0.957	0.977	0.914	0.508	0.386	0.808	0.816	0.578	1.122	0.684
db4	3	0.932	0.964	0.863	0.639	0.465	0.914	0.950	0.832	0.709	0.485
db4	4	**0.960**	**0.979**	**0.920**	**0.487**	**0.390**	**0.922**	**0.952**	**0.843**	**0.684**	**0.501**

associated nodes in each of these layers giving an added versatility (ASCE Task Committee on Application of ANN in Hydrology, 2000a; Xiong and O'Connor, 2002; Yilmaz et al., 2011).

In addition to the superiority of the ANN over M5 Tree model, the present results demonstrated the importance of the IIS algorithm as a data pre-processing (i.e., a suitable feature screening) tool, where performance metrics for IIS optimized (IIS-ANN and IIS-M5 Tree) model (Table 4; Table 5) were noticeably better than their standalone counterparts. The results of our study accede the deduction of earlier work by Galelli and Castelletti (2013a) that found the IIS to be a significantly useful tool for non-redundant input selection skill in different test conditions (e.g., the presence of noise or presence of several redundant variables). The first important understanding obtained from the results is that a proper feature selection should be carefully carried out prior to executing data-driven models, as redundant data can have a great influence on the 'learning' process and eventually, can affect the forecasting accuracy.

With that, carefully selected and the most relevant input variables also means fewer input weights within the input layer which can provide a greater confidence that an overtraining may not happen, leading to a parsimonious and computationally efficient ANN model. IIS was able to demonstrate that the different sets of best predictor data for each site. For example, the one-month antecedent streamflow water level (SWL^1) was useful for all the hydrological sites, but additionally, the zero-lagged precipitation (PCN^0) and surprisingly, the zero-lagged evaporation ($EVAP^0$) which moderates *SWL* was useful for forecasting *SWL* for Gwydir River. For Darling River site, the IIS algorithm depicted one and two months lagged streamflow water level (i.e., SWL^1, SWL^2) as the optimal predictor variables. In fact, the utilization of globally pooled predictor data without an application of IIS led to significantly poor performances, outlaying the importance of input screening for optimal performance of streamflow forecasting models. The interesting finding here is also that there were unique combinations of input variables responsible for prediction of streamflow at the three hydrological sites.

Other than integrating the ANN and M5 Tree models with IIS algorithm, a further enhancement in model accuracy was attained via incorporation of wavelet transform of IIS selected predictor signals that led to the IIS-W-ANN (and IIS-W-M5 Tree) models. The main purpose of wavelet transformation is to identify and isolate the embedded deterministic components of hydro-meteorological time series data and provide a reliable physical basis to the machine learning model to overcome deleterious effects of distinctive topographical conditions within the watershed (even with similar climatic conditions) (Nourani et al., 2014). Wavelet transformation has been deemed imperative in hydrological forecasting (Adamowski and Chan, 2011; Badrzadeh et al., 2016; Cannas et al., 2006; Galelli and Castelletti, 2013a; Maheswaran and Khosa, 2012; Nourani et al., 2011; Okkan, 2012). In this study, maximum overlap discrete wavelet transformation (MODWT) has been implemented for data pre-processing which can effectively diagnose the signal's main frequency components and the abstract local information without any loss of information.

MODWT application as a multi-resolution analysis utility was a major advancement to enhance the forecasting capability of an ANN model. In accordance with the results of other investigations whereby applications of wavelet transformation improved hydrological forecasting e.g. (Badrzadeh et al., 2016; Cannas et al., 2006; Kim and Valdes, 2003; Krishna et al., 2011; Santos et al., 2014; Wang and Ding, 2003), these results clearly showed that the wavelet-based ANN model, the IIS-W-ANN, provided better prediction estimations at all stations in comparison to the corresponding standalone ANN, M5 Tree, and hybrid IIS-W-M5 Tree models as the largest values of correlation coefficient (*r*), *WI*, E_{NS} and lower error values, *RMSE* and *MAE* were registered (Table 4). It is clearly evident that the MODWT decomposition of input data provided greater insights into the physical process (particularly revealing the frequencies therein) (Daubechies, 1990) enabled the learning algorithm in the ANN responsible for mapping the predictors to the predictand (SWL) to effectively capture the deterministic components at various resolution levels, consequently resulted in swift convergence, negligible errors and the apparent improved model performance.

The appropriate number of wavelet decomposition levels, on the other hand, needs to be selected with caution, as using the conventional method, the number of decomposition levels were approximated to be 3 which is dependent on the length of the data series. However, the results obtained in this study are contradictory (Table 6), as db3 with four levels of decomposition yielded best performances by the IIS-W-ANN at Richmond and Gwydir rivers. This shows that all the embedded deterministic features are not clearly revealed to the models using three levels of decompositions. Therefore, it is important to carefully determine the apt wavelet decomposition levels based on the characteristics of the original time series and not solely based on the series length, since unreasonable decomposition of the original series would not provide all pertinent information to the model leading to poor performance of wavelet coupled model.

The success of data-driven models (including ANN and M5 tree) were largely dependent on the quantity and quality of historically measured data for training and validation which ultimately is the basis of their existence. So, to reduce the inadvertent introduction of disastrous biases and common database errors plaguing practical applications (Witten et al., 2011), authentic and reliable meteorological data from SILO were used. However, initial instrumental recording errors in time-series during the time of observation of respective variables (e.g. *PCN*, $T_{\max}$, T_{min}, *Evap*, and *VP*) could have introduced uncertainties in the predictor variable itself. Prior data quality control and the cleansing process needs to be performed on inputs to eliminate these biases and to prevent incorrect conclusions and/or recommendations (Deo et al., 2016e). The data length used for hydrological forecasting could also be a limitation as a shorter period will provide insufficient information, while too long a period could feed in unnecessary information, as such, a rational 40 years of data has been utilized. Since, an ANN is not a routing model, for real-time applications, testing with smaller time-steps such as weekly, daily, and hourly can provide greater understanding that high, moderate and low flow events could be explored independently with shorter time intervals and are recommended in further studies.

6. Conclusion

In this study of monthly streamflow forecasting at three candidate stations (Richmond, Gwydir, and Darling River) within Australia's Murray-Darling Basin, the iterative input selection (IIS) algorithm was first applied to select model's input data, which were later decomposed using maximum overlap discrete wavelet transformation (MODWT). Consequently, the MODWT-decomposed sub-series assisted in the extraction of low and high-frequency fluctuations and trends in the historical data before being utilized as model's input variables while generating a hybrid IIS-W-ANN model. Benchmarking of the model against M5 Tree revealed that apparently, the non-linear framework of ANN provided an edge over the piecewise-linear-functions combinations used in M5 Tree since ANN outperformed M5 Tree applied at all the hydrological sites.

To demonstrate the usefulness of input selection for optimisation of ANN-based streamflow forecasts, initially, a global set of predictor variables was constructed based on a statistically significant lagged combinations of streamflow water level, accompanied by primary meteorological variables related to streamflow, including precipitation, maximum and minimum temperature, solar radiation, vapor pressure and evaporation. Next, the iterative input selection (IIS) algorithm was applied to sieve out the best combination of predictor variables to safeguard the ANN model against input data redundancy. The model's performance metrics demonstrated that the IIS algorithm was a suitable tool for feature selection, as the IIS-optimized model had better performance in comparison to the non-IIS standalone models, with higher values of *WI* and E_{NS}, and lower errors (*RMSE*, *MAE*). The significant variables selected for Richmond River were: PCN^0 and SWL^1, while for Gwydir River, three inputs were selected: SWL^1, $EVAP^0$, and PCN^0 and for Darling River, two significant lags of antecedent streamflow water level, SWL^1, and SWL^2, were selected in developing the hybrid ANN and M5 Tree models.

As clearly verified, the application of non-decimated wavelet transform, MODWT, aimed to enhance the ANN model, by allowing it to cope with the non-stationarity and seasonality features within the input datasets. The resulting IIS-W-ANN outperformed the standalone ANN, IIS-ANN, and tree based M5 Tree, IIS-M5 Tree and hybrid IIS-W-M5 Tree models. Particularly, in comparison of the IIS-W-ANN with IIS-ANN, there was an overall increase of 7.5% and 3.8% in *r* and *WI,* respectively and a decreased by 21.3% in *RMSE*. The best performance indices recorded by IIS-W-ANN models were in the range: $r = 0.908–0.960$, $WI = 0.936–0.979$ and $E_{\mathrm{NS}} = 0.770–0.920$. Congruent with these metrics, *RMSE* and *MAE* values alternated from 0.162–0.487 m and 0.139–0.390 m, respectively. Coefficient of determination (R^2) in scatterplots registered values of 0.825 to 0.921, to confirm that the IIS-W-ANN recorded better forecasting performance at all hydrological sites. Comparison of boxplots and whisker diagrams illustrated that the distribution of the streamflow forecasted by IIS-W-ANN model and observed streamflow water level were comparable and the median values were relatively close. Further validation of the IIS-W-ANN model was provided by the lowest *RRMSE* values (15.65%–21.00%) registered for all hydrological sites. Comparison of relative errors revealed that the IIS-W-ANN model applied at Gwydir River yielded a value of $RRMSE = 15.65\%$ & $MAPE = 14.79\%$, including a promising performance for Richmond River and Darling River sites.

This study advocated that the IIS coupled wavelet-ANN model (IIS-W-ANN) can be a meritorious scientific tool for forecasting streamflow water level, however, the determination of the best mother wavelet that presents an optimal decomposition level still needed an in-depth investigation in hydrological problems. This is because mother wavelets led to different decomposition level and hence, produced disparate results at the candidate sites. Nevertheless, this study supports ongoing research on investigations of purely data-driven models with non-decimated maximum overlap discrete wavelet transform and input selection algorithms to better describe the behavior of stochastic variables used to predict hydrological variables such as streamflow water level.

While our study has clearly stipulated the superiority of input selection, optimization of ANN-based forecasts with iterative input selection (IIS) and maximum overlap discrete wavelet transform (MODWT) including a robust model identification process via combinations of training algorithms and hidden transfer functions (Section 3.2–3.3), improvements can be introduced by using add-on optimiser algorithms (e.g. particle swarm optimisation, PSO or firefly optimiser algorithm, FFA) (Chau, 2006; Emary et al., 2015; Kumar et al., 2013; Olatomiwa et al., 2015; Sedki and Ouazar, 2010). If optimisers are embedded in IIS-W-ANN, they may assist in fine-tuning the hidden layer weights and biases to attain neuronal architectures and forecasts. Finally, the evaluation of forecasts at smaller time steps (e.g. daily or hourly) is also necessary for operational use of ANN model (including its verification at other sites) which could form the subject of a subsequent independent study.

Acknowledgment

The data were acquired from NSW Department of Primary Industries Data Portal, Scientific Information for Land Owners (SILO) Portal developed by Queensland Department of Environment and Resource Management and the Australian Bureau of Meteorology. This study was supported by Office of Research and Graduate Studies University of Southern Queensland Postgraduate Research Scholarship (USQ-PRS) Grant # 2016 to the first author. Corresponding author Dr. R C Deo also thanks the support of USQ Academic Division and Outside Studies program grant (ADOSP, 2016). We thank both reviewers and the Editor-in-Chief for their constructive criticisms that has improved the clarity of the paper.

References

Abbot, J., Marohasy, J., 2012. Application of artificial neural networks to rainfall forecasting in Queensland, Australia. Adv. Atmos. Sci. 29, 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., Chan, H.F., 2011. A wavelet neural network conjunction model for groundwater level forecasting. J. Hydrol. 407, 28–40.

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.

Alvisi, S., Mascellani, G., Franchini, M., B'ardossy, A., 2006. Water level forecasting through fuzzy logic and artificial neural network approaches. Hydrol. Earth Syst. Sci. 10, 1–17.

ASCE, 2000a. Task committee on application of ANN in hydrology, artificial neural networks in hydrology-I-preliminary concepts. J. Hydrol. Eng. 5, 115–123.

Avriel, M., 2003. Nonlinear Programming: Analysis and Methods. Courier Corporation.

Australian Bureau of Statistics, 2008. Water and the Murray-Darling Basin - A Statistical Profile, 2000-01 to 2005-06. 23 May, 15/08/2008, < http://www.abs.gov.au/ausstats/abs@.nsf/mf/4610.0.55.007 > .

Badrzadeh, H., Sarukkalige, R., Jayawardena, A.W., 2016. Improving Ann-based short-term and long-term seasonal river flow forecasting with signal processing techniques. River Res. Appl. 32, 245–256.

Beesley, C., Frost, A., Zajaczkowski, J., 2009. A comparison of the BAWAP and SILO spatially interpolated daily rainfall datasets. In: 18th World IMACS/MODSIM Congress. Cairns, Australia, pp. 17.

Bhattacharya, B., Solomatine, D.P., 2003. Neural networks and M5 model trees in modelling water level discharge relationship for an Indian River. In: European Symposium on Artificial Neural Networks. ESANN'2003 Proceedingspp. 407–412 Bruges (Belgium).

Bhattacharya, B., Solomatine, D.P., 2005. Neural networks and M5 model trees in modelling water level–discharge relationship. Neurocomputing 63, 381–396.

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005. Input determination for neural network models in water resources applications. Part 1—background and methodology. J. Hydrol. 301, 75–92.

Cameron, D., Kneale, P., See, L., 2002. An evaluation of a traditional and a neural net modelling approach to flood forecasting for an upland catchment. Hydrol. Process. 16, 1033–1046.

Campolo, M., Soldati, A., Andreussi, P., 2003. Artificial neural network approach to flood forecasting in the River Arno. Hydrol. Sci. J. 48, 381–398.

Cannas, B., Fanni, A., See, L., Sias, G., 2006. Data preprocessing for river flow forecasting using neural networks: wavelet transforms and data partitioning. Physics and Chemistry of the Earth, Parts A/B/C 31, 1164–1171.

Chai, T., Draxler, R.R., 2014a. Root mean square error (RMSE) or mean absolute error (MAE)?—arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247–1250.

Chai, T., Draxler, R.R., 2014b. Root mean square error (RMSE) or mean absolute error (MAE)? — arguments against avoiding RMSE in the literature. Geosci. Model Dev. 7, 1247–1250.

Chau, K.W., 2006. Particle swarm optimization training algorithm for ANNs in stage prediction of Shing Mun River. J. Hydrol. 329, 363–367.

Chau, K.W., 2007. A split-step particle swarm optimization algorithm in river stage forecasting. J. Hydrol. 346, 131–135.

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, 138–149.

Cornish, C.R., Bretherton, S., Percival, D.B., 2005. Maximal OverlapWavelet Statistical Analysis With Application to Atmospheric Turbulence. Kluwer Academic Publishers, Netherlands.

Cottrill, A., Haendon, H.H., Lim, E.P., Langford, S., Kuleshov, Y., Charles, A., Jones, D., 2012. Seasonal Climate Prediction in the Pacific using the POAMA coupled model forecast. The Centre for Australian Weather and Climate Research (CAWCR) Technical Report: No. 048.

CSIRO and Bureau of Meteorology, 2015. Climate Change in Australia Information for Australia's Natural Resource Management Regions: Technical Report. CSIRO and Bureau of Meteorology, Australia.

Daubechies, I., 1990. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Inf. Theor. 36, 961–1005.

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.

Demirel, M.C., Booij, M.J., Hoekstra, A.Y., 2015. The skill of seasonal ensemble low-flow forecasts in the Moselle River for three different hydrological models. Hydrol. Earth Syst. Sci. 19, 275–291. http://dx.doi.org/10.5194/hess-19-275-2015.

Dennis Jr., J.E., Schnabel, R.B., 1996. Numerical Methods for Unconstrained Optimization and Nonlinear Equations, Siam.

Deo, R.C., Downs, N., Parisi, A., Adamowski, J., Quilty, J., 2017a. 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., 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., Byun, H.-R., Adamowski, J.F., Kim, D.-W., 2015b. A real-time flood monitoring index based on daily effective precipitation and its application to Brisbane and Lockyer Valley flood events. Water Resour. Manag. 1–19. http://dx.doi.org/10.1007/s11269-11015-11046-11263.

Deo, R.C., Kisi, O., Singh, V.P., 2017b. 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., 2016. An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland. Environ. Monit. Assess. 188, 90. http://dx.doi.org/10.1007/s10661-016-5094-9.

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. Sust. Energ. Rev. 72, 828–848.

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., Tiwari, M.K., Adamowski, J.F., Quilty, M.J., 2016b. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. Stoch. Env. Res. Risk A. 1–30.

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.

Dghais, A.A.A., Ismail, M.T., 2013. A comparative study between discrete wavelet transform and maximal overlap discrete wavelet transform for testing stationarity. International Journal of Mathematical, Computational, Physical, Electrical and Computer Engineering 12, 1677–1681.

Emary, E., Zawbaa, H.M., Ghany, K.K.A., Hassanien, A.E., Parv, B., 2015. Firefly optimization algorithm for feature selection. In: Proceedings of the 7th Balkan Conference on Informatics Conference. ACM, pp. 26.

Fazel, S.A.A., Mirfenderesk, H., Tomlinson, R., 2014. Application of neural network to flood forecasting an examination of model sensitivity to rainfall assumptions. In: Ames, D.P., Quinn, N.W.T., Rizzoli, A.E. (Eds.), 7th Intl. Congress on Env. Modelling and Software. International Environmental Modelling and Software Society (iEMSs), San Diego, CA, USA.

Galelli, S., Castelletti, A., 2013a. Tree-based iterative input variable selection for hydrological modeling. Water Resour. Res. 49, 4295–4310.

Galelli, S., Castelletti, A., 2013b. Tree-based iterative input variable selection for hydrological modeling. Water Resour. Res. 49, 4295–4310.

Galelli, S., et al., 2014. An evaluation framework for input variable selection algorithms for environmental data-driven models. Environ. Model. Softw. 62, 33–51.

George, E.I., 2000. The variable selection problem. J. Am. Stat. Assoc. 95, 1304–1308.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Mach. Learn. 63, 3–42.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. 377, 80–91.

HariKumar, R., Vasanthi, N., Balasubramani, M., 2009. Performance analysis of artificial neural networks and statistical methods in classification of oral and breast cancer stages. International Journal of Soft Computing and Engineering (IJSCE) 2.

Hejazi, M.I., Cai, X., 2009. Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm. Adv. Water Resour. 32, 582–593.

Helman, P., 2009. Droughts in the Murrary-Darling Basin Since European Settlement. Griffith Centre for Coastal Management Research Report No 100, Licensed From the Murray-Darling Basin Authority Under a Creative Commons Attribution 3.0 Australia License.

Hora, J., Campos, P., 2015. A review of performance criteria to validate simulation models. Expert. Syst. 32, 578–595.

Huang, H.-Y., 1970. Unified approach to quadratically convergent algorithms for function minimization. J. Optim. Theory Appl. 5, 405–423.

Humphrey, G.B., Gibbs, M.S., Dandy, G.C., Maier, H.R., 2016. A hybrid approach to monthly streamflow forecasting: integrating hydrological model outputs into a Bayesian artificial neural network. J. Hydrol. 540, 623–640.

IPCC, Pachauri, R.K., Meyer, L.A., 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 [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]. Geneva, Switzerland pp. 151.

Jain, A., Srinivasulu, S., 2004. Development of effective and efficient rainfall-runoff models using integration of deterministic, real-coded genetic algorithms and artificial neural network techniques. Water Resour. Res. 40, 1–12.

Jain, S.K., Das, A., Srivastava, D.K., 1999. Application of ANN for reservoir inflow prediction and operation. J. Water Resour. Plan. Manag. 125, 263–271.

Jeffrey, S.J., Carter, J.O., Moodie, K.B., Beswick, A.R., 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environ. Model. Softw. 16, 309–330.

Jekabsons, G., 2010. M5PrimeLab: M5' Regression Tree and Model Tree Toolbox for Matlab/Octave.

Karunanithi, N., Grenney, W.J., Whitley, D., Bovee, K., 1994. Neural networks for river flow prediction. J. Comput. Civ. Eng. 8, 201–220.

Kim, T.-W., Valdes, J.B., 2003. Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. J. Hydrol. Eng. 8, 319–328.

Kim, T.-W., Valdés, J.B., 2003. Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. J. Hydrol. Eng. 8, 319–328.

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.

Krause, P., Boyle, D., Bäse, F., 2005a. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97.

Krause, P., Boyle, D.P., Base, F., 2005b. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97.

Krishna, B., Rao, Y.R.S., Nayak, P.C., 2011. Time series modeling of river flow using wavelet neural networks. Journal of Water Resource and Protection 03, 50–59.

Kumar, D., Prasad, R.K., Mathur, S., 2013. Optimal design of an in-situ bioremediation system using support vector machine and particle swarm optimization. J. Contam. Hydrol. 151, 105–116.

Legates, D.R., McCabe, G.J., 1999a. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241.

Legates, D.R., McCabe, G.J., 1999b. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241.

Liong, S.-Y., Lim, W.-H., Kojiri, T., Hori, T., 2000. Advance flood forecasting for flood stricken Bangladesh with a fuzzy reasoning method. Hydrol. Process. 14, 431–448.

Liong, S.-Y., Sivapragasam, C., 2002. Flood stage forecasting with support vector machines. J. Am. Water Resour. Assoc. 38, 173–186.

Londhe, S.N., Dixit, P.R., 2012. Forecasting stream flow using support vector regression and M5 model trees. International Journal of Engineering Research and Development 2, 1–12.

López, G., Batlles, F., Tovar-Pescador, J., 2005. Selection of input parameters to model direct solar irradiance by using artificial neural networks. Energy 30, 1675–1684.

Maheswaran, R., Khosa, R., 2012. Comparative study of different wavelets for hydrologic forecasting. Comput. Geosci. 46, 284–295.

Maheswaran, R., Khosa, R., 2013. Long term forecasting of groundwater levels with evidence of non-stationary and nonlinear characteristics. Comput. Geosci. 52, 422–436.

Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environ. Model. Softw. 15, 101–124.

Maier, H.R., Jain, A., Dandy, G.C., 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. Environ. Model. Softw. 25, 891–909.

Marquardt, D.W., 1963. An algorithm for least-squares estimation of nonlinear parameters. Journal of the Society for Industrial & Applied Mathematics 11, 431–441.

McBride, J.L., Nicholls, N., 1983. Seasonal relationships between Australian rainfall and the Southern oscillation. Mon. Weather Rev. 111, 1998–2004.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133.

Mehr, A.D., Kahya, E., Şahin, A., Nazemosadat, M.J., 2014. Successive station monthly streamflow prediction using different artificial neural network algorithms. Int. J. Environ. Sci. Technol. 12, 2191–2200.

Mishra, A.K., Singh, V.P., 2011. Drought modeling — a review. J. Hydrol. 403, 157–175.

Moghaddamnia, A., Gousheh, M.G., Piri, J., Amin, S., Han, D., 2009. Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques. Adv. Water Res. 32, 88–97.

Moustris, K.P., Larissi, I.K., Nastos, P.T., Paliatsos, A.G., 2011. Precipitation forecast using artificial neural networks in specific regions of Greece. Water Resour. Manag. 25, 1979–1993.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — a discussion of principles. J. Hydrol. 10, 282–290.

Ni, Q., Wang, L., Ye, R., Yang, F., Sivakumar, M., 2010. Evolutionary modeling for streamflow forecasting with minimal datasets: a case study in the west Malian River, China. Environ. Eng. Sci. 27, 377–385.

Nourani, V., Hosseini Baghanam, A., Adamowski, J., Kisi, O., 2014. Applications of hybrid wavelet–artificial intelligence models in hydrology: a review. J. Hydrol. 514, 358–377.

Nourani, V., Kisi, Ö., Komasi, M., 2011. Two hybrid artificial intelligence approaches for modeling rainfall–runoff process. J. Hydrol. 402, 41–59.

Okkan, U., 2012. Wavelet neural network model for reservoir inflow prediction. Scientia Iranica 19, 1445–1455.

Olatomiwa, L., et al., 2015. A support vector machine–firefly algorithm-based model for global solar radiation prediction. Sol. Energy 115, 632–644.

Onyari, E.K., Ilunga, F.M., 2013. Application of MLP neural network and M5P model tree in predicting Streamflow_South Africa. International Journal of Innovation, Management and Technology 4, 11–15.

Long-memory processes, the Allan variance and wavelets. In: Percival, D.B., Guttorp, P., Georgiou, E.F., Kumar, P. (Eds.), Wavelets in Geophysics. Academic, San Diego, pp. 325–344.

Percival, D.B., Lennox, S.M., Wang, Y.G., Darnell, R.E., 2011. Wavelet-based multiresolution analysis of Wivenhoe Dam water temperatures. Water Resour. Res. 47.

Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis. Cambridge University Press, UK.

Pereira Filho, A.J., dos Santos, C.C., 2006. Modeling a densely urbanized watershed with an artificial neural network, weather radar and telemetric data. J. Hydrol. 317, 31–48.

Phien, H.N., Kha, N.D.A., 2003. Flood Forecasting for the Upper Reach of the Red River Basin North Vietnam. 29. pp. 267–272.

Quilty, J., Adamowski, J., Khalil, B., Rathinasamy, M., 2016. Bootstrap Rank-Ordered Conditional Mutual Information (broCMI)—A Nonlinear Input Variable Selection Method for Water Resources Modeling. Water Resour. Res..

Quinlan, J.R., 1992. Learning with continuous classes. In: Sterling, A. (Ed.), 5th Australian Joint Conference on Artificial Intelligence, Singapore, pp. 343–348.

Rahimikhoob, A., 2014. Comparison between M5 model tree and neural networks for estimating reference evapotranspiration in an arid environment. Water Resour. Manag. 28, 657–669.

Rathinasamy, M., Adamowski, J., Khosa, R., 2013. Multiscale streamflow forecasting using a new Bayesian model average based ensemble multi-wavelet Volterra nonlinear method. J. Hydrol. 507, 186–200.

Rathinasamy, M., et al., 2014. Wavelet-based multiscale performance analysis: an approach to assess and improve hydrological models. Water Resour. Res. 50, 9721–9737.

Sajikumara, N., Thandaveswarab, B.S., 1999. A non-linear rainfall–runoff model using an artificial neural network. J. Hydrol. 216, 32–55.

Santos, C.A.G., Freire, P.K.M.M., Silva, G.B.L., Silva, R.M., 2014. Discrete wavelet transform coupled with ANN for daily discharge forecasting into Três Marias reservoir. In: Proceedings of the International Association of Hydrological Sciences. 364. pp. 100–105.

Sattari, M.T., Pal, M., Apaydin, H., Ozturk, F., 2013. M5 model tree application in daily river flow forecasting in Sohu stream, Turkey. Water Res. 40, 233–242.

Sedki, A., Ouazar, D., 2010. Hybrid particle swarm and neural network approach for streamflow forecasting. Mathematical Modelling of Natural Phenomena 5, 132–138.

Sehgal, V., Tiwari, M.K., Chatterjee, C., 2014. Wavelet bootstrap multiple linear regression based hybrid modeling for daily river discharge forecasting. Water Resour. Manag. 28, 2793–2811.

Shamseldin, A.Y., 1997. Application of a neural network technique to rainfall runoff. J. Hydrol. 199, 272–294.

Shensa, M.J., 1992. The discrete wavelet transform: wedding the a trous and Mallat algorithms. IEEE Trans. Signal Process. 40, 2464–2482.

Solomatine, D.P., Dulal, K.N., 2003. Model trees as an alternative to neural networks in rainfall—runoff modelling. Hydrol. Sci. J. 48, 399–411.

Solomatine, D.P., Siek, M.B.L.A., 2004. Flexible and optimal m5 model trees with applications to flow predictions. In: Liong, P.B. (Ed.), 6th International Conference on Hydroinformatics. World Scientific Publishing Company.

Solomatine, D.P., Xue, Y., 2004. M5 model trees and neural networks application to flood forecasting in the upper reach of the Huai River in China. J. Hydrol. Eng. 9, 491–501.

Steinbuch, I.M., Molengraft, M.J.G.V.D., 2005. Wavelet theory and applications. In: A Literature Study.

Taormina, R., Chau, K.-W., 2015a. ANN-based interval forecasting of streamflow discharges\using the LUBE method and MOFIPS. Eng. Appl. Artif. Intell. 45, 429–440.

Taormina, R., Chau, K.-W., 2015b. Data-driven input variable selection for rainfall-runoff modeling using binary-coded particle swarm optimization and extreme learning machines. J. Hydrol.

Tiwari, M.K., Adamowski, J., 2013. Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models. Water Resour. Res. 49, 6486–6507.

Tiwari, M.K., Chatterjee, C., 2011. A new wavelet-bootstrap-ANN hybrid model for daily discharge forecasting. J. Hydroinf. 13, 500–519.

Tozer, C., Kiem, A., Verdon-Kidd, D., 2012. On the uncertainties associated with using gridded rainfall data as a proxy for observed. Hydrol. Earth Syst. Sci. 16, 1481–1499.

Vogl, T., Mangis, J., Rigler, A., Zink, W., Alkon, D., 1988. Accelerating the convergence of the backpropagation method. Biol. Cybern. 59, 257–263.

Wang, W., Ding, J., 2003. Wavelet network model and its application to the prediction of hydrology. Nat. Sci. 1, 67–71.

Wang, Y., Witten, I.H., 1997. In: Inducing model trees for continuous classes. European Conference on Machine Learning. pp. 128–137 Prague.

Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2, 184–194.

Willmott, C.J., 1984. On the evaluation of model performance in physical geography. In: Gaile, G.L., Willmott, C.J. (Eds.), Spatial Statistics and Models. Springer, pp. 443–460.

Witten, I.H., Frank, E., Hall, M.A., 2011. Data Mining — Practical Machine Learning Tools and Techniques, 3 ed. Morgan Kaufmann Publishers, United States.

Xiong, L., O'Connor, K.M., 2002. Comparison of four updating models for real-time river flow forecasting. Hydrol. Sci. J. 47, 621–639.

Yaseen, Z.M., et al., 2016a. Stream-flow forecasting using extreme learning machines: a case study in a semi-arid region in Iraq. J. Hydrol.

Yaseen, Z.M., et al., 2016b. Stream-flow forecasting using extreme learning machines: a case study in a semi-arid region in Iraq. J. Hydrol. 542, 603–614.

Yilmaz, A.G., Imteaz, M.A., Jenkins, G., 2011. Catchment flow estimation using artificial neural networks in the mountainous Euphrates Basin. J. Hydrol. 410, 134–140.

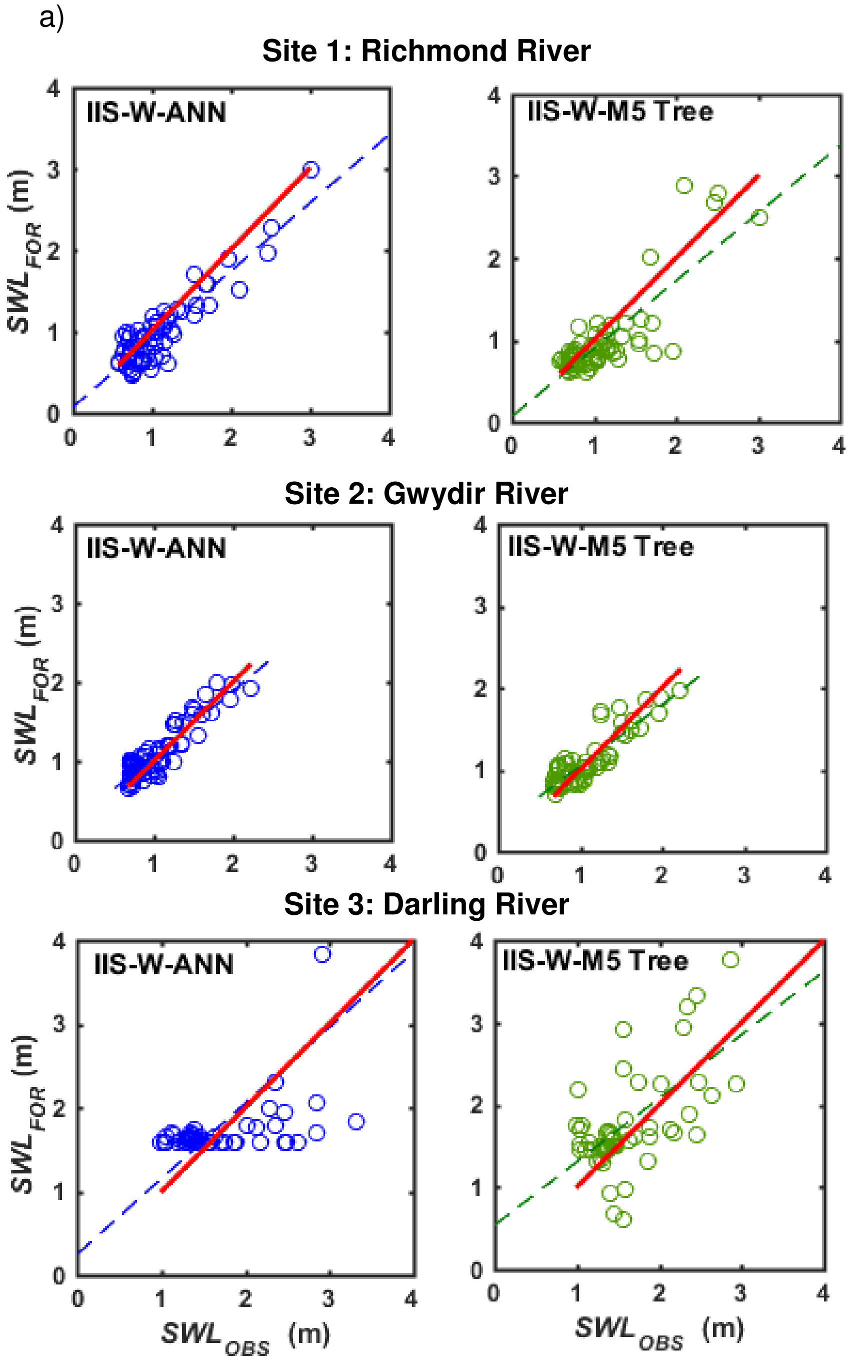
Zajaczkowski, J., Wong, K., Carter, J., 2013. Improved historical solar radiation gridded data for Australia. Environ. Model. Softw. 49, 64–77.

Zhao, M., Hendon, H.H., 2009. Representation and prediction of the Indian Ocean dipole in the POAMA seasonal forecast model. Q. J. R. Meteorol. Soc. 135 (639), 337–352. http://dx.doi.org/10.1002/qj.370.

Supplementary analysis and discussions

The scatter plots with regression lines as presented in the published paper is one way of evaluating the model. An alternative method is to draw an X = Y line [also known as the 1:1 line or the 45° line) and compute the percentage deviations of the forecasted values from this line. Figure S1 (a-b) shows the scatterplot with the 1:1 line plotted in red together with the regression line. Overall, the regression line deviates a lot from the X=Y line for the non-wavelet based models (*i.e.*, IIS-ANN and IIS-M5 Tree models) in forecasting streamflow water level. In addition, the scatterplots of wavelet-based models (*i.e.*, IIS-W-ANN and IIS-W-M5 Tree models) showed that the regression lines were closer to the 1:1 line. Particularly, at Site 2: Gwydir River and Site 3: Darling River, the IIS-W-ANN model performed very well as the 1:1 line and the regression lines were very similar in nature.

A further insight was provided by the computations of percentage deviations from the 1:1 line at all sites from all the models under considerations. Table A1 in the appendix shows the full data on percentage deviations, while Table S1, presented here, summarizes the outcomes of the percentage deviations. A comparison of the total over and under-predictions from IIS-W-ANN and IIS-W-M5 Tree showed that at Site 1-Menindee River and Site 3-Darling River the IIS-W-ANN models were clearly better with a lower number of data points being out of the 5% tolerance range. At Site 2-Gwydir, the IIS-W-ANN model slightly over-predicted (39/70 points) in comparison to the IIS-W-M5 Tree model (32/70 points). These outcomes certainly complement the results presented in the main chapter (published article in *Atmospheric Research* journal (Vol. 197, Pages 42-63)), *i.e.*, the IIS-W-ANN model has a better potential of forecasting monthly streamflow water level values at the three study sites.


a)
Site 1: Richmond River
IIS-W-ANN
IIS-W-M5 Tree
Site 2: Gwydir River
IIS-W-ANN
IIS-W-M5 Tree
Site 3: Darling River
IIS-W-ANN
IIS-W-M5 Tree
SWL_{FOR} (m)
SWL_{OBS} (m)
0
1
2
3
4

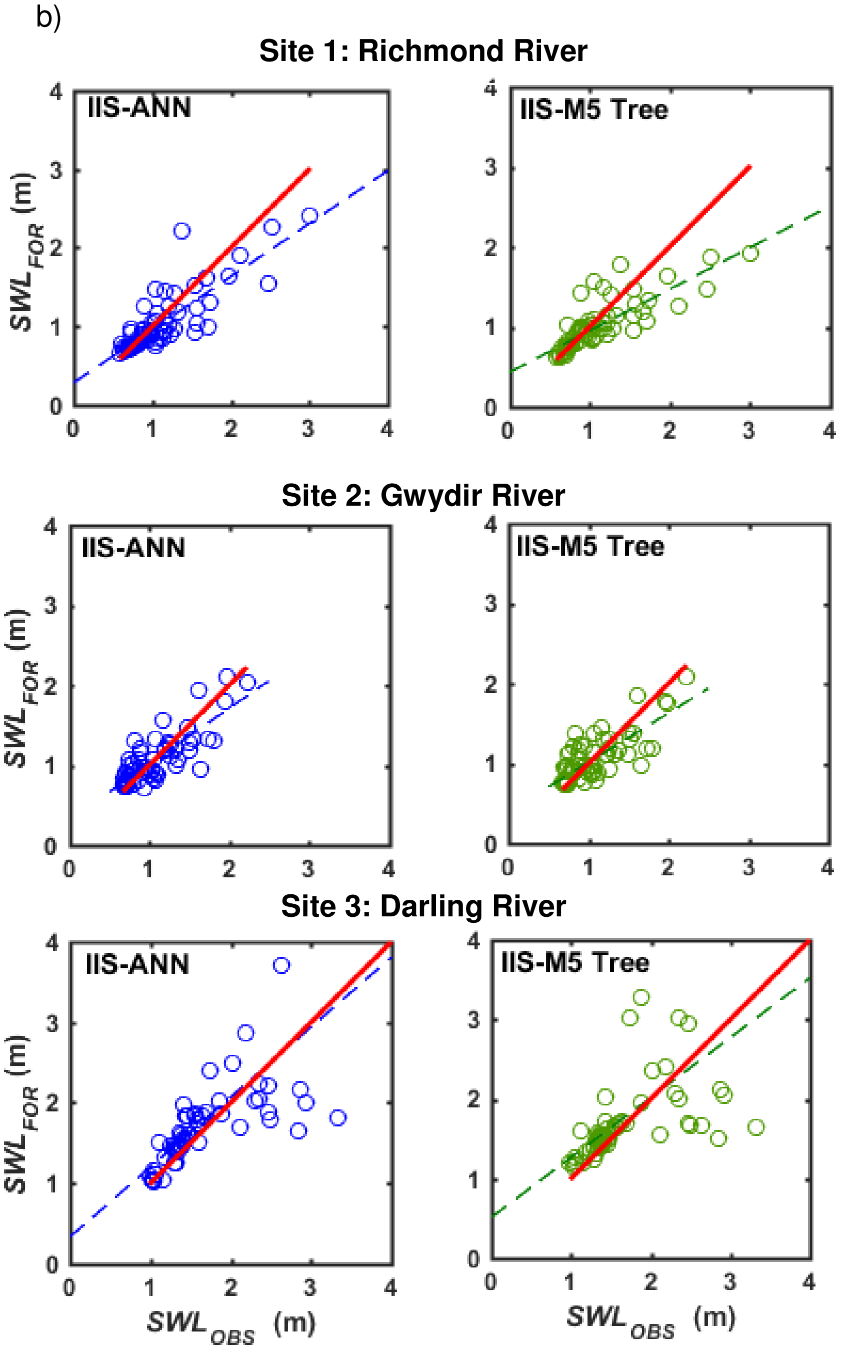


Figure S1 Scatter plots of observed (SWL_{OBS}) and forecasted (SWL_{FOR}) streamflow water level for all the stations from: a) IIS-W-ANN and IIS-W-M5 Tree models, b) IIS-ANN and IIS-M5 Tree models (No wavelet applied). (Note: The dashed line in blue and green is the least-squares fitting line to the respective scatter plots and the solid red line is the 45° or the X = Y line for comparison).

Table S1 Number of points that were under and overpredicted by the wavelet-based models (*i.e.*, IIS-W-ANN and IIS-W-M5 Tree models) with respect to 5% tolerance limit.

	Site 1-Menindee River		**Site 2-Gwydir River**		**Site 3-Darling River**	
	IIS-W-ANN	IIS-W-M5 Tree	IIS-W-ANN	IIS-W-M5 Tree	IIS-W-ANN	IIS-W-M5 Tree
Under-prediction	39	41	18	24	21	23
Over-prediction	16	20	39	32	38	39
Total	55	61	57	56	59	62

Chapter 4: Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition

Foreword

This chapter is an exact copy of the published article in the journal *Geoderma* (Vol. 330, Pages 136-161).

Further to streamflow water level forecasting (Chapter 3), forecasting of another important hydrological variable, the soil moisture, is undertaken in this chapter with the employment of ensemble modelling technique. Two self-adaptive multi-resolution analysis utilities, *viz.*, ensemble empirical mode decomposition (EEMD) and complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) are applied to appropriately unveil and extract entrenched features within the soil moisture time series and address the non-stationarity issues. The EEMD and CEEMDAN utilities resolve the soil moisture time series into a number of intrinsic mode functions (IMFs) and a residual component. These resolved IMFs and residual sub-series are forecasted by respective machine learning models. The machine learning models used for this purpose are extreme learning machine (ELM) which is a single layer feed-forward neural network algorithm and the bootstrap-aggregated random forest (RF) models. In this case, only the soil moisture time series is used to forecast future soil moisture capitalizing on the memory feature of several (lagged) months within the *SM* time-series. Hence, partial autocorrelation function (*PACF*) has been adopted to determine the salient input lags.

The proposed hybrid EEMD-ELM model was extensively evaluated against hybrid CEEMDAN-ELM, and the equivalent random forest hybrid models (*i.e.*, EEMD-RF and CEEMDAN-RF) as well as the standalone ELM and RF models in forecasting upper (0-0.2 m) and lower layer (0.2-1.5 m) relative soil moisture at seven hydrological sites within MDB, Australia.

Geoderma 330 (2018) 136–161

Contents lists available at ScienceDirect

Geoderma

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

Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition

Ramendra Prasad*, Ravinesh C. Deo*, Yan Li, Tek Maraseni

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

ARTICLE INFO

Handling Editor: Morgan Cristine L.S.

Keywords:
Hybrid models
EEMD
CEEMDAN
ELM
Extreme learning machine
Random forest
Soil moisture forecasting
Drought-prone Murray-Darling Basin

ABSTRACT

Soil moisture (*SM*) is an essential component of the environmental and the agricultural system. Continuous monitoring and forecasting of soil moisture is a desirable strategy to understand the soil dynamics for proactive planning and decision-making measures for agriculture and related fields. In this study hybrid data-intelligent, extreme learning machine (ELM) models are designed and explored for monthly *SM* forecasting. The chaotic, complex and dynamical behavior of *SM* can compound the accuracy of data-driven models. Consequently, two versatile, computationally efficient and self-adaptive multi-resolution utilities namely, complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and the ensemble empirical mode decomposition (EEMD) algorithms are utilized to address these data non-stationarity issues, which if not resolved can lead to model prediction inaccuracies. The difference in these approaches is that, during the EEMD process, a Gaussian white noise is added to the intact (*i.e.*, unresolved) time series only, while, the CEEMDAN requires sequential additions at each decomposition phase. Integration of these multi-resolution tools with the ELM model led to the hybrid CEEMDAN-ELM and the EEMD-ELM models, that were benchmarked with random forest (RF) equivalent models. Using *WaterDyn* model's hind-simulated *SM* data, these models were applied (without any climate inputs) to forecast the upper (0.2 m) and the lower layer (0.2–1.5 m depth) soil moisture in Australia's agricultural-hub, the Murray-Darling Basin. The standalone ELM and RF model has similar computation efficiency and model performances. However, despite the implementation of computationally expensive ensemble techniques (*i.e.*, EEMD and CEEMDAN, the hybrid ensembles EEMD-ELM and CEEMDAN-ELM were highly efficient with improved performances. The research outcomes showed that the CEEMDAN-ELM model outperformed the alternative models at three (out of the seven) sites applied for upper layer *SM* forecasts, while the EEMD-ELM hybrid model was superior at all seven sites for the lower layer soil moisture forecasts. The study signifies the important role of the self-adaptive multi-resolution utility (CEEMDAN) hybridized with the ELM algorithm to potentially develop automated prediction systems for forecasting soil moisture, with potential applications in agriculture.

1. Introduction

The structure and functioning of the natural hydrological system is contingent upon soil moisture (*SM*) which is the principal regulating element of groundwater hydrology, biogeochemical balance, partitioning of the mass and energy fluxes in between land-atmosphere system (Brocca et al., 2017; Brocca et al., 2010; Petropoulos, 2014), and nutrient and greenhouse gas fluxes. On the other hand, the agricultural yield is also explicitly dependent on *SM* content and any unprecedented fluctuations could be deleterious for this volatile industry. To devise sustainable planning, and scheduling of specialized agricultural tasks, efficient and effective temporal predictive systems are essential tools. Advanced or forecasted knowledge of this important variable, *SM*, is pivotal for proactive sustainable decisions in efficient irrigation scheduling, grazing scheduling, water quality monitoring, yield predictions (Gill et al., 2006), water resource management (Zhang et al., 2017a) and soil carbon loss prediction (Rey et al., 2017). Intelligent agricultural decision support systems based on artificial intelligence used in monitoring and forecasting *SM* can provide useful and tangible solutions in enhancing sustainability and productivity of farming systems.

Envisioning this, *SM* forecast models have been established that includes empirical formulations, the water balance approach, the dynamic soil-water models, time series models, remote sensing models and neural network models (Huang et al., 2011). However, these models have limitations in practical applications. For instance, water balance, soil-water dynamic, and time series model require an intensive volume of spatial and temporal (measured) data as initialization

* Corresponding authors.
E-mail addresses: ramendra.prasad@usq.edu.au (R. Prasad), ravinesh.deo@usq.edu.au (R.C. Deo).

https://doi.org/10.1016/j.geoderma.2018.05.035
Received 7 December 2017; Received in revised form 2 March 2018; Accepted 30 May 2018

conditions. In addition, the remote sensing model has a poor stability (*i.e.*, plagued by dew) while the empirical model parameters lack practical scope (Huang et al., 2011; Mahmood and Hubbard, 2004; Weimann et al., 1998). The problem is further exacerbated by the perplexing association between *SM* and its derivative factors, such as climate dynamics and geomorphologic properties (*e.g.*, topography, soil properties, vegetation type and density, depth to water table and land use) (Famiglietti et al., 1998; Zhang et al., 2017a). Moreover, sophisticated programs and rigorous optimization techniques are required for model calibration (Jain and Srinivasulu, 2004).

To surmount the difficulties, the preciseness of extreme learning machine (ELM) pioneered by Huang et al. (2004) is evaluated in forecasting *SM*-derived from the physical *WaterDyn* model (AWAP, 2016; Raupach et al., 2009). ELM is a recent state-of-the-art data intelligent model. It is convenient to use single layer feed-forward neural network (SLFN) with better generalization capability (Shamshirband et al., 2015; Sun et al., 2008). ELM has demonstrated high accuracy at a lower computational expense for forecasting water demand (Mouatadid and Adamowski, 2016; Tiwari et al., 2016), stream-flow (Deo and Sahin, 2016; Yaseen et al., 2016), wind speed (Shamshirband et al., 2015), dew-point temperature (Mohammadi et al., 2015) and evapotranspiration (Patil and Deka, 2016). The SLFN modeling framework of ELM is similar to that of the feed-forward neural network with random weights (Schmidt et al., 1992) and random vector functional-links (RVFL) (Pao et al., 1994) where the input weights and biases are also randomly assigned. ELM is occasionally referred to as a variant of RVFL (Cecotti, 2016; Scardapane et al., 2015). Yet, ELM has subtle but important variations. In comparison to feed-forward neural network with random weights, the ELM has added output biases which were lacking in the former model (Huang, 2014; Schmidt et al., 1992). In addition, there is no direct connection in between inputs and outputs in ELM, which is the case with RVFL (Huang, 2014; Pao et al., 1994; Wang and Wan, 2008). ELM also provides added versatility for implementations of various nonlinear activation and kernel functions (Huang, 2014; Shamshirband et al., 2015; Sun et al., 2008). However, the literature shows that ELM has not been fully explored in *SM* forecasting, and hybrid models of ELM integrating multi-resolution analysis are relatively scarce. One study by Liu et al. (2014) forecasted *SM* in Dookie apple orchard, Victoria, Australia using ELM and support vector machines (SVM), revealing the superiority of ELM in *SM* forecasting at a soil depth of 20, 40 and 60 cm. Yet, that study period spanned across a very short period (14 months) and apparently lacked the inclusion of significant seasonal and long-term climate dynamics derived from realistic, physically-based inputs.

To benchmark ELM, a bootstrapped-aggregated tree approach, random forest (RF) has been designed. RF has proven to yield good performance with reasonable prediction accuracy in forecasting hydrometeorological variables, such as temperature variation (Naing and Htike, 2015), wind power (Lahouar and Ben Hadj Slama, 2017) and standardized precipitation index (Chen et al., 2012a). Similar to ELM, RF is uncommon in *SM* forecasting. The study by Matei et al. (2017) used RF to forecast *SM* at soil depths of 10 cm, 30 cm and 50 cm in Transylvania plain, Romania, while no such study has been carried out in Australia so far.

Despite the ability to handle dynamicity and nonlinearity, so far no single data-intelligent approach has been able to provide aptest forecasts under erratic hydrological conditions (Yaseen et al., 2016). The chaotic, complex and dynamical behavior of pedologic and hydrological processes leads to non-stationarities (varying mean) and seasonality (changes in variance) within the model input series (Hu and Si, 2013; Kim and Valdes, 2003; Nourani et al., 2014). This behavior can compound the ability of conventional data-intelligent models in accurately simulating the soil moisture. With the insight to address this issue, two relatively new and advanced versions of empirical mode decomposition (EMD) has been utilized to resolve the embedded frequency information (*i.e.*, related to the physical structure of data) in the model inputs. Multi-resolution analyses (MRA) tool, ensemble-EMD (EEMD), was proposed by Wu and Huang (2009) and the complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) was proposed by Torres et al. (2011). Both aim to segregate higher frequency input series into lower frequency resolved parts to extract and isolate salient features representing the physical structure of the data. Both of these techniques have merits over conventional approaches (*e.g.*, wavelet transform (WT) (Mallat, 1989; Mallat, 1998; Nourani et al., 2014a; Nourani et al., 2009)), singular value decomposition (SVD), singular spectrum analysis (SSA) (Chau and Wu, 2010; Chitsaz et al., 2016) and principal component analysis (PCA) (Hu et al., 2007)). Among these, WT has been widely used (*e.g.*, (Anctil and Tape, 2004; Deo et al., 2017a; Deo et al., 2016; Labat et al., 2000; Nourani et al., 2009; Wen et al., 2016)). In particular, the non-decimated wavelet function (*i.e.*, maximum overlap discrete wavelet transform, MODWT) is able to retain the downsampled values but the choice of the mother wavelet with MODWT is a major concern. There is no explicit rule to select an optimal wavelet other than by an iterative trial and error process (Prasad et al., 2017). The EEMD and CEEMDAN decompositions does not require prescribed frequency bands or imposed basis functions, thus making the decomposition completely self-adaptive. This offers a significant advantage over wavelets. Both EEMD and CEEMDAN solve the 'mode mixing' issue of EMD, achieved by the addition of a Gaussian white noise to the intact (*i.e.*, undecomposed) series. EEMD has been found to reduce the difficulties in the forecasting process, by reducing the complexity of a time series (Di et al., 2014). During CEEMDAN-based decomposition, a Gaussian white noise with unit variance and noise coefficient is added sequentially at each decomposition stage. Although this does have limitations on parallel computing, the reconstruction of CEEMDAN decomposed data is complete and noise-free (Ren et al., 2015; Zhang et al., 2017b). In spite of the advantages and self-adaptability making it suited for practical applications, neither EEMD nor CEEMDAN has been broadly applied in soil moisture forecasting applications.

EEMD-based data-driven models have been explored in forecasting precipitation (Beltran-Castro et al., 2013; Jiao et al., 2016; Ouyang et al., 2016), reservoir inflows (Bai et al., 2015) and daily river data (Seo and Kim, 2016). Although these studies found that the models generated improved forecasts, very limited application of EEMD in *SM* forecasting has been carried out. Basha et al. (2015) carried out forecasting of temperature, precipitation and *SM* patterns for the United Arab Emirates using EEMD coupled Non-Stationary Oscillation Resampling (NSOR) model and compared it with Coupled Model Intercomparison Project phase 5 (CMIP5) projections. They found that the EEMD-NSOR model had a better forecasting capability. Likewise, CEEMDAN has been found to be more effective than EMD to forecast wind speed (Ren et al., 2015; Zhang et al., 2017b), power load (Palaninathan et al., 2016) and electricity markets (Afanasyev and Fedorova, 2016). In terms of estimation error, CEEMDAN was comparable with wavelet-decomposition (Afanasyev and Fedorova, 2016) but to the best of the authors' knowledge, the application of the technique is yet to be explored in forecasting *SM* at large.

The purpose of this research study is to develop a new and precisely tuned hybrid data-intelligent model overcoming non-stationarity issues in forecasting upper and lower layer soil moisture with potential for practical applications. Temporal hind-casted *SM* data generated from the physically-driven hydrological model (*i.e.*, Commonwealth Scientific and Industrial Research Organisation's (CSIRO's) *WaterDyn* model) incorporating climatic forcing (*e.g.*, solar radiation, temperature, rainfall, *etc.*) (AWAP, 2016; Raupach et al., 2009) are utilized. Two self-adaptive multi-resolution analysis methods (*i.e.*, EEMD and CEEMDAN) are embedded into an extreme learning machine (ELM) algorithm to resolve the frequencies and to unveil the physical structure of the input variable before the model is applied in the actual forecasting of soil moisture. The resulting hybrid EEMD-ELM and the CEEMDAN-ELM based ensemble models are designed and then

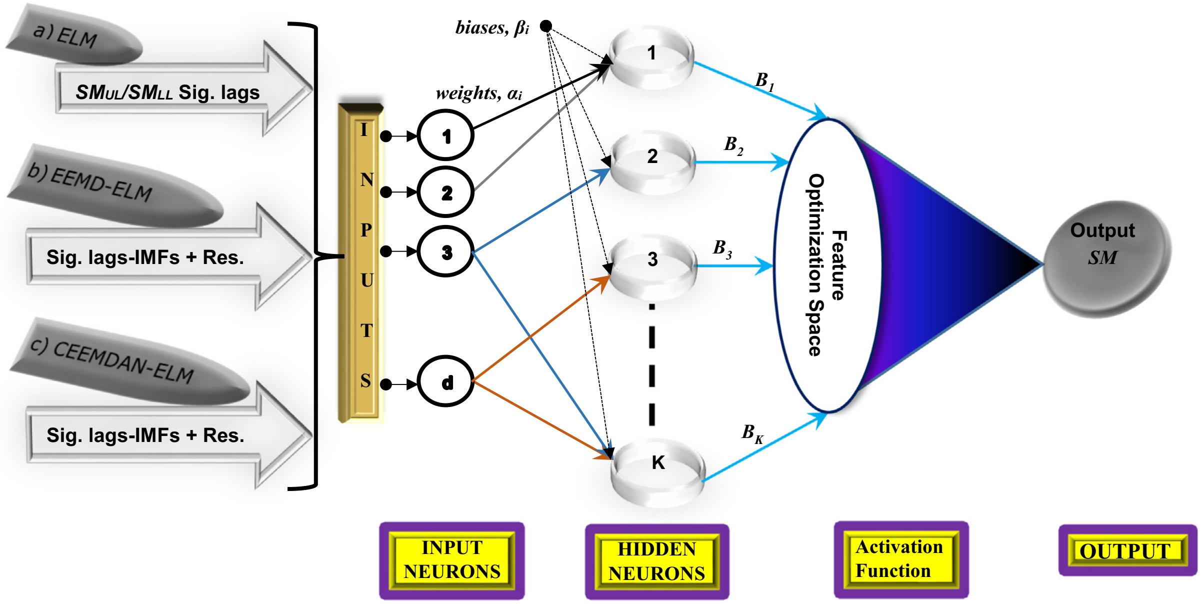


Fig. 1. The architecture of extreme learning machine (ELM) network. Details of input variables are provided in Tables 4a–b, while the modeling framework is given in Table 6a. The hidden neurons from 50 to 200 were used.

evaluated at seven hydrological sites within the Murray-Darling Basin in Australia. To cross-validate the model's versatility in *SM* forecasting, ELM-hybrid models are benchmarked against random forest (RF) equivalent hybrid and standalone RF models, as the first study for *SM* forecasting. Next, we outline the theory of the data-intelligent algorithms followed by the decomposition techniques, methods, data, and the results. An overview of the challenges and the prospects of real-time *SM* forecasting using the proposed EEMD-ELM approach is presented with concluding remarks for closing the paper.

2. Data-intelligent algorithms

2.1. Extreme learning machine

In this paper, ELM model is developed in accordance with single layer feed-forward neural network (SLFN). The input weights in ELM are randomly assigned while the output weights are analytically determined, as depicted by the simplified schematic architecture in Fig. 1. The general output function of the ELM algorithm with K hidden neurons is expressed as (Huang et al., 2004; Huang et al., 2006):

$$\sum_{i=1}^{K} B_i G_i(\alpha_i,\ \beta_i, x_t) = z_t \tag{1}$$

where $x_t \in \boldsymbol{R}^{\boldsymbol{d}}$ are predictor inputs (with d as the input dimension and t as the occurrence instances), $z_t \in \boldsymbol{R}$ represents the model output (forecasted values), $B \in \boldsymbol{R}^K$ represents the output weights, $\alpha_i \in \boldsymbol{R}^K$ are the input weights and $\beta_i \in \boldsymbol{R}$ are the biases and i is the index of the hidden neuron. Consequently, the additive hidden nodes with the activation function of $g(x)$: $\boldsymbol{R} \to \boldsymbol{R}$ is represented as (Huang et al., 2006):

$$G_i(\alpha_i,\ \beta_i, x_t) = g(\alpha_i .\ x_t + \beta_i) \tag{2}$$

where $G_i(\alpha_i, \beta_i, x_t)$ is the output of the i^{th} hidden node corresponding to input x. In hydrology non-linear logistic functions (*e.g.*, logarithmic sigmoid) are the preferred activation functions (Deo and Sahin, 2016; Deo and Şahin, 2015). For the training of the ELM model applied for the forecasting of monthly relative upper and lower layer soil moisture levels, the term $N = 218$ denotes the pairs of the training predictor samples (x_t) and the observed soil moisture (SM^{OBS}): $T = \{(x_t, SM_t^{OBS}) : x_t \in \boldsymbol{R}^{\boldsymbol{d}}, SM_t^{OBS} \in \boldsymbol{R}\}$ with d = the number of predictor inputs (input neurons) and $t = 1,\ 2,\ \ldots N = 218$ monthly data were used. It is important to note that for the standalone ELM models developed in this paper, $x_t \in \boldsymbol{R}^{\boldsymbol{d}}$ are the significant lags of the intact soil moisture time series *i.e.*, the series without EEMD/CEEMDAN analysis. On the other hand, for the ensemble EEMD-ELM/CEEMDAN-ELM models, $x_t \in \boldsymbol{R}^{\boldsymbol{d}}$ represents the significant lagged series of the intrinsic mode functions (IMFs) and the residual component, which result after the EEMD/CEEMDAN transformations of the original input data series. During training of the ELM model, the output (z_t) is replaced with the observed soil moisture values (SM_t^{OBS}) since the observed input-output pairs are used to determine the relevant weights and biases. As such, the general equation (Eq. (1)) becomes:

$$\sum_{i=1}^{K} B_i G_i(\alpha_i,\ \beta_i, x_t) = SM_t^{OBS} \tag{3}$$

Simplifying Eq. (3) with G being the hidden layer output, B being the weights gives:

$$SM^{OBS} = GB \tag{4}$$

which in matrix notations are:

$$G = \begin{bmatrix} g(\alpha_1 .\ x_1 + \beta_1) & \ldots & g(\alpha_K .\ x_1 + \beta_K) \\ \vdots & \ldots & \vdots \\ g(\alpha_i .\ x_N + \beta_1) & \ldots & g(\alpha_K .\ x_N + \beta_K) \end{bmatrix}_{N\times K},\ B = \begin{bmatrix} B_1 \\ \vdots \\ B_K \end{bmatrix}_{K\times 1},$$

$$\text{and } SM^{OBS} = \begin{bmatrix} SM_1^{OBS} \\ \vdots \\ SM_{N=218}^{OBS} \end{bmatrix}_{N\times 1}$$

With suitable number of hidden neurons and randomized allocation of input layer weights, and hidden neurons biases (α and β), the ELM network's output weights (B) are analytically determined to yield zero forecasting errors. Hence, following Eq. (4), the values of B are estimated directly from the N input-output data samples *via* a least-square solution as follows (Huang et al., 2006):

$$\widehat{B} = G^{\dagger} SM^{OBS} \tag{5}$$

where $G^{\dagger}$ is the Moore–Penrose generalized inverse of G.

This process then enables an efficient and random selection of the input weights and the corresponding hidden layer biases are able to reduce the ELM modeling network to a linear prediction system. Thus, the output weights (*i.e.*, the features transferred from the hidden to the

Table 1
Geographical locations and basic physical characteristics of the selected sites in the Murray-Darling Basin, New South Wales (NSW), Australia.

Site no.	Station names	Location			Physical characteristics					
		Long. (°E)	Lat. (°S)	Approx. elev. (m)	Major climate classes (Hijmans et al., 2005)	Land use (Department of Agriculture and Water Resources, 2015)	Soil type (ASRIS, 2014)	Range of agricultural holding (ha) (Australian Bureau of Statistics, 2008)	Population density (persons/km^2) (ABS, 2011)	Soil characteristics (ASRIS, 2014)
1	Menindee	142.15	−32.45	75.3	Desert	Grazing-Native vegetation	Calcarosol	18,700–38,600	0–2	Brown sands with clay substrata underneath.
2	Balranald	143.30	−34.75	65.5	Savannah	Dry-land cropping	Calcarosol	3700–18,700	0–2	Brown calcareous with crusty loamy soils.
3	Wanaaring	144.45	−30.15	101.7	Savannah	Grazing-Native vegetation	Dermosol	18,700–38,600	0–2	Sandy yet loamy red soils with areas of yellow soils mantled with shallow loams.
4	Bobadah	146.75	−32.45	277.3	Savannah	Dry-land cropping	Kandosol	600–3700	0–2	Chief soils are gravelly and non-gravelly neutral red soils on the gently undulating terrain.
5	Moorwatha	146.75	−35.90	211.4	Sub-Tropical	Dry-land cropping	Sodosol	0–600	2–5	Chief soils are hard neutral red soils and loamy red duplex soils.
6	Jerrawa	149.05	−34.75	656.0	Temperate	Grazing-Modified pastures	Sodosol	0–600	5–10	Chief soils are hard neutral yellow and yellow mottled soils.
7	Rocky Creek	150.20	−30.15	689.0	Sub-Tropical	Dry-land cropping	Sodosol	3700–18,700	0–2	Chief soils are hard neutral and acidic yellow and yellow mottled soils.

output layer) of this linear system are analytically determined by a simple generalized inverse operation of the hidden layer output matrices. This makes the ELM model computationally efficient in comparison with the conventional lengthy iterative adjustments required in the case of the network parameters of an artificial neural network model (Deo et al., 2017a). The suitability of ELM has been explored in many fields (Deo and Sahin, 2016; Mouatadid and Adamowski, 2016; Patil and Deka, 2016; Shamshirband et al., 2015; Tiwari et al., 2016; Wang et al., 2017c; Yaseen et al., 2018; Yaseen et al., 2016), however, it is yet to be extensively tested on huge datasets. In this paper, the forecasts of the objective variable (*i.e.*, monthly soil moisture) are then generated and the results are compared with a Random Forest (RF) model, which are described in the next section.

2.2. Random Forest

Random Forest (RF) is a regression tree-based ensemble technique introduced by Breiman (2001) (an extension of bagging (Breiman, 1996)). RF aims to reduce the variance without undue increase in the bias. Employing a bootstrap aggregation (bagging) approach, the RF model is able to overcome the overfitting issue of conventional solitary regression trees. During training, '*n*' bootstrap replicas are taken from the training data-set, using random sampling with replacement. For the standalone RF models developed in this study, the significant lagged soil moisture time series without the EEMD/CEEMDAN analysis were used as the model's inputs. While, for the EEMD-RF/CEEMDAN-RF models, the significant lagged series of the intrinsic mode functions (IMFs) and the residual component, that were obtained after the EEMD/CEEMDAN transformations were applied, were taken to be the model's inputs. The outputs in all the modeling cases were the monthly relative upper and the lower layer soil moisture level, as with the case of ELM-based models. Henceforth, a single tree is constructed on every separate '*n-replicas*' with simultaneous computation of out-of-bag (OOB) errors of respective trees using the data that were not used during training as:

$$\text{OOB error} = \frac{1}{N}\sum_{t=1}^{N}(SM_t^{OBS} - SM_t^{FOR})^2 \tag{6}$$

where SM_t^{OBS} is the t^{th} instance of observed value and SM_t^{FOR} is the corresponding forecasted value.

The single regression trees are put together, whereby the forecasted output is averaged over an ensemble of trees. This stabilizes the outputs to allow for desirable adaptability and better generalization (Lin and Jeon, 2006). At each split in decision trees, a random subset of features is selected for testing. This further improves the accuracy (Breiman, 2001) and overcomes model over-fitting issues (Diaz-Uriarte and Alvarez de Andres, 2006). Only three parameters in RF requires some tuning: i) *m*, the number of randomly assigned predictor variables at each node, ii) *J*, the number of trees in the forest and iii) tree size, the maximum number of terminal nodes/leaf. *J* is not problematic (Breiman, 2001), however, an excessively small value may lead to convergence of the generalization error. The default value of *m*, one-third of the total number of variables, has been reported to be ideal (Liaw and Wiener, 2002) and therefore has been adopted in this study with $J = 200$ and tree size = 5, which has generated optimal results shown later.

2.3. Ensemble empirical mode decomposition (EEMD)

The EEMD is an improvement to overcome the mode mixing issue of empirical mode decomposition (EMD), offering a better time series processing (Wu and Huang, 2009). EEMD features stronger self-adaptability and local variation characteristics (Li et al., 2015) while effectively detecting the non-stationarity and nonlinearity. It separates the embedded oscillations at different scales into intrinsic mode functions (IMFs) and a residual (trend) component (Wu and Huang, 2009).

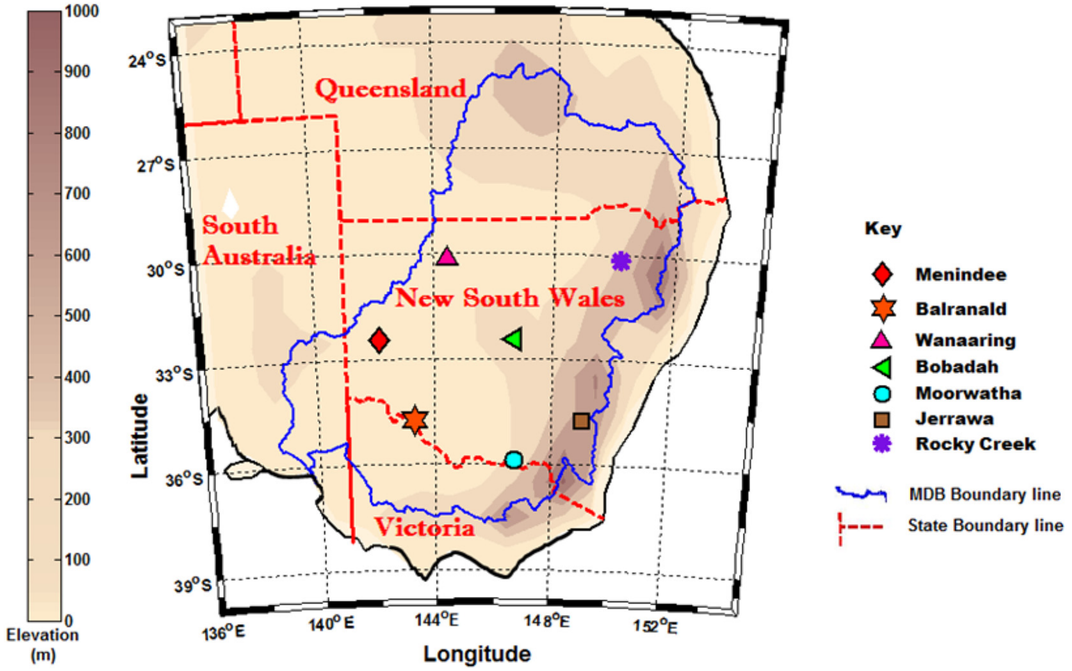


Fig. 2. Map of study region showing the selected stations and its geographical locations. The colored contour gradients show the elevation (in meters) above sea level. (*Refer to the key for the names of sites with respective marker labels.*)

Table 2
Monthly climatic features of upper and lower layer relative soil moisture (SM_{UL} and SM_{LL}) at the selected sites.

Site no.	Station names	SM_{UL} Monthly climatic features					SM_{LL} Monthly climatic features				
		Min.	Max.	Mean	Skew-ness	Kurtosis	Min.	Max.	Mean	Skew-ness	Kurtosis
1	Menindee	0.013	0.434	0.139	0.791	0.058	0.205	0.703	0.343	1.241	1.421
2	Balranald	0.013	0.470	0.159	0.668	−0.182	0.034	0.187	0.092	0.695	−0.533
3	Wanaaring	0.011	0.452	0.141	0.979	0.302	0.098	0.485	0.213	1.302	1.487
4	Bobadah	0.015	0.520	0.197	0.572	−0.272	0.119	0.560	0.290	0.436	−0.840
5	Moorwatha	0.018	0.754	0.288	0.380	−0.635	0.147	0.999	0.571	0.158	−0.695
6	Jerrawa	0.026	0.893	0.302	0.598	0.097	0.210	1.000	0.584	0.446	−0.403
7	Rocky Creek	0.036	0.814	0.285	0.463	0.489	0.145	1.000	0.473	0.442	0.218

[NB: The relative soil moisture values are based on the base climatological reference period: 1961–1990].

A Gaussian white noise is added to the intact (*i.e.*, unresolved) time series to provide a uniform reference frame in gathering the components of the same frequency. The noise is eliminated through averaging of corresponding IMFs and the residual component. A brief realization of the EEMD algorithm is as follows (Wu and Huang, 2009): For an unresolved signal $x(t)$, (1) Add a white noise series $s(t)$ such that $x^{'}(t) = s(t) + x(t)$. (2) Decompose $x'(t)$ into IMFs and residue. (3) Repeat steps 1 and 2 '*p*' numbers of times, with different white noise each time (where '*p*' = ensemble number). (4) Compute the mean of all IMF components (without mode mixing) and the mean of residue components cancelling out the added white noise. Hence, the intact (unresolved) time series is expressed as the sum of IMFs and the residue as:

$$x(t) = \sum_{i=1}^{m} IMF_i(t) + r_m(t) \tag{7}$$

where $IMF_i(t)$ is the intrinsic mode functions, $r_m(t)$ denotes the final residue component, m is the total number of IMFs, and i is the component indices.

Intrinsic mode functions should satisfy two admissibility conditions (Huang et al., 1998; Wu and Huang, 2009); (i) Over its entire length, the number of extrema and the number of zero-crossings must either be equal, or differ at most by one; (ii) At any point, the mean value of the signal defined by the local maxima and the envelope defined by the local minima is zero. Therefore, IMF_1 has a maximum amplitude and highest frequency, while the subsequent IMFs have lower amplitude and frequency. The final component, the residue (or trend) is a slowly varying mode around the long-term average.

2.4. Complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN)

Alternatively, the CEEMDAN technique proposed by Torres et al. (2011) has similar advantages as being self-adaptive and one that avoids mode mixing problems. With that, the reconstructed time series is identical to the intact (unresolved) one (Colominas et al., 2014; Torres et al., 2011). The key dissimilarity is that during CEEMDAN decomposition process, a Gaussian white noise with unit variance and noise coefficient is added at each of the decomposition stages. This noise-added signal is decomposed *via* EMD to obtain the first IMF and the subsequent residual component. Consider an unresolved signal, $x(t)$, and added white noise, $s_n(t)$. To obtain the first IMF by CEEMDAN (*i.e.* IMF_1), for every $n = 1\ldots, N$ decompose each $x_n^{'}(t) = x(t) + \varepsilon s_n(t)$ *via* EMD where: N = ensemble number and ε = amplitude of the added noise. Note that at each subsequent stage, the coefficient ε_p allows an appropriate selection of the signal-to-noise ratio of the white noise. Then collate the first IMFs produced by EMD (*i.e.* d_1) and compute the ensemble average as follows:

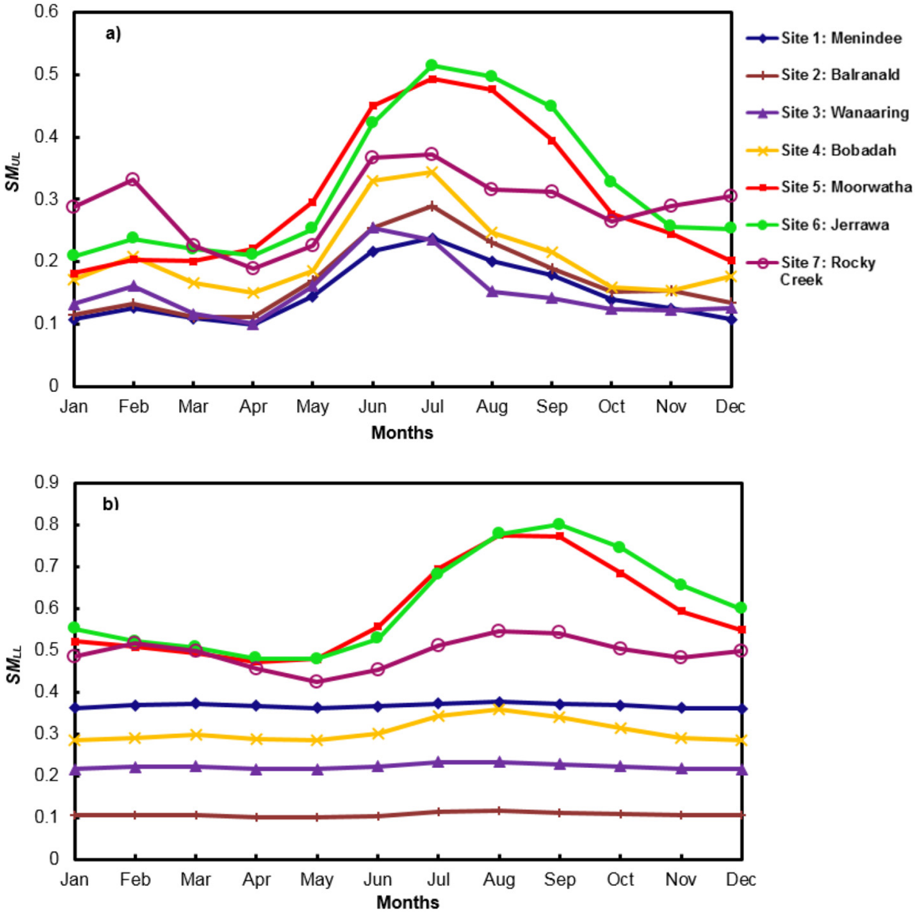


Fig. 3. Monthly variations of a) upper layer (SM_{UL}) and b) lower layer (SM_{LL}) soil moisture. (NB: *SM_{UL} and SM_{LL}* are relative values and are dimensionless.)

Table 3
Data partitions used for model development in this study.

Sites	Period	Number of datum points	Number of input features after lags	Data partition		
				Training	Validation	Testing
All sites and for both soil moisture layers	Jan 1990 to Dec 2016	324	324–12 = 312	70% 218	15% 47	15% 47

$$\mathrm{IMF}_1(t) = \frac{1}{N}\sum_{n=1}^{N} d_1 \tag{8}$$

The remaining component after obtaining IMF_1 is described as:

$$r_1(t) = x(t) - \mathrm{IMF}_1(t) \tag{9}$$

Next, the computation of IMF_2 of the intact signal, $x(t)$, from the remaining component, where $F_j(.)$ is the operator which produces the j^{th} IMF obtained *via* EMD, is expressed as:

$$\mathrm{IMF}_2(t) = \frac{1}{N}\sum_{n=1}^{N} F_1\left[r_1(t) + \varepsilon_1 F_1(s_n(t))\right] \tag{10}$$

Hence, the subsequent IMFs ($p = 2, 3\ldots P$) are expresses as follows:

$$\left.\begin{aligned} r_p(t) &= r_{p-1}(t) - \mathrm{IMF}_p(t) \\ \mathrm{IMF}_{p+1}(t) &= \tfrac{1}{N}\sum_{n=1}^{N} F_1\left[r_p(t) + \varepsilon_p F_p(s_n(t))\right] \end{aligned}\right\} \tag{11}$$

The remaining component, $r_p(t)$, is repeatedly decomposed using EMD to obtain subsequent IMFs until the residue does not satisfy the conditions of IMF's. The final residual (R) component is expressed as:

$$R(t) = x(t) - \sum_{p=1}^{P} \mathrm{IMF}_p(t) \tag{12}$$

3. Materials and method

3.1. Study area and description of model design data

The study area, New South Wales (NSW), hubs Australia's agricultural belt, the Murray-Darling Basin (MDB) that covers 14% of land area and encompasses 67% as agricultural land (Australian Bureau of Statistics, 2010). MDB contributes to 1/3 of Australia's food supply (Welsh et al., 2013) and 2% of the total economic output (Australian Bureau of Statistics, 2014). Accordingly, seven sites were selected as summarized in Table 1 with distinct geophysical characteristics acquired from various sources (ABS, 2011; ASRIS, 2014; Australian Bureau of Statistics, 2008; Department of Agriculture and Water Resources, 2015; Hijmans et al., 2005). Fig. 2 illustrates the physical locations with overlayed elevation shades for better visualization.

Data-intelligent models solely rely on historical information to forecast the future *SM*. In this paper, we adopt the simulated *SM* data from the physically-driven *WaterDyn* hydrological model developed by

Table 4
Input variables for standalone, EEMD and CEEMDAN hybrid extreme learning machine (ELM) and random forest (RF) models based on *PACF* in forecasting: a) relative soil moisture for upper layer (SM_{UL}); b) relative soil moisture for lower layer (SM_{LL}).

a) SM_{UL}	Significant input lag numbers at respective sites						
Sites	1	2	3	4	5	6	7
Standalone models (Intact time series (*i.e.*, without EEMD/CEEMDAN analysis)	1, 2, 5	1, 10, 11	1, 10, 12	1, 10, 11, 12	1, 2, 3, 4, 10, 11	1, 2, 4, 10, 11	1
EEMD (Hybrid ELM/RF Models)							
IMF-1	2–3	1–5	2–4	1–3	1–4	1–6	1–4
IMF-2	1–6, 8	1–6, 8	1–4, 6	1–4, 6	1–4, 6	1–4, 6	1–8, 10
IMF-3	1–8	1–6, 8	1–6	1–6	1–6	1–6	1–6
IMF-4	1–6	1–6	1–6	1–6, 8	1–5, 7	1–6, 8	1–6, 8
IMF-5	1–6, 8	1, 8, 10	1–6	1–6	1–8	1–6	1–8, 10
IMF-6	1–8	1–8	1–6	1–6	1–8	1–6	1–7
IMF-7	1–6	1–8	1–5	1–6	1–7	1–6	1–6
IMF-8		1–7		1–7	1–6		
Residual	1–6	1–12	1–6	1–12	1–12	1–6	1–8
CEEMDAN (Hybrid ELM/RF Models)							
IMF-1	1–2	1–5	1–3, 5	1–3, 5	1–4	1–6	1–4
IMF-2	1–6, 8	1–4, 6	1–4, 6	1–7	1–4, 6–9	1–4, 6–8	1–4, 6–8
IMF-3	1–3, 5–6	1–5	1–3, 5–7	1–6	1–7	1–4, 6–7	1–3, 5–6
IMF-4	1–6, 8	1–6, 8–9	1–6, 8	1–6, 8	1–6, 8–10	1–6, 8–10	1–6, 8
IMF-5	1–6	1–5	1–5	1–6, 8	1–6, 8–11	1–6, 8	1–6, 8–9
IMF-6	1–8, 10	1–8	1–6	1–6	1–7	1–6	1–6
IMF-7	1–6	1–6	1–6	1–6	1–6	1–6	1–5
IMF-8	1–6	1–7	1–7	1–6	1–6		1–6
Residual	1–6	1–6	1–7	1–7	1–7, 9	1–5	1–6

b) SM_{LL}	Significant input lag numbers at respective sites						
Sites	1	2	3	4	5	6	7
Standalone models (Intact time series (*i.e.*, without EEMD/CEEMDAN analysis)	1, 2, 3	1, 2, 3, 12	1, 2	1, 2, 3	1, 2, 3, 8, 9, 12	1, 2, 8, 9	1, 2
EEMD (Hybrid ELM/RF Models)							
IMF-1	2–5	2–5	1–6	2–6	2–6	1–5	2–7
IMF-2	1–6	1–6, 8	1–8, 10–11	1–6	1–7	1–6	1–6
IMF-3	1–8	1–7	1–6, 8	1–6	1–10	1–7	1–7
IMF-4	1–7	1–7	1–7	1–7	1–6	1–6, 8	1–7
IMF-5	1–7	1–8	1–7	1–7	1–7	1–8	1–8
IMF-6	1–6	1–6	1–7	1–6, 8	1–8	1–6	1–6
IMF-7	1–5, 7–10	1–8	1–7	1–5	1–6	1–5	1–5
IMF-8		1–6					
Residual	1–12	1–12	1–5	1–6	1–12	1–6	1–12
CEEMDAN (Hybrid ELM/RF Models)							
IMF-1	2–5	2–7	1–7	2–6	1–6	2–5	2–7
IMF-2	1–2	1–2	1–2, 4–5	1–5, 7–8	1–4	1–4, 6–7	1–4
IMF-3	1–3, 5–8	1–3, 5–7	1–3, 5–7	1–2, 4–6	1–6	1–6, 8	1–3, 5–7
IMF-4	1–6, 8	1–6, 8	1–6, 8	1–6, 8	1–6, 8–11	1–6, 8–9	1–6, 8
IMF-5	1–5	1–5	1–5, 7–8	1–5	1–6	1–5	1–5
IMF-6	1–7	1–7	1–7	1–5, 7	1–6	1–6	1–6
IMF-7	1–5	1–6	1–6	1–6	1–6	1–6	1–7
IMF-8	1–6		1–7	1–6	1–6		
Residual	1–6	1–6	1–7	1–7	1–6	1–6	1–6

the Commonwealth Scientific and Industrial Research Organisation (CSIRO) in collaboration with Australian Bureau of Meteorology (BOM) as part of the Australian Water Availability Project (AWAP) (AWAP, 2016; Raupach et al., 2009). Authentic and reliably gridded monthly historical values of relative (*i.e.*, fractional) soil moisture data bounded by [0, 1] for the upper layer (denoted hereafter as SM_{UL}) and the lower layer (SM_{LL}) over a 0.05° × 0.05° spatial resolution, are used (Raupach et al., 2009). The physically-generated AWAP-based *SM* data provides a realistic account of the soil's hydrological conditions by simulating climatic variables to model terrestrial water balance across continental Australia. Inputs and constraints include meteorological forcing (*i.e.*, solar radiation, precipitation, minimum and maximum daily temperature) and continental parameter maps (*e.g.* albedo, soil characteristics, seasonality of vegetation greenness). Due to this dependence on meteorological inputs, quality controlled meteorological fields generated by BOM's network of rain gauge and weather sites are used by *WaterDyn* model while solar irradiance data is obtained using geostationary satellite imageries (AWAP, 2016; Raupach et al., 2009).

The upper layer *SM* is up to a depth of 0.2 m from the surface and the lower layer is from 0.2–1.5 m depth. The upper layer is usually characterized as surface *SM* and the lower layer as root-zone *SM* (Seneviratne et al., 2010). The *SM* is commonly expressed either as a dimensionless ratio of two masses or two volumes or as a ratio of a mass per unit volume (Petropoulos, 2014). However, the relative values in this study are based on the base climatological reference period [1961–1990] (Raupach et al., 2009). Consequently, the study period is from January 1990–December 2016. The missing data were computed and replaced *via* monthly climatology and interpolation techniques (Beesley et al., 2009; Tozer et al., 2012).

In tandem with Table 2, variations in climatological patterns in upper and lower layer *SM* at geographically-diverse sites are apparent (Fig. 3a–b). SM_{UL} (Fig. 3a) shows a maximum during June–August (winter) and minimum during December–January (summer) and April (autumn). However, with SM_{LL} three sites (Sites 1, 2 and 3) showed no clear trend while Sites 4, 5, 6 and 7 occupied the largest magnitudes during the winter-spring transition periods (August–September) and

Table 5
Performance of EEMD and CEEMDAN-hybridized ELM and random forest models, during model development phase: training and validation periods, based on *r* = Pearson's correlation coefficient; *RMSE* = root mean square error and *MAE* = mean absolute error: a) Relative Soil Moisture - upper layer (SM_{UL}) and b) Relative Soil Moisture - lower layer (SM_{LL}).

a) SM_{UL}	Extreme Learning Machine (ELM)						Random Forest (RF)					
Optimum models at respective sites	Training			Validation			Training			Validation		
	r	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*
SITE: 1												
Standalone	0.641	0.070	0.055	0.591	0.085	0.066	0.829	0.056	0.044	0.607	0.082	0.063
EEMD	0.875	0.044	0.036	0.788	0.062	0.050	0.913	0.039	0.031	0.828	0.058	0.045
CEEMDAN	0.832	0.051	0.041	0.738	0.075	0.058	0.933	0.036	0.028	0.848	0.058	0.047
SITE: 2												
Standalone	0.666	0.074	0.058	0.563	0.092	0.074	0.821	0.060	0.049	0.539	0.092	0.075
EEMD	0.940	0.034	0.027	0.820	0.061	0.046	0.944	0.036	0.028	0.866	0.062	0.049
CEEMDAN	0.944	0.033	0.027	0.637	0.086	0.069	0.946	0.036	0.028	0.842	0.062	0.050
SITE: 3												
Standalone	0.542	0.087	0.067	0.212	0.103	0.082	0.820	0.067	0.052	0.397	0.094	0.072
EEMD	0.877	0.050	0.040	0.799	0.061	0.046	0.901	0.048	0.038	0.759	0.066	0.049
CEEMDAN	0.854	0.055	0.043	0.711	0.071	0.058	0.933	0.042	0.032	0.811	0.061	0.050
SITE: 4												
Standalone	0.527	0.101	0.082	0.287	0.112	0.092	0.839	0.072	0.059	0.301	0.111	0.089
EEMD	0.886	0.055	0.043	0.734	0.079	0.062	0.929	0.051	0.040	0.768	0.076	0.061
CEEMDAN	0.922	0.046	0.036	0.719	0.088	0.061	0.938	0.047	0.038	0.795	0.073	0.058
SITE: 5												
Standalone	0.789	0.101	0.078	0.479	0.126	0.099	0.891	0.080	0.064	0.471	0.125	0.094
EEMD	0.974	0.037	0.029	0.579	0.189	0.106	0.964	0.048	0.038	0.861	0.075	0.059
CEEMDAN	0.973	0.038	0.030	0.556	0.229	0.125	0.968	0.045	0.036	0.854	0.077	0.058
SITE: 6												
Standalone	0.766	0.106	0.084	0.508	0.124	0.089	0.897	0.079	0.064	0.455	0.127	0.091
EEMD	0.957	0.049	0.039	0.691	0.103	0.076	0.96	0.052	0.041	0.756	0.091	0.073
CEEMDAN	0.958	0.048	0.039	0.646	0.111	0.082	0.963	0.051	0.041	0.726	0.095	0.075
SITE: 7												
Standalone	0.394	0.122	0.101	0.489	0.121	0.101	0.643	0.104	0.083	0.404	0.125	0.100
EEMD	0.842	0.072	0.059	0.816	0.078	0.058	0.925	0.055	0.043	0.875	0.072	0.057
CEEMDAN	0.890	0.061	0.049	0.798	0.083	0.062	0.930	0.054	0.044	0.884	0.070	0.056

b) SM_{LL}	Extreme Learning Machine (ELM)						Random Forest (RF)					
Optimum models at respective sites	Training			Validation			Training			Validation		
	r	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*
SITE: 1												
Standalone	0.992	0.013	0.007	0.986	0.024	0.016	0.980	0.020	0.010	0.953	0.046	0.027
EEMD	0.999	0.004	0.003	0.972	0.035	0.019	0.995	0.010	0.005	0.975	0.044	0.029
CEEMDAN	0.975	0.023	0.017	0.980	0.037	0.030	0.995	0.010	0.005	0.971	0.050	0.038
SITE: 2												
Standalone	0.981	0.007	0.005	0.861	0.047	0.035	0.985	0.007	0.004	0.85	0.071	0.046
EEMD	0.998	0.002	0.002	0.991	0.013	0.008	0.994	0.004	0.003	0.932	0.064	0.044
CEEMDAN	0.997	0.003	0.002	0.522	0.105	0.051	0.994	0.004	0.003	0.909	0.056	0.040
SITE: 3												
Standalone	0.984	0.014	0.008	0.977	0.017	0.011	0.985	0.014	0.009	0.965	0.021	0.014
EEMD	0.999	0.003	0.002	0.993	0.010	0.006	0.996	0.007	0.005	0.986	0.016	0.010
CEEMDAN	0.999	0.003	0.003	0.995	0.008	0.005	0.995	0.008	0.005	0.988	0.016	0.010
SITE: 4												
Standalone	0.970	0.027	0.018	0.902	0.037	0.022	0.965	0.029	0.020	0.834	0.047	0.033
EEMD	0.997	0.009	0.006	0.872	0.042	0.020	0.992	0.014	0.010	0.963	0.029	0.022
CEEMDAN	0.996	0.010	0.007	0.904	0.036	0.020	0.993	0.014	0.009	0.960	0.028	0.022
SITE: 5												
Standalone	0.972	0.049	0.035	0.920	0.084	0.061	0.966	0.058	0.044	0.905	0.094	0.076
EEMD	0.997	0.017	0.013	0.982	0.040	0.028	0.990	0.033	0.025	0.971	0.062	0.049
CEEMDAN	0.997	0.017	0.013	0.971	0.051	0.030	0.990	0.031	0.023	0.969	0.058	0.046
SITE: 6												
Standalone	0.971	0.046	0.033	0.826	0.116	0.084	0.957	0.058	0.042	0.878	0.095	0.067
EEMD	0.993	0.023	0.016	0.922	0.092	0.055	0.982	0.037	0.028	0.934	0.093	0.069
CEEMDAN	0.991	0.026	0.017	0.874	0.117	0.061	0.988	0.032	0.024	0.942	0.096	0.074
SITE: 7												
Standalone	0.910	0.068	0.051	0.845	0.095	0.065	0.936	0.059	0.043	0.854	0.091	0.069

(*continued on next page*)

Table 5 (*continued*)

b) SM_{LL}	Extreme Learning Machine (ELM)						Random Forest (RF)					
Optimum models at respective sites	Training			Validation			Training			Validation		
	r	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*
EEMD	0.992	0.021	0.016	0.884	0.087	0.049	0.985	0.030	0.022	0.929	0.069	0.055
CEEMDAN	0.989	0.024	0.018	0.759	0.200	0.088	0.986	0.029	0.022	0.945	0.072	0.056

lowest values in May. Similarly, the least magnitude of upper layer *SM* is registered at Site 3-Wanaaring while Site 6-Jerrawa recorded the highest value (0.893) (Table 2). With, lower layer *SM*, however, the least value is recorded at Site 2-Balranald (0.034). Both Site 6-Jerrawa and Site 7-Rocky Creek recorded the highest levels (1.000). It is clear that upper and lower layer relative *SM* exhibit distinct patterns, and so, would be a worthy hydrological property to be tested through the respective models.

The skewness of upper layer soil moisture at all study sites were closer to zero confirming near-normal distributions. However, for lower layer *SM* except for Site 1-Menindee [skewness = 1.24] and Site 3-Wanaaring [skewness = 1.30], the other sites showed near-normal distributions. Similarly, kurtosis factors (kurt ≤ 3) of SM_{UL} and SM_{LL} at all sites illustrated that the distributions had fewer and less extreme outliers in comparison to normal distributions (Table 2).

3.2. Decomposition algorithms (EEMD & CEEMDAN) initialization parameters

Two important parameters; the ensemble number and the amplitude of added white noise, must be appropriately defined to attain optimum results and also to cancel out the added white noise series from the intact (*i.e.*, unresolved) signal. Given below is the statistical rule used to control the effect of the added white noise (Wu and Huang, 2009):

$$e_n = \frac{\varepsilon}{\sqrt{N}} \tag{13}$$

where the N = ensemble number, ε = amplitude of the added noise, and e_n = final standard deviation (SD), which is the difference between input signal and corresponding IMFs. The recommended amplitude of added white noise is 20% of SD (Wu and Huang, 2009). Consequently, based on previous similar studies (*e.g.*, (Ouyang et al., 2016; Ren et al., 2015; Wang et al., 2013)) for both EEMD and CEEMDAN, e_n = 0.2 while, for EEMD, N = 100 and for CEEMDAN, N = 20 (Torres et al., 2011) are used in this paper.

3.3. Predictive model development

With MATLAB running on Intel *i*7, 3.40 GHz processor, six forecast models namely, ELM, RF, EEMD-ELM, EEMD-RF, CEEMDAN-ELM, and CEEMDAN-RF were constructed to forecast two separate objective variables: monthly upper and lower layer soil moisture. There is no set rule for data partitioning (Deo et al., 2016), so the subsets in our study had training-70%, validation-15% and testing-15% (Table 3). For all model development, the data were serially divided, whereby the first 70% (218 months) were used to train the model, then the next 47 months for model validation and the final 47 months for model testing. The sequential division method was adopted to avoid distortion of the natural embedded frequencies within the soil moisture time series data. This is to allow the multi-resolution analysis utilities EEMD/CEEMDAN to appropriately unveil and extract these entrenched features for respective machine learning models which otherwise would not have been possible. A memory of several (lagged) months in *SM* time-series could result from serial correlation in time-space (*i.e.*, persistence) arising from hydro-meteorological factors (Chiew et al., 1998). Hence, partial autocorrelation function (*PACF*) has been adopted and lagged series with statistically significant relationship (*i.e.*, at 95% confidence interval) were screened as salient inputs (Ouyang et al., 2016; Ren et al., 2015; Seo and Kim, 2016; Wang et al., 2013).

Two different modeling techniques have been adopted for this study. Firstly, the conventional solitary modeling approach was adopted for the standalone ELM and RF models. For standalone models, *PACF* was applied to the monthly intact soil moisture time series (*i.e.*, the time series without EEMD/CEEMDAN analysis) and statistically significant set of predictor variables (Tables 4a-b) were determined. Then these were channeled into the ELM and RF models for forecasting of upper and lower layer soil moisture values at 1-month temporal resolution.

Alternatively, the ensemble modeling process as illustrated in Fig. 3 was used, which can be summarized as follows:

1. *EEMD/CEEMDAN decompositions*: The monthly intact SM_{UL} and SM_{LL} time series data (without EEMD/CEEMDAN analysis) were decomposed into respective monthly IMFs and a residual component using EEMD and CEEMDAN procedures, respectively. An example of the IMFs and the residual component from respective EEMD and CEEMDAN at Site 2, for SM_{UL}, has been illustrated in Fig. 5a–b.
2. *Input matrix creation*: *PACF* was applied to each of the monthly IMFs and residual component time series generated in the above phase. Salient lagged inputs of each IMF and residual component were determined. Individual input matrix was created for each IMF and the residual component containing its respective significant lags as summarized in Tables 4a–b for SM_{UL} and SM_{LL}, respectively.
3. *Ensemble forecasting*: These individual input matrices were used to forecast the respective future IMFs and the residual component using the ELM and RF models at a temporal resolution of 1 month. Then the forecasted IMFs and residual component were integrated at the end to generate the monthly forecasts of either SM_{UL} or SM_{LL} values. It must be noted that the EEMD and CEEMDAN transformations are purely self-adaptive and data dependent multi-resolution techniques. As such, the number of IMFs and the residual component generated are contingent upon the nature of data which in turn determines the ensemble numbers. Then optimal models were averred based on *r*, *RMSE*, and *MAE* during validation phases as described in Tables 5a–b with the least mean square error (*MSE*) for affirmation.
4. *Model testing and evaluation*: The model performances were evaluated using an independent testing data-set at each site using the model evaluation criteria described in Section 3.4.

For ELM, hidden neurons were set from 50 to 200 following earlier studies to avoid overfitting and unnecessarily large or small network architectures (Deo and Sahin, 2016; Deo et al., 2017a; Yaseen et al., 2016). For optimal feature extraction, various combination of transfer functions including sigmoidal, sine, hard-limit, triangular basis, and radial basis were implemented. The resulting models with unique architectures for both objective variables (SM_{UL} or SM_{LL}) are shown in Table 6a.

Equivalent random forest (RF) models were constructed to benchmark the ELM model. During training, the RF model registers three unique parameters: i) Delta Criterion Decision Split (*C*) showing the split criterion contributions; ii) Number of predictor split (N_{p}) showing

Table 6
Modeling frameworks of a) extreme learning machine (ELM) and b) random forest (RF), in forecasting relative soil moisture in the upper and lower layers.

a) ELM Optimum models at respective sites	Upper layer soil moisture (SM_{UL})				Lower layer soil moisture (SM_{LL})			
	No. of neurons			Transfer function	No. of neurons			Transfer function
	Input layer	Hidden layer	Output layer		Input layer	Hidden layer	Output layer	
SITE: 1								
ELM	3	177	1	*hardlim*	3	60	1	*tribas*
EEMD-ELM	50	113	1	*hardlim*	59	69	1	*tribas*
CEEMDAN-ELM	54	75	1	*hardlim*	49	59	1	*hardlim*
SITE: 2								
ELM	3	57	1	*tribas*	4	56	1	*tribas*
EEMD-ELM	63	60	1	sin	65	51	1	*sin*
CEEMDAN-ELM	55	75	1	*tribas*	45	50	1	*sig*
SITE: 3								
ELM	3	193	1	*hardlim*	2	81	1	*tribas*
EEMD-ELM	43	50	1	*tribas*	56	55	1	*sin*
CEEMDAN-ELM	53	61	1	*hardlim*	58	51	1	*sin*
SITE: 4								
ELM	4	75	1	*hardlim*	3	61	1	*tribas*
EEMD-ELM	58	124	1	*hardlim*	49	50	1	*sig*
CEEMDAN-ELM	56	55	1	*sig*	54	50	1	*sin*
SITE: 5								
ELM	6	117	1	*hardlim*	6	56	1	*sig*
EEMD-ELM	62	65	1	*sig*	61	50	1	*sig*
CEEMDAN-ELM	65	76	1	*sig*	56	52	1	*radbas*
SITE: 6								
ELM	5	89	1	*hardlim*	4	88	1	*sig*
EEMD-ELM	48	52	1	*sig*	50	55	1	*radbas*
CEEMDAN-ELM	52	52	1	*sig*	48	51	1	*sin*
SITE: 7								
ELM	1	81	1	*hardlim*	2	115	1	*hardlim*
EEMD-ELM	56	116	1	*hardlim*	57	68	1	*sig*
CEEMDAN-ELM	54	149	1	*hardlim*	47	51	1	*sin*

b) RF Optimum models at respective sites	Upper layer soil moisture (SM_{UL})			Lower layer soil moisture (SM_{LL})		
	Avg. delta criterion decision split (C)	Avg. number of predictor split (N_p)	Avg. permuted predictor delta error (E_D)	Avg. delta criterion decision split (C)	Avg. number of predictor split (N_p)	Avg. permuted predictor delta error (E_D)
SITE: 1						
RF	0.012	2169	0.917	0.022	2170	0.988
EEMD-RF	0.062	2300	0.730	0.119	2229	0.593
CEEMDAN-RF	0.078	2226	0.761	0.107	2405	0.743
SITE: 2						
RF	0.014	2158	0.897	0.004	2981	0.857
EEMD-RF	0.086	2274	0.635	0.077	2242	0.579
CEEMDAN-RF	0.101	2278	0.761	0.047	2265	0.765
SITE: 3						
RF	0.015	2159	0.680	0.020	3259	1.374
EEMD-RF	0.103	2142	0.775	0.139	2432	0.626
CEEMDAN-RF	0.096	2371	0.768	0.124	2398	0.712
SITE: 4						
RF	0.018	2743	0.482	0.025	2167	1.020
EEMD-RF	0.128	2226	0.686	0.148	2350	0.715
CEEMDAN-RF	0.102	2331	0.763	0.089	2306	0.733
SITE: 5						
RF	0.027	1878	0.604	0.055	1920	0.840
EEMD-RF	0.129	2261	0.689	0.117	2250	0.645
CEEMDAN-RF	0.084	2270	0.728	0.106	2135	0.759
SITE: 6						
RF	0.033	2246	0.744	0.064	2844	1.234
EEMD-RF	0.121	2128	0.724	0.123	2279	0.702
CEEMDAN-RF	0.081	2149	0.750	0.094	2211	0.764
SITE: 7						
RF	0.062	6508	0.720	0.080	3222	1.623
EEMD-RF	0.096	2222	0.686	0.130	2214	0.649
CEEMDAN-RF	0.093	2311	0.778	0.116	2268	0.768

Table 7

Performance evaluation of EEMD and CEEMDAN-hybridized ELM and random forest (RF) models in forecasting of the upper layer and lower layer soil moisture during the testing period, based on r = Pearson's correlation coefficient; *RMSE* = root mean square error and *MAE* = mean absolute error. (Maximum r and minimum *RMSE* & *MAE* are **boldfaced**.)

Sites	Upper Layer soil moisture (SM_{UL})						Lower Layer soil moisture (SM_{LL})					
	ELM			RF			ELM			RF		
	r	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*
SITE: 1												
Standalone	0.731	0.084	0.059	0.635	0.097	0.071	0.963	0.018	0.012	0.856	0.037	0.021
EEMD	0.855	0.063	0.045	0.842	0.066	0.050	**0.983**	**0.013**	**0.009**	0.965	0.025	0.017
CEEMDAN	**0.870**	**0.061**	**0.046**	0.845	0.065	0.049	0.909	0.032	0.025	0.835	0.050	0.043
SITE: 2												
Standalone	0.695	0.079	0.056	0.641	0.086	0.067	0.952	0.008	0.005	0.917	0.011	0.009
EEMD	**0.925**	**0.042**	**0.034**	0.875	0.056	0.043	**0.992**	**0.003**	**0.002**	0.866	0.017	0.015
CEEMDAN	0.897	0.049	0.035	0.878	0.059	0.046	0.965	0.008	0.006	0.889	0.018	0.016
SITE: 3												
Standalone	0.602	0.098	0.071	0.590	0.104	0.076	0.977	0.015	0.009	0.948	0.023	0.014
EEMD	**0.862**	**0.060**	**0.050**	0.803	0.074	0.059	**0.993**	**0.008**	**0.005**	0.958	0.037	0.031
CEEMDAN	0.839	0.065	0.051	0.829	0.075	0.059	0.993	0.008	0.006	0.983	0.027	0.022
SITE: 4												
Standalone	0.603	0.117	0.087	0.497	0.126	0.098	0.960	0.034	0.022	0.884	0.064	0.041
EEMD	0.834	0.082	0.063	0.844	0.090	0.072	**0.995**	**0.013**	**0.010**	0.975	0.053	0.035
CEEMDAN	**0.877**	**0.070**	**0.056**	0.864	0.085	0.068	0.990	0.018	0.013	0.969	0.054	0.037
SITE: 5												
Standalone	0.762	0.122	0.090	0.724	0.131	0.102	0.965	0.052	0.041	0.883	0.09	0.076
EEMD	0.942	0.065	0.047	0.921	0.091	0.071	**0.992**	**0.024**	**0.018**	0.973	0.049	0.042
CEEMDAN	**0.948**	**0.061**	**0.042**	0.919	0.086	0.063	0.987	0.031	0.025	0.968	0.058	0.048
SITE: 6												
Standalone	0.686	0.142	0.097	0.686	0.145	0.115	0.943	0.065	0.048	0.914	0.083	0.067
EEMD	**0.943**	0.066	0.056	0.904	0.096	0.076	**0.993**	**0.023**	**0.018**	0.954	0.070	0.056
CEEMDAN	**0.943**	**0.065**	**0.054**	0.896	0.101	0.083	0.987	0.033	0.025	0.961	0.075	0.057
SITE: 7												
Standalone	0.451	0.133	0.101	0.353	0.141	0.105	0.892	0.075	0.051	0.862	0.086	0.066
EEMD	**0.860**	**0.076**	**0.058**	0.833	0.088	0.068	**0.985**	**0.029**	**0.021**	0.958	0.067	0.043
CEEMDAN	0.855	0.081	0.061	0.832	0.089	0.071	0.980	0.033	0.025	0.967	0.059	0.039

the number of decision splits and; iii) Permuted Predictor Delta Error (E_D) showing the variable importance to the prediction error. Table 6b shows the averages of these values from optimum models in forecasting of both SM_{UL} and SM_{LL}.

3.4. Model evaluation benchmarks

Comprehensive and robust model assessment requires both objective and subjective evaluations (Dawson et al., 2007) as no single statistical measure is purely definitive (Chai and Draxler, 2014; Dawson et al., 2007). Thus, a wide range of statistical metrics are used whose equations are as follows (Legates and McCabe, 1999; Legates and McCabe, 2013; Nash and Sutcliffe, 1970; Shamseldin, 1997; Willmott, 1981; Willmott, 1984):

i. Correlation coefficient (r):

$$r = \frac{\sum_{i=1}^{N} (SM_{UL}^{OBS,i} - \overline{SM_{UL}^{OBS}})(SM_{UL}^{FOR,i} - \overline{SM_{UL}^{FOR}})}{\sqrt{\sum_{i=1}^{N} (SM_{UL}^{OBS,i} - \overline{SM_{UL}^{OBS}})^2}\sqrt{\sum_{i=1}^{N} (SM_{UL}^{FOR,i} - \overline{SM_{UL}^{FOR}})^2}}, (-1 \le r \le 1) \tag{14}$$

ii. Root mean square error (*RMSE*):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (SM_{UL}^{FOR,i} - SM_{UL}^{OBS,i})^2}, (0 \le RMSE\le) \tag{15}$$

iii. Mn absolute error (*MAE*):

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |(SM_{UL}^{FOR,i} - SM_{UL}^{OBS,i})|, (0 \le MAE\le) \tag{16}$$

iv. Willmott's Index (*WI*):

$$WI = 1 - \left[\frac{\sum_{i=1}^{N} (SM_{UL}^{OBS,i} - SM_{UL}^{FOR,i})^2}{\sum_{i=1}^{N} (|SM_{UL}^{FOR,i} - \overline{SM_{UL}^{OBS}}| + |SM_{UL}^{OBS,i} - \overline{SM_{UL}^{OBS}}|)^2}\right], (0 \le WI \le 1) \tag{17}$$

v. Nh–Sutcliffe Efficiency (E_{NS}):

$$E_{NS} = 1 - \left[\frac{\sum_{i=1}^{N} (SM_{UL}^{OBS,i} - SM_{UL}^{FOR,i})^2}{\sum_{i=1}^{N} (SM_{UL}^{OBS,i} - \overline{SM_{UL}^{OBS}})^2}\right], (-\infty < E_{NS} \le 1) \tag{18}$$

vi. Legates-McCabe's Index (L):

$$L = 1 - \left[\frac{\sum_{i=1}^{N} |SM_{UL}^{FOR,i} - SM_{UL}^{OBS,i}|}{\sum_{i=1}^{N} |SM_{UL}^{OBS,i} - \overline{SM_{UL}^{OBS}}|}\right], (-\infty < L \le 1) \tag{19}$$

In these equations, ${SM_{UL}}^{OBS}$ = observed upper layer (UL) soil moisture and ${SM_{UL}}^{FOR}$ = forecasted upper layer soil moisture, i = occurrence time and N = total number of data points. (Subscript UL is replaced with LL for lower layer SM.

Table 8
Performance evaluation of EEMD and CEEMDAN-hybridized ELM and random forest (RF) models during the testing period, based on *WI* = Willmott's Index; E_{NS} = Nash–Sutcliffe efficiency and *L* = Legates-McCabe's index, in forecasting upper and lower layer soil moisture. The models with largest *L* at each site have been shown in **boldface**.

Optimum models at respective sites	Upper layer soil moisture (SM_{UL})						Lower layer soil moisture (SM_{LL})					
	ELM			Random Forest			ELM			Random Forest		
	WI	E_{NS}	*L*	*WI*	E_{NS}	*L*	*WI*	E_{NS}	*L*	*WI*	E_{NS}	*L*
SITE: 1												
Standalone	0.593	0.485	0.390	0.300	0.315	0.273	0.973	0.925	0.733	0.831	0.695	0.534
EEMD (Hybrid)	**0.837**	**0.710**	**0.534**	0.813	0.680	0.483	**0.988**	**0.961**	**0.811**	0.934	0.863	0.626
CEEMDAN (Hybrid)	0.854	0.729	0.523	0.828	0.692	0.493	0.902	0.776	0.453	0.761	0.443	0.050
SITE: 2												
Standalone	0.625	0.470	0.387	0.502	0.380	0.275	0.950	0.905	0.755	0.920	0.827	0.616
EEMD (Hybrid)	**0.934**	**0.854**	**0.629**	0.839	0.731	0.529	**0.993**	**0.984**	**0.891**	0.822	0.558	0.331
CEEMDAN (Hybrid)	0.906	0.797	0.617	0.817	0.711	0.505	0.963	0.909	0.749	0.794	0.525	0.291
SITE: 3												
Standalone	0.414	0.317	0.240	0.219	0.242	0.188	0.983	0.953	0.823	0.956	0.893	0.740
EEMD (Hybrid)	**0.893**	**0.743**	**0.470**	0.783	0.616	0.375	**0.995**	**0.986**	**0.906**	0.865	0.726	0.423
CEEMDAN (Hybrid)	0.850	0.699	0.458	0.738	0.605	0.374	0.995	0.986	0.893	0.924	0.847	0.587
SITE: 4												
Standalone	0.446	0.332	0.251	0.208	0.219	0.150	0.970	0.918	0.750	0.836	0.706	0.543
EEMD (Hybrid)	0.813	0.671	0.459	0.713	0.601	0.378	**0.996**	**0.987**	**0.888**	0.892	0.803	0.604
CEEMDAN (Hybrid)	**0.888**	**0.763**	**0.518**	0.770	0.648	0.410	0.992	0.978	0.856	0.885	0.791	0.578
SITE: 5												
Standalone	0.759	0.569	0.430	0.653	0.504	0.357	0.970	0.923	0.752	0.881	0.772	0.535
EEMD (Hybrid)	0.941	0.877	0.699	0.843	0.761	0.548	**0.993**	**0.984**	**0.890**	0.968	0.933	0.747
CEEMDAN (Hybrid)	**0.949**	**0.893**	**0.733**	0.876	0.786	0.599	0.989	0.973	0.849	0.951	0.905	0.707
SITE: 6												
Standalone	0.671	0.469	0.369	0.609	0.450	0.255	0.949	0.889	0.711	0.905	0.824	0.597
EEMD (Hybrid)	0.952	0.887	0.635	0.854	0.757	0.506	**0.994**	**0.986**	**0.890**	0.927	0.873	0.663
CEEMDAN (Hybrid)	**0.953**	**0.889**	**0.648**	0.830	0.731	0.463	0.987	0.972	0.846	0.908	0.854	0.657
SITE: 7												
Standalone	0.385	0.203	0.130	0.322	0.099	0.099	0.919	0.795	0.563	0.874	0.731	0.444
EEMD (Hybrid)	**0.888**	**0.738**	**0.505**	0.797	0.648	0.412	**0.990**	**0.969**	**0.824**	0.914	0.837	0.639
CEEMDAN (Hybrid)	0.892	0.706	0.475	0.782	0.64	0.393	0.986	0.960	0.789	0.938	0.873	0.672

The first evaluation metric, Pearson's correlation coefficient (*r*) provides information on the strength and direction on the agreement between $SM_{UL}^{OBS}(/SM_{LL}^{OBS})$ and $SM_{UL}^{FOR}(/SM_{LL}^{FOR})$, yet it is limited to linear association of forecasted and observed data. The absolute error measures, *RMSE* and *MAE*, glean information on the average discrepancies between forecasted and observed values (Legates and McCabe, 1999). However, *MAE* does not provide information about under/over-predictions, while *RMSE* is oversensitive to peak *SM* levels and insensitive to low levels (Hora and Campos, 2015; Willems, 2009).

With similar bounds [0 ↔ 1] as correlation-based measures goodness-of-fit, *WI* is advantageous, yet, it lacks meaningful zero in providing a convenient reference point (Dawson et al., 2007) causing obscured physical meaning. The most popular metric, E_{NS} is dimensionless and scaled version of mean squared error (Willems, 2009), offering a better physical interpretation of the goodness-of-fit: 1 = perfect model; 0 = no predictive advantage; and negative values when forecasted values diverge (Legates and McCabe, 2013; Mehr et al., 2013). Owing to the squared values of residual terms, both *WI* and E_{NS} are oversensitive to the peak residual values (Legates and McCabe, 1999; Willems, 2009; Willmott, 1981). In comparison, the Legate-McCabe's index (*L*) is not overestimated since it takes absolute values into account and gives errors and differences the appropriate weights (Legates and McCabe, 1999). *L* is also simple, easy to interpret and is acclaimed to yield a relative assessment of model performances (Legates and McCabe, 1999).

We note the absolute error measures (*RMSE* and *MAE*) are in real units which limit their ability to assess the model performances across various case study sites. Hence, the percentage error measures *viz.*, relative root mean square error (*RRMSE*) and mean absolute percentage error (*MAPE*) are used, as shown below:

I. Relative root mean square error (*RRMSE*, %):

$$RRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(SM_{UL}^{FOR,i} - SM_{UL}^{OBS,i})^2}}{\frac{1}{N}\sum_{i=1}^{N}(SM_{UL}^{OBS,i})} \times 100 \tag{20}$$

II. Mean absolute percentage error (*MAPE*; %):

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{(SM_{UL}^{FOR,i} - SM_{UL}^{OBS,i})}{SM_{UL}^{OBS,i}}\right| \times 100 \tag{21}$$

[*NB: The symbols used have the same meaning as mentioned above.*]

RMSE and *MAE* are used to determine model performance with simultaneous monitoring of correlation *r* whereas *WI*, $E_{NS,}$ and *L* provides further goodness-of-fit assessments and eventually, *RRMSE* and *MAPE* compared models at different study sites. However, the limitation of the above mentioned objective metrics is the quantification of assessment in a few numbers (Willems, 2009). Thus, to get a better insight, subjective model performance assessments *via* various diagnostic plots *e.g.*, box-plots, forecasting error histogram, time series graphs and polar plots are also performed.

4. Results and general discussion

This section provides results for an extensive evaluation of the proposed hybrid EEMD-ELM model against standalone ELM, hybrid CEEMDAN-ELM and the equivalent random forest models (*i.e.*, RF,

Table 9
Model comparison at different sites using relative error measures, *RRMSE* and *MAPE*. The optimal model with lowest relative (%) error has been shown in **boldface**.

Optimum models at respective sites	Upper Layer soil moisture (SM_{UL})				Lower Layer soil moisture (SM_{LL})			
	ELM		Random Forest		ELM		Random Forest	
	RRMSE (%)	*MAPE (%)*	*RRMSE (%)*	*MAPE (%)*	*RRMSE (%)*	*MAPE (%)*	*RRMSE (%)*	*MAPE (%)*
SITE: 1								
Standalone	49.80	46.83	57.45	58.17	**4.60**	**2.83**	**9.30**	**4.75**
EEMD (Hybrid)	37.40	34.60	39.28	48.41	**3.31**	**2.04**	**6.23**	**4.02**
CEEMDAN (Hybrid)	36.16	48.73	38.53	49.35	7.97	6.14	12.6	11.00
SITE: 2								
Standalone	46.31	55.37	50.11	67.50	8.17	5.38	11.00	8.41
EEMD (Hybrid)	24.29	36.12	33.01	43.60	3.37	2.44	17.60	17.30
CEEMDAN (Hybrid)	28.64	38.97	34.22	48.42	7.99	5.99	18.20	18.60
SITE: 3								
Standalone	58.29	66.17	61.42	66.72	6.75	3.90	10.20	5.48
EEMD (Hybrid)	35.73	57.02	43.71	59.10	3.74	2.06	16.30	14.70
CEEMDAN (Hybrid)	38.68	45.52	44.33	53.63	**3.75**	**2.48**	12.20	9.88
SITE: 4								
Standalone	51.96	64.54	56.19	66.02	10.90	6.74	20.60	10.70
EEMD (Hybrid)	36.47	54.45	40.16	58.59	4.25	3.18	16.80	9.82
CEEMDAN (Hybrid)	30.96	54.04	37.72	59.23	5.64	3.88	17.30	10.70
SITE: 5								
Standalone	**38.14**	**42.95**	**40.92**	53.13	8.75	7.32	15.10	13.10
EEMD (Hybrid)	20.39	**17.56**	28.43	32.99	4.01	3.14	8.14	7.28
CEEMDAN (Hybrid)	**18.99**	**18.34**	**26.87**	33.20	5.14	4.62	**9.74**	8.40
SITE: 6								
Standalone	41.55	45.67	42.28	50.87	10.50	8.22	13.20	10.90
EEMD (Hybrid)	**19.16**	23.60	**28.12**	33.44	3.72	3.14	11.20	9.24
CEEMDAN (Hybrid)	19.00	24.47	29.58	35.75	5.30	4.20	12.00	8.99
SITE: 7								
Standalone	44.55	46.23	47.37	**47.84**	16.30	11.00	18.6	13.50
EEMD (Hybrid)	25.53	25.36	29.60	**31.25**	6.30	4.73	14.50	8.06
CEEMDAN (Hybrid)	27.05	27.23	29.96	**30.92**	7.21	5.43	12.80	**7.45**

EEMD-RF, and CEEMDAN-RF). The evaluation is carried out after forecasting upper and lower layer soil moisture at seven hydrological sites. EEMD-ELM models are optimized using a combination of activation functions and hidden layer neurons that yielded the lowest *RMSE* and *MAE* during the validation stage to screen the optimal models. Based on statistical metrics in Eqs. (14)–(21) and diagnostic (visual) plots, the justifications of the results are made.

Evaluation of ELM and RF integrated with EEMD and CEEMDAN are provided in Table 7, in terms of conventional metrics *r*, *RMSE*, and *MAE*. In forecasting the upper layer *SM*, three sites (Sites 1, 4 and 5) registered the maximum *r*-value from CEEMDAN-ELM models, while three sites (Sites 2, 3 and 7) had the highest value reported by EEMD-ELM models. Interestingly, Site 6 recorded the same magnitude ($r = 0.943$) from both models. Consequently, the lowest *RMSE* and *MAE* values were attained by the hybrid EEMD-ELM model at three sites (Sites 2, 3 and 7) and the other four sites (Sites 1, 4, 5 and 6) showed the lowest errors from the hybrid CEEMDAN-ELM model. It is evident thus far that the hybridized ensemble ELM and RF models indeed are better in comparison to their standalone counterparts. However, due to the rather unclear outcomes from these metrics, the decision to determine the optimal ensemble model can be obscured. In contrast, for lower layer *SM* forecasting, all three measures (*r*, *RMSE*, and *MAE*) consistently displayed the superiority of the hybrid EEMD-ELM with the largest *r* and lowest *RMSE*, and *MAE* values attained at all seven sites (Table 7). In comparison to the ELM model, the highest percentage increase in *r* was recorded at Site 7 (10.4%), whiles the lowest *RMSE*0.003 and *MAE*0.002, were recorded by the hybrid EEMD-ELM model for Site 2. The accuracy of the other data-intelligent models (including the hybrid CEEMDAN-ELM/CEEMDAN-RF and EEMD-RF) were disparate and confirmed that the hybrid EEMD-ELM model had a better potential to generate accurate SM_{LL} forecasts.

Numerical quantification of model performances using Willmott's Index (*WI*), the Nash–Sutcliffe Efficiency (E_{NS}), and Legates-McCabe's Index (*L*), which ideally are unity for perfect models, showed that the hybrid EEMD-based and CEEMDAN-based ensemble hybrids demonstrated a dramatic improvement in comparison to the standalone models for SM_{UL} (Table 8). Comparing the hybrid EEMD-ELM model with the standalone ELM noted the smallest increment in the value of *L* by about 36.9% at Site 1 and a significant increment by about 288.5% at Site 7. Similarly, the value of *WI* was about 0.503 (*i.e.*, incremented by 130.6%) and E_{NS} was 0.535 (263.5%) at Site 7. The hybrid CEEMDAN-ELM model did register higher values of *L* than the ELM models with the highest percentage increase of about 265.4% (Site 7) and the lowest increase of about 34.1% (Site 1). *WI* and E_{NS} showed a similar increase, however, these percentage increments were comparatively lower than the hybrid EEMD-ELM model. More closely with the value of *L* taking the precedence on the basis of benefits discussed earlier, it can be noted that hybrid EEMD-ELM model has had a better performance in forecasting upper layer soil moisture (SM_{UL}) at four sites (Sites 1, 2, 3 and 7) than the hybrid CEEMDAN-ELM and RF counterparts. Although at the three sites (Sites 4, 5 and 6) the CEEMDAN-ELM had a slightly better performance, the superiority of the hybrid EEMD-ELM in comparison with the hybrid CEEMDAN-ELM was demonstrated by increments in *L* value, of about 2.1%, 1.9%, 2.6% and 5.9% at Sites 1, 2, 3 and 7, respectively. A similar trend was consistently demonstrated by the value of *WI* and E_{NS}, despite their notable drawbacks (Section 3.4). Likewise, in forecasting lower layer soil moisture, the measures of *WI*, $E_{NS,}$ and *L*, (Table 8) unanimously revealed the supremacy in the performance of the hybrid EEMD-ELM at all seven study sites without any contention from the hybrid CEEMDAN-ELM, CEEMDAN-RF and the hybrid EEMD-RF models, while the standalone ELM and RF were beyond question in terms of their

Table 10
Statistical analysis in terms of maximum, minimum, lower quartile-LQ (Q_{25}), median (Q_{50}), upper quartile-UQ (Q_{75}), mean and range of observed and forecasted values during the test datasets of upper and lower soil moisture from the best ELM, EEMD-ELM and CEEMDAN ELM and its RF counterparts. (NB: Being relative values the *SM* is dimensionless.)

	Upper Layer soil moisture (SM_{UL})									Lower Layer soil moisture (SM_{LL})								
	Standalone (Site 5)			EEMD ensemble (Site 6)			CEEMDAN ensemble (Site 5)			Standalone (Site 1)			EEMD ensemble (Site 1)			CEEMDAN ensemble (Site 3)		
	OBS	ELM	RF	OBS	ELM	RF	OBS	ELM	RF	OBS	ELM	RF	OBS	ELM	RF	OBS	ELM	RF
Maximum	0.852	0.692	0.506	0.939	0.872	0.686	0.852	0.743	0.605	0.631	0.647	0.576	0.631	0.663	0.542	0.412	0.395	0.339
Minimum	0.035	0.001	0.135	0.061	0.070	0.117	0.035	0.021	0.101	0.305	0.303	0.306	0.305	0.313	0.319	0.144	0.141	0.172
Upper quartile	0.171	0.185	0.206	0.174	0.192	0.239	0.171	0.174	0.214	0.366	0.366	0.363	0.366	0.367	0.382	0.176	0.177	0.193
Median	0.264	0.283	0.280	0.342	0.304	0.310	0.264	0.249	0.291	0.386	0.382	0.383	0.386	0.387	0.396	0.206	0.204	0.219
Lower quartile	0.471	0.436	0.423	0.458	0.445	0.404	0.471	0.450	0.421	0.418	0.417	0.407	0.418	0.415	0.426	0.252	0.252	0.264
Mean	0.320	0.306	0.305	0.343	0.335	0.330	0.320	0.307	0.318	0.400	0.398	0.390	0.400	0.402	0.403	0.224	0.224	0.231
Range	0.817	0.691	0.371	0.878	0.802	0.570	0.817	0.722	0.504	0.325	0.344	0.270	0.325	0.350	0.223	0.268	0.253	0.166

Table 11
Modeling time of EEMD and CEEMDAN-hybridized extreme learning machine (ELM) and random forest (RF) models for a) Relative Soil Moisture - upper layer (SM_{UL}) and b) Relative Soil Moisture - lower layer (SM_{LL}).

Optimum models at respective sites	a) Modeling time for SM_{UL} (seconds)		b) Modeling time for SM_{LL} (seconds)	
	ELM	RF	ELM	RF
SITE: 1				
Standalone	0.155	2.150	0.121	1.754
EEMD	0.516	17.149	0.544	16.703
CEEMDAN	0.488	19.215	0.484	17.477
SITE: 2				
Standalone	0.228	1.857	0.197	2.012
EEMD	0.599	17.613	0.542	17.817
CEEMDAN	0.651	18.066	0.404	15.565
SITE: 3				
Standalone	0.159	1.844	0.113	1.568
EEMD	0.535	14.491	0.677	15.949
CEEMDAN	0.584	18.722	0.524	18.520
SITE: 4				
Standalone	0.169	1.834	0.107	1.646
EEMD	0.666	17.072	0.469	15.546
CEEMDAN	0.519	18.000	0.539	17.738
SITE: 5				
Standalone	0.116	1.834	0.185	1.893
EEMD	0.490	17.686	0.473	16.464
CEEMDAN	0.465	19.083	0.536	17.476
SITE: 6				
Standalone	0.196	1.836	0.160	1.751
EEMD	0.411	14.924	0.704	15.362
CEEMDAN	0.562	16.557	0.424	15.933
SITE: 7				
Standalone	0.124	1.445	0.142	1.558
EEMD	0.694	15.617	0.429	16.628
CEEMDAN	0.790	18.545	0.431	15.747

comparison with the hybrid-equivalent models. The maximum magnitude of $L = 0.906$ was recorded at Site 2, and the highest $E_{NS} = 0.987$ and $WI = 0.996$ were registered at Site 4, by the respective hybrid EEMD-ELM model. Intriguingly, the hybrid EEMD-ELM model at all the sites ensued $L > 0.800$ with the lowest value of $L = 0.811$ (Site 1). Hence, it is evident that the hybridization based on EEMD acted to enhance the performance of the standalone ELM models for both the upper and lower layer *SM* forecasting.

The statistical performances for all sites were disparate in terms of the range of performance metrics attained. EEMD-ELM model, in forecasting SM_{UL}, exhibited the largest correlation ($r = 0.943$) at Site 6, while the lowest $RMSE = 0.060$ was recorded at Site 3 and the lowest $MAE = 0.034$ at Site 2. Site 6 recorded the largest WI (≈ 0.952), E_{NS} (≈ 0.887) and Site 5 registered the largest L (≈ 0.699). Similarly, the metrics of SM_{LL} were also incongruent with the highest $r = 0.995$ at Site 4 and the least $RMSE = 0.003$ and $MAE = 0.002$ at Site 2. These results ascertain that the performance of the EEMD-ELM (and the RF counterpart) model is not universally similar when the study sites are having different geographical, physical and climatic characteristics as depicted in Tables 1 and 2 and Fig. 2.

To enable model evaluations at geographically diverse sites, the relative measures (*i.e.*, *RRMSE* and *MAPE*) must alternatively be used. Evidently, the results in Table 9 exhibited that the lowest value of *RRMSE* for all seven sites from the ELM model were apparently lower than the RF counterparts. In comparison with the hybrid EEMD-RF models, the hybrid EEMD-ELM showed the largest reduction in *RRMSE* value for Site 6 (−8.96%) while Site 1 (−1.88%) had the lowest reduction. Accordingly, the hybrid EEMD-ELM had the best performance for Site 6 ($RRMSE = 19.16\%$), while the hybrid CEEMDAN-ELM model registered the smallest value of relative error at Site 5

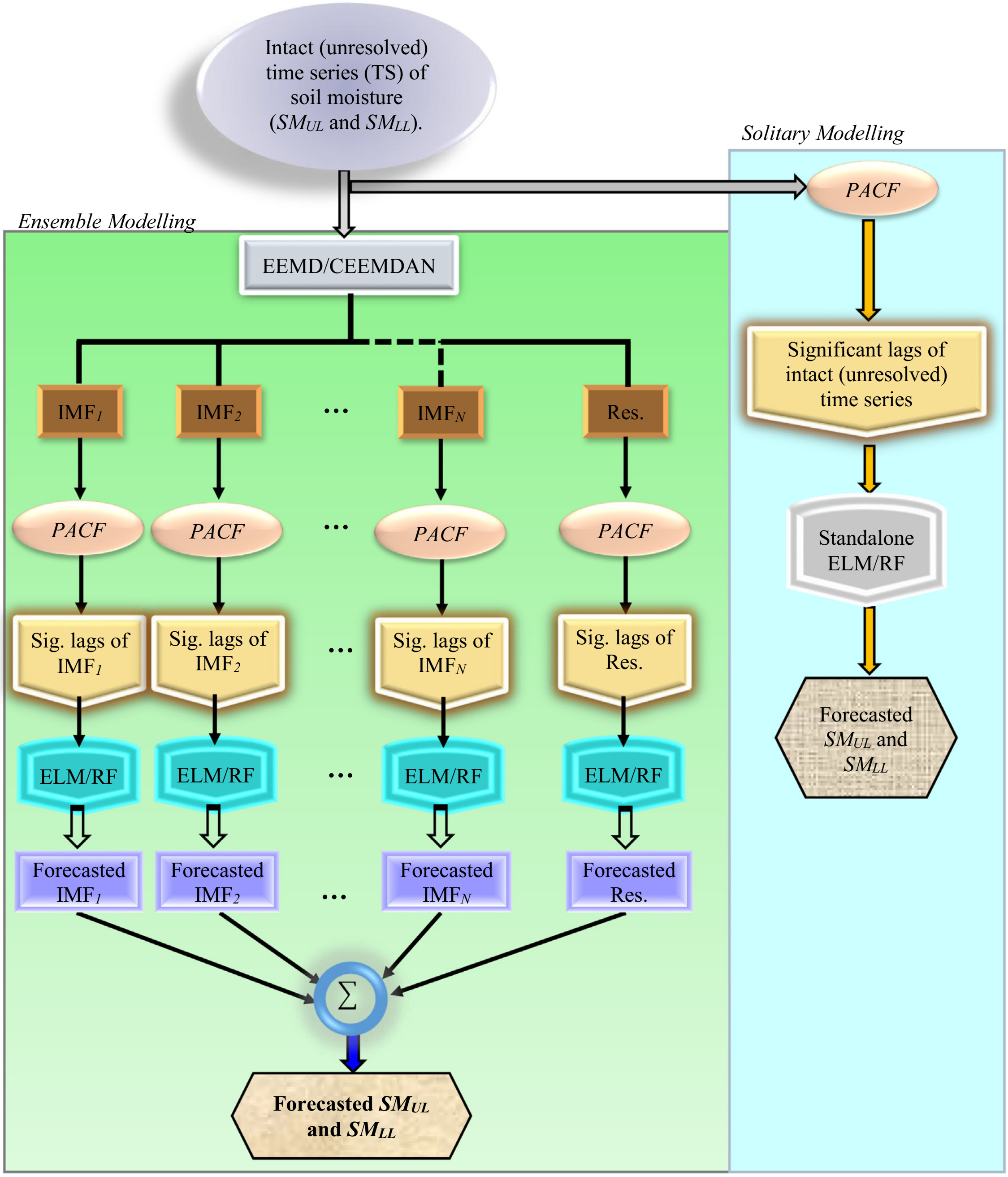


Fig. 4. A schematic view of the model development process. (The definitions of acronyms used here are as follows: SM_{UL} – upper layer soil moisture, SM_{LL} – lower layer soil moisture, IMF – intrinsic mode functions, subscript *N* represents the IMF number(s), *PACF*– partial auto-correlation function, Sig. – significant, Res. – residual, ELM –extreme learning machine, RF – random forest.)

(*RRMSE* = 18.99%) when applied for forecasting upper layer soil moisture. Furthermore, the comparison of models at respective sites in forecasting SM_{LL} revealed that the hybrid EEMD-ELM model is able to generate the best results for Site 1 (*RRMSE* = 3.31% and *MAPE* = 2.04%). The lowest relative errors recorded by the hybrid CEEMDAN-ELM models was for Site 3 (*RRMSE* = 3.75% and *MAPE* = 2.48%), showing unarguably that the EEMD-ELM model can yield very good accuracy. In accordance with the outcomes from absolute measures, the relative measures concur on the suitability of the hybrid EEMD-ELM for both upper and lower layer *SM* forecasting.

Further analysis of the spread of forecasting errors (FE) was carried out to assess the capability of the best hybrid EEMD-ELM model (Site 6) for forecasting upper layer soil moisture in the test period. FE is the difference between the forecasted and observed *SM* (FE = $SM_{UL}^{FOR} - SM_{UL}^{OBS}$). Ideally, FE must be 0, hence a better model is bound to have higher frequencies of the forecasting error value closer to zero. Fig. 6 plots a histogram showing the percentage frequency distribution of FE computed in error brackets of 0.1 step-sizes. This could assist in the understanding of model accuracy for practical applications (Deo et al., 2016). It clearly shows that the ELM models outperformed the respective RF models since the FE values display a narrower and a closer location to the zero frequency distribution. An in-depth examination shows that the hybrid EEMD-ELM registered the highest percentage of FE (91%) in the first bin ($0 < \text{FE} \leq 0.1$) followed by the hybrid CEEMDAN-ELM (89%) and then the standalone ELM model (64%). Accumulation of the percentages reveal that the total (100%) of all FE values from the hybrid EEMD-ELM was below 0.2, while the hybrid CEEMDAN-ELM yielded a total of 98% of percentage

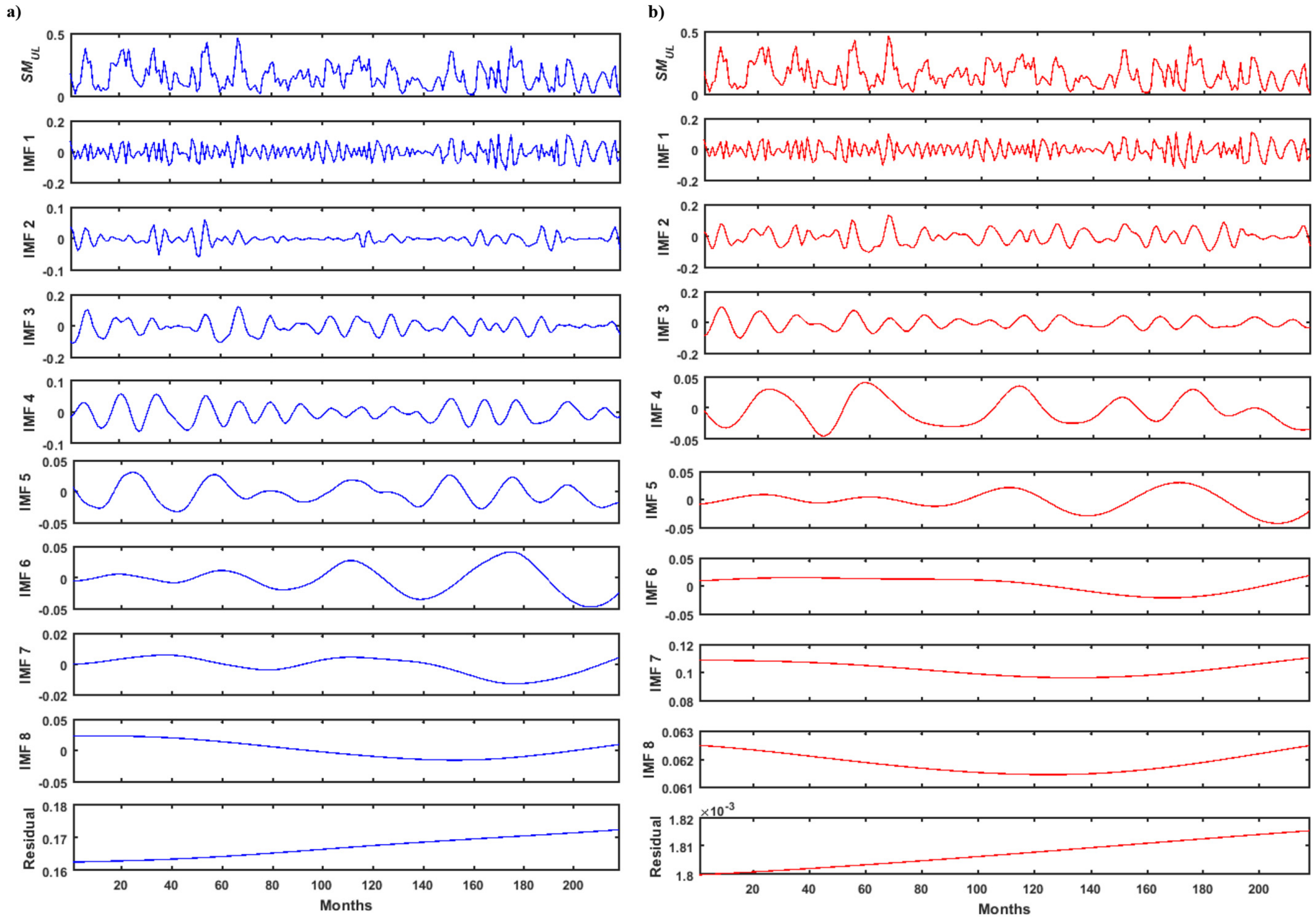


Fig. 5. Temporal waveforms of IMFs and the residual from a) EEMD and b) CEEMDAN transformation of intact (*i.e.*, unresolved) time series (TS) (lag 0) of upper layer soil moisture (SM_{UL}) at Site 2 during the training period. The intact upper layer soil moisture TS has also been plotted for comparison. (The definitions of acronyms used here are as follows: SM_{UL} – upper soil moisture, IMF – intrinsic mode functions.)

errors in this band. This supports the suitability of the hybrid EEMD-ELM model for forecasting upper layer soil moisture.

Next, the evaluation of the best hybrid EEMD-ELM was undertaken with a time-series plot of upper layer *SM* (Fig. 7). These results confirm that the SM_{UL} forecasts from the standalone models (*i.e.*, ELM and RF) divert from the observed values, while the best hybrid ELM (both EEMD and CEEMDAN based) are able to attain a better accuracy. A closer examination showed that in congruence with the key statistical metrics (*i.e.*, Table 8), there was a very good visual agreement between observed data and forecasted SM_{UL} generated by the best hybrid EEMD-ELM model. In addition, the forecasts generated by the best hybrid EEMD-ELM recorded the highest number of points within one standard deviation (*i.e.*, 38 out of 47) to confirm its superiority.

Visual evaluation of forecasted upper layer soil moisture was performed with scatter-plots and the coefficient of determination (R^2) in the testing set (Fig. 8a). The optimal hybrid EEMD-ELM and the hybrid CEEMDAN-ELM models are seen to register identical R^2 values of 0.889 and 0.899, respectively. With that, the gradient (m) of the linear fit which is an alternative model performance metric, was found to be very close to unity (*i.e.*, 0.893) in case of the best hybrid EEMD-ELM model. On the other hand, the y-intercept, which should ideally be zero, registered by the best hybrid EEMD-ELM was 0.029 whereas best hybrid CEEMDAN-ELM registered a y-intercept of 0.027. To further affirm the accuracy in forecasting lower layer soil moisture (SM_{LL}), Fig. 8b displays the scatter plots of the best hybrid EEMD-ELM models (Site 1). With $R^2 = 0.966$, it can be assuredly established that a huge 96.6% of the observed SM_{LL} values were forecasted using the best hybrid EEMD-ELM model. In congruence with the results presented in Table 8, ELM with the implementation of the self-adaptive multi-resolution analysis utility, EEMD, yet again has enhanced forecasting.

Compelling evidence of the superiority of the hybrid EEMD-ELM model, in terms of *SM* forecasting accuracy, has been noted so far. However, for practical applications such as precision agriculture, seasonal accuracy is also imperative. Based on average forecast errors (Fig. 9a), seasonally ELM models have better accuracies, with the hybrid CEEMDAN-ELM achieving accurate forecasts of SM_{UL} during the summer and autumn seasons. Nevertheless, the best hybrid EEMD-ELM attained very comparable accuracies in the winter and spring and has the paramount accuracy with the least average error in October ($\overline{FE} = 0.026$) (Fig. 10a). While, November registered the highest $\overline{FE} = 0.098$ value (the lowest accuracy). Alternatively, all models have better lower layer *SM* forecasting potential during summer (ELM being the best) and perform the worst during spring (Fig. 9b). During summer and autumn, however, both the hybrid CEEMDAN-ELM and the hybrid EEMD-ELM have similar accuracy. On monthly basis (Fig. 10b), our proposed hybrid EEMD-ELM model registered the best performance in January with the least $\overline{FE} = 0.002$.

The model preciseness was assessed using box-plots illustrating the spread of observed and forecasted data with respect to their quartiles. Being non-parametric, box-plots can offer a better understanding of the degree of spread and skewness where whiskers indicate the variability outside of 25th and 75th percentiles. In forecasting upper layer soil moisture (Fig. 11a), the best hybrid EEMD-ELM model established very similar distribution to the observations, however, the best hybrid

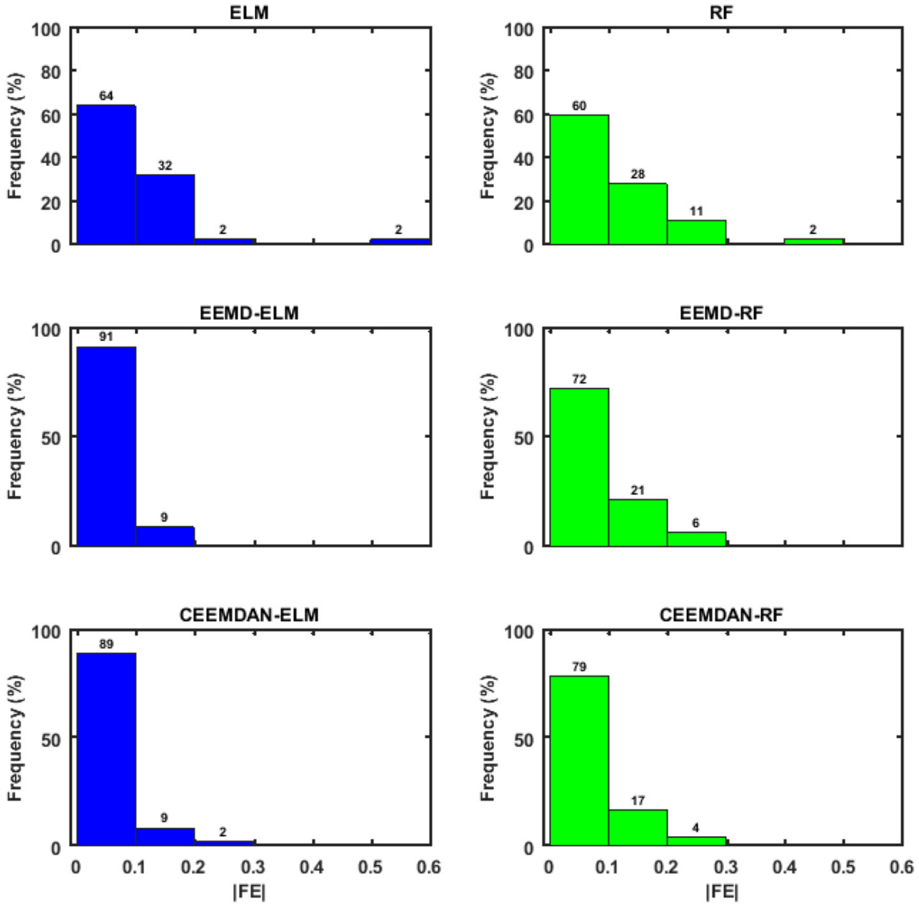


Fig. 6. Histogram illustrating the frequency (in percentages) of absolute forecasting errors (|FE|) of the best performing ELM and random forest (RF) models in forecasting upper layer soil moisture (SM_{UL}). [Best ELM: Site 5; Best EEMD-ELM: Site 6; Best CEEMDAN-ELM: Site 5 and the corresponding RF models].

CEEMDAN-ELM showed distorted spread with overestimated upper quartiles. Similarly, for lower layer *SM*, the best hybrid EEMD-ELM model (Fig. 11b) produced very analogous distribution to that of forecasted values, yet again the best hybrid CEEMDAN-ELM showed skewed distribution towards the upper-end with overestimated upper quartile. Thus, the hybrid EEMD-ELM had better predictive performance on the basis of box-plots and is reinforced by the previous assessment metrics (Table 8).

The measures of spread (Table 10) showed that the best hybrid EEMD-ELM models performed well in forecasting the upper layer soil moisture. They registered the closest forecasted upper quartile (UQ), lower quartile (LQ), maximum and minimum values, further consolidating their superiority. For lower layer *SM* forecasting, the measures of spread including the UQ, median, LQ and mean, of forecasts generated by the best hybrid EEMD-ELM were practically equal to the observed values. However, both maximum and minimum values were slightly over-predicted. Overall the hybrid EEMD-ELM performed with a better precision in terms of the forecast distribution of both upper and lower soil moistures.

To validate the suitability of the hybrid EEMD-ELM models for practical deployment, modeling time (in seconds) of EEMD-ELM models were determined at all sites and compared to the other comparative models as illustrated in Table 11. Modeling time is the time elapsed during training and validation phases. It is evident that the ELM models are faster than the random forest equivalents. In addition, the modeling time for ensemble hybrid models (EEMD-ELM, CEEMDAN-ELM, EEMD-RF, and CEEMDAN-RF) has increased in comparison to their standalone counterparts. Despite this, the execution speeds of the proposed hybrid EEMD-ELM models were far much less in comparison to all RF models. For upper layer *SM*, the EEMD-ELM registered minimum modeling time of 0.411 s (at Site 6) and a maximum of 0.694 s (at Site 7). The overall average modeling time was 0.559 s (559 ms), while for EEMD-RF it was 16.364 s. A comparison of average modeling times revealed that an increase of 0.395 s was noted in moving from standalone ELM to hybrid EEMD-ELM ensemble, while a huge 14.535 s increase was noted in traversing from standalone RF to EEMD-RF. For lower layer *SM*, a minimum modeling time of 0.429 s (at Site 7) and a maximum of 0.704 s (at Site 6) were registered by the EEMD-ELM while average modeling time was 0.549 s (549 ms). However, the equivalent EEMD-RF registered an average of 16.353 s. Once again the increase in average modeling times in moving from standalone ELM to EEMD-ELM ensemble was far less (0.402 s) in comparison to a large 14.612 s increase in moving from standalone RF to EEMD-RF. The outcome of modeling time shows computation efficiency of EEMD-ELM and further affirms its suitability for real-world soil moisture forecasting applications.

5. Further discussion and insights of data-intelligent models

In this study, the suitability of a hybrid model (EEMD-ELM) with a self-adaptive multi-resolution tool, ensemble empirical mode decomposition, coupled with the non-tuned ELM model in forecasting upper and lower layer *SM* was examined and the performance was compared with the equivalent random forest (RF) models. Overall, the ELM models, established on the basis of single layer feed-forward neural network, outperformed the RF models in all approaches under investigation (*i.e.*, standalone, EEMD ensemble and CEEMDAN ensemble) and showed its better capability for modeling and simulating the monthly upper and lower layer soil moisture data derived from the physically-based *WaterDyn* model (AWAP, 2016; Raupach et al., 2009). It is noteworthy that if meteorological forcing and soil parameters are not available in real-time, the data-intelligent models based on EEMD and CEEMDAN ensemble approach with historically simulated soil

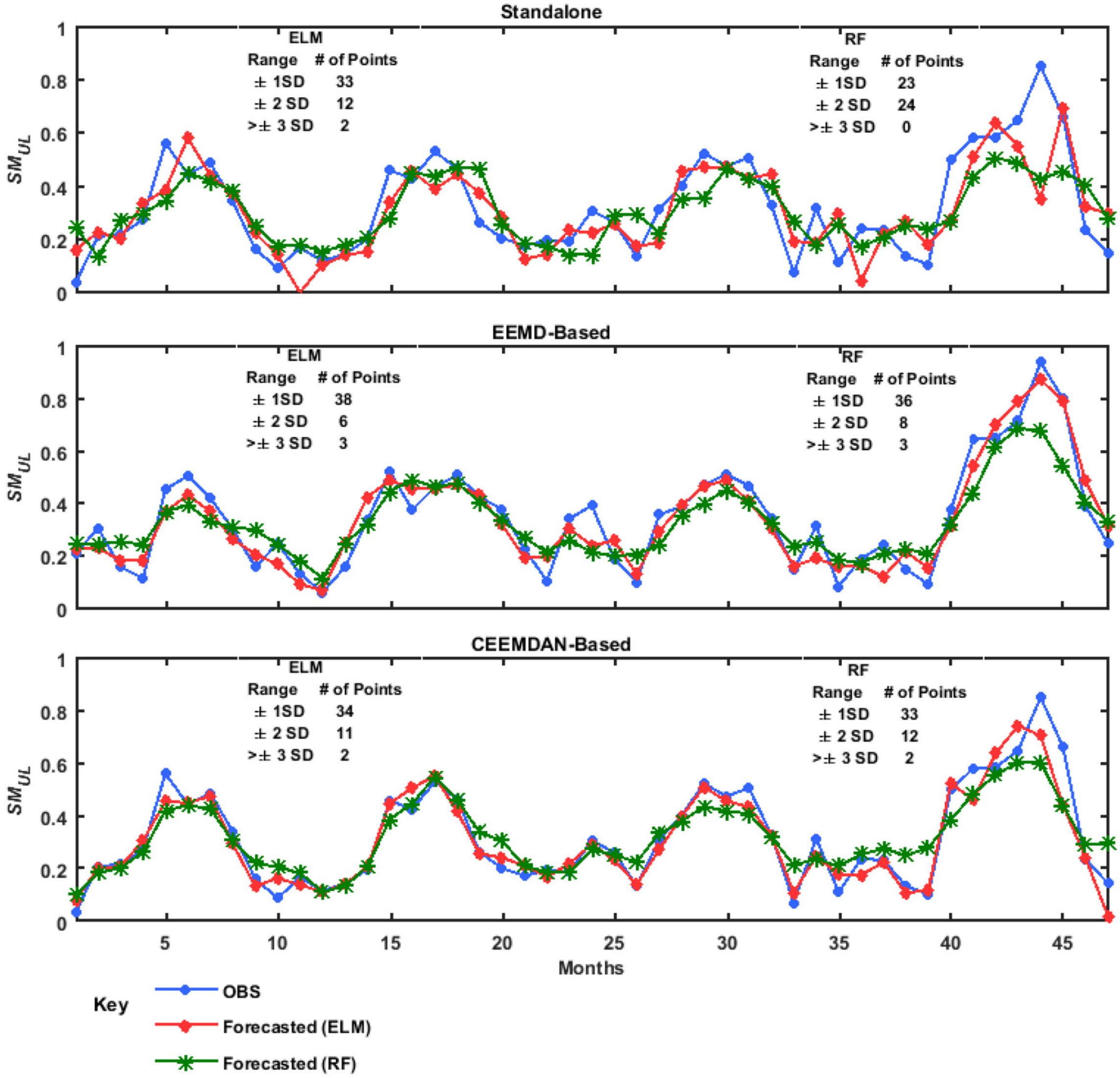


Fig. 7. Observed and forecasted upper layer soil moisture (SM_{UL}) during the test period, from the ELM, EEMD-ELM and CEEMDAN-ELM and its RF counterparts. [Best ELM: Site 5; Best EEMD-ELM: Site 6; Best CEEMDAN-ELM: Site 5 with their corresponding RF models].

moisture can amicably be incorporated into agricultural and environmental decision-making purposes.

Fundamentally, standalone data-intelligent models incur challenges in handling seasonality and non-stationarity of climate-based inputs and the challenges are exacerbated by complex pedologic and hydrological processes (*e.g.*, soil moisture). The multi-resolution analysis utility, EEMD was seen to enhance the data series by extracting the entrenched frequency-based information that otherwise would not be apparent from intact (*i.e.*, unresolved) data-series. Channeling this information to respective models led to the formation of the hybrid EEMD-ELM (EEMD-RF) models. The EEMD decomposition of the *SM* time series evidently facilitated the training algorithm in ELM more effectively to capture the deterministic components at various resolution levels, while effectively mapping the processes in forecasting the *SM* data. This is likely to result in swifter model convergence, negligible errors and improved precision. These desirable outcomes were apparent from the largest values in performance metrics (r, WI, E_{NS}, L) and the lower error values, *RMSE* and *MAE* from EEMD-ELM model (Tables 7 and 8). This finding accedes with the outcomes of other EEMD based studies (*e.g.*, (Bai et al., 2015; Basha et al., 2015; Beltran-Castro et al., 2013; Jiao et al., 2016; Ouyang et al., 2016; Ren et al., 2015; Seo and Kim, 2016). Hence, it is certain that the EEMD process has a good potential to provide a reliable physical basis to data-intelligent models by isolating the embedded deterministic components and physical processes in the time series data.

One obvious feature noted is that the hybrid EEMD-ELM was more accurate for the majority of the sites in forecasting the upper layer soil moisture, while for the lower layer soil moisture, the hybrid EEMD-ELM was unanimously the best option. The upper (surface) layer being at the boundary of Earth-Atmosphere system is in constant interaction with meteorological variations and deep soil hydraulics. The upper layer is also vulnerable to vegetation types, such as trees, crops, grass, or fallow (Ladson et al., 2004). On the contrary, the lower layer (or root zone) *SM* is less susceptible to exterior variations and is influenced mainly by deep percolation, groundwater recharge, and plant uptake. These differences in variations are apparent in the climatological pattern (Fig. 4a–b), with upper layer *SM* showing greater seasonal pattern and variability. Yet, the level of lower *SM* is rather stable. In addition, *SM* is also dependent on soil type (*e.g.*, thickness, texture, bulk density, pedality, and soil organic carbon content), which hugely influences the soil hydraulics and soil water retention capacity (Maraseni et al., 2008; Maraseni and Pandey, 2014). This vertical (within profile) and spatial (between profiles) soil type inconsistency (Ladson et al., 2004) may affect the accuracy of data-driven models. The best hybrid EEMD-ELM

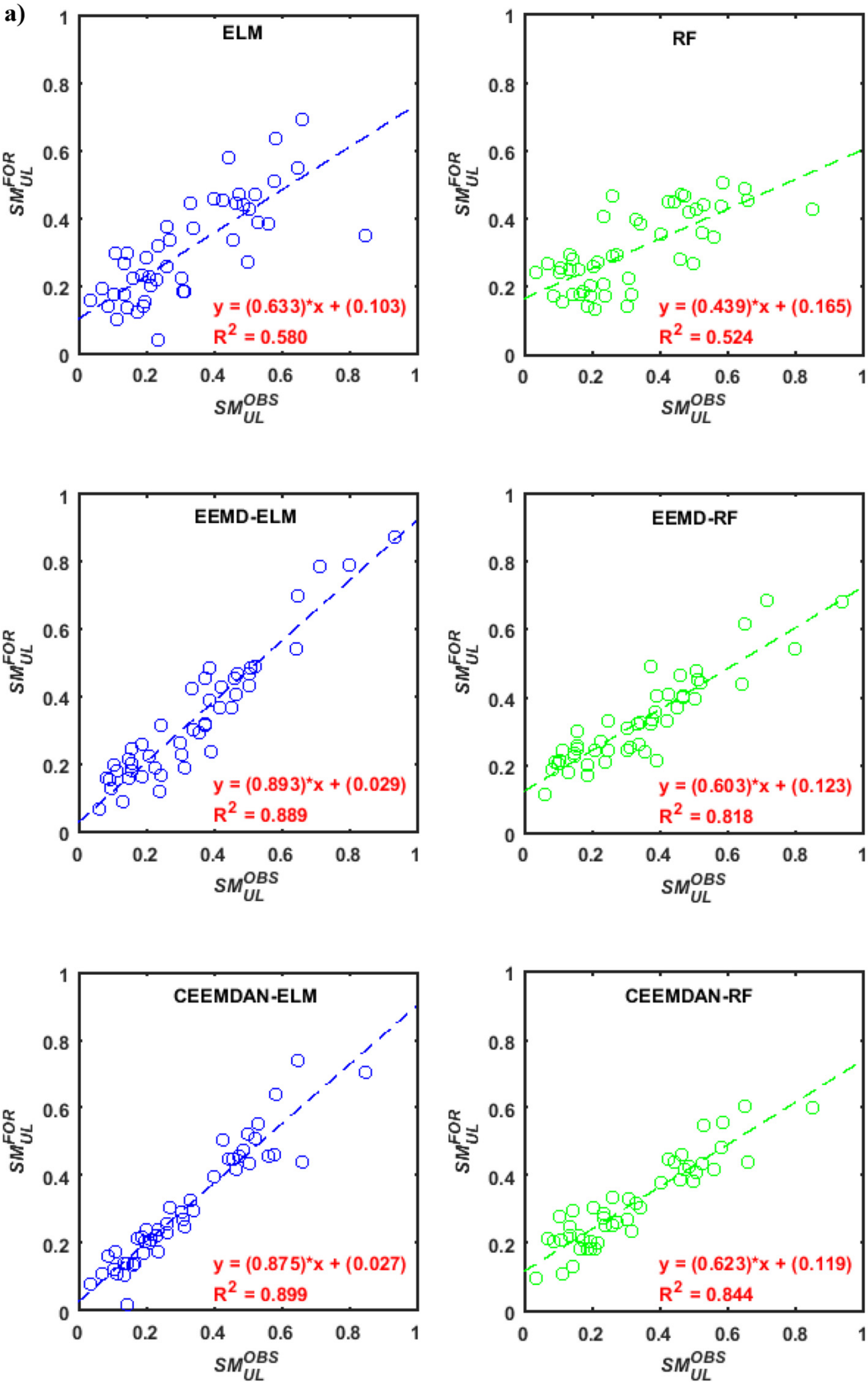


Fig. 8. Scatter plots of the best ELM and random forest (RF) models in forecasting a) upper layer soil moisture (SM_{UL}) [Best ELM: Site 5; Best EEMD-ELM: Site 6; Best CEEMDAN-ELM: Site 5 with their corresponding RF models] and b) lower layer soil moisture (SM_{LL}) [Best standalone-ELM: Site 1; Best EEMD-ELM: Site 1; Best CEEMDAN-ELM: Site 3 with their corresponding RF models].

in forecasting upper layer *SM* was at Site 6-Jerrawa which falls in a temperate climate zone with sodosol soil type and is covered with modified pastures for grazing. Chief soils at Jerrawa are hard neutral yellow and yellow mottled soils. In contrast, in SM_{LL} forecasting, Site 1-Menindee had the best hybrid EEMD-ELM model. Menindee is located in the desert region with calcarosol soil type having brown sands with clay substrata underneath where grazing is predominant and native vegetation covers this site. Thus, the application of hybrid EEMD is capable of overcoming the deleterious effects of distinctive topographical and climatological conditions.

ELM without being constrained to extract pertinent information using non-linearly connecting elements (*i.e.*, neurons) is a meritorious data-intelligent model for designing real-life agricultural and hydrological decision-support systems, as stipulated in studies in different areas (*e.g.*, (Deo et al., 2017b; Guo, 2016; Kaya and Uyar, 2013; Sun et al., 2008; Syed-Abdul et al., 2017; Yadav et al., 2017)). ELM has a better generalization capability and is able to handle large-scale data with computationally fast predictions (Mouatadid and Adamowski,

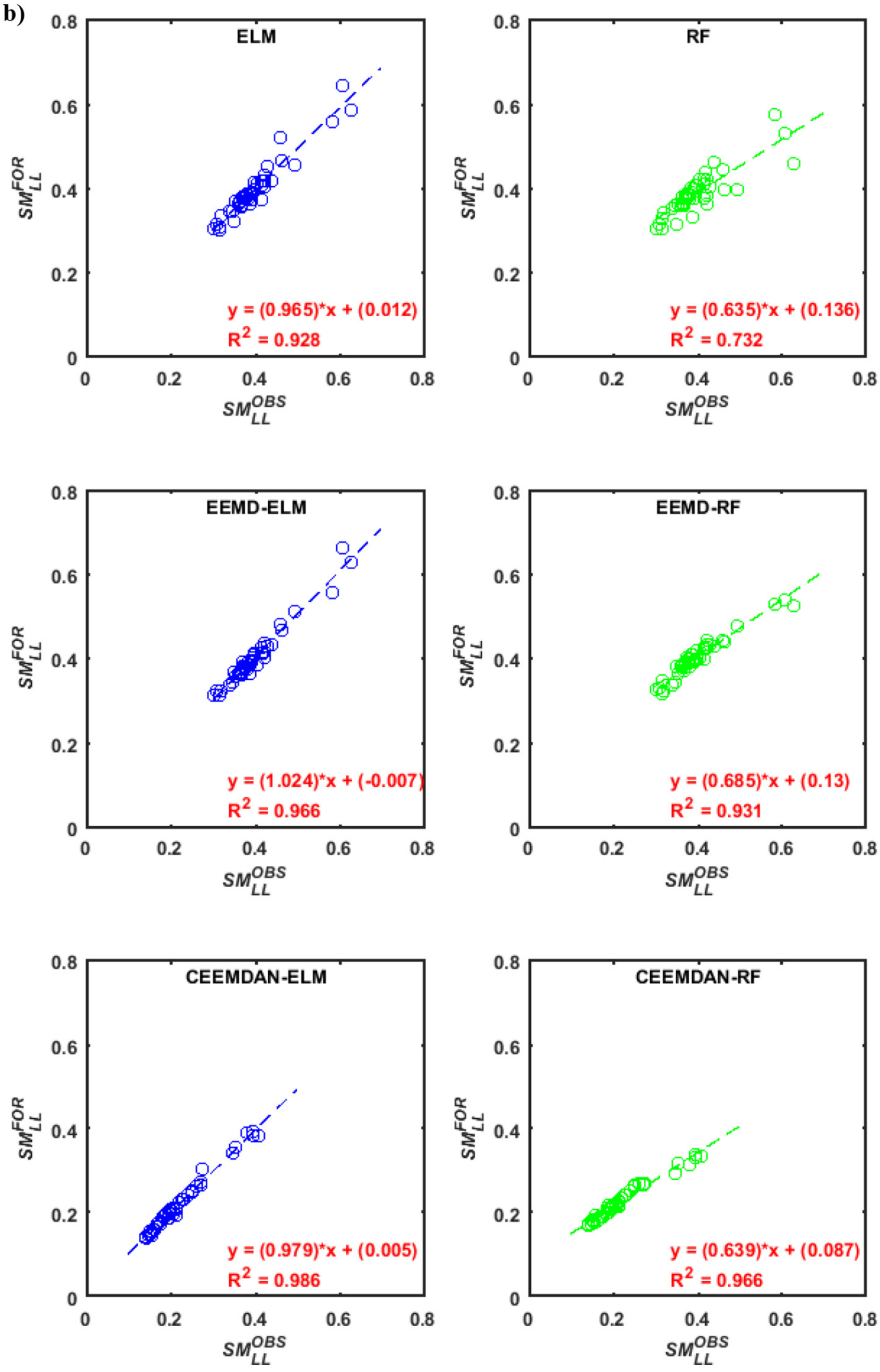


Fig. 8. (*continued*)

2016), giving it an added versatility for such decision-support systems. The promising results from the hybrid EEMD-ELM is a very good beginning for their prospective applications in hydrology, however, it only is worthwhile if the modeling and forecasting are carried out in real-time. The challenges in real-time deployments include computational time and memory since the efficiency and scalability becomes a relevant dimension (Bequé and Lessmann, 2017).

ELM overcomes this due to high efficiency, random input weights and hidden layer biases generation and analytical determination of output weights (Chen et al., 2012b; Deo and Sahin, 2016; Wan et al., 2014; Xu and Wang, 2016). The other challenges are the automated self-optimization capability which the ELM has no issues with, since fewer user-defined parameters are required and the network parameters are automatically generated (Şahin et al., 2014) avoiding issues like, learning rates, learning epochs, stopping criteria, and local optima (Chen et al., 2012b). This allows ELM to have better generalization capability with much faster learning rate (Huang et al., 2015) as clearly illustrated by the modeling times (Table 11). Even when the model is hybridized using the computationally expensive ensemble technique, *viz.* EEMD, the random forest becomes exceedingly slow as the modeling time increased by almost 14.5 s. However, the modeling time of ELM is not affected much as there is a slight increase (≈0.4 s) in the modeling times of EEMD-ELMs in comparison to standalone ELMs at all the sites. This ease of tuning, efficient running capability with reduced

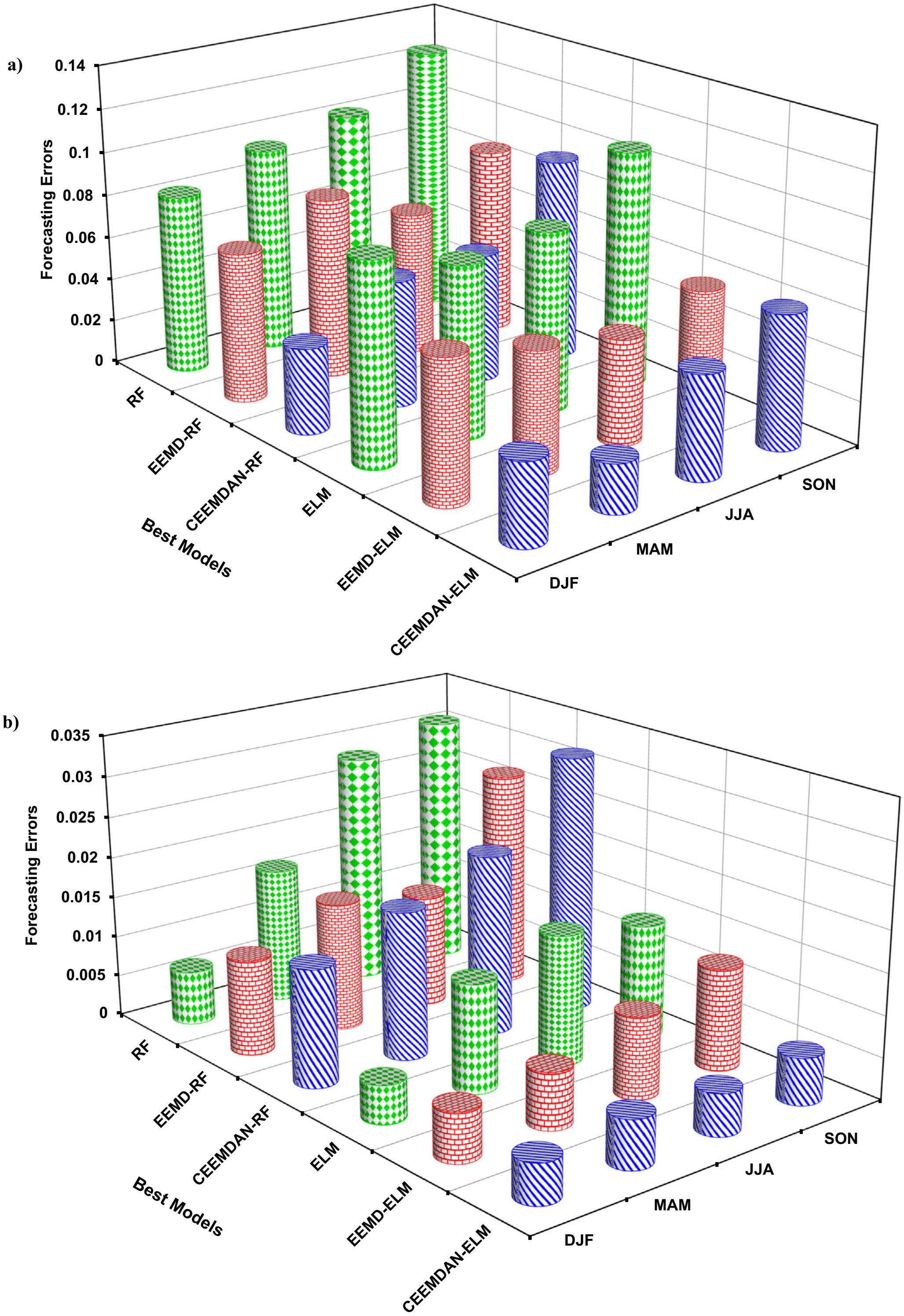


Fig. 9. Bar graphs of average seasonal forecasting errors (Summer-DJF; Autumn-MAM; Winter-JJA; Spring-SON) in forecasting: a) upper layer (SM_{UL}) and b) lower layer (SM_{LL}) soil moisture using the best: ELM, EEMD-ELM and CEEMDAN-ELM models and the corresponding RF models [NB: Best models for SM_{UL} forecasting were ELM: Site 5; EEMD-ELM: Site 6; CEEMDAN-ELM: Site 5 while for forecasting SM_{LL} the best models were as follows ELM: Site 1; EEMD-ELM: Site 1; CEEMDAN-ELM: Site 3].

modeling time, reduced computational complexity and very less human intervention makes ELM well-suited for big data analytics, online systems and efficient real-time applications (Deo et al., 2017b; Frances-Villora et al., 2016; Huang et al., 2015).

In addition, the multi-resolution utility, EEMD is self-adaptive requiring trivial human intervention making the novel hybrid EEMD-ELM model an encouraging prospect for real-time applications in decision-support systems. This technique can be transformed and embedded into hand-held low memory devices (*e.g.*, mobile phones and tablets) and within user-friendly mobile apps. For everyday real-time applications, the technology needs to become portable, smaller, cheaper and more reliable. Particularly with low memory devices, it is important to compute forecasts with low latency where the hybrid EEMD-ELM models certainly provide an edge. This is evident from the reduced

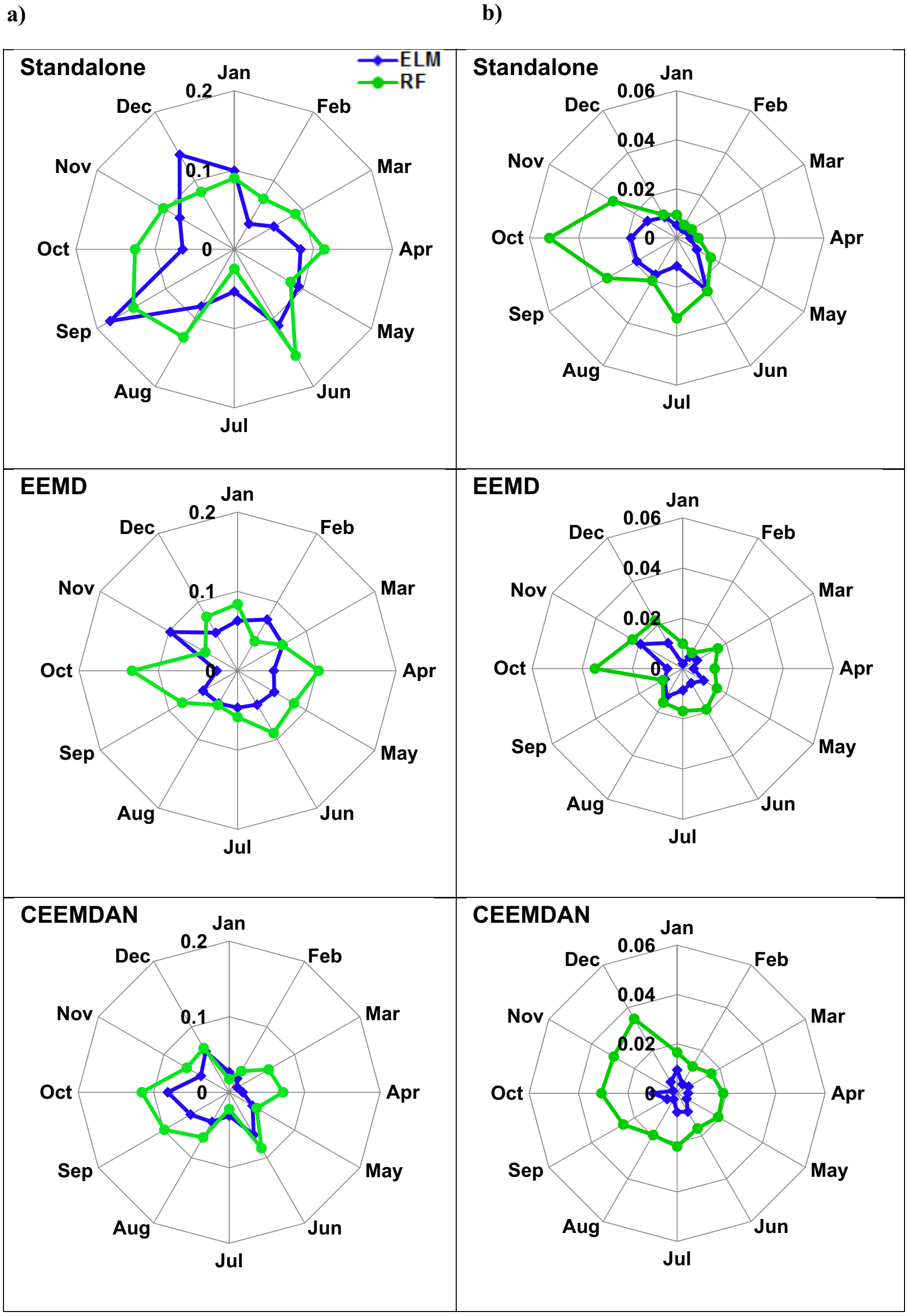


Fig. 10. Polar plots showing monthly average of forecasting errors in forecasting: a) upper layer (SM_{UL}) and b) lower layer (SM_{LL}) soil moisture using the best ELM, EEMD-ELM and CEEMDAN-ELM models and their RF counterparts. [The best models were as follow: for SM_{UL} ELM: Site 5; EEMD-ELM: Site 6; CEEMDAN-ELM: Site 5 while for SM_{LL} ELM: Site 1; EEMD-ELM: Site 1; CEEMDAN-ELM: Site 3].

modeling time (Table 11). The EEMD-ELM being super fast is more adaptable to a real-time forecasting system than the RF counterpart models. Alternatively, external cloud-based servers could be used to execute background sophisticated models, however, this may not be feasible in everyday farming situations. Real-time easy to use apps would be convenient to farmers, farm managers, and the government for precision agriculture, agricultural and hydrological decision support systems (Leeuwen et al., 2011) and flood/drought early warning systems, which are projected to increase in frequency and severity under future warmer climate scenario (IPCC, 2014).

Prior to any real-time applications, the model's testing with smaller time-steps (*e.g.*, weekly, daily, and hourly horizons) could provide more detailed understanding through finer predictions which are desirable decision time-scale for real-life applications (Deo et al., 2017b). Extreme events such as periods of high, moderate and low *SM* levels could also be explored in a follow-up study. In addition, other decomposition techniques like the singular value decomposition (Chitsaz et al., 2016; Wallace et al., 1992), wavelet transforms (Daubechies, 1990), maximum overlap discrete wavelet and Fourier transform (Percival and Walden, 2000) could be trialed for shorter time horizons in independent studies. The recently proposed two-phase decomposition (Wang et al., 2017a; Wang et al., 2017b) and empirical wavelet transform (EWT) (Kedadouche et al., 2016; Peng et al., 2017) could also be tested with hybrid EEMD and hybrid CEEMDAN-based models for forecasting soil moisture data.

The ability of ELM's potential in spatial *SM* forecasting needs to be

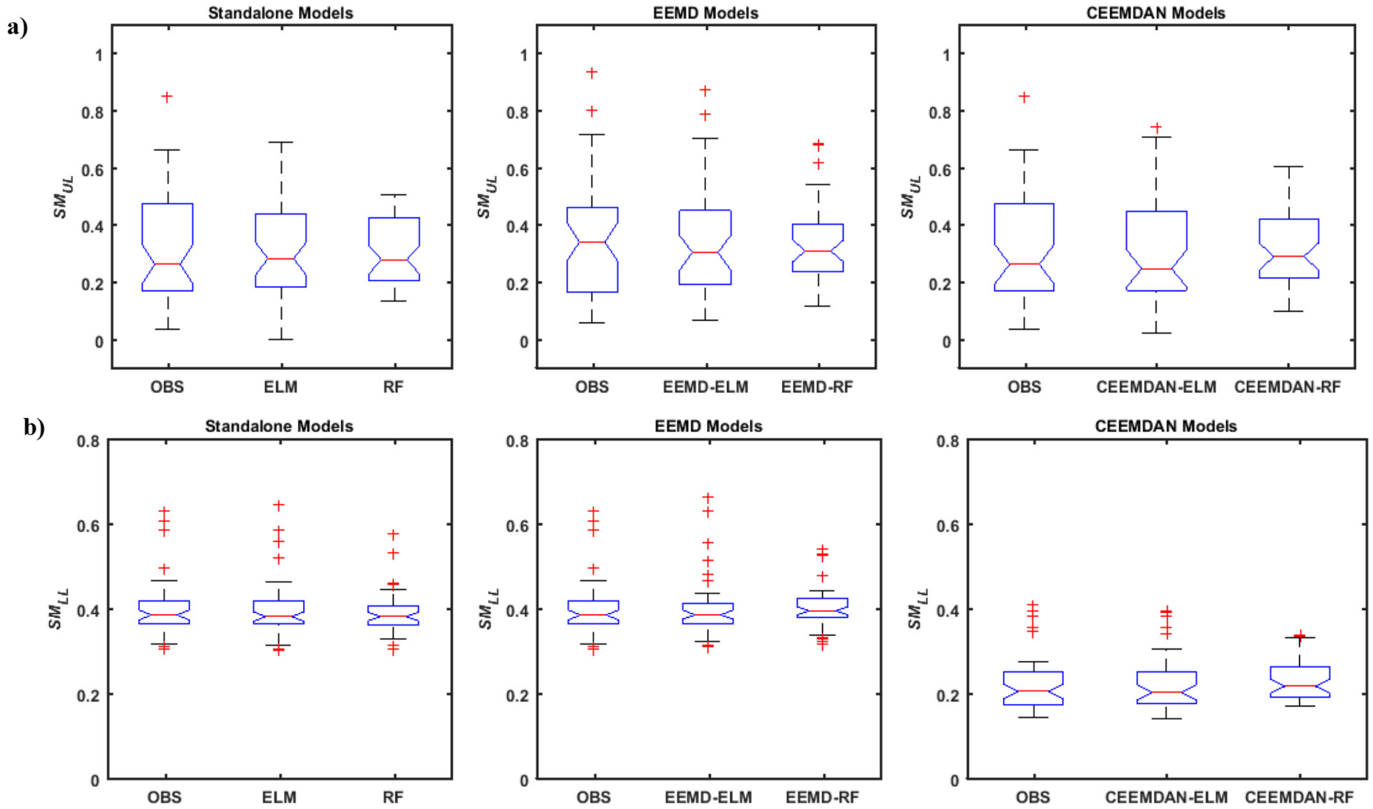


Fig. 11. Box plots of optimal ELM and RF models in forecasting a) upper layer soil moisture (SM_{LL}) [Best ELM: Site 5; EEMD-ELM: Site 6; CEEMDAN-ELM: Site 5] and b) lower layer soil moisture (SM_{LL}) [Best standalone-ELM: Site 1; EEMD-ELM: Site 1; CEEMDAN-ELM: Site 3]. (NB: *SM* are relative values and is dimensionless.

carried out in subsequent independent studies as a key limitation of ELM is that it has not been extensively tested on huge data sets and in extensive spatial forecasting applications. Deepa and Lakshmi (2016) also pointed out that the universal approximation capability of basic ELM and its performance in sparse high-dimensional applications are yet to be answered and suggested an introduction of sparse coding techniques to allow ELM to aptly handle high dimensional data. The progression of ELM has been a continuous process with studies being conducted in the implementation of ELM as deep learning networks or multiple-layer neural networks in solving classification problems (Ding et al., 2015; Kasun et al., 2013; Tang et al., 2016) and time-series applications are underway.

6. Conclusion

Based on the antecedent soil moisture from the physically-based *WaterDyn* (hydrological) modeled data from January 1990–December 2016, upper and lower layer soil moisture has been forecasted in this study, using hybrid data intelligent models tested at seven sites in the MDB region, NSW, Australia. Self-adaptive multi-resolution utilities based on the EEMD and CEEMDAN approaches helped resolve the intact (*i.e.*, undecomposed) time series into intrinsic mode functions (IMF) and a residual component. After determining the significant lagged inputs of corresponding IMFs and residual component *via* a partial-auto-correlation function, inputs were channeled into ELM (and RF) forming the hybrid models. Model performances were assessed using the objective (statistical measures) and subjective (graphical) methods and the findings are as follows:

(1) The ELM performed better than the random forest model in forecasting upper and lower layer *SM* at all study sites.
(2) Incorporation of the ensemble empirical mode decomposition (EEMD) based on the multi-resolution analysis utility led to an enhanced accuracy of the standalone models.
(3) The objective evaluation showed that the hybrid EEMD-ELM model had the best performance in forecasting the upper layer *SM* at four (out of the seven) sites *viz.*, Sites 1, 2, 3 and 7. This is evident in percentage increase in Legates-McCabe's Index (*L*) values ranging between 36.9% and 288.5%, Willmott's Index (*WI*) values ranging between 24% and 130.6%, and the Nash–Sutcliffe Efficiency (E_{NS}) values ranging between 46.4% and 263.5% in comparison to the standalone ELM model. On the other hand, the hybrid CEEMDAN-ELM model performed better at the other 3 study sites. Similarly, in forecasting the lower layer *SM* the hybrid EEMD-ELM model was unanimously the best objective model for all study sites with the highest percentage increase in the value of $L = 46.4\%$, $WI = 7.7\%$, and $E_{NS} = 21.9\%$, in comparison to the standalone ELM model.
(4) Site 6-Jerrawa registered the best hybrid EEMD-ELM with the sigmoid activation function and a neuronal architecture of 48-52-1 (Input-Hidden-Output) for forecasting the upper layer *SM*. For the lower layer *SM* forecasting, Site 1-Menindee had the best hybrid EEMD-ELM model with the triangular basis activation function and a neuronal architecture of 59-69-1. Remarkably, both of these are grazing sites, and therefore, have very important implications for the potential use of the newly designed hybrid models in real-time applications for portable devices (*e.g.*, soil moisture prediction *apps* on mobile phones) for farmers and other decision-makers.
(5) Subjective evaluations using the various diagnostic plots also affirmed the superiority of the hybrid EEMD-ELM model in forecasting both the upper and the lower layer soil moisture. Importantly, monthly evaluations showed that in the forecasting of upper layer *SM*, the best hybrid EEMD-ELM model yielded the highest accuracy for the month of October ($\overline{FE} = 0.026$) while for the lower layer *SM*, the month of January ($\overline{FE} = 0.002$) had the highest accuracy.
(6) Although a computationally expensive ensemble modeling

approach, *viz.* EEMD was adopted there is a slight increase of ≈0.4 on average in the modeling times of EEMD-ELMs in comparison to standalone ELMs at all the sites revealing the better efficiency of the EEMD-ELM. While with the random forest becomes exceedingly slow in the ensemble method.

Overall, the standalone ELM and RF models had similar computation efficiency and model performances. However, despite the computationally expensive ensemble techniques (*i.e.*, EEMD) being implemented, the hybrid ensembles EEMD-ELM was highly efficient with improved performances. Based on the newly designed data-intelligent model incorporating extreme learning machine with multi-resolution technique, it is ascertained that the hybrid EEMD-ELM model has the greatest ability to forecast both the upper and the lower layer soil moisture than its comparative counterparts. However, the forecasting accuracy is sensitive to meteorological factors (especially the upper layer *SM*) and the vertical and spatial soil texture inconsistencies. The capability of data-intelligent models has been explored in many areas, yet, this has been lagging behind in the very important agricultural applications. Further independent studies with the incorporation of meteorological variables and the testing of the hybrid model over smaller time-steps with various MRA utilities is still an open research problem. ELM has the potential for real-time *SM* monitoring and forecasting *via* its user-friendly mobile app applications that could be useful in agricultural decision systems, precision agriculture, flood and drought early warning and adaptive water resources planning that are major challenges exacerbated by the consequences of projected warmer climates in most parts of the world.

Acknowledgement

The authors are grateful to Australian Water Availability Project (AWAP) for providing the relevant soil moisture data for New South Wales region. This study was supported by University of Southern Queensland Office of Research and Graduate Studies (ORGS) Postgraduate Research Scholarship (USQ-PRS) Grant # 2016 awarded to the first author. We thank Editor-in-Chief Professor Christine Morgan and both reviewers whose comments have improved the clarity of the final paper.

References

ABS, 2011. Population estimates and Australia's new statistical geography. In: Australian Bureau of Statistics. 2011 Regional Population Growth, Australia (Australia).

Afanasyev, D.O., Fedorova, E.A., 2016. The long-term trends on the electricity markets: comparison of empirical mode and wavelet decompositions. Energy Econ. 56, 432–442.

Anctil, F., Tape, D.G., 2004. An exploration of artificial neural network rainfall-runoff forecasting combined with wavelet decomposition. J. Environ. Eng. Sci. 3 (S1), S121–S128.

ASRIS, 2014. Australian Soil Resource Information System. Department of Agriculture, Fisheries and Forestry.

Australian Bureau of Statistics, 2008. Water and the Murray-Darling Basin – A Statistical Profile, 2000–01 to 2005–06.

Year book Australia, 2009–10. In: Australian Bureau of Statistics (Ed.), Australian Bureau of Statistics. Commonwealth of Australia Canberra.

Australian Bureau of Statistics, 2014. Value of Agricultural Commodities Produced, Australia. pp. 2012–2013.

AWAP, 2016. Readme File: Australian Water Availability Project (AWAP), CSIRO. CSIRO.

Bai, Y., Wang, P., Xie, J., Li, J., Li, C., 2015. Additive model for monthly reservoir inflow forecast. J. Hydrol. Eng. 20 (7), 04014079.

Basha, G., Ouarda, T.B.M.J., Marpu, P.R., 2015. Long-term projections of temperature, precipitation and soil moisture using non-stationary oscillation processes over the UAE region. Int. J. Climatol. 35 (15), 4606–4618.

Beesley, C.A., Frost, A.J., Zajaczkowski, J., 2009. A comparison of the BAWAP and SILO spatially interpolated daily rainfall datasets. In: 18th World IMACS/MODSIM Congress, Cairns, Australia.

Beltran-Castro, J., Valencia-Aguirre, J., Orozco-Alzate, M., Castellanos-Domınguez, G., Travieso-Gonzalez, C.M., 2013. Rainfall forecasting based on ensemble empirical mode decomposition and neural networks. In: Rojas, I., Joya, G., Cabestany, J. (Eds.), Advances in Computational Intelligence: 12th International Work-Conference on Artificial Neural Networks. Springer, Puerto de la Cruz, Tenerife, Spain, pp. 471–480.

Bequé, A., Lessmann, S., 2017. Extreme learning machines for credit scoring: an empirical evaluation. Expert Syst. Appl. 86, 42–53.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24 (2), 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brocca, L., Melone, F., Moramarco, T., Morbidelli, R., 2010. Spatial-temporal variability of soil moisture and its estimation across scales. Water Resour. Res. 46 (2), 1–14.

Brocca, L., Ciabatta, L., Massari, C., Camici, S., Tarpanelli, A., 2017. Soil moisture for hydrological applications: open questions and new opportunities. Water 9 (2), 140.

Cecotti, H., 2016. Deep Random Vector Functional Link Network for handwritten character recognition. In: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 3628–3633.

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.

Chau, K., Wu, C., 2010. A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. J. Hydroinf. 12 (4), 458–473.

Chen, J., Li, M., Wang, W., 2012a. Statistical uncertainty estimation using random forests and its application to drought forecast. Math. Probl. Eng. 2012, 1–12.

Chen, X., Dong, Z.Y., Meng, K., Xu, Y., Wong, K.P., Ngan, H.W., 2012b. Electricity price forecasting with extreme learning machine and bootstrapping. IEEE Trans. Power Syst. 27 (4), 2055–2062.

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.

Chitsaz, N., Azarnivand, A., Araghinejad, S., 2016. Pre-processing of data-driven river flow forecasting models by singular value decomposition (SVD) technique. Hydrol. Sci. J. 61 (12), 2164–2178.

Colominas, M.A., Schlotthauer, G., Torres, M.E., 2014. Improved complete ensemble EMD: a suitable tool for biomedical signal processing. Biomedical Signal Processing and Control 14, 19–29.

Daubechies, I., 1990. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Inf. Theory 36 (5), 961–1005.

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.

Deepa, M., Lakshmi, M.R., 2016. Survey of deep and extreme learning machines for big data classification. Asian Journal of Research in Social Sciences and Humanities 6 (8).

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., 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), 90.

Deo, R.C., Wen, X., Feng, Q., 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., Tiwari, M.K., Adamowski, J.F., Quilty, J.M., 2017a. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. Stoch. Env. Res. Risk A. 31 (5), 1211–1240.

Deo, R.C., Downs, N., Parisi, A.V., Adamowski, J.F., Quilty, J.M., 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.

Department of Agriculture and Water Resources, 2015. Catchment Scale Land Use of Australia. Agricultural Land Management, Australia.

Di, C., Yang, X., Wang, X., 2014. A four-stage hybrid model for hydrological time series forecasting. PLoS ONE 9 (8), 1–18.

Diaz-Uriarte, R., Alvarez de Andres, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7, 3.

Ding, S., Zhang, N., Xu, X., Guo, L., Zhang, J., 2015. Deep extreme learning machine and its application in EEG classification. Math. Probl. Eng. 2015, 1–11.

Famiglietti, J.S., Rudnicki, J.W., Rodell, M., 1998. Variability in surface moisture content along a hillslope transect: Rattlesnake Hill, Texas. J. Hydrol. 210 (1–4), 259–281.

Frances-Villora, J.V., Rosado-Muñoz, A., Martínez-Villena, J.M., Bataller-Mompean, M., Guerrero, J.F., Wegrzyn, M., 2016. Hardware implementation of real-time Extreme Learning Machine in FPGA: analysis of precision, resource occupation and performance. Comput. Electr. Eng. 51, 139–156.

Gill, M.K., Asefa, T., Kemblowski, M.W., McKee, M., 2006. Soil moisture prediction using support vector machines. J. Am. Water Resour. Assoc. 42 (4), 1033–1046.

Guo, Z., 2016. An extreme learning machine-based intelligent decision-making model for multivariate sales forecasting. In: Intelligent Decision-making Models for Production and Retail Operations. Springer, pp. 295–316.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25 (15), 1965–1978.

Hora, J., Campos, P., 2015. A review of performance criteria to validate simulation models. Expert. Syst. 32 (5), 578–595.

Hu, W., Si, B.C., 2013. Soil water prediction based on its scale-specific control using multivariate empirical mode decomposition. Geoderma 193-194, 180–188.

Hu, T., Wu, F., Zhang, X., 2007. Rainfall–runoff modeling using principal component analysis and neural network. Hydrol. Res. 38 (3), 235–248.

Huang, G.-B., 2014. An insight into extreme learning machines: random neurons, random features and kernels. Cogn. Comput. 6 (3), 376–390.

Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.-C., Tung, C.C., Liu, H.H., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of Royal Society A 454, 903–995.

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2004. Extreme learning machine: a new learning scheme of feedforward neural networks. In: 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), pp. 985–990.

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2006. Extreme learning machine: theory and applications. Neurocomputing 70 (1–3), 489–501.

Huang, C., Li, L., Ren, S., Zhou, Z., 2011. Research of soil moisture content forecast model based on genetic algorithm BP neural network. In: Li, D., Liu, Y., Chen, Y. (Eds.), International Federation for Information Processing 2011. CCTA 2010, Part II, IFIP AICT. 345. pp. 309–316.

Huang, G., Huang, G.B., Song, S., You, K., 2015. Trends in extreme learning machines: a review. Neural Netw. 61, 32–48.

IPCC, 2014. In: Edenhofer, O., Pichs-Madruga, R., Sokona, Y., Farahani, E., Kadner, S., Seyboth, K., Adler, A., Baum, I., Brunner, S., Eickemeier, P., Kriemann, B., Savolainen, J., Schlömer, S., von Stechow, C., Zwickel, T., Minx, J.C. (Eds.), Climate Change 2014: Mitigation of Climate Change. Contribution of Working Group III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Chang. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Jain, A., Srinivasulu, S., 2004. Development of effective and efficient rainfall-runoff models using integration of deterministic, real-coded genetic algorithms and artificial neural network techniques. Water Resour. Res. 40 (4), 1–12.

Jiao, G., Guo, T., Ding, Y., 2016. A new hybrid forecasting approach applied to hydrological data: a case study on precipitation in Northwestern China. Water 8 (9), 367.

Kasun, L.L.C., Zhou, H., Huang, G.-B., Vong, C.M., 2013. Representational learning with extreme learning machine for big data. IEEE Intell. Syst. 28 (6), 31–34.

Kaya, Y., Uyar, M., 2013. A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. Appl. Soft Comput. 13 (8), 3429–3438.

Kedadouche, M., Thomas, M., Tahan, A., 2016. A comparative study between empirical wavelet transforms and empirical mode decomposition methods: application to bearing defect diagnosis. Mech. Syst. Signal Process. 81, 88–107.

Kim, T.-W., Valdes, J.B., 2003. Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. J. Hydrol. Eng. 8 (6), 319–328.

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

Ladson, T., Lander, J., Western, A., Grayson, R., 2004. Estimating extractable soil moisture content for Australian soils. In: Cooperative Research Centre for Catchment Hydrology, (Technical Report).

Lahouar, A., Ben Hadj Slama, J., 2017. Hour-ahead wind power forecast based on random forests. Renew. Energy 109, 529–541.

Leeuwen, W.v., Hutchinson, C., Drake, S., Doorn, B., Kaupp, V., Haithcoat, T., Likholetov, V., Sheffner, E., Tralli, D., 2011. Benchmarking enhancements to a decision support system for global crop production assessments. Expert Syst. Appl. 38 (7), 8054–8065.

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.

Legates, D.R., McCabe, G.J., 2013. A refined index of model performance: a rejoinder. Int. J. Climatol. 33 (4), 1053–1056.

Li, B., Chen, Z., Yuan, X., 2015. The nonlinear variation of drought and its relation to atmospheric circulation in Shandong Province, East China. PeerJ 3, e1289.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22.

Lin, Y., Jeon, Y., 2006. Random forests and adaptive nearest neighbors. J. Am. Stat. Assoc. 101 (474), 578–590.

Liu, Y., Mei, L., Ooe, S.K., 2014. Prediction of soil moisture based on extreme learning machine for an apple orchard. In: Conference on Computational Interdisciplinary Science. IEEE, pp. 400–404.

Mahmood, R., Hubbard, K.G., 2004. An analysis of simulated long-term soil moisture data for three land uses under contrasting hydroclimatic conditions in the Northern Great Plains. J. Hydrometeorol. 5, 160–179.

Mallat, S.G., 1989. A theory for multiresolution signal decomposition: the wavelet representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 11 (7), 674–693.

Mallat, S.G., 1998. A Wavelet Tour of Signal Processing. Academic, New York.

Maraseni, T.N., Pandey, S.S., 2014. Can vegetation types work as an indicator of soil organic carbon? An insight from native vegetations in Nepal. Ecol. Indic. 46, 315–322.

Maraseni, T.N., Mathers, N.J., Harms, B., Cockfield, G., Apan, A., Maroulis, J., 2008. Comparing and predicting soil carbon quantities under different land use systems on the Red Ferrosol soils of Southeast Queensland. J. Soil Water Conserv. 63 (4), 250–257.

Matei, O., Rusu, T., Petrovan, A., Mihuţ, G., 2017. A data mining system for real time soil moisture prediction. Procedia Engineering 181, 837–844.

Mehr, A.D., Kahya, E., Olyaie, E., 2013. Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. J. Hydrol. 505, 240–249.

Mohammadi, K., Shamshirband, S., Motamedi, S., Petković, D., Hashim, R., Gocic, M., 2015. Extreme learning machine based prediction of daily dew point temperature. Comput. Electron. Agric. 117, 214–225.

Mouatadid, S., Adamowski, J., 2016. Using extreme learning machines for short-term urban water demand forecasting. Urban Water J. 1–9.

Naing, W.Y.N., Htike, Z.Z., 2015. Forecasting of monthly temperature variations using random forests. ARPN Journal of Engineering and Applied Sciences 10 (21), 10109–10112.

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.

Nourani, V., Komasi, M., Mano, A., 2009. A multivariate ANN-wavelet approach for rainfall–runoff modeling. Water Resour. Manag. 23 (14), 2877–2894.

Nourani, V., Baghanam, A.H., Adamowski, J., Kisi, O., 2014. Applications of hybrid wavelet–artificial intelligence models in hydrology: a review. J. Hydrol. 514, 358–377.

Ouyang, Q., Lu, W., Xin, X., Zhang, Y., Cheng, W., Yu, T., 2016. Monthly rainfall forecasting using EEMD-SVR based on phase-space reconstruction. Water Resour. Manag. 30 (7), 2311–2325.

Palaninathan, A.C., Qiu, X., Suganthan, P.N., 2016. Heterogeneous ensemble for power load demand forecasting. In: 2016 IEEE Region 10 Conference (TENCON), pp. 2040–2045.

Pao, Y.-H., Park, G.-H., Sobajic, D.J., 1994. Learning and generalization characteristics of the random vector functional-link net. Neurocomputing 6, 163–180.

Patil, A.P., Deka, P.C., 2016. An extreme learning machine approach for modeling evapotranspiration using extrinsic inputs. Comput. Electron. Agric. 121, 385–392.

Peng, T., Zhou, J., Zhang, C., Fu, W., 2017. Streamflow forecasting using empirical wavelet transform and artificial neural networks. Water 9 (6), 406.

Percival, D.B., Walden, A.T., 2000. Wavelet Methods for Time Series Analysis. Cambridge University Press, UK.

Petropoulos, G.P., 2014. Remote Sensing of Energy Fluxes and Soil Moisture Content. CRC Press, Taylor & Francis Group, Boca Raton, FL.

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.

Raupach, M.R., Briggs, P.R., Haverd, V., King, E.A., Paget, M., Trudinger, C.M., 2009. Australian water availability project (AWAP)-CSIRO marine and atmospheric research component-final report for phase 3. In: CAWCR Technical Report No. 013.

Ren, Y., Suganthan, P.N., Srikanth, N., 2015. A comparative study of empirical mode decomposition-based short-term wind speed forecasting methods. IEEE Transactions on Sustainable Energy 6 (1), 236–244.

Rey, A., Oyonarte, C., Morán-López, T., Raimundo, J., Pegoraro, E., 2017. Changes in soil moisture predict soil carbon losses upon rewetting in a perennial semiarid steppe in SE Spain. Geoderma 287, 135–146.

Şahin, M., Kaya, Y., Uyar, M., Yıldırım, S., 2014. Application of extreme learning machine for estimating solar radiation from satellite data. Int. J. Energy Res. 38, 205–212.

Scardapane, S., Panella, M., Comminiello, D., Uncini, A., 2015. Learning from distributed data sources using random vector functional-link networks. Procedia Computer Science 53, 468–477.

Schmidt, W.F., Kraaijveld, M.A., Duin, R.P.W., 1992. Feedforward neural networks with random weights. In: Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol.II. Conference B: Pattern Recognition Methodology and Systems, pp. 1–4.

Seneviratne, S.I., Corti, T., Davin, E.L., Hirschi, M., Jaeger, E.B., Lehner, I., Orlowsky, B., Teuling, A.J., 2010. Investigating soil moisture–climate interactions in a changing climate: a review. Earth Sci. Rev. 99 (3–4), 125–161.

Seo, Y., Kim, S., 2016. Hydrological forecasting using hybrid data-driven approach. Am. J. Appl. Sci. 13 (8), 891–899.

Shamseldin, A.Y., 1997. Application of a neural network technique to rainfall runoff. J. Hydrol. 199, 272–294.

Shamshirband, S., Mohammadi, K., Tong, C.W., Petković, D., Porcu, E., Mostafaeipour, A., Ch, S., Sedaghat, A., 2015. Application of extreme learning machine for estimation of wind speed distribution. Clim. Dyn. 46 (5–6), 1893–1907.

Sun, Z.-L., Choi, T.-M., Au, K.-F., Yu, Y., 2008. Sales forecasting using extreme learning machine with applications in fashion retailing. Decis. Support. Syst. 46 (1), 411–419.

Syed-Abdul, S., Iqbal, U., Jack, L.Y., 2017. The novel use of an Extreme learning machines for clinical decision support systems. Comput. Methods Prog. Biomed. 147, A1.

Tang, J., Deng, C., Huang, G.B., 2016. Extreme learning machine for multilayer perceptron. IEEE Trans Neural Netw Learn Syst 27 (4), 809–821.

Tiwari, M., Adamowski, J., Adamowski, K., 2016. Water demand forecasting using extreme learning machines. Journal of Water and Land Development 28 (1).

Torres, M.E., Colominas, M.A., Schlotthauer, G., Flandrin, P., 2011. A complete ensemble empirical mode decomposition with adaptive noise. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 4144–4147.

Tozer, C.R., Kiem, A.S., Verdon-Kidd, D.C., 2012. On the uncertainties associated with using gridded rainfall data as a proxy for observed. Hydrol. Earth Syst. Sci. 16 (5), 1481–1499.

Wallace, J.M., Smith, C., Bretherton, C.S., 1992. Singular value decomposition of wintertime sea surface temperature and 500mb height anomalies. J. Clim. 5, 561–576.

Wan, C., Xu, Z., Pinson, P., Dong, Z.Y., Wong, K.P., 2014. Probabilistic forecasting of wind power generation using extreme learning machine. IEEE Trans. Power Syst. 29 (3), 1033–1044.

Wang, L.P., Wan, C.R., 2008. Comments on "The extreme learning machine". IEEE Trans. Neural Netw. 19 (8), 1494–1495 (author reply 1495–1496).

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, D., Luo, H., Grunder, O., Lin, Y., Guo, H., 2017a. Multi-step ahead electricity price forecasting using a hybrid model based on two-layer decomposition technique and BP neural network optimized by firefly algorithm. Appl. Energy 190, 390–407.

Wang, D., Wei, S., Luo, H., Yue, C., Grunder, O., 2017b. A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Sci. Total Environ. 580, 719–733.

Wang, X., Li, Y., Chen, T., Yan, Q., Ma, L., 2017c. Quantitative thickness prediction of tectonically deformed coal using Extreme Learning Machine and Principal Component Analysis: a case study. Comput. Geosci. 101, 38–47.

Weimann, A., Von Schonermark, M., Schumann, A., Jorn, P., Gunther, R., 1998. Soil moisture estimation with ERS-1 SAR data in the East-German loess soil area. Int. J. Remote Sens. 19 (2), 237–243.

Welsh, W.D., Vaze, J., Dutta, D., Rassam, D., Rahman, J.M., Jolly, I.D., Wallbrink, P., Podger, G.M., Bethune, M., Hardy, M.J., Teng, J., Lerat, J., 2013. An integrated modeling framework for regulated river systems. Environ. Model. Softw. 39, 81–102.

Wen, X., Feng, Q., Deo, R.C., Wu, M., Si, J., 2016. Wavelet analysis–artificial neural network conjunction models for multi-scale monthly groundwater level predicting in an arid inland river basin, northwestern China. Hydrol. Res. 48 (6), 1710–1729.

Willems, P., 2009. A time series tool to support the multi-criteria performance evaluation of rainfall-runoff models. Environ. Model. Softw. 24 (3), 311–321.

Willmott, C.J., 1981. On the validation of models. Phys. Geogr. 2, 184–194.

Willmott, C.J., 1984. On the evaluation of model performance in physical geography. In: Gaile, G.L., Willmott, C.J. (Eds.), Spatial Statistics and Models. Springer, pp. 443–460.

Wu, Z., Huang, N.E., 2009. Ensemble empirical mode decomposition: a noise-assisted data analysis method. Adv. Adapt. Data Anal. 1 (1), 1–41.

Xu, S., Wang, J., 2016. A fast incremental extreme learning machine algorithm for data streams classification. Expert Syst. Appl. 65, 332–344.

Yadav, B., Ch, S., Mathur, S., Adamowski, J., 2017. Assessing the suitability of extreme learning machines (ELM) for groundwater level prediction. Journal of Water and Land Development 32 (1), 103–112.

Yaseen, Z.M., Jaafar, O., Deo, R.C., Kisi, O., Adamowski, J., Quilty, J., El-Shafie, A., 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., Deo, R.C., Hilal, A., Abd, A.M., Bueno, L.C., Salcedo-Sanz, S., Nehdi, M.L., 2018. Predicting compressive strength of lightweight foamed concrete using extreme learning machine model. Adv. Eng. Softw. 115, 112–125.

Zhang, S., Shao, M., Li, D., 2017a. Prediction of soil moisture scarcity using sequential Gaussian simulation in an arid region of China. Geoderma 295, 119–128.

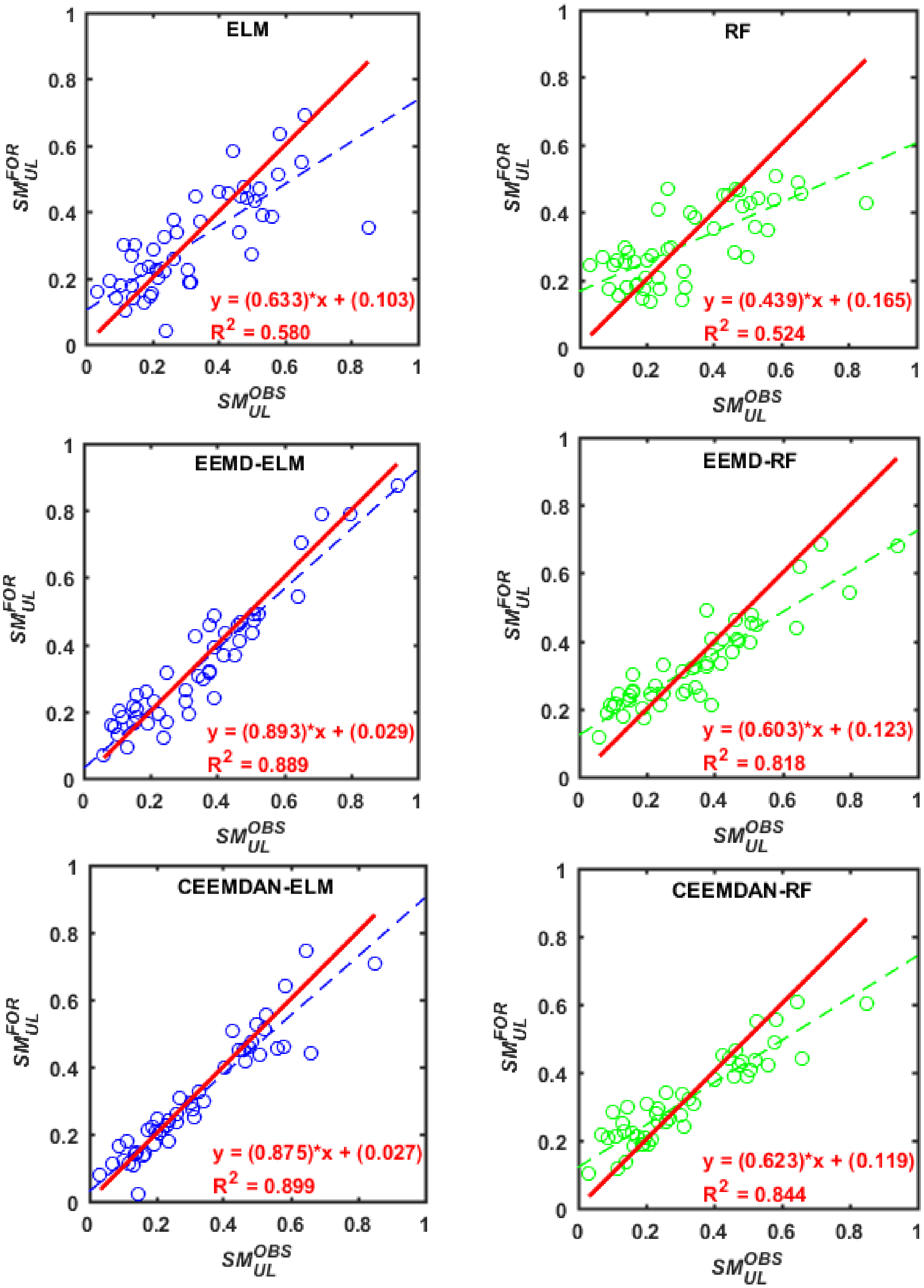
Zhang, W., Qu, Z., Zhang, K., Mao, W., Ma, Y., Fan, X., 2017b. A combined model based on CEEMDAN and modified flower pollination algorithm for wind speed forecasting. Energy Convers. Manag. 136, 439–451.

Supplementary analysis and discussions

The scatterplot with an X = Y line and the percentage deviations of the forecasted values from this line were computed to have a better understanding of the model performances in forecasting upper layer (*SM_{UL}*) and lower layer soil moisture (*SM_{LL}*) values. The scatterplots clearly showed that the ELM models performed better in comparison to the comparative random forest (RF) models in both the *SM_{UL}* (Figure S2a) and *SM_{LL}* (Figure S2b) forecasts. A closer examination showed that it was lucid from the scatterplots (Figure S2 (a-b)) that the hybridized EEMD-ELM and CEEMDAN-ELM models outperformed the other standalone models in forecasting both the *SM_{UL}* and *SM_{LL}*, as the 1:1 lines were very close to the regression lines.

A detailed analysis via the percentage deviations from the 1:1 line from all models was conducted. The appendix Table A2 details the full data on percentage deviations. Summarizing the data on the number of points that deviated from the 5% tolerance limit (Table S2 a-b) showed that in forecasting *SM_{UL}*, the best CEEMDAN-ELM model registered a total of 28/47 points that were over/underpredicted. This was far less in comparison to other competing models in forecasting upper layer soil moisture values. Interestingly, in forecasting *SM_{LL}*, both the hybridized models, EEMD-ELM and CEEMDAN-ELM models registered the least number of over/under predicted values which apparently was 5/47. This result further supports the outcomes of the study presented in the main chapter that was published in the journal *Geoderma* (Vol. 330, Pages 136-161).

a)


ELM
y = (0.633)*x + (0.103)
R^2 = 0.580
RF
y = (0.439)*x + (0.165)
R^2 = 0.524
EEMD-ELM
y = (0.893)*x + (0.029)
R^2 = 0.889
EEMD-RF
y = (0.603)*x + (0.123)
R^2 = 0.818
CEEMDAN-ELM
y = (0.875)*x + (0.027)
R^2 = 0.899
CEEMDAN-RF
y = (0.623)*x + (0.119)
R^2 = 0.844
SM_{UL}^{FOR}
SM_{UL}^{OBS}
0
0.2
0.4
0.6
0.8
1

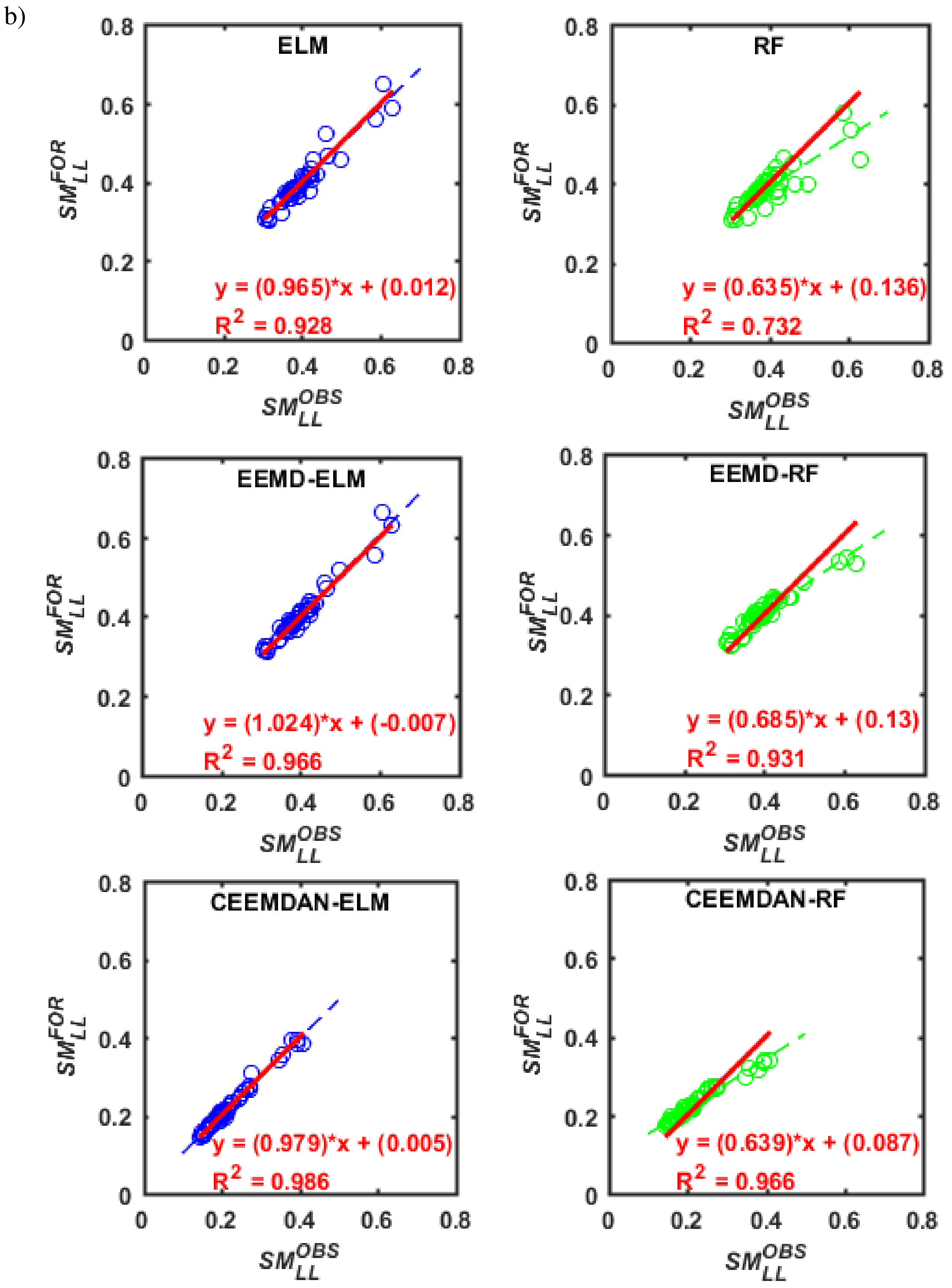


Figure S2 Scatter plots of the best ELM and random forest (RF) models in forecasting: a) upper layer soil moisture (SM_{UL}) [Best ELM: Site 5; Best EEMD-ELM: Site 6; Best CEEMDAN-ELM: Site 5 with their corresponding RF models] and b) lower layer soil moisture (SM_{LL}) [Best standalone-ELM: Site 1; Best EEMD-ELM: Site 1; Best CEEMDAN-ELM: Site 3 with their corresponding RF models].

(Note: The dashed line in blue and green is the least-squares fitting line to the respective scatter plots and the solid red line is 45°, X = Y line for comparison).

Table S2 Number of points that were under and overpredicted by the hybridized EEMD-ELM, CEEMDAN-ELM, EEMD-RF, CEEMDAN-RF, and the standalone ELM and RF models with respect to 5% tolerance limit at the sites that recorded the best performance in forecasting a) SM_{UL} and b) SM_{LL}.

a) SM_{UL}	**ELM Site-30**	**EEMD-ELM Site 43**	**CEEMDAN-ELM Site 30**	**RF Site-30**	**EEMD-RF Site 43**	**CEEMDAN-RF Site-30**
Under prediction	21	22	16	21	21	18
Over prediction	21	19	12	22	17	18
Total	42	41	28	43	38	36

b) SM_{LL}	**ELM Site-30**	**EEMD-ELM Site 43**	**CEEMDAN-ELM Site 30**	**RF Site-30**	**EEMD-RF Site 43**	**CEEMDAN-RF Site-30**
Under prediction	5	2	3	10	4	6
Over prediction	3	3	2	2	13	30
Total	8	5	5	12	17	36

Chapter 5: Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors

Wisdom of crowds

> "*When a group of diverse and independent individuals make a prediction or an estimate about a quantity, then often the mean of these estimates is better than the individual predictions or estimates.*" (Surowiecki, 2004)

Foreword

This chapter is an exact copy of the published article in the *Soil & Tillage Research* journal (Vol. 181, Pages 63-81).

Since soil moisture level is contingent upon many interrelated hydro-meteorological variables, a total of sixty input variables are utilized to forecast upper and lower layer soil moisture in this particular study. The sixty inputs are screened using a two-phase feature selection method. Firstly, the Neighbourhood Component Analysis (NCA) based feature selection algorithm for regression purposes (*fsrnca*) is applied to determine the relative feature weights. Following that, a basic ELM is utilized to determine the optimal set from the *fsrnca* determined feature weights. Inspired by the ideology of '*wisdom of crowds*', the committee of models approach is developed in this chapter. Four different standalone models including second-order Volterra, M5 model tree, random forest, and an ELM are utilized as initial feature extracting expert models from the screened predictor inputs. Following that, a novel ensemble committee of model is developed with the artificial neural network as the basis (ANN-CoM).

The ANN-CoM is evaluated against the standalone models (the second-order Volterra, M5 model tree, random forest, and ELM models) in forecasting monthly upper and lower layer soil moisture at four candidate sites.

Soil & Tillage Research 181 (2018) 63–81

Contents lists available at ScienceDirect

Soil & Tillage Research

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

Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors

Ramendra Prasad*, Ravinesh C. Deo*, Yan Li, Tek Maraseni

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

ARTICLE INFO

Keywords:
Committee of models
Extreme learning machine
Random forest
Volterra
M5 tree
Soil moisture forecasting
Murray-Darling Basin

ABSTRACT

Soil moisture (*SM*) is a key component of the global energy cycle that regulates all domains of the natural environmental and the agricultural system. In this research, the challenge is to develop a low-cost data-intelligent *SM* forecasting model using climate dynamics (i.e., the climate indices, atmospheric and hydro-meteorological parameters) as the model inputs. A newly designed, multi-model ensemble committee machine learning approach based on the artificial neural network (ANN-CoM) is developed to forecast monthly upper layer (~0.2 m from the surface) and the lower layer (~0.2–1.5 m deep) *SM* at four agricultural sites in Australia's Murray-Darling Basin. ANN-CoM model is validated with respect to non-tuned second-order Volterra, M5 model tree, random forest, and an extreme learning machine (ELM) models. To construct the ANN-CoM model, the input variables comprised of the hydro-meteorological data from the Australian Water Availability Project, large-scale climate indices and atmospheric parameters derived from the Interim ERA European Centre for Medium-Range Weather Forecasting ECMWF reanalysis fields leads to a total of 60 potential predictors used for *SM* forecasting. To reduce the model input data dimensionality for accurate forecasts, the Neighborhood Component Analysis (NCA) based feature selection algorithm for regression purposes (*fsrnca*) is applied to determine the relative feature weights related to the targeted variable. The optimal predictor variables are then screened with an ELM model as the fitness function of the *fsrnca* algorithm to identify the set of most pertinent model variables. Extensive performance evaluation using statistical score metrics with visual and diagnostic plots show that the ensemble committee based, ANN-CoM model is able to effectively capture the nonlinear dynamics involved in the modeling of monthly upper and lower layer *SM* levels. Therefore, the ANN-CoM multi-model ensemble-based approach can be considered to be a superior *SM* forecasting tool, portraying as an amicable, integrated (or ensemble) machine learning stratagem that can be explored for soil moisture modeling and applications in agriculture and other hydro-meteorological phenomena.

1. Introduction

Being a vital component of the response loop within a climatic system and gas exchange mechanism, the soil moisture (*SM*) plays an important role in hydrological and agricultural processes (Tian et al., 2017). *SM* controls the partitioning of energy into sensible and latent heat fluxes, and precipitation into evapotranspiration and runoff (Brocca et al., 2017; Munro et al., 1998; Petropoulos, 2014). *SM* is not only important for agricultural production but is also imperative for biomass production, biophysical and ecological processes, runoff potential, soil erosion/slope failure, flood control mechanisms, reservoir management and water quality assessments. The *SM* levels are greatly influenced by vegetation cover, soil characteristics, climate dynamics and land use. In addition, the rising global temperature trend shows significant reductions in projected *SM* level within the Australian Murray-Darling Basin region (Cai et al., 2009; Timbal et al., 2015) which is expected to severely affect the hydrological cycle, agriculture, and human lives. Thus, forecasted *SM* is critical for an assessment and development of sustainable agricultural and hydrological management practices (Tian et al., 2017).

Advancements in measurement and estimation techniques have led to a variety of ways to quantify *SM*, some of which include in-situ measurements, remote sensing, and formulation of physical models. However, the spatial in-situ measurements and now-casts of *SM* are expensive in terms of the installation, calibration, and maintenance issues of apparatus, time-consuming and labor-intensive. This has resulted in limited spatial and temporal monitoring of *SM* using ground-based point measurement techniques (Grayson and Western, 1998; Walker et al., 2003). To increase spatial coverage, remote sensing of *SM* via satellites has been developed. Yet, satellites are only able to

* Corresponding authors.
E-mail addresses: ramendrap23@gmail.com (R. Prasad), ravinesh.deo@usq.edu.au (R.C. Deo).

https://doi.org/10.1016/j.still.2018.03.021
Received 1 February 2018; Received in revised form 30 March 2018; Accepted 31 March 2018

estimate the *SM* in the top few centimeters of soil (1–5 cm) in areas away from large water bodies (e.g., ocean or lake) with low vegetation (Dharssi and Steinle, 2011; Du et al., 2000; Walker et al., 2003). The vegetation water content, dew, radiative fog and soil roughness add uncertainties in the satellite-derived observations (Dharssi and Steinle, 2011). As a result, significant vertical gradients in the *SM* can be overlooked.

Accordingly, the *WaterDyn* physical model has been developed to simulate *SM* and several other hydrological parameters across the Australian continent at a grid resolution of $0.05^{\circ} \times 0.05^{\circ}$ (Raupach et al., 2009, 2012). Developed under the Australian Water Availability Project (AWAP), the *WaterDyn* physical model incorporates meteorological forcing, *i.e.*, solar radiation, precipitation, minimum and maximum temperatures coupled with continental parameter maps, e.g., albedo, soil characteristics, seasonality of vegetation greenness to compute *SM* for the upper layer (up to a depth of 0.2 m from the surface) and the lower layer (0.2–1.5 m depth). The meteorological fields for this model are generated by the Australian Bureau of Meteorology (BOM) from its network of rain gauge and weather stations while solar irradiance data is obtained using geostationary satellites (Raupach et al., 2009, 2012). However, the main constraint faced by this physical model is the high spatial and temporal variability of meteorological data, which may not be appropriate for small-scale applications, such as 'on-farm' decision making. Another drawback is that the *WaterDyn* physical model is accustomed to determine instantaneous ('now-casts') *SM* level at the point in time when meteorological inputs are channeled, requiring a constant supply of input variables. More precisely, this physical model is hindcasting since the system operates using already recorded meteorological data. For instance, monthly *SM* levels are attained after the observation and the accumulation of all the essential meteorological parameters are completed at the end of the month. Despite the advancements in measurement techniques, delayed progress in *SM* forecasting is evident. Particularly, the forecasted value of *SM* at the local scale (e.g., at the farm level) is imperative for key decision making but the current limitations in *SM* forecasting tools present a significant challenge in this respect.

To ameliorate *SM* predictability issues, the forecasting ability of advanced data-driven models offer feasible alternatives at the local scale modeling of *SM*. The predictive models are able to 'learn' from historical data making it advantageous for practical applications (Zhang et al., 1998). Data intelligent models have been successfully applied in agricultural and soil science applications to forecast field capacity and permanent wilting point (Ghorbani et al., 2017), soil water retention and saturated hydraulic conductivity (Merdun et al., 2006; Schaap and Leij, 1998) and soil temperature (Samadianfard et al., 2018). Yet, *SM* forecasting applications are still in their nascent stages (Liu et al., 2014; Matei et al., 2017; Myers et al., 2009; Yang et al., 2017). Researchers argue that forecasting of hydro-climatic variables must explore hybrid (rather than standalone) models building on the strengths of individual data-driven models (Jain and Kumar, 2007; Maier et al., 2010; Tiwari and Adamowski, 2013). Consequently, a new two-stage multi-model ensemble committee of models constructed on the basis of artificial neural networks (ANN) is explored in this study. The notion is to extract the pertinent information simulated by standalone expert models and further optimize it via an ANN for a collective forecast. This overcomes the weaknesses of conventional simple averaging forecast combinations whereby the overall model performance is compromised by the worst performing model(s). This novel multi-model ensemble committee of models approach has to overcome the inherent drawbacks of individual standalone models, building on the aptness, and subsequently surpassing the individual performances (Barzegar et al., 2017; Hatampour, 2013). The key advantage is that the committee based model combination reaps the benefit of all expert models yielding better generalization and performance, i.e., obtains a comparable or lower error than simple averaging and individual best single expert models (Barzegar and Moghaddam, 2016; Barzegar et al., 2015; Chen and Lin, 2006). Although, varying degree of a standalone ANN has been successfully applied in *SM* forecasting (Huang et al., 2010; Yang et al., 2017), model combination techniques have been overlooked in environmental applications (Baker and Ellison, 2008). Related committee modeling approaches were successfully applied in preparing groundwater vulnerability maps (Barzegar et al., 2017), groundwater contamination risk assessment (Barzegar et al., 2015) and groundwater salinity forecasting (Barzegar and Moghaddam, 2016). However, to the best of the authors' knowledge, *SM* forecasting is yet to be performed using the novel two-stage ensemble committee of models.

To collate the relevant features, four-standalone expert data-intelligent models viz., 2nd order Volterra, M5 tree, random forest (RF) and an extreme learning machine (ELM) model have been used. The 2nd order Volterra performed well in forecasting of streamflow (Maheswaran and Khosa, 2012, 2015; Rathinasamy et al., 2013), however*,* *SM* forecasting has not been piloted. Likewise, the M5 model tree has not been applied so far in *SM* forecasting, although a similar regression tree algorithm (Cubist) was noted (Myers et al., 2009). The *SM* forecasting from the bootstrapped-aggregated tree approach, RF, has yielded good performance with a reasonable prediction accuracy in one study in Romania (Matei et al., 2017). Literature shows that ELM is also uncommon in *SM* forecasting as only one study by Liu et al. (2014) in Victoria, Australia applied ELM and support vector machine for a short period (14 months) ignoring the long-term dynamics. Hence, overall, the application of data-driven models in the area of *SM* forecasting has not been fully exploited.

The objective of this study is to develop a low cost (saving labor, time, energy, and money) *SM* forecasting model using climate dynamics, i.e., the climate mode indices, atmospheric and hydro-meteorological drivers as the model inputs. The other factors such as vegetation cover, soil characteristics, i.e., soil texture, soil structure, initial *SM*, hydraulic conductivity, and *SM* pressure and land-use are assumed to be site-specific and constant in this study. Thus, historical hydro-meteorological variables from AWAP, climate indices, and the Interim ERA European Centre for Medium-Range Weather Forecasting reanalysis derived atmospheric data are collated leading to sixty inputs. Consequently, salient inputs are screened using a two-stage feature selection technique via Neighborhood Component Analysis based feature weights and modeled minimum relative error criteria. A novel well-trained two-stage hybrid multi-model ensemble committee based on ANN (ANN-CoM) data intelligent model is developed in forecasting upper and lower layer *SM* within the Murray-Darling Basin region, Australia. The performance of the new ANN-CoM model is evaluated with various statistical measures together and the diagnostic plots and this is benchmarked against the four primary standalone models.

2. Materials and methodology

2.1. Machine learning algorithms used in developing ensemble committee model

2.1.1. 2nd order volterra model

The Volterra model is built upon the Taylor series expansion for nonlinear autonomous causal systems with memory. A second-order representation has been adopted, as substantiated by previous studies (Labat et al., 1999; Maheswaran and Khosa, 2012, 2015; Rathinasamy et al., 2013). With $z(t)$ as the model output and t as the *tth* instances, the 2nd-order Volterra expansion could be expressed as:

$$z(t) = \int_{\tau_1=0}^{\tau_1=t} k_1(\tau_1) X(\tau - \tau_1) d\tau_1 + \int_{\tau_2=0}^{\tau_2=t} \int_{\tau_1=0}^{\tau_1=t} k_2(\tau_1, \tau_2) X(\tau - \tau_1) X(\tau - \tau_2) d\tau_1 d\tau_2 \qquad (1)$$

where $k_1(\tau_1)$ and $k_2(\tau_1, \tau_2)$ are the Volterra kernels. In a condensed notation Eq. (1) gives:

$$z(t) = K_1[x(t)] + K_2[x(t)] \tag{2}$$

where $K_1[x(t)]$ and $K_2[x(t)]$ are the 1st and 2nd order Volterra operators, respectively.

Since *SM* forecasting requires multiple predictor inputs, the Volterra series expansion for a multiple input single output (MISO) system is expressed as:

$$\begin{aligned} z(t) = & \sum_{n=1}^{N}\sum_{\beta=1}^{M} k_1^{(n)}(\beta)x_n(t-\beta) \\ & + \sum_{n=1}^{N}\sum_{\alpha=1}^{M}\sum_{\beta=1}^{M} k_{2s}^{(n)}(\alpha,\beta)x_n(t-\alpha)x_n(t-\beta) \\ & + \sum_{n1}^{N}\sum_{n2=1}^{n1-1}\sum_{i=1}^{M}\sum_{\beta=1}^{M} k_{2\times}^{(n1,n2)}(\alpha,\beta)x_{n1}(t-\alpha)x_{n2}(t-\beta) \end{aligned} \tag{3}$$

where N *is* the number of inputs; M represents the memory length of each significant lagged input variable; the $k_1^{(n)}$ is the first order kernels; $k_{2s}^{(n)}$ is the second order self kernels and $k_{2\times}^{(n1,n2)}$ is the second order cross-kernels.

Finally, the estimation of Volterra kernels was achieved via the principle of orthogonal least squares (OLS) since OLS suitably handles collinearity amongst predictor inputs (Billings et al., 1988; Wei and Billings, 2004).

2.1.2. M5 tree model

The M5 model tree developed by Quinlan (1992) is a hierarchical solitary modeling process with linear regression functions at the leaves having clearly expressed and easily understood decision structures. The algorithm splits the training data into subsets via 'divide-and-conquer' lemma (Bhattacharya and Solomatine, 2005). The splitting is contingent upon the minimization of intra-subset variation in the output variable's values down each branch. To achieve this, the attribute that maximizes the standard deviation reduction (*SDR*) is selected for splitting. For a training data set *T*, the *SDR* is computed as follows (Bhattacharya and Solomatine, 2005; Witten et al., 2011):

$$SDR = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i) \tag{4}$$

where $sd(T)$ is the standard deviation of the class values, T_i: $i = 1, 2, \ldots S$ are the sub-sets resulting from node splits. The splitting process ceases when the standard deviation is less than 5% of the original instance or when only a few instances remain.

A local specialized linear regression model is built, as depicted in Fig. 1a, leading to a large overfitting regression tree. Finally, the tree is pruned back from leaves and a smoothing process is employed to fade away the sharp discontinuities in between adjacent linear models at the leaves.

2.1.3. Random Forest (RF)

Random forest introduced by Breiman (2001) utilizes ensemble bootstrap aggregation (bagging) techniques (Breiman, 1996) to reduce the variance by computing the average of forecasts from several single regression tree models.

From the training data, '*btrp*' numbers of bootstrap replicas are taken through random sampling with replacement. Then '*J*' numbers of individual tree models are constructed and trained on a randomly selected subset of predictors of size '*m*' out of a total number of predictors, '*P*'. Finally, the outputs of single regression trees are put together, forming an ensemble and the forecasts are averaged over the ensemble. A built-in cross validation is carried out through the computation of out-of-bag (OOB) error, using the unutilized training data.

Three parameters require slight tuning i) the number of randomly assigned predictor variables (*m*), ii) the number of trees (*J*), and iii) the maximum number of terminal nodes/leaf (*tree size*). *J* is almost a non-issue, however, a high value diminishes the variance and increases the computational cost (Breiman, 2001). The recommended value of *m*, *i.e.*, one-third of the total number of variables ($m = \frac{1}{3}P$) (Liaw and Wiener, 2002) has been adopted, while, the parameters $J = 500$ and $tree\ size = 5$ generated the optimal results.

2.1.4. Extreme learning machine (ELM)

Developed by Huang et al. (2004), the ELM is a single layer feed-forward neural network (SLFN) acclaimed to have a good generalization capability. Fig. 1b illustrates a simplified architecture. ELM has proven to be computationally efficient overcoming the issues of over-fitting, stopping criteria, learning epochs.

For a training of *N* samples, the mathematical realization of the ELM algorithm can be described as follows (Huang et al., 2004):

$$\sum_{i=1}^{K} B_i\, G_i(\alpha_i,\ \beta_i,\ x_t) = z_t \tag{5}$$

where $z_t \in \boldsymbol{R}$ represents the model output, $B \in \boldsymbol{R}^K$ represents the output weights, $\alpha_i \in \boldsymbol{R}^K$ are the input weights and $\beta_i \in \boldsymbol{R}$ are the biases and; $i = 1, 2, \ldots K$ are the indices of hidden neurons. Compactly re-writing Eq. (5) after replacing the model output (z_t) with observed training data (Y) yields:

$$Y = GB \tag{6}$$

where G is the hidden layer output matrix, B are the weights and Y are the observed target training matrices. With the objective to yield zero forecasting errors, the ELM algorithm can be summarized as follows:

1 Aleatory allocation of input weights (α_i) and the biases (β_i).
2 Computation of hidden layer output matrix *G*.
3 Analytical determination of output weights matrix via a least-square solution as:

$$\widehat{B} = G^{\dagger}Y \tag{7}$$

where $G^{\dagger}$ is the Moore–Penrose generalized inverse of G.

1 Finally, generation of forecasts by feeding in the test dataset as inputs.

2.1.5. Artificial neural network - the basis of committee of models

The artificial neural network (ANN) developed by McCulloch and Pitts (1943) mimics the complex nonlinear structure of the human brain. ANNs have the ability to learn subtle functional relationships among the input-output data without a priori knowledge of the underlying physical system (Zhang et al., 1998). ANNs are nonlinear models proven to be robust, efficient, adaptive in noisy environments and working well with non-Gaussian data (Jain et al., 1999; Sehgal et al., 2014), therefore having been adopted as the basis of the committee of models (CoM) as shown in Fig. 1(c and d).

The multilayer feed-forward neural network comprises of three layers: an input layer, a hidden layer, and an output layer. During the network training, the input data series propagates in a forward direction, layer by layer, with simultaneous construction of the nonlinear relationship between the inputs and output based on the logical input-output mapping system (Fausett, 1994; Haykin, 1999). The interneuron weights and added biases are determined in a logical manner, via a learning algorithm, such that the mean squared error (*MSE*) in between the modeled output and the observed target is minimized.

An early stopping criterion has been adopted to avoid overfitting and finally, the effectiveness of the network was evaluated and verified using new (unseen) data sets in the testing phase. For more information, readers can refer to Haykin (1999).

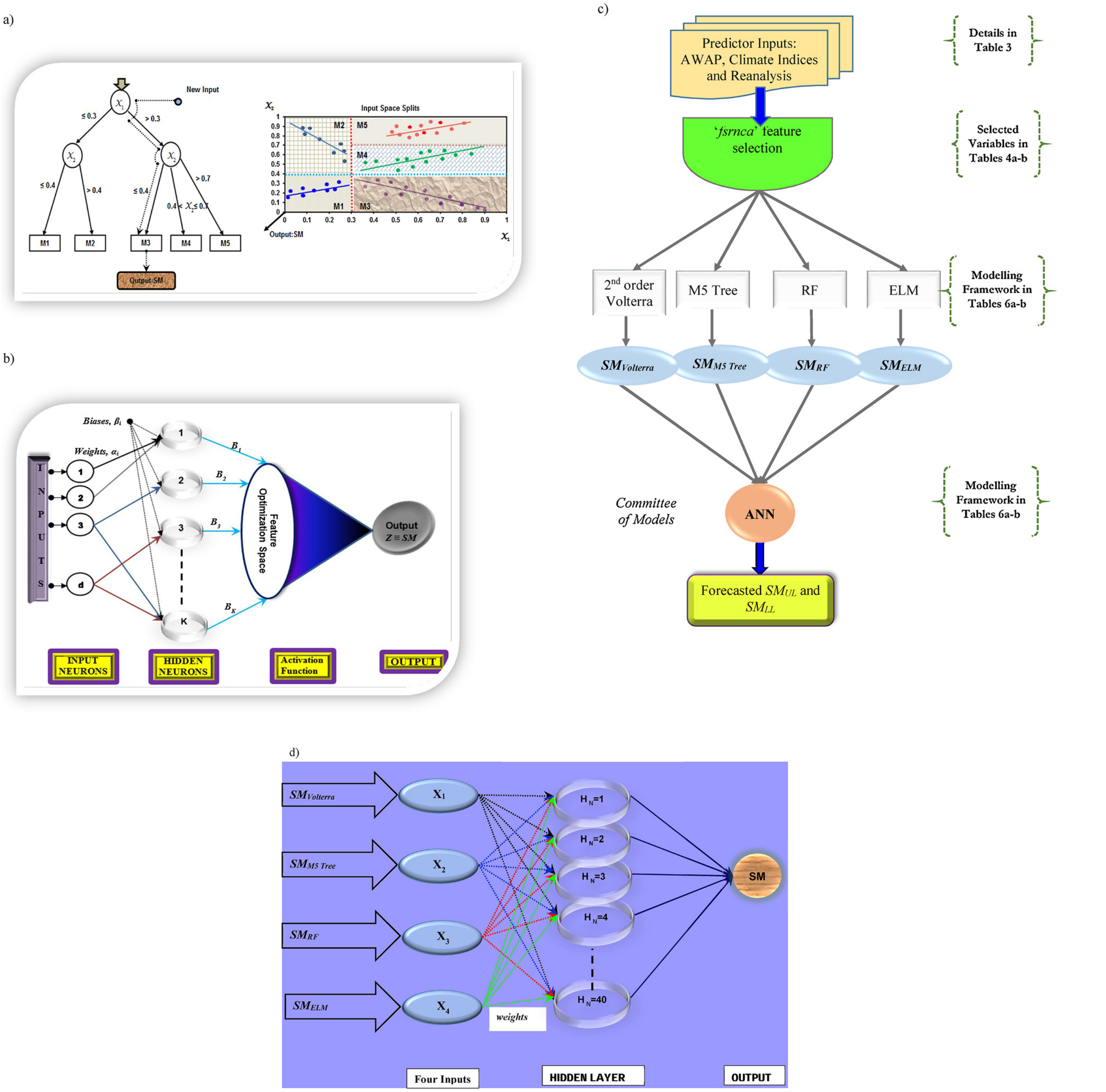


Fig. 1. The architecture of the newly proposed two-stage data-driven model used in relative soil moisture (SM) forecasting: (a) M5 model tree; (b) ELM; (c) Schematic view of the model development steps; (d) Multilayer Feed-forward Neural Network based Committee of Model (ANN-CoM).

2.1.6. Feature selection algorithm based on neighbourhood component analysis (NCA): '*fsrnca*'

Feature selection is an integral component of the model development process in order to minimize input dimensionality, reduce computational complexity, improve the accuracy, and increase the interpretability and understanding of the system dynamics (Bowden et al., 2005; Maier et al., 2010; Yang et al., 2012). The feature selection for regression based on Neighborhood Component Analysis (NCA) called *fsrnca* has been employed to isolate the salient inputs from the initial sixty input variables. Developed by Yang et al. (2012), the *fsrnca* algorithm is simple, efficient, nonlinear and a non-parametric embedded method. The *fsrnca* algorithm uses the training data to perform NCA feature selection with regularization in learning the feature weights via minimization of an objective function that measures the average leave-one-out regression loss.

In brief, for a training data set $T = \{(x_i, y_i): i = 1, 2, 3, \ldots, N\}$ where $x_i \in \boldsymbol{R^P}$ are the feature vectors, $y_i \in \boldsymbol{R}$ are the target and N is the number of samples in the training data set. The *fsrnca* algorithm learns a function $g(x)$: $\boldsymbol{R^P} \rightarrow \boldsymbol{R}$, to predict the response y from predictors, optimizing the nearest distances. The weighted distance (D_w) between two samples (for e.g., x_α and x_β) could be denoted as:

$$D_w(x_\alpha, x_\beta) = \sum_{j=1}^{J} w_j^2 \, |x_{\alpha j} - x_{\beta j}| \tag{8}$$

where w_j is a weight associated with the *jth* feature.

A probability distribution ($p_{\alpha\beta}$) is used to approximate the reference point to maximize its leave-one-out prediction accuracy on the training data set, whereby the probability of x_α selects x_β as its reference point. Subsequently, using a gradient ascent method, the *fsrnca* algorithm

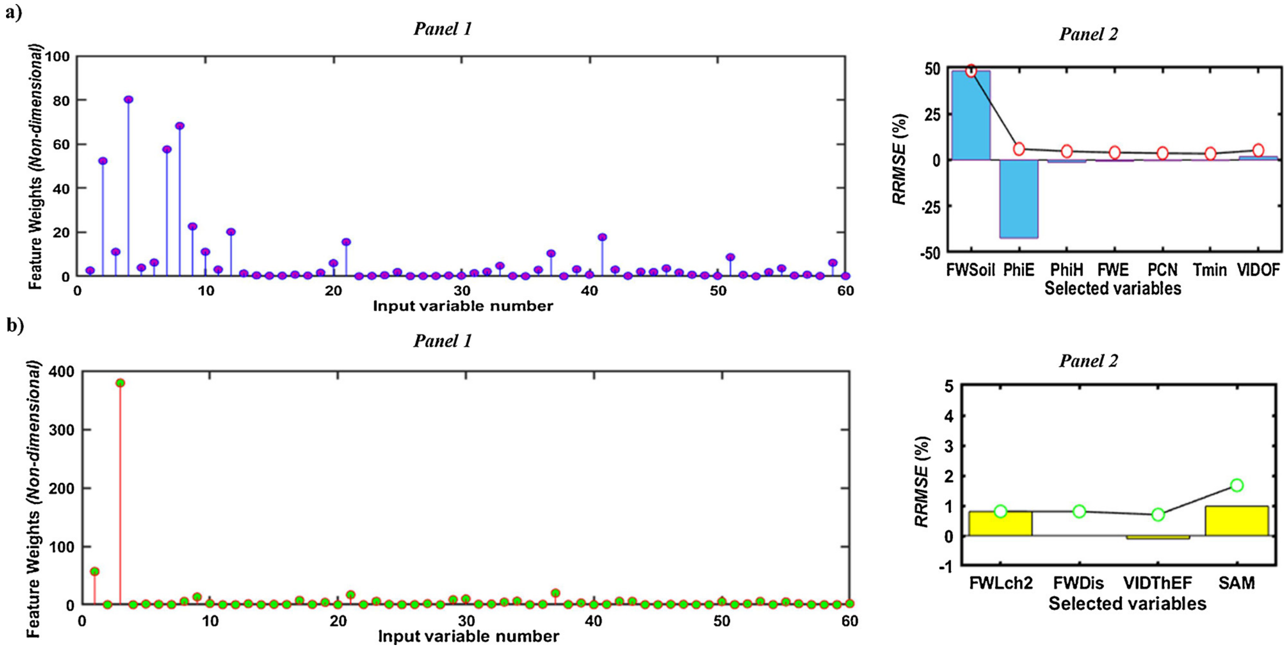


Fig. 2. The feature weights determined by the Neighbourhood Component Analysis for regression feature selection algorithm (*fsrnca*) from a pool of 60 input variables (Panel 1), and the corresponding changes in relative root mean square errors (*RRMSE*) with subsequent addition of each input (in the ascending order determined by the *fsrnca* feature weights) using a basic ELM model (Panel 2). (a) Site 1 upper layer soil moisture and (b) Site 1 lower layer soil moisture. Note the bars in Panel 2 show the decrement contribution in the *RRMSE* value of each variable, while the line graph shows the cumulative decrement in *RRMSE* values. (NB: Refer to Table 3 for full names of selected variables from abovementioned acronyms).

finds a weighting vector ‘***w***’ that offers itself in selecting the feature subset. A regularization parameter is introduced to avoid overfitting. Since the input feature vectors are in different scales and units, all the predictors were standardized before application of *fsrnca* feature selection. Fig. 2a and b illustrates the feature weights of all 60 inputs at Site 1.

2.2. Study region and data description

The study region, New South Wales (NSW), Australia, situated within the Murray-Darling Basin accounted for ∼23% of Australia’s agricultural production by value in the financial year 2015–16 (Australian Bureau of Statistics, 2017). Additionally, agriculture is the most important industry for rural dwellers (Campbell and Scarlett, 2014). Therefore, development and implementation of adequate *SM* forecasting tools are important to continue this thriving industry. As such, four sites (illustrated in Fig. 3) with distinct geophysical conditions including major climate classes (Hijmans et al., 2005), land use (Department of Agriculture and Water Resources, 2015), range of agricultural holding (Australian Bureau of Statistics, 2008) and soil types (ASRIS, 2014) were selected (Table 1) to test the performance of ensemble ANN-CoM under various situations.

The 0.05° × 0.05° spatially gridded monthly relative *SM* data for the upper soil layer (SM_{UL}) (0.2 m from surface) and the lower soil layer (SM_{LL}) (0.2–1.5 m deep) were obtained from the Australian Water Availability Project (AWAP) (Raupach et al., 2009, 2012). In order to create gridded data, AWAP employs an anomaly-based three-dimensional smoothing splines approach (Beesley et al., 2009; Tozer et al., 2012). It must be noted that the AWAP generated relative *SM* values are relative to the base climatological reference period: 1961 to 1990 (Raupach et al., 2009). Hence, the study period has been from January 1990 to December 2016 and the SM_{UL} and SM_{LL} soil layer depths were consistent as abovementioned for possible integrations into decision support systems in the future.

Mean climatological patterns of the SM_{UL} (Fig. 4a) and SM_{LL} (Fig. 4b) showed varied trends at the four candidate sites. The SM_{UL} (Fig. 4a) exhibited vivid maxima during June to August (winter). Three sites *viz.*, Site 1-Menindee, Site 3-Bobadah, and Site 4-Rocky Creek recorded minimum in April (autumn) while Site 2-Balranald recorded a minimum in March. However, the SM_{LL} at Sites 1 and 2 were very stable with no clear monthly or seasonal trend. At Sites 3 and 4, the largest magnitudes occurred during August-September months, *i.e.*, winter-spring transition periods, while the lowest values were recorded in May. The increased through-flow and deep percolation with varying meteorological factors could have been the possible contributing factors for this occurrence.

The monthly statistical features of SM_{UL} and SM_{LL}, shown in Table 2, reveal that Site 1-Menindee and Site 2-Balranald registered the least magnitude of SM_{UL} while Site 4-Rocky Creek recorded the highest value (0.814). Interestingly, the least value of SM_{LL} was recorded at Site 2-Balranald (0.034) while Site 4-Rocky Creek recorded the highest values (1.000). Subsequently, the skewness and kurtosis were computed to determine the characteristics of data distribution. The skewness measures the degree of symmetricity of a distribution and values outside the range of -2 and +2 indicate significant deviations from normality (Esmaeili et al., 2018). While, kurtosis is a measure of whether the data are heavy-tailed or light-tailed relative to a normal distribution and the kurtosis is 3 for a univariate normal distribution (DeCarlo, 1997). The skewness values for SM_{UL} were much closer to zero to confirm approximate symmetric data distributions. In the case of SM_{LL}, all sites except for Site 1-Menindee (skewness = 1.24) showed near-symmetric distributions. At all the four sites and for both SM_{UL} and SM_{LL}, the kurtosis coefficients were less than three (*platykurtic*), revealing that the distributions of both SM_{UL} and SM_{LL} displayed fewer and less extreme outliers (lighter tails and is flatter) (Table 2). Hence, distinctive geographical features are lucid which is bound to offer varying model testing conditions.

2.3. Data inputs and feature selection procedure

Data-driven models are completely dependent on the information from historical data-sets. Therefore, authentic and reliable global sets of

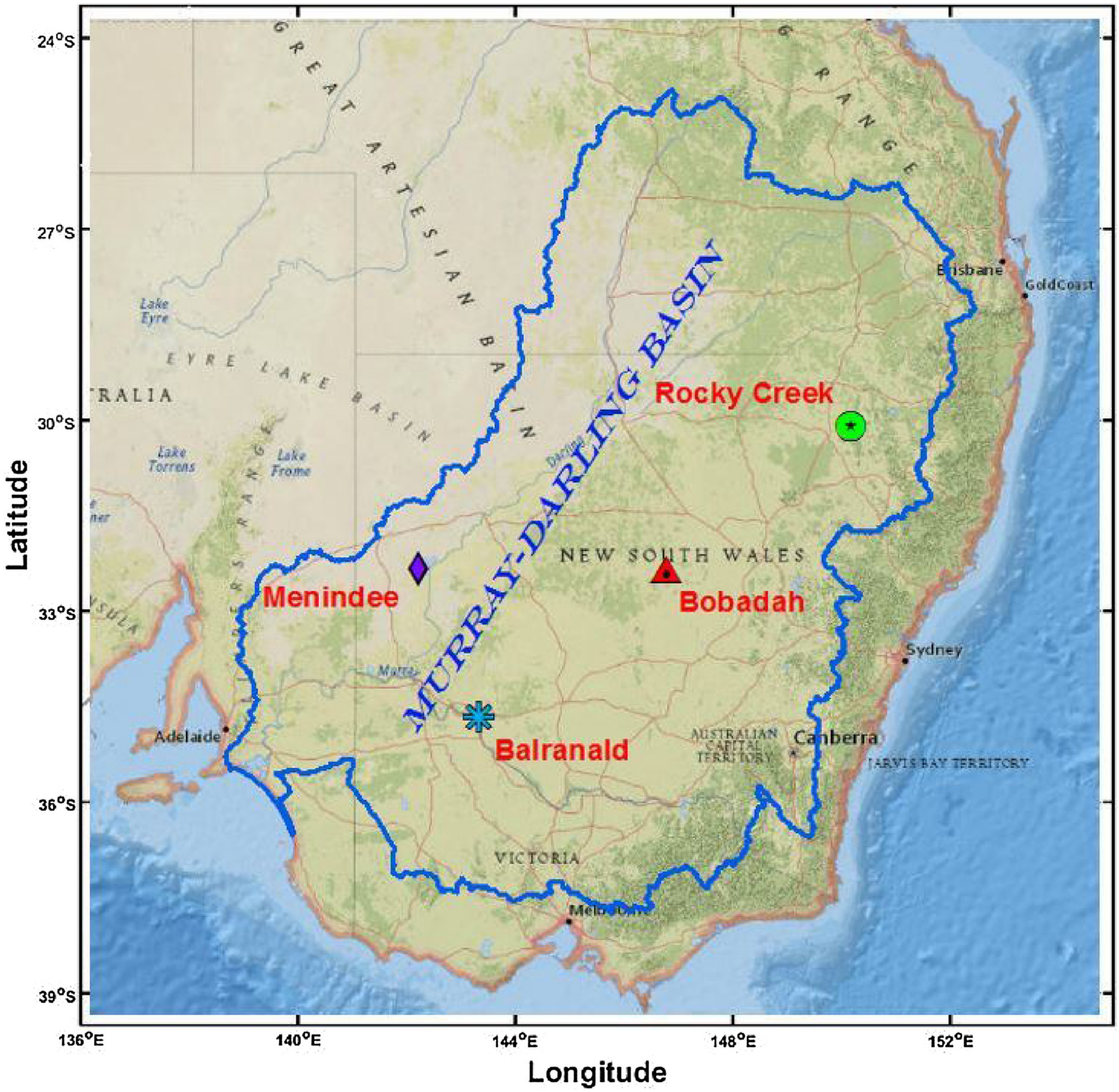


Fig. 3. Map of the study region showing the selected hydrological study sites and their geographical locations within Australian Murray-Darling basin region.

predictors, including hydro-meteorological AWAP data (Raupach et al., 2009, 2012), ECMWF reanalysis derived atmospheric data (Dee et al., 2011) and synoptic scale climate indices were collated, leading to 60 inputs (Table 3).

Next, the global predictor set was screened via Neighborhood Component Analysis feature selection for regression algorithm (*fsrnca*) to select optimal features (Section 2.1.6). The *fsrnca* algorithm computes the relative weights of each predictor inputs in determining the objective variable, *SM* (Yang et al., 2012) (e.g., Fig. 2a and b (left-hand side panels)). However, the problem regarding the optimum threshold weight (above which the input variable needs to be selected) remained unresolved. To address this, an innovative and effective method was developed whereby all the input variables were ranked according to *fsrnca* feature weights. Then a basic-ELM model with 50 hidden neurons and *sigmoid* transfer functions was applied to assess the effectiveness of each historic input variable in predicting the *SM* based on the relative

Table 1
Geographic locations and physical characteristics of the selected sites in the Australian Murray-Darling Basin.

Site No.	Station Names	Location			Physical Characteristics			
		Long. (°E)	Lat. (°S)	Approx. Elevation (m)	Major Climate Classes (Hijmans et al., 2005)	Land Use (Department of Agriculture and Water Resources, 2015)	The range of agricultural holding (ha) (Australian Bureau of Statistics, 2008)	Soil Type (ASRIS, 2014)
1	Menindee	142.15	−32.45	75.3	Desert	Grazing-Native vegetation	18700-38600	Calcarosol
2	Balranald	143.30	−34.75	65.5	Savannah	Dry-land cropping	3700-18700	Calcarosol
3	Bobadah	146.75	−32.45	277.3	Savannah	Dry-land cropping	600-3700	Kandosol
4	Rocky Creek	150.20	−30.15	689.0	Sub-Tropical	Dry-land cropping	3700-18700	Sodosol

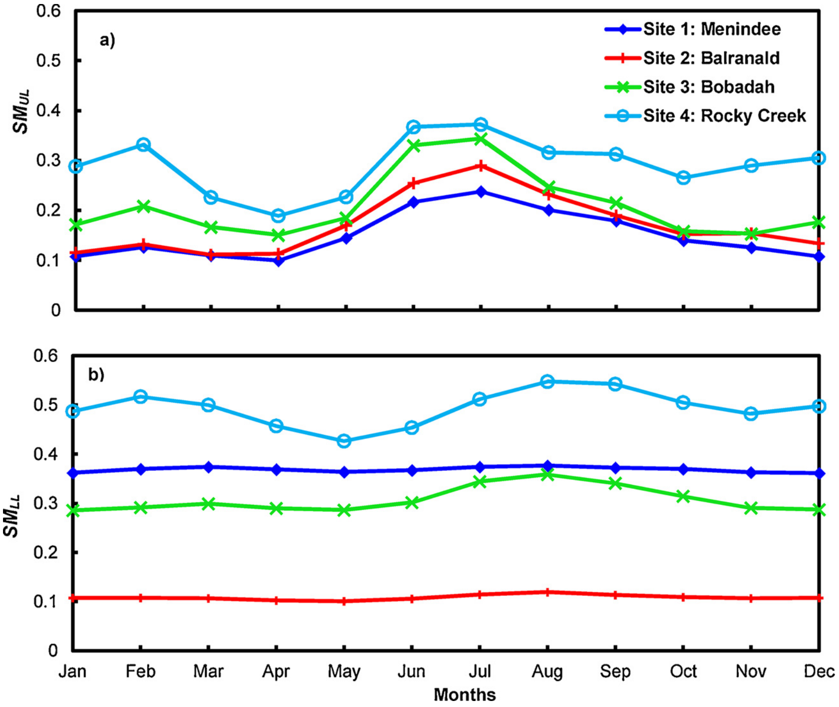


Fig. 4. The monthly variations in (a) upper layer (SM_{UL}) and (b) lower layer soil moisture (SM_{LL}) at the four study sites. (SM_{UL} *and* SM_{LL} are the relative fractional values and the unit is dimensionless).

root mean square error (*RRMSE*). Consequently, the most significant variables (based on *fsrnca* feature weights) were successively added to the input variable set and the basic-ELM was executed with simultaneous monitoring of *RRMSE*. When no significant improvements in the performance were achieved, *i.e.*, when the selection of a further variable led to a decrease of *RRMSE* lower than 0.01%, the algorithm terminated. The Panel 2 (i.e., right-hand side panel) of Fig. 2a and b shows examples of plots of *RRMSE* and the selected variables at Site 1. For SM_{UL}, the cumulative *RRMSE* decreased monotonically with the number of selected variables, up to the sixth variable, minimum temperature (Tmin). When the seventh variable i.e., the Vertical integral of the divergence of ozone flux (VIDOF) is selected, no further significant decrease in *RRMSE* is recorded and the algorithm terminated. Likewise, for SM_{LL}, after selecting the third variable i.e., the Vertical integral of the divergence of thermal energy flux (VIDThEF), the cumulative *RRMSE* increased asserting three significant variables. Table 4a and b show a summary of salient input variables with monthly statistical features. The salient inputs for forecasting SM_{UL} showed varied statistical properties. At Site 1-Menindee, three inputs (FWSoil, PhiE, and PCN) showed *leptokurtic* (kurtosis > 3) distributions, i.e., the central peaks of data distributions were higher, sharper, with longer and fatter tails in comparison to a normal distribution. While, the other three inputs (PhiH, FWE, and Tmin) showed *platykurtic* distribution, i.e., in comparison to a normal distribution, the central peaks were lower and broader with shorter and thinner tails. Interestingly, FWSoil at Site 2-Balranald showed a *mesokurtic* (almost normal) distribution with kurtosis = 3.05. At Site 4-Rocky Creek all inputs were *platykurtic*. For SM_{LL} forecasting, except for VIDThEF (skewness = −0.15: kurtosis = 0.01), the distributions of all other inputs at all sites were highly positively skewed (skewness > +1) and *leptokurtic* (kurtosis > 3).

2.4. Model development

One-month antecedent salient inputs as shown in Table 4a-b were used as predictors to the standalone models, while the target data were the time-series of posterior observed monthly SM_{UL} and SM_{LL}. Prior to training the models, all inputs were normalized to conform to the range (0, 1) (Deo and Sahin, 2016; Deo et al., 2017a,b). Then the data were sequentially divided into subsets having training (70%), validation (15%) and testing (15%) as indicated in Table 5. It is important to note that the independent (validation) data were used to screen the optimal model (from several trained models with ELM and ANN algorithms were used) to ensure that the most accurately trained model was selected. The selected model that performed the best in terms of the root mean square error in the validation set was thus used to forecast soil moisture in the testing phase. Initially, following this procedure, the standalone models were developed and implemented, followed by the ensemble committee of models for the final forecasting of relative soil moisture.

Table 2
Monthly hydrological statistics of the upper and lower layer relative soil moisture at the selected sites. (SM_{UL} *and* SM_{LL} are the relative fractional values and the unit is dimensionless).

Site No.	Station Names	Monthly statistical features of upper layer soil moisture (SM_{UL})					Monthly statistical features of lower layer soil moisture (SM_{LL})				
		Min.	Max.	Mean	Skew-ness	Kurtosis	Min.	Max.	Mean	Skew-ness	Kurtosis
1	Menindee	0.013	0.434	0.139	0.791	0.058	0.205	0.703	0.343	1.241	1.421
2	Balranald	0.013	0.470	0.159	0.668	−0.180	0.034	0.187	0.092	0.695	−0.533
3	Bobadah	0.015	0.520	0.197	0.572	−0.272	0.119	0.560	0.290	0.436	−0.840
4	Rocky Creek	0.036	0.814	0.285	0.463	0.489	0.145	1.000	0.473	0.442	0.218

[NB: The relative soil moisture values are based on the base climatological reference period: 1961–1990, recommended by Australian Bureau of Meteorology].

Table 3
Database of the input variables used to develop the ANN-CoM and the comparative data-driven models in forecasting the SM_{UL} and SM_{LL}.

	Variables	*Acronym*	Units	Source
I	Total Monthly Local Discharge (Runoff + Drainage)	FWDis	mm	Australian Water Availability Project (AWAP) (Raupach et al., 2009, 2012) [Definitions of abbreviations in Units used: mm = millimeters, W = Watts, m = meters, MJ = Mega-Joules, ˚C = degrees Celsius]
N	Total Monthly Total Evaporation (Soil + Vegetation)	FWE	mm	
P	Total Monthly Deep Drainage	FWLch2	mm	
U	Total Monthly Soil Evaporation	FWsoil	mm	
T	Total Monthly Total Transpiration	FWTra	mm	
S	Total Monthly Open Water Evaporation ('pan' equiv.)	FWWater	mm	
	Monthly Average Sensible Heat Flux	PhiH	W/m^2	
	Monthly Average Latent Heat Flux	PhiE	W/m^2	
	Total Monthly Precipitation	PCN	mm	
	Monthly Average Incident Solar Radiation	SolarMJ	MJ/m^2	
	Monthly Average Maximum Temperature	Tmax	˚C	
	Monthly Average Minimum Temperature	Tmin	˚C	
	SST of NINO 1 + 2 region	NINO 1 + 2		Sea Surface Temperature (SST): Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4) - Climate Prediction Centre-NOAA
	SST of NINO3 region	NINO3		
	SST of NINO4 region	NINO4		
	SST of NINO3.4 region	NINO3.4		
	Tripole Index for the Interdecadal Pacific Oscillation	TPI (IPO)		
	Dipole Mode Index (Previously known as IOD)	DMI		Japan Agency for Marine-Earth Science and Technology (JAMSTEC)
	El Nino Modoki Index	EMI		
	Pacific Decadal Oscillation	PDO		Joint Institute of the Study of Atmosphere and Ocean (JISAO)
	Southern Oscillation Index	SOI		BOM-Australia
	Southern Annular Mode Index	SAM		Natural Environment Research Council (NERC)
	Vertical integral of the mass of the atmosphere	VIMA	kg/m^2	Interim ERA European Centre for Medium-Range Weather Forecasting (ECMWF) (Dee et al., 2011) [Definitions of abbreviations in Units used: kg = kilograms, m = meters, K = Kelvin, J = Joules, W = Watts, s = seconds, Pa = Pascals, ˚C = degrees Celsius]
	Vertical integral of temperature	VIT	K kg/m^2	
	Vertical integral of water vapor	VIWA	kg/m^2	
	Vertical integral of cloud liquid water	VICLWA	kg/m^2	
	Vertical integral of ozone	VIO	kg/m^2	
	Vertical integral of kinetic energy	VIKE	J/m^2	
	Vertical integral of thermal energy	VIThE	J/m^2	
	Vertical integral of potential + internal energy	VIPIE	J/m^2	
	Vertical integral of potential + internal + latent energy	VIPILE	J//m^2	
	Vertical integral of total energy	VITotE	J/m^2	
	Vertical integral of energy conversion	VIEC	W/m^2	
	Vertical integral of the divergence of cloud liquid water flux	VIDCLWF	kg/m^{2}s	
	Vertical integral of divergence of mass flux	VIDMF	kg/m^{2}s	
	Vertical integral of divergence of kinetic energy flux	VIDKEF	W/m^2	
	Vertical integral of divergence of thermal energy flux	VIDThEF	W/m^2	
	Vertical integral of divergence of moisture flux	VIDMF	kg/m^{2}s	
	Vertical integral of divergence of geopotential flux	VIDGF	W/m^2	
	Vertical integral of divergence of total energy flux	VIDTotEF	W/m^2	
	Vertical integral of divergence of ozone flux	VIDOF	kg/m^{2}s	
	Vertical integral of mass tendency	VIMT	kg/m^{2}s	
	Surface air pressure	sp	Pa	
	Total column water	tcw	kg/m^2	
	Total column water vapour	tcwv	kg/m^2	
	Soil temperature level 1 (depth: 0.00 - 0.07 m)	stl1	˚C	
	Soil temperature level 2 (depth: 0.07 - 0.28 m)	stl2	˚C	
	Soil temperature level 3 (depth: 0.28 – 1.00 m)	stl3	˚C	
	Soil temperature level 4 (depth: 1.00 – 2.89 m)	stl4	˚C	
	Mean sea level pressure	msl	Pa	
	Total cloud cover	tcc	(0 - 1)	
	10 metre U wind component	u10	m/s	
	10 metre V wind component	v10	m/s	
	2-meter temperature	t2m	˚C	
	2-metre dewpoint temperature	d2m	˚C	
	Surface albedo	al	(0 - 1)	
	Low cloud cover	lcc	(0 - 1)	
	Medium cloud cover	mcc	(0 - 1)	
	High cloud cover	hcc	(0 - 1)	
	Total column ozone	tco3	kg/m^2	
Objective Variables	Relative soil moisture: upper layer (UL) and the lower layer (LL)	SM_{UL} SM_{LL}	Fraction 0 - 1	Australian Water Availability Project (AWAP) (Raupach et al., 2009, 2012)

Table 4
The most salient input variables with respective statistical features applied at the candidate study sites, as determined by the Neighbourhood Component Analysis for regression (*fsrnca*) feature selection algorithm with the minimum relative root mean square error (*RRMSE*) for the forecasting of (a) upper layer soil moisture and (b) lower layer soil moisture. (*SM_{UL} and SM_{LL}* are the relative fractional values and the unit is dimensionless).

a) Upper Layer Soil Moisture (*SM_{UL}*)

Site No. & Station names	Names of Selected Variable	*Acronym (units)*	Monthly statistical features				
			Min.	Max.	Mean	Skew-ness	Kurtosis
Site 1- Menindee	Soil Evaporation	Fwsoil (mm)	2.09	85.29	18.76	1.71	4.36
	Latent Heat Flux	PhiE (W/m^2)	0.31	79.50	13.90	2.09	6.50
	Sensible Heat Flux	PhiH (W/m^2)	3.55	135.86	34.63	1.31	2.46
	Total Evaporation (Soil + Vegetation)	FWE (mm)	64.13	340.36	195.83	−0.10	−1.26
	Precipitation	PCN (mm)	0.00	171.10	19.40	2.81	12.91
	Minimum temperature	Tmin (°C)	2.49	23.63	11.96	0.15	−1.15
Site 2 - Balranald	Soil Evaporation	Fwsoil (mm)	0.38	54.14	11.23	1.76	3.05
	Latent Heat Flux	PhiE (W/m^2)	6.10	119.94	46.08	0.72	0.06
	Sea Surface Temperature (SST) in NINO 1 + 2 region	NINO 1 + 2 (*dimensionless*)	35.61	309.65	171.37	−0.04	−1.37
	Sensible Heat Flux	PhiH (W/m2)	19.69	29.12	23.38	0.27	−1.02
Site 3 - Bobadah	Soil Evaporation	FWSoil (mm)	0.29	71.54	15.94	1.48	2.71
	Latent Heat Flux	PhiE(W/m2)	47.25	333.93	176.29	0.06	−1.17
	Precipitation	PCN (mm)	0.00	210.70	35.52	2.02	5.60
Site 4 - Rocky Creek	Soil Evaporation	FWSoil (mm)	1.37	55.18	18.39	0.91	0.31
	Latent Heat Flux	PhiE (W/m^2)	61.19	305.94	170.09	0.10	−1.06
	Precipitation	PCN (mm)	47.39	302.80	141.47	0.44	−0.53
	Open Water Evaporation ('pan' equiv.)	FWWater (mm)	0.00	314.50	69.17	1.42	2.38
	Maximum temperature	Tmax (°C)	13.85	34.84	24.62	−0.12	−1.26

b) Lower Layer Soil Moisture (*SM_{LL}*)

Site No. & Station names	Names of Selected Variable	*Acronym (units)*	Monthly statistical features				
			Min.	Max.	Mean	Skew-ness	Kurtosis
Site 1- Menindee	Deep Drainage	FWLch2 (mm)	0.08	27.42	0.99	9.52	116.77
	Local Discharge (Runoff + Drainage)	FWDis (mm)	0.08	6.43	0.88	2.46	6.47
	Vertical integral of divergence of thermal energy flux	VIDThEF (W/m^2)	−323.02	854.70	252.31	−0.15	0.01
Site 2 - Balranald	Deep Drainage	FWLch2 (mm)	0.00	0.21	0.01	4.61	26.44
Site 3 - Bobadah	Deep Drainage	FWLch2 (mm)	0.01	40.04	1.18	7.42	62.03
	Local Discharge (Runoff + Drainage)	FWDis (mm)	0.01	6.49	0.64	3.01	12.20
Site 4 - Rocky Creek	Deep Drainage	FWLch2 (mm)	0.02	96.98	6.88	4.07	17.24
	Local Discharge (Runoff + Drainage)	FWDis (mm)	0.02	24.76	3.13	2.94	10.15

2.4.1. *Standalone model development*

The 2nd order Volterra model was the first standalone model to be developed. Then the M5 Tree software package developed by Jekabsons (2010) was utilized to erect, prune and validate using 10-fold cross-validations (Bhattacharya and Solomatine, 2005; Deo et al., 2017a,b). Table 6a-b summarize the number of rules for each optimal model. Following that, random forest (RF) models were built using the '*tree-bagger*' MATLAB functions. Average values of the three unique parameters i) Delta criterion decision split (C), ii) Number of predictor split (N_p), and iii) Permuted predictor delta error (E_D) are shown in Table 6a and b. For ELM, hidden neurons from 50 to 200 were trialed (Deo and Sahin, 2016; Yaseen et al., 2016) with various combinations of transfer functions including sigmoidal, sine, hard-limit, triangular basis, and radial basis. Table 6a-b show resulting best ELM models with unique architectures at all four sites.

2.4.2. *Multi-model ensemble committee of model development*

A second-stage optimized hybrid ensemble model was established by channeling the outputs from the above-mentioned models as inputs to the feed-forward ANN model producing a multi-model ensemble committee of models (ANN-CoM). Fig. 1c illustrates a schematic view of modeling stages. Through determination of appropriate weights and biases, the ANN was used to generate the optimized collective forecasts. Fig. 1d shows a simplified architecture of the ANN-CoM while the modeling frameworks of best ANN-CoM models at respective sites are presented in Table 6a and b. The determination of optimal hidden layer neurons is important to avoid significantly small architecture that could lack sufficient degrees of freedom or unnecessarily large architecture that might cause overfitting (Karunanithi et al., 1994). Thus, hidden neurons from 1 to 40 in increments of 1 were trialed. The newly developed ANN-CoM model was further optimized by trialing various combinations of the hidden transfer and output functions one at a time combined with two training algorithms (i.e., the Levenberg–Marquardt (*trainlm*) and the Quasi-Newtonian Broyden-Fletcher-Goldfarb, and Shanno (*trainbfg*) algorithms). The optimal models in each case were averred based on Pearson's correlation coefficient (r), root mean square

Table 5
Data partitions used in this study.

Sites	Period	Number of datum points	Data Partition		
			Training	Validation	Testing
All sites and for both upper and lower soil moisture	Jan 1990 to Dec 2016	324 – 1 = 323	70% 227 1990-2008	15% 48 2009-2012	15% 48 2013-2016

Table 6
The model development framework for the extreme learning machine (ELM), random forest, M5 model tree and the Multilayer Feed-forward Neural Network based Committee of Model (ANN-CoM) adopted in forecasting (a) SM_{UL}, (b) SM_{LL}.

Lower Layer Soil Moisture (SM_{LL})	M5 Tree	Random Forest			Extreme Learning Machine (ELM)				ANN-based committee of models (ANN-CoM)					
	No. of Rules	Avg. Delta Criterion Decision Split (C)	Avg. Number of predictor split (N_p)	Avg. Permuted Predictor Delta Error (E_D)	No. of Neurons			Transfer Function	No. of Neurons			Hidden transfer function	Output transfer function	Training algorithm
					Input Layer	Hidden Layer	Output Layer		Input Layer	Hidden Layer	Output Layer			
Site 1- Menindee	10	5.34exp -05	146.97	1.07	6	65	1	*sig*	4	26	1	*tansig*	*tansig*	*trainlm*
Site 2 - Balranald	20	8.97exp -05	217.42	1.55	4	74	1	*tribas*	4	16	1	*logsig*	*purelin*	*trainlm*
Site 3 - Bobadah	11	0.0001	166.67	1.78	3	52	1	*sig*	4	6	1	*logsig*	*tansig*	*trainbfg*
Site 4 - Rocky Creek	6	0.0001	176.87	1.18	5	55	1	*sig*	4	13	1	*logsig*	*purelin*	*trainbfg*

Table 7
The performances of ANN-CoM *vs.* the comparative models in the model development (*i.e.*, training and validation) phase, based on the Pearson's correlation coefficient (r), root mean square error ($RMSE$) and the mean absolute error (MAE). Note: (a) upper layer (SM_{UL}), b) lower layer (SM_{LL}) soil moisture. (All statistical metrics are dimensionless).

a) Upper Layer Soil Moisture (SM_{UL})	2nd order Volterra			M5 Tree			Random Forest			ELM			ANN-CoM		
	r	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*
Training Phase															
Site 1- Menindee	0.981	0.018	0.015	0.989	0.014	0.009	0.986	0.017	0.010	0.999	0.004	0.003	0.999	0.004	0.003
Site 2 - Balranald	0.917	0.041	0.028	0.981	0.020	0.015	0.986	0.019	0.012	0.992	0.013	0.010	0.996	0.010	0.007
Site 3 - Bobadah	0.863	0.072	0.052	0.975	0.028	0.020	0.982	0.026	0.019	0.994	0.014	0.010	0.990	0.017	0.013
Site 4 - Rocky Creek	0.948	0.052	0.042	0.982	0.027	0.020	0.981	0.030	0.018	0.996	0.012	0.009	0.995	0.013	0.010
Validation Phase															
Site 1- Menindee	0.982	0.020	0.016	0.986	0.018	0.014	0.973	0.030	0.022	0.950	0.040	0.011	0.988	0.019	0.007
Site 2 - Balranald	0.885	0.050	0.035	0.965	0.028	0.023	0.946	0.039	0.029	0.890	0.060	0.030	0.945	0.035	0.023
Site 3 - Bobadah	0.822	0.078	0.057	0.972	0.028	0.023	0.960	0.035	0.028	0.981	0.024	0.017	0.977	0.024	0.018
Site 4 - Rocky Creek	0.942	0.062	0.048	0.960	0.039	0.026	0.963	0.040	0.029	0.991	0.019	0.014	0.989	0.022	0.015
b) Lower Layer Soil Moisture (SM_{LL})	2nd order Volterra			M5 Tree			Random Forest			ELM			ANN-CoM		
	r	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*	*r*	*RMSE*	*MAE*
Training Phase															
Site 1- Menindee	0.923	0.046	0.032	0.998	0.007	0.005	0.996	0.010	0.005	1.000	0.002	0.002	1.000	0.003	0.002
Site 2 - Balranald	0.933	0.016	0.013	0.994	0.005	0.003	0.997	0.003	0.001	0.998	0.003	0.002	1.000	0.001	0.001
Site 3 - Bobadah	0.932	0.045	0.033	0.997	0.010	0.007	0.996	0.011	0.004	1.000	0.004	0.003	1.000	0.002	0.002
Site 4 - Rocky Creek	0.908	0.080	0.058	0.996	0.016	0.011	0.994	0.019	0.007	0.999	0.006	0.004	1.000	0.005	0.003
Validation Phase															
Site 1- Menindee	0.959	0.043	0.033	0.995	0.015	0.009	0.987	0.026	0.013	1.000	0.002	0.002	0.813	0.092	0.022
Site 2 - Balranald	0.948	0.036	0.024	0.953	0.066	0.031	0.944	0.039	0.019	0.985	0.025	0.010	0.930	0.036	0.021
Site 3 - Bobadah	0.916	0.046	0.033	0.999	0.007	0.006	0.999	0.004	0.002	0.994	0.011	0.004	0.996	0.008	0.004
Site 4 - Rocky Creek	0.923	0.094	0.070	0.997	0.015	0.010	0.995	0.018	0.008	0.916	0.114	0.042	0.998	0.012	0.007

error (*RMSE*) and mean absolute error (*MAE*) during validation phases (Table 7a-b) with the least mean square error (*MSE*) for confirmation. The multi-model ensemble ANN-Committee of models and comparative models (viz., Volterra, M5 model tree, random forest, ELM) were developed on the MATLAB platform running over Intel *i*7, 3.40 GHz processor. Finally, the testing data were utilized to assess the generalization capabilities of ANN-CoM based on the following model evaluation measures.

2.5. Model evaluation measure

Model evaluation is an important process as it confirms the acceptability and reliability of respective models. To carry out a comprehensive assessment, various statistical evaluation measures were used to capitalize on the benefits of the individual measures. The equations are as follows:

i Correlation coefficient (r):

$$r = \frac{\sum_{i=1}^{N}(SM_{UL}^{OBS,i} - \overline{SM_{UL}^{OBS}})(SM_{UL}^{FOR,i} - \overline{SM_{UL}^{FOR}})}{\sqrt{\sum_{i=1}^{N}(SM_{UL}^{OBS,i} - \overline{SM_{UL}^{OBS}})^2}\sqrt{\sum_{i=1}^{N}(SM_{UL}^{FOR,i} - \overline{SM_{UL}^{FOR}})^2}} \tag{9}$$

ii Root mean square error ($RMSE$):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(SM_{UL}^{FOR,i} - SM_{UL}^{OBS,i})^2} \tag{10}$$

iii Mean absolute error (MAE):

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|(SM_{UL}^{FOR,i} - SM_{UL}^{OBS,i})| \tag{11}$$

iv Willmott's Index (*WI*):

$$WI = 1 - \left[\frac{\sum_{i=1}^{N} (SM_{UL}^{OBS,i} - SM_{UL}^{FOR,i})^2}{\sum_{i=1}^{N} (|SM_{UL}^{FOR,i} - \overline{SM_{UL}^{OBS}}| + |SM_{UL}^{OBS,i} - \overline{SM_{UL}^{OBS}}|)^2} \right] \quad (12)$$

v Nash–Sutcliffe Efficiency (E_{NS}):

$$E_{NS} = 1 - \left[\frac{\sum_{i=1}^{N} (SM_{UL}^{OBS,i} - SM_{UL}^{FOR,i})^2}{\sum_{i=1}^{N} (SM_{UL}^{OBS,i} - \overline{SM_{UL}^{OBS}})^2} \right] \quad (13)$$

vi Legates-McCabe's Index (*L*):

$$L = 1 - \left[\frac{\sum_{i=1}^{N} |SM_{UL}^{FOR,i} - SM_{UL}^{OBS,i}|}{\sum_{i=1}^{N} |SM_{UL}^{OBS,i} - \overline{SM_{UL}^{OBS}}|} \right] \quad (14)$$

The, SM_{UL}^{OBS} represents the observed upper layer (*UL*) soil moisture and SM_{UL}^{FOR} is the forecasted upper layer soil moisture, *i* represents the occurrence time/place and *N* is the number of data points. (N.B. Subscript *UL* is replaced with *LL* in the case of lower layer soil moisture). The first metric, Pearson's correlation coefficient (*r*) [$Range = (-1, +1)$; $Ideal\,value = +1$] is absolute and non-dimensional. It quantifies the strength and direction of linear association in between observed SM_{UL}^{OBS} *or* SM_{LL}^{OBS} and forecasted values SM_{UL}^{FOR} *or* SM_{LL}^{FOR}. Yet, mediocre or poor models could achieve high correlations. The root mean square error (*RMSE*) [$Range = (0, +\infty$); $Ideal\,value = 0$] and mean absolute error (*MAE*) [$Range = (0, +\infty)$; $Ideal\,value = 0$] are absolute error measures and cannot be applied to compare the performance of models in different unitary systems/sites (Hora and Campos, 2015). However, both are deemed to provide more information about respective model performances than the relative measures (Legates and McCabe, 1999). A bias towards high *SM* level events is induced in *RMSE* by the square-root of the squared error values. Likewise, the goodness-of-fit measure Willmott's Index (*WI*) or index of agreement [$Range = (0, +1)$; $Ideal\,value = +1$] being a ratio of mean square error to potential error, is better at handling differences in modelled and observed means and variances (Bennett et al., 2013; Willmott, 1984). However, its limitation is the interpretation of physical meaning since zero is rather meaningless and it registers higher values (≥0.65) for poor models. Instead, the Nash–Sutcliffe Efficiency (E_{NS}) [$Range = (-\infty\,, +1)$; $Ideal\,value = +1$] compares the performance of the model to a model using mean of the observed data (Bennett et al., 2013; Legates and McCabe, 2013). $E_{NS} = 0$ indicates that the performance is no better than using the means while negative values indicate that the forecasted values diverge. Both, *WI* and E_{NS}, are sensitive to outliers due to the squaring of the difference terms. In contrast, the Legate-McCabe's index (*L*) [$Range = (-\infty\,, +1)$; $Ideal\,value = +1$] considers absolute values for computation and gives errors and differences the appropriate weights (Legates and McCabe, 1999). Therefore, *L* is not inflated by the squared values and is insensitive to outliers making it simple and easy to interpret.When comparing models at different sites, the relative measures, Relative Root Mean Square Error (*RRMSE*) [$Range = (0, +\infty)$; $Ideal\,value = 0$] and Mean Absolute Percentage Error (*MAPE*) [$Range = (0, +\infty)$; $Ideal\,value = 0$] are used. The advantage of being scale independent makes these relative measures more appropriate (Hyndman and Koehler, 2006). Particularly, *MAPE* does not have to offset positive and negative values of forecasting error (Hora and Campos, 2015). The equations are as follows:

I

I Relative root mean square error (*RRMSE*, %):

$$RRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N} (SM_{UL}^{FOR,i} - SM_{UL}^{OBS,i})^2}}{\frac{1}{N}\sum_{i=1}^{N} (SM_{UL}^{OBS,i})} \times 100 \quad (15)$$

II Mean absolute percentage error (*MAPE*; %):

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} \left| \frac{(SM_{UL}^{FOR,i} - SM_{UL}^{OBS,i})}{SM_{UL}^{OBS,i}} \right| \times 100 \quad (16)$$

[NB: The symbols used have the same meaning as mentioned above.]

The primary measures for model assessments were *RMSE* and *MSE* and *r*. Then *WI*, $E_{NS,}$ and *L* provided further goodness-of-fit assessments whereby *L* was superior and eventually, *RRMSE* and *MAPE* were used to compare models at different sites. In addition, graphical comparisons of data patterns using various diagnostic plots e.g., scatter plots, histograms, box-plots, polar plots, Taylor diagram and bar graphs were used to provide further insights and to avoid traducing model errors in terms of single magnitudes.

3. Results and discussion

The newly proposed hybrid multi-model ensemble ANN based committee of models (ANN-CoM) is evaluated against standalone 2nd order Volterra model, M5 model tree, random forest (RF) and ELM models at four sites within the Murray-Darling Basin. The outcomes of the assessments based on evaluation measures (Eqs. (9)–(16)) and diagnostic plots are as follows.

Performance evaluation with respect to the metrics, Pearson's correlation coefficient (*r*), root mean square error (*RMSE*) and mean absolute error (*MAE*) (Table 8a and b) showed cluttered outcomes. In SM_{UL} forecasting (Table 8a), the largest ***r*** was recorded from ELM at Sites 1 ($r = 0.999$) and 3 ($r = 0.992$), while ANN-CoM had highest values at Sites 2 ($r = 0.994$) and 4 ($r = 0.995$). At Site 1, the least *RMSE* was recorded by both ELM and ANN-CoM. Interestingly, at Site 3 both ELM and ANN-CoM registered $RMSE = 0.019$. In terms of *MAE*, the ELM had the best performance at Site 3, while ANN-CoM had the best performance at Site 2. At Sites 1 ($MAE = 0.005$) and 4 ($MAE = 0.011$), both ELM and ANN-CoM had similar performances. For SM_{LL} forecasting (Table 8b), all three measures unanimously exhibited superior performance of the hybrid ANN-CoM revealing that ANN-CoM has the best potential in generating accurate forecasts. Remarkably, at Sites 3 and 4, the ideal value of $r = 1.000$ was obtained by ANN-CoM. At Site 2, the ANN-CoM and random forest registered equal values of ***r***, *RMSE,* and *MAE*, while at Site 3 ELM had the same values as ANN-CoM. Although, it was certain that the new ANN-CoM potentially outperformed the other standalone models, the precise forecasting capability of ANN-CoM seemed rather concealed.

Next, the assessment in terms of Willmott's Index (*WI*), Nash-Sutcliffe Efficiency (E_{NS},) and the Legate-McCabe's Index (*L*) demonstrated vivid improvements in the performance of hybrid ANN-CoM model (Table 9a and b). Particularly, for SM_{UL} forecasts at Site 1, all three measures showed that ANN-CoM and ELM had identical performances. Considering *WI* only, at the other three sites (Sites 2, 3 and 4) ANN-CoM was the best. Similarly, the Nash-Sutcliffe Efficiency also showed better performance of ANN-CoM at Sites 1, 2 and 4. While capriciously at Site 3, ELM registered equal value of E_{NS} to that of ANN-CoM. From the perspective of Legate-McCabe's Index, which takes

Table 8
The performance ANN-CoM *vs.* the standalone models applied in forecasting: (a) upper layer (SM_{UL}) and (b) lower layer soil moisture (SM_{LL}) in the testing period, based on Pearson's correlation coefficient (r); root mean square error ($RMSE$) and mean absolute error (MAE). The optimal models yielding the lowest $RMSE$ at each site are shown in **boldface**.

Model	Performance Metrics	Upper Layer Soil Moisture (SM_{UL})				Lower Layer Soil Moisture (SM_{LL})			
		Site 1	Site 2	Site 3	Site 4	Site 1	Site 2	Site 3	Site 4
2nd order Volterra	r	0.985	0.923	0.884	0.948	0.917	0.956	0.975	0.968
	$RMSE$	0.021	0.042	0.075	0.059	0.034	0.012	0.037	0.053
	MAE	0.015	0.029	0.058	0.047	0.022	0.010	0.029	0.039
M5 Model Tree	r	0.985	0.985	0.970	0.967	0.997	0.998	0.998	0.998
	$RMSE$	0.023	0.020	0.037	0.041	0.006	0.002	0.008	0.012
	MAE	0.014	0.014	0.027	0.029	0.004	0.002	0.005	0.007
Random Forest	r	0.973	0.978	0.965	0.948	0.994	**0.999**	0.998	0.997
	$RMSE$	0.039	0.026	0.045	0.052	0.007	**0.001**	0.008	0.015
	MAE	0.023	0.019	0.034	0.035	0.006	**0.001**	0.004	0.007
Extreme Learning Machine (ELM)	r	**0.999**	0.993	**0.992**	0.994	0.998	0.998	**1.000**	0.998
	$RMSE$	**0.006**	0.013	**0.019**	0.017	0.004	0.002	**0.003**	0.010
	MAE	**0.005**	0.010	**0.015**	0.011	0.003	0.001	**0.002**	0.007
ANN-Committee of models (ANN-CoM)	r	0.998	**0.994**	**0.991**	**0.995**	**0.999**	**0.999**	**1.000**	**1.000**
	$RMSE$	0.007	**0.012**	**0.019**	**0.016**	**0.003**	**0.001**	**0.003**	**0.005**
	MAE	0.005	**0.009**	**0.014**	**0.011**	**0.003**	**0.001**	**0.002**	**0.003**

the precedence based on benefits discussed earlier, at all four sites ANN-CoM performs better. The percentage increase in L at Sites 2, 3 and 4 in comparison to best standalone model, i.e., ELM was 0.9% (Site 2), 0.46% (Site 3) and 0.55% (Site 4). Yet, ELM and ANN-CoM performed evenly at Site 1. The hybrid ANN-CoM model's precision was impeccable with very high predictor metric values ($WI \geq 0.992$, $E_{NS} \geq 0.982$ and $L \geq 0.878$), at all candidate sites. Similarly, in forecasting SM_{LL}, the measures, Willmott's Index (WI), Nash-Sutcliffe Efficiency (E_{NS},) and the Legate-McCabe's Index (L), (Table 9b) consistently revealed better performance of hybrid ANN-CoM at all four study sites. At Site 2, the performance of random forest was in par with ANN-CoM. In addition, the ELM (Site 3) and ANN-CoM (Sites 3 & 4) registered a $WI = 1.000$ indicating a perfect model fit, which practically is ambiguous due to the inherent drawbacks of this measure. With that, the preeminent indicator, Legate-McCabe's Index, was highest at all the four sites confirming the superior performance of ANN-CoM model. It must be noted that at all sites, the ANN-CoM model registered very high performance indicator values ($WI \geq 0.999$, $E_{NS} \geq 0.998$ and $L \geq 0.944$). Hence, with sufficient certainty, it can be seen that the hybrid ANN-CoM model has enhanced performance in forecasting both SM_{UL} and SM_{LL} values.

To further explore the suitability of ANN-CoM in SM forecasting, diagnostic plots were used to overcome the shortcomings of objective metrics. Fig. 5a and b show scatterplots of the observed and forecasted SM during the test period from all five models at all four candidate sites. For better exemplification, the linear fit equation and the coefficient of determination (R^2) [$Range = (0, +1)$; $Ideal\ value = +1$] which provides a measure on the global adequacy of the model (Hora and Campos, 2015) were included. In SM_{UL} forecasts, the plots clearly show that 2nd order Volterra, M5 tree and the random forest underperformed as the scatter points diverted from the $y = x$ linear form. Conversely, the ELM had very similar performance to the ANN-CoM model with comparable R^2 values. In congruence with the outcomes of Willmott's Index (WI), Nash–Sutcliffe Efficiency (E_{NS}) and Legates-McCabe's index (L), the scatterplots of Sites 1, 2 and 4 confirmed the superior performance of ANN-CoM. At Site 3 the R^2 of ANN-CoM was slightly lower than that of ELM, however, owing to the fact that Willmott's Index and Legates-McCabe's index were greater, the superiority of ANN-CoM is lucid. Likewise, in SM_{LL} forecasts, the ANN-CoM model outperformed at all sites registering larger values of R^2 (Fig. 5b). The minimum $R^2 = 0.998$ was recorded at Site 1 revealing that even at a worst-case scenario an overall 99.8% of the observed SM_{LL} values could be well simulated using the ANN-CoM model. The other values were close to unity ($R^2 = 0.999$). The gradient of the linear fit [$Ideal\ value = +1$] and the y-

Table 9
The performance ANN-CoM *vs.* the standalone models applied in forecasting: (a) upper layer (SM_{UL}) and (b) lower layer soil moisture (SM_{LL}) in the testing period based on Willmott's Index (WI); Nash–Sutcliffe Efficiency (E_{NS}) and Legates-McCabe's index (L). The models with the largest value of the Legates-McCabe's index at each site are in **boldface**.

Model	Performance Metrics	Upper Layer Soil Moisture (SM_{UL})				Lower Layer Soil Moisture (SM_{LL})			
		Site 1	Site 2	Site 3	Site 4	Site 1	Site 2	Site 3	Site 4
2nd order Volterra	WI	0.988	0.924	0.892	0.935	0.919	0.871	0.972	0.968
	E_{NS}	0.969	0.851	0.730	0.840	0.742	0.778	0.900	0.896
	L	0.851	0.685	0.501	0.586	0.510	0.544	0.666	0.664
M5 Tree	WI	0.983	0.984	0.967	0.961	0.997	0.996	0.999	0.998
	E_{NS}	0.963	0.967	0.934	0.923	0.992	0.993	0.996	0.995
	L	0.857	0.849	0.770	0.742	0.912	0.930	0.946	0.940
Random Forest	WI	0.937	0.972	0.948	0.932	0.996	**0.999**	0.999	0.997
	E_{NS}	0.891	0.943	0.904	0.875	0.988	**0.998**	0.996	0.992
	L	0.768	0.797	0.706	0.697	0.876	**0.966**	0.954	0.938
Extreme Learning Machine (ELM)	WI	**0.999**	0.994	0.993	0.994	0.999	0.998	1.000	0.999
	E_{NS}	**0.997**	0.986	0.982	0.987	0.996	0.995	0.999	0.996
	L	**0.953**	0.891	0.874	0.902	0.935	0.938	0.974	0.942
ANN-Committee of models (ANN-CoM)	WI	**0.999**	**0.995**	**0.992**	**0.995**	**0.999**	**0.999**	**1.000**	**1.000**
	E_{NS}	**0.997**	**0.987**	**0.982**	**0.989**	**0.998**	**0.998**	**0.999**	**0.999**
	L	**0.953**	**0.899**	**0.878**	**0.907**	**0.944**	**0.966**	**0.978**	**0.971**

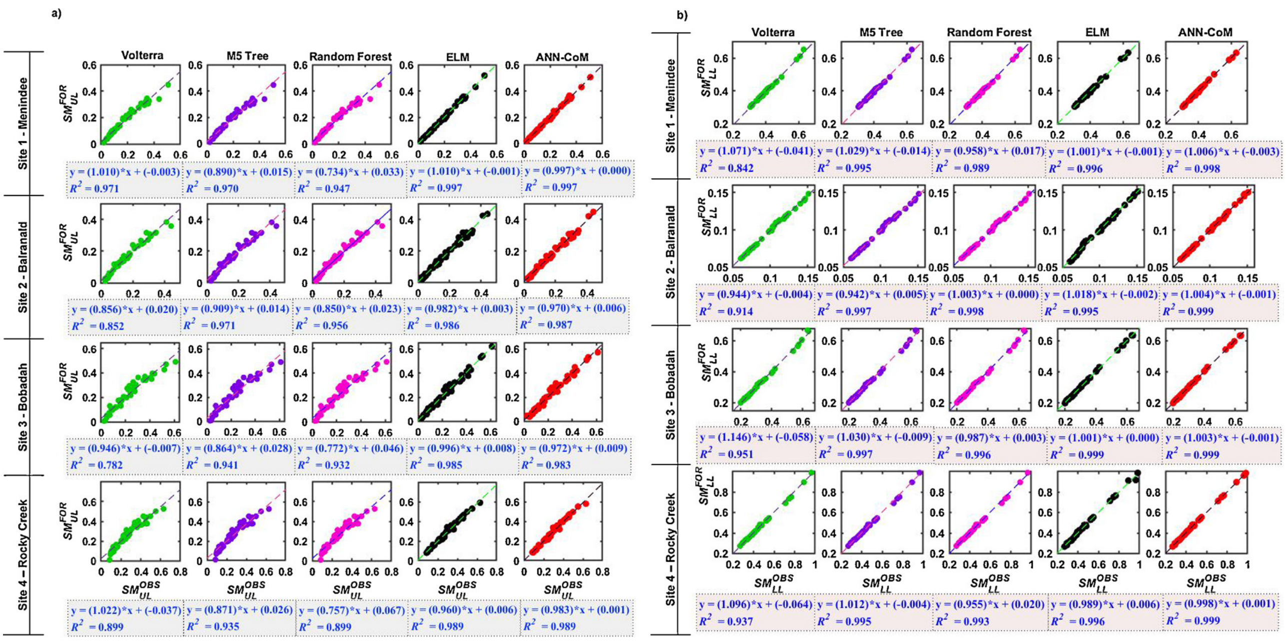


Fig. 5. Scatterplots of the observed (SM_{UL}^{OBS}) and the forecasted (SM_{UL}^{FOR}) soil moisture generated from ANN-CoM *vs.* the comparative standalone models applied at the candidate sites in the testing period (a) upper layer soil moisture (SM_{UL}) (b) lower layer soil moisture (SM_{LL}). Each panel shows a linear regression fit $y = mx + C$, and the coefficient of determination (R^2) denoting the goodness-of-fit. (SM_{UL} *and* SM_{LL} are the relative fractional values and the unit is dimensionless).

intercept [*Ideal value* = 0] for both SM_{UL} and SM_{LL} and at all sites were very close to idyllic magnitudes further reinforcing the outcomes of the scatterplots and the predictor metrics (Table 9a and b).

Additionally, the model evaluations were carried out via the box plots that illustrate the spread of the SM_{UL}^{OBS} and SM_{UL}^{FOR} values with respect to quartiles while the whiskers indicate the variability outside of the 1st and 3rd quartiles, (Fig. 6a and b). For forecasting SM_{UL}, the distributions of the observed (OBS) and the forecasted values from ANN-CoM were congruent, revealing its better performance. Following that was ELM, while the other three models had disparate distributions. The ANN-CoM better captured the high *SM* levels at Sites 1 and 2. In SM_{LL} forecasts, the ANN-CoM outperformed the standalone models in terms of forecast distributions and handled the high *SM* levels better. Hence, forecast distribution also demonstrated that the ANN-CoM has a predictive advantage in comparison to standalone models.

Moreover, the model preciseness was assessed using histograms of forecasting errors (FE) (Fig. 7a-b). FE is the difference between forecasted and observed *SM* during the test period and is computed as $\mathrm{FE} = SM_{UL}^{FOR} - SM_{UL}^{OBS}$; [*Ideal value* = 0]. Hence, a better model is bound to have higher occurrences of FE closer to zero. Four years of data (48 datum points) were used for testing the models, in error brackets of step-size 0.02 commencing from zero. For SM_{UL}, Volterra, M5 tree and random forest showed a larger degree of spread of forecasting errors in-between $-0.18 \le FE \le 0.18$ (Fig. 7a). Conversely, the ANN-CoM registered very small spreads in forecasting error that were closer to zero. At Sites 1 and 2 the inaccuracy ranged between $-0.04 \le \mathrm{FE} \le 0.04$, while at Sites 3 and 4 the range was $-0.06 \le \mathrm{FE} \le 0.06$. On the other hand, for forecasting SM_{LL} at all sites, the ANN-CoM's enhanced forecasting capability was clearly illustrated as all 48 datum points were in the first error bracket ($-0.02 \le \mathrm{FE} \le 0.02$) (Fig. 7b). Accordingly, the forecasting error histogram also confirmed the suitability of ANN-CoM as lower forecasting errors and improved accuracies are apparent.

Since the sites are having different geographical, physical and climatic characteristics (Fig. 3 and Tables 1 and 2), suitable relative measures (i.e., relative root mean square error (*RRMSE*) and mean-absolute percentage error (*MAPE*)) were alternatively used (Table 10a and b) to compare model performances at these sites. In forecasting SM_{UL}, the ANN-CoM model outperformed at two sites with percentage decrease in comparison to ELM (i.e., best standalone model) as follows *RRMSE*|*MAPE* Sites 2 − 3.64% |− 26.56% and 4 − 8.39% |− 0.25% (Table 10a). At the other two sites *i.e.*, Sites 1 and 3, ELM had least *RRMSE* and *MAPE* values. Based on least relative errors, the best model out of all sites was ELM at Site 1 *RRMSE* = 3.89%|*MAPE* = 3.05%. Yet, the correlation-based measures vividly showed that ANN-CoM outperformed at the majority of sites. On the other hand, the results of SM_{LL}, exhibited that the lowest values of both *RRMSE* and *MAPE* registered by the ANN-CoM model at all the four sites were apparently lower than those of standalone counterparts confirming that unarguably the ANN-CoM is the optimal choice (Table 10b). In comparison to the best standalone models, i.e., ELM (Sites 1, 3, & 4) and random forest (Site 2), the percentage decrease in *RRMSE* and *MAPE* values were as follows *RRMSE*|*MAPE*: Site 1 − 17.00% |− 10.00%; Site 2 − 3.77% |− 6.25%; Site 3 − 11.00% |− 16.88%; and Site 4 − 53.21% |− 44.30%. Overall, Site 1 *RRMSE* = 0.83%|*MAPE* = 0.63% had the optimal performance.

So far, the analysis of predictor metrics and various diagnostic plots have provided compelling evidence of the superiority of ANN-CoM model, in terms of the accuracy. Polar plots of monthly averages of absolute forecasting errors ($\overline{|\mathrm{FE}|}$) were then used for assessing the monthly performance of ANN-CoM model (Fig. 8). For brevity, two instances have been selected on the basis of the best (Site 1) and the worst (Site 4) performing ANN-CoM models in forecasting SM_{LL} based on least *RRMSE* (Table 10b). The polar plots revealed that there is a decrease in maximum monthly averages of absolute forecasting error values recorded by all models at the best-case scenario (Site 1) in comparison to the worst-case site (Fig. 8). The best ANN-CoM apparently has the least $\overline{|\mathrm{FE}|}$ in all the months in comparison to its standalone counterparts at Site 1. This reduction evidently suggests that the forecasts generated by the best ANN-CoM are more stable. The hybrid ANN-CoM proved to be the best with $\overline{|\mathrm{FE}|}$ very much closer to zero, in the order of 10^{-3} in all the months. The best ANN-CoM ensued minimum forecasting errors in January, which was consistent with other models at Site 1 while the maximum magnitude was recorded in February. In contrast, worst ANN-CoM model (Site 4) registered dissimilar performance outcomes with June recording the highest $\overline{|\mathrm{FE}|}$ value. For more

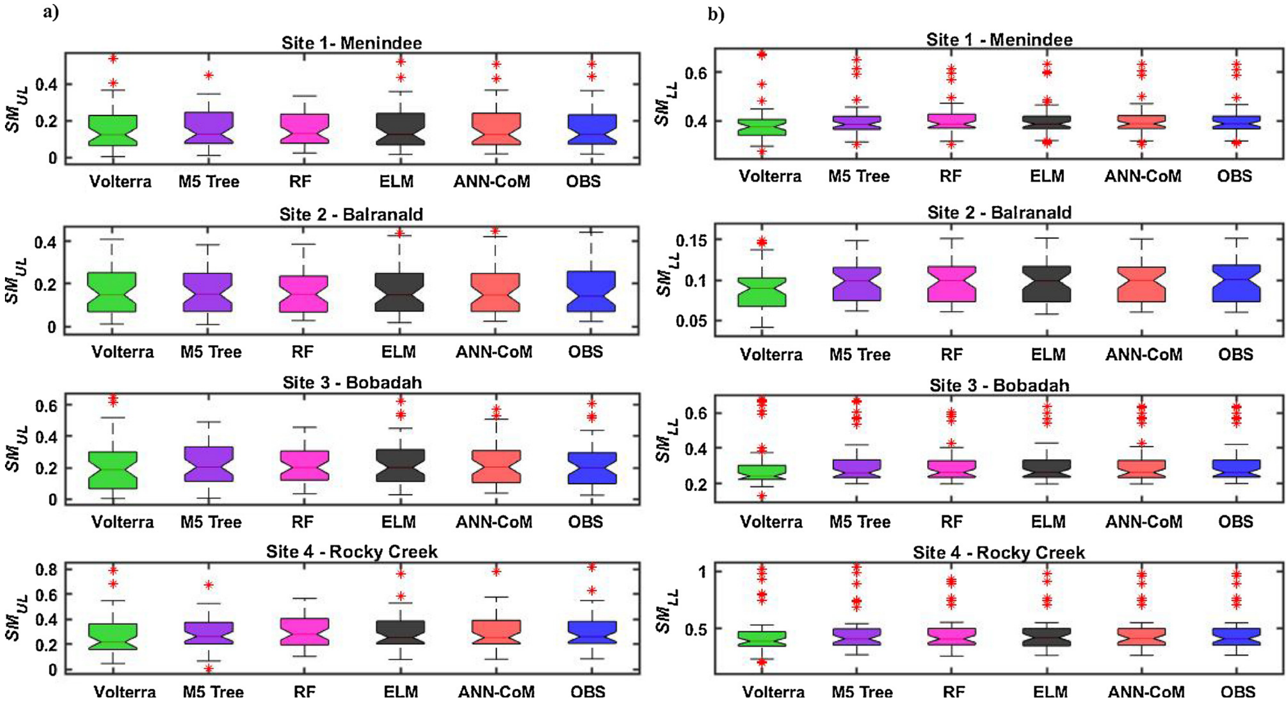


Fig. 6. Box plots of the observed (OBS) *vs.* the forecasted values of (a) the upper layer soil moisture (SM_{UL}) and (b) the lower layer soil moisture (SM_{LL}) generated from the ANN-CoM *vs.* the standalone data-driven models at the candidate study sites. (SM_{UL} *and* SM_{LL} are the relative fractional values and the unit is dimensionless).

perspicacity, the Taylor diagrams were used which illustrate a concise statistical summary of correlation, root-mean-square difference, and the ratio of the model's variances on a single figure (Taylor, 2001). The Taylor plot vividly asserted that in both the best (Fig. 9a) and worst (Fig. 9b) cases the ANN-CoM was better than all the standalone models while the 2nd order Volterra was the most underperforming one further

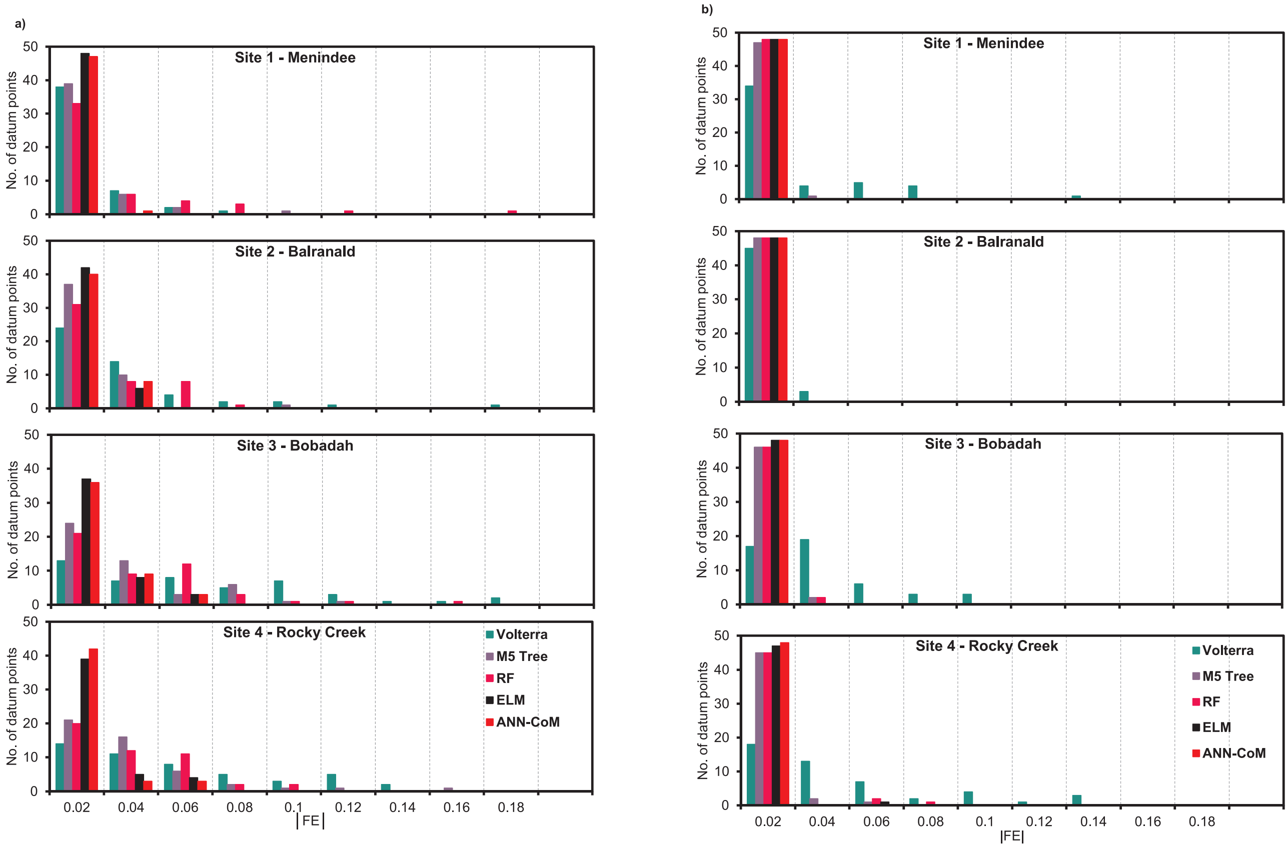


Fig. 7. Histograms illustrating the frequency (*i.e.*, no. of tested points) within each of the absolute forecasting errors (|FE|) generated from the ANN-CoM *vs.* the standalone models for (a) upper layer soil moisture (SM_{UL}), (b) lower layer soil moisture (SM_{LL}).

Table 10
A comparison at the different sites' performances using the relative error measures, relative root mean square error (*RRMSE*) and mean absolute percentage error (*MAPE*) for the forecasting of (a) upper layer soil moisture (SM_{UL}) and (b) lower layer soil moisture (SM_{LL}). The optimal model with lowest relative (%) error is shown in **boldface**.

Model	Performance Metrics (%)	Upper Layer Soil Moisture (SM_{UL})				Lower Layer Soil Moisture (SM_{LL})			
		Site 1	Site 2	Site 3	Site 4	Site 1	Site 2	Site 3	Site 4
2nd order Volterra	*RRMSE*	12.53	25.20	33.89	19.74	8.46	12.40	11.94	11.51
	MAPE	11.64	19.01	36.16	17.76	5.22	10.78	9.13	9.45
M5 Tree	*RRMSE*	13.72	11.87	16.69	13.69	1.44	2.22	2.43	2.50
	MAPE	9.58	10.63	15.64	11.03	0.93	1.44	1.37	1.38
Random Forest	*RRMSE*	23.51	15.52	20.25	17.43	1.83	1.06	2.50	3.25
	MAPE	12.45	12.69	21.98	11.84	1.31	0.80	1.01	1.55
ELM	*RRMSE*	**3.89**	7.69	**8.69**	5.72	1.00	1.89	1.00	2.18
	MAPE	**3.05**	8.66	**8.72**	3.94	0.70	1.74	0.77	1.49
ANN-CoM	*RRMSE*	4.10	**7.41**	8.76	**5.24**	**0.83**	**1.02**	**0.89**	**1.02**
	MAPE	3.15	**6.36**	9.85	**3.93**	**0.63**	**0.75**	**0.64**	**0.83**

establishing the superior forecasting capability of this new ANN-CoM model.

Finally, seasonal forecasting ability of ANN-CoM model was tested since seasonal accuracy is imperative for agricultural productivity and practical model deployment. This was achieved via bar graphs of the average seasonal relative root mean square errors (Fig. 10a and b). Overall, the *RRMSE* in forecasting SM_{LL}, were lower ($<$ 0.9) in comparison to SM_{UL} conveying that ANN-CoM can generate more precise SM_{LL} forecasts. The least *RRMSE* in forecasting SM_{UL} was recorded at Site 1 during spring (SON), while largest magnitude was recorded during summer (DJF) at Site 3 (Fig. 10a). An interesting feature was recorded during winter (JJA) as the *RRMSE* values were consistent across all four sites. In SM_{LL} forecasts the least relative error was registered at Site 3 during spring (SON), while the *RRMSE* were consistently higher during summer (DJF) throughout all sites (Fig. 10b). Henceforth, the ANN-CoM has better performance in winter in SM_{UL} forecasts. Conversely, for SM_{LL} forecasts the models may generate larger uncertainties during summer, yet these uncertainties would be lower than those of SM_{UL} forecasts.

Owing to the fact that, the single data-intelligent models have dissimilar theoretical and mathematical underpinnings, these algorithms capture the predictive features differently. For instance, the 2nd order Volterra model captures the memory effects on the basis of Taylor series while the M5 model tree utilizes smoothed localized regression trees. Although the random forest has regression trees as basic learners, it incorporates a bootstrapped aggregated ensemble approach. The final expert model, ELM employs non-linear elements, i.e., neurons, for feature extraction. These aforementioned differences in modeling structure may lead to some algorithms skipping or overlooking vital features (e.g., seasonality, peaks or extreme *SM* levels) and are apparent from the differences in predictive performances of the standalone models whereby ELM has a better performance and 2nd order Volterra having the worst.

However, this multi-model ensemble committee of models approach based on ANN (ANN-CoM) is able to overcome this by determining a collective forecast. The scaled performance of ANN-CoM (Tables 8 and 9, Figs. 5–9) ascertains that this ensemble committee approach is able to harness the predictive features that otherwise would have been left out in the standalone modeling method. Being nonlinear, the ANN is able to use the internal interconnected multiple neurons and iterative adjustment of feature weights to further improve the forecasts devoid of being constrained to a specific form. Without any assumption of probability distribution like normality or equal dispersion and covariance matrix requirements (Moghaddamnia et al., 2009), the ANN-CoM effectively simulates the stochastic and complex hydrological system of the *SM*. The liberty to select the number of hidden layers and the associated nodes in each of these layers provides ANN with added versatility and robustness (ASCE Task Committee on Application of ANN in Hydrology, 2000; Yilmaz et al., 2011). The best ANN-CoM (Site 1) had a neuronal architecture of (4-24-1: Input-Hidden-Output) with '*trainlm*' as the learning algorithm. In addition, the lower average forecasting errors suggest that the model combination stabilizes the forecasts. Above all, the purpose of combining models is not only to improve accuracy but also to protect against the failure of the individual expert models (Baker and Ellison, 2008), which is very important for real-life applications.

This neural network based multi-model ensemble could possibly be utilized as a forecasting tool for farmers, farm managers, and other decision makers. ANN's capability to operate in the non-stationary environment and subsequently adapt to minor changes in the surrounding makes it more suitable for field deployment as an adaptive system is more likely to remain stable producing a robust performance (Haykin, 1999). In addition, failure in the hardware implementation of ANNs shows a gradual degradation instead of an abrupt malfunction (Haykin, 1999). This further makes the ANN-CoM inherently fault tolerant and well suited for field implementation.

Prior to model development, appropriate feature selection is important to increase forecasting accuracy (Prasad et al., 2017). Two-fold feature selection technique; the Neighborhood Component Analysis feature selection for regression algorithm (*fsrnca*) derived feature weights followed by the innovative modeled minimum *RRMSE* criteria, was pivotal in selecting appropriate inputs in the development of parsimonious models. Fundamentally, the Australian climate is dependent upon many factors, including, rainfall (Abawi et al., 2000; Ummenhofer et al., 2009); Indian Ocean Dipole (IOD) (i.e., the low-frequency coupled ocean-atmosphere variability in the Indian Ocean) (Ashok et al., 2003; Ummenhofer et al., 2009); El Nino Southern Oscillation (ENSO) (Deo et al., 2009); Southern Oscillation Index (SOI) and sea surface temperatures (SST) (Abawi et al., 2000). Correspondingly, sixty predictor input variables were agglutinated. Despite this, feature selection revealed that soil evaporation (FWsoil) and latent heat flux (PhiE) are most important ones for SM_{UL} forecasts while the deep drainage (FWLch2) was important for SM_{LL} forecasting.

Moreover, a contesting platform for the model evaluations was provided by the four sites with distinctive geophysical, topographical and climatological conditions. Besides ANN-CoM generating enhanced *SM* forecasts, it was noted that the SM_{LL} ($0.944 \leq L \leq 0.978$) were better forecasted than SM_{UL} ($0.878 \leq L \leq 0.953$). This difference could have resulted due to different climatological patterns of SM_{UL} and SM_{LL} (Fig. 4a-b). The SM_{UL} (surface layer) showed greater seasonal variability as it is contingent upon surface meteorological variations, vegetation types (trees, crops, grass, or fallow) (Ladson et al., 2004) and the deep soil hydraulics. On the contrary, the SM_{LL} largely depends on deep percolation, groundwater recharge, and plant uptake that are relatively steady across the seasons. In spite of this geophysical disparity amongst sites and vertical soil inconsistencies, the ensemble approach ANN-CoM captured and simulated pertinent physical patterns as is apparent from

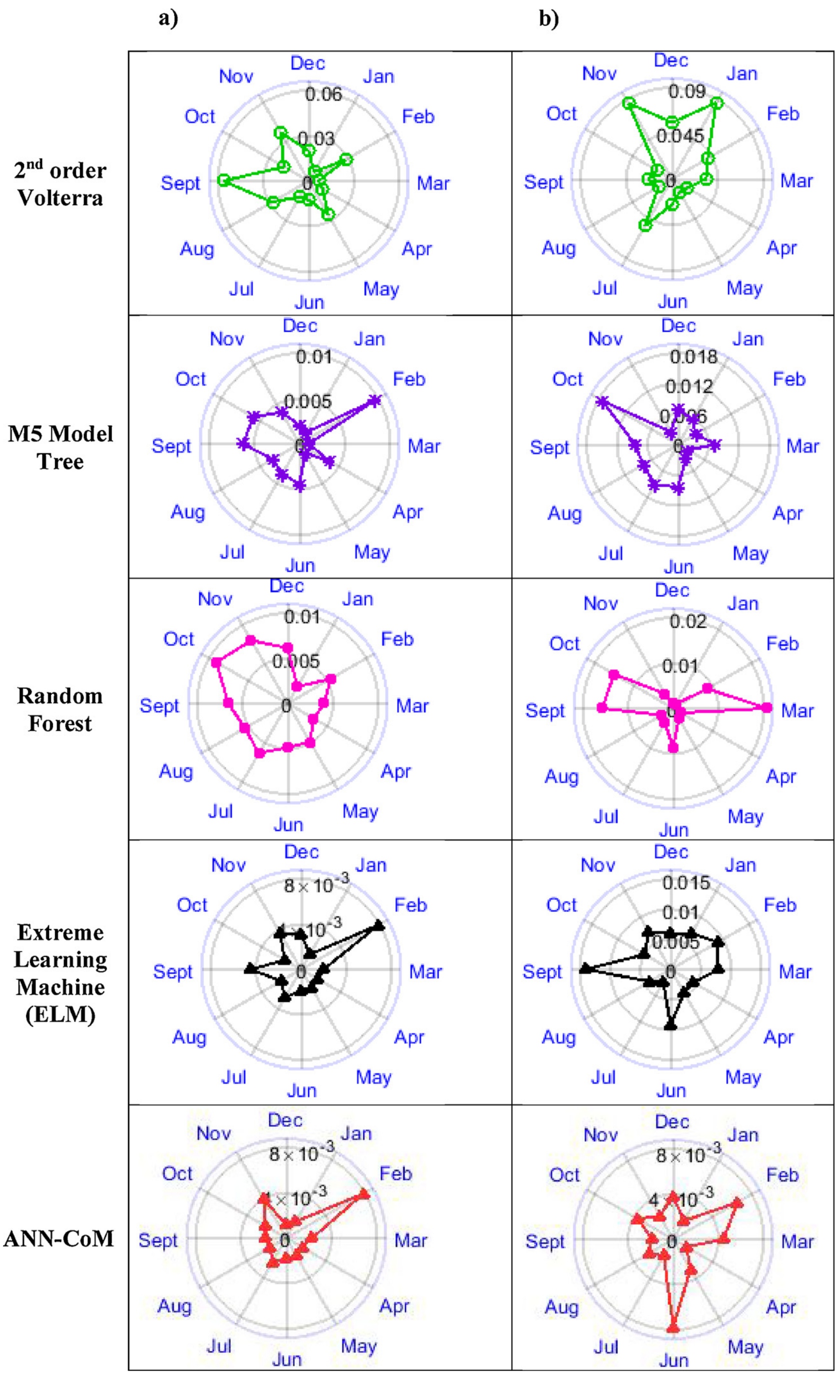


Fig. 8. Polar plots showing the monthly average values of the absolute forecasting error generated from the ANN-CoM vs. the standalone models in forecasting lower layer soil moisture with: a) best ANN-CoM model (Site 1) and b) worst ANN-CoM model (Site 4) based on *RRMSE* values. (*SM_{UL} and SM_{LL}* are the relative fractional values and the unit is dimensionless).

the enhanced performance. However, the context of the present study was limited to time-series forecasting via data-driven ANN based ensemble committee of modelling approach. To gain a better understanding on the capacity ANN based machine-learning techniques into real-life decision support systems, further independent studies with respect to other physical models (e.g., HYDRUS-1D, MACRO, VS2DTI, etc.) need to be performed.

4. Conclusion

In this paper, a new multi-model ensemble committee framework with ANN as the final prediction tool (i.e., ANN-CoM model) was developed and evaluated for forecasting the SM_{UL} (0–0.2 m) and SM_{LL} (0.2–1.5 m). Four study sites within Murray-Darling basin region; Site 1-Menindee, Site 2-Balranald, Site 3-Bobadah, and Site 4-Rocky Creek were selected to assess the model performances. The Neighborhood Component Analysis based feature selection algorithm, *fsrnca*, and a basic ELM (h_n = 50; *'sigmoid'* transfer function) determined the optimal set of predictors from 60 predictor inputs variables.

A holistic evaluation via statistical metrics and diagnostic plots revealed that the ANN-CoM generated superior forecasts in comparison to benchmark standalone models (viz., 2nd order Volterra, M5 model tree, random forest, and extreme learning machine). The site comparisons showed that the ANN-CoM model had the best performance at Site 1

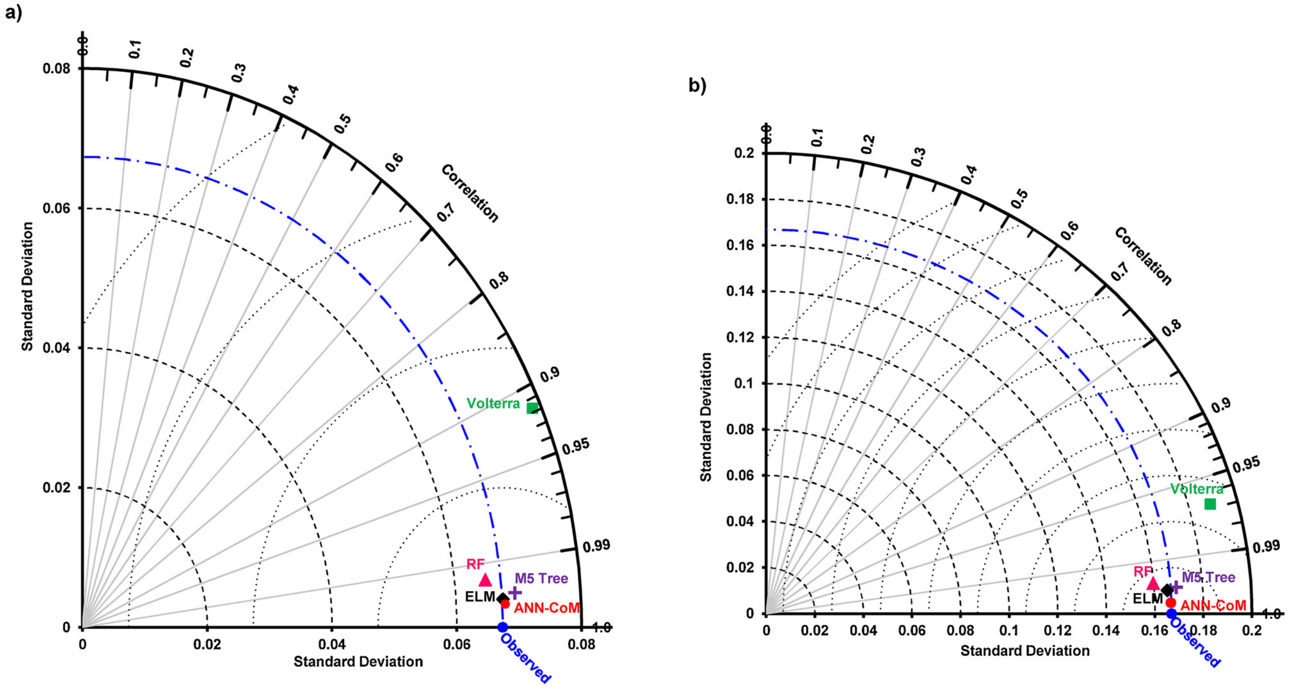


Fig. 9. Taylor plots showing the correlation and standard deviation (SD) of the ANN-CoM *vs.* the standalone models in forecasting SM_{LL} with a) best ANN-CoM model (Site 1) and b) worst ANN-CoM model (Site 4) based on *RRMSE* values. (SM_{UL} *and* SM_{LL} are the relative fractional values and the unit is dimensionless).

$RRMSE = 0.83\%|MAPE = 0.63\%$ in forecasting SM_{LL}. Seasonally, the hybrid ANN-CoM generated better SM_{LL} forecasts than SM_{UL} forecasts.

The findings of this study ascertain that with appropriate input selection (such as *fsrnca* feature weights and the minimum *RRMSE* criteria), the two-stage multi-model ensemble committee based ANN (ANN-CoM) indeed effectively captured the nonlinear dynamics and interactions amongst the input data and SM_{UL} and SM_{LL} in generating optimally combined and stabilized forecasts. The ANN-CoM model is a feasible alternative for SM_{UL} and SM_{LL} forecast implementations in terms of determining the future trends in *SM* levels and could be explored as a data intelligent tool for hydrological and agricultural recourses. In accordance with the performance measures, while the ANN-CoM model was successful in simulating the upper and lower layer soil moisture, there are limitations of the approach that can be addressed in independent follow-up studies.

In this paper, the forecasting horizon was restricted to the monthly scale but for real-time applications such as in the day-to-day agricultural and farming decisions, one also needs to emulate the soil moisture over a much shorter and a practically realistic timescale (e.g., weekly, hourly or sub-hourly). To address this, the application of a set of physically simulated meteorological outputs from the nationally adapted Australian Community Climate and Earth-System Simulator (ACCESS) model and further integrating the data fields into a data intelligent model within a multi-model ensemble committee approach can lead to a new paradigm for real-time soil moisture forecasting. This has an imperative advantage over the present approach as the ACCESS is based on the UK Meteorological Office's Unified Model with a grid resolution of 0.11° (~12 km) and a temporal resolution of 3-hourly up to a lead time of 3-days. Therefore, in a follow-up study, one could utilize ACCESS model simulated meteorological fields coupled with measured soil moisture data in various farming locations to further reprocessed them with the ANN-CoM approach to generate soil moisture estimates for local (e.g., farming) applications.

Another limitation was that the ANN-CoM approach has used a single hidden layer neuronal system optimized by trial and error. To improve the ensemble committee model and to test its ability to be

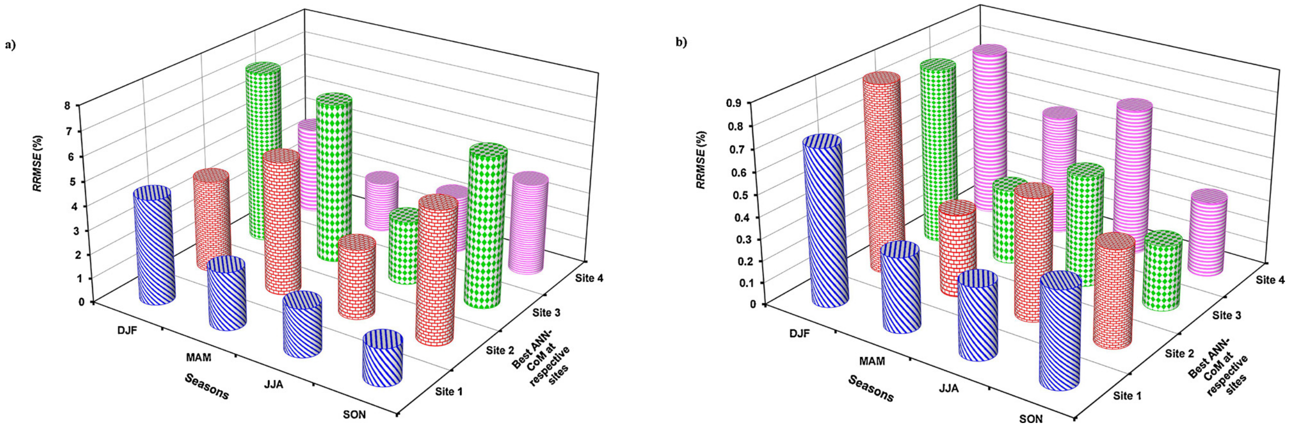


Fig. 10. Bar graphs of the average seasonal relative root mean square errors (*RRMSE*) in forecasting a) upper layer (SM_{UL}) and b) lower layer (SM_{LL}) soil moisture from the best ANN-CoM model at the four candidate sites. The seasons are Summer-DJF; Autumn-MAM; Winter-JJA; Spring-SON. (SM_{UL} *and* SM_{LL} are the relative fractional values and the unit is dimensionless).

embraced as a short-term, real-time prediction tool, a follow-up study could explore algorithms that are more advanced, such a deep learning-based Long Short-Term Memory (LSTM) model where multiple hidden layers in neuronal systems can be incorporated for robust feature extraction. LSTM is becoming a popular tool for prediction where the role of antecedent features are significant estimate a target variable. Deep learning model has a greater ability to capture data patterns and is effective in analyzing the data features related to a given target (e.g., soil moisture). While the use of LSTM was out of the scope of this study, a model based on deep learning can also be explored in digital systems such as mobile phone apps and hand-held devices to provide estimates of soil moisture over a much shorter time-scale than explored in this research paper.

Acknowledgements

The authors are grateful to Australian Water Availability Project (AWAP) and Interim ERA European Centre for Medium-Range Weather Forecasting (ECMWF) for providing the relevant data for this study. This study was supported by University of Southern Queensland Office of Research and Graduate Studies (ORGS) Postgraduate Research Scholarship (USQ-PRS) Grant # 2016 awarded to the first author. The authors sincerely acknowledge all expert reviewers whose comments have improved the paper.

References

Abawi, G., Dutta, S., Harris, T., Ritchie, J., Rattray, D., Crane, A., 2000. The Use of Seasonal Climate Forecasts in Water Resources Management, Hydro 2000: Interactive Hydrology. Barton, A.C.T.: Institution of Engineers, Australia, pp. 447–455.

ASCE Task Committee on Application of ANN in Hydrology, 2000. Artificial neural networks in Hydrology-I-Preliminary concepts. J. Hydrol. Eng. 5, 115–123.

Ashok, K., Guan, Z., Yamagata, T., 2003. Influence of the Indian Ocean Dipole on the Australian winter rainfall. Geophys. Res. Lett. 30.

ASRIS, 2014. Australian Soil Resource Information System. Department of Agriculture, Fisheries and Forestry.

Australian Bureau of Statistics, 2008. Water and the Murray-Darling Basin - A Statistical Profile, 2000-01 to 2005-06.

Australian Bureau of Statistics, 2017. Agricultural census Fact Sheet: Australia, States and Territories. ABS, Canberra, Australia.

Baker, L., Ellison, D., 2008. The wisdom of crowds– ensembles and modules in environmental modelling. Geoderma 147, 1–7.

Barzegar, R., Moghaddam, A.A., 2016. Combining the advantages of neural networks using the concept of committee machine in the groundwater salinity prediction. Model. Earth Systems Environ. 2.

Barzegar, R., Moghaddam, A.A., Baghban, H., 2015. A supervised committee machine artificial intelligent for improving DRASTIC method to assess groundwater contamination risk: a case study from Tabriz plain aquifer, Iran. Stoch. Environ. Res. Risk Assess. 30, 883–899.

Barzegar, R., Moghaddam, A.A., Deo, R., Fijani, E., Tziritis, E., 2017. Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms. Sci. Total Environ. 621, 697–712.

Beesley, C.A., Frost, A.J., Zajaczkowski, J., 2009. A comparison of the BAWAP and SILO spatially interpolated daily rainfall datasets Cairns, Australia. 18th World IMACS / MODSIM Congress.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environ. Model. Softw. 40, 1–20.

Bhattacharya, B., Solomatine, D.P., 2005. Neural networks and M5 model trees in modelling water level–discharge relationship. Neurocomputing 63, 381–396.

Billings, S.A., Korenberg, M.J., Chen, S., 1988. Identification of non-linear output-affine systems using an orthogonal least-squares algorithm. Int. J. Syst. Sci. 19, 1559–1568.

Bowden, G.J., Dandy, G.C., Maier, H.R., 2005. Input determination for neural network models in water resources applications. Part 1—background and methodology. J. Hydrol. 301, 75–92.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Brocca, L., Ciabatta, L., Massari, C., Camici, S., Tarpanelli, A., 2017. Soil moisture for hydrological applications: open questions and new opportunities. Water 9, 140.

Cai, W., Cowan, T., Briggs, P., Raupach, M., 2009. Rising temperature depletes soil moisture and exacerbates severe drought conditions across southeast Australia. Geophys. Res. Lett. 36.

Campbell, R., Scarlett, A., 2014. Economics, Agriculture and Native Vegetation in NSW. The Australia Institute.

Chen, C.-H., Lin, Z.-S., 2006. A committee machine with empirical formulas for permeability prediction. Comput. Geosci. 32, 485–496.

DeCarlo, L.T., 1997. On the meaning and use of kurtosis. Psychol. Methods 2, 292–307.

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.N., Vitart, F., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137, 553–597.

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, 90.

Deo, R.C., Syktus, J.I., McAlpine, C.A., Lawrence, P.J., McGowan, H.A., 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.

Deo, R.C., Tiwari, M.K., Adamowski, J.F., Quilty, J.M., 2017a. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. Stoch. Environ. Res. Risk Assess. 31 (5), 1211–1240.

Deo, R.C., Kisi, O., Singh, V.P., 2017b. Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model. Atmos. Res. 184, 149–175.

Department of Agriculture and Water Resources, 2015. Catchment Scale Land Use of Australia. Agricultural Land Management, Australia.

Dharssi, I., Steinle, P., 2011. Assimilation of satellite derived soil moisture for weather forecasting Monash University. SMOS/SMAP Workshop.

Du, Y., Ulaby, F.T., Dobson, M.C., 2000. Sensitivity to soil moisture by active and passive microwave sensors. IEEE Trans. Geosci. Remote Sens. 105–114.

Esmaeili, S., Asghari Moghaddam, A., Barzegar, R., Tziritis, E., 2018. Multivariate statistics and hydrogeochemical modeling for source identification of major elements and heavy metals in the groundwater of Qareh-Ziaeddin plain, NW Iran. Arabian J. Geoscie. 11.

Fausett, L., 1994. Fundamentals Of Neural Networks. Prentice Hall, New York.

Ghorbani, M.A., Shamshirband, S., Zare Haghi, D., Azani, A., Bonakdari, H., Ebtehaj, I., 2017. Application of firefly algorithm-based support vector machines for prediction of field capacity and permanent wilting point. Soil Tillage Res. 172, 32–38.

Grayson, R.B., Western, A.W., 1998. Towards areal estimation of soil water content from point measurements- time and space stability of mean response. J. Hydrol. 207, 68–82.

Hatampour, A., 2013. Developing a committee machine model for predicting Reservoir porosity from image analysis of thin sections. Middle-East J. Sci. Res. 13, 1438–1444.

Haykin, S., 1999. Neural Networks - A Comprehensive Foundation, 2nd ed. Pearson Education (Singapore) Pte. Ltd. Indian Branch, Delhi, India.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25, 1965–1978.

Hora, J., Campos, P., 2015. A review of performance criteria to validate simulation models. Expert Syst. 32, 578–595.

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2004. Extreme learning machine: a new learning scheme of feedforward neural networks. 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH41). pp. 985–990.

Huang, C., Li, L., Ren, S., Zhou, Z., 2010. Research of soil moisture content forecast model based on genetic algorithm BP neural network Nanchang, China. 4th Conference on Computer and Computing Technologies in Agriculture (CCTA). Springer, IFIP Advances in Information and Communication Technology, AICT-345 309–316.

Hyndman, R.J., Koehler, A.B., 2006. Another look at measures of forecast accuracy. Int. J. Forecast. 22, 679–688.

Jain, A., Kumar, A.M., 2007. Hybrid neural network models for hydrologic time series forecasting. Appl. Soft Comput. 7, 585–592.

Jain, S.K., Das, A., Srivastava, D.K., 1999. Application of ANN for reservoir inflow prediction and operation. J. Water Resour. Plann. Manage. 125, 263–271.

Jekabsons, G., 2010. M5PrimeLab: M5' Regression Tree and Model Tree Toolbox for Matlab/Octave.

Karunanithi, N., Grenney, W.J., Whitley, D., Bovee, K., 1994. Neural networks for river flow prediction. J. Comput. Civil Eng. 8, 201–220.

Labat, D., Ababou, R., Mangin, A., 1999. Linear and nonlinear input/output models for karstic springflow and flood prediction at different time scales. Stoch. Environ. Res. Risk Assess. 13, 337–364.

Ladson, T., Lander, J., Western, A., Grayson, R., 2004. Estimating Extractable Soil Moisture Content for Australian Soils, In: Cooperative Research Centre for Catchment Hydrology (Ed.), Technical Report.

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, 233–241.

Legates, D.R., McCabe, G.J., 2013. A refined index of model performance: a rejoinder. Int. J. Climatol. 33, 1053–1056.

Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22.

Liu, Y., Mei, L., Ooe, S.K., 2014. Prediction of soil moisture based on extreme learning machine for an apple orchard. Conference on Computational Interdisciplinary Science. IEEE. pp. 400–404.

Maheswaran, R., Khosa, R., 2012. Wavelet–Volterra coupled model for monthly stream flow forecasting. J. Hydrol. 450-451, 320–335.

Maheswaran, R., Khosa, R., 2015. Wavelet Volterra coupled models for forecasting of nonlinear and non-stationary time series. Neurocomputing 149, 1074–1084.

Maier, H.R., Jain, A., Dandy, G.C., 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. Environ. Model. Softw. 25, 891–909.

Matei, O., Rusu, T., Petrovan, A., Mihuţ, G., 2017. A data mining system for Real time soil moisture prediction. Procedia Eng. 181, 837–844.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5, 115–133.

Merdun, H., Çınar, Ö., Meral, R., Apan, M., 2006. Comparison of artificial neural network and regression pedotransfer functions for prediction of soil water retention and saturated hydraulic conductivity. Soil Tillage Res. 90, 108–116.

Moghaddamnia, A., Ghafari Gousheh, M., Piri, J., Amin, S., Han, D., 2009. Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques. Adv. Water Resour. 32, 88–97.

Munro, R.K., Lyons, W.F., Shaob, Y., Wood, M.S., Hooda, L.M., Leslie, L.M., 1998. Modelling land surface–atmosphere interactions over the Australian continent with an emphasis on the role of soil moisture. Environ. Model. Softw. 13, 333–339.

Myers, W., Linden, S., Wiener, G., 2009. A data mining approach to soil temperature and moisture prediction Arizona, USA. Seventh Conference on Artificial Intelligence and Its Applications to the Environmental Sciences. American Meteorological Society.

Petropoulos, G.P., 2014. Remote Sensing of Energy Fluxes and Soil Moisture Content. CRC Press, Taylor & Francis Group, Boca Raton, FL.

Prasad, R., Deo, R.C., Li, Y., Maraseni, T., 2017. Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm. Atmos. Res. 197, 42–63.

Quinlan, J.R., 1992. Learning with continuous classes. In: Sterling, A. (Ed.), 5th Australian Joint Conference on Artificial Intelligence. Singapore. pp. 343–348.

Rathinasamy, M., Adamowski, J., Khosa, R., 2013. Multiscale streamflow forecasting using a new bayesian model average based ensemble multi-wavelet Volterra nonlinear method. J. Hydrol. 507, 186–200.

Raupach, M.R., Briggs, P.R., Haverd, V., King, E.A., Paget, M., Trudinger, C.M., 2009. Australian Water Availability Project (AWAP)-CSIRO Marine and Atmospheric Research Component-Final Report for Phase 3.

Raupach, M.R., Briggs, P.R., Haverd, V., King, E.A., Paget, M., Trudinger, C.M., 2012. Australian Water Availability Project. CSIRO Marine and Atmospheric Research, Canberra, Australia. http://www.csiro.au/awap.

Samadianfard, S., Asadi, E., Jarhan, S., Kazemi, H., Kheshtgar, S., Kisi, O., Sajjadi, S., Manaf, A.A., 2018. Wavelet neural networks and gene expression programming models to predict short-term soil temperature at different depths. Soil Tillage Res. 175, 37–50.

Schaap, M.G., Leij, F.J., 1998. Using neural networks to predict soil water retention and soil hydraulic conductivity. Soil Tillage Res. 47, 37–42.

Sehgal, V., Tiwari, M.K., Chatterjee, C., 2014. Wavelet bootstrap multiple linear regression based hybrid modeling for daily River discharge forecasting. Water Resour. Manage. 28, 2793–2811.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. J. Geophys. Res. Atmos. 106, 7183–7192.

Tian, S., Tregoning, P., Renzullo, L.J., Dijk, A.I.J.Mv., Walker, J.P., Pauwels, V.R.N., Allgeyer, S., 2017. Improved water balance component estimates through joint assimilation of GRACE water storage and SMOS soil moisture retrievals. Water Resour. Res. 53, 1820–1840.

Timbal, B., Abbs, D., Bhend, J., Chiew, F., Church, J., Ekström, M., Kirono, D., Lenton, A., Lucas, C., McInnes, K., Moise, A., Monselesan, D., Mpelasoka, F., Webb, L., Whetton, P., 2015. Murray Basin Cluster Report: Climate Change in Australia Projections for Australia's Natural Resource Management Regions: Cluster Reports, eds. Ekström, M. et al., In: Ekström, M., Whetton, P., Gerbing, C., Grose, M., Webb, L., Risbey, J. (Eds.), CSIRO and Bureau of Meteorology, Australia.

Tiwari, M.K., Adamowski, J., 2013. Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models. Water Resour. Res. 49, 6486–6507.

Tozer, C.R., Kiem, A.S., Verdon-Kidd, D.C., 2012. On the uncertainties associated with using gridded rainfall data as a proxy for observed. Hydrol. Earth Syst. Sci. 16, 1481–1499.

Ummenhofer, C.C., England, M.H., McIntosh, P.C., Meyers, G.A., Pook, M.J., Risbey, J.S., Gupta, A.S., Taschetto, A.S., 2009. What causes southeast Australia's worst droughts? Geophys. Res. Lett. 36.

Walker, J.P., Ursino, N., Grayson, R.B., Houser, P.R., 2003. Australian Root Zone Soil Moisture-Assimilation of Remote Sensing Observations, International Congress on Modelling and Simulation (MODSIM). Modelling and Simulation Society of Australia and New Zealand, Inc., Townsville, Australia, pp. 380–385.

Wei, H.L., Billings, S.A., 2004. A unified wavelet-based modelling framework for nonlinear system identification: the WANARX model structure. Int. J. Control 77, 351–366.

Willmott, C.J., 1984. On the evaluation of model performance in physical geography. In: Gaile, G.L., Willmott, C.J. (Eds.), Spatial Statistics and Models. Springer, pp. 443–460.

Witten, I.H., Frank, E., Hall, M.A., 2011. Data Mining - Practical Machine Learning Tools and Techniques, 3 ed. Morgan Kaufmann Publishers, United States.

Yang, W., Wang, K., Zuo, W., 2012. Neighborhood component feature selection for high-dimensional data. J. Comput. 7.

Yang, X., Zhang, C., Cheng, Q., Zhang, H., Gong, W., 2017. A hybrid model for soil moisture prediction by using artificial neural networks. Revista de la Facultad de Ingeniería U.C.V 32, 265–271.

Yaseen, Z.M., Jaafar, O., Deo, R.C., Kisi, O., Adamowski, J., Quilty, J., El-Shafie, A., 2016. Stream-flow forecasting using extreme learning machines: a case study in a semi-arid region in Iraq. J. Hydrol. 542, 603–614.

Yilmaz, A.G., Imteaz, M.A., Jenkins, G., 2011. Catchment flow estimation using artificial neural networks in the mountainous euphrates Basin. J. Hydrol. 410, 134–140.

Zhang, G., Patuwo, B.E., Hu, M.Y., 1998. Forecasting with artificial neural networks: the state of the art. Int. J. Forecast. 14, 35–62.

Supplementary analysis and discussions

The performance of the newly developed hybrid model was diagnosed via a scatterplot with a corresponding linear regression line and the X = Y line. Figure S3 a-b displays the scatter plots of committee of models based on artificial neural networks (ANN-CoM) and the standalone models (including Volterra, M5 Tree, random forest and the extreme learning machine - ELM) in forecasting the upper layer (SM_{UL}) and lower layer soil moisture (SM_{LL}) at four study sites. Comparing the regression fitting line and the 1:1 line, it is clear that three standalone models *viz.* Volterra, M5 Tree, and the random forest did not perform well as the regression line spurned from the 1:1 line in forecasting SM_{UL} (Figure S3a). The ELM performed comparatively well, yet the regression line and the X = Y lines were almost exactly on top of each other for the forecasted values generated from the ANN-CoM model for both the SM_{UL} (Figure S3a). In forecasting SM_{LL} (Figure S3b) the standalone models did perform quite well as the respective regression lines were in close agreement with the X = Y lines. However, the hybrid ANN-CoM outperformed since again the regression lines were on top of the 1:1 lines. The outcomes further ascertain the discussion presented in the main chapter.

With a 5% tolerance limit, the percentage deviations from the 1:1 line for all models at all sites were conducted in forecasting both the upper and lower layer soil moisture values. The full data set on percentage deviations are provided in the Appendix (Table A3). Summarizing the data on under and over-predictions (Table S3 a-b) showed that in forecasting the SM_{UL} (Table S3a), the performance of ELM was in close contention with the ANN-CoM with both ELM and ANN-CoM having the same total number of over/under predictions at Sites 2 (*i.e.*, 26 points) and 3 (*i.e.*, 28 points). At Site 1, the ANN-CoM had an extra point in the total in comparison to ELM, while at Site 4, the ANN-CoM has the least total number of over/under predicted data (*i.e.*, 15 points). For the case of lower layer soil moisture forecasts, the ANN-CoM showed no over/under predictions with respect to the 5% tolerance limit applied in this study. The supplementary result also reinforces the previous discussion on the suitability of the hybrid ANN-CoM for SM_{UL} and SM_{LL} forecasting.

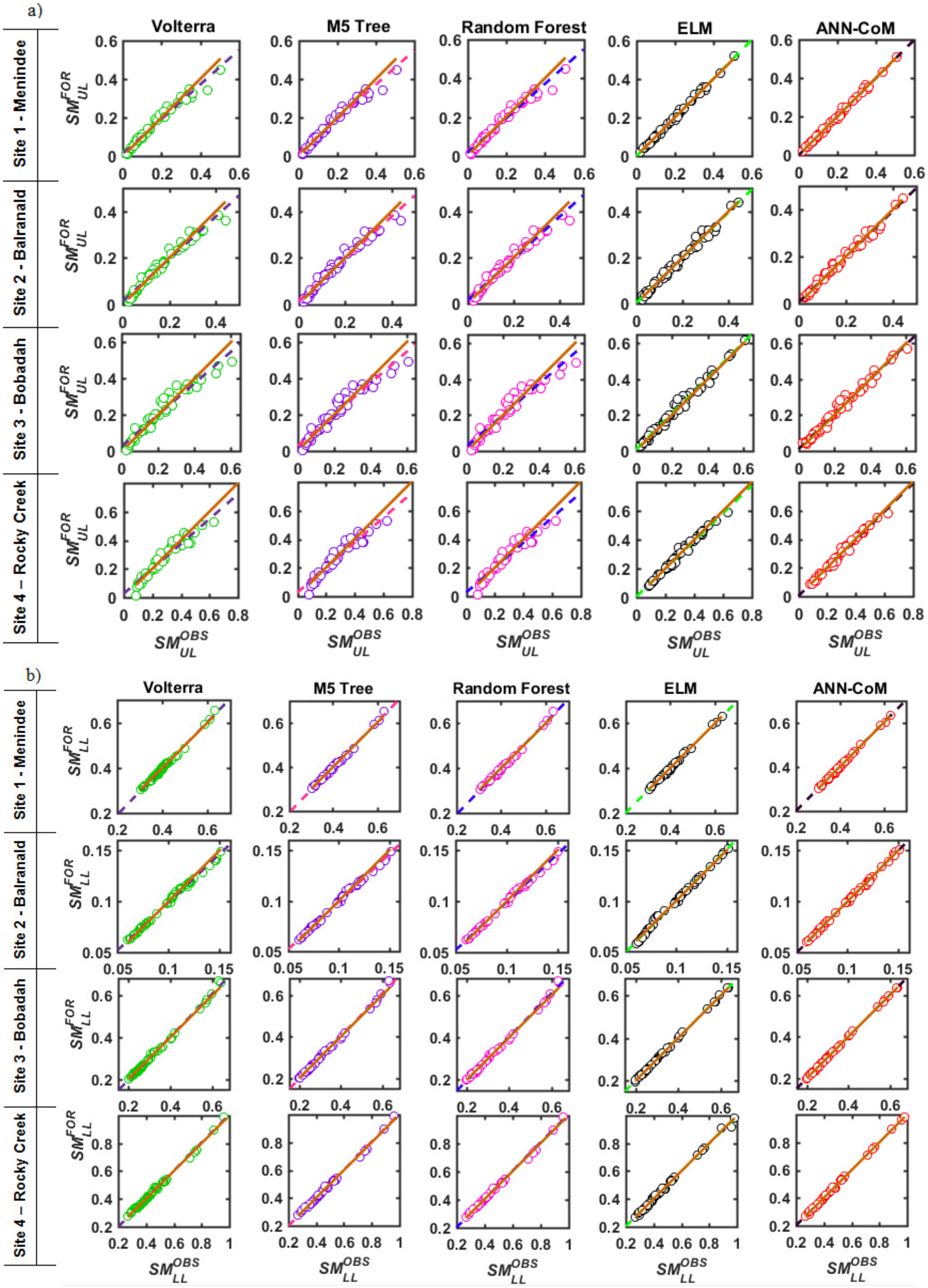


Figure S3 Scatter plots of observed and forecasted values registered by the ANN-CoM and the extreme learning machine (ELM), random forest, M5 Tree and the Volterra in emulating a) SM_{UL} and b) SM_{UL} at four study sites.

(Note: The dashed lines are the least-squares regression line and the solid red line is the 45° or the X = Y line for comparison).

Table S3 Number of data points that were over/underpredicted in comparison to a 5% tolerance limit in forecasting a) upper layer (SM_{UL}) and b) lower layer soil moisture (SM_{LL}) by the ANN-CoM and the contrasting standalone models.

a) SM_{UL}	**Volterra**	**M5 Tree**	**RF**	**ELM**	**ANN-CoM**
			Site 1 - Menindee		
Underprediction	21	16	15	6	8
Overprediction	13	13	22	5	4
Total	34	29	37	**11**	12
			Site 2- Balranald		
Underprediction	22	16	12	13	10
Overprediction	19	19	19	13	16
Total	41	35	31	**26**	**26**
			Site 3 - Bobadah		
Underprediction	24	16	16	6	10
Overprediction	15	18	23	22	18
Total	39	34	39	**28**	**28**
			Site 4 - Rocky Creek		
Underprediction	34	21	15	13	12
Overprediction	7	13	20	3	3
Total	41	34	35	16	**15**

b) SM_{LL}	**Volterra**	**M5 Tree**	**RF**	**ELM**	**ANN-CoM**
			Site 1 - Menindee		
Underprediction	12	0	0	0	0
Overprediction	2	0	0	0	0
Total	14	0	0	0	0
			Site 2- Balranald		
Underprediction	35	0	0	2	0
Overprediction	2	0	0	1	0
Total	37	0	0	3	0
			Site 3 - Bobadah		
Underprediction	30	0	1	0	0
Overprediction	7	1	0	0	0
Total	37	1	1	0	0
			Site 4 - Rocky Creek		
Underprediction	18	0	2	0	0
Overprediction	8	1	1	0	0
Total	26	1	3	0	0

Chapter 6: Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach

Foreword

This chapter is an exact copy of the submitted manuscript to the *Catena* journal.

As outlined in Chapter 5, the soil moisture is dependent upon many factors, which need to be appropriately incorporated into the respective forecasting models. However, if non-stationarity features are not appropriately accounted for, the performance of classical models may degrade. As a result, the advanced and self-adaptive multi-resolution utility, EEMD, is suited for the purpose. Yet currently, EEMD could only be used in a single variable input approach. Hence in this study, a novel multivariate sequential EEMD approach is developed and evaluated in forecasting of near-real-time *i.e.*, weekly soil moisture levels. In addition to methodological improvement, the operational improvement is also achieved with shorter forecasting horizon. Thirteen inputs are decomposed using sequential multivariate EEMD approach into six intrinsic mode functions (IMFs) and a residual component. A two-stage feature selection is utilized to extract the relevant features from the IMFs and residual signals.

The feature selection included cross-correlation function (*CCF*) followed by random forest driven Boruta wrapper-based algorithm. Integration of this with ELM led to the development of hybrid multivariate sequential EEMD-Boruta-ELM, which is evaluated against comparative hybrid multivariate adaptive regression splines (MARS) (EEMD-Boruta-MARS), classical MARS and classical ELM in forecasting weekly soil moisture values at four study sites.

It must be noted that the standard tolerance for *in-situ* soil moisture measuring instruments is ± 3% (Zamora et al., 2011) and for remotely sensed soil moisture retrieval the tolerance of 4 - 5% is often considered as an acceptable level of accuracy (Kornelsen and Coulibaly, 2013), yet so far there has not been any standardized or allowed error range for soil moisture forecasts. As such the key forecasting tolerance in this study is based upon the percentage/relative $RMSE$ ($RRMSE$) criteria set by Li et al. (2013), which has been widely used. In here, the model is "Excellent" ($RRMSE < 10\%$), "Good" ($10\% < RRMSE < 20\%$), "Fair" ($20\% < RRMSE < 30\%$) or "Poor" ($RRMSE \geq 30\%$).

Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach

Ramendra Prasad, Ravinesh C Deo, Yan Li, and Tek Maraseni

University of Southern Queensland, Australia

Abstract

Soil moisture forecasts are vital for understanding climatic change processes, environmental monitoring, health of ecological systems, agriculture and hydrology. In this study, we design a new multivariate sequential predictive model that utilizes the ensemble empirical mode decomposition (EEMD) algorithm hybridized with extreme learning machines (ELM) to forecast soil moisture (*SM*) over weekly horizons. The EEMD data pre-processing utility is a self-adaptive tool, does not require a predefined basis function and avoids frequency-mode mixing issues. The proposed multivariate sequential EEMD model is designed to sequentially demarcate model predictor variables and the target (*SM*) into analogous intrinsic mode functions (IMFs) and a residue component using the EEMD process, to address the complexities and associated non-linearities in hydrologic-based inputs. To validate the new approach, four diversely characterized sites in Australia's Murray-Darling Basin are purposely selected where 13 weekly hind-casted predictors are collated from Australian Water Availability Project *WaterDyn* physical model. After sequential EEMD transformation process, a two-stage feature selection employing cross-correlation and random forest Boruta wrapper algorithm is adopted to extract pertinent features from hydro-meteorological predictor series to construct a hybridized multivariate sequential EEMD-Boruta-ELM model. Comprehensive model evaluation using statistical metrics and diagnostic plots against alternative methods: hybrid multivariate adaptive regression splines (MARS) (EEMD-Boruta-MARS) and classical MARS and ELM, establish the superiority of hybrid EEMD-Boruta-ELM model, yielding relatively low errors and high performance. The study ascertains that the EEMD-Boruta-ELM hybrid model can be explored as a pertinent data-driven tool for relatively short-term soil forecasts, thus advocating its practical use in near real-time hydrological applications.

1.0 Introduction

Despite being a minuscule percentage (0.0001% of the Earth's water), soil moisture (*SM*) is an important terrestrial water reservoir controlling the hydrological and ecological systems from boundary layer dynamics to global energy cycles (Islam and Engman, 1996). Prolonged low *SM* level combined with a lack of strategic planning, can threaten the hydro-meteorological and agricultural process culminating in health and socio-economic issues. Hence, efficient and reliable predictive systems can serve as an integrated tool for *SM* forecasting and realization of future trends. Besides realizing profitable and agile primary industries, the future *SM* information can permit users and respective authorities to make informed decisions in managing this limited resource and the potential risks resulting from reduced or excessive levels (Argent et al., 2015; Zhang et al., 2017).

Data-intelligent models are able to extract pertinent predictive features from historical data. Subsequently, they have been successfully applied in forecasting hydro-meteorological variables like stream-flow (Srinivasulu and Jain, 2006; Londhe and Dixit, 2012; Mehr *et al.*, 2014; Ni *et al.*, 2010; Prasad *et al.*, 2017; Yaseen *et al.*, 2016; Deo and Sahin, 2016), air temperature (Kisi and Sanikhani, 2015), drought (Deo, Tiwari*, et al.*, 2016; Deo, Ravinesh C. *et al.*, 2017), rainfall run-off (Zhang and Govindaraju, 2000; Young *et al.*, 2017; Hosseini and Mahjouri, 2016) and water demand (Tiwari and Adamowski, 2013; Tiwari *et al.*, 2016; Mouatadid and Adamowski, 2016). Few studies pertaining to soil moisture forecasting has also been conducted, with applications of artificial neural network (ANN) (Elshorbagy and Parasuraman, 2008; Kornelsen and Coulibaly, 2014), extreme learning machine (ELM) (Liu *et al.*, 2014), multivariate relevance vector machine (Zaman and McKee, 2014) and random forest (Matei *et al.*, 2017). Yet, these studies implemented classical non-tuned standalone modelling techniques that have innate limitations in terms of generalization capabilities. To address this, scholars have tried heuristic optimizations, particularly with ANNs to improve the forecasting capability including genetic algorithm-GA (Huang *et al.*, 2010) and particle swarm optimization-PSO (Xiaoxia and Chengming, 2016; Yang *et al.*, 2017). Recently, Prasad *et al.* (2018) designed a hybrid ANN based committee ensemble model to forecast *SM* at monthly horizons. However, this study together with the aforementioned optimization studies did not employ any kind of multi-resolution data pre-processing schemes.

A major concern with traditional standalone data-driven models is their inability to handle non-stationarity features if the inputs are not properly pre-processed (Adamowski *et al.*, 2012; Adamowski and Chan, 2011; Wang, D. *et al.*, 2017; Deo, Ravinesh C *et al.*, 2017). Soil moisture and interconnected hydro-meteorological predictors naturally comprise of entrenched non-stationarity features that induce non-normality, bimodality, asymmetric cycles and non-linearities (Wu *et al.*, 2011). To overcome this and simultaneously improve the model performance, it is necessary to incorporate a multi-resolution data pre-processing tool to appropriately unveil the features. The application of classical multi-resolution analysis (MRA) tools, *i.e.*, Fourier transformation, only performs the transformations at frequency resolutions losing the time stamp, which is a major drawback. Alternatively, discrete wavelet transformation (DWT) has been propitiously adopted in forecasting applications (Mallat, 1998, 1989; Nourani *et al.*, 2014; Nourani *et al.*, 2009; Deo, Tiwari*, et al.*, 2016; Deo, Wen*, et al.*, 2016; Krishna *et al.*, 2011) and has been trialled by Yang *et al.* (2017) as DWT-ANN combined with PSO to forecast *SM* in China. However, the key weakness of DWT is the inherent decimation effect that curtails the information and generates half the wavelet coefficients, while the other half of the smooth version is recursively processed at a coarser resolution by high and low pass filters (Rathinasamy *et al.*, 2014). Instead, a more advanced wavelet tool *viz.* maximum-overlap discrete wavelet transformation (MODWT) is preferred that solves the critical decimation issue of DWT (Prasad *et al.*, 2017; Rathinasamy *et al.*, 2014; Cornish *et al.*, 2005; Dghais and Ismail, 2013; Percival *et al.*, 2011). Fundamentally, both DWT and MODWT analysis translate the time-frequency information over the time-variable window and therefore requires a well-suited user-defined basis function/mother wavelet (Chen *et al.*, 2012). The selection of an apt mother wavelet is yet an unresolved concern which is achieved via a rather lengthy trial and error process (Prasad *et al.*, 2017). The other conventional time series decomposition approaches (*e.g.*, singular value decomposition (SVD), singular spectrum analysis (SSA) (Chau and Wu, 2010; Chitsaz *et al.*, 2016), principal component analysis (PCA) (Hu *et al.*, 2007), and empirical decomposition) are contingent upon autocorrelations and are non-local making it unsuitable for extracting physically meaningful information from non-stationary time series (Wu *et al.*, 2011).

Consequently, the empirical mode decomposition (EMD) was developed to segregate higher frequency input series into lower frequency resolved components (Huang *et al.*, 1998). The key merit of EMD technique is that the decompositions do not require prescribed frequency bands or imposed basis functions forming a completely self-adaptive procedure. In addition, EMD is a temporally local decomposition technique that uses extrema information of the riding waves in non-stationary time series to extract and isolate salient features representing the physical structure of the data (Wu *et al.*, 2011). Yet, the end-points extending problem (Qiu *et al.*, 2017) together with an important 'mode mixing' shortcoming impedes the EMD algorithm. Hence, an advanced noise-assisted version, the ensemble-EMD (EEMD) was devised by Wu and Huang (2009), whereby a Gaussian white noise is added to the original (undecomposed) series to efficiently extract the embedded periodic and trend information within a time series. Successful applications of EEMD-based data-driven models has been noted in forecasting precipitation (Jiao *et al.*, 2016; Beltran-Castro *et al.*, 2013; Ouyang *et al.*, 2016), reservoir inflows (Bai *et al.*, 2015), daily river data (Seo and Kim, 2016) and soil moisture (Prasad et al., 2018b). Despite these studies substantiating that EEMD ensemble models generate improved forecasts, *SM* forecasting via this approach has not been extensively explored. The literature shows only one study by Basha *et al.* (2015) utilized EEMD coupled Non-Stationary Oscillation Resampling (NSOR) model to forecast temperature, precipitation and *SM* patterns that were compared with the Coupled Model Intercomparison Project phase 5 (CMIP5) projections. They found improved forecasting capability of the EEMD-NSOR model. In addition to having improved forecasts, the EEMD-based models also reduce the difficulties in the forecasting process by making the time series less intricate (Di *et al.*, 2014).

In spite of the aforementioned enhancements, so far all EEMD-based studies used single predictor forecasting technique whereby the lagged time series of the objective variable was used to forecast the future data (Beltran-Castro *et al.*, 2013; Jiao *et al.*, 2016; Ouyang *et al.*, 2016; Bai *et al.*, 2015; Basha *et al.*, 2015; Seo and Kim, 2016). None of these studies utilized multiple input variables, which is a critical issue since environmental and hydrological variables are driven by many influencing parameters that may have been left out. For instance, the *SM* level is naturally contingent upon the antecedent condition of soil evaporation, evapotranspiration, surface runoff and deep percolation (*i.e.*, groundwater recharge) (Van Loon, 2015) and therefore, these parameters need to be appropriately incorporated into the respective models.

To utilize several predictors and subsequently extract most, if not all, possible relevant predictive features, a new multivariate sequential EEMD hybridized modelling approach is developed in this study and applied to forecast soil moisture at the weekly horizon. Twelve hydro-meteorological predictor time series and a lagged *SM* series (thirteen in total) hindcasted by the *WaterDyn* physical model, developed under the Australian Water Availability Project (AWAP), are acquired. These inputs are transformed into respective intrinsic mode functions (IMFs) and a residual component in a sequential manner. Then the cross-correlation function followed by Boruta feature selection algorithm (a random forest-based wrapper process) is further implemented to reduce the input dimensions and optimize the forecasts. Boruta input selection is an easy to tune wrapper algorithm developed by Kursa *et al.* (2010) and has been strongly recommended as a feature selection tool for predictive model applications (Li *et al.*, 2016; Christa *et al.*, 2017). With random forest (RF) as the underlying instrument model, Boruta incorporates the interactions between features and iteratively removes the irrelevant input(s). Although Boruta has been successfully applied as a feature selection utility in modeling and predicting forest biodiversity (Leutner *et al.*, 2012), asymptomatic stress (Poona and Ismail, 2014), seabed hardness (Li *et al.*, 2016), sponge species richness (Li *et al.*, 2017) and $PM_{2.5}$ air quality (Lyu *et al.*, 2017), this technique has been not been explored in hydrological applications anywhere.

In this study, we hybridize extreme learning machines (ELM) with feature selection and multi-resolution utilities, leading to the design of a newly proposed EEMD-Boruta-ELM hybrid model. This is the first study to utilize the multivariate sequential EEMD hybrid technique in forecasting *SM*, and evaluating its predictive capability in a drought-prone region. ELM model used in this study is a robust, convenient to use and a computationally efficient nonlinear artificial intelligence algorithm (Shamshirband *et al.*, 2015; Huang *et al.*, 2015; Xu and Wang, 2016) developed by Huang *et al.* (2004). Alternatively, a comparative multivariate sequential EEMD hybridized multivariate adaptive regression splines (MARS) model (*i.e.*, EEMD-Boruta-MARS) is also developed. MARS is a nonlinear model that utilizes partitioned input space to develop piecewise associations within the training data into basis functions to generate forecasts (Friedman, 1991). Finally, the performance of the hybrid EEMD-Boruta-ELM model is validated against the comparative hybridized EEMD-Boruta-MARS and standalone ELM and MARS models. The next section outlines the study area and data followed by the description of hybrid data-intelligent model development procedure. After that, the empirical study and results are presented and finally, the paper is concluded encapsulating the findings and key considerations for future study.

2.0 Brief accounts of data-driven modeling frameworks

This section provides a brief account of data-driven modeling algorithms used in this study.

2.1 Extreme learning machine (ELM)

The ELM is a single layer feed-forward neural network (SLFN) proven to have good generalization capability with computationally inexpensive easy to tune network (Ahila et al., 2015; Huang et al., 2006; Huang et al., 2015). The mathematical realization of the ELM network can concisely be outlined as follows (Huang et al., 2004; Huang et al., 2006):

$$\sum_{j=1}^{J} W_j\, h\left(c_j \cdot x_i + \varepsilon_j\right) = O_i \qquad (1)$$

where $\{(x_i, y_i): x_i \in \boldsymbol{R}^{\boldsymbol{P}}, y_i \in \boldsymbol{R}\}$ are $i = 1, 2, \dots N$ distinct samples of training data; P = the number of input neurons; $O_i \in \boldsymbol{R}$ represents the model output, $W \in \boldsymbol{R}^J$ represents the weights in between '*J*' hidden layers and output node, $c_j \in \boldsymbol{R}$ are the weights between the input layer and hidden layer and ε_j are the learning parameters of the hidden layers; $h(\cdot)$ is the activation function, and; $j = 1, 2, \dots J$ are the indices of hidden neurons.

In a concise form with '*H*' as the hidden layer output matrix, '*W*' as the weights and '*Y*' as the training output, Eq. 1 can be rewritten as:

$$Y = HW \qquad (2)$$

After randomly assigning suitable input weights (c_j) and corresponding biases, the algorithm analytically determines the hidden layer output matrix, H. Subsequently, the network establishes a linear system whereby the output weights matrix is determined via a least-square solution to yield zero forecasting errors (between O_i *and* y_i) as:

$$\widehat{W} = H^{\dagger}Y \tag{3}$$

where $H^{\dagger}$ is the Moore–Penrose generalized inverse of H.

Finally, the forecasts are generated using these aleatory assigned input weights and algorithm computed output weights.

2.2 Multivariate adaptive regression splines (MARS)

The MARS algorithm, introduced by Friedman (1991), is a non-parametric combination of additive and/or interactive simple linear functions. Over an equivalent interval, the algorithm partitions the training data into several splines, which again are split into several subgroups separated by knots. Subsequently, a pair of basis functions (*BF*) describing the associations between the predictor variable and the target are created at respective knots to produce continuous models with continuous derivatives (Friedman, 1991). Considering a knot of the i^{th} subgroup at position k, the output (O) with predictor input training vector, $\boldsymbol{X}: x_t \in \boldsymbol{R}$, and corresponding target vector, $\boldsymbol{Y}: y_t \in \boldsymbol{R}$, where t=1, 2 …N is the number of datum points, could be formulated as:

$$\left.\begin{array}{c} O = \mathrm{BF}_i(x) = \max(0, x-k) \\ \textit{and its mirror}: \\ O = \mathrm{BF}_i(x) = \max(0, k-x) \end{array}\right\} \tag{4}$$

The final modeled output, ***Y'***, is the summation of a series of *BF*s, as:

$$\boldsymbol{Y}' = \alpha + \sum_i^I (\rho_i \times BF_i) \tag{5}$$

where α is a general constant, ρ is a unique constant for respective BF and I represents the maximum number of subgroups.

Initially, a forward stepwise approach maximizes the number of potential knots creating a large overfitting model. Consequently, a backward deletion phase iteratively prunes the *BF*s that contribute the least towards model fit based on minimum Generalized Cross Validation (*GCV*) computed as follows (Craven and Wahba, 1979; Deo et al., 2017a; Friedman and Silverman, 1989):

$$GCV = \frac{\mathrm{MSE}}{(1 - \frac{\mathrm{C(M)}}{\mathrm{N}})^2} \tag{6}$$

where *MSE* is mean squared error and *C(M)* is the penalty factor for *M* number of *BF*s. Eventually, the optimal model with least *GCV* is selected.

3.0 Materials and Methods

3.1 Study area

To develop and effusively investigate the performance of the proposed hybrid model in forecasting weekly soil moisture, the agricultural Murray-Darling basin region (MDB) with an area of 1,042,730 km² (14% of mainland Australia) is considered (The Murray–Darling Basin Authority, 2010). The region of focus is the state of New South Wales (NSW), located on the east coast of Australia. The key agricultural commodities for export growth in NSW over the last 5 years were beef, vegetables, and fruit (NSW-Department of Industry, 2017). NSW accounted for ~ $^1/_4$ of Australia's wine exports by volume and 38% of Australian total sheep and lamb flock size in the last financial year (2015-2016) (Australian Bureau of Statistics, 2017) asserting that NSW is one of the most significant agricultural states in Australia. Consequently, four sites (Site 1-Menindee, Site 2-Cooinbil, Site 3-Fairfield, and Site 4-Bodangora) were selected as illustrated in Figure 1.

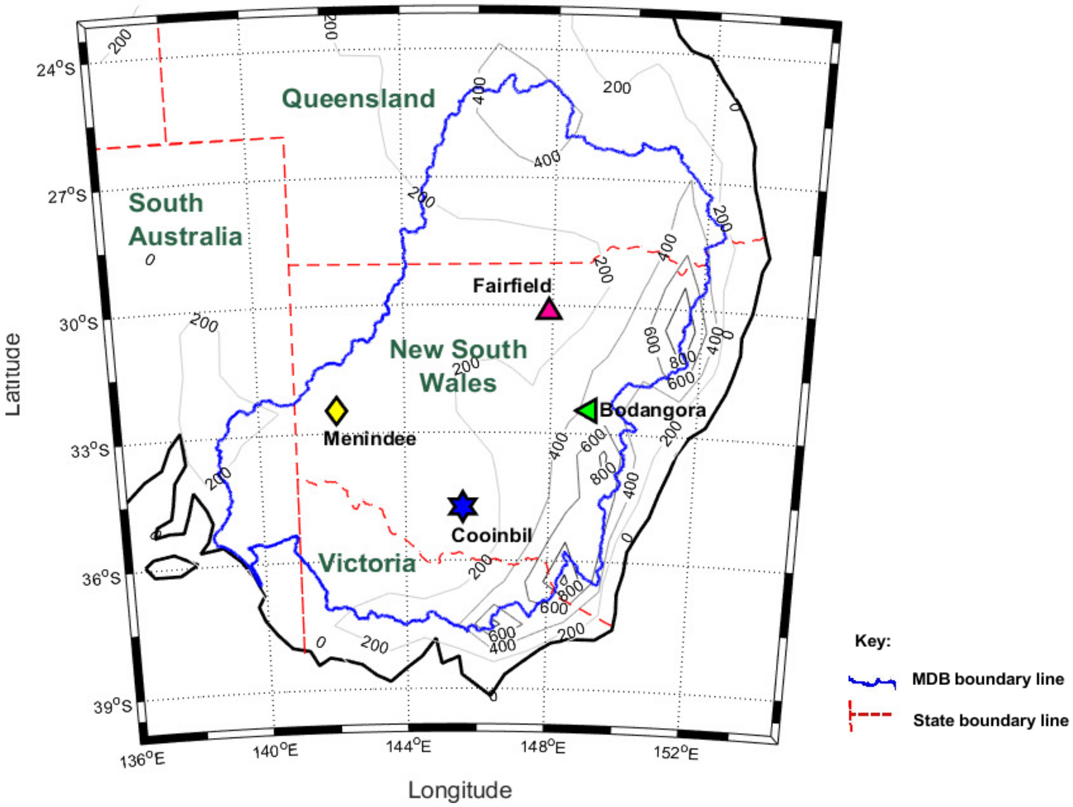


Figure 1 The study region showing the candidate test sites and their geographical locations within the Australian Murray-Darling Basin overlayed with elevation contours (grey lines).

The study sites show disparate geophysical features as demonstrated by the primary climate classes (Hijmans et al., 2005), land use (Department of Agriculture and Water Resources, 2015), soil types (ASRIS, 2014) and range of agricultural holding (Australian Bureau of Statistics, 2008) data combined with differing elevations (Table 1).

Table 1 Geographic and physical characteristics of the tested study sites where the hybrid EEMD-Boruta-ELM model is evaluated against the EEMD-Boruta-MARS, MARS, and ELM for forecasting soil moisture over weekly horizons.

Site No.	**Site Name**	**Geographic characteristics**			**Physical characteristics**			
		Long. (°E)	Lat. (°S)	Elev. (m)	Primary climate classes (Hijmans *et al.*, 2005)	Land-use (Department of Agriculture and Water Resources, 2015)	Soil type (ASRIS, 2014)	The range of agricultural holding (ha) (Australian Bureau of Statistics, 2008)
1	Menindee	142.15	32.45	75.3	Desert	Grazing-Native vegetation	Calcarosol	18700-38600
2	Cooinbil	145.60	34.75	111.4	Savannah	Grazing-modified pastures	Sodosol	600-3700
3	Fairfield	147.90	30.15	131.0	Savannah	Dry-land cropping	Vertosol	3700-18700
4	Bodangora	149.05	32.45	486.7	Sub-Tropical	Dry-land cropping	Sodosol	600-3700

The weekly data for the study were sourced from Australian Water Availability Project (AWAP) that commenced weekly data generation at 0.05° × 0.05° grid resolution in January 2007 (Raupach et al., 2009; Raupach et al., 2012). Accordingly, the study period is from January 2007 to December 2016 and the cut-off was 01 January 2017 to account for final week overlap. A total of 13 AWAP derived predictors were collated, including twelve weekly hydro-meteorological inputs and a lagged *SM* data series as shown in Table 2, while the target was successive future *SM* data. AWAP utilizes the *WaterDyn* physical model to simulate the soil hydrological parameters including *SM* level after incorporating previously recorded meteorological data, soil characteristics, vegetation greenness, solar irradiance and albedo (Raupach et al., 2009; Raupach et al., 2012). The *SM* data are relative values bounded by [0, 1] computed with respect to the base climatological reference period (from 1961 to 1990) and is up to a depth of 0.20 m from the surface (Raupach et al., 2009) while the hydro-meteorological inputs are in their respective standard units (Table 2).

The stochastic nature of weekly *SM*,that warrants the utilization of a multi-resolution data pre-processing method, is apparent from Figure 2, supported by descriptive statistics (Table 3). The data at Sites 1 (Skew = 1.140) and 3 (Skew = 1.450) are positively skewed while the other two show symmetrical distributions. The kurtosis factors show that distributions at all sites have fewer and less extreme outliers (Kurt<3). In addition, lower levels of soil moisture were recorded at Site 1-Menindee with the lowest value of 0.012. This site falls in the desert climate class (Table 1), while Site 4-Bodangora is in sub-tropical class and Sites 2 (Cooinbil) and 3 (Fairfield) lie in the savannah climate class.

Table 2 Database of the weekly input variables for the study period January 2007–December 2016 adopted for developing the multivariate sequential hybrid EEMD-Boruta-ELM, hybrid EEMD-Boruta-MARS, ELM, and MARS models. Source: Australian Water Availability Project-AWAP (Raupach *et al.*, 2012, 2009).

	Variables	***Acronym***	**Units**
INPUTS	Total weekly local discharge (Runoff + Drainage)	FWDis	mm
	Total weekly total evaporation (Soil + Vegetation)	FWE	mm
	Total weekly deep drainage	FWLch2	mm
	Total weekly soil evaporation	FWsoil	mm
	Total weekly total transpiration	FWTra	mm
	Total weekly open water evaporation ('pan' equiv.)	FWWater	mm
	Weekly average sensible heat flux	PhiH	W/m^2
	Weekly average latent heat flux	PhiE	W/m^2
	Total weekly precipitation	PCN	mm
	Weekly average incident solar radiation	SolarMJ	MJ/m^2
	Weekly average maximum temperature	Tmax	°C
	Weekly average minimum temperature	Tmin	°C
	Antecedent weekly relative soil moisture	SM_{t-1}	Fraction 0 - 1
Objective variable	Relative soil moisture (0-0.20 m deep)	*SM*	Fraction 0 - 1

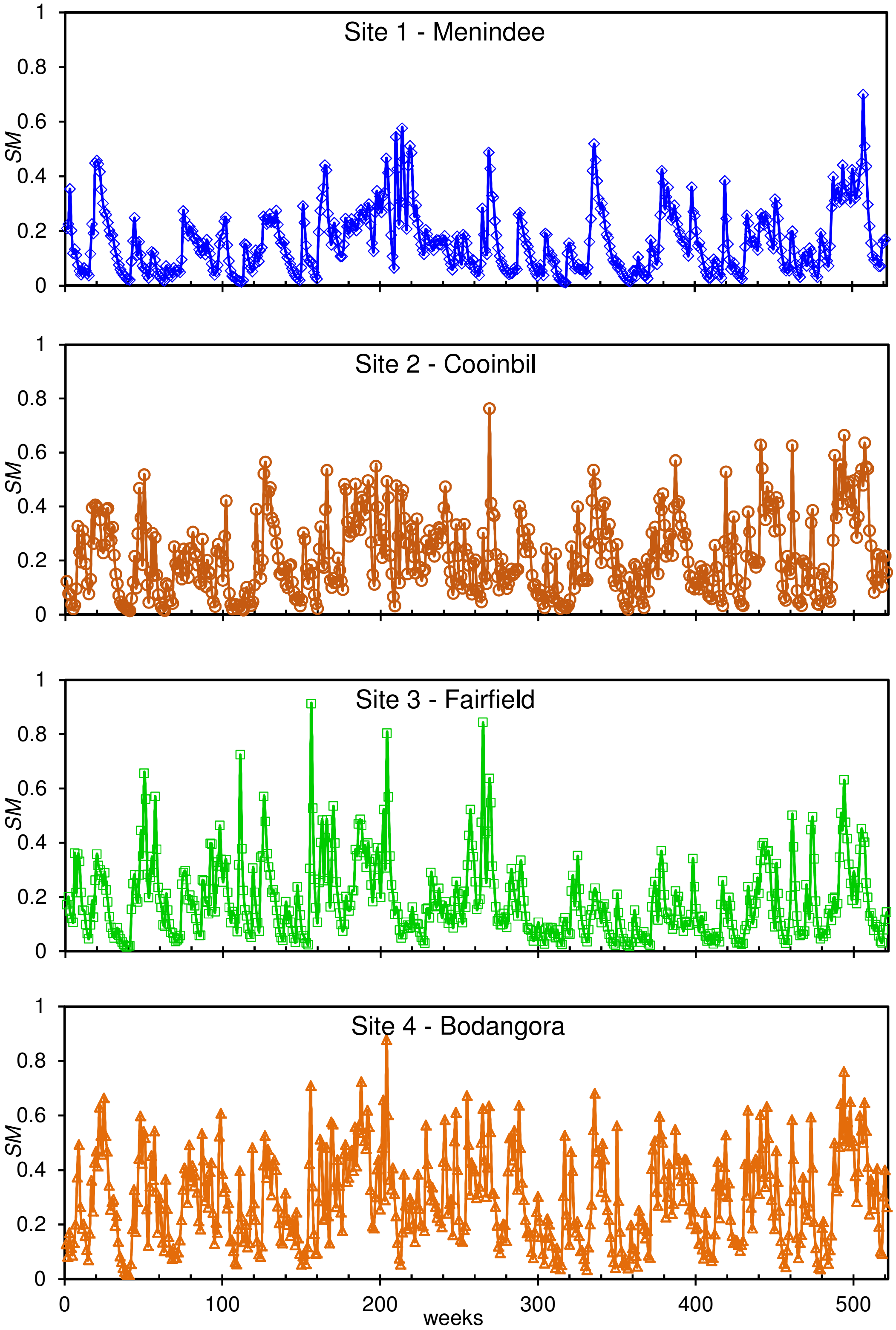


Figure 2 Time-series of the normalized weekly soil moisture (*SM*) at the respective sites showing the stochastic nature of the hydrological variable.

Table 3 Weekly hydrological statistics of the relative soil moisture at the candidate test sites.

Site No.	Site Name	Weekly statistical features of relative soil moisture				
		Minimum	Maximum	Mean	Skewness	Kurtosis
1	Menindee	0.012	0.699	0.160	1.140	1.185
2	Cooinbil	0.014	0.762	0.217	0.739	0.024
3	Fairfield	0.016	0.913	0.188	1.450	2.872
4	Bodangora	0.015	0.878	0.291	0.475	-0.500

Note: *The relative soil moisture values are based on a base climatological reference period: 1961–1990, recommended by the Australian Bureau of Meteorology.*

3.2 The proposed hybrid multivariate sequential EEMD-Boruta-ELM approach

A hybrid multivariate sequential EEMD transformation technique integrated with Boruta feature selection and ELM modeling algorithm for weekly soil moisture forecasting is proposed. The sequential EEMD ensemble-modeling scheme using multivariate inputs is illustrated in Figure 3 and can be described as follows:

a) *Sequential EEMD stage*: Initially, all 13 predictor inputs (Table 2) and the target (*SM*) were partitioned into training (40% *i.e.*, 207 datum points), validation (30% *i.e.*, 155 datum points) and testing (30% *i.e.*, 155 datum points) after excluding five points to allow for five weekly lags from a total of 522 weekly datum points. This was done to prevent any inclusion of future data into the training and validation sets. Then, all 13 predictor inputs and the target data were sequentially transformed using EEMD into six IMFs and a residual component. Algorithm detail of EEMD is provided in the following subsection 3.4.

b) *IMF Collation*: The IMFs of similar nature were pooled together, *i.e.*, all IMF 1s (of inputs and target) were pooled into one set, then all IMF 2's were pooled into the next set until the sixth one, then finally all residuals were pooled into the final set.

c) *Feature selection and significant inputs*: Besides reducing the input dimensions, the important feature selection stage also increases the model efficiency, optimizes the model performances and could provide an insight into the underlying physical processes without altering the data (Bennasar et al., 2015).

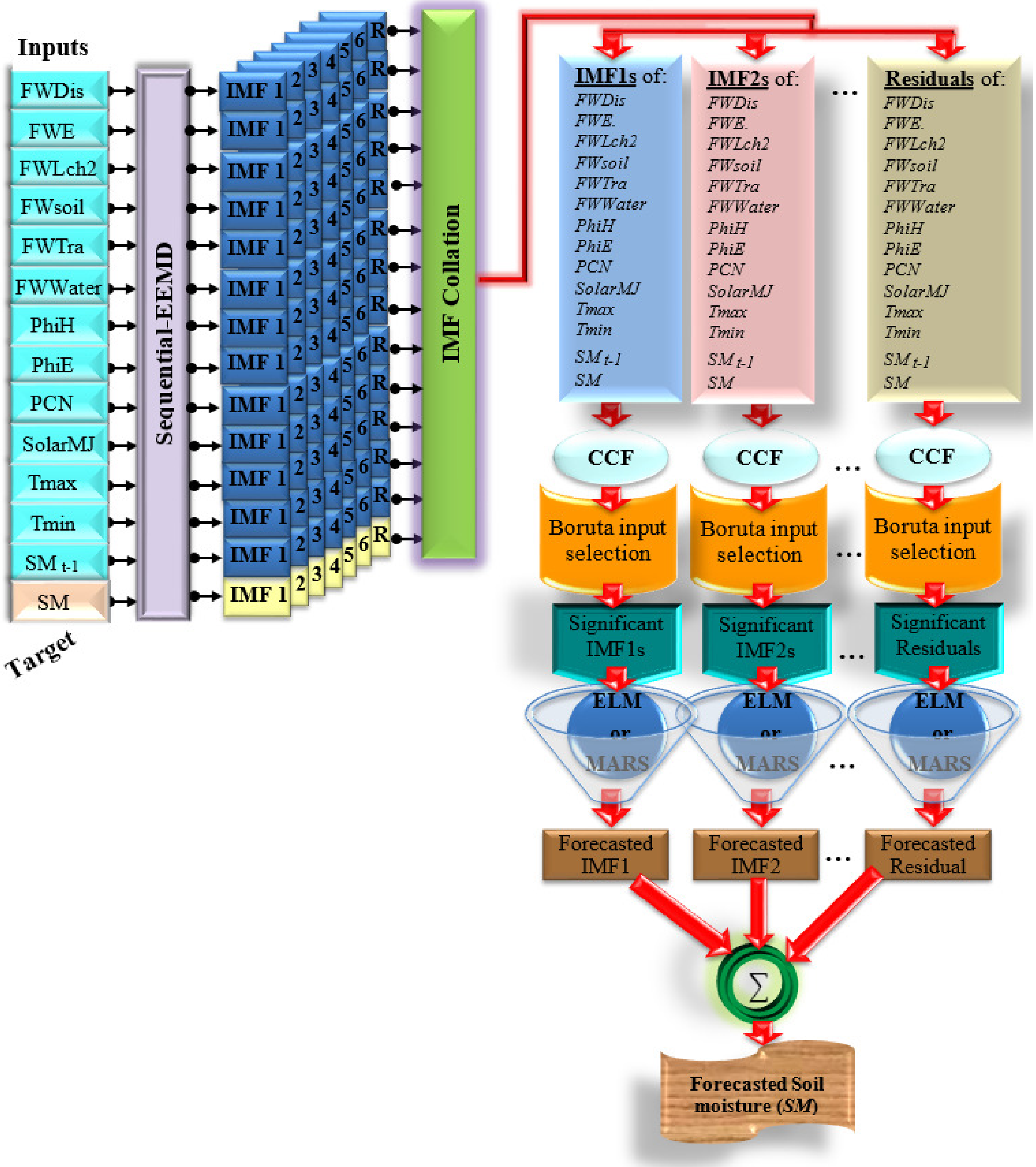


Figure 3 Schematic of the two-phase hybrid multivariate sequential ensemble empirical mode decomposition-extreme learning machine model optimized with the Boruta wrapper-based feature selection (*i.e.*, hybrid EEMD-Boruta-ELM) and the comparative EEMD-Boruta-MARS model constructed for weekly soil moisture forecasting. [For model input names, see Table 2].

A two-stage feature selection was carried out to reduce the input dimensionality. First, the cross-correlation function (*CCF*) was applied to determine the significant lags of analogous IMFs and residual inputs. Later the significant lags of IMFs and residuals were subjected to Boruta feature selection algorithm for a final screening process. Using an underlying random forest wrapper mechanism, Boruta computes the Z-scores of each predictor inputs relative to the shadow attribute (Boruta algorithm detail is provided in subsection 3.5). The distribution of Z-score metrics determines the importance factors (Kursa et al. (2010). A minimal-optimal feature selection strategy was adopted by ranking the salient IMFs and residual based on the Boruta determined importance factors and then a stepwise model building was carried. Table 4 illustrates the salient model inputs determined after the two-phase feature selection for all models at all sites. Both the hybrids EEMD-Boruta-ELM and EEMD-Boruta-MARS models required the same number of inputs at Site 1 (3 significant inputs) and Site 3 (10 significant inputs). In comparison to the hybrid EEMD-Boruta-ELM, the EEMD-Boruta-MARS required one and four more features to reach its peak performance at Sites 2 and 4, respectively. To get all data on a uniform scale, prior to feature selection the data normalization was carried out to confine in the range of 0 and 1.

d) *Ensemble forecasting*: Channelling the screened multiple predictor inputs (as in part c), into the ELM or MARS models, respective IMFs and the residual component were forecasted. Brief accounts of ELM and MARS data-driven modeling frameworks are outlined in section 2.

e) *Ensemble averaging*: Then the forecasted IMFs and the residual component were integrated at the end to generate the forecasted *SM* values.

Table 4 The salient model input variables, determined by cross-correlation and Boruta feature selection. [For standalone models the inputs were the statistically significant lags of intact or undecomposed time series whereas, for hybrid EEMD-Boruta models, the lagged intrinsic mode functions (IMFs) and residual components were inputs. Table 2 contains input variable names while *t* – 1, *t* – 2… *t* – 5 are the weekly lagged data series].

Model platform		ELM	MARS
Site 1 - Menindee			
Standalone model		FWDis$_{t-1}$, FWE$_{t-1}$, FWLch2$_{t-1}$, FWPT$_{t-1}$, FWSoil$_{t-1}$, FWTra$_{t-1}$, FWWater$_{t-1}$, PhiE$_{t-1}$, PhiH$_{t-1}$, PCN$_{t-1}$, SolarMJ$_{t-1}$, Tmax$_{t-1}$, Tmin$_{t-1}$	
Hybrid Model **EEMD-Boruta**		**3 significant inputs**	**3 significant inputs**
	IMF-1	SM_{t-2}, PCN_{t-1}, $SolarMJ_{t-1}$	SM_{t-2}, PCN_{t-1}, $SolarMJ_{t-1}$
	IMF-2	SM_{t-1}, PCN_{t-1}, SM_{t-2}	SM_{t-1}, PCN_{t-1}, SM_{t-2}
	IMF-3	SM_{t-1}, SM_{t-2}, $PhiH_{t-1}$	SM_{t-1}, SM_{t-2}, $PhiH_{t-1}$
	IMF-4	SM_{t-1}, SM_{t-2}, SM_{t-3}	SM_{t-1}, SM_{t-2}, SM_{t-3}
	IMF-5	SM_{t-1}, SM_{t-2}, SM_{t-3}	SM_{t-1}, SM_{t-2}, SM_{t-3}
	IMF-6	SM_{t-1}, SM_{t-2}, SM_{t-3}	SM_{t-1}, SM_{t-2}, SM_{t-3}
	Residual	SM_{t-2}, SM_{t-1}, SM_{t-3}	SM_{t-2}, SM_{t-1}, SM_{t-3}
Site 2 - Cooinbil			
Standalone model		FWDis$_{t-1}$, FWLch2$_{t-1}$, FWPT$_{t-1}$, FWWater$_{t-1}$, PhiE$_{t-1}$, PhiH$_{t-1}$, PCN$_{t-1}$, SolarMJ$_{t-1}$, Tmax$_{t-1}$, Tmin$_{t-1}$, SM$_{t-1}$	
Hybrid Model **EEMD-Boruta**		**7 significant inputs**	**8 significant inputs**
	IMF-1	SM_{t-2}, $FWLch2_{t-1}$, $FWTra_{t-1}$, PCN_{t-1}, $FWSoil_{t-1}$, FWE_{t-1}, $PhiE_{t-1}$	SM_{t-2}, $FWLch2_{t-1}$, $FWTra_{t-1}$, PCN_{t-1}, $FWSoil_{t-1}$, FWE_{t-1}, $PhiE_{t-1}$, $PhiE_{t-2}$
	IMF-2	SM_{t-1}, PCN_{t-1}, $FWLch2_{t-2}$, $FWLch2_{t-1}$, SM_{t-4}, $PhiH_{t-1}$, SM_{t-3}	SM_{t-1}, PCN_{t-1}, $FWLch2_{t-2}$, $FWLch2_{t-1}$, SM_{t-4}, $PhiH_{t-1}$, SM_{t-3}, $FWSoil_{t-1}$
	IMF-3	SM_{t-1}, SM_{t-2}, PCN_{t-1}, $PhiH_{t-1}$, SM_{t-3}, $FWLch2_{t-5}$, $FWSoil_{t-1}$	SM_{t-1}, SM_{t-2}, PCN_{t-1}, $PhiH_{t-1}$, SM_{t-3}, $FWLch2_{t-5}$, $FWSoil_{t-1}$, $FWLch2_{t-3}$
	IMF-4	SM_{t-1}, SM_{t-2}, SM_{t-3}, SM_{t-4}, $FWLch2_{t-5}$, $FWTra_{t-2}$, $FWTra_{t-4}$	SM_{t-1}, SM_{t-2}, SM_{t-3}, SM_{t-4}, $FWLch2_{t-5}$, $FWTra_{t-2}$, $FWTra_{t-4}$, PCN_{t-1}
	IMF-5	SM_{t-1}, SM_{t-2}, SM_{t-3}, $PhiE_{t-4}$, SM_{t-4}, SM_{t-5}, $PhiE_{t-5}$	SM_{t-1}, SM_{t-2}, SM_{t-3}, $PhiE_{t-4}$, SM_{t-4}, SM_{t-5}, $PhiE_{t-5}$, $FWLch2_{t-5}$
	IMF-6	SM_{t-1}, SM_{t-2}, SM_{t-3}, $PhiH_{t-5}$, SM_{t-4}, $PhiH_{t-4}$, $PhiH_{t-3}$	SM_{t-1}, SM_{t-2}, SM_{t-3}, $PhiH_{t-5}$, SM_{t-4}, $PhiH_{t-4}$, $PhiH_{t-3}$, SM_{t-5}
	Residual	$FWTra_{t-3}$, $FWTra_{t-5}$, SM_{t-1}, $FWTra_{t-4}$, SM_{t-2}, SM_{t-3}, $FWTra_{t-1}$	$FWTra_{t-3}$, $FWTra_{t-5}$, SM_{t-1}, $FWTra_{t-4}$, SM_{t-2}, SM_{t-3}, $FWTra_{t-1}$, SM_{t-5}
Site 3 - Fairfield			

Standalone model			$FWDis_{t-1}$, FWE_{t-1}, $FWPT_{t-1}$, $FWSoil_{t-1}$, $FWTra_{t-1}$, $FWWater_{t-1}$, $PhiE_{t-1}$, $PhiH_{t-1}$, PCN_{t-1}, $SolarMJ_{t-1}$, $Tmax_{t-1}$, SM_{t-1}	
Hybrid Model	**EEMD-Boruta**		**10 significant inputs**	**10 significant inputs**
		IMF-1	*$FWLch2_{t-1}$, $FWSoil_{t-1}$, $PhiE_{t-1}$, PCN_{t-2}, PCN_{t-1}, $FWDis_{t-2}$, $FWSoil_{t-2}$, FWE_{t-1}, $SolarMJ_{t-1}$, PCN_{t-3}*	*$FWLch2_{t-1}$, $FWSoil_{t-1}$, $PhiE_{t-1}$, PCN_{t-2}, PCN_{t-1}, $FWDis_{t-2}$, $FWSoil_{t-2}$, FWE_{t-1}, $SolarMJ_{t-1}$, PCN_{t-3}*
		IMF-2	*PCN_{t-1}, SM_{t-1}, $PhiH_{t-1}$, SM_{t-4}, SM_{t-5}, $SolarMJ_{t-1}$, $PhiE_{t-1}$, PCN_{t-2}, $FWSoil_{t-1}$, $FWLch2_{t-2}$*	*PCN_{t-1}, SM_{t-1}, $PhiH_{t-1}$, SM_{t-4}, SM_{t-5}, $SolarMJ_{t-1}$, $PhiE_{t-1}$, PCN_{t-2}, $FWSoil_{t-1}$, $FWLch2_{t-2}$*
		IMF-3	*SM_{t-1}, SM_{t-2}, $PhiH_{t-1}$, $FWLch2_{t-4}$, $FWLch2_{t-3}$, PCN_{t-1}, $FWLch2_{t-5}$, $PhiE_{t-1}$, PCN_{t-2}, FWE_{t-1}*	*SM_{t-1}, SM_{t-2}, $PhiH_{t-1}$, $FWLch2_{t-4}$, $FWLch2_{t-3}$, PCN_{t-1}, $FWLch2_{t-5}$, $PhiE_{t-1}$, PCN_{t-2}, FWE_{t-1}*
		IMF-4	*SM_{t-1}, SM_{t-2}, SM_{t-3}, $FWLch2_{t-5}$, SM_{t-4}, $FWLch2_{t-4}$, FWE_{t-1}, $FWLch2_{t-2}$, $FWLch2_{t-3}$, $FWLch2_{t-1}$*	*SM_{t-1}, SM_{t-2}, SM_{t-3}, $FWLch2_{t-5}$, SM_{t-4}, $FWLch2_{t-4}$, FWE_{t-1}, $FWLch2_{t-2}$, $FWLch2_{t-3}$, $FWLch2_{t-1}$*
		IMF-5	*SM_{t-1}, SM_{t-2}, SM_{t-3}, FWE_{t-3}, FWE_{t-1}, FWE_{t-4}, $FWLch2_{t-3}$, FWE_{t-2}, $Tmax_{t-1}$, $FWLch2_{t-5}$*	*SM_{t-1}, SM_{t-2}, SM_{t-3}, FWE_{t-3}, FWE_{t-1}, FWE_{t-4}, $FWLch2_{t-3}$, FWE_{t-2}, $Tmax_{t-1}$, $FWLch2_{t-5}$*
		IMF-6	*SM_{t-1}, SM_{t-4}, SM_{t-5}, SM_{t-3}, SM_{t-2}, $FWSoil_{t-1}$, $FWSoil_{t-2}$, $PhiH_{t-5}$, $PhiH_{t-4}$, $PhiH_{t-3}$*	*SM_{t-1}, SM_{t-4}, SM_{t-5}, SM_{t-3}, SM_{t-2}, $FWSoil_{t-1}$, $FWSoil_{t-2}$, $PhiH_{t-5}$, $PhiH_{t-4}$, $PhiH_{t-3}$*
		Residual	*SM_{t-1}, SM_{t-2}, SM_{t-4}, SM_{t-3}, PCN_{t-1}, SM_{t-5}, $FWDis_{t-1}$, PCN_{t-2}, PCN_{t-3}, $PhiE_{t-1}$*	*SM_{t-1}, SM_{t-2}, SM_{t-4}, SM_{t-3}, PCN_{t-1}, SM_{t-5}, $FWDis_{t-1}$, PCN_{t-2}, PCN_{t-3}, $PhiE_{t-1}$*

Site 4 - Bodangora

Standalone model			$FWDis_{t-1}$, $FWLch2_{t-1}$, $FWPT_{t-1}$, $FWWater_{t-1}$, $PhiH_{t-1}$, PCN_{t-1}, $SolarMJ_{t-1}$, $Tmax_{t-1}$, SM_{t-1}	
Hybrid Model	**EEMD-Boruta**		**6 significant inputs**	**10 significant inputs**
		IMF-1	*$FWLch2_{t-1}$, $FWSoil_{t-1}$, SM_{t-2}, SM_{t-1}, FWE_{t-1}, PCN_{t-1}*	*$FWLch2_{t-1}$, $FWSoil_{t-1}$, SM_{t-2}, SM_{t-1}, FWE_{t-1}, PCN_{t-1}, $PhiE_{t-1}$, $FWTra_{t-1}$, PCN_{t-3}, $PhiH_{t-2}$*
		IMF-2	*SM_{t-1}, PCN_{t-1}, SM_{t-4}, $PhiH_{t-1}$, SM_{t-3}, $SolarMJ_{t-1}$*	*SM_{t-1}, PCN_{t-1}, SM_{t-4}, $PhiH_{t-1}$, SM_{t-3}, $SolarMJ_{t-1}$, $FWLch2_{t-1}$, $FWPT_{t-1}$, $FWWater_{t-1}$, $PhiH_{t-4}$*
		IMF-3	*SM_{t-1}, PCN_{t-1}, SM_{t-2}, PCN_{t-2}, $PhiH_{t-1}$, $FWSoil_{t-1}$*	*SM_{t-1}, PCN_{t-1}, SM_{t-2}, PCN_{t-2}, $PhiH_{t-1}$, $FWSoil_{t-1}$, SM_{t-3}, $FWLch2_{t-5}$, $FWLch2_{t-4}$, $FWLch2_{t-3}$*
		IMF-4	*SM_{t-1}, SM_{t-2}, SM_{t-3}, PCN_{t-1}, SM_{t-4}, PCN_{t-2}*	*SM_{t-1}, SM_{t-2}, SM_{t-3}, PCN_{t-1}, SM_{t-4}, PCN_{t-2}, $PhiE_{t-2}$, SM_{t-5}, $PhiE_{t-1}$, $PhiE_{t-3}$*
		IMF-5	*SM_{t-1}, SM_{t-2}, SM_{t-3}, SM_{t-4}, SM_{t-5}, $FWTra_{t-4}$*	*SM_{t-1}, SM_{t-2}, SM_{t-3}, SM_{t-4}, SM_{t-5}, $FWTra_{t-4}$, $FWTra_{t-5}$, $FWTra_{t-3}$, $FWDis_{t-1}$, $FWSoil_{t-1}$*
		IMF-6	*SM_{t-1}, SM_{t-2}, $FWSoil_{t-1}$, $FWSoil_{t-2}$, SM_{t-3}, FWE_{t-1}*	*SM_{t-1}, SM_{t-2}, $FWSoil_{t-1}$, $FWSoil_{t-2}$, SM_{t-3}, FWE_{t-1}, PCN_{t-5}, $PhiE_{t-1}$, $FWSoil_{t-3}$, SM_{t-4}*
		Residual	*SM_{t-1}, SM_{t-2}, SM_{t-3}, SM_{t-4}, $PhiE_{t-1}$, PCN_{t-1}*	*SM_{t-1}, SM_{t-2}, SM_{t-3}, SM_{t-4}, $PhiE_{t-1}$, PCN_{t-1}, SM_{t-5}, $FWTra_{t-1}$, FWE_{t-1}, $FWSoil_{t-3}$*

3.3 Model development procedure

In the forecasting of weekly *SM* (*i.e.*, the predictant), the MATLAB platform running on Intel *i*7, 3.40 GHz processor was utilized to develop all models. Firstly, the ELM models with hidden neurons varying from 1 to 200 were developed. In addition, different transfer functions including sigmoidal, sine, hard-limit, triangular basis, and radial basis were trialed. Then using the ARESLab toolbox (version 1.13.0) (Jekabsons, 2016), a piecewise cubic MARS model was developed in two phases, *i.e.*, forward selection and backward deletion. During the forward selection phase, the algorithm iteratively adds the basis function pairs to the initial intercept term in order to minimize the objective function (*i.e.*, *MSE*) creating a large model. This model is prone to overfitting, therefore a backward deletion phase is executed whereby the model is pruned backward with the elimination of functions one by one until the intercept term is left. The final, best-performing model's selection was contingent upon least Generalized Cross-Validation (*GCV*) value.

For the case of hybrid multivariate sequential-EEMD based models (*i.e.*, EEMD-Boruta-ELM and EEMD-Boruta-MARS), the significant IMFs and residuals were used as inputs, while in developing the standalone models (ELM and MARS), intact or undecomposed hydro-meteorological time series were used as inputs. During experimentation, the validation data set was utilized to determine the optimal models based on Pearson's correlation coefficient (*r*), root mean square error (*RMSE*) and mean absolute error (*MAE*) that are presented in Table 5. Correspondingly, the architectures of best-performing hybrid EEMD-Boruta-ELM and standalone ELM models at all candidate sites are shown in Table 6. Similarly, Table 7 illustrates the basis functions (BF) and *GCV* statistics for the EEMD-Boruta-MARS and standalone MARS at all sites. The maximum number of BFs for standalone MARS model was 8 at Site 4-Bodangora, while to forecast individual IMFs, up to 15 BFs were required.

3.4 Ensemble empirical mode decomposition (EEMD)

Besides effectively overcoming the mode mixing issue of EMD, the multi-resolution analysis utility EEMD features a strong self-adaptability and local variation characteristics (Li et al., 2015; Wu and Huang, 2009). With an added Gaussian white noise providing a uniform reference frame in the time-frequency domain, EEMD detects and separates the embedded oscillations at different scales into intrinsic mode functions (IMFs) and the trend/residual component (Wu and Huang, 2009). The band-limited IMFs must fulfill two acceptability conditions (Huang et al., 1998; Sharpley and Vatchev, 2005; Wu and Huang, 2009) *i.e.*, (a) it needs to have exactly one zero between any two consecutive local extrema and (b) have a zero "local mean".

Table 5 Evaluation of the multivariate sequential hybrid EEMD-Boruta-ELM *vs*. the hybrid EEMD-Boruta-MARS, MARS and ELM models during model-development phase (*i.e.*, training and validation) using the *r* = Pearson's correlation coefficient, *RMSE* = root mean square error and *MAE* = mean absolute error.

Model	**ELM**						**MARS**					
	Training			**Validation**			**Training**			**Validation**		
	r	***RMSE***	***MAE***	***r***	***RMSE***	***MAE***	***r***	***RMSE***	***MAE***	***r***	***RMSE***	***MAE***
Site 1 - Menindee												
Stand-alone model	0.871	0.054	0.033	0.842	0.072	0.044	0.864	0.055	0.035	0.844	0.064	0.039
Hybrid: EEMD-Boruta	0.956	0.033	0.020	0.781	0.082	0.052	0.954	0.033	0.021	0.825	0.070	0.051
Site 2 - Cooinbil												
Stand-alone model	0.760	0.090	0.065	0.628	0.098	0.068	0.760	0.090	0.064	0.685	0.092	0.064
Hybrid: EEMD-Boruta	0.879	0.066	0.046	0.850	0.067	0.044	0.923	0.054	0.039	0.828	0.075	0.053
Site 3 - Fairfield												
Stand-alone model	0.766	0.099	0.070	0.845	0.071	0.055	0.791	0.094	0.067	0.822	0.072	0.055
Hybrid: EEMD-Boruta	0.922	0.060	0.044	0.830	0.074	0.048	0.909	0.064	0.047	0.884	0.063	0.046
Site 4 - Bodangora												
Stand-alone model	0.770	0.107	0.080	0.689	0.122	0.096	0.765	0.108	0.083	0.622	0.144	0.113
Hybrid: EEMD-Boruta	0.887	0.078	0.058	0.886	0.075	0.059	0.918	0.067	0.051	0.908	0.069	0.049

Table 6 Model development framework of multivariate sequential hybrid EEMD-Boruta-ELM and standalone ELM applied in forecasting the relative soil moisture.

Models	Number of Neurons			Transfer function
	Input layer	Hidden layer	Output layer	
Site 1 - Menindee				
ELM	13	14	1	*sigmoid*
multivariate sequential hybrid EEMD-Boruta-ELM	3	5	1	*sine*
Site 2 - Cooinbil				
ELM	11	8	1	*sine*
multivariate sequential hybrid EEMD-Boruta-ELM	7	6	1	*radial basis*
Site 3 - Fairfield				
ELM	12	6	1	*sine*
multivariate sequential hybrid EEMD-Boruta-ELM	10	17	1	*sine*
Site 4 - Bodangora				
ELM	9	16	1	*sine*
multivariate sequential hybrid EEMD-Boruta-ELM	6	7	1	*sigmoid*

Table 7 Model development framework of (a) standalone MARS and (b) hybrid EEMD-Boruta-MARS including the respective model equations with the basis functions (BF) and generalized cross-validation (*GCV*) statistic during the training period.

a) MARS	Model equations	Optimal basis functions	*GCV*
Site 1 - Menindee	$y = 0.348 - 1.107BF_1 + 1.601BF_2 - 0.451BF_3 - 1.158BF_4 + 0.377BF_5$	5	0.00347
Site 2 - Cooinbil	$y = 0.287 - 0.861BF_1 + 0.232BF_2 + 0.397BF_3$	3	0.00887
Site 3 - Fairfield	$y = 0.263 - 1.295BF_1 + 0.400BF_2 + 0.252BF_3 + 0.321BF_4$	4	0.01009
Site 4 - Bodangora	$y = 0.724 - 0.559BF_1 - 0.861BF_2 + 0.371BF_3 - 1.395BF_4 + 0.531BF_5 - 0.330BF_6 + 1.828BF_7 - 0.412BF_8$	8	0.01397

b) hybrid EEMD-Boruta-MARS

Site 1 - Menindee	IMF1	$IMF1^{FOR} = 0.493 - 1.653BF_1 + 0.965BF_2$	2	0.00381
	IMF2	$IMF2^{FOR} = 0.538 + 1.121BF_1 - 1.783BF_2 + 0.747BF_3 - 0.304BF_4 + 0.555BF_5 - 0.450BF_6 - 0.347BF_7$	7	0.00107
	IMF3	$IMF3^{FOR} = 0.425 - 2.932BF_1 + 1.697BF_2 - 0.884BF_3$	3	0.00129
	IMF4	$IMF4^{FOR} = 0.521 - 0.204BF_1 + 0.477BF_2 + 0.914BF_3 - 1.495BF_4$	4	0.00023

	IMF5	$IMF5^{FOR} = 0.669 - 0.911BF_1 + 1.015BF_2$	2	0.00894
	IMF6	$IMF6^{FOR} = 0.178 + 1.871BF_1 - 25.371BF_2 + 0.052BF_3 + 38.040BF_4 - 10.812BF_5$	5	0.01179
	Residual	$Res^{FOR} = 0.203 + 0.669BF_1 - 0.544BF_2$	2	0.00155
	Overall-ensemble	$SM^{FOR} = IMF1^{FOR} + IMF2^{FOR} + IMF3^{FOR} + IMF4^{FOR} + IMF5^{FOR} + IMF6^{FOR} + Res^{FOR}$	-	0.00410
Site 2 - Cooinbil	IMF1	$IMF1^{FOR} = 1.393 - 1.691BF_1 + 2.381BF_2 - 0.787BF_3 + 0.561BF_4 + 1.456BF_5 - 1.065BF_6 - 1.702BF_7$	7	0.01877
	IMF2	$IMF2^{FOR} = 0.450 + 0.812BF_1 - 0.456BF_2 - 9.954BF_3 + 7.445BF_4 - 0.696BF_5$	5	0.00263
	IMF3	$IMF3^{FOR} = 1.367 - 0.968BF_1 + 0.946BF_2 + 1.463BF_3 - 0.122BF_4 + 2.136BF_5 - 1.922BF_6$	6	0.00059
	IMF4	$IMF4^{FOR} = 0.877 + 2.175BF_1 - 1.598BF_2 + 0.095BF_3 - 0.685BF_4 + 0.524BF_5 + 0.487BF_6 + 0.335BF_7 - 0.180BF_8 - 0.445BF_9 + 0.176BF_{10} - 0.464BF_{11} - 0.638BF_{12}$	12	0.00018
	IMF5	$IMF5^{FOR} = 1.632 - 4.428BF_1 + 4.418BF_2 + 1.259BF_3 - 0.223BF_4 + 3.387BF_5 - 3.505BF_6 - 3.054BF_7 + 2.239BF_8 - 1.140BF_9 + 0.303BF_{10} - 0.331BF_{11} + 0.222BF_{12} - 1.874BF_{13} + 3.289BF_{14} - 2.276BF_{15}$	15	0.00009
	IMF6	$IMF6^{FOR} = 0.806 + 1.052BF_1 - 0.923BF_2 - 0.491BF_3 + 0.855BF_4 - 0.164BF_5 - 0.133BF_6$	6	0.00005
	Residual	$Res^{FOR} = 0.182 + 1.270BF_1 - 0.917BF_2 - 0.203BF_3 + 0.343BF_4$	4	0.00001
	Overall-ensemble	$SM^{FOR} = IMF1^{FOR} + IMF2^{FOR} + IMF3^{FOR} + IMF4^{FOR} + IMF5^{FOR} + IMF6^{FOR} + Res^{FOR}$	-	0.00319
Site 3 - Fairfield	IMF1	$IMF1^{FOR} = 0.498 - 4.602BF_1 - 0.525BF_2 - 0.876BF_3 + 4.272BF_4$	4	0.01159
	IMF2	$IMF2^{FOR} = 0.131 + 0.397BF_1 - 0.578BF_2 - 0.305BF_3 + 0.533BF_4 + 0.718BF_5 - 1.070BF_6 - 0.258BF_7 + 0.141BF_8 + 0.526BF_9$	9	0.00462
	IMF3	$IMF3^{FOR} = -0.255 + 1.865BF_1 - 1.850BF_2 - 0.769BF_3 + 0.999BF_4 + 0.427BF_5 - 0.081BF_6$	6	0.00090
	IMF4	$IMF4^{FOR} = -0.041 - 1.780BF_1 + 2.004BF_2 + 2.002BF_3 - 2.199BF_4 + 0.228BF_5 - 1.067BF_6 + 0.020BF_7 + 0.267BF_8 + 0.732BF_9 - 0.142BF_{10} + 0.770BF_{11} - 1.155BF_{12} - 1.082BF_{13} + 1.424BF_{14} + 0.402BF_{15}$	15	0.00038
	IMF5	$IMF5^{FOR} = 0.407 + 0.627BF_1 - 0.652BF_2 - 0.126BF_3 + 0.426BF_4 - 0.335BF_5 + 0.530BF_6 + 0.037BF_7$	7	0.00006
	IMF6	$IMF6^{FOR} = 1.808 + 0.573BF_1 - 0.637BF_2 + 2.739BF_3 - 2.654BF_4 - 2.272BF_5 + 2.392BF_6$	6	0.00001
	Residual	$Res^{FOR} = 0.658 + 1.265BF_1 - 0.887BF_2 + 0.099BF_3 - 0.176BF_4$	4	0.00000
	Overall-ensemble	$SM^{FOR} = IMF1^{FOR} + IMF2^{FOR} + IMF3^{FOR} + IMF4^{FOR} + IMF5^{FOR} + IMF6^{FOR} + Res^{FOR}$	-	0.00251

Site 4 - Bodangora	IMF1	$IMF1^{FOR} = 0.613 - 0.396BF_1 - 4.242BF_2 + 4.046BF_3 - 0.640BF_4$	4	0.02006
	IMF2	$IMF2^{FOR} = 0.371 - 0.826BF_1 + 0.907BF_2 - 0.796BF_3 + 0.477BF_4 + 0.449BF_5 - 3.172BF_6 + 0.274BF_7 + 3.815BF_8 + 0.264BF_9 + 0.946BF_{10} - 1.293BF_{11} - 4.837BF_{12} + 3.496BF_{13}$	13	0.00420
	IMF3	$IMF3^{FOR} = 0.227 + 0.477BF_1 - 1.230BF_2 + 0.289BF_3 + 0.876BF_4 - 0.249BF_5 - 0.227BF_6 + 0.149BF_7 + 1.157BF_8 - 1.502BF_9$	9	0.00102
	IMF4	$IMF4^{FOR} = 0.564 + 0.556BF_1 - 0.680BF_2 + 0.407BF_3 - 0.807BF_4 + 0.497BF_5 + 0.084BF_6 + 2.342BF_7 - 2.695BF_8 - 1.492BF_9 + 1.850BF_{10}$	10	0.00018
	IMF5	$IMF5^{FOR} = -0.239 - 3.615BF_1 + 4.336BF_2 + 4.928BF_3 - 5.187BF_4$	4	0.00117
	IMF6	$IMF6^{FOR} = 1.086 + 1.687BF_1 - 1.697BF_2 - 0.519BF_3 + 0.619BF_4 - 0.045BF_5 + 0.259BF_6$	6	0.00002
	Residual	$Res^{FOR} = 0.165 + 1.342BF_1 - 1.459BF_2 - 0.362BF_3 + 0.350BF_4$	4	0.00001
	Overall-ensemble	$SM^{FOR} = IMF1^{FOR} + IMF2^{FOR} + IMF3^{FOR} + IMF4^{FOR} + IMF5^{FOR} + IMF6^{FOR} + Res^{FOR}$	-	0.00381

Note: The '$IMF(n)^{FOR}$' is the respective forecasted IMFs from n = 1 to 6. 'SM^{FOR}' is the forecasted soil moisture and *GCV* for "Overall-ensemble" is an average value.

Considering an undecomposed time-series $z(t)$, the concise realization of the EEMD algorithm is as follows (Wu and Huang, 2009):

(1) Add a Gaussian white noise series $g(t)$ to get $z'(t) = z(t) + g(t)$.

(2) Decompose $z'(t)$ into IMFs and repeat this procedure with altered white noise series each time until the maximum number of repetitions/ensemble number is reached.

(3) Finally, the ensemble mean of all IMFs and the mean of residue components are computed eliminating the added Gaussian noise.

After extensive decompositions, the time series can be expressed as:

$$z(t) = \sum_{i=1}^{p} IMF_i(t) + R_p(t) \tag{7}$$

where $IMF_i(t)$ is the intrinsic mode functions, $R_p(t)$ denotes the final residue component, p is the total number of IMFs, and i is the component indices.

The output optimization is achieved via a statistical rule devised by Wu and Huang (2009) that is based on the ensemble number (N) and the amplitude of the added white noise (ε) with e_n as the final standard deviation as below:

$$e_n = \frac{\varepsilon}{\sqrt{N}} \tag{8}$$

The recommended values of respective parameters; $e_n = 0.20$ and N=100, were used (Ouyang et al., 2016; Ren et al., 2015; Wang et al., 2013; Wu and Huang, 2009).

3.5 Feature selection algorithm: Boruta input selection

For a set of T distinct samples of predictors ($x_t \in \boldsymbol{R}^n$) and target ($y_t \in \boldsymbol{R}$) with n = the number of inputs and t =1, 2, …T, the algorithm can briefly be outlined as follows (Kursa et al., 2010; Kursa and Rudnicki, 2010):

1) Create a permuted (*i.e.*, randomly ordered) shadow (duplicated) variable, x_t' for respective input vector, x_t, to add randomness and remove correlations between shadow inputs and the target (y_t).
2) Using a random forest model predict the target (y_t) using both x_t' and x_t as inputs.
3) Compute the variable importance measures *i.e.*, permutation importance or Mean Decrease Accuracy (MDA) for every input (x_t) and the respective shadow input (x_t') over all trees (m_{tree} = 500 in this study) as (Hur et al., 2017; Strobl et al., 2008):

$$MDA = \frac{1}{m_{tree}} \sum_{m=1}^{m_{tree}} \frac{\sum_{t \in OOB} I(y_t = f(x_t)) - \sum_{t \in OOB} I(y_t = f(x_t^n))}{|OOB|} \tag{9}$$

where $I(\bullet)$ is the indicator function; OOB (Out-of-Bag) is the prediction error of each of the training samples utilizing bootstrap aggregation; $(y_t = f(x_t))$ are forecasted values before permuting; and $(y_t = f(x_t^n))$ are forecasted values after permuting.

4) Calculate the Z-scores as:

$$\text{Z-score} = \frac{MDA}{SD} \tag{10}$$

such that *SD* is the standard deviation of accuracy losses, and then determine the maximum Z-score among shadow attributes (*MZSA*).

5) Following that, the Z-scores of the inputs are compared with the corresponding shadows and evaluated using a variable importance distribution. The inputs with Z-scores < *MZSA* are tagged "Unimportant" and permanently removed while inputs having Z-scores > *MZSA* are tagged "Confirmed".
6) During each iteration, new shadows are created and the algorithm stops when all inputs are either "Confirmed", or the iteration threshold (maxRuns *i.e.*,

500 in this study) is reached. The unassigned inputs after reaching maxRuns are classed as "Tentative" and are either confirmed or rejected by comparing the respective median Z-scores with the median Z-scores of the best shadow input.

An example of the 'importance plot' from Boruta input selection algorithm at Site 4-Bodangora, is shown in Figure 4. It illustrates that precipitation (PCN) is the most important input followed by 1 week antecedent *SM* (*SM*(t-1)).

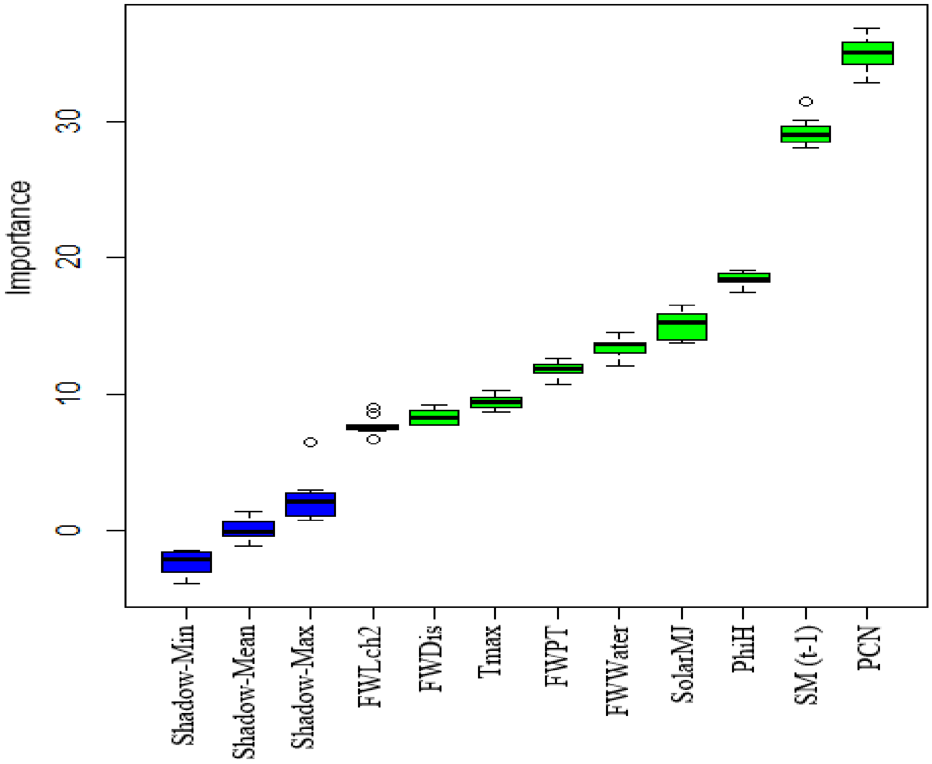


Figure 4 Box plots of the Z-scores registered by the Boruta input selection algorithm (Site 4-Bodangora as an example) used in determining significant antecedent original time-series data used for weekly soil moisture forecasting. Blue corresponds to the shadow inputs while the green represents the Z-score distributions of confirmed inputs with a notably large importance. [For the names of input variables, see Table 2.]

3.6 Model performance comparison measures

Although at least one of the goodness-of-fit measure and a relative error measure and/or at least one absolute error measure is recommended (Legates and McCabe, 1999), a robust evaluation of hybrid EEMD-Boruta-ELM model in forecasting weekly soil moisture was performed. To achieve this, a diverse range of model evaluation metrics have been used, as outlined below (Bennett et al., 2013; Legates and McCabe, 1999; Legates and McCabe, 2013; Nash and Sutcliffe, 1970; Shamseldin, 1997; Willmott, 1981; Willmott, 1984), that concurrently avoids selective interpretation of the statistical outcomes by building on the advantages of each metric. The equations of metrics are presented below with a brief discussion of each. Note that in the equations, SM^{OBS} represents observed soil moisture, SM^{FOR} is forecasted soil moisture values, *i* represents the time-stamp and *N* is the total number of data points.

The first measure, Pearson's correlation coefficient (r) $[Ideal\ value = +1]$ describes the degree of collinearity in between forecasted (SM^{FOR}) and observed data (SM^{OBS}) (Moriasi et al., 2007). However, *r* is absolute and based on linear relations.

$$r = \frac{\sum_{i=1}^{N}\left(SM^{OBS,i} - \overline{SM^{OBS}}\right)\left(SM^{FOR,i} - \overline{SM^{FOR}}\right)}{\sqrt{\sum_{i=1}^{N}\left(SM^{OBS,i} - \overline{SM^{OBS}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(SM^{FOR,i} - \overline{SM^{FOR}}\right)^2}}, (-1 \le r \le 1) \qquad (11)$$

In terms of absolute forecasting error measures, the root mean square error (*RMSE*) $[Ideal\ value = 0]$ and mean absolute error (*MAE*) $[Ideal\ value = 0]$ provides an assessment of the actual forecasting errors with respect to the total number of observations. Yet, the involvement of the squared term in *RMSE* induces a bias towards high *SM* level, while the absolute computations (no squared values) in *MAE* reduces the biases (Chai and Draxler, 2014; Legates and McCabe, 1999; Roy et al., 2016):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(SM^{FOR,i} - SM^{OBS,i}\right)^2}, (0 \le RMSE < +\infty) \qquad (12)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\left(SM^{FOR,i} - SM^{OBS,i}\right)\right|, (0 \leq MAE < +\infty) \quad (13)$$

Additionally, the ratio of mean square error to potential error *i.e.*, the Willmott's Index (*WI*) or index of agreement $[Ideal\ value = +1]$ can detect additive and proportional differences in the observed and forecasted means and variances (Legates and McCabe, 1999; Moriasi et al., 2007).

$$WI = 1 - \left[\frac{\sum_{i=1}^{N}\left(SM^{OBS,i} - SM^{FOR,i}\right)^2}{\sum_{i=1}^{N}\left(\left|SM^{FOR,i} - \overline{SM^{OBS}}\right| + \left|SM^{OBS,i} - \overline{SM^{OBS}}\right|\right)^2}\right], (0 \leq WI \leq 1) \quad (14)$$

With that, the normalized statistic Nash–Sutcliffe Efficiency (E_{NS}) $[Ideal\ value = +1]$ compares the residual variance to the observed data variance and indicates the fitting of observed and forecasted data to 1:1 line (Bennett et al., 2013; Moriasi et al., 2007; Nash and Sutcliffe, 1970).

$$E_{NS} = 1 - \left[\frac{\sum_{i=1}^{N}\left(SM^{OBS,i} - SM^{FOR,i}\right)^2}{\sum_{i=1}^{N}\left(SM^{OBS,i} - \overline{SM^{OBS}}\right)^2}\right], (-\infty < E_{NS} \leq 1) \quad (15)$$

Both WI and E_{NS} utilize squaring of the difference terms making them sensitive to extreme values and registering relatively high metrics.

Alternatively, the Legate-McCabes index (L) $[Ideal\ value = +1]$ is insensitive to extreme SM levels and not inflated since the goodness-of-fit is computed using absolute values with ease of interpretation (Legates and McCabe, 1999).

$$L = 1 - \left[\frac{\sum_{i=1}^{N}\left|SM^{FOR,i} - SM^{OBS,i}\right|}{\sum_{i=1}^{N}\left|SM^{OBS,i} - \overline{SM^{OBS}}\right|}\right], (-\infty < L \leq 1) \quad (16)$$

The performance characterization of models at different sites is an integral component for which relative measures instead of the absolute measures are germane. The Relative Root Mean Square Error (*RRMSE*) and Mean Absolute

Percentage Error ($MAPE$) are used for this purpose since $RRMSE$ and $MAPE$ provide an overall relative forecasting accuracy of respective models and are always positive with $[Ideal\ value = 0]$.

$$RRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(SM^{FOR,i} - SM^{OBS,i}\right)^2}}{\frac{1}{N}\sum_{i=1}^{N}\left(SM^{OBS,i}\right)} \times 100, (0 \le RRMSE < +\infty) \tag{17}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\left(SM^{FOR,i} - SM^{OBS,i}\right)}{SM^{OBS,i}}\right| \times 100, (0 \le MAPE < +\infty) \tag{18}$$

4.0 Results and discussions

This section provides an account of the empirical results of the experiments carried out and the assessments of the performance of the novel multivariate sequential EEMD technique combined with Boruta feature selection in forming the hybrid EEMD-Boruta-ELM model to forecast weekly *SM*. Hence, an equivalent MARS based model (EEMD-Boruta-MARS) and standalone ELM and MARS were developed to benchmark the hybrid EEMD-Boruta-ELM model. The performance evaluation of the multivariate sequential hybrid EEMD-Boruta-ELM model in forecasting *SM* for the short-term *i.e.*, weekly forecast horizon in comparison to an equivalent EEMD-Boruta-MARS and the traditional models; ELM and MARS during the testing period was carried out using the abovementioned statistical metrics (Eq. 11-18), at four hydrological sites within Australian MDB. Using the testing data-sets, the non-normalized metrics, *r*, *RMSE,* and *MAE* were initially used as performance measures due to their wide usage and ease of communication (Bennett et al., 2013). Next, the normalized *WI*, $E_{NS,}$ and *L* characterized the goodness-of-fit and *RRMSE* and *MAPE* provided model comparisons at different sites. Despite the use of numerical assessment metrics, holistic evaluation is imperative for model acceptance and adoption. Hence, graphical assessments using scatter plots, histograms, box-plots, and Taylor plots are performed.

The initial evaluation of the forecasting capability of the hybridized multivariate sequential EEMD-Boruta-ELM model based on non-normalized measures; *r*, *RMSE,* and *MAE* are presented in Table 8.

Table 8 Evaluation of multivariate sequential hybrid EEMD-Boruta-ELM *vs*. the hybrid EEMD-Boruta-MARS, MARS and ELM models in the testing phase. *r* = Pearson's correlation; *RMSE* = root mean square error; *MAE* = mean absolute error; *WI* = Willmott's Index; E_{NS} = Nash–Sutcliffe efficiency, *L* = Legates-McCabe's index. The maximum *L* & minimum *RMSE* and the corresponding measures for each site are **boldfaced**.

	ELM						**MARS**					
Model Type	**Non-normalized measures**			**Normalized measures**			**Non-normalized measures**			**Normalized measures**		
	r	***RMSE***	***MAE***	***WI***	E_{NS}	***L***	***r***	***RMSE***	***MAE***	***WI***	E_{NS}	***L***
Site 1 - Menindee												
Stand-alone model	0.895	0.057	0.038	0.896	0.790	0.617	0.872	0.061	0.041	0.888	0.758	0.592
Hybrid EEMD-Boruta	0.948	0.040	0.027	0.952	0.893	0.735	**0.956**	**0.036**	**0.026**	**0.961**	**0.913**	**0.746**
Site 2 - Cooinbil												
Stand-alone model	0.816	0.094	0.064	0.801	0.658	0.525	0.804	0.098	0.068	0.759	0.628	0.499
Hybrid EEMD-Boruta	**0.912**	**0.068**	**0.052**	**0.912**	**0.821**	**0.615**	0.860	0.094	0.072	0.866	0.655	0.471
Site 3 - Fairfield												
Stand-alone model	0.834	0.074	0.059	0.831	0.662	0.414	0.820	0.078	0.066	0.834	0.617	0.342
Hybrid EEMD-Boruta	**0.938**	**0.050**	**0.037**	**0.920**	**0.843**	**0.634**	0.925	0.059	0.043	0.877	0.784	0.568
Site 4 - Bodangora												
Stand-alone model	0.805	0.099	0.077	0.772	0.644	0.447	0.721	0.115	0.092	0.687	0.514	0.340
Hybrid EEMD-Boruta	**0.909**	**0.072**	**0.052**	**0.881**	**0.812**	**0.628**	0.902	0.073	0.056	0.905	0.804	0.598

It shows that at three sites (Site 2-Cooinbil, Site 3-Fairfield, and Site 4-Bodangora), the hybrid EEMD-Boruta-ELM models had better performances yielding lowest error measures (*RMSE*, and *MAE*) and largest correlation values when compared to the EEMD-Boruta-MARS (the MARS hybrid counterpart) and the traditional (MARS and ELM) models. At Site 1-Menindee, the EEMD-Boruta-MARS model outperformed the other models. A significant decrease in *RMSE* and *MAE* were noted when hybrid EEMD-Boruta-ELM models are compared to the standalone ELM models at all sites. For example, the *RMSE*/*MAE* values were lower by 29.82%/28.95% (Site 1); 27.66%/18.75% (Site 2); 32.43%/37.29% (Site 3); 27.27%/32.46% (Site 4). Similarly, the correlation coefficient, *r*, was relatively larger for hybrid EEMD-Boruta-ELM models at all hydrological sites. The advantage of the multivariate sequential EEMD scheme with the MARS model *i.e.*, EEMD-Boruta-MARS over the classic MARS model is also noticeable in Table 8 with EEMD-Boruta-MARS having scaled performances.

In congruence, the normalized goodness-of-fit indicators (Willmott's Index (*WI*), Nash-Sutcliffe Efficiency ($E_{NS,}$) and the Legate-McCabe's Index (*L*)) justified a better utility of the hybridized multivariate sequential EEMD-Boruta-ELM model compared to the competing models. In accordance with Table 8, the model hybridization (EEMD-Boruta-ELM) led to dramatic improvements in the values of all three metrics (*WI*, $E_{NS,}$ and *L*) in comparison to the classical ELM model at all hydrological sites. For instance, the magnitude of *WI* increased from 0.896 to 0.952 at Site 1-Menindee, it increased from 0.801 to 0.912 at Site 2-Cooinbil, it increased from 0.831 to 0.920 at Site 3-Fairfield and it increased from 0.772 to 0.881 at Site 4-Bodangora. Likewise, the value of E_{NS} also increased by 0.103 (Site 1-Menindee), 0.163 (Site 2-Cooinbil), 0.181 (Site 3-Fairfield) and 0.168 (Site 4-Bodangora). The final normalized metric Legate-McCabe's Index (*L*) that can be considered a better measure, on the basis of its advantages as discussed earlier, also registered large increases. The percentage increases in *L* at all sites achieved by hybrid EEMD-Boruta-ELM model in comparison to standalone ELM model were 19.12% (Site 1), 17.14% (Site 2) and a huge 53.14% at Site 3 and 40.49% at Site 4. The outcomes of the model assessment metrics confirms that the multivariate sequential EEMD based ELM modeling scheme with Boruta feature selection unveiled and aptly selected the entrenched features within the multiple inputs with the lowest errors (*RMSE* $\leq$ 0.072,

$MAE \leq 0.052$) and high performances metrics ($r \geq 0.909$, $WI \geq 0.881$, $E_{NS} \geq 0.812$ and $L \geq 0.615$) realized during the testing period.

Next, the scatterplot of the observed (SM^{OBS}) *vs.* the forecasted (SM^{FOR}) weekly soil moisture values from the four models depicted the concurrent comparisons of SM^{OBS} and SM^{FOR} during the testing period at all four sites (Figure 5). The scatter plots also include the least squares fitting line and the corresponding equation; $SM^{FOR} = (m \times SM^{OBS}) + C$, where m is the gradient of the regression line, and C is the y-intercept. A perfectly fitting model is ought to have a unity-slope/gradient (m), zero y-intercepts (C) and the regression line through the origin with scatter points distributed in the very close proximity of the regression line denoting least discrepancies between SM^{OBS} and SM^{FOR} (Bennett et al., 2013). The gradient (m) together with the coefficient of determination (R^2) are alternative model performance metrics $[Ideal\ value = +1]$. The magnitudes registered from the hybrid EEMD-Boruta-ELM models were close to unity which in pairs ($m|R^2$) are 0.862|0.837 for Site 1-Menindee, 0.747|0.832 for Site 2-Cooinbil, 0.772|0.884 for Site 3-Fairfield and 0.749|0.827 for Site 4-Bodangora whereas for the case of EEMD-Boruta-MARS models, the R^2 values were 0.914, 0.739, 0.856 and 0.813 at these sites, respectively. Alternatively, the y-intercept $[Ideal\ value = 0]$, were found to be close to naught *i.e.*, 0.025 (Site 1); 0.071 (Site 2); 0.020 (Site 3); and 0.063 (Site 4). Probably due to the nature of the data series, at Site 1 there were three outlier points in hybrid EEMD-Boruta-ELM model's scatterplot, which may have degraded its performance relative to EEMD-Boruta-MARS model. The scatterplots display the 1:1 lines together with the linear regression fitting lines of the observed and forecasted values. Comparing these two lines also show that the EEMD-Boruta-ELM had a better performance in comparison to the competing models as the lines were similar in nature and regression line was in very close agreement with the 1:1 lines. Hence, in tandem with the predictor metrics (Table 8), the scatter plots also supported the suitability of hybridized EEMD-Boruta-ELM in forecasting of weekly SM.

The boxplots illustrating the data distribution of the observed (OBS) and forecasted SM values from the four models is shown in Figure 6. Boxplots gives a clear visualization of the data distribution with respect to quartiles distinctly indicating the outliers. The lower and upper horizontal lines of the box represent the lower (Q_{25}) and upper (Q_{75}) quartiles, while the median (Q_{50}) is denoted by the middle line. At all sites, the distributions of forecasted *SM* values generated from the hybrid EEMD-Boruta-ELM models are very similar to that of the observed values. Even the observed and hybrid EEMD-Boruta-ELM model-generated forecasts have a similar number of outlier points. While the forecast distributions of EEMD-Boruta-MARS, standalone ELM, and MARS model are disparate, in comparison to the distribution of observed values.

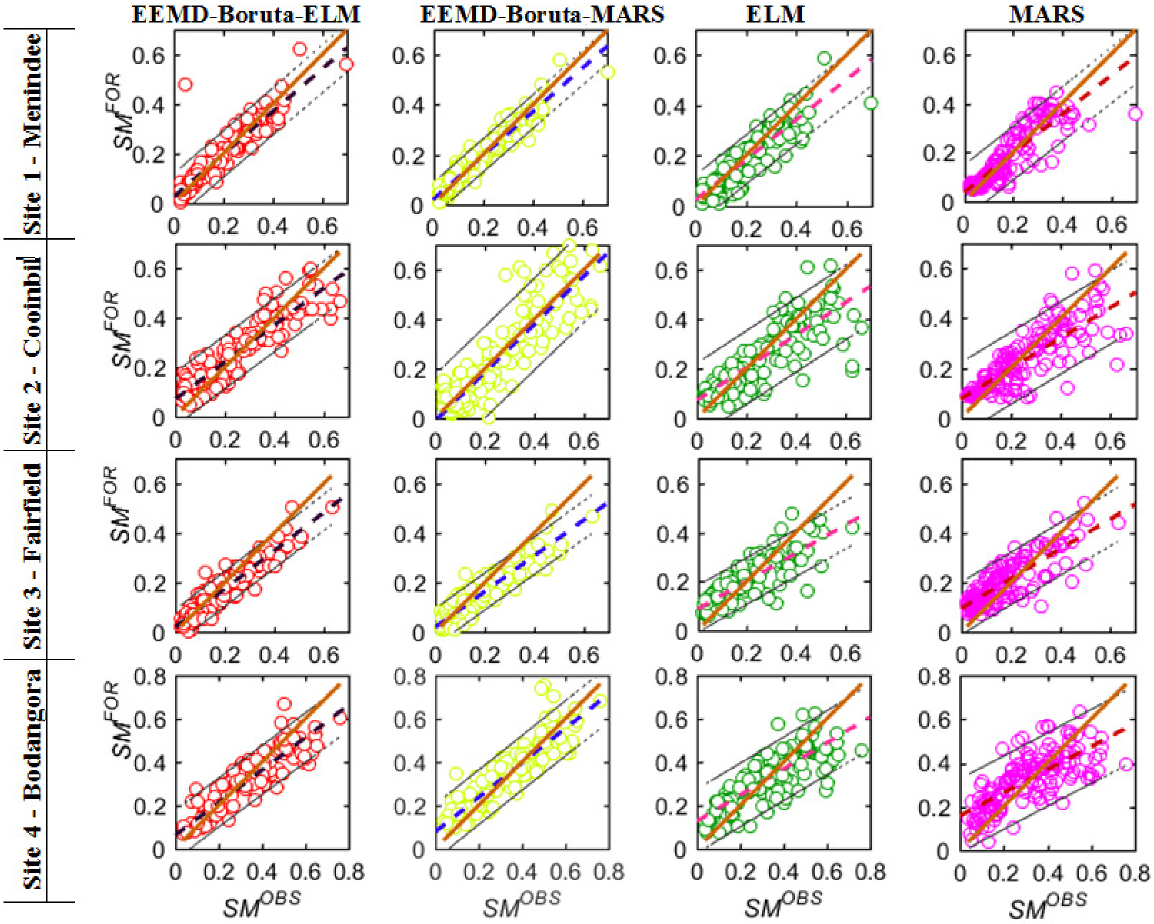


Figure 5 Scatterplot of the observed (SM^{OBS}) *vs*. the forecasted (SM^{FOR}) weekly normalized soil moisture generated from hybrid EEMD-Boruta-ELM, compared with three other data-driven models (*i.e.*, EEMD-Boruta-MARS, MARS, and ELM) in the testing phase. A perfect model linear fit $y = x$ (middle dashed) with upper and lower bounds of 95% prediction intervals, a linear regression fit $y = mx + C$, and the coefficient of determination (R^2) are displayed in each panel. (Note: The dashed colored lines are the least-squares fit line to the respective scatter plots and the solid orange line is 45°, X = Y line for comparison).

In considering the under and over predictions, percentage deviations from the 1:1 line were calculated and the data is available in the Appendix, Table A4. A succinct table is presented here (Table 9) that outlines the number of over and under predicted data points with respect to 5% tolerance limit. Capriciously, at Site 1, EEMD-MARS has a slightly lower total number of over and under predicted data points (*i.e.*, 117/155). Yet, at the other three sites, the hybridized multivariate sequential EEMD-ELM has a least total number of over and under predicted data points showing superior performance of the multivariate sequential EEMD-ELM model in comparison to the similar MARS model.

Table 9 Number of data points that were over/underpredicted in comparison to a 5% tolerance limit in forecasting upper layer soil moisture (SM_{UL}) by the multivariate sequential EEMD-ELM, EEMD-MARS, and the standalone ELM and MARS models.

	MARS	**ELM**	**EEMD-Boruta-MARS**	**EEMD-Boruta-ELM**
		Site 1 - Menindee		
Underprediction	45	62	59	58
Overprediction	84	73	58	65
Total	129	135	**117**	123
		Site 2 - Cooinbil		
Underprediction	66	55	85	53
Overprediction	76	79	58	78
Total	142	134	143	**131**
		Site 3 - Fairfield		
Underprediction	37	40	104	87
Overprediction	110	101	37	49
Total	147	141	141	**136**
		Site 4 - Bodangora		
Underprediction	49	45	47	62
Overprediction	86	92	81	61
Total	135	137	128	**123**

For more perspicacity, the values of maximum, minimum, Q_{75}, Q_{50}, Q_{25}, mean and range of observed and forecasted *SM* generated by the respective models are compared in Table 10. At Sites 2, 3 and 4, the above statistics of *SM* forecasts generated by the hybridized EEMD-Boruta-ELM model were on par with the observed values reinforcing the results in Table 8 and Figures 5 and 6. At Site 1, despite EEMD-Boruta-MARS model recording slightly better values of Q_{75}, Q_{50}, and minimum, the hybrid EEMD-Boruta-ELM model better captured the lower quartile (Q_{25}), mean, range, and maximum values, affirming the better accuracy of hybrid EEMD-Boruta-ELM model.

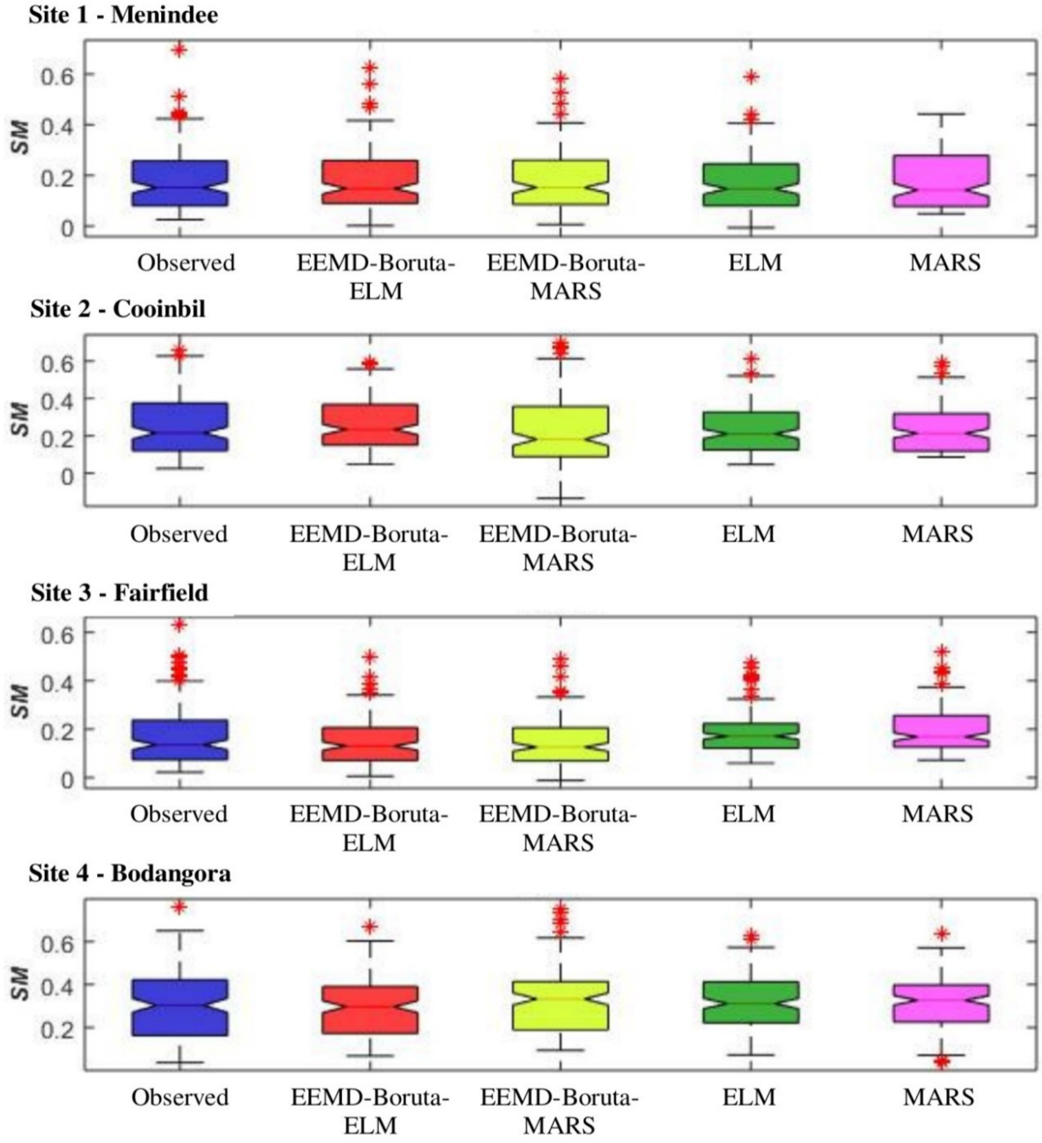


Figure 6 Box plots of the observed *vs*. the forecasted weekly normalized soil moisture generated by the hybrid EEMD-Boruta-ELM *vs*., the comparative models EEMD-Boruta-MARS, ELM and MARS models. [Soil moisture (*SM*) is quantified as relative fractional value and is dimensionless].

Table 10 The testing phase statistics: maximum, minimum, upper quartile (Q_{75}), median (Q_{50}), lower quartile (Q_{25}), mean and range of observed and forecasted soil moisture generated by hybrid EEMD-Boruta-ELM *vs*. the comparative EEMD-Boruta-MARS, MARS and ELM models. (Note: *SM* is dimensionless.)

	Site 1 - Menindee					**Site 2 - Cooinbil**				
Statistic Name	**Standalone**		**Hybridized: EEMD-Boruta**		**OBS**	**Standalone**		**Hybridized: EEMD-Boruta**		**OBS**
	MARS	**ELM**	**MARS**	**ELM**		**MARS**	**ELM**	**MARS**	**ELM**	
Maximum	0.443	0.586	0.579	0.621	0.699	0.591	0.613	0.699	0.594	0.664
Minimum	0.049	-0.006	0.006	0.002	0.026	0.086	0.046	-0.135	0.047	0.025
Q_{75}	0.078	0.080	0.085	0.090	0.080	0.120	0.125	0.090	0.151	0.120
Q_{50}	0.142	0.147	0.152	0.148	0.152	0.213	0.211	0.181	0.234	0.215
Q_{25}	0.278	0.246	0.260	0.257	0.257	0.317	0.325	0.358	0.368	0.372
Mean	0.180	0.167	0.178	0.180	0.180	0.230	0.238	0.235	0.259	0.251
Range	0.394	0.592	0.572	0.619	0.673	0.504	0.567	0.834	0.546	0.639

	Site 3 - Fairfield					**Site 4 - Bodangora**				
Statistic Name	**Standalone**		**Hybridized: EEMD-Boruta**		**OBS**	**Standalone**		**Hybridized: EEMD-Boruta**		**OBS**
	MARS	**ELM**	**MARS**	**ELM**		**MARS**	**ELM**	**MARS**	**ELM**	
Maximum	0.522	0.477	0.492	0.501	0.631	0.636	0.628	0.751	0.670	0.762
Minimum	0.072	0.059	-0.012	0.005	0.022	0.041	0.072	0.094	0.068	0.037
Q_{75}	0.127	0.121	0.069	0.071	0.074	0.227	0.223	0.189	0.174	0.163
Q_{50}	0.168	0.171	0.126	0.130	0.136	0.327	0.311	0.333	0.297	0.303
Q_{25}	0.255	0.223	0.201	0.207	0.236	0.398	0.412	0.413	0.391	0.422
Mean	0.199	0.185	0.143	0.152	0.171	0.320	0.313	0.323	0.294	0.308
Range	0.450	0.417	0.504	0.496	0.610	0.595	0.556	0.657	0.602	0.724

Further to that, the absolute values of the weekly forecasting error were calculated as $|\mathrm{FE}| = SM^{FOR} - SM^{OBS}$ and a histogram was designed to display the percentage frequency of $|\mathrm{FE}|$ (Figure 7). In comparison to standalone models, both the hybridized multivariate EEMD models (*i.e.*, EEMD-Boruta-ELM and EEMD-Boruta-MARS) have a narrower range of $|\mathrm{FE}|$ showing a better accuracy. A closer examination of the percentage $|\mathrm{FE}|$ from the hybrid multivariate EEMD models further strengthens the suitability of hybrid EEMD-Boruta-ELM model with larger percentages (98% at Site 1, 86% at Site 2, 93% at Site 3, and 84% at Site 4) in the first error bracket $(-0.1 \leq |\mathrm{FE}| \leq 0.1)$. Although, the EEMD-Boruta-MARS recorded 1% more error values than hybridized EEMD-Boruta-ELM at Site 1 in the first two error bracket $(-0.2 \leq |\mathrm{FE}| \leq 0.2)$, the differences in percentages for the EEMD-Boruta-MARS were lower by 3% (Site 2) and 1% (Site 4) in comparison to the proposed hybrid EEMD-Boruta-ELM model. While at Site 3 both the hybrids EEMD-Boruta-ELM and EEMD-Boruta-MARS had 100% of errors within $-0.2 \leq |\mathrm{FE}| \leq 0.2$. Consequently, a better realization of the performance comparison was achieved with the Taylor diagram (Figure 8). Using a single figure, the Taylor diagram concisely provides an angular statistical summary of root-mean-square difference, correlation coefficients, and the ratio of the model's variances by plotting standard deviation against the correlations on the polar and radial axis, respectively (Taylor, 2001). Once more, at Sites 2, 3, and 4, the hybridized EEMD-Boruta-ELM model clearly outperformed attaining the closest proximity with respect to the observed statistics. At Sites 3 and 4, the hybrid EEMD-Boruta-ELM model yielded larger *r* with little variance and the EEMD-Boruta-MARS model yielded close results providing good contention. At Site 2-Cooinbil, however, the performance of EEMD-Boruta-MARS was very contrasting in comparison to hybrid EEMD-Boruta-ELM which had the optimal performance. Yet, at Site 1 the correlation noted from EEMD-Boruta-MARS were higher. On the other hand, the angular statistics of standalone ELM and MARS models were farther away than the expected/observed magnitude showing lower performances.

In terms of accuracy, the superiority of multivariate sequential EEMD decomposition prescribed *SM* models is clearly evident. However, the result ascertains that the performances of hybrid EEMD-Boruta-ELM (and the comparative EEMD-Boruta-MARS) model are not universally similar when the study sites with different geophysical and pedologic conditions are taken into account as in Tables 1 and 3; Figures 1 and 2. The geographical signature is reflected in the performance of the models at these sites with a varied range of statistics being conceded (Table 8; Figures 5-8).

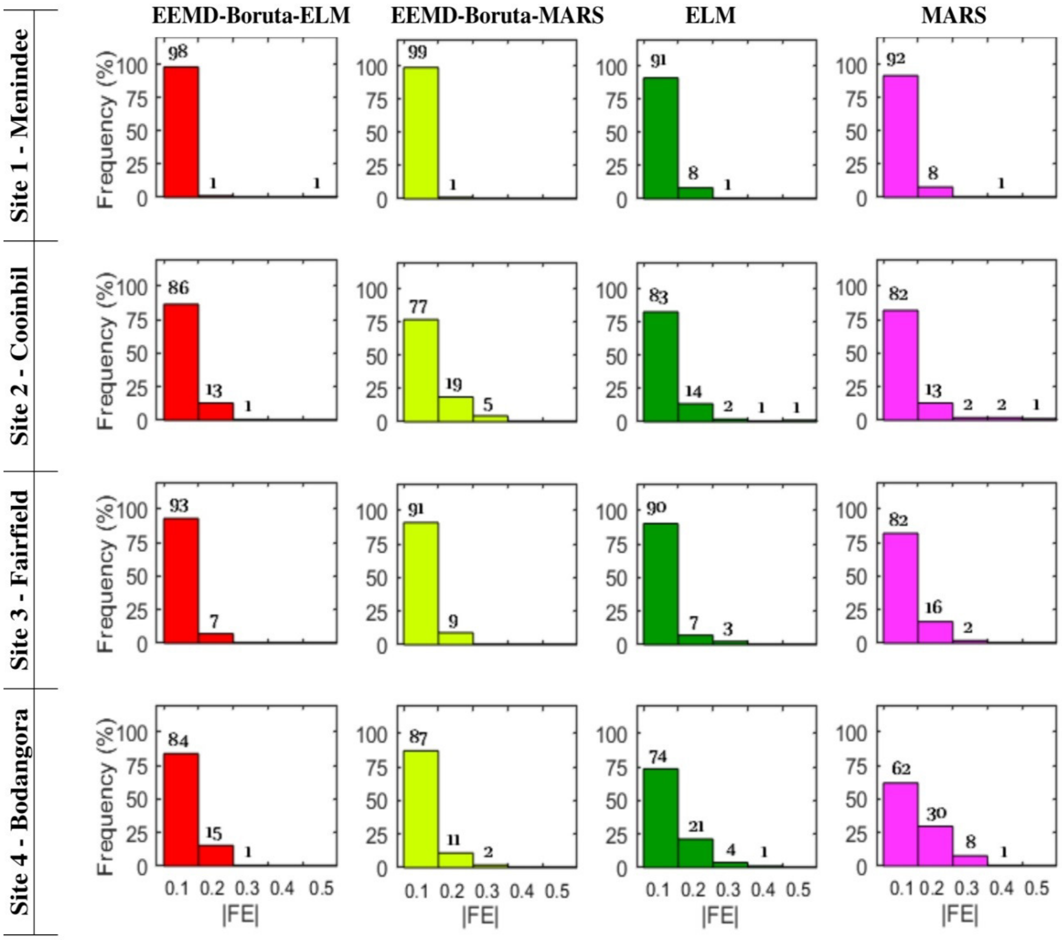


Figure 7 Histograms illustrating the percentage frequency of the absolute value of weekly forecasting error (|FE|) generated from the hybrid EEMD-Boruta-ELM, *vs.* the EEMD-Boruta-MARS, ELM, and MARS models.

To overcome the shortcoming of the absolute measures, Table 11 compares the alternative relative error measures (*i.e.*, $RRMSE$ and $MAPE$). Usually, the $RRMSE$ is used to categorize the model performances as “Excellent” ($RRMSE < 10\%$), “Good” ($10\% < RRMSE < 20\%$), “Fair” ($20\% < RRMSE < 30\%$) or “Poor” ($RRMSE \geq 30\%$) (Li et al., 2013). Following this classification, the performance of hybrid EEMD-Boruta-ELM model at all sites was “Fair” since the $RRMSE$ recorded by the hybrid EEMD-Boruta-ELM models at all sites were between 22.39% and 29.32%. While, the EEMD-Boruta-MARS models at Site 2 and 3 were “Poor” with $RRMSE$ values of 37.46% and 34.35%, respectively. Based on $MAPE$ the hybrid EEMD-Boruta-ELM model had the best performance at Site 4-

Bodangora (*MAPE* = 20.61% and *RRMSE* = 23.26%), followed by Site 1-Menindee. The corresponding model architecture of this best hybrid EEMD-Boruta-ELM model was 6-7-1 (input–hidden–output layer combinations) with the sigmoid transfer function (Table 6). Similarly, this site features Sodosol soil, with dry land cropping and has sub-tropical climatic characteristics (Table 1).

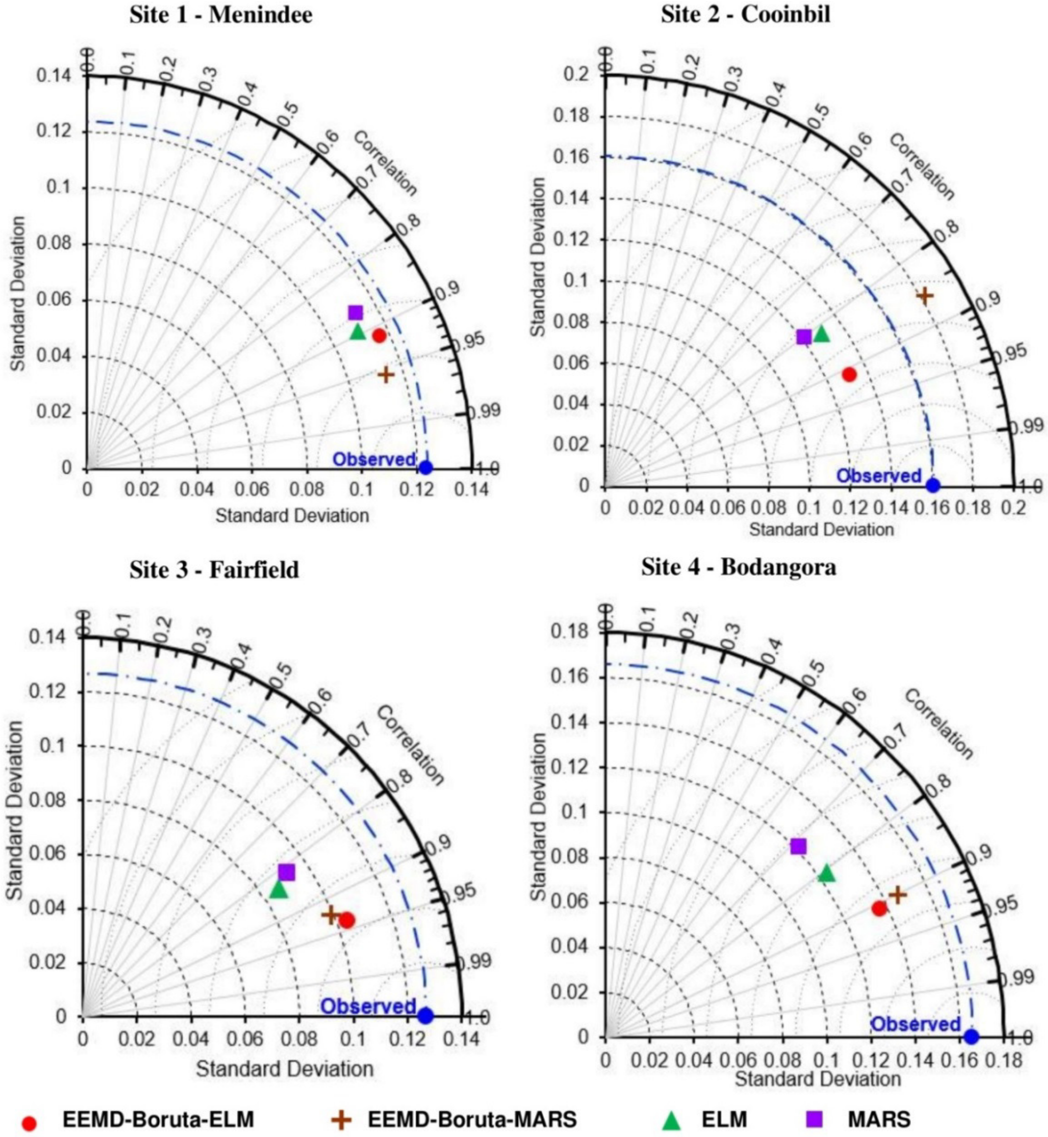


Figure 8 Taylor plots indicating the correlation coefficient and standard deviation (SD) in the testing phase based on the hybrid EEMD-Boruta-ELM, *vs.* the EEMD-Boruta-MARS, ELM and MARS models for forecasting weekly normalized soil moisture at the candidate study sites.

Table 11 Model comparison at different sites using relative error in testing phase: *RRMSE* and *MAPE*. The optimal model with lowest relative (%) error at each site has been shown in **boldface**.

	ELM		MARS	
Sites	*RRMSE* *(%)*	*MAPE* *(%)*	*RRMSE* *(%)*	*MAPE* *(%)*
	Site 1 - Menindee			
Standalone model	31.41	28.01	33.68	28.64
Hybridized: EEMD-Boruta	22.39	20.81	**20.27**	**18.71**
	Site 2 - Cooinbil			
Standalone model	37.27	37.99	38.91	37.72
Hybridized: EEMD-Boruta	**26.98**	**39.01**	37.46	43.43
	Site 3 - Fairfield			
Standalone model	42.97	58.98	45.73	72.62
Hybridized: EEMD-Boruta	**29.32**	**24.77**	34.35	32.43
	Site 4 - Bodangora			
Standalone model	32.03	34.73	37.44	43.11
Hybridized: EEMD-Boruta	**23.26**	**20.61**	23.74	29.19

Since two contesting modelling frameworks were used in this study, a comparison of standalone ELM and MARS models at all hydrological sites evidently showed that the ELM model outperformed the MARS models. The ELM model being a purely non-linear SLFN model was able to better extract the predictive features than the MARS model that is a combination of additive and/or interactive simple linear functions. For hybridized model comparisons, the superior performance of hybrid EEMD-Boruta-ELM models further illustrated that ELM is apparently more robust in simulating *SM* in comparison to the MARS model. For feature selection, the two-step method; *CCF* followed by Boruta were integral in determining the salient inputs. During the final feature selection stage, an equal opportunity was provided to both the models as the best Boruta ranked inputs were iteratively used in the modelling process. At Site 2, the hybridized EEMD-Boruta-ELM model required 7 significant variables while the EEMD-Boruta-MARS required 8 significant variables to reach its peak performance. Likewise, at Site 4 EEMD-Boruta-MARS model required 4 additional significant variables to reach optimum performance. Yet, the optimum performances of these MARS established models (*i.e.*, EEMD-Boruta-MARS) were significantly lower than the ELM established hybrid models (*i.e.*, EEMD-Boruta-ELM).

The outcomes also revealed that the self-adaptive MRA utility, EEMD, has proven to be a valuable tool for detecting and isolating non-linear signal properties into subseries *i.e.*, intrinsic mode functions (IMFs) and the residual component. This is evident from the improved forecasting accuracies of multivariate sequential EEMD based models (Tables 8-11; Figures 5-8). However, a common practice with EEMD forecasting is to utilize only significant lagged IMFs and residual of *SM* to forecast future *SM* values (Jiao *et al.*, 2016; Beltran-Castro *et al.*, 2013; Ouyang *et al.*, 2016; Bai *et al.*, 2015; Seo and Kim, 2016). This is the first study that clearly ascertains that in order to forecast weekly *SM*, more than lagged *SM* time series is required (Table 4) as the multivariate sequential EEMD approach outperformed the classical models (Tables 8-11; Figures 5-8). Table 4 depicts that a minimum of three significant inputs are needed in this multivariate EEMD forecasting. For instance, at Site 1, to forecast IMF 1 the inputs were SM_{t-2}, PCN_{t-1}, $SolarMJ_{t-1}$, and to forecast IMF 2 the inputs were SM_{t-1}, PCN_{t-1}, SM_{t-2}, while the forecasting of the residual component required *SM* only (SM_{t-2}, SM_{t-1}, SM_{t-3}). A similar result is seen at other sites as well (Table 4), however, interestingly at Site 3-Fairfield none of the lags of *SM* time-series were selected to forecast IMF 1 despite a large number (10) of inputs were required. So, if only *SM* were used in all these cases, many important predictive features may have been left out affecting the overall forecasting performances. Additionally, the *SM* forecasting was performed at a shorter forecast horizon to concur with the real-life decision support systems that require precise forecasting within a specified time limit (Yu *et al.*, 2011), which in this case was weekly. The short-time weekly forecasts would potentially allow for the development of such systems for near real-time agricultural and hydrological applications.

5.0 Future scope

Despite the efficacy of the approach proposed in this study, in further studies, the multivariate EEMD technique can be improved by combining the tool with the variational mode decomposition (VMD) into a two-stage decomposition scheme as demonstrated by Wang, Deyun *et al.* (2017) and Wang, D. *et al.* (2017). Studies with the variants of EMD including complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) (Torres *et al.*, 2011), improved complete ensemble empirical mode decomposition with adaptive noise (ICEEMDAN) (Colominas *et al.*, 2014) and empirical wavelet transform (EWT) (Kedadouche *et al.*, 2016; Peng *et al.*, 2017) could also provide greater insight into the performance of the multivariate ensemble modelling approach. With that, alternative optimization techniques in terms of the feature determination could also be explored. For instance, iterative input selection (IIS) (Galelli and Castelletti, 2013), Neighborhood Component Analysis feature selection for regression algorithm (*fsrnca*) (Yang *et al.*, 2012), modified minimum redundancy maximum relevance (mMRMR) algorithm (Hejazi and Cai, 2009), joint mutual information maximisation feature selection (JMIM) (Bennasar *et al.*, 2015) or bootstrap rank-ordered conditional mutual information (broCMI) (Quilty *et al.*, 2016) can be explored as alternative tools to improve the proposed method. Moreover, an extensive set of model inputs could be incorporated from various sources like several atmospheric parameters derived from satellite measurements, Interim ERA European Centre for Medium-Range Weather Forecasting (ECMWF) (Dee *et al.*, 2011) and climate indices (*e.g.*, Sea Surface Temperatures-SSTs, Dipole Mode Index-DMI, Pacific Decadal Oscillation-PDO, etc.) of longer time series than investigated in this paper.

6.0 Conclusions

Multivariate sequential EEMD scheme is proposed to address naturally embedded non-stationary features within multivariate hydro-meteorological predictor inputs in forecasting weekly soil moisture. Optimization with Boruta feature selection algorithm led to the establishment of hybrid EEMD-Boruta-ELM model. The performance of the hybrid multivariate sequential EEMD-Boruta-ELM model in emulating weekly soil moisture at four hydrological sites within Australian MDB was benchmarked with a comparative multivariate sequential EEMD-Boruta-MARS and standalone ELM and MARS models. Using an independent testing dataset, several numerical assessment metrics were determined including Pearson's correlation coefficient (r), mean absolute error (MAE), root mean square error ($RMSE$), Willmott's Index (WI), the Nash-Sutcliffe coefficient (E_{NS}), the Legates-McCabe Index (L) and relative errors (*i.e.*, $MAPE$ and $RRMSE$).

The performance of the hybridized multivariate sequential EEMD-Boruta-ELM model at all sites was "Fair" with $22.39\% \leq RRMSE \leq 29.32\%$. The best hybrid multivariate sequential EEMD-Boruta-ELM model was established at Site 4-Bodangora ($MAPE$ = 20.61% and $RRMSE$ = 23.26%) that featured Sodosol soil, with dry land cropping and has sub-tropical climatic characteristics. A comprehensive evaluation via numerical assessment metrics and diagnostic plots revealed that the hybrid multivariate sequential EEMD-Boruta-ELM model outperformed the comparative multivariate sequential EEMD-Boruta-MARS and classical models in forecasting soil moisture at a weekly forecasting horizon.

Further improvements in this multivariate ensemble modelling approach could be achieved by cascading it into a two-stage decomposition scheme using VMD process. In addition, different EMD variants could be considered with various feature selection algorithms and a larger dataset need to be explored in future independent studies. Despite these limitations, the hybrid EEMD-Boruta-ELM model proved to be an effective tool in capturing the non-linear dynamics in forecasting weekly soil moisture. This hybrid multivariate sequential EEMD-Boruta-ELM model can amicably be embedded in designing of hydrological and precision agricultural (PA) applications as well as for drainage mapping, drought stress identification, crop yield prediction and other emerging intelligent and autonomous systems.

References

Adamowski, J., 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., 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).

Ahila, R., Sadasivam, V., Manimala, K., 2015. An integrated PSO for parameter determination and feature selection of ELM and its application in classification of power system disturbances. Applied Soft Computing 32, 23-37.

Argent, R.M., Western, A.W., Lillc, A., 2015. Towards operational forecasting of agricultural soil water in Australia., 21st International Congress on Modelling and Simulation, Gold Coast, Australia, pp. 2437-2443.

ASRIS, 2014. Australian Soil Resource Information System. Department of Agriculture, Fisheries, and Forestry.

Australian Bureau of Statistics, 2008. Water and the Murray-Darling Basin - A Statistical Profile, 2000-01 to 2005-06

Australian Bureau of Statistics, 2017. Agricultural census fact sheet: Australia, states, and territories. ABS, Canberra, Australia.

Bai, Y., Wang, P., Xie, J., Li, J., Li, C., 2015. Additive Model for Monthly Reservoir Inflow Forecast. Journal of Hydrologic Engineering 20(7), 04014079.

Basha, G., Ouarda, T.B.M.J., Marpu, P.R., 2015. Long-term projections of temperature, precipitation and soil moisture using non-stationary oscillation processes over the UAE region. International Journal of Climatology 35(15), 4606-4618.

Beltran-Castro, J., Valencia-Aguirre, J., Orozco-Alzate, M., Castellanos-Domınguez, G., Travieso-Gonzalez, C.M., 2013. Rainfall forecasting based on ensemble empirical mode decomposition and neural networks. In: I. Rojas, G. Joya, J. Cabestany (Eds.), Advances in Computational Intelligence: 12th International Work-

Conference on Artificial Neural Networks. Springer, Puerto de la Cruz, Tenerife, Spain, pp. 471-480.

Bennasar, M., Hicks, Y., Setchi, R., 2015. Feature selection using Joint Mutual Information Maximisation. Expert Systems with Applications 42(22), 8520-8532.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising performance of environmental models. Environmental Modelling & Software 40, 1-20.

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 Development 7(3), 1247-1250.

Chau, K., Wu, C., 2010. A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. Journal of Hydroinformatics 12(4), 458-473.

Chen, J., Heincke, B., Jegen, M., Moorkamp, M., 2012. Using empirical mode decomposition to process marine magnetotelluric data. Geophysical Journal International 190(1), 293-309.

Chitsaz, N., Azarnivand, A., Araghinejad, S., 2016. Pre-processing of data-driven river flow forecasting models by singular value decomposition (SVD) technique. Hydrological Sciences Journal 61(12), 2164-2178.

Christa, M., Kempa-Liehrb, A.W., Feindta, M., 2017. Distributed and parallel time series feature extraction for industrial big data applications. Neurocomputing.

Colominas, M.A., Schlotthauer, G., Torres, M.E., 2014. Improved complete ensemble EMD: A suitable tool for biomedical signal processing. Biomedical Signal Processing and Control 14, 19-29.

Cornish, C.R., Bretherton, S., Percival, D.B., 2005. Maximal OverlapWavelet Statistical Analysis with Application to atmospheric turbulence. Kluwer Academic Publishers, Netherlands.

Craven, P., Wahba, G., 1979. Smoothing noisy data with spline functions. Estimating the correct degree of smoothing by the method of generalized cross-validation. Numerische Mathematik 31(377-403).

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.J., Park, B.K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.N., Vitart, F., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Quarterly Journal of the Royal Meteorological Society 137(656), 553-597.

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. Atmospheric Research 184, 149-175.

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), 90.

Deo, R.C., Tiwari, M.K., Adamowski, J.F., Quilty, J.M., 2016a. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. Stochastic Environmental Research and Risk Assessment.

Deo, R.C., Tiwari, M.K., Adamowski, J.F., Quilty, M.J., 2017b. Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model. Stochastic Environmental Research and Risk Assessment 31(5), 1211–1240.

Deo, R.C., Wen, X., Qi, F., 2016b. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. Applied Energy 168, 568-593.

Department of Agriculture and Water Resources, 2015. Catchment Scale Land Use of Australia. Agricultural Land Management, Australia.

Dghais, A.A.A., Ismail, M.T., 2013. A Comparative Study between Discrete Wavelet Transform and Maximal Overlap Discrete Wavelet Transform for Testing

Stationarity. International Journal of Mathematical, Computational, Physical, Electrical, and Computer Engineering 12(7), 1677-1681.

Di, C., Yang, X., Wang, X., 2014. A Four-Stage Hybrid Model for Hydrological Time Series Forecasting. PLoS ONE 9(8), 1-18.

Elshorbagy, A., Parasuraman, K., 2008. On the relevance of using artificial neural networks for estimating soil moisture content. Journal of Hydrology 362(1-2), 1-18.

Friedman, J.H., 1991. Multivariate adaptive regression splines. The Annals of Statistics 19(1), 1-141.

Friedman, J.H., Silverman, B.W., 1989. Flexible parsimonious smoothing and additive modeling. Technometrics 31(1), 3-21.

Galelli, S., Castelletti, A., 2013. Tree-based iterative input variable selection for hydrological modeling. Water Resources Research 49(7), 4295-4310.

Hejazi, M.I., 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.

Hijmans, R.J., Cameron, S.E., Parra, J.L., Jones, P.G., Jarvis, A., 2005. Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25(15), 1965-1978.

Hu, T., Wu, F., Zhang, X., 2007. Rainfall–runoff modeling using principal component analysis and neural network. Hydrology Research 38(3), 235-248.

Huang, C., Li, L., Ren, S., Zhou, Z., 2010. Research of Soil Moisture Content Forecast Model Based on Genetic Algorithm BP Neural Network. In: D. Li, Y. Liu, Y. Chen (Eds.), 4th Conference on Computer and Computing Technologies in Agriculture (CCTA). Computer and Computing Technologies in Agriculture IV. Springer, International Federation for Information Processing (IFIP) Advances in Information and Communication Technology, AICT-345, Nanchang, China, pp. 309-316.

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2004. Extreme learning machine: a new learning scheme of feedforward neural networks, 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), pp. 985-990.

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2006. Extreme learning machine: Theory and applications. Neurocomputing 70(1-3), 489-501.

Huang, G., Huang, G.B., Song, S., You, K., 2015. Trends in extreme learning machines: a review. Neural Netw 61, 32-48.

Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., Yen, N.-C., Tung, C.C., Liu, H.H., 1998. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of Royal Society A 454, 903–995.

Hur, J.-H., Ihm, S.-Y., Park, Y.-H., 2017. A Variable Impacts Measurement in Random Forest for Mobile Cloud Computing. Wireless Communications and Mobile Computing 2017, 1-13.

Islam, S., Engman, T., 1996. Why Bother for 0.0001% of Earth's Water-Challenges for Soil Moisture Research. Eos, Trans. Am. Geophys. Union. 77(43), 420.

Jekabsons, G., 2016. Adaptive Regression Splines toolbox for Matlab/Octave, ver. 1.13.0, pp. 1-33.

Jiao, G., Guo, T., Ding, Y., 2016. A New Hybrid Forecasting Approach Applied to Hydrological Data: A Case Study on Precipitation in Northwestern China. Water 8(9), 367.

Kedadouche, M., Thomas, M., Tahan, A., 2016. A comparative study between Empirical Wavelet Transforms and Empirical Mode Decomposition Methods: Application to bearing defect diagnosis. Mechanical Systems and Signal Processing 81, 88-107.

Kisi, O., Shiri, J., Tombul, M., 2013. Modeling rainfall-runoff process using soft computing techniques. Computers & Geosciences 51, 108-117.

Kornelsen, K.C., Coulibaly, P., 2014. Root-zone soil moisture estimation using data-driven methods. Water Resources Research 50(4), 2946-2962.

Krishna, B., Rao, Y.R.S., Nayak, P.C., 2011. Time Series Modeling of River Flow Using Wavelet Neural Networks. Journal of Water Resource and Protection 03(01), 50-59.

Kursa, M.B., Jankowski, A., Rudnicki, W.R., 2010. Boruta – A System for Feature Selection. Fundamenta Informaticae 101, 271–285.

Kursa, M.B., Rudnicki, W.R., 2010. Feature Selection with the Boruta Package. Journal of Statistical Software 36(11).

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation. Water Resources Research 35(1), 233-241.

Legates, D.R., McCabe, G.J., 2013. A refined index of model performance: a rejoinder. International Journal of Climatology 33(4), 1053-1056.

Leutner, B.F., Reineking, B., Müller, J., Bachmann, M., Beierkuhnlein, C., Dech, S., Wegmann, M., 2012. Modelling Forest α-Diversity and Floristic Composition — On the Added Value of LiDAR plus Hyperspectral Remote Sensing. Remote Sensing 4(12), 2818-2845.

Li, B., Chen, Z., Yuan, X., 2015. The nonlinear variation of drought and its relation to atmospheric circulation in Shandong Province, East China. PeerJ 3, e1289.

Li, J., Alvarez, B., Siwabessy, J., Tran, M., Huang, Z., Przeslawski, R., Radke, L., Howard, F., Nichol, S., 2017. Application of random forest, generalised linear model and their hybrid methods with geostatistical techniques to count data: Predicting sponge species richness. Environmental Modelling & Software 97, 112-129.

Li, J., Tran, M., Siwabessy, J., 2016. Selecting Optimal Random Forest Predictive Models: A Case Study on Predicting the Spatial Distribution of Seabed Hardness. PLoS One 11(2), e0149089.

Li, M.-F., Tang, X.-P., Wu, W., Liu, H.-B., 2013. General models for estimating daily global solar radiation for different solar radiation zones in mainland China. Energy Conversion and Management 70, 139-148.

Liu, Y., Mei, L., Ooe, S.K., 2014. Prediction of soil moisture based on extreme learning machine for an apple orchard, Conference on Computational Interdisciplinary Science. IEEE, pp. 400-404.

Londhe, S.N., Dixit, P.R., 2012. Forecasting Stream Flow using Support Vector Regression and M5 model Trees. International Journal of Engineering Research and Development 2(5), 1-12.

Loon, A.F.V., Laaha, G., 2015. Hydrological drought severity explained by climate and catchment characteristics. Journal of Hydrology 526, 3-14.

Lyu, B., Zhang, Y., Hu, Y., 2017. Improving PM2.5 Air Quality Model Forecasts in China Using a Bias-Correction Framework. Atmosphere 8(12).

Mallat, S.G., 1989. A theory for multiresolution signal decomposition: the wavelet representation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 11(7), 674-693.

Mallat, S.G., 1998. A wavelet tour of signal processing. Academic, New York.

Matei, O., Rusu, T., Petrovan, A., Mihuţ, G., 2017. A Data Mining System for Real Time Soil Moisture Prediction. Procedia Engineering 181, 837-844.

Moriasi, D.N., Arnold, J.G., Liew, M.W.V., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. American Society of Agricultural and Biological Engineers 50(3), 885–900.

Mouatadid, S., Adamowski, J., 2016. Using extreme learning machines for short-term urban water demand forecasting. Urban Water Journal, 1-9.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. Journal of Hydrology 10(3), 282-290.

Nourani, V., Baghanam, A.H., Adamowski, J., Kisi, O., 2014. Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review. J. Hydrol. 514, 358-377.

Nourani, V., Komasi, M., Mano, A., 2009. A multivariate ANN-wavelet approach for rainfall–runoff modeling. Water. Resour. Manag. 23(14), 2877-2894.

NSW-Department of Industry, 2017. Export from New South Wales: Agribusiness and food. New South Wales: Department of Industry, NSW, Australia.

Ouyang, Q., Lu, W., Xin, X., Zhang, Y., Cheng, W., Yu, T., 2016. Monthly Rainfall Forecasting Using EEMD-SVR Based on Phase-Space Reconstruction. Water Resources Management 30(7), 2311-2325.

Peng, T., Zhou, J., Zhang, C., Fu, W., 2017. Streamflow Forecasting Using Empirical Wavelet Transform and Artificial Neural Networks. Water 9(6), 406.

Percival, D.B., Lennox, S.M., Wang, Y.G., Darnell, R.E., 2011. Wavelet-based multiresolution analysis of Wivenhoe Dam water temperatures. Water Resources Research 47(5).

Poona, N.K., Ismail, R., 2014. Using Boruta-Selected Spectroscopic Wavebands for the Asymptomatic Detection of <italic>Fusarium Circinatum</italic> Stress. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 7(9), 3764-3772.

Prasad, R., Deo, R.C., Li, Y., Maraseni, T., 2017. Input selection and performance optimization of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm. Atmospheric Research 197, 42-63.

Prasad, R., Deo, R.C., Li, Y., Maraseni, T., 2018a. Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors. Soil and Tillage Research 181, 63-81.

Prasad, R., Deo, R.C., Li, Y., Maraseni, T., 2018b. Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. Geoderma 330, 136-161.

Qiu, X., Ren, Y., Suganthan, P.N., Amaratunga, G.A.J., 2017. Empirical Mode Decomposition based ensemble deep learning for load demand time series forecasting. Applied Soft Computing 54, 246-255.

Quilty, J., Adamowski, J., Khalil, B., Rathinasamy, M., 2016. Bootstrap rank-ordered conditional mutual information (broCMI): A nonlinear input variable selection method for water resources modeling. Water Resources Research 52(3), 2299-2326.

Rathinasamy, M., Khosa, R., Adamowski, J., ch, S., Partheepan, G., Anand, J., Narsimlu, B., 2014. Wavelet-based multiscale performance analysis: An approach to assess and improve hydrological models. Water Resources Research 50(12), 9721-9737.

Raupach, M.R., Briggs, P.R., Haverd, V., King, E.A., Paget, M., Trudinger, C.M., 2009. Australian water availability project (AWAP)-CSIRO marine and atmospheric research component-final report for phase 3. CAWCR Technical Report No. 013.

Raupach, M.R., Briggs, P.R., Haverd, V., King, E.A., Paget, M., Trudinger, C.M., 2012. Australian Water Availability Project. CSIRO Marine and Atmospheric Research, Canberra, Australia. http://www.csiro.au/awap.

Ren, Y., Suganthan, P.N., Srikanth, N., 2015. A comparative study of empirical mode decomposition-based short-term wind speed forecasting methods. IEEE Transactions on Sustainable Energy 6(1), 236–244.

Roy, K., Das, R.N., Ambure, P., Aher, R.B., 2016. Be aware of error measures. Further studies on validation of predictive QSAR models. Chemometrics and Intelligent Laboratory Systems 152, 18-33.

Seo, Y., Kim, S., 2016. Hydrological Forecasting Using Hybrid Data-Driven Approach. American Journal of Applied Sciences 13(8), 891-899.

Shamseldin, A.Y., 1997. Application of a neural network technique to rainfall runoff. Journal of Hydrology 199, 272-294.

Shamshirband, S., Mohammadi, K., Tong, C.W., Petković, D., Porcu, E., Mostafaeipour, A., Ch, S., Sedaghat, A., 2015. Application of extreme learning machine for estimation of wind speed distribution. Climate Dynamics 46(5-6), 1893-1907.

Sharpley, R.C., Vatchev, V., 2005. Analysis of the Intrinsic Mode Functions. Constructive Approximation 24(1), 17-47.

Strobl, C., Boulesteix, A.L., Kneib, T., Augustin, T., Zeileis, A., 2008. Conditional variable importance for random forests. BMC Bioinformatics 9, 307.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research: Atmospheres 106(D7), 7183–7192.

The Murray–Darling Basin Authority, 2010. Guide to the proposed Basin Plan Technical background Part 1.

Tiwari, M., Adamowski, J., Adamowski, K., 2016. Water demand forecasting using extreme learning machines. Journal of Water and Land Development 28(1).

Tiwari, M.K., Adamowski, J., 2013. Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models. Water Resources Research 49(10), 6486-6507.

Torres, M.E., Colominas, M.A., Schlotthauer, G., Flandrin, P., 2011. A complete ensemble empirical mode decomposition with adaptive noise, 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 4144-4147.

Van Loon, A.F., 2015. Hydrological drought explained. Wiley Interdisciplinary Reviews: Water 2(4), 359-392.

Wang, D., Luo, H., Grunder, O., Lin, Y., Guo, H., 2017a. Multi-step ahead electricity price forecasting using a hybrid model based on two-layer decomposition technique and BP neural network optimized by firefly algorithm. Applied Energy 190, 390-407.

Wang, D., Wei, S., Luo, H., Yue, C., Grunder, O., 2017b. A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine. Sci Total Environ 580, 719-733.

Wang, W.-c., Xu, D.-m., Chau, K.-w., Chen, S., 2013. Improved annual rainfall-runoff forecasting using PSO–SVM model based on EEMD. Journal of Hydroinformatics 15(4), 1377-1390.

Wang, W.C., Chau, K.W., Qiu, L., Chen, Y.B., 2015. Improving forecasting accuracy of medium and long-term runoff using artificial neural network based on EEMD decomposition. Environ Res 139, 46-54.

Willmott, C.J., 1981. On the validation of models. Physical Geography 2, 184-194.

Willmott, C.J., 1984. On the evaluation of model performance in physical geography. In: G.L. Gaile, C.J. Willmott (Eds.), Spatial Statistics and Models. Springer, pp. 443-460.

Wu, Z., Huang, N.E., 2009. Ensemble empirical mode decomposition: A noise-assisted data analysis method. Advances in 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. Climate Dynamics 37(3-4), 759-773.

Xiaoxia, Y., Chengming, Z., 2016. A soil moisture prediction algorithm base on improved BP, 2016 Fifth International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Tianjin, China.

Xu, S., Wang, J., 2016. A fast incremental extreme learning machine algorithm for data streams classification. Expert Systems with Applications 65, 332-344.

Yang, W., Wang, K., Zuo, W., 2012. Neighborhood Component Feature Selection for High-Dimensional Data. Journal of Computers 7(1).

Yang, X., Zhang, C., Cheng, Q., Zhang, H., Gong, W., 2017. A Hybrid Model for Soil Moisture Prediction by Using Artificial Neural Networks. Revista de la Facultad de Ingeniería U.C.V. 32(5), 265-271.

Yaseen, Z.M., Jaafar, O., Deo, R.C., Kisi, O., Adamowski, J., Quilty, J., El-Shafie, A., 2016. Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq. Journal of Hydrology 542, 603-614.

Yu, Y., Choi, T.-M., Hui, C.-L., 2011. An intelligent fast sales forecasting model for fashion products. Expert Systems with Applications 38(6), 7373-7379.

Zaman, B., McKee, M., 2014. Spatio-Temporal Prediction of Root Zone Soil Moisture Using Multivariate Relevance Vector Machines. Open Journal of Modern Hydrology 04(03), 80-90.

Zhang, B., Govindaraju, R.S., 2000. Prediction of watershed runoff using Bayesian concepts and modular neural networks. Water Resources Research 36(3), 753-762.

Zhang, S., Shao, M., Li, D., 2017. Prediction of soil moisture scarcity using sequential Gaussian simulation in an arid region of China. Geoderma 295, 119-128.

Chapter 7: Synthesis and future scope

7.1 Synthesis

In this study, an attempt is made to advance the science of hydrological prediction by developing accurate and high precision data intelligent models using hybridized machine learning or computational intelligence techniques for streamflow water level and soil moisture forecasting within the Murray-Darling Basin, Australia. The streamflow water level was forecasted at monthly forecast horizon. While the soil moisture forecasting was commenced from monthly down to weekly forecast horizon to realize near real-time forecasting. In improving the hydrological forecasting of streamflow water level and the upper and lower layer soil moisture levels, hybridized models were developed with new methodological approaches. The machine learning algorithms that were utilized to design the hybrid models included, 2nd order Volterra, M5 Model Tree, random forest, multivariate adaptive regression splines, extreme learning machine, and the artificial neural networks.

Two important issues were addressed in this study i) the problem of selection of non-redundant predictor inputs from sets of multivariate input in hydrological forecasting and ii) non-stationarity and non-linearity issue. Input selection algorithms resolved the first issue of feature optimization, while the latter issue was resolved by time-scale multi-resolution representation of the respective hydrological input time series.

In the first objective (Chapter 3), the iterative input selection algorithm screened the salient inputs. Then an advanced and non-decimated wavelet transformation known as the maximum overlap discrete wavelet transformation (MODWT) was utilized in addressing non-stationarity problem whilst designing high precision streamflow water level forecasting at monthly forecast horizons. Hybridization led to the formation of IIS-W-ANN model that outperformed the comparative M5 Tree based model (IIS-W-M5 Tree), IIS-ANN, IIS-M5 Tree and the standalone ANN and M5 Tree models.

In addition, two self-adaptive techniques that do not require any basis function or pre-defined mother wavelet were utilized in Chapter 4 to further address the non-stationarity and non-linearity issues (Objective 2). This included the ensemble empirical mode decomposition (EEMD) and complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN). The monthly upper and lower layer soil moisture time series were resolved using these multi-resolution utilities into intrinsic mode functions and a residual component. Then partial-auto correlation function (*PACF*) was utilized to determine the input lagged sub-series in the designing of the subsequent hybrid machine learning approaches (EEMD-ELM and CEEMDAN-ELM). The EEMD-ELM model was found to have better performances in emulating upper and lower soil moisture compared to the CEEMDAN-ELM, and the random forest-based hybrid models (EEMD-RF, CEEMDAN-RF) and the standalone ELM and RF models.

Moreover, for Objective 3 an ensemble model combination method called the ensemble committee of models based on ANN (ANN-CoM) was developed and explored for its preciseness in emulating upper and lower layer soil moisture using sixty potential inputs (Chapter 5). A two-stage feature optimization was formulated by employing the Neighbourhood Component Analysis based regression feature (*fsrnca*) selection algorithm and a basic ELM model with *sine* transfer function and 50 hidden neurons. In this objective, the ANN-CoM model was found to outperform the standalone second order Volterra, M5 Model Tree, random forest and ELM models in monthly upper and lower layer soil moisture forecasting.

Finally, in the fourth objective (Chapter 6), forecasting of near-real-time *i.e.*, weekly soil moisture levels was achieved by designing and employing a novel multivariate sequential EEMD approach. This technique was developed to permit the utilization of multiple predictor inputs in EEMD-based modelling approaches. A total of thirteen predictor inputs were collated with two-stage feature optimization via cross-correlation function (*CCF*) followed by Boruta wrapper algorithm was adopted in developing the hybrid multivariate sequential EEMD-Boruta-ELM. The EEMD-Boruta-ELM proved to be better in forecasting weekly soil moisture in comparison to the MARS counterpart (EEMD-Boruta-MARS) and the standalone ELM and MARS models.

The findings clearly showed improved performances of hybridized models developed with respect to standalone counterparts. A further elaboration of the research outcomes are as follows:

a) Feature selections

- The iterative input selection (IIS) algorithm served as an important input determination procedure since enhanced performance by the IIS-based models in forecasting of streamflow water level was lucid. The key important feature of IIS is that the algorithm iteratively selects the most significant inputs in a forward selection method then uses an extra trees-based ranking method to estimate the relative contribution of each candidate input(s) in explaining the output. Finally, the underlying model assesses the effectiveness of each input variable in predicting the output and successively adds the most significant ones in multiple input single output (MISO) technique. When the performance of the MISO model starts to decrease, the IIS algorithm stops executing and the most significant inputs averred at this epoch.
- The *fsrnca* feature selection also played a key role. It utilizes an embedded Neighbourhood component analysis model to determine the relative feature weights of each of the inputs in emulating the target variable. Yet, *fsrnca* only provides respective feature weights and an additional add-on such as basic ELM that was developed in this study is required to determine the threshold feature weight in order to obtain an optimal set of inputs.
- With that, the Boruta input selection is a wrapper algorithm that uses random forest model as the underlying learning algorithm. The principal of Boruta input selection method is the maximum-optimal feature selection that generates a large predictor set. Hence, a reduction of inputs becomes necessary which in this study was achieved by a stepwise model building process.

Feature selection or input determination is a critical process in the development of data-intelligent models. Appropriate feature selection is necessary in order to develop parsimonious yet peak performing models.

Another additional characteristic noted during feature selection for streamflow forecasting (Objective 1 – Chapter 3) and soil moisture forecasting (Objectives 2 & 3 – Chapter 4 & 5) was that precipitation, which is traditionally understood to be an important indicator of hydrological parameters (*SWL* and *SM*), was not selected as the significant input for all study sites.

For *SWL*, soil type could probably have played an important role in soil water retention and subsequently affecting the runoff and streamflow. For instance, if the catchment is largely consisting of sandy soil type that has high drainage coefficient, the response time of rainfall to runoff is reduced. In other words, there is bound to be a higher direct correlation between rainfall and *SWL* at that site. However, if the catchment has a mixture of soil types and more of clay and loamy soil that has more water retention capacity, then there will be low dependence of *SWL* on rainfall.

On the other hand, for soil moisture forecasting study, the vegetation cover and land-use are also major determinants of *SM*. If the land is bare, then solar radiation will directly cause evaporation, which would be a key influential factor. For regions that are shaded by vegetation, they could have a greater chance of retaining water due to rainfall and hence *SM* would have more dependence on rainfall.

In this study, geographically distinct sites with dissimilar hydro-physical features were selected for robust model evaluations.

b) Multi-resolution analysis utilities

i) Maximum overlap discrete wavelet transformation (MODWT)

- Integration of MODWT multi-resolution analysis into the proposed hybrid ANN models (IIS-W-ANN) substantially improved the performances in forecasting of monthly streamflow water level.
- Determination of apt mother wavelet improves the performance of MODWT-hybrid models. Daubechies wavelet was adopted in this study. The db2 (2-vanishing moment) did not yield satisfactory results, while db3 (3-vanishing moment) and db4 (4-vanishing moment) were the most effective ones.
- The determination of optimal wavelet decomposition level is also vital and is usually done using a formula (Chapter 3: Eq. 7). The formula

determined level was three, yet in this study, four decomposition levels achieved better forecasting accuracy. The wavelet decomposition level is apparently data-dependent and a cautious approach is required when determining the optimal level. Different numbers of decomposition via trial and error may be tested.

Overall, the MODWT feature resolving is indeed beneficial, provided optimal mother wavelet and decomposition levels are selected, which essentially are dependent on the nature of the data-set.

ii) Ensemble empirical mode decomposition (EEMD) and Complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN)

- Both EEMD and CEEMDAN are self-adaptive multi-resolution methods hence the number of IMFs and residual component (*i.e.*, resolved frequencies) are contingent upon the embedded features within the data sets.
- Both EEMD and CEEMDAN improved the model performances with respect to the standalone models.
- EEMD models outperformed the CEEMDAN model in forecasting both upper and lower layer soil moisture values. The EEMD-ELM outperformed the alternative models at four (out of the seven) sites applied for upper layer *SM* forecasts and the hybrid the EEMD-ELM model was superior at all seven sites for the lower layer soil moisture forecasts.

The key benefit of the self-adaptive MRA tool, EEMD, integrated with ELM is that the hybrid EEMD-ELM model requires trivial human interventions. This has the prospects 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.

7.2 Novel contributions of the study

This study makes novel contributions in the development of data-intelligent predictive models for hydrological forecasting. In addition to the development of hybridized data-intelligent models, the further novel methodological improvements are as follows:

i) Ensemble committee of models

A new committee-based modelling approach has been developed in this study, which is a major contribution. Generally, model combinations are lacking in the hydrological and environmental applications.

- The ANN-based ensemble committee of models (ANN-CoM) was able to achieve better performance than the non-ensemble counterparts.
- The ANN-CoM was able to further optimize and stabilize the forecasts, since ANN generated appropriate weights, rather than simply averaging out the magnitudes, in forecasting of upper and lower layer soil moisture.
- This ensemble committee of models have huge potential and could possibly be integrated with the global climate models.

ii) Multivariate sequential ensemble empirical mode decomposition

- This is an important new development in this study since previously EEMD was only used as a single-variable forecasting tool.
- The forecasting performance increased with the integration of multivariate sequential EEMD hybridization approach to forecast weekly soil moisture values.

- ➢ Another important finding is that the predictive performances of hydrological models are highly data-sensitive and site dependent due to geographical influences.
- ➢ In addition, near-real-time forecasting was achieved with a gradual reduction in the forecast horizon from monthly to weekly (near-real-time) with evaluation of respective models at the shorter forecast horizon.

The innovative new 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.

7.3 Limitations of the current study and recommendations for future research

This subsection outlines the limitations of the current study and suggests recommendations that need to be properly addressed in future independent studies.

- In streamflow water level forecasting study, the key limitation was the unavailability of concurrently recorded streamflow water level and hydro-meteorological data at the same hydrological station. In future studies, the use of concurrently observed data is recommended that may improve the accuracy of the respective models.
- Monthly forecasting horizon was adopted in streamflow water level forecasting study while down to weekly forecasting was carried out for soil moisture level. However, for operational applications, testing with smaller time-steps such as daily, and hourly are recommended that can provide greater understanding.
- Individual forecasts of high, moderate and low streamflow events and *SM* level events could also be explored independently.
- Integration of add-on optimizer algorithms (*e.g.*, particle swarm optimization (PSO), firefly optimizer algorithm (FFA), or ant colony optimization (ACO)) could also be applied in these hydrological models.
- Studies with additional multi-resolution analysis utilities such as improved complete ensemble empirical mode decomposition with adaptive noise (ICEEMDAN), empirical wavelet transformation (EWT) and variational mode decomposition (VMD) are also suggested, which could provide greater insight into the performance of such data-intelligent hydrological models.
- Alternative feature selection algorithms, modified minimum redundancy maximum relevance (mMRMR) algorithm, joint mutual information maximization feature selection (JMIM) or bootstrap rank-ordered conditional mutual information (broCMI) can further be explored.
- In soil moisture forecasting, the hydro-meteorological data at a horizontal resolution of 5 km × 5 km from AWAP was utilized. An increase in data resolution may assist in better forecasting at localized farm levels. Studies pertaining to higher spatial resolution are also recommended.

- Incorporation of additional potential predictor variables and satellite-based data such as from Giovanni or MODIS are also recommended in future forecasting studies.
- For data analysis, the scatter plots with regression lines were presented in this study, which is one way of evaluating the model. An alternative method is recommended in further studies, whereby an X = Y line [also known as the 1:1 line or the 45° line) is to be drawn together with the linear regression fitting line on the respective scatterplots. In order to extract the pertinent information on over and under predictions, it is also recommended to compute the percentage deviations of the forecasted values from this 1:1 line.

In closing, this study has made novel contributions towards the practical problem of hydrological forecasting using hybridized machine learning techniques. The easy-to-implement, hybridized machine learning data-intelligent forecasting models used in this study have high computational efficiency, and low latency. This could revolutionize the streamflow water level and soil moisture level modelling and forecasting, concurrently serving as an important contrivance for water resource management and agricultural management applications.

References

Note that the references presented here do not include the references from the published articles (Chapters 3 to 5) and the submitted manuscript (Chapter 6). These references are provided in the reference sections of the respective articles.

Abawi, G, Dutta, S, Harris, T, Ritchie, J, Rattray, D and Crane, A 2000, 'The Use of Seasonal Climate Forecasts in Water Resources Management', in *Proceedings of Hydro 2000: Interactive Hydrology*, Barton, A.C.T.: Institution of Engineers, Australia, pp. 447-455. http://search.informit.com.au/documentSummary;dn=295458320665164;res=IELENG>.

Abram, NJ, Gagan, MK, Cole, JE, Hantoro, WS and Mudelsee, M 2008, 'Recent intensification of tropical climate variability in the Indian Ocean', *Nature Geoscience*, vol. 1, no. 12, pp. 849-853, doi:10.1038/ngeo357.

Adamowski, J and Chan, HF 2011, 'A wavelet neural network conjunction model for groundwater level forecasting', *Journal of Hydrology*, vol. 407, no. 1-4, pp. 28-40, doi:10.1016/j.jhydrol.2011.06.013.

Adamowski, J, Fung Chan, H, Prasher, SO, 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*, vol. 48, no. 1, doi:10.1029/2010wr009945.

Ashok, K, Behera, SK, Rao, SA, Weng, H and Yamagata, T 2007, 'El Niño Modoki and its possible teleconnection', *Journal of Geophysical Research*, vol. 112, no. C11, doi:10.1029/2006jc003798.

Australian Bureau of Statistics 2010, *Year Book Australia, 2009–10*, Commonwealth of Australia Canberra, <http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/1301.0Chapter3042009%E2%80%9310>

Australian Bureau of Statistics 2014, *Value of Agricultural Commodities Produced, Australia, 2012-13.*, http://www.abs.gov.au/ausstats/abs@.nsf/mf/7503.0>

Australian Bureau of Statistics 2017, *Agricultural census factsheet: Australia, states and territories*, ABS, Canberra, Australia.

AWAP 2016, *Readme File: Australian Water Availability Project (AWAP)*, CSIRO, CSIRO.

Bai, Y, Wang, P, Xie, J, Li, J and Li, C 2015, 'Additive Model for Monthly Reservoir Inflow Forecast', *Journal of Hydrologic Engineering*, vol. 20, no. 7, p. 04014079, doi:10.1061/(asce)he.1943-5584.0001101.

Baker, L and Ellison, D 2008, 'The wisdom of crowds — ensembles and modules in environmental modelling', *Geoderma*, vol. 147, no. 1-2, pp. 1-7, doi:10.1016/j.geoderma.2008.07.003.

Barzegar, R, Moghaddam, AA, Deo, R, Fijani, E and Tziritis, E 2017, 'Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms', *Sci Total Environ*, vol. 621, pp. 697-712, <https://www.ncbi.nlm.nih.gov/pubmed/29197289>
doi:10.1016/j.scitotenv.2017.11.185.

Basha, G, Ouarda, TBMJ and Marpu, PR 2015, 'Long-term projections of temperature, precipitation and soil moisture using non-stationary oscillation processes over the UAE region', *International Journal of Climatology*, vol. 35, no. 15, pp. 4606-4618, doi:10.1002/joc.4310.

Beesley, CA, Frost, AJ and Zajaczkowski, J 2009, 'A comparison of the BAWAP and SILO spatially interpolated daily rainfall datasets', presented at *18th World IMACS / MODSIM Congress*, Cairns, Australia, http://mssanz.org.au/modsim09>, http://mssanz.org.au/modsim09

Beltran-Castro, J, Valencia-Aguirre, J, Orozco-Alzate, M, Castellanos-Domınguez, G and Travieso-Gonzalez, CM 2013, 'Rainfall forecasting based on ensemble empirical mode decomposition and neural networks', in *Proceedings of Advances in Computational Intelligence: 12th International Work-Conference on Artificial Neural Networks*, Springer, Puerto de la Cruz, Tenerife, Spain, pp. 471-480.

Breiman, L 2001, 'Random Forests', *Machine Learning*, vol. 45, pp. 5-32.

Brocca, L, Melone, F and Moramarco, T 2008, 'On the estimation of antecedent wetness conditions in rainfall-runoff modelling', *Hydrological Processes*, vol. 22, no. 5, pp. 629-642, doi:10.1002/hyp.6629.

Brocca, L, Melone, F, Moramarco, T and Morbidelli, R 2010, 'Spatial-temporal variability of soil moisture and its estimation across scales', *Water Resources Research*, vol. 46, no. 2, pp. 1-14, doi:10.1029/2009wr008016.

Bureau of Meteorology 2018, *National Water Account 2015: Murray–Darling Basin-Region overview*, viewed 14/06/2018, <http://www.bom.gov.au/water/nwa/2015/mdb/index.shtml>

Campbell, R and Scarlett, A 2014, 'Economics, agriculture and native vegetation in NSW', *The Australia Institute.*

CSIRO 2012, 'Climate and water availability in south-eastern Australia: A synthesis of findings from Phase 2 of the South Eastern Australian Climate Initiative (SEACI)', *CSIRO, Australia, September 2012,* p. 41.

Dee, DP, Uppala, SM, Simmons, AJ, Berrisford, P, Poli, P, Kobayashi, S, Andrae, U, Balmaseda, MA, Balsamo, G, Bauer, P, Bechtold, P, Beljaars, ACM, van de Berg, L, Bidlot, J, Bormann, N, Delsol, C, Dragani, R, Fuentes, M, Geer, AJ, Haimberger, L, Healy, SB, Hersbach, H, Hólm, EV, Isaksen, L, Kållberg, P, Köhler, M, Matricardi, M, McNally, AP, Monge-Sanz, BM, Morcrette, JJ, Park, BK, Peubey, C, de Rosnay, P, Tavolato, C, Thépaut, JN and Vitart, F 2011, 'The ERA-Interim reanalysis: configuration and performance of the data assimilation system', *Quarterly Journal of the Royal Meteorological Society*, vol. 137, no. 656, pp. 553-597, doi:10.1002/qj.828.

Deo, RC and Şahin, 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', *Atmospheric Research*, vol. 161-162, pp. 65-81, doi:10.1016/j.atmosres.2015.03.018.

Deo, RC, Byun, H-R, Adamowski, JF and Kim, D-W 2015, 'A Real-time Flood Monitoring Index Based on Daily Effective Precipitation and its Application to

Brisbane and Lockyer Valley Flood Events', *Water Resources Management*, pp. 1-19 (DOI: 10.1007/s11269-11015-11046-11263).

Deo, RC, Byun, H-R, Adamowski, JF and Begum, K 2016, 'Application of effective drought index for quantification of meteorological drought events: a case study in Australia', *Theoretical Applied Climatology*, vol. 122, no. 3-4, doi:10.1007/s00704-015-1706-5.

Deo, RC, Tiwari, MK, Adamowski, JF and Quilty, MJ 2017, 'Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model', *Stochastic Environmental Research and Risk Assessment*, vol. 31, no. 5, pp. 1211–1240, <http://dx.doi.org/10.1007/s00477-016-1265-z> doi:10.1007/s00477-016-1265-z.

Deo, RC, Syktus, J, McAlpine, C, Lawrence, P, McGowan, H and Phinn, SR 2009, 'Impact of historical land cover change on daily indices of climate extremes including droughts in eastern Australia', *Geophysical Research Letters*, vol. 36, no. 8.

Friedman, JH 1991, 'Multivariate adaptive regression splines', *The Annals of Statistics*, vol. 19, no. 1, pp. 1-141.

Gill, MK, Asefa, T, Kemblowski, MW and McKee, M 2006, 'Soil moisture prediction using support vector machines', *JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION*, vol. 42, no. 4, pp. 1033-1046.

Hatampour, A 2013, 'Developing a Committee Machine Model for Predicting Reservoir Porosity from Image Analysis of Thin Sections', *Middle-East Journal of Scientific Research*, vol. 13, no. 11, pp. 1438-1444, doi:10.5829/idosi.mejsr.2013.13.11.1349.

Hejazi, MI and Cai, X 2009, 'Input variable selection for water resources systems using a modified minimum redundancy maximum relevance (mMRMR) algorithm', *Advances in Water Resources*, vol. 32, no. 4, pp. 582-593, doi:10.1016/j.advwatres.2009.01.009.

Henley, BJ, Gergis, J, Karoly, DJ, Power, S, Kennedy, J and Folland, CK 2015, 'A Tripole Index for the Interdecadal Pacific Oscillation', *Climate Dynamics*, vol. 45, no. 11-12, pp. 3077-3090, doi:10.1007/s00382-015-2525-1.

Ho, M, Kiem, AS and Verdon-Kidd, DC 2012, 'The Southern Annular Mode: a comparison of indices', *Hydrology and Earth System Sciences*, vol. 16, no. 3, pp. 967-982, doi:10.5194/hess-16-967-2012.

Huang, B, Banzon, VF, Freeman, E, Lawrimore, J, Liu, W, Peterson, TC, Smith, TM, Thorne, PW, Woodruff, SD and Zhang, H-M 2015, 'Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4). Part I: Upgrades and Intercomparisons', *Journal of Climate*, vol. 28, no. 3, pp. 911-930, doi:10.1175/jcli-d-14-00006.1.

Huang, G-B 2015, 'What are Extreme Learning Machines? Filling the Gap Between Frank Rosenblatt's Dream and John von Neumann's Puzzle', *Cognitive Computation*, vol. 7, no. 3, pp. 263-278, doi:10.1007/s12559-015-9333-0.

Huang, G, Huang, GB, Song, S and You, K 2015, 'Trends in extreme learning machines: a review', *Neural Netw*, vol. 61, pp. 32-48, <http://www.ncbi.nlm.nih.gov/pubmed/25462632>
doi:10.1016/j.neunet.2014.10.001.

Huang, NE, Shen, Z, Long, SR, Wu, MC, Shih, HH, Zheng, Q, Yen, N-C, Tung, CC and Liu, HH 1998, 'The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis', *Proceedings of Royal Society*, vol. A 454, pp. 903–995.

Humphrey, GB, Gibbs, MS, Dandy, GC and Maier, HR 2016, 'A hybrid approach to monthly streamflow forecasting: Integrating hydrological model outputs into a Bayesian artificial neural network', *Journal of Hydrology*, vol. 540, pp. 623-640, doi:10.1016/j.jhydrol.2016.06.026.

IPCC 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 [Core Writing Team, R.K. Pachauri and L.A. Meyer (eds.)]*, Geneva, Switzerland.

Jeffrey, SJ, Carter, JO, Moodie, KB and Beswick, AR 2001, 'Using spatial interpolation to construct a comprehensive archive of Australian climate data', *Environmental Modelling & Software*, vol. 16, no. 2001, pp. 309-330.

Jiao, G, Guo, T and Ding, Y 2016, 'A New Hybrid Forecasting Approach Applied to Hydrological Data: A Case Study on Precipitation in Northwestern China', *Water*, vol. 8, no. 9, p. 367, doi:10.3390/w8090367.

Jones, DA, Wang, W and Fawcett, R 2009, 'High-quality spatial climate data-sets for Australia', *Australian Meteorological and Oceanographic Journal*, vol. 58, pp. 233-248.

Liaw, A and Wiener, M 2002, 'Classification and Regression by random forest', *R News*, vol. 2, pp. 18-22.

Liu, W, Huang, B, Thorne, PW, Banzon, VF, Zhang, H-M, Freeman, E, Lawrimore, J, Peterson, TC, Smith, TM and Woodruff, SD 2015, 'Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4): Part II. Parametric and Structural Uncertainty Estimations', *Journal of Climate*, vol. 28, no. 3, pp. 931-951, doi:10.1175/jcli-d-14-00007.1.

Loon, AFV and Laaha, G 2015, 'Hydrological drought severity explained by climate and catchment characteristics', *Journal of Hydrology*, vol. 526, pp. 3-14, doi:10.1016/j.jhydrol.2014.10.059.

Maheswaran, R and Khosa, R 2012, 'Wavelet–Volterra coupled model for monthly streamflow forecasting', *Journal of Hydrology*, vol. 450-451, pp. 320-335, doi:10.1016/j.jhydrol.2012.04.017.

Maheswaran, R and Khosa, R 2015, 'Wavelet Volterra Coupled Models for forecasting of nonlinear and non-stationary time series', *Neurocomputing*, vol. 149, pp. 1074-1084, doi:10.1016/j.neucom.2014.07.027.

Maier, HR, Jain, A, Dandy, GC and Sudheer, KP 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*, vol. 25, no. 8, pp. 891-909, doi:10.1016/j.envsoft.2010.02.003.

McAlpine, C, Syktus, J, Ryan, J, Deo, RC, McKeon, G, McGowan, H and Phinn, S 2009, 'A continent under stress: interactions, feedbacks and risks associated with impact of modified land cover on Australia's climate', *Global change biology*, vol. 15, no. 9, pp. 2206-2223.

Mishra, AK and Singh, VP 2010, 'A review of drought concepts', *Journal of Hydrology*, vol. 391, no. 1-2, pp. 202-216, doi:10.1016/j.jhydrol.2010.07.012.

Mouatadid, S and Adamowski, J 2016, 'Using extreme learning machines for short-term urban water demand forecasting', *Urban Water Journal*, pp. 1-9, doi:10.1080/1573062x.2016.1236133.

Newman, M, Alexander, MA, Ault, TR, Cobb, KM, Deser, C, Di Lorenzo, E, Mantua, NJ, Miller, AJ, Minobe, S, Nakamura, H, Schneider, N, Vimont, DJ, Phillips, AS, Scott, JD and Smith, CA 2016, 'The Pacific Decadal Oscillation, Revisited', *Journal of Climate*, vol. 29, no. 12, pp. 4399-4427, doi:10.1175/jcli-d-15-0508.1.

NSW-Department of Industry 2017, *Export from New South Wales: Agribusiness and food*, New South Wales: Department of Industry, NSW, Australia, viewed 5th October, <https://www.industry.nsw.gov.au/export-from-nsw/key-industry-sectors/agribusiness-and-food >

NSW Department of Primary Industries-Office of Water 2016, *Rivers and streams*, 25/05/2016, <http://www.water.nsw.gov.au/realtime-data/hydro-rivers>

NSW Department of Primary Industries (DPI) Office of Water 2018, *Rivers and streams*, viewed 28/03/2018, <http://www.water.nsw.gov.au/realtime-data/hydro-rivers>

Ouyang, Q, Lu, W, Xin, X, Zhang, Y, Cheng, W and Yu, T 2016, 'Monthly Rainfall Forecasting Using EEMD-SVR Based on Phase-Space Reconstruction', *Water Resources Management*, vol. 30, no. 7, pp. 2311-2325, doi:10.1007/s11269-016-1288-8.

Petropoulos, GP 2014, *Remote Sensing of Energy Fluxes and Soil Moisture Content*, CRC Press, Taylor & Francis Group, Boca Raton, FL.

Rathinasamy, M, Khosa, R, Adamowski, J, ch, S, Partheepan, G, Anand, J and Narsimlu, B 2014, 'Wavelet-based multiscale performance analysis: An approach to assess and improve hydrological models', *Water Resources Research*, vol. 50, no. 12, pp. 9721-9737, doi:10.1002/2013wr014650.

Raupach, MR, Briggs, PR, Haverd, V, King, EA, Paget, M and Trudinger, CM 2009, *Australian water availability project (AWAP)-CSIRO marine and atmospheric research component-final report for phase 3*, CAWCR Technical Report No. 013.

Raupach, MR, Briggs, PR, Haverd, V, King, EA, Paget, M and Trudinger, CM 2012, *Australian Water Availability Project*, CSIRO Marine and Atmospheric Research, Canberra, Australia. http://www.csiro.au/awap, http://www.csiro.au/awap>

Renaud, O, Starck, JL and Murtagh, F 2002, *Wavelet-based forecasting of short and long-term memory time series*, Institute of Economics and Econometrics, Geneva School of Economics and Management, University of Geneva., http://EconPapers.repec.org/RePEc:gen:geneem:2002.04.>

Saji, NH, Goswami, BN, Vinayachandran, PN and Yamagata, T 1999, 'A dipole mode in the tropical Indian Ocean', *nature*, vol. 401, p. 360, <http://dx.doi.org/10.1038/43854> doi:10.1038/43854.

Seneviratne, SI, Corti, T, Davin, EL, Hirschi, M, Jaeger, EB, Lehner, I, Orlowsky, B and Teuling, AJ 2010, 'Investigating soil moisture-climate interactions in a changing climate: A review', *Earth-Science Reviews*, vol. 99, no. 3-4, pp. 125-161, doi:10.1016/j.earscirev.2010.02.004.

Seo, Y and Kim, S 2016, 'Hydrological Forecasting Using Hybrid Data-Driven Approach', *American Journal of Applied Sciences*, vol. 13, no. 8, pp. 891-899, doi:10.3844/ajassp.2016.891.899.

Surowiecki, J 2004, *The wisdom of crowds : why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*, Anchor Books, New York, USA.

Tang, J, Deng, C and Huang, GB 2016, 'Extreme Learning Machine for Multilayer Perceptron', *IEEE Trans Neural Netw Learn Syst*, vol. 27, no. 4, pp. 809-821, <https://www.ncbi.nlm.nih.gov/pubmed/25966483> doi:10.1109/TNNLS.2015.2424995.

Taschetto, AS and England, MH 2009, 'El Niño Modoki Impacts on Australian Rainfall', *Journal of Climate*, vol. 22, no. 11, pp. 3167-3174, doi:10.1175/2008jcli2589.1.

The Murray–Darling Basin Authority 2010, *Guide to the proposed Basin Plan Technical background Part 1*, <http://www.mdba.gov.au/sites/default/files/archived/guide_pbp/Guide-to-proposed-BP-vol2-0-12.pdf>

Timbal, B, Abbs, D, Bhend, J, Chiew, F, Church, J, Ekström, M, Kirono, D, Lenton, A, Lucas, C, McInnes, K, Moise, A, Monselesan, D, Mpelasoka, F, Webb, L and Whetton, P 2015, *Murray Basin Cluster Report: Climate Change in Australia Projections for Australia's Natural Resource Management Regions: Cluster Reports, eds. Ekström, M. et al.*, CSIRO and Bureau of Meteorology, Australia.

Torres, ME, Colominas, MA, Schlotthauer, G and Flandrin, P 2011, 'A complete ensemble empirical mode decomposition with adaptive noise', in *Proceedings of 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4144-4147.

Tozer, CR, Kiem, AS and Verdon-Kidd, DC 2012, 'On the uncertainties associated with using gridded rainfall data as a proxy for observed', *Hydrology and Earth System Sciences*, vol. 16, no. 5, pp. 1481-1499, doi:10.5194/hess-16-1481-2012.

Ummenhofer, CC, England, MH, McIntosh, PC, Meyers, GA, Pook, MJ, Risbey, JS, Gupta, AS and Taschetto, AS 2009, 'What causes southeast Australia's worst droughts?', *Geophysical Research Letters*, vol. 36, no. 4, doi:10.1029/2008gl036801.

van Dijk, AIJM, Beck, HE, Crosbie, RS, de Jeu, RAM, Liu, YY, Podger, GM, Timbal, B and Viney, NR 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*, vol. 49, no. 2, pp. 1040-1057, doi:10.1002/wrcr.20123.

Visbeck, M 2009, 'A Station-Based Southern Annular Mode Index from 1884 to 2005', *Journal of Climate*, vol. 22, no. 4, pp. 940-950, doi:10.1175/2008jcli2260.1.

Wang, D, Wei, S, Luo, H, Yue, C and Grunder, O 2017, 'A novel hybrid model for air quality index forecasting based on two-phase decomposition technique and modified extreme learning machine', *Sci Total Environ*, vol. 580, pp. 719-733,

<http://www.ncbi.nlm.nih.gov/pubmed/27989476> doi:10.1016/j.scitotenv.2016.12.018.

Welsh, WD, Vaze, J, Dutta, D, Rassam, D, Rahman, JM, Jolly, ID, Wallbrink, P, Podger, GM, Bethune, M, Hardy, MJ, Teng, J and Lerat, J 2013, 'An integrated modelling framework for regulated river systems', *Environmental Modelling & Software*, vol. 39, pp. 81-102, doi:10.1016/j.envsoft.2012.02.022.

White, I, Falkland, T and Scott, D 1999, *Droughts in Small Coral Islands: Case Study, South Tarawa, Kiribati*, International Hydrology Programme-V Technical Documents in Hydrology No. 26, UNESCO, Paris.

Wilhite, DA and Glantz, MH 1985, 'Understanding the Drought Phenomenon: The Role of Definitions', *Water International*, vol. 10, no. 3, pp. 111–120, <http://digitalcommons.unl.edu/droughtfacpub/20>, http://digitalcommons.unl.edu/droughtfacpub/20

Wu, Z and Huang, NE 2009a, 'Ensemble empirical mode decomposition: a noise-assisted data analysis method', *Advances in Adaptive Data Analysis*, vol. 01, no. 01, pp. 1-41, <http://www.worldscientific.com/doi/abs/10.1142/S1793536909000047> doi:10.1142/s1793536909000047.

Wu, Z and Huang, NE 2009b, 'Ensemble empirical mode decomposition: A noise-assisted data analysis method', *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1-41, doi:10.1142/S1793536909000047.

Yang, X, Zhang, C, Cheng, Q, Zhang, H and Gong, W 2017, 'A Hybrid Model for Soil Moisture Prediction by Using Artificial Neural Networks', *Revista de la Facultad de Ingeniería U.C.V.*, vol. 32, no. 5, pp. 265-271.

Zajaczkowski, J, Wong, K and Carter, J 2013, 'Improved historical solar radiation gridded data for Australia', *Environmental Modelling & Software*, vol. 49, pp. 64-77, doi:10.1016/j.envsoft.2013.06.013.

Appendix

Table A1 Percentage deviations of forecasted values from the X=Y line from the IIS-W-ANN, IIS-W-M 5 Tree, IIS -ANN, IIS-M 5 Tree models at respective sites.

	Site 1-Menindee River				Site 2-Gwydir River				Site 3-Darling River			
Data point reference nos.	IIS-W-ANN %	IIS-W-M5 Tree %	IIS-ANN %	IIS-M5 Tree %	IIS-W-ANN %	IIS-W-M5 Tree %	IIS-ANN %	IIS-M5 Tree %	IIS-W-ANN %	IIS-W-M5 Tree %	IIS-ANN %	IIS-M5 Tree %
1	-38.08	-18.71	-3.13	-1.20	-7.89	-8.15	-22.75	-19.98	13.92	-33.19	9.45	11.19
2	-34.71	17.01	-2.95	0.24	-14.72	-14.48	-5.18	-6.35	9.84	-52.72	5.38	-0.88
3	-24.67	-51.91	-25.22	-23.95	-20.05	-16.13	-9.05	-6.69	-27.55	62.40	-42.06	-46.75
4	-1.17	-18.56	10.06	1.86	-1.69	-6.96	1.73	1.96	18.87	-13.79	-17.05	-40.68
5	-21.05	8.26	-37.44	-40.42	4.46	-0.97	-10.46	-6.32	4.37	-20.15	-5.69	-24.77
6	-1.17	-17.50	-19.88	-36.36	6.00	-3.40	-15.09	-24.98	-3.45	-19.57	-2.61	-13.30
7	-7.22	-25.73	2.65	-6.33	9.56	2.35	-27.29	-34.44	-8.94	-23.07	-11.42	-14.43
8	-10.55	-22.70	-9.05	-11.42	16.84	23.39	35.07	36.21	-8.48	-13.15	-6.40	-10.24
9	-20.56	-32.12	12.60	10.71	39.55	44.46	27.72	33.84	-15.63	-14.58	225.28	220.45
10	-2.43	-32.48	-27.21	-26.61	35.73	17.65	4.44	7.95	-39.56	-20.07	40.70	-36.76
11	-25.68	-40.12	-24.89	-23.74	32.54	15.78	6.35	2.18	-9.18	30.45	36.60	72.07
12	-45.76	-26.64	-9.07	-11.56	37.82	11.69	7.75	7.75	-6.56	-5.23	10.54	-0.58
13	-49.45	-2.22	-16.94	-13.66	22.82	8.04	0.83	3.51	2.27	57.17	19.30	6.86
14	-35.43	-4.59	-8.48	-7.02	-10.36	-9.70	-20.45	-21.10	0.65	-38.47	10.71	3.51
15	-27.37	-18.60	-8.03	8.40	-2.28	9.50	21.60	24.34	1.94	-61.66	16.99	5.68
16	-29.63	-33.71	-7.74	2.81	-9.71	-18.61	-25.93	-15.93	0.84	85.06	13.11	4.09
17	-22.92	-31.00	-13.90	-3.66	8.54	-16.22	-21.76	-15.18	-45.00	39.42	-45.34	-50.51
18	-4.06	-56.75	-16.55	-16.68	22.92	9.72	31.31	29.41	8.95	-4.69	-14.80	-40.88
19	-5.85	20.24	-3.54	-29.39	12.38	-2.74	-42.60	-41.18	-3.19	-8.56	-5.53	-8.48
20	-4.18	-22.47	-10.94	-15.66	22.17	2.49	2.43	28.53	-10.14	-10.30	-9.68	-14.79
21	6.09	-26.84	24.53	27.67	4.63	-13.66	-0.13	2.77	-7.12	-34.34	-2.22	-6.95
22	2.46	-6.94	-20.36	-17.70	6.04	-3.80	-14.58	-15.59	4.82	1.42	15.97	17.46
23	-21.81	-34.70	-1.47	-5.63	12.70	-2.27	26.90	19.63	-15.22	-8.57	-0.25	74.41
24	-26.27	-24.71	-19.13	-15.93	-10.62	-8.69	-25.47	-26.76	-0.54	13.99	-5.15	5.72
25	-29.60	-7.38	-0.32	-5.99	9.34	2.91	28.13	16.02	-36.06	-8.45	-27.89	-33.30
26	-21.83	-9.83	-1.48	-0.11	-15.55	-14.62	-9.49	-16.75	-27.08	-24.33	31.24	10.39
27	-11.94	21.29	4.65	-1.51	-8.58	-12.36	-9.78	-16.79	-14.95	-29.38	8.69	4.74
28	-1.93	4.32	10.79	9.09	-5.78	-8.32	22.74	29.19	-35.22	-33.27	-23.88	-31.69
29	5.87	5.46	2.92	3.36	-7.54	-13.40	-23.32	-32.16	-41.13	31.36	-24.53	-25.72
30	-16.06	-36.44	-33.11	-26.57	-0.95	-5.08	5.83	-11.16	-21.07	34.70	-10.61	19.69
31	-9.44	11.17	-10.35	-26.13	-2.44	-15.33	1.09	8.99	-13.54	28.05	-11.54	-9.81
32	-28.62	36.55	-10.03	-40.16	9.34	2.68	0.26	7.40	-2.05	35.10	-12.56	-15.12
33	-9.08	-12.52	60.80	29.38	30.58	16.94	34.95	30.73	30.73	-23.42	-31.61	-30.19
34	14.63	4.38	-10.67	-4.21	39.63	22.96	5.08	5.89	-24.22	-20.35	-5.01	27.94
35	-9.58	-13.24	-25.89	-14.32	25.49	7.39	13.66	-0.36	12.24	15.39	38.27	42.97
36	9.99	-38.87	-40.84	-37.99	16.33	2.81	3.47	2.78	11.27	5.22	27.64	19.72
37	-25.24	-14.31	7.09	10.80	9.59	0.38	11.74	9.77	14.30	3.75	14.79	7.74
38	-33.25	-9.80	-3.89	0.72	-23.10	-17.97	-20.02	-14.71	13.17	17.64	7.08	1.60

39	-18.56	10.48	6.78	0.04	-20.93	-12.65	3.04	-9.52	11.81	9.60	9.39	9.02
40	6.51	-0.01	24.39	8.36	-15.24	-9.20	-13.66	-10.80	8.64	1.77	7.79	1.22
41	31.14	8.23	0.54	-1.50	-9.34	-13.76	-7.59	-9.35	-17.13	-19.74	-19.85	-27.61
42	17.96	-4.04	4.06	10.39	-13.59	-11.72	-8.94	-6.46	-11.93	11.13	22.55	16.33
43	14.92	14.28	8.78	1.10	-2.55	-8.23	20.89	14.86	1.05	-4.79	12.15	5.22
44	6.27	-7.93	4.84	7.89	10.77	20.69	17.29	48.53	13.39	10.12	27.18	11.64
45	22.03	14.60	0.70	10.07	33.89	17.80	25.04	35.51	13.51	2.70	9.80	4.10
46	30.43	-4.17	1.60	2.84	41.59	19.33	2.42	6.17	18.52	8.45	18.73	10.33
47	44.10	-0.82	6.22	-1.33	29.10	17.63	5.74	4.88	18.64	8.00	7.67	4.41
48	44.08	8.73	13.10	-4.93	20.33	13.76	6.98	9.14	18.95	7.21	7.79	4.45
49	0.36	-22.57	39.39	57.96	11.25	17.84	5.84	8.54	17.06	5.19	5.49	12.57
50	1.30	11.45	3.98	3.89	-12.12	-19.34	-11.39	-12.67	20.76	8.15	12.39	18.18
51	2.08	26.89	27.50	9.43	-0.85	-9.33	-1.24	-7.63	20.09	7.11	3.71	4.18
52	8.76	28.22	13.33	4.46	2.00	4.83	33.87	23.85	21.03	9.00	6.98	11.65
53	2.03	11.09	33.18	38.37	9.94	4.41	-21.06	-13.89	22.91	11.46	6.65	12.47
54	-2.77	43.85	0.33	-3.80	12.01	17.33	-2.32	-5.68	22.75	11.35	2.77	8.33
55	-7.78	-29.54	-42.85	-38.54	15.74	-9.31	-0.06	-6.70	21.34	20.68	5.50	8.80
56	-34.29	16.33	1.39	-1.16	7.73	36.29	42.53	49.30	25.30	9.73	10.07	14.84
57	-36.99	-1.99	-19.09	-14.36	-0.70	16.80	11.05	6.54	26.03	21.89	4.94	1.91
58	-16.32	-20.84	-21.97	-18.49	4.27	26.48	9.79	12.96	52.28	30.29	36.46	43.16
59	6.23	-14.71	8.15	13.04	0.32	20.15	16.63	9.72	58.35	38.65	-1.68	16.85
60	8.57	-15.09	-7.31	-5.73	-5.15	16.77	12.93	9.10	44.99	34.78	-8.26	3.58
61	-0.91	15.64	1.57	-0.45	-3.74	24.58	3.25	6.20	25.35	14.55	-4.18	-3.45
62	-4.39	12.92	-0.68	-0.82	19.22	33.23	20.76	12.38	28.45	5.25	17.94	8.74
63	-1.54	-13.89	0.10	-7.37	19.57	42.48	-4.56	12.65	21.64	-0.91	-4.54	0.45
64	-5.72	-29.15	39.70	48.01	33.38	45.24	21.70	25.72	23.84	4.24	9.81	6.99
65	-12.62	-25.76	-22.29	-18.05	16.29	33.39	1.40	-25.70	34.16	22.82	11.73	13.25
66	-30.76	-13.14	3.28	9.64	20.37	35.62	-10.87	3.61	52.46	45.38	10.64	22.36
67	-20.45	-24.37	-27.50	-19.94	28.15	28.40	58.54	66.98	52.78	63.73	-3.28	13.53
68	-10.60	-17.26	-1.51	2.66	16.51	40.62	19.04	49.50	60.48	76.52	10.25	19.46
69	-15.96	-11.80	-2.08	6.31	8.06	32.32	15.68	9.65	54.85	111.45	0.82	11.94
70	-27.60	-9.71	3.61	-6.13	-14.55	-4.26	-5.93	-7.08	50.31	64.83	5.99	11.02

Table A2 Percentage deviations of forecasted values from the X=Y line from the Best EEMD-ELM, CEEMDAN-ELM and the comparative RF models in forecasting a) upper layer soil moisture and b) lower layer soil moisture.

a) upper layer soil moisture (*SM_{UL}*)

Data point reference nos.	ELM (Site 30) %	EEMD-ELM (Site 43) %	CEEMDAN-ELM (Site 30) %	RF (Site 30) %	EEMD-RF (Site 43) %	CEEMDAN-RF (Site 30) %
1	356.57	8.67	130.88	595.68	17.22	189.08
2	7.24	-24.75	-3.65	-35.90	-19.98	-11.73
3	-6.74	15.11	-2.82	24.89	61.29	-6.04
4	22.76	61.37	12.30	8.54	112.95	-2.94
5	-31.41	-19.02	-18.64	-38.33	-18.96	-25.05
6	30.50	-14.36	1.06	0.91	-22.12	-0.86
7	-9.35	-12.10	-2.33	-13.93	-20.92	-11.88
8	8.39	-12.98	-13.56	11.84	2.19	-10.59
9	37.23	31.44	-17.02	53.55	91.45	36.88
10	55.49	-31.64	83.08	92.24	-1.40	131.10
11	-100.78	-30.70	-16.52	6.96	36.83	10.36
12	-12.12	15.09	-4.52	31.68	91.29	-2.05
13	-3.80	57.04	-0.57	23.26	57.85	-6.79
14	-22.89	26.48	4.68	3.27	-4.22	4.94
15	-26.53	-6.41	-2.40	-38.86	-14.78	-15.52
16	6.70	21.46	18.94	5.08	30.46	5.26
17	-26.86	-1.07	4.15	-17.17	0.44	3.24
18	-4.92	-7.60	-10.34	1.06	-6.00	-0.65
19	42.03	1.55	-1.85	78.02	-4.23	28.76
20	39.20	-15.07	19.51	27.08	-11.12	50.72
21	-28.45	-14.81	21.96	6.94	19.22	21.94
22	-26.83	89.45	-12.62	-12.03	102.84	-4.38
23	24.02	-10.52	16.45	-24.70	-23.26	-0.57
24	-27.70	-39.21	-4.18	-54.17	-45.74	-10.72
25	-1.91	36.94	-11.36	9.80	6.70	-3.73
26	28.46	31.99	3.90	116.46	106.75	66.15
27	-40.54	-17.75	-11.92	-27.09	-32.81	7.85
28	13.74	0.73	-0.83	-12.33	-8.48	-5.58
29	-10.39	-0.87	-2.56	-31.87	-14.78	-16.67
30	-0.77	-4.52	-3.59	-2.12	-11.57	-11.41
31	-15.39	-12.85	-14.13	-15.59	-13.73	-19.37
32	34.53	-11.29	-1.15	20.43	-4.81	-2.70
33	167.76	8.59	54.31	272.61	60.16	199.26
34	-41.14	-39.12	-21.42	-43.46	-19.73	-24.66
35	165.87	90.71	58.80	127.98	123.28	88.51
36	-83.04	-13.66	-25.86	-27.77	-8.85	7.13
37	-5.83	-50.07	-3.38	-11.52	-13.34	18.65
38	97.02	45.37	-20.36	85.47	52.78	86.64

39	69.46	66.03	16.52	131.97	125.74	168.71
40	-45.54	-16.03	4.96	-46.72	-13.80	-22.93
41	-12.20	-15.79	-20.70	-25.09	-31.87	-16.57
42	8.65	7.96	9.24	-13.56	-4.92	-4.83
43	-15.07	10.13	14.41	-24.67	-4.08	-6.84
44	-58.80	-7.06	-17.13	-49.90	-27.45	-29.15
45	4.51	-1.15	-33.29	-31.09	-32.08	-33.44
46	36.39	24.51	2.55	72.35	3.60	23.33
47	103.25	28.13	-85.40	91.22	33.81	104.06

b) lower layer soil moisture (*SM_{LL}*)

Data point reference nos.	ELM (Site 30) %	EEMD-ELM (Site 43) %	CEEMDAN-ELM (Site 30) %	RF (Site 30) %	EEMD-RF (Site 43) %	CEEMDAN-RF (Site 30) %
1	-0.52	2.39	1.14	1.32	-0.43	-0.45
2	-0.50	0.51	-1.96	1.88	-0.47	-1.70
3	0.11	1.55	0.84	1.40	0.33	2.74
4	1.09	5.13	1.24	2.69	2.95	5.61
5	-9.90	-0.45	-0.92	-9.02	-4.04	4.71
6	-0.19	0.45	1.30	-14.35	-5.09	4.58
7	13.21	4.60	0.01	-2.90	-3.79	3.72
8	-4.37	-1.11	0.97	5.14	-1.60	5.43
9	0.19	-1.13	1.16	4.46	0.84	6.66
10	1.26	-4.81	0.15	3.99	-0.23	6.65
11	-0.72	1.91	2.36	3.04	0.00	7.63
12	-0.05	-0.30	2.97	1.02	1.18	9.07
13	0.42	0.58	3.17	0.42	2.16	9.37
14	1.39	2.59	0.49	1.37	5.30	7.08
15	-4.38	0.29	2.08	-3.72	1.11	5.72
16	-4.16	3.46	4.68	-9.28	0.74	8.81
17	5.71	-0.18	-1.16	-5.97	1.21	8.22
18	2.18	-1.82	2.51	-0.35	2.21	10.88
19	0.32	2.23	3.57	0.82	2.34	13.38
20	-1.44	-0.31	-0.16	-1.37	2.29	10.90
21	4.04	4.01	-0.52	2.21	5.88	11.21
22	-2.48	0.32	-1.25	3.89	5.85	13.67
23	1.31	1.35	2.93	3.04	7.01	21.80
24	-0.20	-0.38	-2.26	0.16	5.67	17.36
25	1.16	-1.56	-4.04	-0.80	4.43	14.88
26	2.37	0.33	-2.40	2.85	5.69	17.56
27	0.99	0.48	-1.72	3.72	5.80	19.82
28	0.56	0.77	0.36	1.10	4.84	20.87
29	1.07	1.17	0.09	-0.32	6.12	20.81
30	-2.45	-0.24	-7.91	-1.97	3.67	0.97
31	0.55	0.62	-2.38	-0.99	4.23	4.69

32	-2.70	-2.31	-2.54	-1.38	0.96	6.46
33	3.79	1.01	-4.58	2.35	2.96	7.34
34	-0.53	-0.62	-2.46	3.73	-0.81	9.89
35	1.24	-1.27	1.19	2.91	-1.28	12.82
36	4.57	0.97	-8.84	7.39	1.15	2.21
37	-2.91	-1.13	0.66	4.61	0.71	3.58
38	1.21	4.28	5.81	1.64	6.91	9.58
39	0.26	2.80	4.70	0.53	7.82	14.56
40	-4.42	-0.52	2.42	-3.39	9.74	7.19
41	-8.05	5.69	10.73	-10.20	8.83	-2.15
42	-6.83	-6.14	-1.72	-14.16	5.39	-15.42
43	-1.14	-4.45	0.07	-14.09	4.64	-10.17
44	-8.09	3.67	-2.80	-19.65	-3.34	-14.35
45	-7.01	0.06	-6.52	-27.25	-16.26	-18.30
46	6.35	9.03	-0.24	-12.45	-10.88	-16.19
47	-4.66	-5.12	2.45	-1.79	-9.78	-17.34

Table A3 Percentage deviations of forecasted values from the X=Y line from the ANN-CoM and the competing standalone models (Volterra, M5 tree, random forest (RF) and extreme learning machine (ELM)) in forecasting a) upper layer soil moisture and b) lower layer soil moisture at the four study sites.

a) upper layer soil moisture (*SM_{UL}*)

Data point reference nos.	Site 1- Menindee					Site 2- Balranald				
	Volterra %	M5 Tree %	RF %	ELM %	ANN -CoM %	Volterra %	M5 Tree %	RF %	ELM %	ANN-CoM %
1	-6.86	-50.22	20.16	5.45	-4.93	-50.81	-60.22	50.96	-26.88	-1.16
2	-17.75	-3.81	15.02	-1.64	-3.22	-58.73	-31.88	-6.41	-13.57	-23.04
3	-3.48	6.46	2.52	0.13	1.25	8.53	13.53	20.02	12.03	16.98
4	0.02	14.28	-3.05	0.45	-0.58	15.11	29.16	17.47	3.95	7.75
5	-6.63	-6.01	-10.16	-6.62	-5.32	28.01	6.15	8.21	0.17	3.27
6	-16.61	-22.62	-26.61	-1.53	-1.76	-37.34	-18.81	-12.48	-1.11	1.72
7	1.34	-4.52	-12.32	1.08	0.12	-9.06	0.49	2.27	3.92	1.67
8	22.09	12.64	4.39	-1.52	2.62	33.54	6.22	13.87	1.59	7.63
9	2.11	3.19	36.08	-1.66	-3.76	26.12	19.26	32.35	9.33	17.94
10	-45.51	-8.44	35.64	-0.74	0.84	-29.94	5.78	42.74	-2.12	7.72
11	-74.00	-27.05	31.51	-10.27	2.53	-27.20	-5.74	12.91	19.98	6.59
12	11.23	-7.64	10.86	0.65	0.08	17.81	5.29	3.98	1.61	1.00
13	16.77	-19.03	22.08	-4.81	-6.44	4.60	-10.62	10.36	-4.18	5.12
14	-22.11	20.91	-3.13	-3.16	-3.29	-15.87	-4.11	-2.61	-8.99	-5.71
15	-7.51	5.68	0.06	-0.39	0.16	7.44	6.99	10.92	9.61	10.28
16	17.01	-0.58	-17.84	3.19	5.41	7.43	-4.41	-15.94	11.42	-3.34
17	2.68	1.39	-21.91	5.78	0.17	-8.73	10.93	-1.28	11.35	8.09
18	-6.66	4.00	-9.83	-0.62	-2.67	-26.91	-8.47	-17.34	-4.46	-5.70
19	12.71	17.65	4.20	2.69	3.56	13.77	-0.65	2.88	-0.04	1.13
20	16.79	14.38	9.80	6.23	9.37	64.22	13.15	20.53	-11.80	0.37
21	7.91	4.37	5.16	-1.19	-0.44	17.49	43.41	37.25	19.09	25.33
22	-22.99	-7.42	12.73	-3.39	-3.53	-40.30	5.21	28.57	25.24	7.30
23	-20.35	-7.33	9.29	5.05	5.68	-4.70	-1.54	-1.45	-2.98	-1.45
24	8.97	-21.63	10.74	-2.01	-5.18	14.86	-6.76	-2.60	-7.08	-1.56
25	4.90	-0.04	8.47	-7.54	-9.03	15.19	-0.32	-3.59	-0.54	-4.53
26	-18.67	20.75	2.31	-3.93	-3.09	-7.64	-4.14	3.68	-0.65	1.20
27	-25.18	2.51	11.91	1.82	4.60	-45.03	13.09	19.91	47.62	15.99
28	-4.69	-4.36	-6.69	-1.44	-2.46	-10.38	-7.97	-16.17	-3.28	-8.05
29	3.03	12.26	0.37	3.14	1.78	12.72	19.69	14.98	4.01	11.23
30	-4.71	2.32	-6.92	-3.53	-5.12	-19.16	-5.31	-17.20	-6.29	-6.66
31	-5.95	-2.88	-4.12	0.06	-0.13	-8.38	-5.68	-13.06	-8.27	-7.63
32	2.89	7.62	10.71	1.99	1.18	30.26	3.09	4.23	0.13	0.36
33	-0.86	-5.99	-5.30	-5.35	-5.29	-9.42	11.54	12.06	-7.83	-0.30
34	-15.97	-0.85	17.96	1.26	2.21	-27.87	5.46	18.05	-13.06	-3.20
35	0.89	1.63	8.85	2.51	1.99	3.11	-4.23	-0.33	-1.04	-0.45
36	10.25	-6.37	18.95	2.34	2.25	7.93	-29.94	14.06	-18.77	10.00
37	7.94	4.55	0.46	2.21	3.72	16.94	-1.66	-12.62	7.71	-4.48

38	-9.30	13.99	4.38	-7.93	-6.32	-0.94	0.09	9.24	-3.36	3.04
39	-9.34	-3.19	6.28	-5.45	-5.64	-1.15	6.70	-2.88	11.17	3.87
40	-3.09	11.62	8.21	2.66	3.27	-14.57	21.05	-1.46	-13.33	-5.20
41	8.15	-4.94	-13.67	3.33	2.46	-11.06	-4.96	-15.27	-3.01	-7.55
42	-12.09	-14.73	-16.85	-3.57	-3.90	-29.80	-8.12	-14.04	-8.43	-7.65
43	-8.30	-10.85	-14.70	-0.38	-1.62	-13.15	-8.30	-3.09	-4.64	-4.46
44	0.05	-10.55	-16.16	-1.90	-2.23	14.85	-5.43	-3.22	1.12	-0.57
45	6.03	-11.78	-34.14	2.48	-0.01	-0.65	-6.99	-18.65	3.26	2.44
46	-5.64	-2.58	-15.78	1.87	-0.06	-16.21	-6.66	-8.60	-8.51	-8.42
47	-9.85	1.17	8.37	3.31	3.88	4.60	0.62	2.56	9.83	5.30
48	10.62	11.00	20.96	6.04	5.95	22.83	10.40	4.90	17.37	11.11

Data point reference nos.	**Site 3- Bobadah**					**Site 4 – Rocky Creek**				
	Volterra %	M5 Tree %	RF %	ELM %	ANN-CoM %	Volterra %	M5 Tree %	RF %	ELM %	ANN-CoM %
1	-13.21	-74.60	57.01	-6.96	29.01	-36.48	3.67	10.89	15.46	13.76
2	147.05	12.10	129.11	19.13	-17.67	-5.58	6.83	-7.88	4.46	2.85
3	-3.15	3.91	-3.47	4.38	1.88	18.28	7.28	13.25	6.10	5.66
4	-95.60	21.57	-2.46	22.13	15.41	-40.50	-0.34	0.53	-2.44	-7.82
5	35.36	23.39	7.98	-0.54	-2.37	-12.29	18.96	25.34	-4.20	0.07
6	15.44	-11.49	-15.25	1.76	-0.32	-19.69	-3.69	-3.27	-5.80	-3.15
7	-40.29	-10.39	-14.88	2.65	0.24	-22.13	-18.72	-11.23	0.21	1.94
8	-86.29	-4.57	-20.39	-2.71	-0.27	-18.77	13.59	0.46	-0.10	1.14
9	24.07	16.30	20.50	8.86	10.47	-0.90	9.79	10.15	7.52	5.63
10	-70.08	-24.22	-4.89	-16.85	-15.49	-25.24	-9.37	6.21	-0.60	-5.58
11	-67.16	-76.97	35.36	8.12	87.34	6.39	-3.89	32.76	-2.47	1.37
12	29.01	-27.92	97.01	18.52	12.93	-22.05	-1.21	-21.46	-7.11	-6.07
13	1.68	-4.24	21.51	12.05	7.10	-43.58	-87.72	65.30	-6.37	0.24
14	26.92	-4.46	9.45	-2.70	-3.13	-23.22	-7.45	7.69	-6.06	-5.54
15	19.57	2.92	-0.98	6.26	4.48	0.02	-7.68	5.51	-1.14	-0.18
16	-26.59	19.90	-15.73	6.34	-0.56	-35.11	-6.10	-1.69	2.71	-0.84
17	-25.40	28.05	20.60	20.83	22.03	-28.15	-3.71	-1.41	-2.67	-3.95
18	-4.18	-6.96	-4.45	-0.53	0.88	-17.34	9.58	4.63	-0.85	1.90
19	-75.03	-10.34	-16.35	-2.57	1.11	-35.19	-3.70	-2.62	-2.17	-4.44
20	3.49	7.48	10.06	4.42	6.30	18.79	15.10	16.67	-0.68	2.62
21	-39.68	1.07	-6.00	-3.26	-12.76	-13.85	-0.80	-6.97	-2.34	-2.22
22	-32.79	0.05	6.38	-5.73	-12.41	-27.02	-25.75	4.01	-6.55	-11.64
23	-3.72	-0.79	27.94	1.50	1.25	-33.70	-27.36	12.61	-10.60	-11.85
24	16.78	11.50	8.82	8.23	8.89	-1.30	5.73	17.47	0.41	0.71
25	4.16	6.29	-1.91	6.92	3.19	3.92	6.04	-3.96	0.41	0.58
26	-31.25	-1.37	-4.66	2.23	-6.68	-28.50	1.03	-9.59	1.86	1.07
27	-55.57	11.90	13.41	15.95	9.49	-7.59	1.53	9.62	-5.94	-5.07
28	38.24	23.43	24.63	16.87	19.78	-2.22	-13.25	-8.95	0.69	-0.22
29	-19.35	12.37	11.76	10.60	9.85	-21.40	-3.31	7.69	1.03	-1.28
30	-22.42	-6.75	-9.17	-4.31	-4.09	-8.81	-1.71	0.22	-1.25	0.48
31	-15.98	-15.55	-13.89	-9.12	-8.09	-20.57	-10.61	-1.03	-0.44	1.40

32	-32.77	3.25	4.07	0.54	0.01	5.92	5.20	9.83	-0.88	-0.35
33	-70.58	-19.60	-20.02	-13.70	-18.82	-12.83	-0.68	-7.73	-3.35	-3.01
34	84.30	4.39	67.25	5.59	-7.93	-8.74	-13.06	20.95	-2.59	-2.81
35	-20.56	-9.70	-17.99	-9.77	-11.04	-5.05	-6.05	-5.32	-11.97	-11.70
36	25.18	5.30	43.91	16.44	8.45	-7.48	1.41	15.14	-0.97	0.28
37	36.14	16.20	19.46	12.66	10.30	11.15	8.34	12.43	2.94	4.82
38	-82.91	3.41	-18.86	-2.83	-3.54	-49.53	-10.14	-22.75	-11.30	-12.08
39	-0.56	31.91	41.45	30.50	15.37	-20.20	-23.39	4.31	4.68	-4.69
40	159.55	81.70	74.25	19.32	-4.50	-28.39	-13.35	3.49	-6.45	-7.41
41	1.74	-5.79	-4.40	6.88	6.60	-4.47	22.85	14.93	1.78	3.91
42	5.31	-19.54	-25.23	1.45	-6.38	-18.14	-16.54	-15.34	-6.43	-8.12
43	-30.48	-9.77	-15.61	-2.26	-3.69	-23.85	-17.48	-8.88	0.01	2.09
44	-12.69	-1.50	-5.75	-4.63	-4.04	10.33	-9.34	-13.17	-0.07	-0.68
45	-0.04	-17.61	-19.86	2.07	-2.11	-2.92	-17.15	-30.54	-6.87	-4.10
46	-31.48	-4.92	-13.74	-3.48	-0.72	25.01	-8.84	-17.27	-3.67	0.06
47	-10.82	5.49	6.47	8.36	12.23	-9.30	-8.41	-19.03	-10.45	-8.89
48	41.10	27.59	20.88	24.91	21.83	-10.69	11.78	16.19	-4.02	-2.28

b) lower layer soil moisture (*SM_{LL}*)

Data point reference nos.	Site 1- Menindee					Site 2- Balranald				
	Volterra %	M5 Tree %	RF %	ELM %	ANN -CoM %	Volterra %	M5 Tree %	RF %	ELM %	ANN-CoM %
1	1.44	0.85	-0.84	0.50	0.68	-17.01	-1.97	0.79	0.44	-0.05
2	-4.03	-2.83	-2.92	-2.66	-2.44	-24.06	-2.92	-2.14	-2.12	-2.66
3	1.81	0.09	0.87	0.17	0.70	-11.90	0.59	0.06	0.33	-0.27
4	0.01	-1.13	-0.26	-0.59	-0.22	-8.72	0.76	-0.03	-0.05	0.17
5	0.64	-0.29	0.59	0.39	0.69	-6.12	0.01	0.57	0.03	0.61
6	-2.54	0.10	-1.01	-0.88	-0.21	-7.11	-1.64	1.74	0.57	0.64
7	-3.75	-2.18	1.21	1.91	0.88	-0.41	-3.76	0.27	0.36	1.06
8	3.96	-1.94	1.46	0.75	0.47	-0.16	-3.36	0.47	0.75	1.07
9	-30.23	-1.95	4.08	-1.60	-0.16	-7.24	-3.70	0.31	-0.05	0.35
10	3.90	-0.14	3.27	0.01	0.65	-8.26	-1.18	0.14	0.93	-0.37
11	-19.07	-0.47	-3.46	-1.53	-1.77	-10.18	0.40	-1.07	-0.53	-0.92
12	0.58	0.16	-0.47	0.05	0.38	-16.16	2.24	0.56	0.30	0.60
13	1.65	-0.11	0.75	0.11	0.65	-14.10	1.06	0.24	0.98	0.65
14	-11.77	-3.08	-1.29	-2.51	-2.30	-11.93	-2.84	-2.50	-2.81	-2.49
15	0.82	-0.29	1.19	0.34	0.81	-9.72	1.27	1.14	0.88	1.05
16	-4.31	-1.21	-0.99	-0.92	-0.77	-13.57	-0.20	-1.53	-0.04	-0.66
17	-11.38	-0.59	1.69	0.49	0.36	-11.59	0.60	-0.69	-0.07	-0.64
18	1.13	-1.61	1.22	-0.27	-0.45	-11.28	-1.91	-0.72	-0.56	-1.54
19	4.31	-0.57	3.27	-0.03	0.39	-9.91	-1.51	-0.16	0.60	-0.73
20	3.11	-0.15	2.07	0.17	0.55	-12.51	0.70	2.26	0.68	0.77
21	-3.97	0.23	-1.37	-0.16	-0.06	-13.99	-0.52	0.27	-0.52	-0.22
22	1.14	0.37	-0.68	0.46	0.54	-13.40	-0.61	1.63	1.44	-0.01
23	0.18	-1.10	-0.80	-0.69	-0.35	-9.22	-1.49	-1.96	3.82	-1.36
24	-5.63	-0.28	-0.05	0.29	0.21	0.87	0.10	0.57	4.60	0.23

25	-1.33	-0.40	-0.37	0.19	0.26	3.45	0.13	1.15	3.54	0.21
26	-13.43	-3.08	-0.80	-2.01	-2.05	-0.28	-1.18	-2.64	-1.11	-2.42
27	-4.05	-0.28	0.53	0.89	0.58	-14.50	1.72	0.20	-0.29	-0.02
28	-4.11	-1.14	0.94	-0.15	-0.04	-26.79	0.20	-1.01	-3.59	-1.36
29	-2.76	-0.26	1.46	0.73	0.79	-23.62	1.32	0.21	-1.18	-0.23
30	-6.20	-1.07	1.35	-0.36	-0.09	-12.35	0.32	-0.71	-2.19	-1.31
31	-1.82	-0.37	0.54	0.55	0.55	-16.14	-0.36	-0.01	4.83	-0.26
32	-2.50	-0.32	0.66	0.51	0.54	-20.50	-0.60	-0.63	5.10	-0.42
33	-1.25	-1.17	-0.55	-0.43	-0.30	-22.28	1.43	-0.48	-0.14	-0.34
34	-3.07	0.02	0.44	0.79	0.63	-37.48	1.61	0.45	-3.41	-0.03
35	-2.07	-1.61	-1.00	-0.04	-0.81	-24.54	1.17	-0.60	-5.15	-1.00
36	-6.15	-0.69	0.76	0.81	0.09	7.70	2.56	0.64	-4.21	0.19
37	-3.28	-0.12	0.23	0.93	0.14	-1.08	1.14	0.60	-2.42	-0.47
38	0.06	-1.28	-0.68	-1.13	-1.58	-1.79	-0.46	-1.54	-4.68	-2.21
39	0.22	0.53	2.00	0.85	0.38	1.64	1.58	0.23	-3.77	-0.33
40	-3.57	-0.65	-1.57	-0.17	-0.78	3.40	1.44	0.35	-5.03	-0.59
41	-13.56	-0.06	1.60	0.45	-0.01	6.60	1.36	-0.06	-3.95	-0.59
42	-3.59	-1.90	-1.45	-0.35	-1.16	0.12	-1.83	-2.00	3.74	-1.34
43	1.48	-0.32	1.02	-0.12	0.52	-5.51	-0.61	0.55	-0.65	0.25
44	-16.72	-0.70	0.95	0.50	-0.01	-8.22	0.94	-0.70	-0.03	-0.42
45	-14.54	-2.22	-0.10	-1.71	1.05	-9.63	-2.30	-1.11	-0.19	-1.56
46	6.95	3.42	-2.92	0.31	0.22	-1.21	-1.77	-0.06	0.26	-0.43
47	10.43	0.71	-1.91	-0.97	-0.74	-6.03	-4.36	-0.39	-0.26	-0.44
48	-5.88	0.72	-3.38	1.43	0.36	-13.41	-3.60	-0.11	0.34	-0.66

Data point reference nos.	**Site 3- Bobadah**					**Site 4 – Rocky Creek**				
	Volterra %	M5 Tree %	RF %	ELM %	ANN-CoM %	Volterra %	M5 Tree %	RF %	ELM %	ANN-CoM %
1	-14.16	-0.80	-0.13	0.19	0.04	-35.39	-0.84	0.23	4.09	0.04
2	-10.54	-3.42	-2.16	-2.58	-2.71	-4.65	-1.59	-2.29	-2.89	-2.46
3	-5.17	-1.22	-0.36	-0.30	-0.34	-3.53	-1.81	1.68	0.39	0.21
4	-9.31	-1.03	-0.46	-0.25	-0.63	-2.50	0.86	-0.19	-0.29	-0.01
5	3.78	1.21	0.45	1.88	1.10	0.56	0.36	0.47	2.13	1.11
6	-8.49	3.43	0.78	2.26	1.79	-5.23	-1.41	-1.93	0.67	-0.70
7	-3.66	-2.56	0.52	0.22	0.38	-4.40	-1.61	-0.48	0.36	-0.11
8	-0.25	-2.10	-0.38	1.11	0.71	-3.33	-1.11	0.95	0.58	0.49
9	-12.30	-1.74	-0.30	-0.08	-0.53	-3.37	0.59	-0.79	-0.40	-0.26
10	-8.26	2.31	0.78	1.00	0.74	0.59	0.61	0.22	2.74	1.36
11	-8.69	-1.32	-1.35	-0.15	-0.33	-18.28	0.24	-1.67	-1.92	-1.59
12	-5.89	-0.71	0.19	0.54	0.56	8.99	1.80	1.02	-0.93	0.50
13	-7.90	-1.02	0.04	-0.08	-0.09	-19.85	3.15	-0.90	2.33	2.19
14	-24.82	-3.15	-2.86	-2.36	-3.30	-21.85	1.57	-2.79	-0.97	-0.59
15	-21.18	-0.89	-0.24	0.00	-0.14	-16.80	2.79	22.80	-4.27	-2.80
16	-8.80	-0.84	0.07	-0.92	-0.64	-3.26	-0.63	-1.25	0.57	-0.52
17	-18.61	-0.13	-0.14	0.24	0.04	0.38	0.06	0.97	1.39	0.83
18	-8.46	-0.71	-1.53	-0.75	-1.16	0.14	-1.57	-1.86	-1.42	-1.21

19	-8.11	-0.78	0.53	0.75	0.44	-23.77	-0.40	-0.86	0.52	-0.52
20	-8.95	1.67	-0.29	0.92	0.34	-0.59	-0.65	-0.93	-0.60	-0.39
21	-7.73	-0.88	-0.26	0.00	0.05	-2.37	0.27	1.28	0.19	0.06
22	-4.23	-0.63	0.22	0.33	0.35	2.58	0.29	0.98	0.48	0.81
23	-1.12	-0.62	-0.42	0.00	-0.29	-10.06	2.42	0.27	-0.89	0.05
24	-41.25	1.13	0.34	1.79	0.03	-30.21	2.79	-0.25	2.61	2.72
25	-10.34	1.26	-0.74	1.18	0.42	-17.81	2.28	0.10	-1.07	-0.03
26	0.54	-1.70	-2.05	-0.69	-1.58	9.28	-0.46	-2.40	-3.38	-1.92
27	4.68	0.81	1.27	2.85	1.24	12.71	3.05	-1.21	1.61	0.96
28	6.92	-0.21	-0.89	-1.31	-1.12	6.23	-0.33	-0.02	-2.18	-0.67
29	6.07	0.56	0.12	0.35	-0.05	3.91	-0.07	0.29	-0.78	0.09
30	-1.64	-0.39	-0.76	1.36	-0.07	0.89	0.43	0.59	3.11	1.32
31	-5.43	-1.09	0.24	-0.12	0.05	-1.54	-0.89	0.90	1.07	0.55
32	-8.62	-0.45	0.17	0.31	0.36	-1.19	-0.99	-0.02	1.11	0.42
33	-7.88	-1.32	-0.71	0.11	0.00	-6.93	-1.22	0.39	1.03	0.41
34	-10.71	-0.38	0.94	0.49	0.24	-12.25	1.67	1.07	1.16	1.21
35	-8.56	-1.20	-1.20	-0.51	-0.97	-27.19	-0.11	0.76	-0.22	-0.16
36	1.18	1.84	0.06	1.13	1.04	-25.14	0.36	-0.03	2.75	0.94
37	-14.61	-1.18	-0.04	-0.24	-0.42	-19.25	0.56	0.07	-0.28	0.18
38	-7.08	-2.87	-1.78	-1.94	-1.88	-10.23	-0.56	-1.65	-1.71	-1.32
39	-2.67	-0.74	1.49	0.14	0.40	-6.67	0.58	0.77	2.88	1.33
40	1.66	0.91	-1.24	0.53	0.07	5.87	0.19	0.41	-1.63	-0.20
41	-9.23	-0.99	-0.11	-0.07	-0.19	12.74	3.20	0.97	-0.25	1.27
42	-11.14	-0.56	1.67	1.88	1.96	-16.22	4.27	3.71	3.43	3.89
43	16.32	0.09	2.56	-0.97	-0.08	12.31	-3.05	-0.55	0.00	-0.16
44	7.43	1.75	-0.88	0.96	0.15	4.76	-2.34	-0.35	-0.59	-0.46
45	-7.26	3.57	-5.45	-0.53	-0.74	-3.24	2.43	-5.32	-4.96	-0.40
46	6.92	5.70	-4.46	0.55	0.22	-0.01	5.89	-5.18	0.19	-0.04
47	16.01	-0.74	1.87	-0.02	-0.14	14.34	0.26	0.55	2.05	0.36
48	13.78	-1.09	3.04	-0.20	0.76	-0.22	-1.66	-0.11	0.28	0.23

Table A4 Percentage deviations of forecasted values from the X=Y line from the multivariate sequential EEMD-ELM, EEMD-MARS and the standalone models MARS and ELM models in forecasting of upper layer soil moisture at the four study sites.

Data point reference nos.	Site 1 - Menindee				Site 2 - Cooinbil			
	MARS %	ELM %	EEMD-MARS %	EEMD-ELM %	MARS %	ELM %	EEMD-MARS %	EEMD-ELM %
1	46.78	6.63	12.68	957.52	265.27	316.86	422.07	166.01
2	42.78	44.05	1.32	10.99	-46.02	-63.62	18.34	3.23
3	105.54	117.84	-9.71	87.39	30.12	21.15	46.85	-41.47
4	113.70	34.25	79.33	37.71	41.23	18.49	39.43	20.61
5	1.20	-19.36	26.73	11.02	-51.67	-45.20	-17.65	-9.24
6	-29.55	-1.22	-20.63	-7.45	0.76	3.47	11.30	-7.60
7	12.04	-11.62	22.45	14.39	-30.96	-48.30	17.70	5.90
8	-4.62	-4.37	-7.69	9.42	-39.89	-27.65	-18.63	-10.05
9	15.13	25.50	-2.70	11.33	25.03	1.36	24.25	0.02
10	14.13	39.73	29.54	49.12	6.12	43.07	8.91	5.59
11	-43.17	-32.78	22.69	20.92	-72.50	-63.74	-34.09	-38.66
12	-43.84	-44.80	8.51	-3.33	7.11	8.75	8.57	6.63
13	-30.73	-38.88	-15.55	-22.16	-31.34	-23.73	-9.93	-7.03
14	17.51	12.06	-2.16	0.90	18.19	24.55	9.04	15.15
15	25.87	10.32	23.48	10.72	-8.41	5.23	10.84	-0.04
16	-1.34	-20.74	0.14	-4.52	-28.19	-20.84	-25.42	-2.39
17	9.05	-7.89	-14.47	-11.02	4.82	-2.51	-17.79	-2.45
18	38.16	9.64	-5.94	-7.90	-4.73	-2.10	-12.81	-8.88
19	41.79	6.40	4.92	-4.58	15.29	16.42	22.46	21.40
20	13.46	-1.62	1.77	-1.96	-43.64	-47.39	-7.52	-13.75
21	2.77	-12.56	-3.43	-11.38	-9.63	-18.63	-15.69	-7.66
22	25.74	8.55	2.40	-0.94	-5.35	7.91	11.58	23.43
23	24.17	6.13	-6.68	-3.87	-10.99	-11.32	-14.38	0.88
24	21.24	2.63	-6.22	-5.72	-9.37	-0.26	-20.95	2.75
25	9.83	8.86	-6.13	-5.75	2.00	8.51	-24.51	3.95
26	3.15	5.06	1.34	-3.81	-1.69	6.74	-21.58	16.52
27	-20.85	-5.58	-6.18	-13.33	-7.42	-3.78	-26.24	4.81
28	-19.26	9.70	-12.99	-12.54	14.76	14.46	-12.92	22.29
29	10.57	-0.48	-14.97	-12.18	21.07	18.75	-12.06	38.00
30	-3.85	-7.55	17.36	15.78	36.66	28.23	-36.61	35.06
31	-60.63	-60.88	10.02	2.16	64.83	45.67	-70.70	49.61
32	-55.03	-45.26	-15.31	-17.88	26.62	4.56	-73.22	10.45
33	30.10	10.73	26.93	23.70	80.55	50.60	-49.87	57.93
34	8.73	-16.54	4.37	-0.66	14.64	-14.27	-69.44	22.70
35	36.28	14.42	12.83	10.09	29.09	11.72	-60.65	28.05
36	27.10	0.76	8.74	11.49	52.49	38.19	-54.97	47.99
37	-21.52	-32.47	-3.60	-5.69	-34.30	-34.12	-58.68	-6.23
38	13.39	2.63	10.87	9.71	36.77	15.30	-32.21	27.98
39	19.10	-13.27	-0.17	5.17	-9.90	-8.89	-41.52	42.22
40	30.22	16.18	-5.18	29.49	-29.09	-24.08	-54.85	6.19

41	28.17	-24.10	1.04	1.74	63.18	71.28	-38.95	53.63
42	35.13	13.09	4.62	48.53	37.95	58.62	-77.33	52.05
43	98.64	16.88	11.72	15.23	61.94	85.35	-70.33	55.25
44	112.16	-74.00	18.81	-33.15	49.43	57.41	-55.11	45.56
45	36.36	34.98	-14.48	6.88	21.71	30.95	-15.90	8.99
46	-15.97	-13.21	-26.28	-12.79	-57.58	-47.32	-46.96	-31.24
47	3.21	-6.58	-14.19	-13.25	42.07	36.57	-24.58	11.65
48	5.34	17.60	-33.09	-17.91	15.06	56.62	-40.41	12.19
49	35.71	45.25	-48.35	-1.60	54.37	125.27	-9.92	52.09
50	60.95	21.07	-36.80	-15.23	68.83	109.82	12.66	107.20
51	156.99	108.42	168.94	197.74	204.25	280.26	283.89	446.77
52	-51.46	-60.70	46.01	20.89	-67.98	-64.62	-3.08	1.29
53	-41.30	-29.33	-32.36	-22.18	-32.35	-21.38	-30.33	-17.50
54	33.36	28.74	19.09	29.54	42.20	15.89	6.62	37.77
55	34.09	35.28	54.23	49.80	-5.53	15.58	-71.09	29.21
56	24.27	7.72	81.22	49.02	43.02	53.76	94.74	139.74
57	23.85	5.10	51.49	50.77	-67.74	-59.60	-16.10	-5.52
58	-23.35	-40.21	13.18	-11.44	-6.28	-16.14	-38.90	-15.33
59	-1.51	14.70	-14.92	-24.10	-11.98	-12.37	-35.92	-16.55
60	26.86	45.96	-26.90	-26.94	32.64	55.40	-57.03	2.87
61	59.39	64.13	-38.21	-17.38	35.86	93.00	49.11	52.39
62	61.79	-5.76	-47.06	-80.88	80.01	98.32	41.56	79.80
63	108.01	40.78	-78.28	-92.70	174.75	172.85	58.85	134.34
64	158.72	133.76	-66.91	-36.08	218.89	182.03	52.53	175.52
65	21.80	49.63	-13.35	10.04	14.32	-13.59	-44.93	2.55
66	-48.71	-55.75	-7.95	-23.86	-18.74	-20.41	-35.00	-11.36
67	-22.47	-17.22	-14.53	-19.58	-5.36	-18.62	-27.92	-19.32
68	21.46	1.85	4.95	-1.62	-12.56	-1.24	-8.07	1.35
69	35.30	16.00	9.96	2.99	-0.56	15.82	0.41	10.03
70	25.62	15.68	9.78	4.04	5.87	9.73	-18.34	-11.34
71	-5.94	-5.85	-1.03	-13.95	10.78	13.33	-28.22	-22.03
72	9.45	12.75	-4.04	-15.58	33.22	22.03	-34.29	-8.02
73	37.37	39.56	-8.64	-7.40	18.08	7.52	-5.92	1.07
74	-13.63	11.66	6.43	-6.18	30.55	18.68	56.70	39.80
75	-39.47	-36.61	-18.31	-24.63	-65.68	-69.13	-27.84	-36.37
76	-16.44	-10.40	-7.72	-15.40	9.22	13.40	4.32	8.78
77	4.79	-1.36	-6.94	-7.96	11.29	14.95	15.92	17.35
78	2.65	-0.57	0.75	-1.48	-19.42	-6.24	24.25	15.14
79	-8.68	-13.65	-12.22	-13.07	-26.57	-29.16	-26.65	-11.60
80	-9.80	4.43	-4.32	-11.14	-8.75	-3.05	-14.88	-3.82
81	20.07	8.94	-9.35	-10.27	-4.78	-1.41	1.31	-3.28
82	12.61	2.58	-9.46	-14.58	-14.00	-9.53	-2.79	-3.35
83	1.72	25.31	-4.26	-2.56	-22.19	-4.70	-8.94	-6.18
84	0.66	7.08	22.20	9.82	9.07	16.25	11.39	7.72
85	-53.47	-50.10	-24.60	-29.69	-47.45	-39.50	-28.96	-23.74
86	10.45	1.69	4.80	-1.37	-1.46	0.93	-15.06	-2.53
87	31.63	12.08	6.01	7.58	-0.58	-5.32	-10.46	2.31
88	38.89	18.11	24.07	15.01	7.83	11.32	-9.37	11.28
89	28.34	0.26	25.14	14.09	17.11	23.32	-15.92	3.22

90	8.18	-7.34	22.58	12.95	57.04	66.60	-27.01	70.12
91	4.29	-46.60	23.35	8.46	70.79	2.78	57.66	138.68
92	17.88	54.20	15.59	33.24	-47.66	-56.08	-35.58	-36.80
93	-13.86	-21.50	-18.83	-26.09	30.63	22.35	-4.45	-10.94
94	31.27	57.50	31.64	36.66	-7.57	-2.38	17.84	30.35
95	-67.53	-65.20	-27.97	-34.20	-65.40	-66.32	-29.79	-30.52
96	-4.58	-1.56	-4.44	-4.65	26.27	7.93	3.35	36.21
97	31.52	36.40	17.22	30.27	16.16	-8.35	-36.25	47.15
98	15.23	-5.71	43.44	27.69	38.12	13.43	72.65	144.53
99	35.52	30.93	27.30	21.14	105.32	114.40	107.52	283.06
100	70.64	43.78	2.26	-13.76	77.34	89.19	109.24	217.26
101	145.22	182.25	45.73	76.80	221.78	274.57	371.08	291.31
102	-16.51	19.14	-5.99	-3.68	-56.60	-57.58	-17.49	-26.23
103	-14.93	1.03	3.27	-15.25	40.99	43.77	0.53	-12.63
104	17.59	70.08	1.21	34.69	-11.91	8.82	-38.94	-30.10
105	22.88	26.48	17.30	8.86	38.79	100.90	40.00	50.78
106	-15.22	18.12	0.99	9.14	-27.23	7.50	70.61	48.98
107	-16.46	27.36	-4.93	4.20	-37.41	-33.13	-1.41	-7.86
108	3.19	18.74	-3.70	-1.03	-16.70	-1.54	-10.77	-10.14
109	33.60	38.32	7.63	18.72	63.69	29.33	19.85	36.40
110	15.88	39.30	3.79	22.81	45.37	50.32	-27.81	89.81
111	70.98	47.71	20.66	39.57	122.53	135.08	197.73	205.43
112	103.14	113.72	61.26	81.77	140.56	95.33	136.29	160.24
113	-21.79	-15.66	11.82	6.09	96.98	33.02	95.30	55.06
114	-30.65	17.99	-20.40	-13.81	-12.35	-34.95	-8.04	-38.58
115	26.67	29.84	1.69	5.85	15.86	21.33	-12.74	-31.63
116	18.11	12.09	1.24	4.99	10.11	46.34	-26.27	-34.63
117	31.20	32.59	1.77	16.48	38.93	57.08	-33.43	-38.81
118	15.02	8.84	-15.85	1.99	87.90	65.09	9.31	-20.98
119	21.83	38.01	17.21	42.99	151.19	91.29	80.25	63.46
120	-46.19	-37.71	5.85	12.11	29.20	-13.95	79.90	62.89
121	-40.97	-35.59	0.03	-7.55	-6.11	-33.10	26.43	4.33
122	-14.11	-27.03	-8.37	-8.32	-41.79	-49.60	-26.94	-26.25
123	28.07	9.77	16.39	22.11	33.04	35.97	38.33	31.98
124	4.54	-17.18	-1.66	0.03	-35.48	-29.72	5.60	-8.16
125	5.57	-9.72	-7.97	-8.93	-21.90	-26.32	-6.89	-22.18
126	21.27	6.10	-8.67	-0.44	-4.13	-6.00	1.44	-21.30
127	3.31	-10.33	-2.05	0.89	-8.25	13.39	29.99	-9.70
128	-17.86	-30.31	-13.90	-12.99	-49.08	-45.29	-7.67	-29.84
129	-1.34	4.64	9.36	9.76	30.29	37.43	34.35	25.64
130	4.27	-10.35	-6.23	-1.95	-30.86	-19.17	-4.36	-22.93
131	1.30	-0.74	-7.58	-1.77	-7.88	-6.74	-4.34	-14.46
132	-4.19	-18.08	-17.90	-17.68	-15.48	-1.41	13.65	-23.01
133	30.61	8.10	-9.33	-3.90	-3.70	-3.79	38.04	-5.10
134	-24.29	-37.42	-23.09	-23.45	-37.26	-28.71	23.19	-16.39
135	-5.48	-7.20	-3.27	-4.69	22.70	29.47	80.13	18.41
136	22.94	1.77	2.19	4.11	-20.50	-1.05	102.35	8.41
137	3.19	-20.32	-19.35	-14.56	-21.67	-25.62	62.90	12.82
138	-9.86	-22.24	-20.12	-20.56	-49.07	-37.17	16.41	-10.12

139	-7.91	-17.70	-23.47	-18.44	-8.10	-6.24	30.08	-0.22
140	-21.49	-16.87	-2.23	-7.26	-14.34	-4.84	39.40	-7.31
141	-48.64	-41.80	-24.39	-20.00	-47.46	-34.71	6.57	-18.92
142	-38.32	14.89	13.40	21.74	-13.97	-1.91	16.33	8.56
143	-21.89	1.47	10.40	7.80	-21.25	-10.60	-25.58	-0.79
144	17.80	21.13	20.95	15.63	16.76	37.33	-58.33	-4.97
145	38.81	5.37	10.07	29.53	-29.70	-8.14	-113.28	-31.81
146	26.72	-19.08	25.74	37.75	28.34	42.33	-102.91	45.10
147	15.80	-75.85	51.09	52.09	25.46	2.79	28.35	186.47
148	-2.70	-90.50	61.22	37.21	-29.45	-63.03	14.31	101.28
149	-3.56	-85.43	53.17	33.30	-4.95	-32.90	-26.58	7.57
150	-5.26	-88.37	44.74	21.04	-37.29	-22.64	-74.87	-5.96
151	29.19	-65.53	56.34	15.19	45.53	59.53	-117.69	59.33
152	4.42	-108.61	40.18	26.25	-2.70	16.78	-121.14	134.51
153	-38.81	-70.08	-25.63	-32.93	-48.65	-45.91	-97.74	26.84
154	-3.98	-28.69	-28.92	-30.49	27.71	25.97	-114.03	15.89
155	-31.94	-70.86	-46.71	-50.96	-18.46	12.90	-186.40	9.42

Data point reference nos.	Site 3 - Fairfield				Site 4 - Bodangora			
	MARS %	ELM %	EEMD-MARS %	EEMD-ELM %	MARS %	ELM %	EEMD-MARS %	EEMD-ELM %
1	74.43	129.60	-34.60	-59.87	25.43	46.34	2.08	20.59
2	242.05	106.93	-102.78	-64.51	40.77	69.59	-24.84	-1.85
3	218.80	222.93	58.49	-17.58	0.21	27.13	-5.25	-15.19
4	328.58	175.72	124.34	14.56	133.52	154.06	22.24	11.80
5	407.79	245.70	284.93	142.67	62.23	51.99	177.09	179.38
6	-15.49	2.86	-24.55	-38.94	-36.72	-47.89	-12.61	-7.82
7	7.88	2.43	-24.69	-29.59	-15.79	-20.68	-21.90	-22.42
8	18.36	19.13	-30.36	-26.13	4.02	-1.41	-13.14	-8.36
9	75.23	52.30	-1.84	-9.84	19.55	16.75	-5.21	1.99
10	86.64	67.18	37.37	42.20	35.60	23.89	23.22	24.99
11	-41.92	-48.12	-23.90	-24.06	-45.40	-55.68	-23.11	-33.19
12	-6.78	-8.85	-33.35	-24.96	-0.52	-1.24	8.84	3.53
13	13.20	18.02	-22.54	-15.97	-10.95	-10.00	-12.70	-11.75
14	29.42	19.25	-11.04	-8.71	12.04	14.07	-5.06	-0.98
15	54.20	28.37	3.66	10.01	49.25	28.50	21.71	19.77
16	29.04	29.73	0.16	5.83	-4.49	-13.51	11.31	-3.89
17	31.31	42.77	-9.65	2.93	-17.05	-3.43	-9.48	-15.97
18	20.26	34.45	-17.87	-5.29	9.60	11.10	-1.94	-1.52
19	62.74	80.26	10.55	18.96	51.48	35.78	11.30	25.09
20	9.97	44.39	15.68	11.38	12.54	2.16	20.74	20.78
21	-33.16	-0.47	-37.25	-33.01	-33.90	-20.19	-20.15	-23.60
22	24.39	-4.23	-11.79	-5.22	-12.36	0.14	10.84	6.33
23	10.98	6.06	-23.89	-15.70	-13.43	-2.66	-8.02	-4.99
24	50.37	23.45	-0.48	14.00	-2.66	2.77	0.30	0.27
25	61.89	44.91	-10.32	5.64	1.34	9.74	-0.89	-3.34

26	60.09	52.06	7.45	24.60	23.89	21.90	22.91	14.71
27	-19.87	-20.21	-35.34	-23.39	-28.36	-29.18	-19.69	-22.52
28	22.77	39.89	-19.24	-1.41	-8.44	3.21	-3.00	-8.11
29	34.72	63.10	14.67	19.75	0.09	9.88	-3.64	-4.59
30	12.74	2.18	26.65	35.24	37.56	20.14	5.81	10.91
31	-15.96	-5.88	26.55	25.58	37.64	25.11	28.05	20.65
32	-33.97	-26.32	-35.81	-28.13	-35.50	-17.42	-16.57	-27.99
33	16.01	2.80	-0.80	11.99	36.34	29.68	20.38	4.53
34	44.58	12.89	13.10	30.21	39.70	37.08	14.80	6.31
35	57.99	23.57	29.10	37.61	-3.64	10.84	3.90	-13.05
36	78.47	44.48	39.97	48.49	53.56	89.47	15.75	-5.89
37	-26.71	-39.66	-26.53	-15.21	-13.51	-0.70	2.73	-19.39
38	27.92	45.88	-32.40	-4.34	38.88	77.36	40.11	6.37
39	98.26	58.69	-16.04	3.83	80.10	71.70	71.56	38.91
40	135.73	86.31	-27.34	-0.41	-58.13	-58.13	-31.81	-45.16
41	126.99	62.28	5.36	58.30	45.09	59.23	6.64	-18.09
42	174.31	91.70	-24.08	14.22	111.96	110.98	27.94	-0.06
43	72.82	84.69	-51.38	-26.41	83.43	64.37	53.84	22.44
44	210.79	116.77	-36.79	-22.20	86.52	79.25	45.64	22.41
45	415.84	268.51	-41.30	11.51	112.53	104.63	74.27	46.76
46	212.43	117.95	-53.26	-1.04	0.22	-5.86	39.89	21.24
47	70.22	114.38	-63.11	-85.11	-22.44	-16.32	-0.09	-11.03
48	173.04	163.89	-1.58	-90.14	7.30	5.75	-5.05	-14.03
49	271.00	186.43	121.22	7.86	48.45	49.89	23.35	20.00
50	-27.62	-38.08	-26.31	-21.33	-28.78	-28.24	-17.94	-14.89
51	25.65	23.29	-35.44	3.42	42.41	39.28	8.36	7.41
52	77.32	34.77	2.19	-23.72	-24.93	-20.41	-13.15	-24.40
53	47.59	43.10	-36.04	-23.39	-14.18	-14.84	0.31	-9.65
54	105.53	78.76	16.04	36.27	19.65	17.32	29.18	29.14
55	-10.43	-20.05	-26.79	-14.04	-11.39	-7.90	-12.03	-15.05
56	60.52	35.75	-29.57	-10.56	45.02	31.06	-1.57	-17.73
57	124.86	63.48	-10.76	3.58	27.96	21.86	-0.58	-27.17
58	174.10	106.79	-48.73	12.27	99.23	89.83	40.21	27.55
59	218.18	190.85	-43.10	28.84	43.12	19.08	9.76	2.04
60	354.74	284.64	-64.04	-0.85	87.34	67.34	-0.66	-9.91
61	199.97	134.49	-68.80	-21.03	54.03	20.79	-2.55	-22.35
62	229.51	199.80	-62.47	-21.90	114.05	67.78	17.15	-1.74
63	381.69	309.99	-38.50	-11.61	81.09	9.52	20.98	-3.62
64	154.78	134.99	-31.54	-0.88	44.27	11.13	54.95	28.86
65	-10.45	70.30	-48.42	-47.12	-33.76	-34.23	-20.31	-30.00
66	16.99	13.71	-10.70	-17.53	38.01	31.79	35.12	18.13
67	-34.22	-21.84	-45.24	-40.83	-32.12	-35.79	-28.86	-36.91
68	31.51	10.45	-10.90	-6.30	0.35	10.51	17.54	11.70
69	48.75	32.58	1.23	4.66	50.29	26.48	26.89	19.39
70	62.01	35.94	23.43	27.45	63.09	37.76	78.60	48.32
71	-12.72	-14.40	-17.61	-24.93	-42.60	-43.16	-17.59	-28.23
72	1.35	19.53	-17.52	-14.27	1.13	7.69	-12.64	-17.91
73	-14.14	-6.45	-28.96	-28.27	-7.53	-1.33	-10.00	-6.85
74	30.25	-2.77	3.05	10.71	22.40	7.60	24.50	25.41

75	-33.57	-44.90	-23.78	-19.95	-42.67	-49.45	-27.14	-31.26
76	-4.12	-14.14	-23.13	-18.26	-10.57	4.39	3.75	1.68
77	-11.25	-21.75	-17.79	-11.89	1.15	5.07	3.90	9.26
78	3.37	-16.79	-7.55	-0.16	6.01	-3.23	28.78	20.62
79	-11.06	-27.77	-15.30	-7.27	-42.38	-35.22	-21.08	-30.12
80	-7.94	-29.78	-21.78	-16.97	-31.42	-12.28	-1.27	-11.94
81	14.03	-0.62	-20.70	-17.07	-1.71	12.93	-5.59	-6.26
82	31.80	-4.55	-8.61	-1.90	25.91	13.75	12.04	2.95
83	55.95	10.02	4.85	12.08	38.68	30.61	27.96	13.35
84	56.36	22.60	67.51	69.45	42.32	43.79	55.72	28.56
85	-68.69	-65.43	-46.40	-47.27	-54.20	-42.40	-25.28	-40.47
86	59.35	46.40	-11.11	1.96	10.15	15.98	8.87	-4.02
87	53.64	20.30	9.61	14.41	33.39	22.72	2.52	-0.66
88	67.33	23.36	-21.53	-12.68	68.11	60.54	39.22	11.08
89	81.51	72.66	-16.30	-7.02	47.22	88.04	67.52	10.43
90	88.99	39.58	10.95	-3.93	-17.05	44.06	146.40	45.06
91	192.40	145.28	76.62	17.63	59.82	75.19	197.89	84.34
92	129.02	84.31	102.04	74.82	-2.97	11.09	29.57	6.89
93	4.79	-4.58	43.06	48.40	-14.63	0.60	24.40	6.31
94	-21.40	-14.45	13.95	32.29	-21.87	-3.62	19.64	7.05
95	-45.92	-48.38	-37.59	-26.88	-43.28	-39.72	-14.25	-25.12
96	18.80	24.01	-13.40	-16.51	-3.51	-5.08	1.28	-8.22
97	22.29	30.48	2.38	-3.61	8.83	9.28	-3.70	-6.64
98	99.00	53.25	-14.43	-17.15	86.10	88.26	23.83	0.80
99	132.12	53.63	17.87	45.33	25.19	82.66	131.67	50.87
100	138.32	95.20	33.94	81.33	-69.60	-25.86	40.90	5.65
101	111.79	84.62	18.43	22.72	-2.70	39.75	16.08	13.63
102	134.71	98.97	73.12	-8.27	-41.22	-28.06	-32.12	-30.41
103	68.23	119.22	77.74	-5.99	41.91	44.80	-3.28	6.14
104	-47.31	-28.98	-38.31	-40.54	13.27	7.34	-18.05	0.60
105	75.21	58.91	13.24	-22.17	23.45	25.12	41.15	35.09
106	88.92	78.57	89.97	96.08	17.24	24.31	41.00	34.69
107	-55.19	-63.71	-43.14	-24.50	-40.82	-38.09	-20.93	-28.18
108	5.51	-8.60	-34.04	-26.99	18.31	11.17	6.67	-3.82
109	0.01	21.00	-30.26	-10.84	49.26	42.29	-10.08	-2.39
110	54.12	41.16	-34.06	-35.84	88.16	96.16	58.74	21.72
111	109.94	58.59	7.90	20.11	122.62	114.08	148.63	40.22
112	170.70	111.27	-18.36	33.41	255.22	167.95	215.74	69.50
113	203.00	179.16	45.98	74.96	288.41	134.47	230.35	170.64
114	79.32	92.92	-12.24	12.40	4.68	-43.19	-16.37	-17.12
115	105.80	179.07	1.48	1.14	46.75	54.88	-8.60	-17.12
116	26.02	27.03	-27.34	-28.74	100.69	57.01	-18.83	-24.22
117	77.78	130.48	-18.79	-12.10	175.72	99.77	8.29	-3.89
118	-22.08	-4.37	-47.03	-36.70	299.39	138.10	106.83	36.03
119	12.19	20.72	-35.16	-22.78	42.60	-31.86	51.32	8.01
120	55.81	56.78	12.06	14.55	28.62	18.60	57.10	20.96
121	-27.02	-4.18	-35.23	-34.53	-16.05	-4.55	-8.79	-18.14
122	9.82	12.38	-24.85	-24.78	-19.29	-31.22	-17.33	-29.86
123	49.03	41.65	26.05	23.53	31.94	35.64	29.67	19.60

124	-10.46	0.70	22.40	17.82	12.29	-6.22	19.99	14.43
125	-35.19	-34.70	-20.26	-21.53	-22.74	-19.40	0.87	-4.90
126	-11.46	-20.94	-30.89	-23.89	-32.67	-24.42	-12.03	-18.67
127	-18.76	-27.64	-6.40	-7.31	-32.96	-11.75	17.84	8.66
128	-30.21	-33.10	-26.67	-20.66	-48.17	-40.01	-10.12	-20.86
129	-18.24	-13.83	3.70	5.47	-35.44	-2.88	48.35	32.36
130	-13.33	-30.26	-14.46	-6.60	-28.33	-25.75	-11.59	-11.79
131	12.90	-6.49	-5.56	-5.32	-7.12	7.47	10.95	-1.82
132	21.42	-14.16	-7.25	-4.69	-27.24	-33.98	-9.62	-27.40
133	39.27	18.47	-1.26	5.86	11.37	25.78	51.29	18.69
134	-39.32	-55.74	-26.47	-22.59	6.19	-6.38	-6.06	-9.42
135	19.06	11.44	-10.43	-13.97	48.25	51.40	16.68	9.12
136	27.31	-0.02	22.61	17.46	79.88	34.18	11.74	8.57
137	-34.48	-37.62	-10.68	-13.28	16.36	-14.00	9.35	-9.54
138	-32.78	-40.13	-26.33	-30.07	-10.03	-23.20	-3.15	-19.25
139	-3.64	-7.00	-27.02	-22.88	16.52	15.08	28.31	4.89
140	-33.51	-31.38	-28.74	-27.84	-0.62	-5.50	-2.49	0.32
141	-14.03	-16.09	-36.69	-33.72	-20.95	-17.36	-0.14	-13.86
142	9.42	12.27	-8.89	6.30	-14.35	2.69	13.32	-6.15
143	9.88	-6.16	-15.68	-9.64	1.91	12.32	-8.96	-9.83
144	57.30	33.28	-23.75	-27.75	56.22	33.51	3.53	0.17
145	-1.24	-21.89	-39.46	-39.95	-19.78	-35.79	-22.45	-38.05
146	46.71	52.78	-30.34	-24.87	44.74	8.94	9.29	-15.50
147	80.65	86.75	-10.72	-4.73	65.69	5.99	27.87	-4.78
148	-3.26	-14.76	-55.31	-35.79	-4.77	-47.57	12.01	-10.33
149	66.44	105.81	-63.71	-44.34	38.64	-4.33	2.65	-16.35
150	118.78	114.12	-75.78	-43.07	139.29	47.24	5.01	2.37
151	168.36	192.03	-132.71	-41.21	207.82	107.66	135.02	109.88
152	335.94	294.50	-111.12	8.70	216.79	129.32	269.22	192.16
153	21.74	71.41	-90.49	-41.12	-13.31	-35.68	23.48	4.01
154	21.24	62.39	-76.99	-56.80	42.53	15.80	-8.51	-23.17
155	34.92	21.42	-63.63	-64.62	88.99	45.43	-29.65	-26.73