EVAPORATION AND SOIL MOISTURE PREDICTION WITH ARTIFICIAL INTELLIGENCE AND DEEP LEARNING METHODS

A Thesis submitted by

W.J.M. LAKMINI PRARTHANA JAYASINGHE

B.Sc., M.Phil.

For the award of

Doctor of Philosophy

2023

ABSTRACT

Understanding future changes and predicting hydrological variables well in advance is practically useful in water resources and drought management measures. This doctoral thesis presents the new methodologies and the findings based on three primary objectives that aim to build artificial intelligence and deep learning hybrid models to forecast drought-related hydrological variables comprised of evaporation, evapotranspiration, and soil moisture within the key drought-prone regions in Queensland, Australia. Data preprocessing techniques that involve feature selection and data decomposition to reveal the patterns or trends in modeling data are used in the model hybridization stage where standalone models are integrated with these techniques and the significance of their influence in enhancing the model performances are tested. In the first objective, the Long Short-Term Memory (LSTM) predictive model is hybridized with the Neighborhood Component Analysis (NCA) feature selection technique to enhance the model's predictive efficacy that aims to accurately predict pan evaporation (*Ep*). The second objective aims to develop novel methods to forecast reference evapotranspiration (*ET*) and is achieved by hybridizing the LSTM model with Boruta-Random Forest (Boruta) feature selection technique and the Multivariate Empirical Mode Decomposition (MEMD) technique to further improve the efficacy. In the third objective, the 1-, 14-, and 30-days ahead soil moisture (*SM*) within the topsoil layer (0–10 cm depth) is forecasted by employing a hybrid deep learning forecasting model built using LSTM network coupled with Maximum Overlap Discrete Wavelet Transform (moDWT) data decomposition method and Least Absolute Shrinkage and Selection Operator (Lasso) feature selection method. When compared with benchmark models, all the hybrid models developed in this study registered a comparatively high performance with low error performance metrics to demonstrate their usefulness in forecasting *Ep, ET,* and *SM* values in the present study region. In the practical sense, as the models developed in this study provide accurate estimations, their capabilities can undoubtedly be employed to successfully manage water resources and drought events. Further, this doctoral study shows that artificial intelligence and deep learning models developed in this study could be a significant forward step in contributing to the advancement of data-driven hydrological forecasting methods that may be useful for understanding the future trend of hydrological variables. The outcomes and implications thus contributed to the advancement of science while creating socio-economic benefits due to their usefulness in water resources and drought event management.

CERTIFICATION OF THESIS

I W.J.M. Lakmini Prarthana Jayasinghe declare that the PhD Thesis entitled "*Evaporation and soil moisture prediction with artificial intelligence and deep learning methods"* is not more than 100,000 words in length including quotes and exclusive of tables, figures, appendices, bibliography, references, and footnotes.

This Thesis is the work of W.J.M. Lakmini Prarthana Jayasinghe except where otherwise acknowledged, with the majority of the contribution to 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.

Date: 06/06/2023

Endorsed by:

Professor Ravinesh C. Deo
Principal Supervisor

Dr. Nawin Raj
Associate Supervisor

Dr. Afshin Ghahramani
External Supervisor

Dr Sujan Ghimire
External Supervisor

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

STATEMENT OF CONTRIBUTION

The doctoral research thesis has produced three quartile 1 (*Q*1 ranked) journal publications completed during the PhD candidature.

Field of Research (FOR): The focus of this doctoral thesis is on the national priority area: 050205 - *Environmental Management*; 461103 - *Deep learning*; 460501 - *Data engineering and data science*.

Articles 1, 2, and 3 are the primary (core) parts of this thesis. The following presents the student contributions and the contributions of the co-authors of the publications.

Article 1: Chapter 4

Jayasinghe W. J. M. L. P., Deo R.C., Ghahramani A., Ghimire S., Raj N. (2022). Development and evaluation of hybrid deep learning long short-term memory network model for pan evaporation estimation trained with satellite and ground-based data, *Journal of Hydrology*, 127534. (https://doi.org/10.1016/j.jhydrol.2022.127534) (**Scopus Ranked *Q1; Impact Factor: 6.708, SNIP: 1.857; 94th percentile* in category: Water Science and Technology**).

The percentage contributions for this paper are W. J. M. Lakmini Prarthana Jayasinghe 75%, Ravinesh C. Deo 10%, Nawin Raj 5%, Afshin Ghahramani 5%, and Sujan Ghimire 5%.

Author	Task performed
W.J.M. Lakmini Prarthana Jayasinghe (PhD Candidate)	Exploring the methodologies in literature, data collection and analysis, programming, model development and implementation, preparation of tables and figures, writing and revising of the manuscript.
Ravinesh C Deo **(**Principal Supervisor**)**	Supervising and assisting with developing model concepts, providing beneficial advice, information, and comments, editing and preparing the manuscript for submission, guidance for selecting suitable journals, and holding the co-authorship for the manuscript.

Nawin Raj (Associate Supervisor)	Editing and proofreading of the manuscript, holding the co-authorship for manuscript
Afshin Ghahramani (External Supervisor)	Editing and proofreading of the manuscript, holding the co-authorship for manuscript
Sujan Ghimire (External Supervisor)	Editing, advice in methods, proofreading, interpretation of results, holding the co-authorship for manuscript

Article 2: Chapter 5

Jayasinghe W. J. M. L. P., Deo R.C., Ghahramani A., Ghimire S., Raj N. (2021). Deep Multi-Stage Reference Evapotranspiration Forecasting Model: Multivariate Empirical Mode Decomposition Integrated with the Boruta-Random Forest Algorithm, *IEEE, Access*, 166695. (https://doi.org/10.1109/ACCESS.2021.3135362) (***Scopus Ranked Q1; Impact Factor: 3.37 and SNIP 1.326; 97th percentile* in category: General Engineering)**.

The percentage contributions for this paper are W. J. M. Lakmini Prarthana Jayasinghe 75%, Ravinesh C. Deo 10%, Nawin Raj 5%, Afshin Ghahramani 5%, and Sujan Ghimire 5%.

Author	Task performed
W.J.M. Lakmini Prarthana Jayasinghe (PhD Candidate)	Exploring the methodologies in literature, data collection and analysis, programming, model development and implementation, preparation of tables and figures, writing and revising of the manuscript.
Ravinesh C Deo **(**Principal Supervisor**)**	Supervising and assisting with developing model concepts, providing beneficial advice, information, and comments, editing and preparing the manuscript for submission, guidance for selecting suitable journals, and holding the co-authorship for the manuscript.
Nawin Raj (Associate Supervisor)	Editing and proofreading of the manuscript, holding the co-authorship for manuscript
Afshin Ghahramani (External Supervisor)	Editing and proofreading of the manuscript, holding the co-authorship for manuscript

Sujan Ghimire (External Supervisor)	Editing, advice in methods, proofreading, interpretation of results, holding the co-authorship for manuscript

Article 3: Chapter 6

Jayasinghe W. J. M. L. P., Deo R.C., Ghahramani A., Ghimire S., Raj N. Soil moisture forecasting at 1 day, 14 days and 30 days ahead horizon with 3-phase deep learning Long Short-Term Memory network, wavelet, and Lasso regression moDWT-Lasso-LSTM approach. This paper is submitted to the Journal of *Stochastic Environmental Research and Risk Assessment* and is under review process. (***Scopus Ranked Q1; Impact Factor:4.2; 82nd percentile*** **in category: Engineering)**.

The percentage contributions for this paper are W. J. M. Lakmini Prarthana Jayasinghe 75%, Ravinesh C. Deo 10%, Nawin Raj 5%, Afshin Ghahramani 5%, and Sujan Ghimire 5%.

Author	Task performed
W.J.M. Lakmini Prarthana Jayasinghe (PhD Candidate)	Exploring the methodologies in literature, data collection and analysis, programming, model development and implementation, preparation of tables and figures, writing and revising of the manuscript.
Ravinesh C Deo **(**Principal Supervisor**)**	Supervising and assisting with developing model concepts, providing beneficial advice, information, and comments, editing and preparing the manuscript for submission, guidance for selecting suitable journals, and holding the co-authorship for the manuscript.
Nawin Raj (Associate Supervisor)	Editing and proofreading of the manuscript, holding the co-authorship for manuscript
Afshin Ghahramani (External Supervisor)	Editing and proofreading of the manuscript, holding the co-authorship for manuscript
Sujan Ghimire (External Supervisor)	Editing, advice in methods, proofreading, interpretation of results, holding the co-authorship for manuscript

ACKNOWLEDGEMENTS

Firstly, I want to convey my sincere gratitude to Prof. Ravinesh C Deo, my principal research supervisor for directing me to the field of Artificial Intelligence and Data Science, a new stunning pathway to success, that I never touched and never thought of when I started this doctoral study. Further, he continuously motivated me on this big journey, while providing essential technical support and very quick responses to all my queries. I am highly impressed with the role of my supervisory team and would like to offer very special thanks to associate supervisor Dr. Nawin Raj, and external supervisors Dr. Afshin Ghahramani and Dr. Sujan Ghimire who spent their valuable time for proofreading and upgrading the quality of the research papers and thesis based on this study. Further, I kindly appreciate the big help given by Dr. Barbara Hormes of the English Angels Program, University of Southern Queensland, Australia, for making language corrections in my research papers and she generously spent her valuable time when I need her proficiency.

Further, I would like to express my heartfelt gratitude to my parents, husband, two daughters, and all other relatives who patiently tolerate my failure to be with them enough during this study period while further appreciating their encouragement and generous support. I also respect all my work colleagues at the Wayamba University of Sri Lanka and Advanced Data Analytics: Environmental Modelling & Simulation Research Group, including Dr. Masrur Ahamed for their wonderful and unforgettable support, encouragement, and motivation throughout my studies.

I would like to extend my heartfelt gratitude to the Scientific Information for Landowners (SILO-Queensland), QLD, Bureau of Meteorology, Australia, and GIOVANNI for continuously collecting valuable data over the years, maintaining big data banks and providing free access to those data for many research including the current study.

Finally, I extend my gratitude to the University of Southern Queensland (UniSQ) for providing me with the International PhD Fee and Research Training Program (RTP) Stipend Scholarships (2020-2023) and to Wayamba University of Sri Lanka (WUSL) for granting me paid study leave. It is a great privilege for me to complete my higher studies at the University of Southern Queensland, which is a world-class distinguished university with great academic professionals. This experience indeed will allow me to explore many avenues around the globe to grow up in my career in the future and so thanking for giving me this great opportunity.

DEDICATION

I would like to dedicate this work to my wonderful parents, my husband and two little daughters who understand my hunger for learning and make many sacrifices for achieving this big goal, and all my teachers, who generously give their knowledge to me and always show me the right direction, all my friends and relations who always encouraging and supporting me in my studies, my principal supervisor and other supervisors who always watching my research work very closely while advising me to make this research work successful.

LIST OF PUBLICATIONS AND OTHER RESEARCH OUTCOMES

Publications

1. **Jayasinghe W. J. M. L. P.,** Deo R.C., Ghahramani A., Ghimire S., Raj N. Development and evaluation of hybrid deep learning long short-term memory network model for pan evaporation estimation trained with satellite and ground-based data. *Journal of Hydrology* 2022, 127534. (https://doi.org/10.1016/j.jhydrol.2022.127534) (**Scopus Ranked *Q1; Impact Factor: 6.708, SNIP: 1.857; 94th percentile* in category: Water Science and Technology**).

2. **Jayasinghe W. J. M. L. P.,** Deo R.C., Ghahramani A., Ghimire S., Raj N. Deep Multi-Stage Reference Evapotranspiration Forecasting Model: Multivariate Empirical Mode Decomposition Integrated with the Boruta-Random Forest Algorithm. *IEEE, Access* 2021, 166695. (https://doi.org/10.1109/ACCESS.2021.3135362) (***Scopus Ranked Q1; Impact Factor: 3.37 and SNIP 1.326; 97th percentile* in category: General Engineering)**.

3. **Jayasinghe W. J. M. L. P.,** Deo R.C., Ghahramani A., Ghimire S., Raj N. Soil moisture forecasting at 1 day, 14 days and 30 days ahead horizon with 3-phase deep learning Long Short-Term Memory network, wavelet, and Lasso regression moDWT-Lasso-LSTM approach. This paper is re-submitted to the Journal of *Stochastic Environmental Research and Risk Assessment and is under review process. (**Scopus Ranked Q1; Impact Factor:4.2; 82nd percentile in category: Engineering**)*.

Symposiums

1. **Jayasinghe W. J. M. L. P.,** Deo R.C., Ghahramani A., Ghimire S., Raj N**.** Deep hybrid Long Short-Term Memory network algorithm for Pan Evaporation prediction with Neighborhood Component Analysis. *Higher Degree Research Symposium, School of Science, University of Southern Queensland, Australia* **7 December 2020**.

2. **Jayasinghe W. J. M. L. P.,** Deo R.C., Ghahramani A., Ghimire S., Raj N**.** Daily deep multi-stage Reference Evapotranspiration forecasting model. *Higher Degree Research Symposium, School of Science, University of Southern Queensland, Australia* **6 December 2021**.

3. **Jayasinghe W. J. M. L. P.,** Deo R.C., Ghahramani A., Ghimire S., Raj N. Development of novel hybridized three-phase deep Soil Moisture forecasting model. *Higher Degree Research Symposium, School of Science, University of Southern Queensland, Australia* **17 October 2022**.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ANN	Artificial Neural Network
AIRS	Atmospheric Infrared Sounder
APB	Absolute Percentage Bias
Boruta	Boruta-Random Forest
BOM	Australian Bureau of Meteorology
CNN	Convolutional Neural Network
DWT	Discrete Wavelet Transformation
DNN	Deep Neural Network
DL	Deep Learning
DT	Decision Tree
ECDF	Empirical Cumulative Distribution Function
ELM	Extreme Learning Machine
Ep	Pan Evaporation
EEMD	Ensemble Empirical Mode Decomposition
EMD	Empirical Mode Decomposition
ET	Reference Evapotranspiration
FE	Forecasting Error
GIOVANNI	Geospatial Online Interactive Visualization & Analysis Infrastructure
GLDAS	Global Land Data Assimilation System
IMF	Intrinsic Mode Function
KGE	Kling-Gupta Efficiency
Lasso	Least Absolute Shrinkage and Selection Operator
LM	Legate and McCabe Index
LSTM	Long- short term memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MEMD	Multivariate Empirical Mode Decomposition
MODIS	Moderate Resolution Imaging Spectroradiometer
moDWT	Maximum Overlap Discrete Wavelet Transform
MSE	Mean Squared Error
NS	Nash–Sutcliffe Index

NCA	Neighbourhood Component Analysis
QLD	Queensland
R	Pearson’s Correlation Coefficient
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
RRMSE	Relative Root Mean Square Error
SD	Standard Deviation
SILO	Scientific Information for Landowners
SM	Soil Moisture
SPI	Standardized Precipitation Index
SPEI	Standardized Precipitation and Evapotranspiration Index
SST	Sea Surface Temperature
TRMM	Tropical Rainfall Measuring Mission
WI	Willmott’s Index

MODEL NOTATIONS

Chapter 4 (Published Article 1)

NCA-LSTM	Two-phase hybrid model integrating the NCA feature selection algorithm with LSTM.

Chapter 5 (Published Article 2)

MEMD-Boruta-LSTM	Deep learning hybrid model integrating the MEMD and Boruta with LSTM
MEMD-Boruta-DNN	Deep learning hybrid model integrating the MEMD and Boruta with DNN
MEMD-Boruta-DT	Deep learning hybrid model integrating the MEMD and Boruta with DT

Chapter 6 (Under review Article 3)

moDWT-Lasso-LSTM	Multi-step three-phase hybrid model integrating the moDWT and Lasso feature selection algorithms with LSTM.
moDWT-Lasso-DNN	Multi-step three-phase hybrid model integrating the moDWT and Lasso feature selection algorithm with DNN.
moDWT-Lasso-ANN	Multi-step three-phase hybrid model integrating the moDWT and Lasso feature selection algorithms with ANN.
Lasso-LSTM	Multi-step two-phase hybrid model integrating the Lasso feature selection algorithm with LSTM.
Lasso-DNN	Multi-step two-phase hybrid model integrating the Lasso feature selection algorithm with DNN.
Lasso-ANN	Multi-step two-phase hybrid model integrating the Lasso feature selection algorithm with ANN.

CHAPTER 1: INTRODUCTION

1.1 Background

Freshwater resources are essential for the existence of human beings as it used for many purposes such as drinking, bathing, irrigating crops, hydropower generation, and other recreational activities. It is also a basic need that is required to ensure the existence of wildlife, flora, and fauna. Furthermore, it is closely related to developing recurrent drought conditions which make a huge impact on the environment creating adverse disasters like bushfires. Although 70% of our planet is covered by water, only 3% of it will be available as fresh water. Furthermore, two-thirds of freshwater exists in unavailable forms such as frozen glaciers or is unreachable to humans and other living beings in different ways. It is estimated that approximately 1.1 billion people globally have very limited access to usable water while a total of 2.7 billion people are confronted with water scarcity at least one month of the year (Fund, 2022). Further, competition for freshwater resources is likely to increase because of population expansion, urbanization, and climate change with a greater impact on high water-demanding sectors like agriculture. By 2050, the population is projected to reach over 10 billion, and this will require food and water for survival. It is predicted that agricultural production will need to increase by almost 70% to fulfill the needs of this increasing population (Bank, 2020). Therefore, with continuously increasing demand, in the future, freshwater is going to be very limited, scarce, and could become a rare resource in the future.

Among water-demanding activities of human beings like drinking, bathing, recreation, and hydropower generation, the agriculture sector can be recognized as one of the main sectors responsible for consuming higher volumes of fresh water and thereby possibly leading to unnecessary wastages. The usage of water for power generation can be minimized in the future with many alternatives, especially solar energy. However, fresh water is massively and essentially used as a major input, especially in irrigated agriculture which plays a vital role in ensuring food security in the world and no alternatives exist to replace this requirement. The land extent acquired by irrigated agriculture out of the total land area cultivated is approximately 20 percent, contributing to 40 percent of global total food production (Bank, 2020). Further, the productivity per unit land area under irrigated agriculture is considered at least two times higher on average than that of rainfed agriculture and therefore providing more opportunities for increasing and diversifying crop production.

On average, agriculture is responsible for 70 percent of worldwide freshwater withdrawals (Bank, 2020).

Under such background, management, conservation, and early identification of excess and short supplies of freshwater resources is very important to ensure uninterrupted water supply to essential operations and activities. Also, it will be very helpful in the management of natural disasters like drought and bushfires while conserving wildlife and the environment. Hence, a better understanding and precise forecasting of variations and future trends of hydrological parameters like rainfall, relative humidity, *SPEI*, *SPI*, evaporation, evapotranspiration, and soil moisture in advance will be very helpful. Therefore, this study focussed on the development of hybrid deep learning models to predict three important hydrological parameters: pan evaporation (*Ep*), reference evapotranspiration (*ET*), and soil moisture (*SM*) which is undoubtedly useful in water resources management and early identification of developing drought conditions and bushfire hazards.

Evaporation is the process through which a substance changes from a liquid or solid state to a vapor (Brutsaert, 2013). The evaporative process depletes the earth's surface water resources, and the pan evaporation (*Ep*) method is the most popular technique used to quantify this evaporative water loss. Water loss from on-farm storage and earth surface by evaporation is crucial as low soil moisture impacts to crop and pasture development, particularly in drought-prone areas. For instance, it is estimated that evaporation can account for up to 40% of storage volume loss annually in northern New South Wales and Queensland in Australia. In the long run, the evaporative process can be significantly accountable for the depletion of water storage used for drinking, bathing, irrigating crops, hydropower generation, and other recreational activities. Also, the evaporation process can accelerate the drying of natural water bodies and consequently deprive the drinking water for wildlife while excessive evaporation conditions particularly in dry spells develop drought conditions and natural disasters like bushfires. Therefore, predicting evaporation is a crucial factor to be considered in the current situation in the world.

Evapotranspiration is the combination of two distinct processes whereby water is lost from the crop via transpiration and from the soil surface by evaporation respectively (Sobrino et al., 2005). Around the world, evapotranspiration is a topic of intense research because this process significantly depletes the soil moisture causing water stresses particularly to the crops and developing drought and bushfire conditions. Therefore, predicting evapotranspiration in advance is highly useful in climatic characterization, water management, designing and operating irrigation projects, and figuring out crop water requirements. Furthermore, it is credited with helping to get early knowledge of natural disasters like bushfires, drought, and the importance of water as a crucial component for the sustainability of life (Abdullah et al., 2015).

Soil moisture (*SM*) refers to the water that is present in the soil and is crucial for sustaining plant growth as part of the soil-plant-atmosphere water cycle (Liao et al., 2018). Monitoring *SM* gives the knowledge for developing management methods that will best protect natural ecosystems from the threat of climate change while also minimizing the harm caused by precipitation deficiencies. In addition to this, *SM* information can greatly help geoscientists and the appropriate authorities to manage finite water resources needed for agriculture and other human activities, and manage the possible problems associated with decreased *SM* levels (Zhang et al., 2017). For instance, it can aid with drought monitoring, bushfire, and flood forecasting activities and enable more precise water, energy, and carbon budgeting (tern, 2022). So, early evaluation of *SM* reserves and monitoring of changes in available *SM* could help in developing risk reduction strategies and ensures the successful execution of government initiatives (McNairn et al., 2012).

Precise evaporation, evapotranspiration, and soil moisture forecasting models under climate change, especially in agricultural regions, can help stakeholders to make better decisions about water planning and resource management. Also, the projected above information is crucially important for early warning system design as well as controlling hydrological and agricultural drought situations. Additionally, the ability to predict evaporation, evapotranspiration, and soil moisture at the micro-scale and having advanced or projected knowledge of these variables would help farmers and farm managers to make proactive, sustainable decisions for effective irrigation, grazing, and water quality monitoring at ground level. Also, this knowledge could be used to create a knowledge-based system for tracking water resources and enhancing precision agriculture while it can have a significant impact on predicting bushfire hazards in advance and helping for reducing fire risk and prevalence (Marcar et al., 2006).

Predictive models based on machine learning can now be used in many different contexts because of the recent improvements in computing power. Deep learning, which is an advancement of machine learning algorithms incorporated with climatic and hydrological variable forecasting will provide a better understanding of the risks and repercussions of climate change and provide important information for mitigating such risks. Free and easy access and availability of big data sets also accelerated the popularizing and utilization of deep learning technologies in forecasting model development. Since predicting is crucial to the sustainability of climatology, hydrology, and agriculture, it is an active topic of research. Most technological advancements have relied on a systematic layered improvement approach, which is also how novel models for hydrological and agricultural applications are being developed. Therefore, this research is exploring new and sophisticated deep learning predictive models hybridized with feature optimization and multi-resolution analysis methodologies to predict pan evaporation, reference evapotranspiration, and soil moisture across Queensland, Australia.

1.2 Statement of the problem

Water shortages are a growing reality, particularly in arid and semi-arid areas. It is intensified further since rainfall is becoming less frequent and less predictable due to climate change’s alteration of long-established weather patterns. Furthermore, it causes dry seasons to turn into droughts which is a socioeconomic risk that poses serious challenges to groundwater reservoirs, resulting in water scarcity, failed crops, damaged habitats, unprecedented climate crises like bushfires, wildlife threats, and lost social or recreational opportunities (Mpelasoka et al., 2008, Riebsame et al., 2019). For instance, in 2019, a bushfire catastrophe burned around 20% of forests and claimed the lives of nearly a billion wild animals in Australia (Society, 2022). Under drought conditions, water losses are increased due to natural phenomena like evaporation, evapotranspiration, and thereby intensifying the magnitude of water scarcity-which causes significant impacts on high water-demanding agricultural activities and consequently reduces crop production and resulting extinction of natural wildlife habitats. The best approach to manage the challenges rising due to water scarcity is to use weather and climate data to make significant decisions while taking anticipated climate change into account (Government, 2019). Weather and climatic data can be strategically utilized for ensuring the smooth running of agricultural operations, water resource management, and strategic planning under water-scarce conditions.

Queensland is the second largest land area of states in Australia with more than 84 % of the land being utilized for water-demanding agricultural activities. But a large portion of Queensland experiences drought, land degradation, decreased profitability, increased debt, and human hardship due to less rainfall, especially during the ENSO-EI Nino period (Government, 2019). In 2020, the Queensland government declared that drought is expected to hit 67.4% of Queensland's geographical areas (Queensland, 2020). Also, the unprecedented bushfire crisis is impacted many parts of Queensland, particularly in summer. For example, in Queensland, a bushfire from November to December 2018 damaged a large number of homes, several buildings and vehicles, wildlife, crops, and pastures and burned 1.4 million hectares of land (Agency, 2022). Because of that, it is crucial to take action to mitigate the existing situation using a reliable method. The Queensland government proposed that regional climatic variations and climate predictions are vital solutions to planning and managing agricultural land (Government, 2019). In this scenario, evaporation, evapotranspiration, streamflow, radiation, soil moisture, and drought-affected factors are essential considerations when managing and taking strategic planning for existing problems.

Since these uncertainties directly affect income and food security, the government and policymakers need stronger forecasting models making them to determine any possible future reductions and associated dangers to food security. Such forecasting systems will be promisingly helpful in implementing strategic plans to avoid reductions in water resources, crop yields, and dangers to food security. This demonstrates the critical importance of advanced artificial intelligence models, which can aid in decision-making in water resource-depleting conditions, farming systems, precision agriculture, climate change, and natural disasters by generating predictions more precisely.

Reliable artificial intelligence predictive models with higher accuracy could be an important avenue for predicting drought-connected factors like evaporation, evapotranspiration, and soil moisture. However, the most crucial and pressing concern with choosing the non-redundant and most significant input (predictor) data that is still a challenge in developing forecasting models. This is because the usage of irrelevant inputs might introduce unnecessary problems during the model's execution, that is increasing the model's complexity while decreasing the model's forecasting accuracy. To overcome this problem this study uses feature selection algorithms such as NCA, Boruta, and Lasso in all models developed that can identify the best input parameters using comparison with real features.

Climatic and hydrological variables exhibit complicated temporal behaviour with non-stationarity aspects such as trends, seasonal changes, periodicity, and leaps in time series, which may impair the accuracy of data-driven models (Adamowski and Chan, 2011). Deep learning is an effective and novel method in artificial intelligence and machine learning that is widely used in all science and industrial fields in the big data era (Emmert-Streib et al., 2020). The DL model can successfully be used for time series prediction and providing solutions for issues related to utilizing climatic and hydrological variables. Literature proves that, among DL methodologies, Long Short-Term Memory (LSTM) network is widely used in the prediction of hydrometeorological and other variables due to its remarkable performances. To further improve the prediction model performance capability and overcome issues existing in time series big data sets, this study uses advanced data decomposition techniques, MEMD, and moDWT in all its model development efforts.

1.3 Objectives

The main purpose of this doctoral research is to develop hybrid DL forecasting models for, E_P, ET and SM using three different approaches in Queensland based on satellite and ground-based datasets to produce high-quality journal articles. It precisely targets to accomplish the following goals:

Objective 1: Developing an Evaporation Prediction Model

This objective focus to develop and evaluate the deep learning NCA-LSTM model, a combined approach where the LSTM prediction model coupled with the NCA feature selection technique to forecast daily E_P using satellite, and ground-based data and comparing it with standalone LSTM, DNN, RF, ANN, and DT models in Queensland. No evidence is found in the literature to confirm that the LSTM model along with NCA proposed in this objective to forecast E_P has been employed in Queensland, Australia. Furthermore, the NCA algorithm has shown a lack of sensitivity to the increased number of irrelevant features and good performances with high-dimensional data sets (Wei Yang, 2012). Since data for this objective are mainly extracted from satellite (AIRS spectrometer) and ground (SILO-Queensland) data sources and they are high-dimensional (Liu, 2015); NCA is an ideal feature selection algorithm for this objective. Also, LSTM is selected under this objective because it performed well in time series forecasting models as it can continuously update from the rid system to the next forecast using its input, output, and forget gate information in respective memory blocks (Ghimire et al., 2019). Hence, the proposed NCA-LSTM model will be a precise DL predictive model to forecast daily E_P. **This work was published in the *Journal of Hydrology* (Scopus Quartile 1).**

Objective 2: Deep Multi-Stage Reference Evapotranspiration Forecasting Model

This objective aims to develop and evaluate a three-phase hybrid MEMD-Boruta-LSTM model to forecast ET using satellite data and to compare with hybrid MEMD-Boruta-DNN, MEMD-Boruta-DT, and a standalone LSTM, DNN, and DT model in Queensland. No evidence has been found in the literature to prove that the three-phase hybrid LSTM model with MEMD and Boruta has been employed to predict ET in Queensland, Australia. Therefore, the proposed model in this objective will fill an important knowledge gap. The other reason for training a three-phase hybrid model for this purpose is that it yields high performances with relatively low errors (Al-Musaylh et al., 2018). The MEMD and Boruta data pre-processing techniques are employed here for further model improvement because they are the most powerful and enhanced signal decomposition and feature selection techniques used for nonlinear or intermittent time-series analysis (Ren et al., 2014). **This work was published in the *Journal of IEEE Access* (Scopus Quartile 1).**

Objective 3: Development of a novel three-phase hybridized deep soil moisture forecasting model

This objective entails constructing a multi-step hybrid moDWT-Lasso-LSTM soil moisture (SM) forecasting model in the 0-10 cm depth for 1 day, 14 days, and 30 days in advance with satellite data from NASA-Giovanni and ground data from SILO data sources in Queensland, Australia. Due to the nonstationary and nonlinear features of the obtained data, the extracted data were pre-processed using the Maximum Overlap Discrete Wavelet Transform (moDWT) decomposition method and the Least Absolute Shrinkage and Selection Operator (Lasso) feature selection algorithm. Then, the suggested three-phase hybrid moDWT-Lasso-LSTM forecasting model was created using the deep learning Long Short-Term Memory (LSTM) algorithm. The performance of the suggested moDWT-Lasso-LSTM model was statistically compared with benchmarked alternative machine learning models to confirm its viability. **This paper is submitted to the *Journal of Stochastic***

***Environmental Research and Risk Assessment* and is under review process. (Scopus Quartile 1)**

1.4 Significance of the research

This study produced highly reliable and accurate hybrid DL models for forecasting E_P , ET, and SM mainly based on satellite and ground-based data; the findings will be very useful in drought event management, water resources management, and strategic planning to prepare for drought, and water scarcity, and to practice sustainable agriculture in Queensland. Pan evaporation provides a very close estimation of water loss as a height measurement due to evaporation from soil, vegetation, and water resources used for irrigation activities, drinking purposes, recreation activities, and hydropower generation. By multiplying E_P value with the surface area of water storages, the volume of water loss due to evaporation (which is one of the major causes of water loss from water storages) can be calculated. Early identification of evaporative loss is very useful in planning and implementing irrigation schedules. Furthermore, in the long run, it is helpful in crop and land use planning and genetic improvements of commercial crops. In addition, ET gives a very close estimation of water loss from vegetation by evaporation and transpiration. If ET can be predicted precisely, farmers can have a better understanding of the amount of water to be added to their crops through irrigation and avoid unnecessary water losses. Furthermore, SM gives a sound understanding to farmers about the water availability of the soil and helps them in making decisions for better crop plans. The SM can offer timely information for quick decision-making during the growing season, such as types of crops to be grown, prioritizing the crops to be irrigated and accurately determining the total area to be cultivated. Therefore, as early warning decision support systems, the precise predictions of E_P, ET, and SM assist farmers in developing their short-term irrigation and crop plans as well as policymakers and government authorities in implementing better long-term strategic plans for trade development, managing disaster conditions, and securing rural livelihood. Furthermore, long-term predictions are useful in future strategic planning such as genetic improvement of crops, infrastructure upkeep, monitoring and revaluation of the farm's capability and management plan, awareness of animal welfare concerns and community expectations, financial record keeping, and analysis. This will significantly aid in the development of drought preparedness strategies and will lessen the risks associated with drought and water resource management.

In addition to the above socio-economic benefits expected, this study will also fill an important research gap in science and technology as all models proposed here to predict E_P, ET, and SM in Queensland are hybrid DL networks. In comparison to competing machine learning and DL models, these new hybridized sophisticated model architectures have outperformed them in terms of offering more sensible answers to challenges encountered in the real world. This study mainly uses data extracted from satellite and ground sources, and evidence has not been found in the literature to confirm that the approaches proposed in this study to forecast E_P, ET and SM have been used for any past study for Queensland, Australia.

1.5 Thesis layout

The schematic representation of the overview of the thesis is shown in Figure 1. It clearly defines the graphical abstract for easier understanding and the need for an accurate and reliable predictive tool for evaporation, evapotranspiration, and soil moisture. In this thesis, there are seven chapters makeup as follows:

Chapter 1

The objectives of this study are presented in this chapter along with the background information, problem statement and significance for the research.

Chapter 2

This chapter briefly explains previous research works conducted to use machine learning and artificial intelligence models for predicting E_p, ET and SM. It also covers the research gaps in predicting E_p, ET and SM by using artificial intelligence models.

Chapter 3

Chapter 3 establishes the context for the next chapters by describing the study region, data, and general approach used in this investigation. While the specific study area, data, and methods are discussed in the corresponding chapters, this chapter offers general viewpoints.

Chapter 4

This chapter includes the journal paper that has been published in a top-tier journal in hydrology (https://doi.org/10.1016/j.jhydrol.2022.127534). To predict one of the main water loss parameters, pan evaporation in the drought-prone region of Queensland, Australia, this chapter covers the construction of a hybrid Long Short-Term Memory (LSTM) predictive model paired with Neighbourhood Component Analysis (NCA) for feature selection. It compares the developed hybrid model (NCA-LSTM) with competitive benchmark models. This chapter covers the first objective of this study.

Chapter 5

This chapter includes the published paper in the journal IEEE Access, (https://doi.org/10.1109/ACCESS.2021.3135362). This chapter focuses on a unique three-phase deep Long Short-Term Memory (LSTM) forecasting model with Boruta-Random Forest (Boruta) and Multivariate Empirical Mode Decomposition (MEMD) algorithms to forecast evapotranspiration in drought-prone regions. This chapter covers the second objective of this study.

Chapter 6

This chapter includes the article submitted to the Journal of Stochastic Environmental Research and Risk Assessment and is under review process. This chapter focuses on developing three phase hybrid deep (0-10 cm) depth *SM* forecasting model using the Maximum Overlap Discrete Wavelet Transform (moDWT) method, the Least Absolute Shrinkage and Selection Operator (Lasso), and Long Short-Term Memory (LSTM) network for 1, 14 and 30 days in advance.

Chapter 7

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

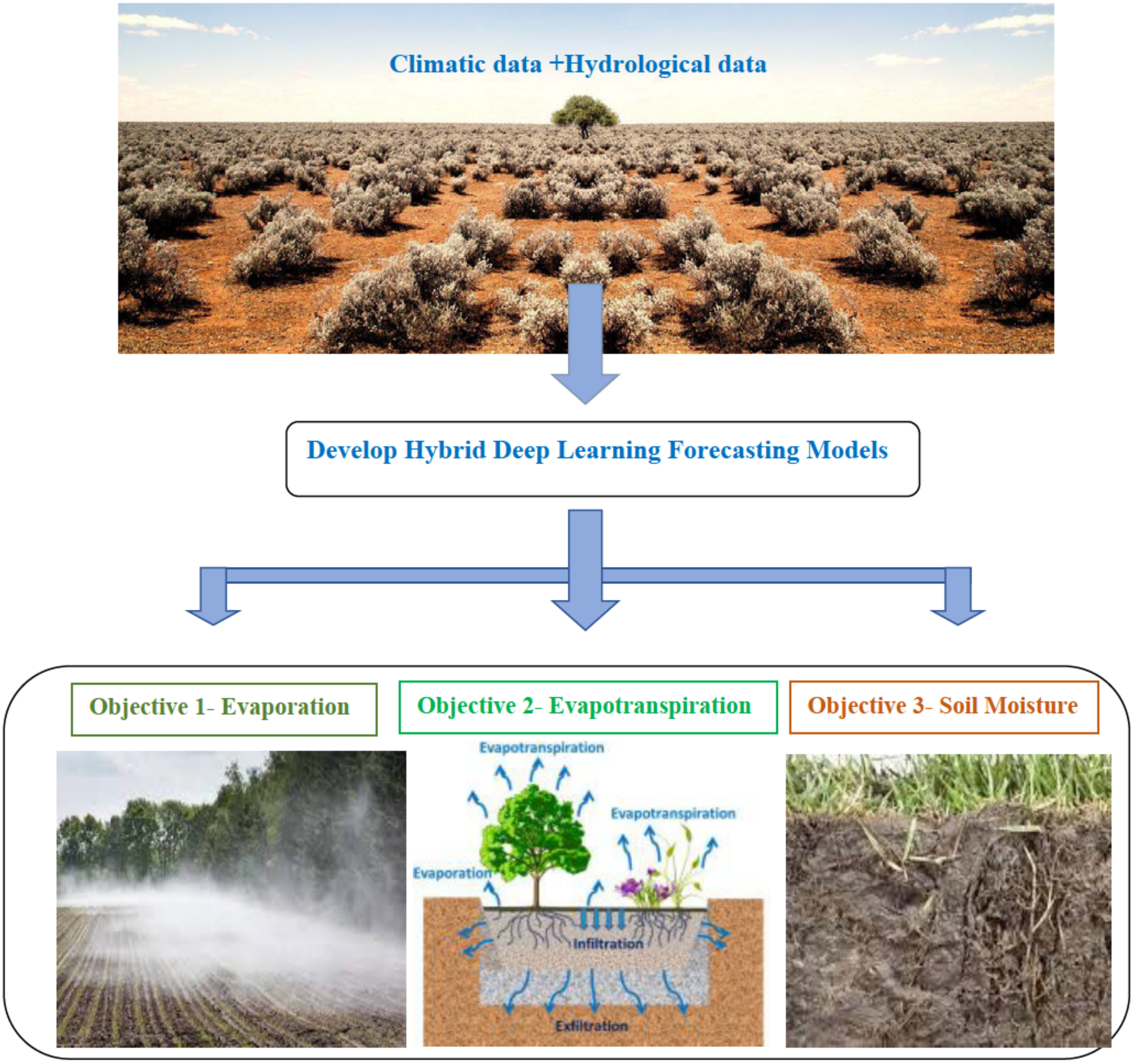


Figure 1: **Schematic view of the doctoral research thesis**

CHAPTER 2: LITERATURE REVIEW

This chapter briefly discusses the previous studies conducted to forecast E_p, ET and SM using machine learning and deep learning methodologies with data pre-processing techniques and the research gaps.

2.1 Previous studies in E_p prediction and research gaps

Many past research studies have experimented to employ data-driven machine learning techniques to predict E_p using various parameters. Goyal et al. (2014) developed Artificial Neural Network (ANN), Least Square Support Vector Regression (LSSVR), Fuzzy Logic (FL), and Adaptive Neuro-Fuzzy Inference Systems (ANFIS) models to predict daily E_p, and the results are evaluated against empirical methods proposed by Hargreaves and Samani (HGS) and the Stephens–Stewart (SS). The findings of this study have shown that FL and LSSVR techniques are superior to the traditional approaches in daily evaporation estimations. Deo et al. (2016) developed Relevance Vector Machine (RVM), Extreme Learning Machine (ELM), and Multivariate Adaptive Regression Spline (MARS) models to predict monthly evaporative losses using meteorological parameters as predictor variables for Amberley weather station, Australia. According to the results, the RVM model appeared to be more accurate in the prediction of evaporation loss. Kisi et al. (2016) developed decision tree-based machine learning methods such as Chi-square Automatic Interaction Detector (CHAID) and Classification and Regression Tree (CART) to predict daily E_p in Turkey and compared that with the neural network model. This study revealed that, neural networks performed better compared to the decision tree-based machine learning models. Wang et al. (2017) developed Fuzzy Genetic (FG), LSSVR, MARS, M5 model tree (M5Tree), and Multiple Linear Regression (MLR) for eight stations around Dongting Lake basin in China to estimate daily E_p and results showed that FG and LSSVR outperform over other machine learning models. Malik et al. (2017) developed Multi-Layer Perceptron Neural Network (MLPNN), Co-Active Neuro-Fuzzy Inference System (CANFIS), Radial Basis Neural Network (RBNN) and Self-Organizing Map Neural Network (SOMNN) models to predict monthly E_p in the Indian central Himalayas region, and it has revealed the superiority of CANFIS over other techniques. However, none of the above research has been tried hybridizing of machine learning models with advanced feature selection methods for E_p prediction.

Recently, many researchers tend to use deep learning AI techniques to develop predicting models because of its high learning capability from big data. Majhi et al. (2020) developed LSTM, Multilayer Artificial Neural Networks, and empirical methods like Hargreaves and Blaney–Criddle model for E_p prediction. In this study, the LSTM model was able to show its superior capability to predict daily evaporative losses against selected benchmark models. Abed et al. (2021) developed Extreme Gradient Boosting, Elastic Net Linear Regression, and LSTM models to predict monthly E_p and used two empirical techniques namely Stephens-Stewart and Thornthwaite for the performance assessment. The results showed that LSTM offered the most precise monthly E_p prediction from all the studied models for both stations in Malaysia. Abed et al. (2022) developed Convolutional Neural Network (CNN), Deep Neural Network (DNN), and Random Forest (RF) to estimate monthly E_p of Malaysian weather stations. The results showed that the CNN approach was an acceptable model than other comparison models. Kisi et al. (2022) developed LSTM model with grey wolf optimization (GWO), single LSTM, and advanced machine learning methods for E_p prediction using limited climatic variables as input. The outcomes showed that the LSTM-GWO model performed well than other models. Although above research used advanced deep learning models to forecast E_p, those deep learning models are very rarely hybridized with data pre-processing techniques like feature selection in E_p prediction studies which can further improve the model performances with big time series data.

Furthermore, we are unaware of any research employing NCA algorithm incorporated with deep learning to predict daily E_p or using the deep learning NCA-LSTM hybrid model for any other purposes. Therefore, the current study attempts to build a hybrid E_p forecasting model by employing LSTM network coupled with Neighbourhood Component Analysis (NCA) feature selection technique using satellite and ground-based data. This study selected LSTM as the forecasting algorithm since the literature demonstrates that among DL techniques, the Long Short-Term Memory (LSTM) network is frequently used in the prediction of hydro-meteorological and other variables because of its exceptional performances. NCA is selected as the feature selection technique in this study since its remarkable capabilities shown in previous works are likely to increase the overall forecasting skill of a predictive model. This study is a novel experience in the data science field as it is found to be the first time that LSTM is being hybridized with NCA and employed in daily E_p predictions using satellite and ground-based data.

2.2 Previous studies in ET prediction and research gaps

Researchers also have developed data-driven machine learning models to forecast ET and these models have shown superior performances despite the non-linear behaviour of ET (Wu et al., 2020). For instance, Fan et al. (2018) developed tree-based RF, M5Tree, gradient boosting decision tree (GBDT), and extreme gradient boosting (XGBoost) models to predict daily ET. According to the results, the XGBoost and GBDT models have been recommended for daily ET estimation in different climatic zones of China. Tikhamarine et al. (2019) developed ANN-embedded grey wolf optimizer (ANN-GWO), multi-verse optimizer (ANN-MVO), particle swarm optimizer (ANN-PSO), whale optimization algorithm (ANN-WOA) and ant lion optimizer (ANN-ALO) hybrid models to forecast monthly ET in India and Algeria. The results showed that ANN-GWO model provided better performance at both study stations. Nourani et al. (2020) employed ensemble MLR, SVR, ANFIS, ANN, and MLR models for ET forecasting and the results showed that ensemble MLR model performed well compared to all other models. However, none of the above research has been tried hybridizing of machine learning models with advanced data decomposition technique to predict ET.

Saggi and Jain (2019) developed Deep Learning-Multilayer Perceptrons (DL-MLP), Generalized Linear Model (GLM), RF, and Gradient-Boosting Machine (GBM) models to predict ET in the Indian districts of Hoshiarpur and Patiala, The results showed that DL-MLP model outperformed the others comparative models. Yin et al. (2020) developed a hybrid bi-directional LSTM model to forecast daily ET in three meteorological stations in central Ningxia, China. The performance of the hybrid Bi-LSTM model was evaluated by the Penman-Monteith method and the results showed that the hybrid Bi-LSTM model provides the best forecast performance at the selected meteorological stations. Ferreira and da Cunha (2020) developed a DL multi-step ET forecasting model with hybrid CNN-LSTM for 53 weather stations located in Minas Gerais, Brazil and assessed in comparison with standalone LSTM, CNN and traditional machine learning models (ANN and RF). According to the performance analysis, the hybrid CNN-LSTM model outperformed all the comparison models. Salam and Islam (2020) developed Random Tree (RT), Bagging and Random Subspace (RS), RF, and SVM models to predict daily ET in Bangladesh. Considering high prediction accuracy, RT and RF models have been suggested for daily ET prediction of Bangladesh. Above literature does not provide evidence for use of two-phase multistep deep hybrid models coupled with data pre-processing for ET prediction.

Moreover, multi-stage deep neural network-based ET forecasting has not yet been investigated. This project is focused on creating a novel multi-stage hybridized MEMD-Boruta-LSTM deep neural network to anticipate daily ET based on satellite and ground data to fill this knowledge gap.

2.3 Previous studies in *SM* prediction and research gaps

Data-driven predictive models have shown comparatively higher competency in soil moisture prediction (Prasad et al., 2019) and many researchers have conducted experiments to forecast soil moisture using data-driven models. For instance, Jamei et al. (2022) developed Extreme Gradient Boosting (XGBoost) and Categorical Boosting (CatBoost), two modern ensemble-based ML models, integrated with the Empirical Wavelet Transform (EWT) to predict daily root zone soil moisture (RZSM) in Ardabil and Minab regions (highly cold semi-arid and highly warm semi-humid regions and their performances were compared with rival models. The results have demonstrated the superior performance of the EWT-CatBoost and EWT-XGBoost models over the other counterpart models in forecasting multi-step ahead RZSM at Ardabil and Minab sites, respectively. Jamei et al. (2023) developed bidirectional gated recurrent unit (Bi-GRU), cascaded forward neural network (CFNN), adaptive boosting (AdaBoost), genetic programming (GP), and classical multilayer perceptron neural network (MLP) models using, Boruta gradient boosting decision tree (Boruta-GBDT) feature selection and multivariate variational mode decomposition (MVMD) techniques to predict daily Surface Soil Moisture (SSM) models in Iran's dry and semi-arid regions. According to the results, MVMD-Boruta-GBDT-CFNN outperformed all other hybrid models in one and seven days ahead soil moisture forecasting in all tested sites. Basak et al. (2023) two data-driven models based on Naive Accumulative Representation (NAR) and Additive Exponential Accumulative Representation (AEAR) developed at post-wildfire site in southern California. According to the results, AEAR model provided more accurate forecasts than existing models for time horizons of 10–24 hours. Above studies have not tried multi-step *SM* forecasting using machine learning model incorporated with feature selection and data decomposition algorithms.

Cai et al. (2019) developed Deep Learning Regression Network (DNNR), Linear Regression (LR), SVM, and ANN models to predict *SM* in Beijing. The results showed that the DNNR model performed well related to other models. ElSaadani et al. (2021) developed ConvLSTM, CNN, and LSTM models in south Louisiana in the United States to predict *SM* and results showed that the ConvLSTM model performed well than other comparative models. Suebsombut et al. (2021) developed LSTM-based models to forecast *SM* values in Chiang Mai province, Thailand and its results show that the LSTM-based model performs well in predicting soil moisture. Li et al. (2022) developed residual learning encoder-decoder (EDT-LSTM), LSTM, and encoder-decoder LSTM models to predict *SM* of 13 sites spread across different countries. The target EDT-LSTM model offered a new tool to predict *SM* better than other models. Zeynoddin and Bonakdari (2022) developed genetic and teacher–learner-based algorithms (GA and TLA) coupled with LSTM for *SM* forecasting in Quebec, Canada and results showed that TLA-LSTM found to be more computationally effective than GA-LSTM model. Although many research are conducted based on LSTM network as mentioned above, three phase hybrid LSTM models have not been developed in any of those studies.

This is a fresh experience as literature is not providing any evidence for using lasso feature selection and moDWT data decomposition techniques in *SM* prediction works. Additionally, this study has implemented solutions to address boundary condition-related problems that, in real-world scenarios, add errors to forecasts and which have not been adequately addressed in many recent hydrological research works that used wavelet transform techniques for data decomposition. That is also an initiative step in prediction studies that uses wavelet transform data decomposition procedures. Furthermore, this proposed algorithmic combination, referred to as the moDWT-Lasso-LSTM model, filling a research gap in soil moisture prediction as it has not yet been assessed in any other geographic location in the world.

CHAPTER 3: DATA AND METHODOLOGY

This chapter gives a summary of the study locations, description of data, and brief account of the methodology used to develop the hybrid deep learning predictive models. Different study locations were accomplished within the study region for each of the objectives which are explained in depth in each of the chapters. When the general methodology is provided in this chapter, distinct model development methodologies are discussed in respective chapters. The study area is described next followed by data description and general methodology employed in this work for the development of hybrid deep predictive models.

3.1 Study area

This study is undertaken in Queensland, Australia, where around 84% of the state's land resources are utilized for agriculture (DOAWE, 2020). Diverse sites were selected in the arid and semi-arid regions in the study site, Queensland, Australia. Figure 2 shows the map of selected sites in this study. Land resources of these selected sites are mainly used for farming operations to produce a wide range of agricultural products.

3.2 Data description

Since this study is based on a prediction of E_p, *ET*, and *SM*, a range of climatic and hydrological data are used to develop predictive models. Particularly, satellite data mainly from NASA's Goddard Online Interactive Visualization and Analysis Infrastructure (GIOVANNI) database while ground data from Scientific Information for Land Owners (SILO-Queensland) database were used to extract daily data in this study. Table 1 describes the data that was used to execute each objective, along with the sources they obtained and other pertinent information.

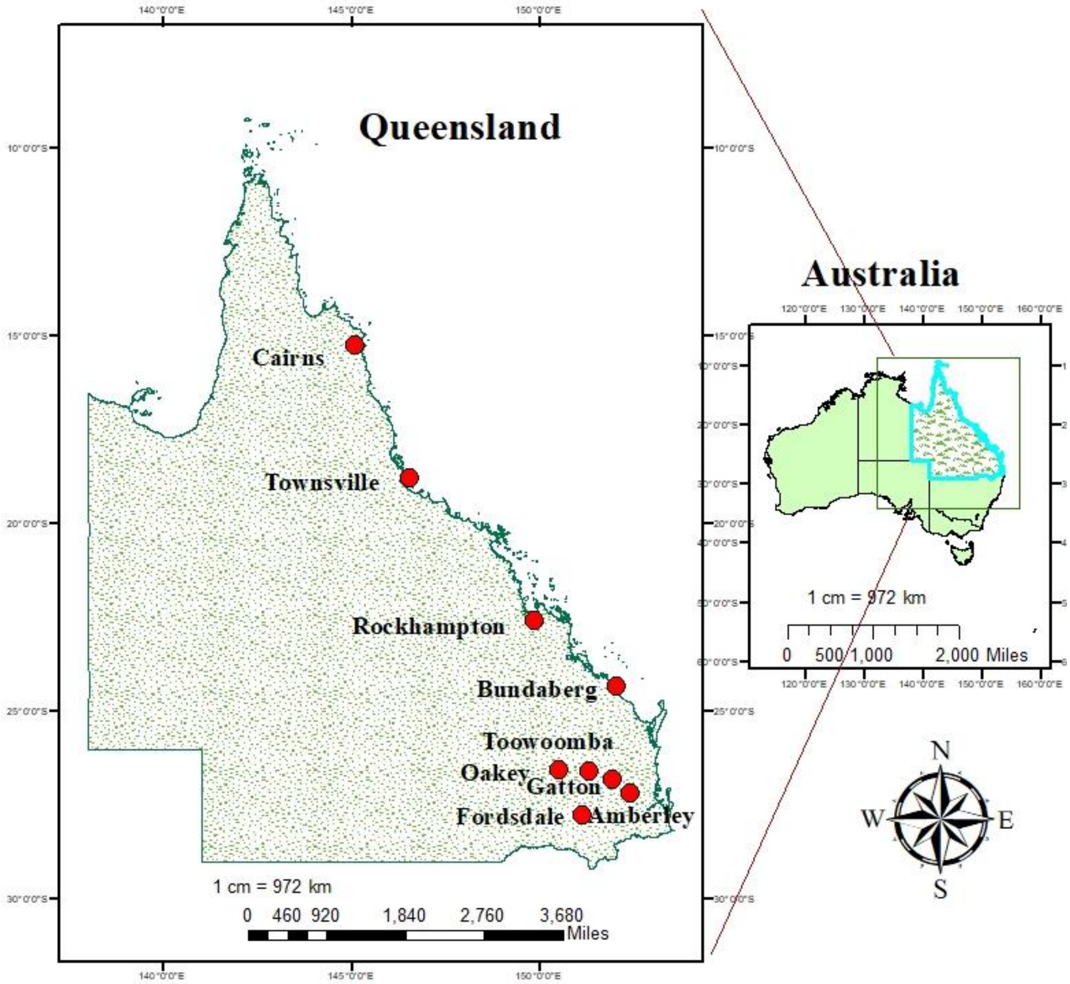


Figure 2: Study sites used to develop forecasting models in Queensland, Australia

3.2.1 Satellite data

NASA's Goddard Online Interactive Visualization and Analysis Infrastructure (GIOVANNI) is a remote sensing database and can be sourced from several platforms/instruments with various spatial and temporal resolutions, observations, disciplines, and measurements (NASA, 2022). Giovanni offers a clear and user-friendly approach to access, view, and analyze an enormous amount of earth science remote sensing data. In this study, Atmospheric Infrared Sounder (AIRS) system, Global Land Data Assimilation System (GLDAS) model, and Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS) platforms were used to extract predictor variables for the E_p, ET, and SM forecasting models development.

3.2.2 *Ground-based data*

The SILO data source provides Australian climate data from 1889 to the present which is operationally managed by Queensland Government (SILO-Queensland, 2022). SILO offers daily meteorological data for a variety of climate variables in gridded and ground-based data formats. In this study, ground-based data for predictor variables and target variables (E_p and *ET*) were extracted from the SILO database for further improvement of models' performances.

Table 1: Specifics about all the data used in this study

		Data	Source	Study period	Forecasted Horizon
Objective 1	Paper1	Predictors: **Meteorological satellite and ground variables** Target: **Pan Evaporation**	Atmospheric Infrared Sounder (AIRS) spectrometer +SILO	31 August. 2002 to 22 September 2020	Daily
Objective 2	Paper2	Predictors: **Meteorological satellite and ground variables** Target: **Evapotranspiration**	Atmospheric Infrared Sounder (AIRS) and Global Land Data Assimilation System (GLDAS) model+SILO	01 February 2003 to 19 April 2011	Daily
Objective3	Paper3	Predictors: **Meteorological satellite and ground variables** Target: **Soil Moisture**	GLDAS and Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS)+SILO	01 January 2005 to 31 December 2020	1 day, 14 days, and 30 days

3.3 General methodology

The proposed novel E_p, *ET*, and *SM* models were developed using an Intel Core i7 @ 3.3 GHz and 16 GB memory computer; configured with freely available DL libraries: Keras (Ketkar, 2017) and TensorFlow (Abadi et al., 2016) in Python (Sanner, 1999). The data pre-processing methods like data decomposition and feature selection were implemented using MATLAB R2019b and R software packages, while ‘‘matplotlib’’ and ‘‘seaborn’’ tools in Python were used for visualizations. All target models were superior based on deep LSTM networks that can capture higher-order nonlinear features in predictor datasets (Majhi et al., 2020).

Before developing deep learning and machine learning models, data pre-processing was carried out to work efficiently with nonlinear and nonstationary time series input data. Data pre-processing is widely used in artificial intelligence model hybridizing and various research studies have shown that it helps to enhance the model’s performance. In this study, feature selection and data decomposition techniques were employed as data pre-processing tools. The data pre-processing techniques used in this study include Neighbourhood Component Analysis (NCA), Boruta-Random Forest (Boruta), and Least Absolute Shrinkage and Selection Operator (Lasso) feature selection methods and Multivariate Empirical Mode Decomposition (MEMD), and Maximum Overlap Discrete Wavelet Transform (moDWT) data decomposition methods. Additionally, suitable scaling or normalization is necessary to prevent the dominance of inputs with wide numeric ranges, which could counteract the impacts of values with a smaller range. In this study, data are scaled to common values using normalization. The results are unaffected by the normalization because the data are normalized between [0,1], which is an invertible range (Hsu et al., 2003). The normalization is done by using Eq. (1) (García et al., 2016);

$$\boldsymbol{X_n = \frac{X_{actual} - X_{min}}{X_{max} - X_{min}}} \tag{1}$$

,where X_n, X_{actual}, X_{max}, and X_{min} represent the normalized, actual, maximum, and minimum values of predictor variable data, respectively.

After processing data, target hybrid predicting models were developed on deep LSTM neural network. In this research, several forecasting models are taken into consideration to assess the target models' performances in forecasting evaporation, evapotranspiration, and soil moisture since it is very important to evaluate and confirms the target models' viability in utilization over other existing models. Models used for evaluation purposes are standalone Long Short-Term Memory network (LSTM), Artificial Neural Network (ANN), Deep Neural Network (DNN), Decision Tree (DT), Random Forest (RF), two-phase Neighbourhood Component Analysis (NCA) based LSTM (NCA-LSTM), Boruta-Random Forest (Boruta) based LSTM (Boruta-LSTM), Boruta based DNN (Boruta-DNN), Boruta based DT (Boruta-DT), Lasso based LSTM (Lasso-LSTM), Lasso based DNN (Lasso-DNN) and Lasso based ANN (Lasso-ANN) and three phase Multivariate Empirical Mode Decomposition (MEMD) and Boruta-Random Forest (Boruta) based LSTM (MEMD-Boruta-LSTM), MEMD and Boruta based DNN (MEMD-Boruta-DNN), MEMD and Boruta based DT (MEMD-Boruta-DT), Maximum Overlap Discrete Wavelet Transform (moDWT) and Least Absolute Shrinkage and Selection Operator (Lasso) based LSTM (moDWT-Lasso-LSTM), moDWT and Lasso based DNN (moDWT-Lasso-DNN), moDWT and Lasso based ANN (moDWT-Lasso-ANN). Figure 3 illustrates a brief overview of artificial intelligence (AI) based on all hybrids deep learning and machine learning models and data preprocessing techniques used in this doctoral research thesis.

The developed models were evaluated by using a wide variety of statistical metrics such as Pearson's correlation coefficient (*r*), Determination of Coefficient (R^2), Mean Squared Error (*MSE*), Root Mean Square Error (*RMSE*), Mean Absolute Error (*MAE*), Willmott's Index (*WI*), Nash–Sutcliffe Efficiency (*NS*), and the Legates-McCabe's index (*LM*). Diagnostic plots, such as box plots, scatter diagrams, Taylor plots, stem plots, and time series plots are also used for a thorough review in addition to the use of numerical assessment measures.

All models' development, performances of the target, and comparative models using metrics and diagnostic plots were discussed accordingly in the respective chapters.

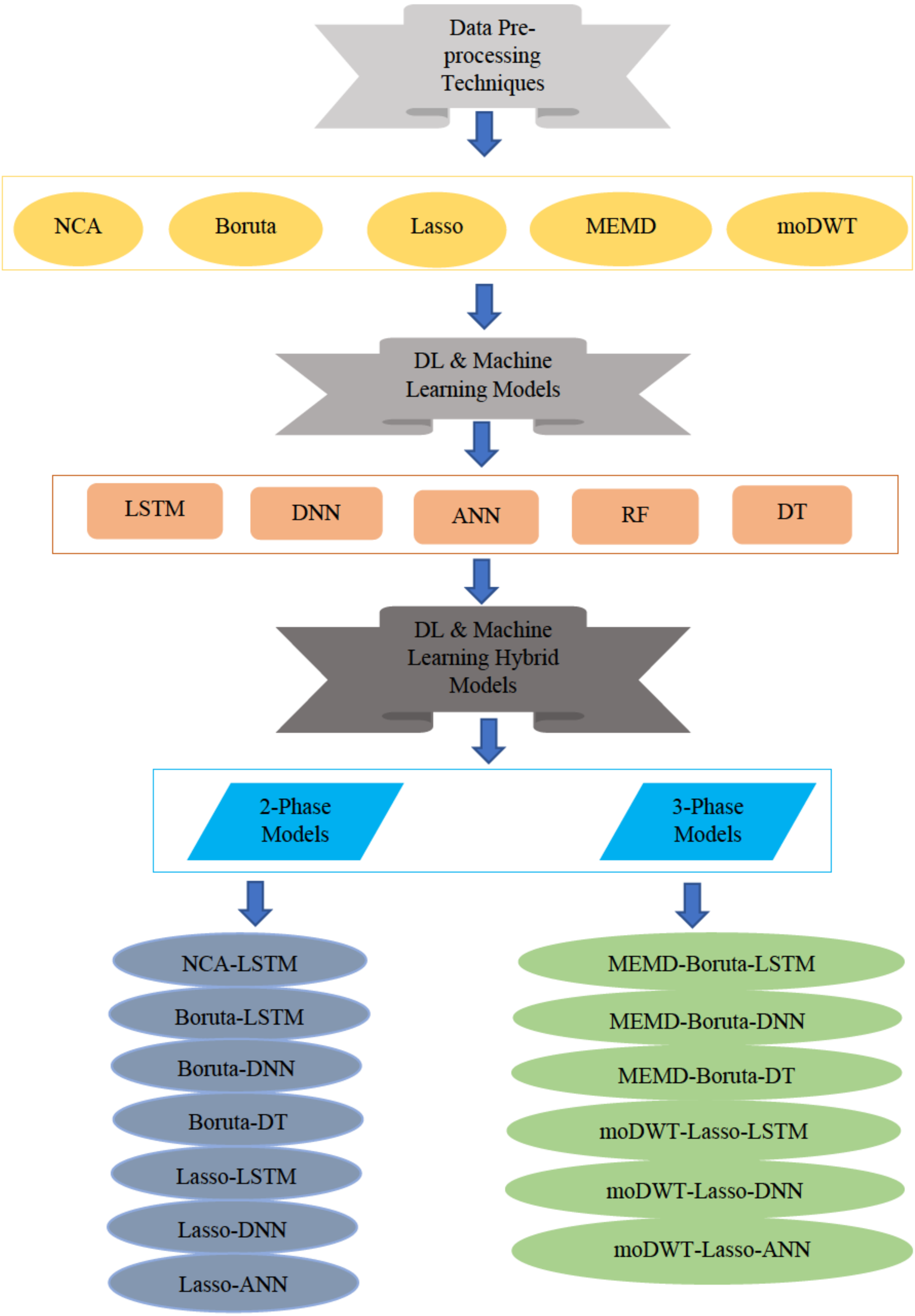


Figure 3: **Overview of all hybrids deep learning, machine learning models, and data preprocessing techniques used in this study**

CHAPTER 4: PAPER 1 - DEVELOPMENT AND EVALUATION OF HYBRID DEEP LEARNING LONG SHORT-TERM MEMORY NETWORK MODEL FOR PAN EVAPORATION ESTIMATION TRAINED WITH SATELLITE AND GROUND-BASED DATA

4.1 Introduction

This chapter is an identical replication of the article that was published in the *Journal of Hydrology*, Volume 607, April 2022.

This work aims to construct a precise deep hybrid artificial intelligence model to predict pan evaporation (*Ep*). To develop a target predictive model, satellite, and ground-based daily-scale big data in drought-prone regions in Queensland, Australia was utilized to train, validate, and test the model which was constructed on deep Long Short-Term Memory (LSTM) network. Model accuracy was increased by selecting significant predictor variables to target variable *Ep* with Neighbourhood Component Analysis (NCA) feature selection technique before training the model. The proposed target LSTM model coupled with NCA denoted as NCA-LSTM model performances were evaluated against competitive benchmark models, i.e., standalone LSTM, other types of DL models, single hidden layer neuronal architecture and decision tree-based method using statistical metrics and analytical plots in the testing phase. Concerning the predictive efficiency, the proposed NCA-LSTM hybrid model, improved with feature selection, outperforms all benchmark models, indicating its future utility in the prediction of daily *Ep*.

4.3 Links and implications

Pan evaporation (*Ep*) measures the evaporative loss from the earth's surface and water storage. The evaporative process is one of the major natural phenomena that is responsible for depleting the usable water resources utilized for agricultural production, drinking water supply, recreation activities, and hydropower generation. This evaporative process also can develop drought conditions and bushfire threats in severe dry spells and can adversely affect the existence of wildlife and the environment. Prediction of *Ep* in advance gives many opportunities for making strategic plans to battle with consequences created by water scarcity conditions due to evaporative losses in short and long-run contexts. Therefore, developing deep learning predicting model for precise prediction of *Ep* is highly beneficial and the proposed *Ep* predicting model developed in this research work will be gap filling great initiative for future use. However, it is suggested that the efficiency of the proposed hybrid deep learning model can be enhanced by signal decomposition techniques (E.g., Ensemble Empirical Mode Decomposition (EEMD), Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN), etc). So future researchers can combine the proposed hybrid NCA-LSTM model in this study with appropriate signal decomposition technique and it will promisingly enhance the current proposed model performances and will be a useful predictive tool in the field of hydrology. And also, future researchers can further develop this proposed model training methodology for forecasting *Ep* in long-run scenarios (E.g., One month ahead of *Ep* forecasting) which will be more useful in long-run strategic planning.

However, *Ep* only gives an estimate of water loss due to the evaporative process and it does not account for the water loss due to the transpiration process of plants and trees which makes the vegetative cover on the earth's surface. So, considering only the *Ep* for quantifying water loss will give an underestimate and can be insufficient in many situations. Therefore, this study in its second objective focused to develop a deep learning model to forecast Evapotranspiration *(ET)* which is a hydrological parameter quantifying water loss due to both evaporative and transpiration processes. The next chapter will explain the research outcome of this second objective in detail.

CHAPTER 5: PAPER 2 - DEEP MULTI-STAGE REFERENCE EVAPOTRANSPIRATION FORECASTING MODEL: MULTIVARIATE EMPIRICAL MODE DECOMPOSITION INTEGRATED WITH THE BORUTA-RANDOM FOREST ALGORITHM

5.1 Introduction

This chapter is an identical replication of the article that was published in the Journal of *IEEE Access*, Volume 9, December 2021.

This work aims to design a novel multi-stage deep learning hybrid Long Short-Term Memory (LSTM) predictive model that is coupled with Multivariate Empirical Mode Decomposition (MEMD) and Boruta-Random Forest (Boruta) algorithms to forecast evapotranspiration (*ET*) in the drought-prone regions of Queensland, Australia. Daily satellite and ground-based big data was used to build the proposed multi-stage deep learning hybrid model, i.e., MEMD-Boruta-LSTM, and the performance of the model was compared against competing benchmark models including hybrid MEMD-Boruta-DNN, MEMD-Boruta-DT, and standalone LSTM, DNN, and DT models in testing phase using statistical metrics and diagnostic plots. The testing results showed that the target MEMD-Boruta-LSTM hybrid model attained the lowest relative error and the highest efficiency relative to benchmark models for all study sites. Thus, the proposed multi-stage deep hybrid MEMD-Boruta-LSTM model surpassed all other benchmark models in terms of predictive efficacy and proved its value in the forecasting of the daily *ET*.

IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received November 23, 2021, accepted December 10, 2021, date of publication December 14, 2021, date of current version December 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3135362

Deep Multi-Stage Reference Evapotranspiration Forecasting Model: Multivariate Empirical Mode Decomposition Integrated With the Boruta-Random Forest Algorithm

W. J. M. LAKMINI PRARTHANA JAYASINGHE[1], (Member, IEEE), RAVINESH C. DEO[2], (Senior Member, IEEE), AFSHIN GHAHRAMANI[3], SUJAN GHIMIRE[4], AND NAWIN RAJ[2]

[1]School of Sciences, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[2]School of Sciences, University of Southern Queensland, Springfield, QLD 4300, Australia
[3]Centre for Sustainable Agricultural Systems, Institute for Life Sciences and the Environment, University of Southern Queensland, Toowoomba, QLD 4350, Australia
[4]School of Sciences, University of Southern Queensland, Toowoomba, QLD 4350, Australia

Corresponding authors: Ravinesh C. Deo (ravinesh.deo@usq.edu.au) and W. J. M. Lakmini Prarthana Jayasinghe (lakmini.mudiyanselage@usq.edu.au)

This work was supported in part by the School of Science, University of Southern Queensland, Australia; and in part by the Wayamba University of Sri Lanka.

ABSTRACT Evapotranspiration, as a combination of evaporation and transpiration of water vapour, is a primary component of global hydrological cycles. It accounts for significant loss of soil moisture from the earth to the atmosphere. Reliable methods to monitor and forecast evapotranspiration are required for decision-making. Reference evapotranspiration, denoted as *ET*, is a major parameter that is useful in quantifying soil moisture in a cropping system. This article aims to design a multi-stage deep learning hybrid Long Short-Term Memory (LSTM) predictive model that is coupled with Multivariate Empirical Mode Decomposition (MEMD) and Boruta-Random Forest (Boruta) algorithms to forecast *ET* in the drought-prone regions (*i.e.*, Gatton, Fordsdale, Cairns) of Queensland, Australia. Daily data extracted from NASA's Goddard Online Interactive Visualization and Analysis Infrastructure (GIOVANNI) and Scientific Information for Land Owners (SILO) repositories over 2003–2011 are used to build the proposed multi-stage deep learning hybrid model, *i.e.*, MEMD-Boruta-LSTM, and the model's performance is compared against competitive benchmark models such as hybrid MEMD-Boruta-DNN, MEMD-Boruta-DT, and a standalone LSTM, DNN and DT model. The test MEMD-Boruta-LSTM hybrid model attained the lowest Relative Root Mean Square Error ($\leq$17%), Absolute Percentage Bias ($\leq$12.5%) and the highest Kling-Gupta Efficiency ($\geq$0.89) relative to benchmark models for all study sites. The proposed multi-stage deep hybrid MEMD-Boruta-LSTM model also outperformed all other benchmark models in terms of predictive efficacy, demonstrating its usefulness in the forecasting of the daily *ET* dataset. This MEMD-Boruta-LSTM hybrid model could therefore be employed in practical environments such as irrigation management systems to estimate evapotranspiration or to forecast *ET*.

INDEX TERMS Reference evapotranspiration forecasting, deep learning, multivariate empirical mode decomposition, boruta-random forest algorithm, long short-term memory network.

I. INTRODUCTION

Evapotranspiration estimation is involved in water resource management, hydrological studies, irrigation scheduling, crop modelling, and computing drought indices. Reference evapotranspiration (*ET*) and crop coefficient [1] are mostly used to estimate evapotranspiration related to a particular crop. *ET* can be directly measured by using the lysimeter method. Several empirical methods such as the Hargreaves equation, Priestley–Taylor equation, Ritchie equation, and

The associate editor coordinating the review of this manuscript and approving it for publication was Haiyong Zheng.

the PMF-56 equation have also been developed to estimate *ET* using climatic data. Among them, the PMF-56 equation is widely used due to its accuracy and stability [2]. Other than empirical methods, many researchers have developed data-driven Artificial Intelligence (AI) models to forecast *ET* and these models have shown superior performances despite non-linear behaviour of *ET* [3]. For instance, Nourani, *et al.* [4] employed ensemble Multiple Linear Regression (MLR), Support Vector Regression (SVR), Adaptive Neuro-Fuzzy Inference System, Artificial Neural Network (ANN), and Multiple Linear Regression (MLR) models for *ET* forecasting and the ensemble MLR model has shown the best performance. Tikhamarine, *et al.* [5] examined the comparative potential of ANN-Embedded Grey Wolf Optimizer, Multi-Verse Optimizer, Particle Swarm Optimizer, Whale Optimization Algorithm and Ant Lion Optimizer to predict monthly *ET* in India and Algeria.

Deep Learning (DL) techniques such as the Temporal Convolution Network [6] and the ensemble of Convolutional Neural Networks (CNN) [7] which are comparatively more advanced and precise than the above traditional machine learning methods have also been recently employed to predict *ET*. The Long Short-Term Memory (LSTM) network is a DL neural technique that has been used to predict hydrological variables like water quality [8], solar radiation [9], and streamflow [10], and rainfall-runoff [11]. Several recent studies have shown the exceptional performance of LSTM model in predicting hydrological time series [67]–[70]. The key advantage of the LSTM model is its ability in using sequential data as inputs instead of independent training samples and this feature ensures the model's capability in dealing with more extended historic hydrologic observations with temporal dependence [71], which is a common characteristic related with many types of hydrological time series [66]. However, less research has been carried out to predict *ET* using LSTM based models. Yin, *et al.* [12] proposed a new hybrid bi-directional LSTM model to forecast short term daily *ET* in data scarce regions.

In recent years, use of AI models have become more popular in resolving problems related to many various hydrological aspects [72]. For instance, DT model has been used to map the flood susceptible areas in Kelantan, Malaysia which performed with greater accuracy in comparison with frequency ratio (FR) and logistic regression (LR) methods [74]. The DNN model has been employed in water resource management e.g. development of spatial-temporally continuous evapotranspiration model [75], development of model for mapping suitable groundwater extraction location [76] and shown better performances compared to benchmark models. LSTM is also extensively used for flood forecasting [77], and predicting water table [8], etc.

To further enhance the forecasting model capabilities, hybrid models have been developed in the recent past by many researchers. Ferreira and da Cunha [13] developed a DL multi-step *ET* forecasting model with hybrid CNN-LSTM and assessed in comparison with standalone LSTM, CNN and traditional machine learning models (ANN and RF). According to the performance analysis, the hybrid CNN-LSTM model outperformed all the comparison models.

In addition, two-phase hybrid models which are capable of yielding high performances with relatively low errors are explored [14]. For example, Prasad, *et al.* [15] developed a two-phase hybrid Extreme Learning Machine (ELM) model to forecast soil moisture coupled with the Ensemble Empirical Mode Decomposition (EEMD) data pre-processing method and the Boruta-random forest optimizer (Boruta) feature selection method. This model was superior to the other comparative models and yielded a relatively accurate performance with a small number of errors.

The Multivariate Empirical Mode Decomposition (MEMD) is a data pre-processing method that is an improved extension of standard Empirical Mode Decomposition (EMD) for multichannel data [16] and works efficiently in time series nonlinear and nonstationary signal data pre-processing [17]. For instance, Prasad, *et al.* [15] and Ali, *et al.* [18] proposed new multi-stage models coupled with MEMD to forecast solar radiation and drought thereby showing superior performance when compared with other models.

Boruta-random forest (Boruta) is a feature selection technique [19] that can identify significant input parameters using a comparison with real features to those of random probes [20]. Boruta has been utilized successfully as a feature selection technique in hybrid models to forecast soil moisture [20], [21], streamflow [22], [23], and air quality [24].

However, *ET* forecasting based on multi-stage deep neural networks is yet to be explored. To address this research gap, this study is focused on developing a novel multi-stage MEMD-Boruta-LSTM deep neural network to forecast daily *ET* based on satellite and ground data. DT and DNN models which have been widely employed in prediction of various hydrological parameters are selected for model performance comparison with target model in this study.

II. THEORETICAL OVERVIEWS

In this section, the MEMD, Boruta, and LSTM are described in detail. The models used for comparison purposes in this study: Deep Neural Network (DNN) [9] and Decision Tree (DT) [25] are not explained in detail as they are well-known algorithms.

A. MULTIVARIATE EMPIRICAL MODE DECOMPOSITION METHOD

The MEMD is an advanced version of EMD proposed by Rehman and Mandic [16] which is capable of dealing with multivariate signals and resolved the mode mixing issue by using white Gaussian noises [26]. The MEMD method can be described as follows [16]:

I. Generate a suitable number of direction vectors.
II. Calculate projections of the multiple inputs along with different directions in an *n*-dimensional space.

III. Identify local maxima projections and obtain multivariate envelope curves through them and subsequently calculate the mean.
IV. Extract the detail using the difference of the mean envelope curve and original signal until the stopping criteria is satisfied for a multivariate Intrinsic Mode Function (IMF) [27].

The mathematical formulae of the MEMD can be found elsewhere [16], [28].

B. FEATURE SELECTION: BORUTA-RANDOM FOREST OPTIMIZER ALGORITHM

The algorithm can be briefly explained as follows [19], [29]:

Let $x_t \in R^n$ be the group of predictors for the set of T and $y_t \in R$ be the target for n number of inputs, where $t = 1, 2, \ldots, T$.

I. Create a randomly ordered duplicated (shadow) variable, x_t' for x_t and then predict the target y_t.
II. Calculate Mean Decrease Accuracy (MDA) for every x_t and x_t' over all trees, m_{tree} (=500 in this study) [1], [30]:

$$MDA = \frac{1}{m_{tree}} \sum_{m=1}^{m_{tree}} \times \frac{\sum_{t\epsilon OOB} I\left(y_t = f\left(x_t\right)\right) - \sum_{t\epsilon OOB} I\left(y_t = f\left(x_t^n\right)\right)}{|OOB|}, \quad (1)$$

where $I(*)$ is indicated function, OOB (Out-of-Bag) is a predictive error, $y_t = f(x_t)$ is predicted value before permuting and, $y_t = f\left(x_t^n\right)$ is predicted value after permuting.
III. Compute the Z-score as:

$$Z - score = \frac{MDA}{SD} \quad (2)$$

where, SD is the standard deviation of accuracy loss and, then maximum Z-score (Z_{max}) is determined among duplicated attributes.
IV. Following that, predictors are identified as "Unimportant" when $Z - score < Z_{max}$ and "Confirmed" as important when $Z - score > Z_{max}$ during the process.

C. TIME SEQUENTIAL PREDICTIVE METHOD: LONG SHORT-TERM MEMORY NETWORK

The LSTM is a special Recurrent Neural Network (RNN) [32] related to conventional artificial neural networks that are mainly used to identify patterns in sequences of data [33]. The LSTMs operates with special units, denoted as memory blocks that consist of input, output, and forget gates and these memory blocks continuously update and control the information flow [34]. The calculations are described in 4 steps as follows [35]:

I. The LSTM layer decides which information should be forgotten or remembered, based on the last hidden layer output h_{t-1} and the new input x_t by using "forget gate" f_t:

$$f_t = \sigma\left(w_f\left[h_{t-1}, x_t\right] + b_f\right) \quad (3)$$

where w_f is the weight matrix; b_f is the bias vector and $\sigma(\ldots)$ is the logistic sigmoid function.
II. The LSTM layer decides what information needs to be stored in the new cell state c_t that is represented by the new candidate cell state $\overline{C}_t$ after updating information by using "input gate" i_t:

$$\overline{C}_t = tanh\left(w_C\left[h_{t-1}, x_t\right] + b_C\right) \quad (4)$$
$$i_t = \sigma\left(w_i\left[h_{t-1}, x_t\right] + b_i\right), \quad (5)$$

where $tanh(\ldots)$ is the hyperbolic tangent function.
III. The old cell state C_{t-1} updates to C_t by the "forget gate" f_t to remove unnecessary information and the "input gate" i_t to get a new candidate cell state $\overline{C}_t$:

$$C_t = f_t * C_{t-1} + i_t * \overline{C}_t \quad (6)$$

IV. Finally, the output h_t is derived using "output gate" o_t and the cell state C_t:

$$o_t = \sigma\left(w_o\left[h_{t-1}, x_t\right] + b_o\right) \quad (7)$$
$$h_t = o_t * tanh\left(C_t\right) \quad (8)$$

III. MATERIALS AND METHOD

A. STUDY REGION AND DATASET

This study is centred in Queensland (QLD) Australia, where 84% of the total land resources are used for agricultural operations [36]. The Queensland government declared 67.4% of the land area of Queensland drought-affected in the year 2020 [37]. Therefore, developing a precise model to forecast water losses due to *ET* is useful for strategic planning in water resources management in the state.

The three examined sites located in arid and semi-arid areas in QLD, Australia selected for this study are Gatton −152.34°E, 27.54°S, Fordsdale −152.12°E, 27.72°S and Cairns −145.75°E, 16.87°S (see Figure 1). The land resources of these selected sites are mainly used for agricultural purposes.

To construct a target hybrid model, data for eight daily predictive climatic variables for the period 01 February 2003 to 19 April 2011 were extracted from the databases of NASA's Goddard Online Interactive Visualization and Analysis Infrastructure (GIOVANNI) - Atmospheric Infrared Sounder (AIRS) and GLDAS model satellite and Scientific Information for Land Owners (SILO). The GIOVANNI provides easy and user-friendly access to visualize and analyse the vast amount of Earth Science-related remote sensing data [38] that can be extracted easily without the requirement for advanced prior knowledge of complex remote sensing datasets. In addition, SILO data source provides ground-based data for predictor variables and it assists to further improve the model's performance. This database is operationally managed by the Queensland Government [39].

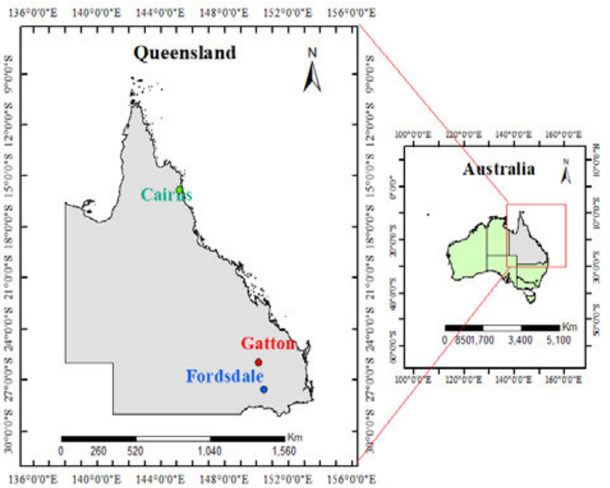


FIGURE 1. Study sites in Queensland, Australia where the proposed MEMD-Boruta-LSTM model was implemented.

TABLE 1. List of satellite-based Goddard Online Interactive Visualization and Analysis Infrastructure (GIOVANNI) and the ground-based Scientific Information for Land Owners (SILO) predictor variables used to forecast daily Reference Evapotranspiration (*ET*). Note: Atmospheric Infrared Sounder (AIRS) and GLDAS model are the two platforms in the GIOVANNI data source.

Data source		Name of input variable	Acronym	Unit
GIOVANNI-Satellite data	AIRS	Surface Temperature-Day	Tsd	°C
		Surface Temperature-Night	Tsn	°C
		Air Pressure	Pa	hPa
	GLDAS Model	Bare Soil Evaporation	Ebs	$kgm^{-2}s^{-1}$
		Transpiration	TR	$kgm^{-2}s^{-1}$
SILO-Ground based data		Maximum Temperature	Tmax	°C
		Minimum Temperature	Tmin	°C
		Radiation	Ra	MJm^{-2}

Missing data due to instrumental and equipment failures were filled with daily mean data of previous years [40]. Table 1 shows a summary of predictive variables and sources of data. For the target variable that is daily *ET*, point-based data is extracted from the SILO database.

B. DEVELOPMENT OF THE PROPOSED MULTI-STAGE DEEP HYBRID MEMD-BORUTA-LSTM MODEL

The proposed multi-stage MEMD-Boruta-LSTM model was developed using an Intel Core i7 @ 3.3 GHz and 16 GB memory computer; built using freely available DL libraries: Keras [46] and TensorFlow [47] in Python [48]. The MEMD data pre-processing method and Boruta feature selection method were implemented using MATLAB R2019b and R respectively, while "*matplotlib*" and "*seaborn*" tools in Python were used for visualizations. The MEMD-Boruta-LSTM hybrid model was developed using historical time series inputs as follows:

Stage 1: In this study, before performing MEMD, firstly, all nine variables (eight predictors + target) (see Table 1) were partitioned into 50% for training (*i.e.*, 1500 data points) and other 50% for testing (*i.e.*, 1500 data points) for all study sites [49] to avoid having a different number of Intrinsic Mode Functions (IMFs). Deo, *et al.* [50] pointed out that, if the complete dataset (training, cross-validation, and testing) is decomposed together without partitioning as explained above, future data (that is testing and yet unseen data by the forecasting model at a particular time step) would unintentionally add bias into the forecast. Thus, it is an important requirement during the decomposition stage to avoid incorporating future datasets that are to be used in the testing phase with the calibration dataset i.e., training and cross-validation in this study.

The MEMD was performed in the decomposition process independently for each training, and testing data partitions for both predictor and target variables for all three sites. In this process, the recommended predefined parameters: ensemble number ($N = 500$) and amplitude of the added white noise ($\varepsilon = 0.2$) were applied [51]–[54]. All the first IMFs of predictor and target variables were pooled into one set. All the second IMFs of predictor and target variables were pooled into one set. This pooling was carried out until the ith IMFs including residuals.

Stage 2: Boruta-random forest is a feature selection technique available in R. Random Forest tree-based algorithm is embedded in this feature selection technique [21]. This feature selection algorithm is used to identify the significantly corelated predictor variables to the target variable in each IMFs and residuals using historical lagged data at (t-1).

Stage 3: After identifying the significantly corelated predictor variables for the model development, respective data of those variables were normalized to remove the variance of features [55] by converting them into $(0 - 1)$ range using equation (9):

$$X_n = \frac{X_{actual} - X_{min}}{X_{max} - X_{min}} \tag{9}$$

where X_{actual}, X_{max}, and X_{min} represent input data for actual, maximum, and minimum values respectively.

Stage 4: The LSTM model was employed to forecast daily *ET* in each IMF and residual using significantly corelated predictor variable data at (t-1) lag. To prepare the best model design, hyperparameters for the target model (MEMD-Boruta-LSTM) were identified using the *Hyperopt* library in Python [56], [57]. *Hyperopt* is one of the hyperparameter optimization algorithms that performed better than the *Grid search* and *Random search* algorithms as it ensures comparatively less time in the model training process while increasing the accuracy of the model [58]. Thereby optimal architecture of the hybrid MEMD-Boruta-LSTM model was used to predict daily *ET*. Finally forecasted *ET* in each

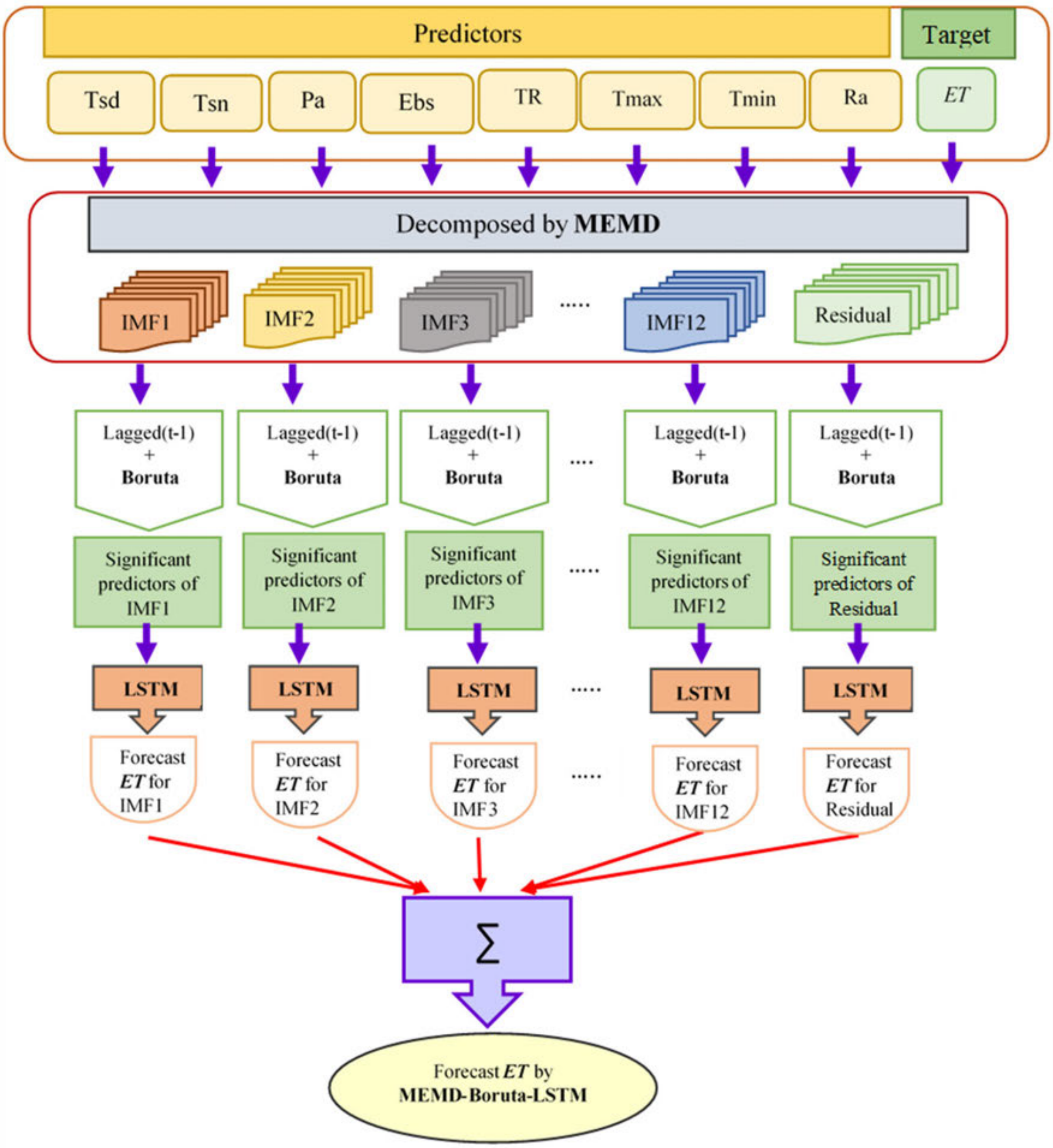


FIGURE 2. Workflow diagram detailing the necessary steps taken to design proposed deep hybrid MEMD-Boruta-LSTM model for daily evapotranspiration (*ET*) forecast. Note: *ET* = Evapotranspiration, MEMD = Multivariate Empirical Mode Decomposition, IMF = Intrinsic Mode Function, LSTM = Long Short-Term Memory. The details of predictors are given in TABLE 1.

IMF and a residual were cumulated to calculate forecasted daily *ET* for each study site. Figure 2 presents the workflow of the proposed multi-stage MEMD-Boruta-LSTM model. The same procedure is followed to develop hybridized DNN and DT with MEMD-Boruta (*i.e.*, MEMD-Boruta-DNN and MEMD-Boruta-DT models). Developed standalone LSTM, DNN, and DT models and hybrid MEMD-Boruta-DNN and MEMD-Boruta-DT were used as benchmark models for the model performance comparison.

C. MODEL PERFORMANCE EVALUATION

Model performances are evaluated using the statistical metrics [41]–[45] given below to confirm whether the target predictive model is superior to other benchmark models and is sufficiently qualified for *ET* prediction in QLD,

I. Correlation Coefficient (r): The correlation coefficient measures the strength of the relationship between two variables and the values range between -1.0 and 1.0 [62]. The value given for perfect forecasting models is equal to $+1$ indicating strong positive relationship of forecasted values derived from the model with actual values, (10) as shown at the bottom of the next page.

II. Root Mean Square Error $\left(RMSE; mmday^{-1}\right)$: This measures the average model-performance error between predicted value (E_p^{FOR}) and observed value

(E_p^{OBS}) [63]. The $RMSE$ value can range from 0 to ∞ and it becomes zero for the best predictive models.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(ET^{FOR,i} - ET^{OBS,i}\right)^2}, \quad 0 \le RMSE < \infty \quad (11)$$

III. Mean Absolute Error $\left(MAE; mmday^{-1}\right)$: This error value provides an assessment of the actual forecasting errors in terms of the total number of observations [21]. MAE can range from 0 to ∞ and it becomes zero for best predictive models. The MAE gives a more precise measure of average model error than the $RMSE$ since it is not influenced by extreme outliers [41].

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|ET^{FOR,i} - ET^{OBS,i}\right|, \quad 0 \le MAE < \infty \quad (12)$$

IV. Relative Root Mean Squared Error ($RRMSE$): The $RRMSE$ is used to measure overall forecasting accuracy of the models and always gives positive values [21]. If the value for $RRMSE$ is less than 10%, model performance is considered to be outstanding, while model performance is considered to be good if it is lying between 10% to 20%. If the value for $RRMSE$ error lies between 20% to 30%, model performance is considered as fair. If the value for $RRMSE$ error is higher than 30% model performance is considered to be poor [64].

$$RRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(ET^{FOR,i} - ET^{OBS,i}\right)^2}}{\frac{1}{N}\sum_{i=1}^{N}ET^{OBS,i}} \times 100 \quad (13)$$

V. Relative Mean Absolute Percentage Error ($RMAE$): The relative mean absolute percentage error measures the size of the error in percentage terms.

$$RMAE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{ET^{FOR,i} - ET^{OBS,i}}{ET^{OBS,i}}\right| \times 100 \quad (14)$$

VI. Nash-Sutcliffe Index (NS): The NS [43] measures how well the plotted line between observed data and simulated data fits into 1:1. The NS is equal to 1, if the model forecasted data is perfectly matched to the observed data. NSE = 0 indicates that the model predictions are as accurate as the mean of the observed data while, Inf < NSE < 0 indicates that observed mean is a better predictor than the model [62].

$$NS = 1 - \left[\frac{\sum_{i=1}^{N}\left(ET^{OBS,i} - ET^{FOR,i}\right)^2}{\sum_{i=1}^{N}\left(ET^{OBS,i} - \overline{ET^{OBS}}\right)^2}\right], \quad -\infty < E_{NS} \le 1 \quad (15)$$

VII. Willmott's Index (WI): Willmott index is a standardized measure of the degree of model prediction error and the value for WI ranges from 0 to 1, whereas this value equals 1 for best predictive models, (16) as shown at the bottom of the page.

VIII. Legate and McCabe Index (LM): The LM is an advanced assessment index based on WI and NS values. This index can be used to assess the goodness-of-fit of a hydrologic or hydro climatic model and is more effective than correlation and correlation-based measures (e.g., the Coefficient of Determination (r^2), WI and NS) [41]. The value for LM ranges from $-\infty$ to 1, whereas this value equals one for best predictive models.

$$LM = 1 - \left[\frac{\sum_{i=1}^{N}\left|ET^{FOR,i} - ET^{OBS,i}\right|}{\sum_{i=1}^{N}\left|\left(ET^{OBS,i}, -\overline{ET^{OBS,i}}\right)\right|}\right], \quad -\infty < LM \le 1 \quad (17)$$

IX. Absolute Percentage Bias ($APB\%$): The APB gives the error of forecasted values as a percentage concerning the observed values. The optimal value for APB is zero and lower-magnitude values closer to zero reflect good accuracy of the model [65].

$$APB = \left[\frac{\sum_{i=1}^{N}\left(ET^{OBS,i} - ET^{FOR,i}\right) \times 100}{\sum_{i=1}^{N}ET^{OBS,i}}\right] \quad (18)$$

TABLE 2. Summarized results of MEMD process.

Study site	Parameters used in MEMD: Ensemble number (N)	Parameters used in MEMD: Amplitude of the added white noise (ε)	Number of initial predictor variables	Number of IMFs & residual	Total number of predictor variables after MEMD
Gatton	500	0.2	8	13	104 (13×8)
Fordsdale	500	0.2	8	13	104 (13×8)
Cairns	500	0.2	8	12	96 (12×8)

$$r = \frac{\sum_{i=1}^{N}\left(ET^{OBS,i} - \overline{ET^{OBS,i}}\right)\left(ET^{FOR,i} - \overline{ET^{FOR}}\right)}{\sqrt{\sum_{i=1}^{N}\left(ET^{OBS,i} - \overline{ET^{OBS,i}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(ET^{FOR,i} - \overline{ET^{FOR,i}}\right)^2}}, \quad -1 \le r \le 1 \quad (10)$$

$$WI = 1 - \left[\frac{\sum_{i=1}^{N}\left(ET^{OBS,i} - ET^{FOR,i}\right)^2}{\sum_{i=1}^{N}\left(\left|\left(ET^{FOR,i} - \overline{ET^{OBS,i}}\right)\right| + \left|\left(ET^{OBS,i} - \overline{ET^{OBS,i}}\right)\right|\right)^2}\right], \quad 0 \le WI \le 1 \quad (16)$$

TABLE 3. Summarized results of Boruta feature selection process.

Study site	Number of initial predictor variables in each IMF and residual	Step 1: MEMD	Step2: Boruta Feature Selection													
			Number of selected predictor variables in each IMF and residuals													Total number of predictor variables identified after feature selection
		Total number of predictor variables after MEMD	IMF1	IMF2	IMF3	IMF4	IMF5	IMF6	IMF7	IMF8	IMF9	IMF10	IMF11	IMF12	Residual	
Gatton	8	104 (13×8)	6	6	7	8	8	8	8	8	8	8	8	8	8	99
Fordsdale	8	104 (13×8)	6	6	6	8	8	8	8	8	8	8	8	8	8	98
Cairns	8	96 (12×8)	6	6	6	6	8	8	8	8	8	8	8	-	8	88

TABLE 4. List of selected predictor variables identified in each IMF and residual used to develop hybrid MEMD-Boruta-LSTM, MEMD-Boruta-DNN and MEMD-Boruta-DT models in Cairns site.

		LSTM	DNN	DT
Standalone model		Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	Pa_{t-1}, Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	Pa_{t-1}, Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}
Hybrid model	IMF1	**6 significant inputs** Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**6 significant inputs** Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**6 significant inputs** Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}
	IMF2	**6 significant inputs** Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**6 significant inputs** Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**6 significant inputs** Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}
	IMF3	**6 significant inputs** Tsd_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**6 significant inputs** Tsd_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**6 significant inputs** Tsd_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}
	IMF4	**6 significant inputs** Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**6 significant inputs** Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**6 significant inputs** Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}
	IMF5	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}
	IMF6	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}
MEMD-Boruta	IMF7	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}
	IMF8	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}
	IMF9	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}
	IMF10	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}
	IMF11	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}
	Residual	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}	**8 significant inputs** Pat-1,Tsd_{t-1}, Tsn_{t-1}, Ebs_{t-1}, TR_{t-1}, $Tmax_{t-1}$, $Tmin_{t-1}$, Ra_{t-1}

X. Kling-Gupta Efficiency (*KGE*): The *KGE* measures the goodness-of-fit of the model. This metric can be decomposed into the contribution of mean, variance, and correlation on the model performance [45]. Perfect models will give value one for the *KGE* index [65].

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{CV_{FOR}}{CV_{OBS}}\right)^2 + \left(\frac{ET^{FOR,i}}{ET^{OBS,i}} - 1\right)^2} \quad (19)$$

where $CV =$ Coefficient of Variation, where $ET^{OBS,i}$ and $ET^{FOR,i}$ are observed and forecasted ith value of the evapotranspiration ET, $\overline{ET^{OBS,i}}$ and $\overline{ET^{FOR,i}}$ are the observed and forecasted average of ET and N is the total number of data points of the test dataset.

IV. RESULTS AND DISCUSSIONS

In the decomposition process, training and testing datasets for Gatton and Fordsdale sites were decomposed into 12 IMFs and a residual component (i.e. 104 (=**8** × **13**) predictors) whereas 11 IMFs and a residual component (i.e. 96 (=**8** × **12**) predictors) were generated by decomposing training and testing datasets for the Cairns site (see Table 2). In the Boruta feature selection process, 99 predictor variables were identified as significantly corelated to the target variable ET in all IMFs and the residual of Gatton, while 98 and 88 predictor variables were identified for Fordsdale, and Cairns sites respectively (see Table 3). Table 4 shows the selected final predictor variables in each IMF and the residual used to develop target hybrid MEMD-Boruta-LSTM model and benchmark models in Cairns site. Identified hyperparameters for the LSTM in target model through the hyperparameter optimization process are listed in Table 5.

The performance of the multi-stage deep MEMD-Boruta-LSTM model and other comparative models: MEMD-Boruta-DNN, MEMD-Boruta-DT, LSTM, DNN and, DT in the testing phase were assessed using statistical metrics calculated using equations (10) to (19), visual graphs, and error distributions between forecasted and observed ET.

Table 6 shows the results derived for statistical metrics: Correlation Coefficient (r), Root Mean Squared Error ($RMSE$; $mm\ day^{-1}$), Mean Absolute Error (MAE; $mm\ day^{-1}$), Willmott's Index (WI), Nash-Sutcliffe coefficient (NS), and Legates and McCabe's (LM). According to the results shown in table 6, the proposed multi-stage deep MEMD-Boruta-LSTM model has yielded the highest r, WI, NS, and LM and lowest $RMSE$ and MAE values over the other benchmark models at all study sites. For instance, values scored for r, WI, NS, and LM by this proposed model for the Gatton site where it showed the best performances among all study sites are 0.9668, 0.9723, 0.8960, and 0.6996 respectively and higher than the respective values scored by other benchmark models. Furthermore, for the same site, this proposed model scored 0.5307 and 0.4204 for $RMSE$ and MAE respectively, and these are the lowest recorded values. These results indicate that the proposed multi-stage deep hybrid MEMD-Boruta-LSTM model can be confidently employed for forecasting daily ET and for achieving higher forecasting accuracy compared to counterpart models (MEMD-Boruta-DNN and, MEMD-Boruta-DT) and standalone models (LSTM, DNN, and DT).

In terms of the Absolute Percentage Bias ($APB\%$) error and Kling-Gupta Efficiency (KGE) calculated in the testing phase, Figure 3(a) and 3(b) show that the proposed deep multi-stage MEMD-Boruta-LSTM model generates better performance in terms of $APB\%$ error percentage and KGE respectively. Figure 3(a) illustrations that the proposed MEMD-Boruta-LSTM model has scored the lowest APB error percentage (9.2-12.3%) while other all comparative models' APB error percentages are within (11.6-19.7%) range for all sites. According to Figure 3(b), the proposed MEMD-Boruta-LSTM model has yielded the highest KGE values (0.89-0.91) while, KGE values are less than 0.86 for other all benchmark models for all sites. These results also provide strong evidence to recognize the superior potentiality of the proposed multi-stage MEMD-Boruta-LSTM model in daily ET forecasting over the other benchmark models.

TABLE 5. List of hyperparameters for the LSTM model. The optimal parameters used for all sites are boldfaced (in blue). Note: ReLU, Uniform, He_uniform, Glorot_uniform, and adam stand for the rectified linear units, uniform initializer, He uniform variance scaling initializer, Glorot uniform initializer, and adaptive moment estimation respectively.

Model hyperparameter Name	Search space for optimal hyperparameter
LSTM Layer 1	[50,**100**,150]
LSTM Layer 2	[50,**100**,150]
LSTM Layer 3	[50,100,**150**]
LSTM Layer 4	[10,20,30,40,**50**]
LSTM Layer 5	[10,20,**30**,40,50]
LSTM Layer 6	[10,20,30,40,**50**]
Epochs	[30, 500, 100,**2000**]
Activation Function	[relu, **tanh**, sigmoid]
Weight Initializer	[uniform , he_uniform, **glorot_uniform**]
Recurrent Activation Function	[**relu**, tanh, sigmoid]
Optimizer	[**adam**]
Dropout Ratio	[0.1, **0.2**, 0.3]
Batch Size	[10,20,**30**]

Architecture of the backpropagation algorithm	
Alpha, α	0.001
Epsilon, ε	0.0000001
Beta, β_1, β_2	0.9, 0.999

α = Learning rate

ε = Small number to prevent any division by zero

β_1, β_2 = 1st, 2nd moment estimation exponential decay rate

The radar plots in Figure 4 demonstrate the proposed deep multi-stage MEMD-Boruta-LSTM model yielded the lowest values for $RRMSE,\%$ and $RMAE,\%$ for all sites (12.59% and 10.89% at Gatton, 16.21% and 15.06% at Fordsdale, 12.47% and 11.29% at Cairns respectively). Further, all values scored for $RRMSE$ and $RMAE$ for all sites by this proposed deep

TABLE 6. Performance evaluation of the proposed hybrid MEMD-Boruta-LSTM model in the testing phase for the comparative counterpart models in terms of the Pearson's Correlation Coefficient (r), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Willmott's Index (WI), Nash Sutcliffe coefficient (NS) and Legates and McCabe's (LM). The best model is boldfaced (in blue).

Study Site	Predictive Model	Model Performance Metrics					
		r	*RMSE* (*mm day^{-1}*)	*MAE* (*mm day^{-1}*)	*WI*	*NS*	*LM*
Gatton	**MEMD-Boruta-LSTM**	**0.9668**	**0.5307**	**0.4204**	**0.9723**	**0.8960**	**0.6996**
	MEMD-Boruta-DNN	0.8978	0.7637	0.5855	0.9277	0.7841	0.5807
	MEMD-Boruta-DT	0.9195	0.6495	0.4904	0.9547	0.8439	0.6488
	LSTM	0.8614	0.8627	0.6671	0.9200	0.7254	0.5231
	DNN	0.8527	0.9638	0.7471	0.8827	0.6562	0.465
	DT	0.8282	0.9431	0.7128	0.8972	0.6708	0.4895
Fordsdale	**MEMD-Boruta-LSTM**	**0.9343**	**0.5773**	**0.4380**	**0.9609**	**0.8424**	**0.6404**
	MEMD-Boruta-DNN	0.8427	0.8791	0.6824	0.8906	0.6348	0.4403
	MEMD-Boruta-DT	0.8936	0.7171	0.5549	0.9329	0.7575	0.5449
	LSTM	0.8364	0.8853	0.6366	0.8959	0.6295	0.4773
	DNN	0.8134	0.8836	0.7006	0.8628	0.6311	0.4252
	DT	0.7990	0.8827	0.6692	0.8818	0.6318	0.4511
Cairns	**MEMD-Boruta-LSTM**	**0.9044**	**0.5218**	**0.4130**	**0.9307**	**0.7655**	**0.5369**
	MEMD-Boruta-DNN	0.8108	0.6382	0.4881	0.8731	0.6482	0.4519
	MEMD-Boruta-DT	0.8240	0.6131	0.4663	0.9014	0.6754	0.4763
	LSTM	0.7261	0.7521	0.5610	0.8157	0.5128	0.3709
	DNN	0.7043	0.7700	0.5980	0.7855	0.4879	0.3285
	DT	0.6890	0.7827	0.5939	0.8033	0.4709	0.3331

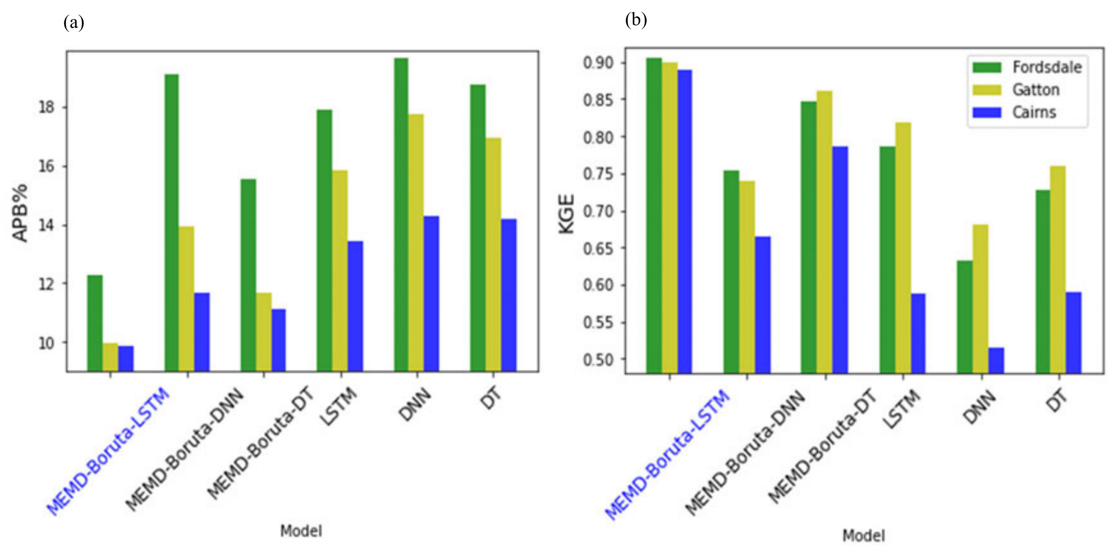


FIGURE 3. Bar graphs show the comprehensive assessment of the performance of the proposed MEMD-Boruta-LSTM model against the counterpart models, based on the (a) Absolute Percentage Bias (*APB*, %) error and (b) Kling-Gupta Efficiency (*KGE*) in the testing phase for the all study sites. The best model for all sites is boldfaced (in blue).

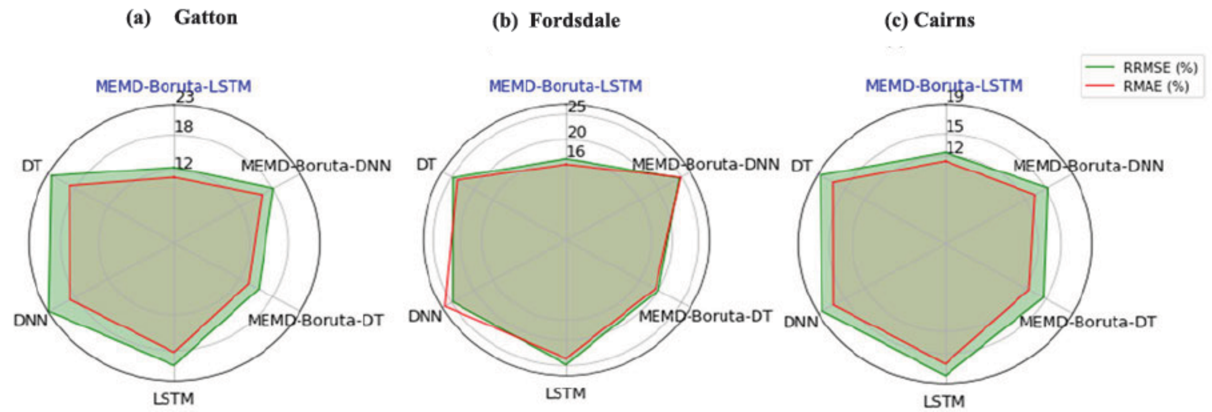


FIGURE 4. The radar plots showing the Relative Root Mean Squared Error (*RRMSE* %) and Relative Mean Absolute Error (*RMAE* %) of the MEMD-Boruta-LSTM hybrid model and comparative models constructed for 1-day evapotranspiration forecasting in the testing phase. The best model is boldfaced (in blue).

MEMD-Boruta-LSTM model is lying within the range of 10%-20%. Therefore, this proposed model can be categorized under the good model group having lower model errors of less than 20% [59], [60].

To further validate the proposed MEMD-Boruta-LSTM model the absolute Forecasting Errors (|FE| = |Observed *ET* – Forecasted *ET*|; mm) of this proposed model and all other benchmark models are compared. The box plots in Figure 5 depict the distribution of |FE| in the testing phase with their upper, median and, lower quartiles for all models and weather stations. The results of box plots shown in Figure 5 indicate that the proposed multi-stage MEMD-Boruta-LSTM model presented the smallest quartiles for |FE| for all sites followed by MEMD-Boruta-DNN, MEMD-Boruta-DT, LSTM, DNN, and DT. These results also clearly indicate that the proposed deep multi-stage MEMD-Boruta-LSTM model is superior to the other benchmark models.

The Empirical Cumulative Distribution Function (ECDF, Figure 6) is also used to illustrate the forecasting skills in terms of the absolute Forecasting Error, |FE| (mm) at each site. Forecasting errors of good models should be closer to zero. The all-hybrid MEMD-Boruta-LSTM, MEMD-Boruta-DNN, and MEMD-Boruta-DT models performed better than standalone LSTM, DNN, and DT models. Based on the forecasting error (0 to ±4 *mm*), Figure 6 visibly depicts that the proposed MEMD- Boruta-LSTM model is the most accurate compared to all other benchmark models.

In summary, the proposed multi-stage hybrid DL model (*i.e.*, MEMD-Boruta-LSTM) provided significant high performance with the lowest values of the absolute and relative errors i.e., *APB*, *RMSE*, *MAE*, *RRMSE*, and *RMAE*, including the highest *r*, *WI*, *NS*, *LM* and *KGE* in respect to the other benchmark models. Consequently, it is promising that the results confirm the deep multi-stage MEMD-Boruta-LSTM model has the potential to forecast daily *ET* and its performance exceeds that of all other comparative hybrid DL and standalone models for all the study sites in Queensland.

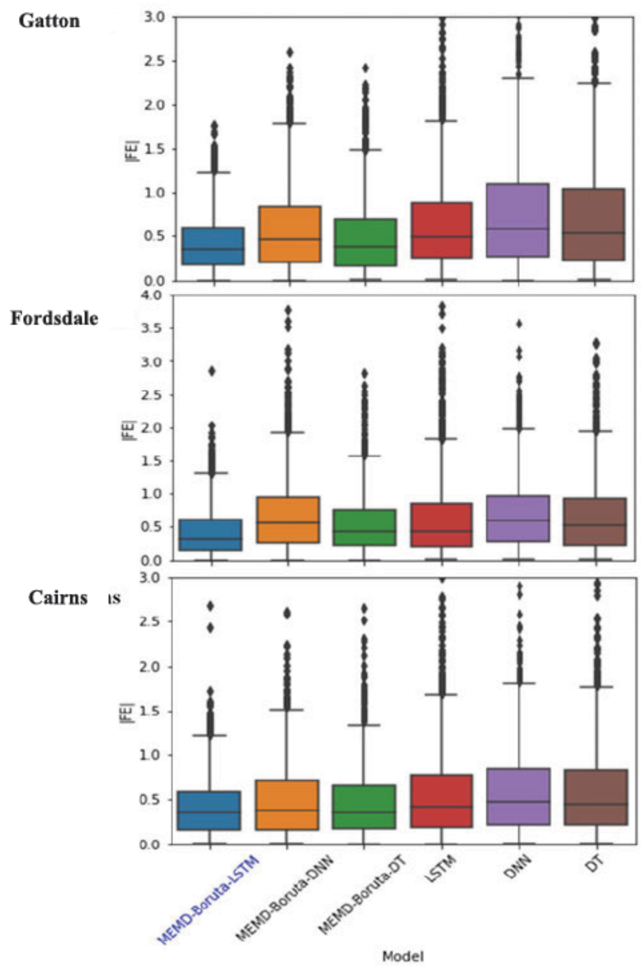


FIGURE 5. The box plots of the absolute value of the Forecasting Errors (|FE|) in the testing phase, generated by the hybrid MEMD-Boruta-LSTM model compared to that of the other predictive models implemented at all study sites. The best model is boldfaced (in blue).

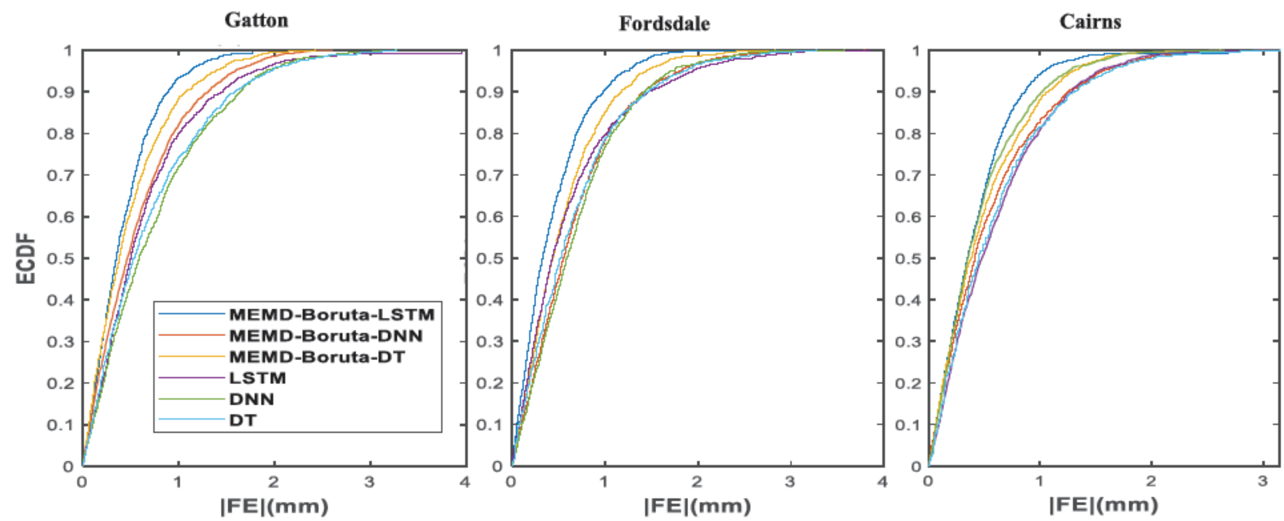


FIGURE 6. Empirical Cumulative Distribution Function (ECDF) of absolute forecasting error, |FE| (mm) of the testing data using MEMD-Boruta-LSTM *vs*. MEMD-Boruta-DNN, MEMD-Boruta-DT, and standalone LSTM, DNN, and DT models in forecasting *ET* for all study sites.

V. CONCLUSION

This study aims to design a novel deep learning multi-stage hybrid MEMD-Boruta-LSTM model as a practical tool to forecast daily *ET* using satellite and ground-based variables.

The Multivariate Empirical Mode Decomposition (MEMD) is incorporated with LSTM to decompose predictor variable data into IMFs and residuals and the Boruta-Random Forest (Boruta) feature selection method has been employed to screen the most correlated predictor variables to target variable *ET* in each IMFs and residuals. The daily predictor and target variable data (01 February 2003 to 19 April 2011) were extracted from GIOVANNI-AIRS, GLDAS model satellites, and the SILO ground database of the Queensland government. The test sites included Gatton, Fordsdale, and Cairns, which are located in drought-prone regions in Queensland, Australia. The integration of LSTM with MEMD and Boruta resulted in a novel multi-stage deep learning MEMD-Boruta-LSTM hybrid model whose performance was evaluated using statistical score metrics and compared with the other hybrid and standalone models namely, MEMD-Boruta-DNN, MEMD-Boruta-DT, LSTM, DNN, and DT based approaches.

The MEMD-Boruta-LSTM hybrid model yielded the highest values for normalized performance metrics: *r, NS, WI, LM* (see Table 6) and the lowest values for *RMSE, MAE, RRMSE*, and *APB* for all sites. Meanwhile, the results also revealed that all hybrid models (MEMD-Boruta-LSTM, MEMD-Boruta-DNN, MEMD-Boruta-DT) remarkably outperformed in comparison with the standalone models (LSTM, DNN, DT) in forecasting *ET* at all study sites (see Table 6). This comparison provides strong evidence to verify that MEMD decomposition and Boruta feature selection methods can be used effectively to improve the forecasting accuracy of any model. All findings of this study confirm that the proposed multi-stage deep hybrid MEMD-Boruta-LSTM model outperformed the comparative hybrid and standalone models in forecasting *ET* at a daily forecasting horizon.

The novel proposed deep hybrid MEMD-Boruta-LSTM model can be practically employed for precise forecasting of *ET*. Evapotranspiration is the main causative natural phenomenon that contributes to the water losses from croplands. By multiplying forecasted *ET* with the relevant crop factor, which is a unique value for individual crops, the water loss due to evapotranspiration can be estimated in advance, which will be helpful in planning precise irrigation schedules for the future while avoiding the wastage of water resources in drought-prone areas. In addition, this multi-stage deep learning hybrid model used for forecasting *ET* is likely to lead to significant financial benefits to the farmers, in arid and semi-arid regions where agricultural practices are adversely affected by the scarcity of water resources.

VI. LIMITATION AND FUTURE RESEARCH

For this model development, data were extracted only for three sites within Queensland (as a case study) as it is impracticable to select more sites representing the whole drought-affected region in Australia or elsewhere. However, this pioneering study has produced a new modelling framework for *ET* forecasting and paves the way for future studies with a wider scope. For example, the geographic consistency of the MEMD-Boruta-LSTM hybrid model, together with its accuracy can be considered in future research. Moreover, the potential use of multi-stage MEMD-Boruta-LSTM for multi-step ahead daily *ET* (7 days, 15 days, 30 days) forecasting can be researched. Further, instead of the MEMD technique for data pre- Decomposition (VMD) technique [61] can be used with Boruta-LSTM to build up a new two-stage deep forecasting model to forecast *ET*.

ACKNOWLEDGMENT

The authors would like to thank NASA's Goddard Online Interactive Visualization and Analysis Infrastructure (GIOVANNI) satellite and Scientific Information for Landowners (SILO) for meteorological data used and the University of Southern Queensland (USQ), Australia, and Wayamba University of Sri Lanka for funding this research, and also would like to thank to the editor and all reviewers for their valuable suggestions in improving the quality of the paper.

REFERENCES

[1] L. S. Pereira, R. G. Allen, M. Smith, and D. Raes, "Crop evapotranspiration estimation with FAO56: Past and future," *Agricult. Water Manage.*, vol. 147, pp. 4–20, Jan. 2015.

[2] G. Huang, L. Wu, X. Ma, W. Zhang, J. Fan, X. Yu, W. Zeng, and H. Zhou, "Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions," *J. Hydrol.*, vol. 574, pp. 1029–1041, Jul. 2019, doi: 10.1016/j.jhydrol.2019.04.085.

[3] M. Wu, Q. Feng, X. Wen, R. C. Deo, Z. Yin, L. Yang, and D. Sheng, "Random forest predictive model development with uncertainty analysis capability for the estimation of evapotranspiration in an arid oasis region," *Hydrol. Res.*, vol. 51, no. 4, pp. 648–665, Aug. 2020, doi: 10.2166/nh.2020.012.

[4] V. Nourani, G. Elkiran, and J. Abdullahi, "Multi-step ahead modeling of reference evapotranspiration using a multi-model approach," *J. Hydrol.*, vol. 581, Feb. 2020, Art. no. 124434, doi: 10.1016/j.jhydrol.2019.124434.

[5] Y. Tikhamarine, A. Malik, A. Kumar, D. Souag-Gamane, and O. Kisi, "Estimation of monthly reference evapotranspiration using novel hybrid machine learning approaches," *Hydrol. Sci. J.*, vol. 64, no. 15, pp. 1824–1842, Nov. 2019, doi: 10.1080/02626667.2019.1678750.

[6] Z. Chen, S. Sun, Y. Wang, Q. Wang, and X. Zhang, "Temporal convolution-network-based models for modeling maize evapotranspiration under mulched drip irrigation," *Comput. Electron. Agricult.*, vol. 169, Feb. 2020, Art. no. 105206, doi: 10.1016/j.compag.2019.105206.

[7] P. de Oliveira e Lucas, M. A. Alves, P. C. de Lima e Silva, and F. G. Guimarães, "Reference evapotranspiration time series forecasting with ensemble of convolutional neural networks," *Comput. Electron. Agricult.*, vol. 177, Oct. 2020, Art. no. 105700, doi: 10.1016/j.compag.2020.105700.

[8] J. Zhang, Y. Zhu, X. Zhang, M. Ye, and J. Yan, "Developing a long short-term memory (LSTM) based model for predicting water table depth in agricultural areas," *J. Hydrol.*, vol. 561, pp. 918–929, Jun. 2018.

[9] S. Ghimire, R. C. Deo, N. Raj, and J. Mi, "Deep learning neural networks trained with MODIS satellite-derived predictors for long-term global solar radiation prediction," *Energies*, vol. 12, no. 12, p. 2407, Jun. 2019.

[10] M. Fu, T. Fan, Z. Ding, S. Q. Salih, N. Al-Ansari, and Z. M. Yaseen, "Deep learning data-intelligence model based on adjusted forecasting window scale: Application in daily streamflow simulation," *IEEE Access*, vol. 8, pp. 32632–32651, 2020.

[11] M. Gauch, F. Kratzert, D. Klotz, G. Nearing, J. Lin, and S. Hochreiter, "Rainfall-runoff prediction at multiple timescales with a single long short-term memory network," *Hydrol. Earth Syst. Sci.*, vol. 25, no. 4, pp. 2045–2062, 2021.

[12] J. Yin, Z. Deng, A. V. M. Ines, J. Wu, and E. Rasu, "Forecast of short-term daily reference evapotranspiration under limited meteorological variables using a hybrid bi-directional long short-term memory model (Bi-LSTM)," *Agricult. Water Manage.*, vol. 242, Dec. 2020, Art. no. 106386, doi: 10.1016/j.agwat.2020.106386.

[13] L. B. Ferreira and F. F. da Cunha, "Multi-step ahead forecasting of daily reference evapotranspiration using deep learning," *Comput. Electron. Agricult.*, vol. 178, Nov. 2020, Art. no. 105728, doi: 10.1016/j.compag.2020.105728.

[14] M. S. AL-Musaylh, R. C. Deo, Y. Li, and J. F. Adamowski, "Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting," *Appl. Energy*, vol. 217, pp. 422–439, May 2018, doi: 10.1016/j.apenergy.2018.02.140.

[15] R. Prasad, M. Ali, P. Kwan, and H. Khan, "Designing a multi-stage multivariate empirical mode decomposition coupled with ant colony optimization and random forest model to forecast monthly solar radiation," *Appl. Energy*, vol. 236, pp. 778–792, Feb. 2019.

[16] N. ur Rehman and D. P. Mandic, "Multivariate empirical mode decomposition," *Proc. R. Soc. Lond. A, Math., Phys. Eng. Sci.*, vol. 466, no. 2117, pp. 1291–1302, 2009.

[17] X. Lang, Q. Zheng, Z. Zhang, S. Lu, L. Xie, A. Horch, and H. Su, "Fast multivariate empirical mode decomposition," *IEEE Access*, vol. 6, pp. 65521–65538, 2018.

[18] M. Ali, R. C. Deo, T. Maraseni, and N. J. Downs, "Improving SPI-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridized with simulated annealing and kernel ridge regression algorithms," *J. Hydrol.*, vol. 576, pp. 164–184, Sep. 2019.

[19] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, "Boruta—A system for feature selection," *Fundamenta Informaticae*, vol. 101, no. 4, pp. 271–285, 2010.

[20] A. M. Ahmed, R. C. Deo, A. Ghahramani, N. Raj, Q. Feng, Z. Yin, and L. Yang, "LSTM integrated with Boruta-random forest optimiser for soil moisture estimation under RCP4. 5 and RCP8. 5 global warming scenarios," *Stochastic Environ. Res. Risk Assessment*, vol. 35, pp. 1–31, Jan. 2021.

[21] R. Prasad, R. C. Deo, Y. Li, and T. Maraseni, "Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and boruta-random forest hybridizer algorithm approach," *Catena*, vol. 177, pp. 149–166, Jun. 2019.

[22] J. Qu, K. Ren, and X. Shi, "Binary grey wolf optimization-regularized extreme learning machine wrapper coupled with the Boruta algorithm for monthly streamflow forecasting," *Water Resour. Manage.*, vol. 35, no. 3, pp. 1029–1045, Feb. 2021.

[23] A. A. M. Ahmed, R. C. Deo, Q. Feng, A. Ghahramani, N. Raj, Z. Yin, and L. Yang, "Deep learning hybrid model with boruta-random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity," *J. Hydrol.*, vol. 599, Aug. 2021, Art. no. 126350.

[24] B. Lyu, Y. Zhang, and Y. Hu, "Improving PM2.5 air quality model forecasts in China using a bias-correction framework," *Atmosphere*, vol. 8, no. 12, p. 147, Aug. 2017.

[25] I. Craig, A. Green, M. Scobie, and E. Schmidt, "Controlling evaporation loss from water storages," Nat. Centre Eng. Agricult., Univ. Southern Queensland, Toowoomba, QLD, Australia, Tech. Rep. 1000580/1, 2005.

[26] N. Ur Rehman and D. P. Mandic, "Filter bank property of multivariate empirical mode decomposition," *IEEE Trans. Signal Process.*, vol. 59, no. 5, pp. 2421–2426, May 2011.

[27] N. E. Huang, M.-L. C. Wu, S. R. Long, S. S. P. Shen, W. Qu, P. Gloersen, and K. L. Fan, "A confidence limit for the empirical mode decomposition and Hilbert spectral analysis," *Proc. R. Soc. Lond. A, Math. Phys. Eng. Sci.*, vol. 459, no. 2037, pp. 2317–2345, 2003.

[28] W. Hu and B. C. Si, "Soil water prediction based on its scale-specific control using multivariate empirical mode decomposition," *Geoderma*, vol.s. 193–194, pp. 180–188, Feb. 2013.

[29] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010.

[30] J.-H. Hur, S.-Y. Ihm, and Y.-H. Park, "A variable impacts measurement in random forest for mobile cloud computing," *Wireless Commun. Mobile Comput.*, vol. 2017, pp. 1–13, Oct. 2017.

[31] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinform.*, vol. 9, no. 307, pp. 1–11, 2008.

[32] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[33] N. K. Manaswi, *Deep Learning with Applications Using Python: Chatbots and Face, Object, and Speech Recognition with TensorFlow and Keras*, 1 ed. Berkeley, CA, USA: Apress L, 2018.

[34] J. Chen, G.-Q. Zeng, W. Zhou, W. Du, and K.-D. Lu, "Wind speed forecasting using nonlinear-learning ensemble of deep learning time series prediction and extremal optimization," *Energ. Convers. Manage*, vol. 165, pp. 681–695, Jun. 2018.

[35] X. Zhang, Q. Zhang, G. Zhang, Z. Nie, Z. Gui, and H. Que, "A novel hybrid data-driven model for daily land surface temperature forecasting using long short-term memory neural network based on ensemble empirical mode decomposition," *Int. J. Environ. Res. Public Health*, vol. 15, no. 5, p. 1032, May 2018, doi: 10.3390/ijerph15051032.
[36] *About my Region—Queensland*, DOAWE, Dept. Agricult., Water Environ, Canberra, ACT, Australia, 2021.
[37] Queensland. *Drought Declarations*. Accessed: Aug. 20, 2021. [Online]. Available: https://www.longpaddock.qld.gov.au
[38] W. Teng, H. Rui, B. Vollmer, R. de Jeu, F. Fang, G.-D. Lei, and R. Parinussa, "NASA Giovanni: A tool for visualizing, analyzing, and intercomparing soil moisture data," *Remote Sens. Terrestrial Water Cycle*, vol. 206, p. 331, Oct. 2014.
[39] A. Morshed, J. Aryal, and R. Dutta, "Environmental spatio-temporal ontology for the linked open data cloud," in *Proc. 12th IEEE Int. Conf. Trust, Secur. Privacy Comput. Commun.*, Jul. 2013, pp. 1907–1912.
[40] S. Ghimire, R. C. Deo, N. Raj, and J. Mi, "Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms," *Appl. Energy*, vol. 253, Nov. 2019, Art. no. 113541, doi: 10.1016/j.apenergy.2019.113541.
[41] D. R. Legates and G. J. McCabe, "Evaluating the use of 'goodness-of-fi' measures in hydrologic and hydroclimatic model validation," *Water Resour. Res.*, vol. 35, no. 1, pp. 233–241, Jan. 1999, doi: 10.1029/1998wr900018.
[42] D. R. Legates and G. J. McCabe, "A refined index of model performance: A rejoinder," *Int. J. Climatol.*, vol. 33, no. 4, pp. 1053–1056, Mar. 2013, doi: 10.1002/joc.3487.
[43] J. E. Nash and J. V. Sutcliffe, "River flow forecasting through conceptual models part I—A discussion of principles," *J. Hydrol.*, vol. 10, no. 3, pp. 282–290, Apr. 1970, doi: 10.1016/0022-1694(70)90255-6.
[44] C. J. Willmott, S. M. Robeson, and K. Matsuura, "A refined index of model performance," *Int. J. Climatol.*, vol. 32, no. 13, pp. 2088–2094, 2012.
[45] H. Kling and H. Gupta, "On the development of regionalization relationships for lumped watershed models: The impact of ignoring sub-basin scale variability," *J. Hydrol.*, vol. 373, nos. 3–4, pp. 337–351, Jul. 2009.
[46] N. Ketkar, "Introduction to keras," in *Deep Learning With Python*. CA, USA: Springer, 2017, pp. 97–111.
[47] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, and M. Kudlur, "Tensorflow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Operating Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.
[48] M. F. Sanner, "Python: A programming language for software integration and development," *J. Mol. Graph Model.*, vol. 17, no. 1, pp. 57–61, 1999.
[49] J. Quilty and J. Adamowski, "Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models for real-world applications with best practices and a new forecasting framework," *J. Hydrol.*, vol. 563, pp. 336–353, Aug. 2018.
[50] R. C. Deo, M. K. Tiwari, J. F. Adamowski, and J. M. Quilty, "Forecasting effective drought index using a wavelet extreme learning machine (W-ELM) model," *Stochastic Environ. Res. Risk Assessment*, vol. 31, no. 5, pp. 1211–1240, Jul. 2017.
[51] Q. Ouyang, W. Lu, X. Xin, Y. Zhang, W. Cheng, and T. Yu, "Monthly rainfall forecasting using EEMD-SVR based on phase-space reconstruction," *Water Resour. Manage.*, vol. 30, no. 7, pp. 2311–2325, May 2016.
[52] Y. Ren, P. N. Suganthan, and N. Srikanth, "A comparative study of empirical mode decomposition-based short-term wind speed forecasting methods," *IEEE Trans. Sustain. Energy*, vol. 6, no. 1, pp. 236–244, Dec. 2014.
[53] W. C. Wang, D. M. Xu, K. W. Chau, and S. Chen, "Improved annual rainfall-runoff forecasting using PSO-SVM model based on EEMD," *J. Hydroinform.*, vol. 15, no. 4, pp. 1377–1390, 2013.
[54] Z. Wu and N. E. Huang, "Ensemble empirical mode decomposition: A noise-assisted data analysis method," *Adv. Adapt. Data Anal.*, vol. 1, no. 1, pp. 1–41, 2008.
[55] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: Methods and prospects," *Big Data Analytics*, vol. 1, no. 1, pp. 1–22, Dec. 2016, doi: 10.1186/s41044-016-0014-0.
[56] B. Komer, J. Bergstra, and C. Eliasmith, "Hyperopt-sklearn," in *Automated Machine Learning*. Cham, Switzerland: Springer, 2019, pp. 97–111.
[57] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, "Hyperopt: A Python library for model selection and hyperparameter optimization," *Comput. Sci. Discovery*, vol. 8, no. 1, Jul. 2015, Art. no. 014008.
[58] S. Putatunda and K. Rama, "A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost," in *Proc. Int. Conf. Signal Process. Mach. Learn. (SPML)*, 2018, pp. 6–10.
[59] K. Mohammadi, S. Shamshirband, M. H. Anisi, K. A. Alam, and D. Petković, "Support vector regression based prediction of global solar radiation on a horizontal surface," *Energy Convers. Manage.*, vol. 91, pp. 433–441, Feb. 2015.
[60] K. Mohammadi, S. Shamshirband, C. W. Tong, M. Arif, D. Petković, and S. Ch, "A new hybrid support vector machine–wavelet transform approach for estimation of horizontal global solar radiation," *Energy Convers. Manage.*, vol. 92, pp. 162–171, Mar. 2015.
[61] K. Dragomiretskiy and D. Zosso, "Variational mode decomposition," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 531–544, Feb. 2014.
[62] *AgriMetSoft (2019). Online Calculators*, AgriMetSoft, South Korea, 2020.
[63] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," *Climate Res.*, vol. 30, no. 1, pp. 79–82, 2005.
[64] C. Ertekin and O. Yaldiz, "Comparison of some existing models for estimating global solar radiation for Antalya (Turkey)," *Energy Convers. Manage.*, vol. 41, no. 4, pp. 311–330, 2000, doi: 10.1016/s0196-8904(99)00127-2.
[65] S. Ghimire, R. C. Deo, N. Raj, and J. Mi, "Deep learning neural networks trained with MODIS satellite-derived predictors for long-term global solar radiation prediction," *Energies*, vol. 12, no. 12, p. 2407, Jun. 2019.
[66] T. Yang, L. Zhang, T. Kim, Y. Hong, D. Zhang, and Q. Peng, "A large-scale comparison of artificial intelligence and data mining (AI&DM) techniques in simulating reservoir releases over the upper Colorado region," *J. Hydrol.*, vol. 602, Nov. 2021, Art. no. 126723.
[67] H. Apaydin, H. Feizi, M. T. Sattari, M. S. Colak, S. Shamshirband, and K.-W. Chau, "Comparative analysis of recurrent neural network architectures for reservoir inflow forecasting," *Water*, vol. 12, no. 5, p. 1500, 2020.
[68] H. Fan, M. Jiang, L. Xu, H. Zhu, J. Cheng, and J. Jiang, "Comparison of long short term memory networks and the hydrological model in runoff simulation," *Water*, vol. 12, no. 1, p. 175, Jan. 2020.
[69] I.-F. Kao, Y. Zhou, L.-C. Chang, and F.-J. Chang, "Exploring a long short-term memory based encoder-decoder framework for multi-step-ahead flood forecasting," *J. Hydrol.*, vol. 583, Apr. 2020, Art. no. 124631.
[70] B. B. Sahoo, R. Jha, A. Singh, and D. Kumar, "Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting," *Acta Geophysica*, vol. 67, no. 5, pp. 1471–1481, 2019.
[71] Y. Wu, Y. Ding, Y. Zhu, J. Feng, and S. Wang, "Complexity to forecast flood: Problem definition and spatiotemporal attention LSTM solution," *Complexity*, vol. 2020, pp. 1–13, Mar. 2020.
[72] T. Yang, L. Zhang, T. Kim, Y. Hong, D. Zhang, and Q. Peng, "A large-scale comparison of artificial intelligence and data mining (AI&DM) techniques in simulating reservoir releases over the upper Colorado region," *J. Hydrol.*, vol. 602, Nov. 2021, Art. no. 126723.
[73] A. R. S. Kumar, M. K. Goyal, C. S. P. Ojha, R. D. Singh, P. K. Swamee, and R. K. Nema, "Application of ANN, fuzzy logic and decision tree algorithms for the development of reservoir operating rules," *Water Resour. Manag.*, vol. 27, no. 3, pp. 911–925, Feb. 2013.
[74] T. M. Shafapour, B. Pradhan, and M. N. Jebur, "Spatial prediction of flood susceptible areas using rule based decision tree (DT) and a novel ensemble bivariate and multivariate statistical models in GIS," *J. Hydrol.*, vol. 504, no. 8, pp. 69–79, Nov. 2013.
[75] Y. Cui, L. Song, and W. Fan, "Generation of spatio-temporally continuous evapotranspiration and its components by coupling a two-source energy balance model and a deep neural network over the Heihe river basin," *J. Hydrol.*, vol. 597, Jun. 2021, Art. no. 126176.
[76] S. Lee, D. Kaown, E. H. Koh, K. S. Ko, and K. K. Lee, "Delineation of groundwater quality locations suitable for target end-use purposes through deep neural network models," *Wiley Online Library*, vol. 50, pp. 0047–2425, Mar. 2021.
[77] X. H. Le, H. V. Ho, G. Lee, and S. Jung, "Application of long short-term memory (LSTM) neural network for flood forecasting," *Water*, vol. 11, no. 7, p. 1387, Jul. 2019.

W. J. M. LAKMINI PRARTHANA JAYASINGHE (Member, IEEE) received the B.Sc. degree (Hons.), in 2008, and the M.Phil. degree in mathematical sciences from the University of Kelaniya, Sri Lanka, in 2014. She is currently pursuing the Ph.D. degree in artificial intelligence with the University of Southern Queensland, Australia. She has been a Senior Lecturer with the Department of Mathematical Sciences, Wayamba University of Sri Lanka, since 2010. Her current research work is co-funded by the University of Southern Queensland, and the Wayamba University of Sri Lanka. Her research interests include deep learning, remote sensing, machine leaning, and mathematical modeling. She is a member of ADSN, NSF, and SLAAS. She has won Physical Sciences Award, in 2008, provided by SLAAS, Sri Lanka.

RAVINESH C. DEO (Senior Member, IEEE) is currently a Professor with the University of Southern Queensland, Australia. His research interests include data science, knowledge & data engineering, deep learning, artificial intelligence (AI), and machine learning (ML) for decision systems. Due to his research leadership, he was awarded internationally competitive fellowships, such as Queensland Smithsonian, Australia–China, JSPS, Chinese Academy of Science Presidential Fellowship, Australia–India Strategic Fellowship and Endeavour Fellowship. He is a 2021 Clarivate Highly Cited Researcher. He has supervised over 25 research degrees, received Employee Excellence Award, such as an Excellence in Research, an Excellence in Postgraduate Supervision, Publication Excellence and Teaching Commendation. He is on the Editorial Board of *Stochastic Environmental Research & Risk Assessment*, IEEE ACCESS, *Remote Sensing, Energies* and *Journal of Hydrologic Engineering*. He has published over 230 articles, *incl*. 170 journals, seven books in Elsevier, Springer and IGI and 23 book chapters with cumulative citations exceeding eight,100 with an H-index of 48 and SciVal Field-Weighted Citation Index over 3.0.

AFSHIN GHAHRAMANI received the master's and Ph.D. degrees from the Tokyo University of Agriculture and Technology. He is currently a Senior Researcher with the University of Southern Queensland, Australia. Previously, he worked as a Postdoctoral Fellow and a Research Scientist at CSIRO and the University of Tsukuba, Japan. He has published more than 40 research papers and have led more than ten research projects. His research interests include sustainable and efficient water and soil management by undertaking multidisciplinary research in the areas of water/soil/agricultural systems & modeling.

SUJAN GHIMIRE received the degree in mechanical engineering from Kathmandu University, Nepal, in 2002, the M.S. degree in renewable energy from the Institute of Engineering, Nepal, in 2007, and the Ph.D. degree in global solar radiation modeling and prediction using artificial intelligence from the University of Southern Queensland, Australia. He is currently an Adjunct Lecturer with University of Southern Queensland and is working as the Territory Customer Support Manager with John Deere Ltd., Australia. His current research interests include optimization algorithms, nature-inspired metaheuristics, machine learning, and feature selection problem for real world problems.

NAWIN RAJ received the B.Sc., B.Ed., PGDMA, and M.Sc. degrees from the University of the South Pacific in the area of computational fluid dynamics, and the Ph.D. degree from the School of Sciences, University of Southern Queensland (USQ), QLD, Australia, in 2015. From 2007 to 2010, he was a Lecturer with Fiji National University. From 2015 to 2016, he worked as a Learning Advisor at USQ, where he is currently a Lecturer. He is a member of Australian Mathematical Society and the Queensland College of Teachers. His research interests include artificial intelligence, deep learning, non-linear oscillation, computational fluid dynamics, and oceanography.

5.3 Links and implications

Evapotranspiration is the main causative natural phenomenon that contributes to the water losses from croplands by evaporation and transpiration. By multiplying forecasted *ET* with the relevant crop factor, which is a unique value for individual crops, the water loss due to evapotranspiration can be estimated in advance. This will help to make precise irrigation schedules, drought event management, and long-term strategic planning for the future in drought-prone areas. All these usefulnesses are ultimately likely to bring significant financial benefits to the farmers, in arid and semi-arid regions where agricultural practices are adversely affected by the scarcity of water resources. Further, this initiative, which has produced a new modelling methodology for *ET* forecasting paving the way for future studies with a wider scope of investigating the terrestrial consistency of the proposed MEMD-Boruta-LSTM hybrid model, together with its accuracy. Moreover, the potential use of multi-stage MEMD-Boruta-LSTM for multi-step ahead long-term *ET* like one month, six months, or one year ahead forecasting can be researched. Further, instead of the MEMD technique for data decomposition, the Variation Mode Decomposition (VMD) technique can be used with Boruta-LSTM to build up a new two-stage deep forecasting model to forecast *ET*.

However, E_P and *ET* which are discussed in objective 1 and 2 respectively give an estimate of water loss. Water availability in soil is also a crucial factor to be considered simultaneously with the water loss due to evaporation and evapotranspiration in water resources management, drought monitoring, and early identification of bushfires and flood disasters. Soil moisture (*SM)* is a hydrological parameter that gives the knowledge and estimate of water availability in soil and is almost equally useful as E_P and *ET* in drought event management. Therefore, the third objective of this PhD study focused to develop a deep learning model to forecast soil moisture *(SM)* on topsoil (0-10 mm depth). The next chapter will explain the research outcome of this third objective in detail.

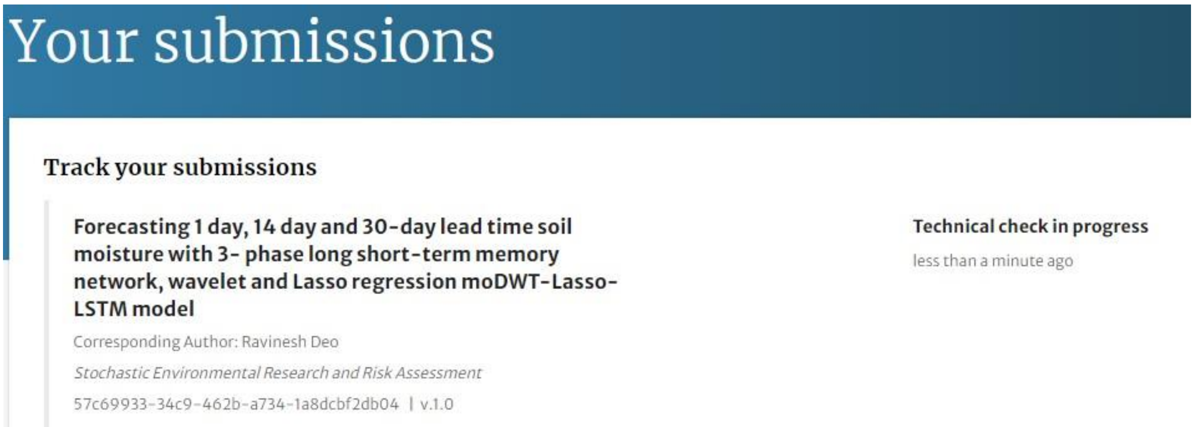
CHAPTER 6: PAPER 3 - SOIL MOISTURE FORECASTING AT 1 DAY, 14 DAYS, AND 30 DAYS AHEAD HORIZON WITH 3-PHASE DEEP LEARNING LONG SHORT-TERM MEMORY NETWORK, WAVELET, AND LASSO REGRESSION moDWT-Lasso-LSTM APPROACH.

6.1 Introduction

This chapter is an identical replication of the article that is submitted to Journal of *Stochastic Environmental Research and Risk Assessment.*

This study develops a multi-step forecasting model for soil moisture (*SM*) in the 0-10 cm depth using a data-driven deep learning hybrid approach by incorporating satellite and ground data. Due to the nonstationary and nonlinear characters of the collected data, the original data were decomposed using the Maximum Overlap Discrete Wavelet Transform (moDWT) decomposition and then selected its features using the Least Absolute Shrinkage and Selection Operator (Lasso). The deep learning Long Short-Term Memory (LSTM) algorithm was then employed to construct the target proposed 3-phase hybrid moDWT-Lasso-LSTM model for 1 day, 14 days, and 30 days ahead *SM* forecasting in Bundaberg Queensland, Australia. This proposed model's performance was statistically compared to benchmarked alternative machine learning models to confirm its viability. Statistical metrics and forecasting error plots were used to assess the performance of the target model against alternative models. The results revealed that, in comparison to other techniques, the 3-phase hybrid deep moDWT-Lasso-LSTM is showing comparatively low errors. This study ascertains that the suggested 3-phase hybrid deep multi-step moDWT-Lasso-LSTM model can be successfully employed as a viable data-driven device for multi-step *SM* forecasting in the topsoil layer (0-10 cm depth).

6.2 Paper under review



Forecasting 1 day, 14 day and 30-day lead time soil moisture with 3-phase long short-term memory network, wavelet and Lasso regression moDWT-Lasso-LSTM model

W.J.M. Lakmini Prarthana Jayasinghe [a,*,c], Ravinesh C. Deo [a,*] , Nawin Raj [a], Sujan Ghimire [a], Afshin Ghahramani [b]

[a] *School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, QLD 4300, Australia*

[b] *Centre for Sustainable Agricultural Systems, University of Southern Queensland, Toowoomba, QLD 4500, Australia*

[c] *Department of Mathematical Sciences, Wayamba University of Sri Lanka, Sri Lanka.*

Abstract

To develop agricultural risk management strategies, identifying water deficits early in the growing cycle is beneficial. Using a data-driven deep learning hybrid approach, this study develops a multi-step soil moisture forecasting model for 1 day, 14 days, and 30 days in the Bundaberg region in Queensland, Australia. To develop the proposed model, Geospatial Interactive Online Visualization and Analysis Infrastructure (satellite) data are combined with ground observations. Due to the periodicity, transientity, and trends in soil moisture in the top layer, time series datasets are relatively complex. Therefore, we decomposed these using the Maximum Overlap Discrete Wavelet Transform (moDWT) method to identify the best correlated wavelet and scaling coefficients of the predictor variables with the target top layer moisture, while the proposed 3-phase hybrid moDWT-Lasso-LSTM model is fully trained using the Least Absolute Shrinkage and Selection Operator (Lasso) method. Using Hyperopt algorithm, the optimal hyperparameters of the model were identified using a deep learning LSTM method and compared with benchmarked machine learning models. In total, nine models were developed, including three standalone models (e.g., LSTM), three models with feature selection (e.g., Lasso-LSTM), and three hybrid models with wavelet decomposition and feature selection (e.g., moDWT-Lasso-LSTM). To compare the target model with alternative models, we used statistical metrics such as Correlation Coefficient, Determination of Coefficient, Mean Absolute Error, Mean Absolute Scaled Error, Symmetric Mean Absolute Percentage Error, Root Mean Squared Error, Nash-Sutcliffe Index, Willmott Index, Legates and McCabe Index, scatter plots, and forecasting error plots. In comparison to alternative methods, the hybrid deep moDWT-Lasso-LSTM produced relatively few errors. Based on this study, we demonstrate that the moDWT-Lasso-LSTM 3-phase hybrid model can be successfully used as a data-driven device for forecasting multistep soil moisture in Bundaberg, Queensland, Australia.

1. Introduction

Soil moisture, as part of the soil-plant-atmosphere water cycle, refers to the water that is present in the soil and is essential for maintaining plant growth Liao et al. (2018). It is a key factor in determining irrigation water requirements (Chang et al., 2015). Forecasting soil moisture is very useful in understanding the future trends of soil moisture levels in advance and accordingly managing water stress conditions affecting crops and planning the irrigation schedules while conserving limited water resources. The land resources of the Bundaberg region in Queensland, Australia, the region considered in this study for developing a soil moisture forecasting model are extensively used for growing commercial crops. Thus, such a soil moisture forecasting model will be very beneficial for agricultural operations in this region.

Data-driven predictive models have shown comparatively higher competency in soil moisture prediction (Prasad et al., 2019a) and other hydro-meteorological variables prediction like evaporation (Jayasinghe et al., 2022, Ghorbani et al., 2018), precipitation (Ortiz-García et al., 2014, Silverman and Dracup, 2000), drought (Khan et al., 2020, Belayneh et al., 2016), evapotranspiration (Jayasinghe et al., 2021, Zhu et al., 2020) and river flow (Deo and Şahin, 2016, Huang et al., 2014). Jamei et al. (2022) employed data-driven predictive tools to forecast soil moisture and in their work, Extreme Gradient Boosting (XGBoost) and Categorical Boosting (CatBoost), two modern ensemble-based ML models, were integrated with the Empirical Wavelet Transform (EWT) to predict long-term multi-step ahead daily root zone soil moisture (RZSM) in highly cold semi-arid and highly warm semi-humid regions (Ardabil and Minab, respectively) and their performances were compared with rival models. The results have demonstrated the superior performance of the EWT-CatBoost and EWT-XGBoost models over the other counterpart models in forecasting multi-step ahead RZSM at Ardabil and Minab sites, respectively. Jamei et al. (2023) again in 2023, constructed multi-level pre-processing model frameworks using NASA's Soil Moisture Active Passive (SMAP)-satellite datasets and apply it to multi-step (one and seven days ahead) daily forecasting of Surface Soil Moisture (SSM) in Iran's dry and semi-arid regions. In this experiment, Boruta Gradient Boosting Decision Tree (Boruta-GBDT) feature selection and Multivariate Variational Mode Decomposition (MVMD) techniques are integrated with advanced Machine Learning (ML) models, that are Bidirectional Gated Recurrent Unit (Bi-GRU), Cascaded Forward Neural Network (CFNN), Adaptive Boosting (AdaBoost), Genetic Programming (GP), and classical Multilayer Perceptron neural network (MLP). According to the results, MVMD-Boruta-GBDT-CFNN outperformed over all other hybrid models in one and seven days ahead soil moisture forecasting in all tested sites. The study done by Basak et al. (2023) is also an another recent examples for data driven approach to forecast soil moisture and in this study, two data-driven models based on Naive Accumulative Representation (NAR) and the Additive Exponential Accumulative Representation (AEAR) are developed and tested.

Among data intelligence approaches, Deep Learning (DL), which is the latest generation of artificial intelligence systems is now becoming a popular category and is employed with great performance in industrial and scientific research (Emmert-Streib et al., 2020). The superiority of DL techniques in learning complex nonlinear functions of input data with low-level information allows them to successfully capture and extract the detailed features of big row input data sets, accumulated over decades, that are easily available for research initiatives. The LSTM algorithm is one of the DL artificial intelligence approaches which is being utilized to forecast hydrological and other variables like water quality (Zhang et al., 2018a), solar radiation (Ghimire et al., 2019a), rainfall-runoff (Gauch et al., 2021), and streamflow (Fu et al., 2020), and some studies has been conducted to recognize the feasibility of using LSTM-based model in predicting *SM*. In south Louisiana in the United States, ElSaadani et al. (2021) investigated that, among the spatial-temporal models tested, the ConvLSTM outperformed other Convolutional Neural Network (CNN) and LSTM-based models in *SM* prediction. To improve the soil moisture prediction accuracy, Li et al. (2022) experimented with unique residual learning encoder-decoder model (EDT-LSTM). This trial utilized data from 13 sites spread across in different countries, and the model demonstrated improved accuracy in 1,3,5,7 and 10 days ahead forecasting of moisture levels in 5 cm deep surface soil layers. Suebsombut et al. (2021) has developed Long-Short Term Memory (LSTM)-based models to forecast *SM* values in Chiang Mai province, Thailand and its results shown that, LSTM-based model performs well in predicting soil moisture. Another recent study conducted by Zeynoddin and Bonakdari (2022) proposed two DL methods which are Genetic and Teacher–Learner-based Algorithms (GA and TLA) coupled with LSTM for *SM* forecasting in Quebec, Canada and results shown that TLA-LSTM found to be more computational-effective and therefore the better option than GA-LSTM.

To further enhance forecasting model capabilities, many researchers have been developing hybrid models in the recent past. It is common for researchers to combine data pre-processing techniques with forecasting models when designing hybridised models. The pre-processing methods work well with nonlinear and nonstationary time series data. In artificial intelligence model hybridization, feature selection is a popular data pre-processing method and a variety of research studies have shown that it enhances the model's performance. The purpose of this process is to reduce the high dimensionality of input data by screening out the most corelated input data sets to the target variable data set as a first step in advanced data-driven model development (Jayasinghe et al., 2022). For example, Iterative Input Selection (IIS) to forecast streamflow (Prasad et al., 2017), Boruta-random forest (Boruta) to forecast evapotranspiration (Jayasinghe et al., 2021), soil moisture(Ahmed et al., 2021a), and streamflow (Ahmed et al., 2021c), and Neighbourhood Component Analysis (NCA) to forecast pan evaporation (Jayasinghe et al., 2022), and soil moisture (Ahmed et al., 2021b) are used as feature selection techniques in developing hydrological prediction models. The research published by Jamei et al. (2023), that explained above in detail also has employed Boruta-GBDT feature selection technique. The Lasso feature selection method which is used in this study also has been employed in hydrological forecasting studies. For instance, Alizadeh et al. (2020) in their study to develop Support Vector Regression (SVR) based model for monthly stream flow prediction at the Karaj River in Iran, Lasso and Particle Swarm Optimization-Artificial Neural Networks (PSO-ANN) feature selection methods are used to select mostly corelated input variables to the target variable. The results indicated that Lasso input selection is more accurate over the PSO-ANN algorithm and therefore improve the accuracy of model forecast. Chu et al. (2020) has also employed Lasso feature selection technique along with Fuzzy C-means (FCM) classification and Deep Belief Networks (DBN) deep learning model (Lasso-FCM-DBN) to forecast streamflow at gauge stations in the Tennessee River catchment, USA and found that Lasso-FCM-DBN approach enhance the performance of streamflow prediction compare to ANN. However, this feature selection technique has not so far been employed with any deep learning approach in soil moisture forecasting model development.

Along with feature selection, wavelet decomposition is a common data pre-processing step in data intelligence model hybridization. Because of periodicities, transients, and trends, hydrological and water resources time series data are complex. This complex data can be decomposed into sub-time series data by using wavelet transform algorithms, which are more interpretable for data-driven models. As a result, wavelet decomposed data often improve model performance and are therefore widely used in hydrological and water resources-related prediction applications. Jamei et al. (2022)'s study explained above in detail is a recent research example that employed wavelet decomposition as a data-pre-processing technique. EWT has been employed to perform wavelet decomposition in this experiment. The wavelet decomposition methods widely used in recent model hybridization works are Discrete Wavelet Transformation (DWT), Maximum Overlap Discrete Wavelet Transform with Multi Resolution Analysis (moDWT-MRA), Maximum Overlap Discrete Wavelet Transform (moDWT), and *á trous* (AT) algorithm (Quilty and Adamowski, 2018). For instance, Prasad et al. (2017) employed moDWT in their hybrid IIS-moDWT-ANN model designed for forecasting streamflow and it has shown better accuracy than the counterpart single and hybrid benchmark models. Adib et al. (2021), in their study for predicting one-day-ahead snow depth (SD) in the North Fork Jocko snow telemetry (SNOTEL) station situated in the city of Missoula, Montana State of the United States, tested different wavelet transform (WT) approaches including discrete wavelet transform (DWT), maximal overlap discrete wavelet transform (MODWT), and multiresolution-based MODWT (MODWT-MRA) along with autoregressive integrated moving average (ARIMA), and artificial intelligence (AI) models. In comparison to standalone ARIMA and AI models, hybrid ARIMA-AI models were found to produce more accurate results showing the wavelet technique's capacity to enhance the model performances.

It is important to note that DWT and moDWT-MRA can add errors to the forecast due to boundary condition-related issues and can provide better results than realistically achievable in the actual world. Therefore, they cannot be used in real-world situations. By using moDWT and AT wavelet transform algorithms with correct practices, boundary condition related issues can be resolved (Quilty and Adamowski, 2018). These boundary condition issues, their impact to the model forecast and remedies to overcome them will be discussed later in detail under the theoretical overview section of this paper. However, many recent hybrid forecasting model development studies, including the above examples that employed wavelet transform techniques to decompose hydrological and water resources related data, have not adequately considered above constraints, and instead have used DWT and moDWT-MRA regardless of their shortcomings. Furthermore, moDWT and AT wavelet transform algorithms, which do not add errors to model forecasts due to boundary condition issues, are not much used in hydrological predation as DWT and moDWT-MRA, so they still need to be explored.

In this study, time series data from satellites and ground stations are combined. The methodology section provides detailed information about the types of data collected and their resolutions and sources. Data of satellite sensor variables can lower the accuracy of hydrological variable predictions (Nikolopoulos et al., 2013, Yong et al., 2012) and this issue can be minimized by integrating ground-based and satellite-based data together, as this study does. Ghimire et al. (2018) have used data from Goddard's Online Interactive Visualization and Analysis Infrastructure (GIOVANNI) combined with reanalysis data from the European Centre for Medium Range Weather Forecasting (ECMWF) to forecast long-term solar radiation. Additionally, Ahmed et al. (2021b) used a combination of satellite GLDAS data, ground Scientific Information for Landowners (SILO) data, and meteorological indices to predict soil moisture.

Due to the high dimensionality of hydrological time series extracted in large volumes, the data for this study require feature selection and wavelet decomposition data pre-processing techniques. Thus, this hybridizing excise used moDWT and Lasso algorithms for wavelet decomposition and feature selection, respectively, along with LSTM data-driven DL network. This is a novel experience as no evidence found in literature explaining use of lasso feature selection and moDWT data decomposition techniques in *SM* prediction works. Further, this study has taken remedies to overcome boundary condition related issues which are adding errors to the forecasts in real world situations. That is also a forwarding step in prediction studies that uses wavelet transform data decomposition procedures. Furter, this proposed combination of algorithms that abbreviated as moDWT-Lasso-LSTM model has not yet been tested in another geographic location and thus fills a gap in soil moisture prediction research.

The objectives in this study are threefold:

(1) To develop deep learning methods for forecasting soil moisture (SM) at 10 cm depth, integrating moDWT data decomposition methods with Lasso methods as feature selection procedures to produce a prediction model based on LSTM utilizing satellite data from GIOVANNI and ground data from SILO.
(2) To employ the hybrid moDWT-Lasso-LSTM model in multi-step *SM* forecasting, *i.e.*, 1 day (*t+1*), 14 days (*t+14*) and 30 days (*t+30*) ahead *SM* forecasting.
(3) To compare the objective model with benchmark models: LSTM, DNN, and ANN (standalone models), Lasso-LSTM, Lasso-DNN, and Lasso-ANN (2- phase hybrid models) and moDWT-Lasso-DNN and, moDWT-Lasso-ANN (3-phase hybrid models).

Above objectives have been established in this study to design a precise *SM* forecasting model for short-, medium- and long-term *SM* predictions and to confirm its comparative advantage. *SM* as a major form of water resource exists on the earth is influencing the agricultural production and consequently affecting food security. Like the few other forms of water resources available in the globe, *SM* is also a limited resource and having growing demand due to expansion of agricultural production. Under *SM* depleted conditions, demand for water from water storages for irrigation purposes is increased while restricting water for other purposes like drinking and recreational activities. Presently on average, agriculture is accountable for 70 percent of total worldwide freshwater withdrawals (Bank, 2020). Precise *SM* predictions will be very helpful in early identification of moisture stress to the crops and actual irrigation water requirements in advance. Furthermore, precise *SM* predictions will be helpful in minimizing water wastage in irrigation activities, early notifying of crop production fluctuations and at last conserving valuable water resources. Considering above benefits of having precise *SM* forecasting tool, this study sets its primary objective to design *SM* forecasting model using LSTM deep learning algorithm with Lasso feature selection and moDWT wavelet transform data decomposition algorithm. Further, this study aims to employ this proposed model in 1 day (*t+1*), 14 days (*t+14*) and 30 days (*t+30*) multi step *SM* forecasting scenarios. This will give an opportunity to observe its usefulness in short-, medium- and long-term forecasting time horizons. Wide range of forecasting time horizons are important in implementing remedial actions against *SM* stress conditions in different levels. For instance, short term *SM* predictions may be important in taking prompt actions against potential sudden crop failures due to moisture stress while long term *SM* predictions may require in making strategic plans to cope with future drought conditions, conserving water resources and ensuring stable crop production in long run. In addition, by comparing the proposed model with competitive rival models, this study aims to recognize the performance improvement without overestimating the proposed model capabilities. The research objectives in this study will make way forward in further improvement of precision of *SM* prediction and thereby adding valuable contribution to the *SM* prediction studies.

2. Theoretical overview

This section describes the moDWT, Lasso, and LSTM algorithms used in the current study to build up the model. This study used ANN and DNN as benchmark models for assessing the target model's performance, which are relatively very recent machine learning models with neural networks like LSTM. These benchmark models are intentionally selected as they are advanced and therefore best possible competitive rivals to the data driven forecasting algorithm used in this study for the proposed model, i.e., LSTM. Use of such newer and advanced benchmark models for comparison purpose is very important for evaluating the proposed model performance without overestimation and overconfidence.

The theoretical foundation of the single neural layer ANN machine learning model is described in earlier research publications by Deo et al. (2018), Deo and Şahin (2017). In the discipline of hydrology, ANN is an extensively utilized algorithm and previous studies revealed its competency in prediction tasks. Prasad et al. (2018), for instance, developed an ANN-CoM based multi-model ensemble committee machine learning strategy to forecast monthly soil moisture at four farming locations in Murray-Darling Basin, Australia. Volterra, Random Forest, M5 tree, and ELM models are used for ANN-CoM model validation. Compared to the other models, the ANN-CoM model has shown high competency in capturing the nonlinear dynamics of soil moisture level. Shirsath and Singh (2010) constructed ANN and Multiple Linear Regression (MLR) models, as well as Penman, Priestley-Taylor and Stephens and

Stewart models for pan evaporation estimation and the estimation results were statistically compared with observed pan evaporation. The comparison reveals that the ANN model outperforms other models. Ghimire et al. (2019b) has described the theoretical background and mathematical formulae of DNN algorithm in detail in a previous study. DNN algorithm is a further improvement of ANN which is also progressively used by researchers in the field of hydrology. It consists of multiple neural layer network architecture and categorized under DL subset of machine learning family. El Bilali et al. (2023) built up an interpretable based ML framework to forecast daily pan evaporation utilizing hourly climate datasets and used DNN along with Extra Tree, XGBoost, SVR models in their exercise. Interpretability of models in predicting daily pan evaporation has been evaluated by employing the Shapely Additive explanations (SHAP), Sobol-based sensitivity analysis, and Local Interpretable Model-agnostic Explanations (LIME). The results shown good consistency of the ML model performances with the real hydro-climatic process of evaporation in a semi-arid environment. Sezen et al. (2019) has employed DNN, ANN, combined conceptual model and regression tree (RT) data driven models to model daily rainfall-runoff in karst Ljubljanica catchment and its sub-catchments in Slovenia with different geological attributes. The results of the study demonstrated that combined conceptual model yielded better modelling performance. Furthermore, Jayasinghe et al. (2022) and Ghimire et al. (2021) have used DNN as a benchmark model for evaluating respective target models in their research works to forecast evaporation and streamflow respectively.

2.1 Decomposition method: Maximum Overlap Discrete Wavelet Transform (moDWT)

Maximum Overlap Discrete Wavelet Transform (moDWT) decomposition method decompose complex time series data with multiple periodicities, transients, and trends into high and low frequency sub time series which is termed as wavelet and scaling coefficients. Those wavelets and scaling coefficients resulted from moDWT are defined as follows (Quilty and Adamowski, 2018):

$$W_{j,t} = \sum_{l=0}^{L-1} h_l \, X_{j-1,t-2^{j-1}l \bmod N} \tag{1}$$

$$V_{j,t} = \sum_{l=0}^{L-1} g_l \, X_{j-1,t-2^{j-1}l \bmod N} \tag{2}$$

,where X is a time series input vector with N values; $j = 1,2,\dots,J$, and J represents the level of decomposition at the time t; the j^{th} level wavelet $\left(W_{j,t}\right)$ and scaling $\left(V_{j,t}\right)$ filters of moDWT are represented as h_l and g_l, respectively, and L is the j^{th} level filters' width.

The moDWT can overcome some issues that can be seen related with other data decomposition algorithms such as DWT and moDWT multi resolution analysis (moDWT-MRA). In some situations, when decomposing data using various wavelet transforms, output values of decomposition process (Coefficients) cannot be calculated correctly (without adding errors) due to unavailability of time series observations needed in the calculation relative to a particular time point considered. The sources accountable for adding such errors termed as boundary conditions. For instance, DWT and moDWT-MRA which has been adapted in earlier research works has a boundary condition that arise due to its need for future data at a particular time point considered in calculating its ultimate decomposed output values termed as detail and approximation coefficients. When historical time series data is used, future data is available relative to a particular data point considered which detail and approximation coefficients are to be calculated. However, future data is not accessible in real world scenario for correctly calculating the detail and approximation components and therefore models developed using DWT and moDWT-MRA process will be unable to do accurate forecast of *SM* in practical implementations. The moDWT is a good remedy to address this boundary condition issue related with future data in calculating detailed and approximation coefficients in real world scenario connected with DWT and moDWT-MRA leading to produce inaccurate forecast in real world situations. The moDWT only uses the current time data of time series observations related to the considering data point along with the past time series data and not involve with future data when calculating its decomposition outputs: wavelet and scaling coefficients (Quilty and Adamowski, 2018). However, moDWT process cannot correctly calculate its decomposition outputs, i.e., wavelet and scaling coefficients for the data points at the beginning of time series data set as this calculation process need past time series data relative to the particular data point considered. As past time series data are not available for data points at the beginning of the data set, all wavelet and scaling coefficients calculated for early data points are incorrect and termed as boundary condition affected wavelet and scaling coefficients. The number of incorrect or boundary condition affected wavelet and scaling coefficients is dependent on decomposition level and wavelet filter used in this process. The total number of incorrect wavelet and scaling coefficients can be calculated using the $Eq.$(3) (Quilty and Adamowski, 2018) and according to this equation high decomposition levels and wavelet filters with higher lengths tend to increase the total number of incorrect wavelet and scaling coefficients.

$$L_J = (2^J - 1)(L - 1) + 1 \tag{3}$$

, where L_J represents the number of wavelet and scaling coefficients affected by the boundary condition for decomposition level J and a wavelet filter of length L.

In order to improve model forecasting accuracy, it is necessary to remove all these boundary condition affected incorrect wavelet and scaling coefficients that are derived at the beginning of the data set. High decomposition levels and lengthy wavelet filters will result in more incorrect wavelet and scaling coefficients that must be removed from the data set, resulting in an inadequate number of correct wavelet and scaling coefficients for model training. In order to further improve the model performance, appropriate selection of decomposition levels and wavelet filters is essential. Quilty and Adamowski (2018), Percival and Walden (2000) describe future data issues in detail. The optimal decomposition level and wavelet filter type cannot be found by using any thumb rule. The number of boundary conditions affected (incorrect) wavelet and scaling coefficients should not be increased unnecessarily, as it leaves inadequate correct wavelet and scaling coefficients to run the model. However, the $Eq.$(4) can be used for calculating the maximum decomposition level (J) that can be adapted (Al-Musaylh et al., 2020, Ghimire et al., 2019b):

$$J = int(log_2 N) \tag{4}$$

2.2 Feature selection method: Least Absolute Shrinkage and Selection Operator (Lasso)

In this study, the Lasso algorithm (Tibshirani, 1996) is employed as a feature selection technique after decomposition of input time series variables by the moDWT algorithm. Suppose that the dataset consists of p input variables and N observations. Let $X = [x_1, x_2, \dots, x_p] \in \mathbb{R}^{N\times p}$ is the input data matrix, in which each column denotes an input variable and $Y = [y_1, y_2, \dots, y_N]^T \in \mathbb{R}^N$ is the response variable where the response value at observation j is represented by y_j and x_j is a vector containing p characteristics. Lasso resolves (Karevan and Suykens, 2016),

$$\hat{\beta} = argmin\|y - X^T\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{5}$$

$$L_1 = \lambda \sum_{j=1}^{p} |\beta_j| \tag{6}$$

By applying a L_1- penalty for the regression coefficients, the Lasso technique degrades least-squares by shrinking the regression coefficients ($\hat{\beta}$) to zero. The variables are chosen to be included in the model during this feature selection procedure if their coefficients after the shrinking step are still non-zero. This process minimizes the prediction error by reducing the complexity of the model.

2.3 Data driven forecasting model: Long Short-Term Memory network (LSTM)

The LSTM is a unique type of Recurrent Neural Network (RNN) (Cho et al., 2014) in connection with traditional artificial neural networks that can recognize intrinsic characteristics of time sequence predictors and targets, considering the recurrent patterns and tendencies throughout long stretches of time (Manaswi, 2018). Input, output, and forget gates are the main components of the special units, or memory blocks, that the LSTMs use to operate and these memory blocks regulate the flow of information and are continuously updated (Chen et al., 2018). The 4 steps calculations are described as follows (Zhang et al., 2018b):

I. The forget gate f_t is used by the LSTM layer to determine which data should either be discarded or retained depending on the most recent hidden layer output h_{t-1}, and the new input x_t:

$$f_t = \sigma\left(w_f[h_{t-1}, x_t] + b_f\right) \quad (7)$$

, where w_f stands for weight matrix; b_f stands for bias vector and $\sigma(\ldots)$ stands for sigmoid logistic function.

II. After information is updated by utilising a “input gate” i_t , the LSTM layer determines which signal must be kept in the newly formed cell state c_t, that is denoted as the new candidate cell state $\bar{C}_t$:

$$\bar{C}_t = tanh(w_C[h_{t-1}, x_t] + b_C) \quad (8)$$

$$i_t = \sigma(w_i[h_{t-1}, x_t] + b_i) \quad (9)$$

, where hyperbolic tangent function is denoted by $tanh(\ldots)$

III. The “forget gate” f_t removes unwanted information from the old cell state C_{t-1} to C_t and the “input gate” i_t obtains a new candidate cell state $\bar{C}_t$:

$$C_t = f_t * C_{t-1} + i_t * \bar{C}_t \quad (10)$$

IV. The cell state C_t and the "output gate" o_t are then used to calculate the output h_t:

$$o_t = \sigma(w_o[h_{t-1}, x_t] + b_o) \quad (11)$$

$$h_t = o_t * tanh(C_t) \quad (12)$$

3. Methodology

3.1. Study region

Bundaberg region-$152.32^{\circ}E, 24.91^{\circ}S$ *is* the targeted site in this study for which the proposed model is designed. This land extends 6,444 square km located in Wide-Bay Burnett region of Queensland state, Australia. Bundaberg region has subtropical climate with warm wet summers and mild winters. In this region, the average annual temperature and rainfall is around 20°C and 774mm respectively and the majority of the rain falls in the summer. Average daily maximum temperature during the hot summer which prevails from November to March is above 28°C. January is the warmest month of the year in Bundaberg and the average maximum and minimum temperatures during this month are 30°C and 23°C respectively. The average minimum and maximum temperatures during July which is the coldest month of the year in Bundaberg are 14°C and 21°C respectively. The seasonal fluctuation of Bundaberg monthly rainfall is significant and receiving its highest rain fall during February with an average of 120 mm. September is reported to be the month that Bundaberg receives lowest rainfall in the year, and it is 28 mm in average. The perceived humidity varies greatly in this region while experiencing mild seasonal variation in the average hourly wind speed throughout the year (Spark, 2023, Government, 2023). According to Bundaberg Regional Council population statistics estimates, this area's total resident population has reached up to 100,118 in year 2021 with a population density of 15.54 persons per square km. The total worthiness of agricultural, forestry and fishing sector in this region is considered to be approximately $1.2 billion. This region is regarded as the food bowl capital in Australia representing 12% of Queensland's total agriculture production. Due to this region's fertile soils, favourable climate and steady water supply, well diversified agricultural operations are carried out and wide range of crops are grown. For instance, this region contributes to produce 50 per cent of Australia's macadamia production and it represents the largest proportion of country's macadamias production. This region is also leading in terms of avocado production in Australia becoming the region that allocate largest land extent for avocado farming in Australia. Further, this region's contribution for mandarin, sweet potato, passionfruit and pasture production are highly significant (Bundaberg-Regional-Council, 2023, bundaberg-agtech-hub, 2023, Growers, 2023). Above information confirms that, Bundaberg region is providing very welcoming platform for the agricultural industries while allowing this sector to be dominant in the region. Therefore, evolving a precise forecasting model to predict the soil moisture for 1,14, and 30 days ahead is strategically important in early identification of water deficit and surplus conditions affecting crop production in the region. Further, it will be helpful in employing precision irrigation practices in the region which consequently contributing to preserve the valuable water resources for future and other water demanding activities. Thus, Bundaberg region is selected for this study which aims to develop a deep learning artificial intelligence model to forecast soil moisture.

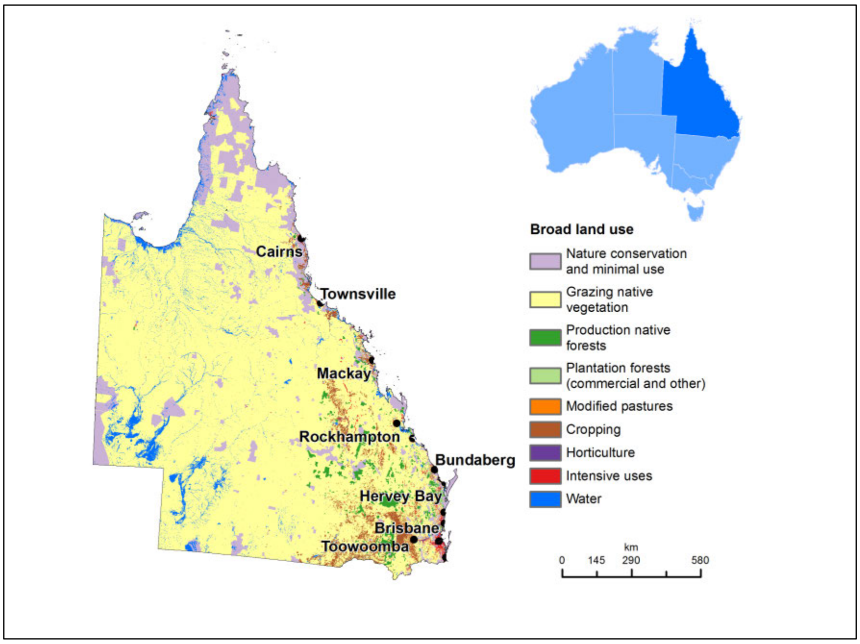


Figure 1. Study site geographical location and land use of the region and surrounding areas (pinterest, 2023)

3.2. Data collection

To conduct this research, satellite and ground based daily climatic data of 15 predictive and target variables from January 1, 2005, to December 31, 2020 are collected for the selected study site. This whole time period consists total of 5844 data points. Satellite-based data including data for target variable, i.e., Soil Moisture (*SM*) (0-10cm depth) are collected from two data platforms of Goddard Online Interactive Visualization and Analysis Infrastructure (GIOVANNI) namely, Global Land Data Assimilation System (GLDAS) and Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS) with 0.01

degree spatial resolution. Giovanni is a web interface that facilitate various users to analyse gridded data captured from various satellite and surface observations by National Aeronautics and Space Administration (NASA), United State of America. The GIOVANNI offers simple access to examine and analyse a massive amount of remote sensing data relevant to Earth Science (Teng et al., 2014). The ground-based data used for this study is collected from the Scientific Information for Landowners (SILO) database for the same time frame. The Queensland Government handles the operation of this database (Morshed et al., 2013). A list of the data sources and predictor variables used in this study, together with their corresponding acronyms, are shown in Table 1.

Table 1. Satellite-based Goddard Online Interactive Visualization and Analysis Infrastructure (GIOVANNI) Global Land Data Assimilation System (GLDAS) spectrometer satellite and Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS) spectrometer with Scientific Information for Land Owners (SILO) ground-based predictor variables used to develop the proposed hybrid moDWT-Lasso-LSTM model and other benchmark models

Data Source		**Name of Predictor Variable**	**Acronym**	**Unit**
GIOVANNI-Satellite data	**FLDAS Model**	Soil Temperature (0-10cm depth)	ST0-10	*K*
		Soil Temperature (10-40cm depth)	ST10-40	*K*
		Soil Temperature (40-100cm depth)	ST40-100	*K*
		Soil Moisture (10-40cm depth)	SM10-40	*kgm-2*
		Soil Moisture (40-100cm depth)	SM10-40	*kgm-2*
		Soil Moisture (100-200cm depth)	SM10-40	*kgm-2*
	GLDAS Model	Ground Water Storage	GWS	*mm*
SILO-Ground based data		Maximum Temperature	max-temp	℃
		Minimum Temperature	min-temp	℃
		Solar radiation	radiation	*MJm-2*
		Relative humidity at the time of maximum temperature	rh-tmax	%
		Relative humidity at the time of minimum temperature	rh-tmin	%
		Mean sea level pressure	mslp	*hPa*
		Rainfall	rain	*mm*
		Reference Evapotranspiration	ET	*mm*

Above 15 predictive variables are selected based on results of correlation matrix and trial runs excluding and including predictor variables having different levels of correlation with the target variable. Those trials, shown that the predictor variables having a weaker correlation with the target variable reduces the forecasting accuracy of all models tested. Therefore, to improve the forecasting accuracy of the models, predictor variables that shown high or reasonable correlation with the target variable are selected. In that sense other layer soil moisture data (SM10-40, SM10-40, SM10-40) which shown good correlation with target layer (SM 0-10) soil moisture data are considered.

3.3. Computers and software used in the study

The proposed multi-stage moDWT-Lasso-LSTM model and all other benchmark models are developed using a computer configured with an Intel Core i7 @ 3.3 GHz processor and 16 GB of memory loaded with freely downloadable deep learning libraries of Keras (Ketkar, 2017) and TensorFlow (Abadi et al., 2016) in Python. The moDWT data decomposing algorithm and Lasso feature selection algorithm are run on MATLAB R2019b and Python respectively. Further, ‘‘matplotlib’’ along with ‘‘seaborn’’ tools in Python are utilised for producing graphical illustrations to visualize the result in depth.

3.4. Data lagging

Row data of all 15 predictor variables are time lagged against row data of target variable, i.e., *SM* accordance with forecasting lead times *t+1*, *t+14*, and *t+30* respectively. In case of lagging data for t+1 *SM* forecasting, all data are stacked in a way that, data of predictor variables at each time point in predictor data sequence are always coinciding with I day ahead data of target variable. Similarly, in case of t+14 and t+30 *SM* forecasting, data of predictor variables at each time point are always coinciding with 14 and 30 days ahead target variable data respectively.

3.5. Data decomposition using moDWT for developing three phase hybrid models

This research adapted moDWT as the data decomposing algorithm to decompose lagged data of predictor variables in the case of developing three-phase hybrid models named as moDWT-Lasso-ANN, moDWT-Lasso-DNN and moDWT-Lasso-LSTM. However, data of target variable are not decomposed using moDWT as it is not providing additive reconstruction function (Quilty and Adamowski, 2018, Percival and Walden, 2000).

Considering there are no rules to determine the optimal decomposition level and wavelet filter type for a decomposition process, trial and error procedures are used in this study, as is common practice in similar studies. However, the *Eq*. (4) which is discussed in theoretical overview section of this paper is used for calculating the maximum decomposition level and in this research, it gives value 9. According to the *Eq*. (3) such higher decomposition level increases the number of incorrect wavelet and scaling coefficients, and it further increases when such higher decomposition level combines with wavelet filters with longer wavelength. Therefore, in this study three different decomposition levels i.e., 2, 4 and 6 which are below this highest decomposition level are selected for trial-and-error process. In this study, commonly used 7 wavelet filters having different wavelet lengths belonging to three different wavelet families are used and they are as follows: Haar (wavelet length equal to 2), db2, db4 and db6 (wavelet lengths equal to 4, 8 and 12 respectively) belonging to Daubechies family and fk4, fk8 and fk14 (wavelet lengths equal to 4, 8 and 14 respectively) belonging to Fejer-Korovkin family. Thus, 21 trails are carried out to find out the best combination of decomposition level and wavelet filter for each 3-phase hybrid model in a particular lead time. As three lead times (*t+1*, *t+14* & *t+30*) and three 3-phase hybrid models (moDWT-Lasso-ANN, moDWT-Lasso-DNN and moDWT-Lasso-LSTM) are tested in this study, a total of 189 trials are conducted to find out the best combinations of decomposition level and wavelet filter relevant to each scenario.

Therefore, although many wavelet filters belonging to many wavelet families are available, current study limits to use only above 7 wavelet filters for trials due to time constraints and to avoid complexity of the study. The lengthiest wavelet filter considered is fk14 and wavelet filters with higher lengths than that are not considered as they tend to increase the number of boundary condition affected wavelet and scaling coefficients. When this filter is used with decomposition level six, i.e., combination of heights decomposition level and highest wavelet filter length in this study scenario, according to the *Eq.* (3), the number of boundary condition affected, or incorrect wavelet and scaling coefficients will equal to 820. Although this value is differed with different combinations of wavelet filter and decomposition levels, 820 wavelet and scaling coefficients (that is the maximum possible number of boundary condition affected wavelet and scaling coefficients) are removed to ensure that all trials that are distinguished each other due to different wavelet filter, decomposition level and forecasting model combinations get the same data set. Similarly, 820 data points are removed from the beginning of the data sets that are used for standalone and 2-phase hybrid models.

3.6. Feature selection process using Lasso for developing 2 phase and 3 phase hybrid models

Feature selection is carried out using Lasso feature selection algorithm to find the mostly correlated predictor variables to the target variable for the case of developing 2-phase hybrid models, i.e., Lasso-ANN, Lasso-DNN and Lasso-LSTM. For this purpose, undecomposed predictor variable data is used for each lead time scenario separately and only the undecomposed data of selected predictor variables are chosen to feed the 2-phase hybrid forecasting models. Further, the Lasso feature selection algorithm is employed to find the mostly corelated wavelet and scaling coefficient data series derived from original predictor variable data series in the data decomposition process carried out using moDWT for development of 3-phase hybrid models, i.e., moDWT-Lasso-ANN, moDWT-Lasso-DNN and moDWT-Lasso-LSTM. This task is performed for each model and lead time scenarios separately.

3.7. Data normalization

In this study, the data ranges for each predictor variable in the data sets prepared for forecasting models vary across all model scenarios. Thus, variables with larger data ranges can be unnecessarily favoured in model forecasting over inputs with narrow ranges regardless of their intrinsic relationship. Before the data driven models are fed with data, data normalization is carried out using Eq (13) to scale the data within 0-1 range. In data normalization, the training and testing data partitions of a particular model scenario is taken together as training model parameters will not be able to generalize the unseen data if they are done separately.

$$X_n = \frac{X_{actual} - X_{min}}{X_{max} - X_{min}} \quad (13)$$

, where X_{actual}, X_{max}, and X_{min} denotes the input data for actual, maximum, and minimum values respectively.

3.8. Hyperparameter optimization

To construct best forecasting model designs, *Hyperopt* hyperparameter optimization algorithm which is available in the Python *Hyperopt* library (Bergstra et al., 2015, Komer et al., 2019) is used to identify the target and all other benchmark model's hyperparameters for each lead time forecast separately and training data partitions are used in this process. In comparison to *Grid search* and *Random search*, the *Hyperopt* hyperparameter optimization technique performs better since it can speed up the model training process while improving model accuracy (Putatunda and Rama, 2018). The list of hyperparameters and their search space used in hyperparameter optimization processes are given in Table 2 while optimal hyperparameters which are identified through the hyperparameter optimization process for designing the target LSTM and all other benchmark model architectures are given in Table 3.

Table 2. List of hyperparameters and their search space used in hyperparameter optimization process Note: ReLU and Adam stand for the Rectified Linear Units and Adaptive Moment Estimation respectively.

Model	Name of Model Hyperparameters	Search Space for Optimal Hyperparameters
LSTM	LSTM Layer 1	[50, 70, 100, 150]
	LSTM Layer 2	[50, 70, 100, 150]
	LSTM Layer 3	[50, 70, 100, 150]
	Dense Layer	[1]
	Epochs	[100, 200, 500]
	Activation Function	[ReLU]
	Optimizer	[Adam]
	Dropout Ratio	[0.1, 0.2]
	Batch Size	[5,10,20,30]
DNN	Hidden neuron 1	[10, 20, 30]
	Hidden neuron 2	[10, 15, 25]
	Hidden neuron 3	[5, 10, 20]
	Dense Layer	[1]
	Epochs	[30, 50, 100, 200]
	Activation Function	[ReLU]
	Optimizer	[Adam]
	Dropout Ratio	[0.1, 0.2, 0.3,0.4,0.5]
	Batch Size	[3, 5, 10]
ANN	Hidden neuron	[10, 20, 30]
	Dense Layer	[1]
	Epochs	[30, 50, 100,300,1000,2000]
	Activation Function	[sigmoid, tanh, ReLU]
	Optimizer	[Adam]
	Dropout Ratio	[0.3, 0.4, 0.5]
	Batch Size	[3,5,10]

Table 3. List of optimal hyperparameters selected by hyperparameter optimization process for LSTM, DNN and ANN models designing at *t+1*, *t+14* and *t+30* lead times.

Lead Time (Days)	Model	Layer 1			Layer 2			Layer 3			Batch Size	Epochs
		No. of Neurons	Activation Function	Dropout ratio	No. of Neurons	Activation Function	Dropout ratio	No. of Neurons	Activation Function	Dropout ratio		
t+1	MoDWT-Lasso-LSTM	50	ReLU	0.1	150	ReLU	0.1	50	ReLU	0.1	20	500
	MoDWT-Lasso-DNN	20	ReLU	0.3	10	ReLU	0.1	5	ReLU	0.1	10	100
	MoDWT-Lasso-ANN	20	ReLU	0.3							10	100
	Lasso-LSTM	50	ReLU	0.1	150	ReLU	0.1	50	ReLU	0.1	20	500
	Lasso-DNN	20	ReLU	0.3	10	ReLU	0.1	5	ReLU	0.1	10	100
	Lasso-ANN	20	ReLU	0.3							10	100
	LSTM	50	ReLU	0.1	150	ReLU	0.1	50	ReLU	0.1	30	500
	DNN	20	ReLU	0.3	10	ReLU	0.1	5	ReLU	0.1	10	100
	ANN	20	ReLU	0.3							10	100
t+14	MoDWT-Lasso-LSTM	100	ReLU	0.3	150	ReLU	0.2	100	ReLU	0.1	10	500
	MoDWT-Lasso-DNN	20	ReLU	0.3	10	ReLU	0.1				5	200
	MoDWT-Lasso-ANN	20	ReLU	0.3							10	100
	Lasso-LSTM	100	ReLU	0.3	150	ReLU	0.1	50	ReLU	0.1	30	500
	Lasso-DNN	20	ReLU	0.4	10	ReLU	0.1	5	ReLU	0.1	10	100
	Lasso-ANN	20	ReLU	0.3							10	100
	LSTM	50	ReLU	0.1	100	ReLU	0.2	50	ReLU	0.1	10	200
	DNN	20	ReLU	0.3	10	ReLU	0.1	5	ReLU	0.1	10	50
	ANN	30	ReLU	0.3							10	100
t+30	MoDWT-Lasso-LSTM	50	ReLU	0.2	100	ReLU	0.2	50	ReLU	0.1	10	200
	MoDWT-Lasso-DNN	30	ReLU	0.5	20	ReLU	0.2	10	ReLU	0.1	5	300
	MoDWT-Lasso-ANN	20	ReLU	0.3							10	100
	Lasso-LSTM	50	ReLU	0.2	100	ReLU	0.2	50	ReLU	0.1	10	200
	Lasso-DNN	20	ReLU	0.3	10	ReLU	0.1	5	ReLU	0.1	5	300
	Lasso-ANN	10	ReLU	0.2							10	100
	LSTM	50	ReLU	0.2	100	ReLU	0.2	50	ReLU	0.1	10	200
	DNN	10	ReLU	0.5	25	ReLU	0.1	5	ReLU	0.3	3	300
	ANN	20	ReLU	0.3							10	100

3.9. Data partitioning and data feeding to models

In this study for all model scenarios, first 75 % of respective data set is allocated for training purpose while the rest, 25 % is allocated for testing purpose and that allows both training and testing data partitions get adequate data for successful model running. Although total of 5844 data points are initially considered, due to the above explained data pre-processing works (data lagging, data decomposition and data removal) number of data points finally utilized at model running stages for each lead time scenario is reduced. So that in case of *t+1* lead time *SM* forecasting, all models are fed with 5023 data points while in cases of *t+14* and *t+30* lead time *SM* forecasting 5010 and 4994 data points are fed to the forecasting models respectively. As the first 75 % of total data set is used for the training purpose in all cases, number of data points used in training phase in *t+1*, *t+14* and *t+30* forecasting scenarios are 3767, 3757 and 3745 respectively. So that, 1256, 1253 and 1249 data points (last 25 % of the entire data set) are left for testing phase in *t+1*, *t+14* and *t+30* forecasting scenarios respectively. For instance, in *t+1* lead time case, daily data points from 01/04/2007 to 23/07/2017 are used for training purpose while, daily data points from 24/07/2017 to 30/12/2020 are used for testing purpose.

Original undecomposed lagged data of predictor variables and data of target variable are used to training and testing the standalone models, i.e., ANN, DNN and LSTM for each lead times. In case of developing 2 phase hybrid models, i.e., Lasso-ANN, Lasso-DNN and Lasso-LSTM, lagged data of predictor variables selected by Lasso feature selection algorithm along with target variable data are used. Lagged decomposed data of predictor variables selected by Lasso feature selection algorithm along with undecomposed target variable data are used for developing 3-phase hybrid models, i.e., moDWT-Lasso-ANN, moDWT-Lasso-DNN and moDWT-Lasso-LSTM. In the training phase of all model development cases, the model can see both input and output variable data. During the testing phase, however, the model can see only the input variable data and has no access the target variable data in the forecasting process. As the testing phase time point range is also historical with respect to the current time, realistically, future data of target variable with respect to all testing phase time points are available. For setting up a situation exactly similar to the real-world application of the model, target variable data are not made available for the forecasting process and instead let the model to forecast values for the target variable for each lead time with respect to each testing phase time point using the respective historical data of input variables using the skills developed in the training phase. Forecasted values of target variable are then compared with real future values of target variable available for all testing phase time points and evaluated the accuracy using statistical and graphical tools. Figure 2 illustrates the schematic view of the all model development process including the 3-phase hybrid moDWT-Lasso-LSTM model for multi-step *SM* forecasting at *t+1*, *t+14* and *t+30* lead times.

3.10 Performance evaluation

When developing machine learning models, evaluating the model performance is crucial. It determines whether a model is suitable for certain applications, compares it with rival models, and identifies areas for improvement (Pearce and Ferrier, 2000). As a result, for SM forecasting at selected sites for the same datasets, the proposed moDWT-Lasso-LSTM model and other benchmark models are evaluated considering forecasting accuracy and errors.

(i) ***Pearson's Correlation Coefficient*** (r)

The following equation ($Eq.$14) is used to derive the value of r,which expresses how closely forecasted (SM^{FOR}) and observed (SM^{OBS}) values are coincided (Moriasi et al., 2007).

The values given for this metric are always floating in between -1 to +1 and it equals +1 when perfectly strong and positive correlation exist between two variables (such as the forecasted and observed *SM*). In contrast, perfectly strong and negative correlations exist between two variables gives value of -1. The value *r* will be equal to zero if there is no relation between any two variables. However, in this instance, there should be a high and positive correlation between the estimated values by the forecasting model and observed values to consider the forecasting model to be competent enough in prediction works, thus *r* value should close or equal to +1 (Van Vuren, 2020).

$$r = \frac{\sum_{i=1}^{N}\left(SM^{OBS,i} - \overline{SM^{OBS}}\right)\left(SM^{FOR,i} - \overline{SM^{FOR}}\right)}{\sqrt{\sum_{i=1}^{N}\left(SM^{OBS,i} - \overline{SM^{OBS}}\right)^2}\sqrt{\sum_{i=1}^{N}\left(SM^{FOR,i} - \overline{SM^{FOR}}\right)^2}}, -1 \leq r \leq 1 \quad (14)$$

(ii) ***Determination of Coefficient*** (R^2)

The determination of coefficient (R^2) can be explained as the proportion of the variance in the dependent variable that is predicted by the independent variables (Chicco et al., 2021). it ranges between -∞ and +1. +1 is considered as the best value.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(SM^{FOR,i} - SM^{OBS,i}\right)^2}{\sum_{i=1}^{N}\left(\overline{SM^{OBS}} - SM^{OBS,i}\right)^2}, -\infty \leq r \leq 1 \quad (15)$$

(iii) ***Root Mean Square Error*** ($RMSE; kgm^{-2}$)

Regression model performances are typically evaluated using the *RMSE* (*Eq*.16). This metric computes the average of prediction error generated by forecasting models, that is the average difference among the forecasted value (SM^{FOR}) and the observed value (SM^{OBS}) (Willmott and Matsuura, 2005). The value of *RMSE* can be anywhere between 0 and ∞, but as model performance increases, the value of *RMSE* is shifting towards zero.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(SM^{FOR,i} - SM^{OBS,i}\right)^2}, 0 \leq RMSE < +\infty \quad (16)$$

(iv) ***Mean Absolute Error*** ($MAE; kgm^{-2}$)

The *MAE* (*Eq*.17) is measuring the actual forecasting errors in relation to the total number of observations (Prasad et al., 2019b); *MAE* value is fluctuating between 0 and ∞, however for ideal predictive models, it becomes zero. As the value given for *MAE* is unaffected by extreme outliers it provide more reliable estimation of the model's average error relative to the *RMSE* (Legates and McCabe Jr, 1999).

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|\left(SM^{FOR,i} - SM^{OBS,i}\right)\right| \quad ,0 \leq MAE < +\infty \tag{17}$$

(v) ***Mean Absolute Scaled Error*** *(MASE)*

The *MASE* (*Eq.*18) proposed by Hyndman and Koehler (2006) is also can be used as a measurement of forecast accuracy and major advantage of this statistical tool is that, the result is independent of the scale of the data. This measures the accuracy of a forecasting model in terms of the in-sample *MAE* value generated by one period a head naïve forecast method. When the forecasting model performance is better than the average one-step, naïve forecast computed in sample, the value for *MASE* will be less than 1 and contrarywise, it is greater than 1 if the forecast is inferior than the in-sample average one-step, naïve forecast (Hyndman, 2006)

$$MASE = \frac{1}{N}\left(\frac{\sum_{i=1}^{N}\left|SM^{FOR,i} - SM^{OBS,i}\right|}{\frac{1}{N-m}\sum_{i=m+1}^{N}\left|SM^{OBS,i} - SM^{OBS,i-m}\right|}\right) \tag{18}$$

(vi) ***Symmetric Mean Absolute Percentage Error (SMAPE)***

The *SMAPE* was first proposed by Armstrong and Forecasting (1985) and it is a modification of Mean Absolute Percentage Error (*MAPE*) to avoid the issue of being infinite or undefined due to zeros in the denominator (Makridakis et al., 2008). Like *MASE*, *SMAPE* is also a scale-independent metrics and thus ideal for comparing performances of forecasting algorithms (Hyndman and Koehler, 2006). Smaller percentage values indicate high levels of accuracy in the forecasting models.

$$SMAPE = \frac{200}{N}\sum_{i=1}^{N}\frac{\left|SM^{FOR,i} - SM^{OBS,i}\right|}{\left(\left|SM^{FOR,i}\right| + \left|SM^{OBS,i}\right|\right)}\,\% \tag{19}$$

(vii) ***Willmott's Index*** (WI)

This index (*Eq.*20) is applicable to a variety model performances issues since it is relatively flexible and more logically measures the model precision than other existing indices (Willmott et al., 2012). This value is ranging from 0 to 1, although the optimum predictive models give value of 1 for this metric.

$$WI = 1 - \left[\frac{\sum_{i=1}^{N}\left(SM^{OBS,i} - SM^{FOR,i}\right)^2}{\sum_{i=1}^{N}\left(\left|\left(SM^{FOR,i} - \overline{SM^{FOR}}\right)\right| + \left|\left(SM^{OBS,i} - \overline{SM^{OBS}}\right)\right|\right)^2}\right], 0 \leq WI \leq 1 \quad (20)$$

(viii) ***Nash-Sutcliffe Index*** (NS)

The value of *NS* (*Eq.*21) (Nash and Sutcliffe, 1970) shows how closely the depicted line between the predicted values and observed values fits within 1:1 ratio. If the predicted data from the model and observed data match exactly, the *NS* will be equal to 1. While -*Inf* < *NS* < 0, implies that the model is not a better predictor than the observed mean, the *NS* = 1, implies that the model estimations match the observed data's mean in terms of accuracy (AgriMetSoft, 2019).

$$NS = 1 - \left[\frac{\sum_{i=1}^{N}\left(SM^{OBS,i} - SM^{FOR,i}\right)^2}{\sum_{i=1}^{N}\left(SM^{OBS,i} - \overline{SM^{OBS}}\right)^2}\right], -\infty < NS \leq 1 \quad (21)$$

(ix) ***Legate and McCabe Index* (*LM*)**

The *LM* value (*Eq.*22) is more advanced evaluation metric compared to *WI* and *NS* values. When assessing the quality of a hydrologic or hydroclimatic model's fit, this index is more helpful than correlation based metrics like WI, Coefficient of Determination (R^2), and *NS* (Legates and McCabe, 1999). Optimal predictive models will give value of one for *LM*, while it ranges between -∞ and 1.

$$LM = 1 - \left[\frac{\sum_{i=1}^{N}\left|SM^{FOR,i} - SM^{OBS,i}\right|}{\sum_{i=1}^{N}\left(\left|\left(SM^{FOR,} - \overline{SM^{FOR}}\right)\right| + \left|\left(SM^{FOR} - \overline{SM^{OBS}}\right)\right|\right)^2}\right], -\infty < LM \leq 1 \quad (22)$$

In Equations (14-22), SM^{OBS} is daily observed soil moisture (0-10cm depth) and SM^{FOR} is daily forecasted soil moisture (0-10 cm depth), $\overline{SM^{OBS}}$ and $\overline{SM^{FOR}}$ are the mean of the values of SM^{OBS} and SM^{FOR} respectively, i is the time of the occurrence, and *N* denotes the overall quantity of data points used in the testing phase.

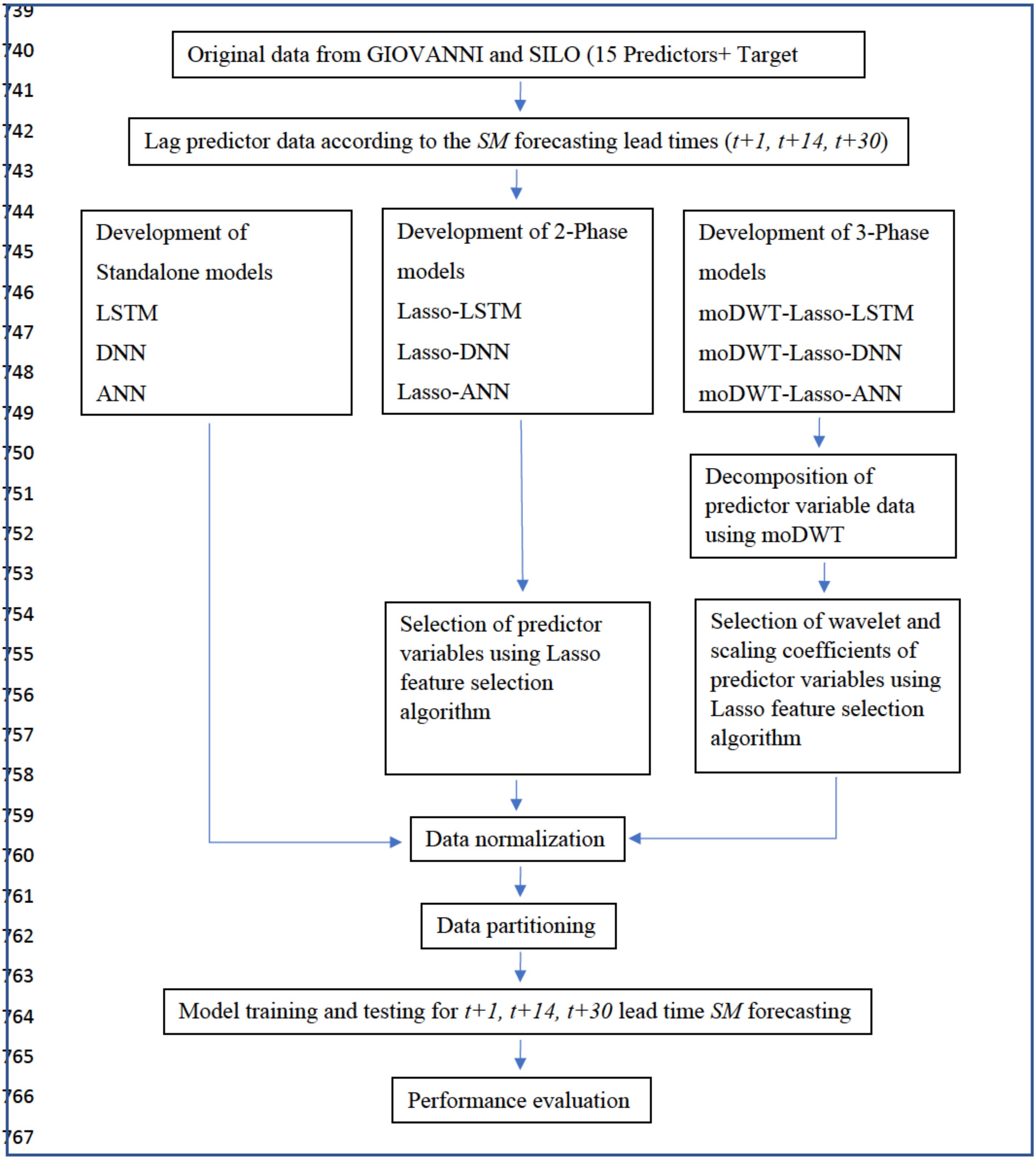


Figure 2. Schematic view of the development of benchmark models and proposed 3-phase hybrid moDWT-Lasso-LSTM model for multi-step *SM* forecasting at *t+1, t+14* and *t+30* lead times.

4. Results and discussion

The summary of the descriptive statistics values of all predictor and target variable data is given in Table 4 and Goos and Meintrup (2015), Brown Breslin et al. (2020) discuss the calculations and interpretations of those descriptive statistics in detail. Descriptive statistics provide information about central tendency (mean, median) and variability (standard deviation) of the data set and shape and the frequency of data distribution. The mean and median values of data sets of many of the variables used in this study are almost coinciding to each other indicating that data values of each dataset are very symmetrically distributed. However, the difference between mean and median values of rh-tmin, SM100-200 and SM40-100 data sets are slightly higher compared to that of other data sets reflecting slight skewness in their data distribution. ET data set is having the lowest standard deviation value indicating that, its data values are more clustered around the mean and is having narrowest range of data dispersion. SM40-100 is having the highest standard deviation value indicating the widest range of data dispersion among all variables considered in this study. In addition to SM40-100, data values of SM100-200, GWS, rh-tmax, rh-tmin, rain and SM10-40 data sets are also spread in a relatively wider range compared to the other variables. Skewness of the data sets of this study is also calculated to interpret the row data distribution. If the skewness value is less than -0.5, the distribution is said to be left-skewed or negatively skewed, with the data points concentrating on the right side and the tail being longer on the left. If the skewness value is more than 0.5, the distribution is considered as positively skewed or right skewed with data points cluster on the left side of the distribution and the tail is longer on the right side. If the skewness value is between -0.5 and 0.5, data distribution is considered to be roughly symmetric and normally distributed. Based on above criteria, the data set of rain is exceptionally right-skewed or positively skewed and data are more clustered around the left tail while right side tail of the distribution is longer. The data set of SM100-200 is showing very slightly right skewed distribution. The data set of rh-tmin is left-skewed or negatively skewed where data points cluster on the right side and the tail is longer on the left side of the distribution. However, the skewness values of other variables indicates that their data sets are more symmetrical and normally distributed. To further understand the row data distribution, Kurtosis of input and target variable data sets also calculated. The Kurtosis value will be close to three for the symmetric and normal data distributions. Such distributions are referred to as mesokurtic distributions. In circumstances, such the Kurtosis value is lower than three, the data distribution is termed as Platykurtic distribution. In such distributions, less data points will be located along the tail with low presence of extreme values relative to the normal distribution. If the Kurtosis value is greater than 3, data distribution is referred as Leptokurtic data distribution. In such situations, data distribution contains more extreme values at the tails. Among the data sets used in this study, Rain and the rh-tmin data sets scored Kurtosis values of 132.1285 and 5.6520 respectively and higher than value 3 indicating that those data sets having more outliers than data sets of other variables. Further, according to the above criteria used for interpreting data sets using Kurtosis, all other data sets can be recognized as data distributions with less outliers. Depending on descriptive statistics discussed above, many data sets used in this study can be identified as data sets closer to the normal and symmetrical distributions.

Table 4. The summary of the descriptive statistics values of all predictors and target variable data

Variable	Mean	Median	Standard Deviation	Skewness	Kurtosis
SM	22.6082	21.5707	3.8999	0.3591	-1.2382
max-temp	27.3613	27.7000	3.5232	-0.3439	-0.1983
min-temp	16.6750	17.3000	4.7370	-0.4436	-0.5403
radiation	18.5636	18.7000	5.9020	-0.2629	-0.6063
rh-tmax	50.4473	50.6000	11.8674	-0.0096	1.1172
rh-tmin	91.5993	96.0000	11.5819	-2.1140	5.6520
ET	3.9792	3.9000	1.3571	0.1223	-0.8611
mslp	1017.5165	1017.7000	4.9808	-0.2450	-0.1818
ST40-100	300.7142	301.8777	4.6184	-0.3690	-1.3350
ST10-40	300.8481	302.3685	5.2255	-0.3879	-1.2737
ST0-10	300.8051	302.5193	5.8458	-0.3975	-1.2068
rain	2.6492	0.0000	10.9918	9.3525	132.1285
SM10-40	80.7710	78.4010	10.3539	0.2873	-1.3730
SM100-200	253.8943	248.9959	19.3150	0.5340	-0.9635
SM40-100	150.0634	145.4409	21.7223	0.3255	-1.3311
GWS	939.5838	939.7291	14.6149	0.0610	-0.3491

Table 5 summarizes the results of trial-and-error to find the best combination of decomposition level and wavelet filter for 3-phase hybrid model development. In most cases, best suited combinations differ from each other except in moDWT-Lasso-LSTM and moDWT-Lasso-DNN at t+1. The best model forecasts are obtained using decomposition level 4 and wavelet filter "haar" for those two cases.

Table 5. Summary of best decomposition levels and wavelet filters resulted from trial-and-error process for 3-phase hybrid models at t*+1, t+14* and *t+30* lead times.

Model	*t+1*		*t+14*		*t+30*	
	Decomposition level	**Wavelet filter**	**Decomposition level**	**Wavelet filter**	**Decomposition level**	**Wavelet filter**
moDWT-Lasso-LSTM	4	haar	4	fk4	2	fk4
moDWT-Lasso-DNN	4	haar	4	db4	4	haar
moDWT-Lasso-ANN	2	haar	4	db6	4	db4

When decomposition level 4 is employed, time series data of each predictor variables are decomposed in to 4 wavelet coefficient data series and 1 scaling coefficient data series regardless of which wavelet filter is combined. Figure 3 graphically illustrates the decomposition results given for the predictor variable: SM10-40 when decomposition level 4 and wavelet filter "haar" are used. (i.e., decomposition level and wavelet filter combination which confirms the best model performances in moDWT-Lasso-LSTM and moDWT-Lasso-DNN at t+1 lead time). The total number of predictor variables increased up to 75 (=15 (no. original predictor variables) $\times$ 4 (no. of wavelet coefficients) + 15 (no. original predictor variables) $\times$ 1 (no. of scaling coefficient)) when decomposition level 4 is used for data decomposition. When decomposition level 2 is used for data decomposition, total number of predictor variables is increased up to 45 (=15 $\times$ 2+ 15 $\times$ 1). However, Lasso feature selection algorithm which is employed next to identify the mostly corelated predictor variables to the target variable (*SM*) reduces the number of wavelet and scaling coefficients data series being qualified for using in the forecasting model training and testing. Different wavelet and scaling coefficients data series are selected by Lasso algorithm with respective to each decomposed data sets derived using different combinations of decomposition level and wavelet filters that used in 3-phase model development. Further, majority of coefficient data series selected by Lasso feature selection algorithm are scaling coefficients of predictor variables. Table 6 shows the summery of wavelet and scaling coefficient data series selected by Lasso feature selection algorithm with respect to all 3-phase hybrid models at *t+1* lead time. According to the summery explained in this table (Table 6), in case of predictor variable: SM10-40, second, third and fourth wavelet coefficient data series (W2, W3, W4) and scaling coefficient data series (V) depicted in Figure 3 are selected by Lasso feature selection algorithm for model development of moDWT-Lasso-LSTM and moDWT-Lasso-DNN at *t+1* lead time.

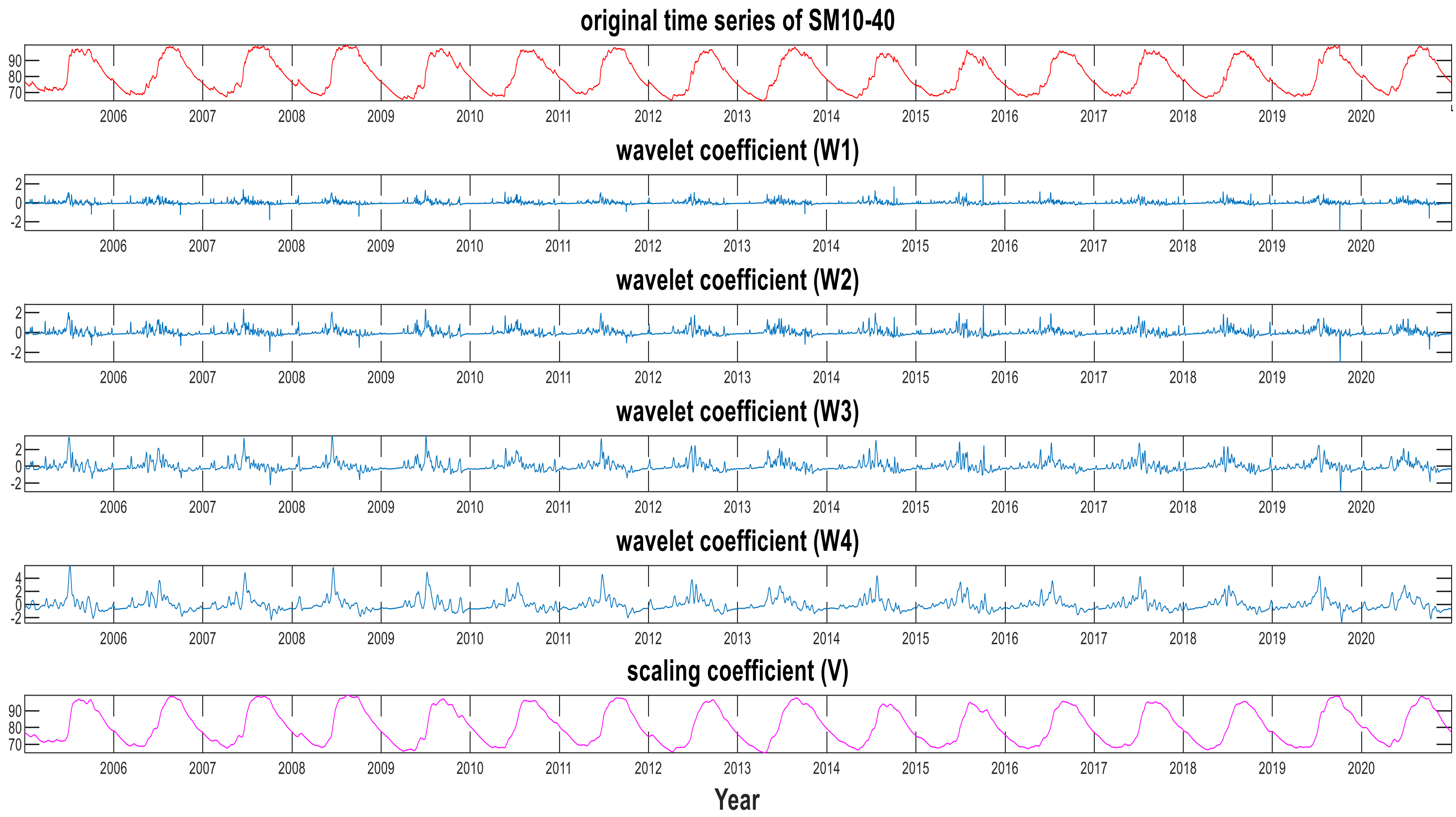


Figure 3. Wavelet and scaling data series resulted from moDWT decomposition process given for the predictor variable: SM10-40 when decomposition level 4 and wavelet filter "haar" is used at *t+1* lead time.

Table 6. Summary of selected wavelet and scaling coefficients by Lasso feature selection technique at *t+1* lead time for 3-phase hybrid model development.

Model	Predictor variables of which **wavelet coefficients** data series selected by Lasso	**Wavelet coefficients (W)** data series selected by Lasso	Predictor variables of which **scaling coefficients** (V) data series selected by Lasso	Total no. of **wavelets and scaling coefficients** data series selected
moDWT-Lasso-LSTM	SM10-40 SM100-200 GWS	W2,W3,W4 W4 W4	min-temp radiation ST0-10 rain SM10-40 SM100-200 GWS	12
moDWT-Lasso-DNN	SM10-40 SM100-200 GWS	W2,W3,W4 W4 W4	min-temp radiation ST0-10 rain SM10-40 SM100-200 GWS	12
moDWT-Lasso ANN	rh-tmin SM10-40	W2 W2	min-temp radiation rh-tmax ST0-10 rain SM10-40 SM100-200 SM40-100 GWS	11

In case of developing 2-phase hybrid models (i.e., Lasso-LSTM, Lasso-DNN and Lasso-ANN) number of predictor variables selected by Lasso feature selection algorithm for *t+1*, *t+14* and *t+30* lead times are 10 (i.e., min-temp, radiation, rh-tmax, rh-tmin, ST0-10, rain, SM10-40, SM100-200, SM40-100 and GWS), 13 (i.e., max-temp, min-temp, radiation, rh-tmax, rh-tmin, mslp, ST40-100, ST0-10, rain, SM10-40, SM40-100, SM100-200 and GWS) and 12 (i.e., max-temp, min-temp, radiation, rh-tmin, mslp, ST40-100, ST0-10, rain, SM10-40, SM100-200, SM40-100 and GWS) respectively.

Table 7. Values scored in the testing phase for statistical metrics used to **e**valuate the proposed hybrid moDWT-Lasso-LSTM and benchmark models for lead times *t+1*, *t+14* and *t+30*. The best values scored for relevant statistical metrics are boldfaced.

Model	*t+1*							
	r	***R^2***	***RMSE***	***MAE***	***MASE***	***SMAPE (%)***	***LM***	***WI***
moDWT-Lasso-LSTM	**0.97290**	**0.92469**	**0.97808**	**0.76623**	**4.39700**	**3.48910**	**0.78021**	**0.98270**
moDWT-Lasso-DNN	0.97243	0.90801	1.05142	0.83664	4.80102	4.28050	0.76069	0.97023
moDWT-Lasso ANN	0.96755	0.87927	1.25829	0.99296	5.69808	4.32120	0.71597	0.97211
Lasso-LSTM	0.96916	0.86992	1.24185	0.99203	5.69274	4.29820	0.71543	0.97145
Lasso-DNN	0.96398	0.78780	1.49764	1.22490	7.02904	5.26880	0.64963	0.95672
Lasso-ANN	0.96310	0.86976	1.30536	1.02215	5.86556	4.45690	0.70762	0.96990
LSTM	0.96728	0.89932	1.08161	0.85262	4.89270	3.76450	0.75543	0.97789
DNN	0.96628	0.66048	1.70606	1.38637	7.95562	5.81720	0.60344	0.93937
ANN	0.95478	0.81781	1.58090	1.25067	7.17693	5.51210	0.62531	0.95712
	t+14							
	r	***R^2***	***RMSE***	***MAE***	***MASE***	***SMAPE (%)***	***LM***	***WI***
moDWT-Lasso-LSTM	0.96012	**0.89224**	**1.18054**	0.96482	0.79649	**4.01170**	0.72280	**0.97491**
moDWT-Lasso-DNN	**0.96149**	0.87398	1.19683	**0.94721**	**0.78195**	4.13590	**0.72846**	0.97264
moDWT-Lasso ANN	0.95139	0.85932	1.29359	1.06966	0.88304	4.67810	0.69336	0.96854
Lasso-LSTM	0.93999	0.87467	1.34597	1.05395	0.87006	4.30280	0.69719	0.96878
Lasso-DNN	0.95380	0.88453	1.20330	0.96490	0.79655	4.73540	0.72340	0.97344
Lasso-ANN	0.95167	0.85824	1.30455	1.06954	0.88293	4.65450	0.69340	0.96818
LSTM	0.94245	0.86700	1.36678	1.05309	0.86935	4.71900	0.69744	0.96750
DNN	0.95413	0.77293	1.48918	1.18204	0.97581	5.06000	0.66115	0.95493
ANN	0.93540	0.77400	1.59018	1.28029	1.05692	5.54800	0.63298	0.95122
	t+30							
	r	***R^2***	***RMSE***	***MAE***	***MASE***	***SMAPE (%)***	***LM***	***WI***
moDWT-Lasso-LSTM	**0.96497**	**0.91564**	**1.13674**	**0.91126**	**0.45417**	**3.98600**	**0.73774**	**0.97849**
moDWT-Lasso-DNN	0.95820	0.88818	1.15259	0.95784	0.47738	4.31100	0.72481	0.97516
moDWT-Lasso ANN	0.95528	0.88467	1.22855	1.00449	0.50063	4.44910	0.71140	0.97286
Lasso-LSTM	0.95051	0.88685	1.22393	0.96703	0.48196	4.22980	0.72169	0.97307
Lasso-DNN	0.95665	0.85161	1.26631	0.99443	0.49562	4.30100	0.71429	0.96852
Lasso-ANN	0.93237	0.81717	1.46481	1.22684	0.61145	5.34670	0.64752	0.95895
LSTM	0.95436	0.87888	1.20148	0.97581	0.48634	4.32890	0.71917	0.97277
DNN	0.95139	0.77771	1.47242	1.15331	0.57480	4.91300	0.66865	0.95562
ANN	0.93926	0.83699	1.40469	1.16230	0.57928	5.09100	0.66607	0.96294

Table 7 displays the calculated values of statistical metrics: Pearson's Correlation Coefficient (*r*), Coefficient of Determination (R^2), Root Mean Squared Error (*RMSE*; kgm^{-2}), Mean Absolute Error (*MAE*; kgm^{-2}), Mean Absolute Scaled Error (*MASE*), Symmetric Mean Absolute Percentage Error (*SMAPE*), Legates and McCabe Index (*LM*) and Willmott's Index (*WI*) which are used to evaluate the performance of the target model (moDWT-Lasso-LSTM) and other benchmark models. The proposed deep moDWT-Lasso-LSTM model has produced the highest values for *r*, R^2 *LM* and *WI* while producing the lowest values for *RMSE*, *MAE*, *MASE* and *SMAPE* in comparison to that of all the benchmark models, as evidenced by the testing phase results provided in Table 7 in *t+1* and *t+30* lead time *SM* forecasting. In case of *t+14* lead time *SM* forecasting, the proposed moDWT-Lasso-LSTM model has been able to score the highest values for R^2 and *WI* and lowset value for *RMSE* and *SMAPE* while scoring the second highest values for *r* and *LM* and second lowest value for *MAE* and *MASE*. In the same lead time, moDWT-Lasso-DNN has scored highest values for *r* and *LM* and lowest value for *MAE* and *MASE* while scoring the second highest values for R^2 and *WI* and second lowest value for *RMSE and SMAPE*. i.e., moDWT-Lasso-LSTM and moDWT-Lasso-DNN has alternatively scores the best and second-best values in *t+14* lead time *SM* forecasting. Above results in general confirms that the proposed moDWT-Lasso-LSTM model outperforms the other benchmark models used in this study. Further, it has shown comparatively higher consistence in securing its position as the best model across all lead times than any other models tested. Although, moDWT-Lasso-DNN has demonstrated performances very parallel to the moDWT-Lasso-LSTM in case of *t+14* lead time, it has been unable to shown consistency as the best model across all lead times.

In case of t+1 lead time, the values given for *MASE* for all the tested models is greater than 1 indicating that accuracy of all the models including the proposed model are inferior to the in-sample average one-step, naïve forecast. However, the proposed model scored the nearest value to the value 1, i.e., 4.39700 in t+1 lead time confirming that it is the best model out of all other models tested in this study in terms of *MASE*. But in case of t+14 and t+30 lead times, values scored for *MASE* by all the tested models in this study are less than 1 indicating that, accuracy of all models are better than the in-sample average one-step naïve forecast. It is showing that, competent *SM* forecasting tool is needed to make accurate *SM* predictions in long term ahead forecasting situations. Therefore, the proposed model by current research is worthful, as it is shown more capabilities than the other benchmark models in many scenarios.

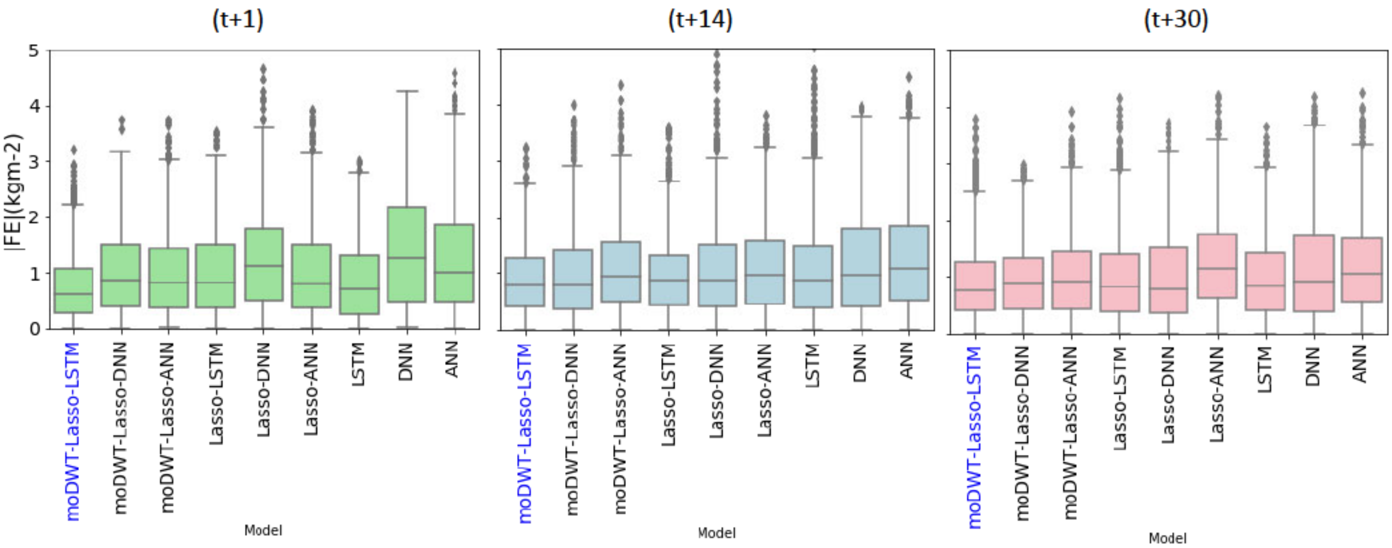


Figure 4. The boxplot of the forecasting errors in the testing phase, generated by the moDWT-Lasso-LSTM hybrid model and other benchmark models at *t+1*, *t+14* and *t+30* lead time *SM* forecasting.

To further affirm the dominance of proposed moDWT-Lasso-LSTM model in terms of prediction competency over the other benchmark models, this suggested model's and all benchmark models' absolute forecasting errors (|FE| = |observed *SM* – forecasted *SM*|; kgm^{-2}) are contrasted. The distribution of |FE| during testing phase, including the upper, median, and lower quartiles for each model for *t+1*, *t+14* and *t+30* lead time *SM* forecasting are illustrated in the boxplots in Figure 4. According to these box plots, the multi-step moDWT-Lasso-LSTM model provided the fewest quartiles for |FE| in *t+1*, *t+14* and *t+30* lead time cases. These results which show narrow error distribution in comparison to the benchmark models further indicate how well the deep multi-step moDWT-Lasso-LSTM model is suited for *SM* forecasting compared to the other benchmark models. Figure 5 shows the stem plots for Nash-Sutcliffe Coefficient (*NS*) calculated for target moDWT-Lasso-LSTM model and benchmark models in testing phase for *t+1, t+14,* and *t+30* lead times *SM* forecasting. These graphs show that the moDWT-Lasso-LSTM model exhibits the highest values of *NS* for all lead times. Further, scatter plots are drawn for the t+30 lead time for all the models tested (Figure 6). With relative to the scatter plots of other forecasting models, data points are more uniformly distributed along the whole 45-degree line with less outliers and less deviations in the scatter plot of moDWT-Lasso-LSTM model showing a strongly positive correlation between observed and forecasted *SM* values. As per the above analysis, the proposed moDWT-Lasso-LSTM model can be identified as the best *SM* forecasting model among all other models tested in this study and therefore it will be a useful tool for *SM* forecasting in Bundaberg in Australia.

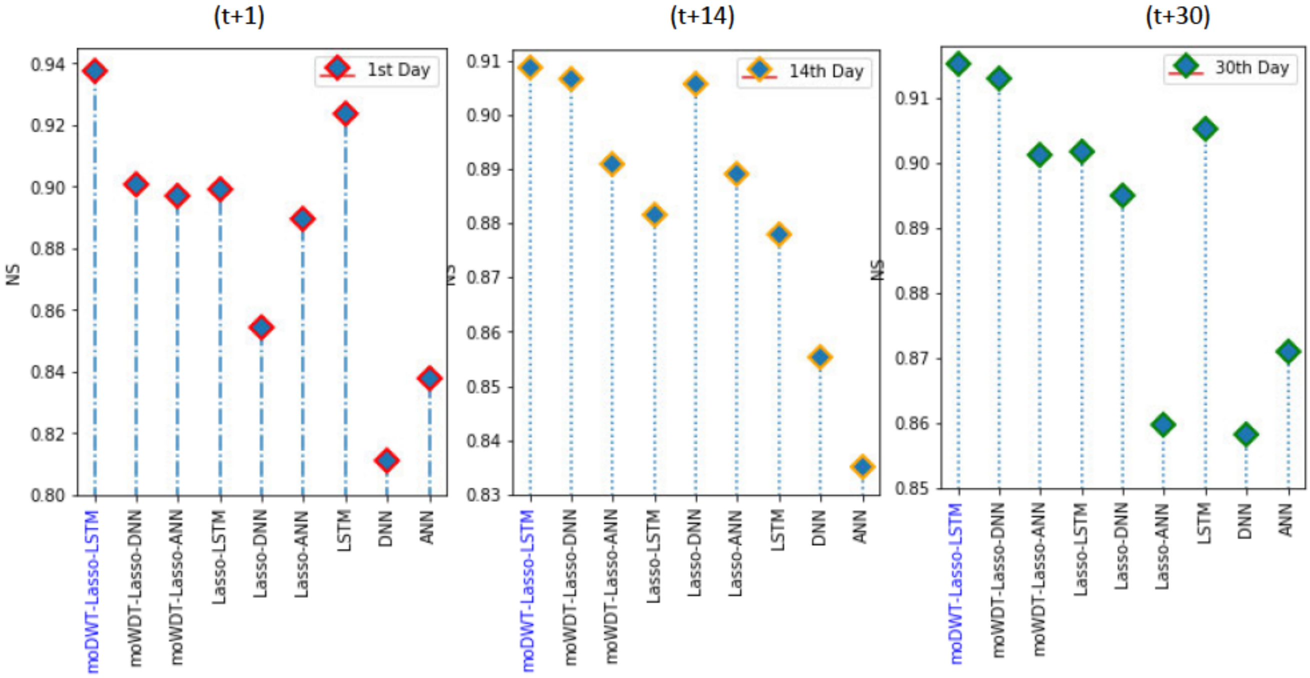


Figure 5. Stem plots of Nash-Sutcliffe Coefficient (*NS*) for hybrid moDWT-Lasso-LSTM model benchmark models in testing phase at *t+1*, *t+14* and *t+30* lead time *SM* forecasting.

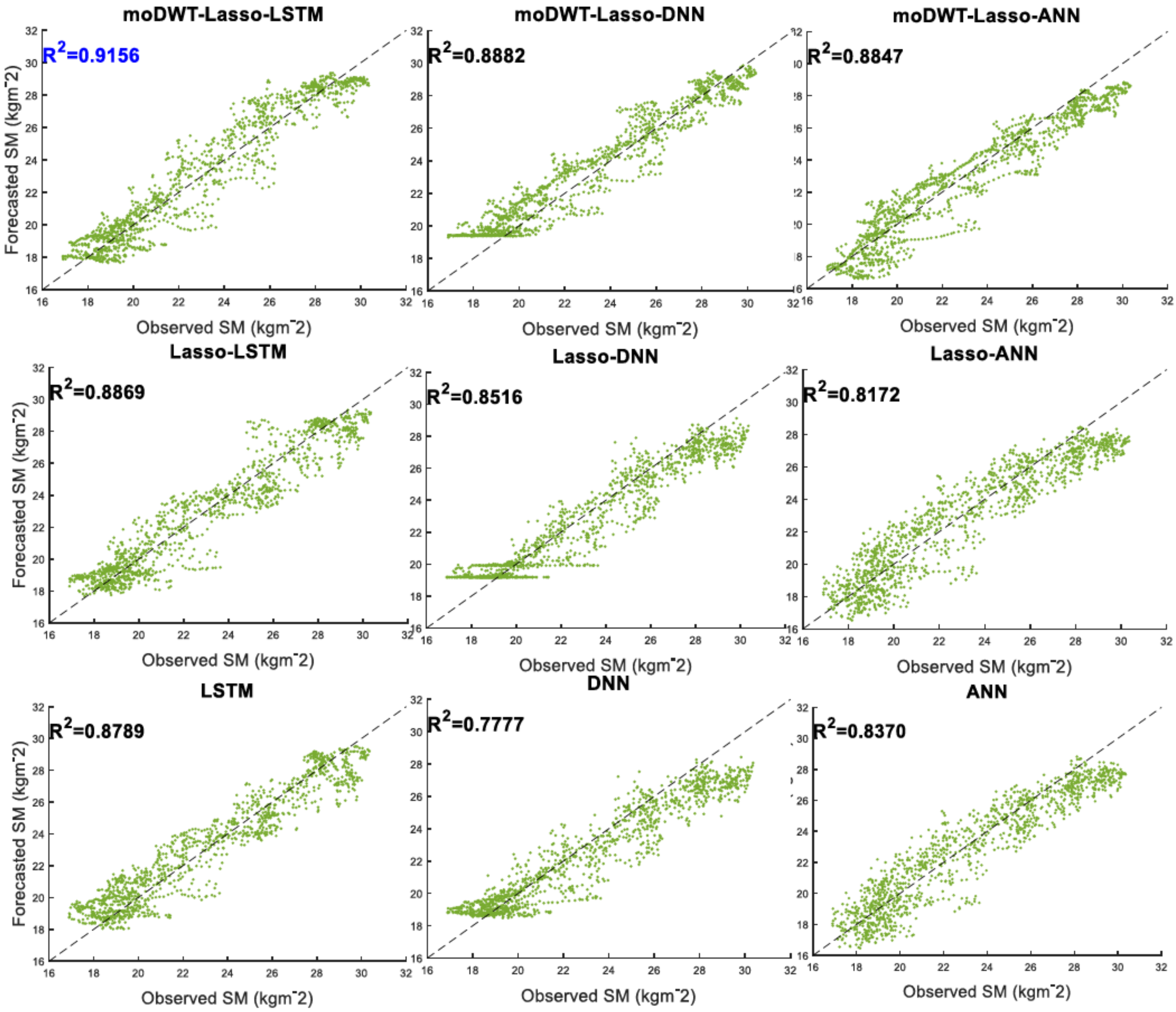


Figure 6. Scatter plots of moDWT-Lasso-LSTM model and other benchmark models in testing phase at *t+30* lead time *SM* forecasting.

5. Conclusion, limitations and, suggestions for future research

Agricultural decision-making especially needs reliable information on climatic and hydrological variables. For instance, farmers at present time very commonly use rainfall forecasts for making decisions relevant to crop establishment, crop harvesting, fertilizer application, land preparation activities etc. The proposed model is designed to forecast *SM* which is also a very important hydrological variable. This information is useful in deciding the need and the appropriate time for irrigation and deciding the accurate quantity of irrigation water needed on a particular day. If adequate moisture is available in the soil, irrigation is not necessary and can be skipped. If soil moisture is not adequate, only the deficit should be compensated via irrigation. Reliable *SM* information is very useful in such decision-making. Further, *SM* forecasts are important to be considered in the application of fertilizer. Containing adequate amount of moisture in soil is essential for dissolving nutrients in fertilizers and make them available to the plant. *SM* ease the plant to absorb important nutrients and thereby maximize the fertilizer use efficiency while reducing the wastage. Especially in areas where there is no access to irrigation water and farming is totally rely on rainfall, knowing reliable information of moisture level in soils beforehand some activities like land preparation, planting and fertilizer applications will be very useful. Further, it is forwarding step in current day context as many farmers are moving towards precision agriculture to cut down cost of production, minimize wastage and conserve resources and inputs.

Under such background, this study has designed a multi-step wavelet 3-phase hybrid deep learning soil moisture forecasting (moDWT-Lasso-LSTM) model using Lasso regression optimization and moDWT decomposition algorithms for soil moisture forecasts in Bundaberg in Queensland, Australia. The daily input data period from January 1, 2005, to December 31, 2020 were obtained from satellite data bases of NASA's (GIOVANNI)- Global Land Data Assimilation System (GLDAS) and Land Data Assimilation System (FLDAS) and ground data base of SILO. To attain an accurate model, extracted data was decomposed by moDWT and then selected features using Lasso algorithm for 1(*t+1*), 14 (*t+14*), and 30 (*t+30*) days ahead. With the incorporation of LSTM, moDWT, and Lasso, the proposed deep learning multi-step moDWT-Lasso-LSTM hybrid model was created, and its performance was evaluated using statistical score measures and contrasted with the performance of the other eight comparison models namely, moDWT-Lasso-DNN, moDWT-Lasso-ANN, Lasso-LSTM, Lasso-DNN, Lasso-ANN, LSTM, DNN, and ANN.

The proposed moDWT-Lasso-LSTM hybrid model yielded improved performance in forecasting *SM* for 1, 14 and, 30 days ahead relative to the other benchmark models and this was particularly clear for *t+1* and *t+30* lead times. But in case of *t+14* lead time, moDWT-Lasso-DNN has shown performances very parallel to that of moDWT-Lasso-LSTM according to the statistical metrics discussed in Table 7. However, when visualizing results of box plots of (|FE|) and stem plots of *NS*, the suggested moDWT-Lasso-LSTM model accomplished better performances than moDWT-Lasso-DNN and any other benchmark models in all lead times. This reaffirmed the usefulness of the suggested moDWT-Lasso-LSTM model over the other benchmark models in predicting *SM*.

However, with respect to the statistical matrix: *MASE*, all models including the proposed model is showing the weakest performances at t+1 lead time compared to t+14 and t+30 lead times. But most of models scored best values for most of the other statistical tools in t+1 lead time compared to t+14 and t+30 lead times. However, as *MASE* is considered to be more reliable tool for assessing forecasting models, what is interpreted by other statistical tools can be excluded and it is better to make the conclusion based upon *MASE* values. Although the proposed model has shown comparatively higher accuracy in t+1 lead time than other models, as discussed earlier, *MASE* value reflects that, its accuracy is still lower than naïve forecast accuracy at t+1. Naïve forecasting is doing predictions in a simple way, i.e., it uses the actual observed values from the last time step as the forecast of the next time step without considering any other factors and adjustments. In real world condition, with respect to our study, it uses today's *SM* value as the forecasted value of *SM* for tomorrow. So it implies that, it is more reliable to use today's actual *SM* value as a clue for tomorrow's (t+1 lead time) *SM* value rather than trusting on *SM* values forecasted by sophisticated models in case of practical situation that anyone needs to know one day ahead *SM* forecast. Realistically, it can be expected that, one day ahead *SM* value can be very similar to current day *SM* value as variables like *SM* may not be remarkably change in very short time unless it is influenced by any other climatic factor like sudden rain. Therefore, relying on proposed model or any other benchmark models used in this work for t+1 lead time *SM* forecasting cannot be recommended according to the current study. But in case of t+14 and t+30 *SM* forecasting, proposed model accuracy is higher than naïve forecast accuracy according to the *MASE* value and further it has shown higher performances against other benchmark models. So that, the proposed model can be practically employed in t+14 and t+30 *SM* forecasting in the selected study region. Furthermore, this research only considers 1 day, 14 days and 30 days ahead *SM* forecasting as an initial step. The number of lead times used in this study will not be enough to visualize any consistent trend of forecasting model performances across the lead times with the increase of lead time length. However, further increasing the lead time length can causes consistent and significant changes in model performances. So that, future researchers can carry out new studies to find out such trends of forecasting ability of models with extended forecasting periods (with increased lead time lengths) and to find solutions for them. Furthermore, future research can consider developing *SM* prediction tools to forecast *SM* for long time durations (For instance *SM* forecast for one month duration) and that can be more important than short duration *SM* forecasting in water resecures management strategic planning.

The input data values used in this study are continuously being recorded by data collecting institutions and they are up to date until current time and will be up to date at any time point considered in the future and therefore model has access to the required historical input data at any real time. Further, the wavelet transform data decomposition procedure followed in this study does not need future data to calculate its wavelet and scaling co-efficient which is used to feed the forecasting algorithm instead of row data. Some wavelets transform data decomposition procedures are needing future data being available to them for calculating their coefficient values and thereby making the forecasting models less useful in the real-time scenario. So that, the proposed model can be practically implemented in real-time situations as required historical input data can be accessed at any time point in future and also it is trained to forecast *SM* with zero involvement with future data.

Further, applying of moDWT data decomposition algorithm to the time series data set used in this study has generally shown a trend of increasing the data driven model performances. So that, future studies which uses lengthy time series climatic data can trial employing moDWT or any other wavelet transform methods to convert complex data patterns into simplified high and low frequency wavelet time series. Further, it is noticed that data driven model performances are varied based on selection of decomposition levels and wavelet filters that distinguished upon wavelet family and filter length. So that, it can be recommended to do trials and error procedure considering time and cost constraints to find out best suited decomposition levels and wavelet filters specific to relevant study scenarios, if this type of data decomposition algorithms is used.

Current study's new modelling strategy for 1, 14, and 30 days ahead *SM* forecasting has brought forward other several potential directions for future research with a wider focus. For instance, this proposed 3-phase hybrid moDWT-Lasso-LSTM model has developed targeting Bundaberg region in Australia and has shown a promising success to employ this in the region for *SM* forecasting. As it is not realistic to consider whole Queensland or Australia due to time and other resources constraints this study has to confine into such a region covering relatively smaller geographical area. Therefore, this methodology or similar concept can be tested in other regions in Australia or elsewhere in the world to examine the geographical consistency.

Further, this model is developed to forecast soil moisture in topsoil layer, i.e., 0-10 cm depth. The depth of the active root zone of crops mainly varied with crop type or genetics, the development stage of the crop, and soil properties. Some crops are having tap roots that penetrate deeply into the soil, while other crops develop many shallow lateral roots. Annual crops own shallow root systems, and their depth varies rapidly in a short time with their growth. Even, perennial crops in their early growth stages are having shallow roots absorbing moisture from the topmost soil layers and they are gradually penetrating to deep soil layers. Further, the crops grown in soils with unfavourable soil properties and conditions (E.g., soils with shallow bedrock or hard layers (Clay), soils compacted due to heavy use of machinery) will tend to have root systems mainly concentrated in topmost soil layers. The root system of many crops is concentrated in the top layers of the soil and near to the base of the plant and further, in most plants, the concentration of moisture-absorbing roots is usually high in the upper top quarter of the root zone. Due to these characteristics, under good soil conditions with no restrictions for moisture and nutrient absorption and no disturbances for root development, soil moisture extraction by plants typically follow a conical water uptake pattern. That is 40 % of total moisture uptake is absorbed from the first one-fourth of the total crop rooting depth, while 30%, 20%, and 10% of total moisture uptake is absorbed from the 2nd, 3rd, and 4th quarters of total crop rooting depth respectively. So, moisture extraction is most rapid in the topmost layers of the soil. Further, the evaporative water loss is also very high in the upper few inches of the soil. So, the topmost soil layer is more vulnerable to the rapid diminution of water creating a high soil water potential gradient.(Lincoln, 2023, Nebraska, 2023, Technology, 2023). Therefore, due to the above reasons, this study deliberately focused to design a *SM* forecasting model for the top-most soil layer (0-10 cm). However, future researchers can consider examining the usefulness of proposed methodology in forecasting *SM* in more deeper soil layers as it is also equally important in water resources management.

Further, another decomposition method called *á trous* (AT) algorithm (Quilty and Adamowski, 2018) (which also a good remedy for future data issue) instead of the moDWT technique coupled with Lasso or any other feature selection technique combined with LSTM can be used to create a novel three-stages deep hybrid *SM* predicting model. Moreover, future researchers can experiment the potentiality of the suggested moDWT-Lasso-LSTM model in prediction of important drought indices such as Palmer drought severity index (PDSI), standardized precipitation index (SPI), and standardized precipitation and evaporation index (SPEI).

Acknowledgement

The authors are grateful to NASA and SILO for providing free access to GIOVANNI satellite and ground-based meteorological data. The authors would like to express their gratitude to the University of Southern Queensland (UniSQ), Australia, and the Wayamba University of Sri Lanka for generously funding this research. In addition, the authors would like to thank Dr. Thong Nguyen-Huy for his useful advice on acquiring gridded data and Dr. Barbara Harmes for proofreading this paper.

References

ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., IRVING, G. & ISARD, M. {TensorFlow}: A System for {Large-Scale} Machine Learning. 12th USENIX symposium on operating systems design and implementation (OSDI 16), 2016. 265-283.

ADIB, A., ZAERPOUR, A. & LOTFIRAD, M. 2021. On the reliability of a novel MODWT-based hybrid ARIMA-artificial intelligence approach to forecast daily snow depth (Case study: the western part of the Rocky Mountains in the USA). *Cold Regions Science and Technology,* 189, 103342.

AGRIMETSOFT. 2019. *AgriMetSoft (2019). Online Calculators* [Online]. AgriMetSoft [Accessed].

AHMED, A., DEO, R. C., GHAHRAMANI, A., RAJ, N., FENG, Q., YIN, Z. & YANG, L. 2021a. LSTM integrated with Boruta-random forest optimiser for soil moisture estimation under RCP4. 5 and RCP8. 5 global warming scenarios. *Stochastic Environmental Research and Risk Assessment,* 35, 1851-1881.

AHMED, A., DEO, R. C., RAJ, N., GHAHRAMANI, A., FENG, Q., YIN, Z. & YANG, L. 2021b. Deep learning forecasts of soil moisture: convolutional neural network and gated recurrent unit models coupled with satellite-derived MODIS, observations and synoptic-scale climate index data. *Remote Sensing,* 13, 554.

AHMED, A. M., DEO, R. C., FENG, Q., GHAHRAMANI, A., RAJ, N., YIN, Z. & YANG, L. 2021c. Deep learning hybrid model with Boruta-Random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. *Journal of Hydrology,* 599, 126350.

AL-MUSAYLH, M. S., DEO, R. C. & LI, Y. 2020. Electrical energy demand forecasting model development and evaluation with maximum overlap discrete wavelet transform-online sequential extreme learning machines algorithms. *Energies,* 13, 2307.

ALIZADEH, Z., SHOURIAN, M. & YASEEN, Z. M. 2020. Simulating monthly streamflow using a hybrid feature selection approach integrated with an intelligence model. *Hydrological Sciences Journal,* 65, 1374-1384.

ARMSTRONG, J. S. & FORECASTING, L.-R. 1985. From crystal ball to computer. *New York ua,* 348.

BANK, T. W. 2020. *Water in Agriculture* [Online]. The World Bank. Available: https://www.worldbank.org/en/topic/water-in-agriculture [Accessed 8/9/2022 2022].

BASAK, A., SCHMIDT, K. M. & MENGSHOEL, O. J. 2023. From data to interpretable models: machine learning for soil moisture forecasting. *International Journal of Data Science and Analytics,* 15, 9-32.

BELAYNEH, A., ADAMOWSKI, J., KHALIL, B. & QUILTY, J. 2016. Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. *Atmospheric research,* 172, 37-47.

BERGSTRA, J., KOMER, B., ELIASMITH, C., YAMINS, D. & COX, D. D. 2015. Hyperopt: a Python library for model selection and hyperparameter optimization. *Computational science & discovery,* 8, 014008.

BROWN BRESLIN, A. M., ATKINSON, P., DELAMONT, S., CERNAT, A., SAKSHAUG, J. W. & WILLIAMS, R. A. 2020. *Descriptive statistics,* London, SAGE Publications Ltd.

BUNDABERG-AGTECH-HUB. 2023. *Our Bundaberg Region* [Online]. Bundaberg Regional Council. Available: https://www.ourbundabergregion.com.au/bundaberg-agtech-hub [Accessed 2023].

BUNDABERG-REGIONAL-COUNCIL. 2023. *Economic development* [Online]. Bundaberg Regional Council, QLD. Available: https://www.bundaberg.qld.gov.au/us/economic-development/3 [Accessed 2023].

CHANG, X., ZHAO, W. & ZENG, F. 2015. Crop evapotranspiration-based irrigation management during the growing season in the arid region of northwestern China. *Environmental Monitoring and Assessment,* 187, 1-15.

CHEN, J., ZENG, G.-Q., ZHOU, W., DU, W. & LU, K.-D. 2018. Wind speed forecasting using nonlinear-learning ensemble of deep learning time series prediction and extremal optimization. *Energy conversion and management,* 165, 681-695.

CHICCO, D., WARRENS, M. J. & JURMAN, G. 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science,* 7**,** e623.

CHO, K., VAN MERRIËNBOER, B., GULCEHRE, C., BAHDANAU, D., BOUGARES, F., SCHWENK, H. & BENGIO, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

CHU, H., WEI, J. & WU, W. 2020. Streamflow prediction using LASSO-FCM-DBN approach based on hydro-meteorological condition classification. *Journal of Hydrology,* 580**,** 124253.

DEO, R. C., GHIMIRE, S., DOWNS, N. J. & RAJ, N. 2018. Optimization of windspeed prediction using an artificial neural network compared with a genetic programming model. *Handbook of Research on Predictive Modeling and Optimization Methods in Science and Engineering.* IGI Global.

DEO, R. C. & ŞAHIN, M. 2016. An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland. *Environmental monitoring and assessment,* 188**,** 1-24.

DEO, R. C. & ŞAHIN, 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. *Renewable and Sustainable Energy Reviews,* 72**,** 828-848.

EL BILALI, A., ABDESLAM, T., AYOUB, N., LAMANE, H., EZZAOUINI, M. A. & ELBELTAGI, A. 2023. An interpretable machine learning approach based on DNN, SVR, Extra Tree, and XGBoost models for predicting daily pan evaporation. *Journal of Environmental Management,* 327**,** 116890.

ELSAADANI, M., HABIB, E., ABDELHAMEED, A. M. & BAYOUMI, M. 2021. Assessment of a Spatiotemporal Deep Learning Approach for Soil Moisture Prediction and Filling the Gaps in Between Soil Moisture Observations. *Frontiers in artificial intelligence,* 4.

EMMERT-STREIB, F., YANG, Z., FENG, H., TRIPATHI, S. & DEHMER, M. 2020. An introductory review of deep learning for prediction models with big data. *Frontiers in Artificial Intelligence,* 3**,** 4.

FU, M., FAN, T., DING, Z. A., SALIH, S. Q., AL-ANSARI, N. & YASEEN, Z. M. 2020. Deep learning data-intelligence model based on adjusted forecasting window scale: application in daily streamflow simulation. *IEEE Access,* 8**,** 32632-32651.

GAUCH, M., KRATZERT, F., KLOTZ, D., NEARING, G., LIN, J. & HOCHREITER, S. 2021. Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrology and Earth System Sciences,* 25**,** 2045-2062.

GHIMIRE, S., DEO, R. C., DOWNS, N. J. & RAJ, N. 2018. Self-adaptive differential evolutionary extreme learning machines for long-term solar radiation prediction with remotely-sensed MODIS satellite and Reanalysis atmospheric products in solar-rich cities. *Remote Sensing of Environment,* 212**,** 176-198.

GHIMIRE, S., DEO, R. C., RAJ, N. & MI, J. 2019a. Deep learning neural networks trained with MODIS satellite-derived predictors for long-term global solar radiation prediction. *Energies,* 12**,** 2407.

GHIMIRE, S., DEO, R. C., RAJ, N. & MI, J. 2019b. Wavelet-based 3-phase hybrid SVR model trained with satellite-derived predictors, particle swarm optimization and maximum overlap discrete wavelet transform for solar radiation prediction. *Renewable and Sustainable Energy Reviews,* 113**,** 109247.

GHIMIRE, S., YASEEN, Z. M., FAROOQUE, A. A., DEO, R. C., ZHANG, J. & TAO, X. 2021. Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks. *Scientific Reports,* 11**,** 17497.

GHORBANI, M., DEO, R. C., YASEEN, Z. M., H KASHANI, M. & MOHAMMADI, B. 2018. Pan evaporation prediction using a hybrid multilayer perceptron-firefly algorithm (MLP-FFA) model: case study in North Iran. *Theoretical and applied climatology,* 133**,** 1119-1131.

GOOS, P. & MEINTRUP, D. 2015. *Statistics with JMP : graphs, descriptive statistics and probability,* West Sussex, England, Wiley.

GOVERNMENT, Q. 2023. Climate change in the Wide Bay-Burnett Region. *In:* GOVERNMENT, Q. (ed.). Queensland Government.

GROWERS, B. F. V. 2023. *Bundaberg Friuts & Vegetable Growers* [Online]. bfvg. Available: https://bfvg.com.au/ [Accessed].

HUANG, S., CHANG, J., HUANG, Q. & CHEN, Y. 2014. Monthly streamflow prediction using modified EMD-based support vector machine. *Journal of Hydrology,* 511**,** 764-775.

HYNDMAN, R. J. 2006. Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting,* 4**,** 43-46.

HYNDMAN, R. J. & KOEHLER, A. B. 2006. Another look at measures of forecast accuracy. *International journal of forecasting,* 22**,** 679-688.

JAMEI, M., ALI, M., KARBASI, M., SHARMA, E., JAMEI, M., CHU, X. & YASEEN, Z. M. 2023. A high dimensional features-based cascaded forward neural network coupled with MVMD and Boruta-GBDT for multi-step ahead forecasting of surface soil moisture. *Engineering Applications of Artificial Intelligence,* 120**,** 105895.

JAMEI, M., KARBASI, M., MALIK, A., JAMEI, M., KISI, O. & YASEEN, Z. M. 2022. Long-term multi-step ahead forecasting of root zone soil moisture in different climates: Novel ensemble-based complementary data-intelligent paradigms. *Agricultural Water Management,* 269**,** 107679.

JAYASINGHE, W. L. P., DEO, R. C., GHAHRAMANI, A., GHIMIRE, S. & RAJ, N. 2021. Deep Multi-Stage Reference Evapotranspiration Forecasting Model: Multivariate Empirical Mode Decomposition Integrated With the Boruta-Random Forest Algorithm. *IEEE Access,* 9**,** 166695-166708.

JAYASINGHE, W. L. P., DEO, R. C., GHAHRAMANI, A., GHIMIRE, S. & RAJ, N. 2022. Development and evaluation of hybrid deep learning long short-term memory network model for pan evaporation estimation trained with satellite and ground-based data. *Journal of Hydrology,* 607**,** 127534.

KAREVAN, Z. & SUYKENS, J. Spatio-temporal feature selection for black-box weather forecasting. Proc. of the 24th european symposium on artificial neural networks, computational intelligence and machine learning, 2016. 611-616.

KETKAR, N. 2017. Introduction to keras. *Deep learning with Python.* Springer.

KHAN, N., SACHINDRA, D., SHAHID, S., AHMED, K., SHIRU, M. S. & NAWAZ, N. 2020. Prediction of droughts over Pakistan using machine learning algorithms. *Advances in Water Resources,* 139**,** 103562.

KOMER, B., BERGSTRA, J. & ELIASMITH, C. 2019. Hyperopt-sklearn. *Automated Machine Learning.* Springer, Cham.

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**,** 233-241.

LEGATES, D. R. & MCCABE JR, G. J. 1999. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water resources research,* 35**,** 233-241.

LI, Q., LI, Z., SHANGGUAN, W., WANG, X., LI, L. & YU, F. 2022. Improving soil moisture prediction using a novel encoder-decoder model with residual learning. *Computers and Electronics in Agriculture,* 195**,** 106816.

LIAO, R., YANG, P., WANG, Z., WU, W. & REN, S. 2018. Development of a soil water movement model for the superabsorbent polymer application. *Soil Science Society of America Journal,* 82**,** 436-446.

LINCOLN, U. O. N. 2023. *vegetable and Fruit Production* [Online]. University of Nebraska Lincoln. Available: https://extensionpubs.unl.edu/, https://extensionpublications.unl.edu/assets/pdf/g2189.pdf [Accessed].

MAKRIDAKIS, S., WHEELWRIGHT, S. C. & HYNDMAN, R. J. 2008. *Forecasting methods and applications*, John wiley & sons.

MANASWI, N. K. 2018. *Deep Learning with Applications Using Python: Chatbots and Face, Object, and Speech Recognition with TensorFlow and Keras,* Berkeley, CA, Apress L. P.

MORIASI, D. N., ARNOLD, J. G., VAN LIEW, M. W., BINGNER, R. L., HARMEL, R. D. & VEITH, T. L. 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE,* 50**,** 885-900.

MORSHED, A., ARYAL, J. & DUTTA, R. Environmental spatio-temporal ontology for the Linked open data cloud. 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, 2013. IEEE, 1907-1912.

NASH, J. E. & SUTCLIFFE, J. V. 1970. River flow forecasting through conceptual models part I—A discussion of principles. *Journal of hydrology,* 10**,** 282-290.

NEBRASKA, U. O. 2023. *Corn Soil-Water Extraction and Effective Rooting Depth in a Silt-Loam Soil* [Online]. Available: https://extensionpublications.unl.edu/assets/pdf/g2245.pdf [Accessed].

NIKOLOPOULOS, E. I., ANAGNOSTOU, E. N. & BORGA, M. 2013. Using high-resolution satellite rainfall products to simulate a major flash flood event in northern Italy. *Journal of Hydrometeorology,* 14**,** 171-185.

ORTIZ-GARCÍA, E., SALCEDO-SANZ, S. & CASANOVA-MATEO, C. 2014. Accurate precipitation prediction with support vector classifiers: A study including novel predictive variables and observational data. *Atmospheric research,* 139**,** 128-136.

PEARCE, J. & FERRIER, S. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. *Ecological modelling,* 133**,** 225-245.

PERCIVAL, D. B. & WALDEN, A. T. 2000. *Wavelet methods for time series analysis*, Cambridge university press.

PINTEREST. 2023. *Land Use Map of Queensland*.

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. 2018. 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. 2019a. Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach. *Catena,* 177**,** 149-166.

PRASAD, R., DEO, R. C., LI, Y. & MARASENI, T. 2019b. Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach. *Catena (Giessen),* 177**,** 149-166.

PUTATUNDA, S. & RAMA, K. A comparative analysis of hyperopt as against other approaches for hyper-parameter optimization of XGBoost. Proceedings of the 2018 International Conference on Signal Processing and Machine Learning, 2018. 6-10.

QUILTY, J. & ADAMOWSKI, J. 2018. Addressing the incorrect usage of wavelet-based hydrological and water resources forecasting models for real-world applications with best practices and a new forecasting framework. *Journal of hydrology,* 563**,** 336-353.

SEZEN, C., BEZAK, N., BAI, Y. & ŠRAJ, M. 2019. Hydrological modelling of karst catchment using lumped conceptual and data mining models. *Journal of Hydrology,* 576**,** 98-110.

SHIRSATH, P. B. & SINGH, A. K. 2010. A comparative study of daily pan evaporation estimation using ANN, regression and climate based models. *Water Resources Management,* 24**,** 1571-1581.

SILVERMAN, D. & DRACUP, J. A. 2000. Artificial neural networks and long-range precipitation prediction in California. *Journal of applied meteorology,* 39**,** 57-66.

SPARK, W. 2023. Weather Spark.

SUEBSOMBUT, P., SEKHARI, A., SUREEPHONG, P., BELHI, A. & BOURAS, A. 2021. Field data forecasting using LSTM and bi-LSTM approaches. *Applied Sciences,* 11**,** 11820.

TECHNOLOGY, N. C. F. A. 2023. Water Management.

TENG, W., RUI, H., VOLLMER, B., DE JEU, R., FANG, F., LEI, G.-D. & PARINUSSA, R. 2014. NASA Giovanni: A Tool for Visualizing, Analyzing, and Intercomparing Soil Moisture Data. *Remote Sensing of the Terrestrial Water Cycle,* 206**,** 331.

TIBSHIRANI, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological),* 58**,** 267-288.

VAN VUREN, T. 2020. Modeling of transport demand–analyzing, calculating, and forecasting transport demand: by VA Profillidis and GN Botzoris, Amsterdam, Elsevier, 2018, 472 pp.,

$125 (paperback and ebook), eBook ISBN: 9780128115145, Paperback ISBN: 9780128115138. Taylor & Francis.

WILLMOTT, C. J. & MATSUURA, K. 2005. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate research,* 30**,** 79-82.

WILLMOTT, C. J., ROBESON, S. M. & MATSUURA, K. 2012. A refined index of model performance. *International journal of climatology,* 32**,** 2088-2094.

YONG, B., HONG, Y., REN, L. L., GOURLEY, J. J., HUFFMAN, G. J., CHEN, X., WANG, W. & KHAN, S. I. 2012. Assessment of evolving TRMM-based multisatellite real-time precipitation estimation methods and their impacts on hydrologic prediction in a high latitude basin. *Journal of Geophysical Research: Atmospheres,* 117.

ZEYNODDIN, M. & BONAKDARI, H. 2022. Structural-optimized sequential deep learning methods for surface soil moisture forecasting, case study Quebec, Canada. *Neural Computing and Applications,* 34**,** 19895-19921.

ZHANG, J., ZHU, Y., ZHANG, X., YE, M. & YANG, J. 2018a. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. *Journal of hydrology,* 561**,** 918-929.

ZHANG, X., ZHANG, Q., ZHANG, G., NIE, Z., GUI, Z. & QUE, H. 2018b. A Novel Hybrid Data-Driven Model for Daily Land Surface Temperature Forecasting Using Long Short-Term Memory Neural Network Based on Ensemble Empirical Mode Decomposition. *International journal of environmental research and public health,* 15**,** 1032.

ZHU, B., FENG, Y., GONG, D., JIANG, S., ZHAO, L. & CUI, N. 2020. Hybrid particle swarm optimization with extreme learning machine for daily reference evapotranspiration prediction from limited climatic data. *Computers and Electronics in Agriculture,* 173**,** 105430.

Statements & Declarations

Declaration of Interests

The authors declare that research work presented in this paper is not influenced by any of their known financial interests or personal relationships.

Authorship Contribution Statement

W.J.M. Lakmini Prarthana Jayasinghe: Writing – original draft, Conceptualization, Methodology, Software, Writing – review & editing, Investigation. **Ravinesh C Deo**: Conceptualization, Writing – review & editing, Supervision. **Afshin Ghahramani**: Writing – review & editing. **Sujan Ghimire**: Conceptualization, Writing – review & editing, Investigation, Supervision. **Nawin Raj**: Writing – review & editing.

6.3 Links and implications

Soil moisture describes the water availability in soil which is essential for sustainable plant growth. It is also can be used as a useful indicator for understanding trends, chances, and magnitudes of drought conditions. Soil moisture directly affects the growth of vegetative cover in natural habitats of wildlife like forests, jungles, and bushes. The growth of vegetative cover is very important as it supplies food for wildlife's existence. In addition, it directly affects the commercial crop production. Low soil moisture levels will eventually develop drought conditions which can cause bushfire threats and adversely affect the existence of wildlife as it led to an inadequate supply of water and food. Therefore, the research work done under 3rd objective of this PhD study to design a deep learning forecasting model to predict *SM* in different soil layers in 1,14 and 30 days ahead is highly advantageous, and the proposed model will be a future useful tool in managing disaster conditions caused by water resources scarcities. However, this research was carried out as a case study extracting data only from the Bundaberg region in Queensland. Due to time constraints, it was not feasible to consider more sites that would represent a larger geographical area in Australia or elsewhere in the world. But this study has developed a fresh modelling strategy for 1, 14, and 30 days ahead *SM* forecasting and it is showing several potential directions for future research with a wider focus. For instance, future studies might be researched on the terrestrial consistency and accuracy of this proposed moDWT-Lasso-LSTM hybrid model. Further, other decomposition methods (For instance, Multivariate Variational Mode Decomposition (MVMD) algorithm (ur Rehman and Aftab, 2019)) instead of the moDWT technique coupled with Lasso or any other feature selection technique combined with LSTM can be used to create a novel three-stages deep hybrid *SM* predicting model. Moreover, future researchers can experiment with the potentiality of the suggested moDWT-Lasso-LSTM model in the prediction of other drought-related parameters such as precipitation and relative humidity and other important drought indices such as Palmer drought severity index (PDSI), standardized precipitation index (SPI), and standardized precipitation and evaporation index (SPEI).

CHAPTER 7: CONCLUSION AND FUTURE SCOPE

7.1 Synthesis

This thesis has enhanced the science of hydrological prediction by developing highly precise hybrid deep learning artificial intelligence models empowered by computational optimization methods. These have been focused on developing more precise E_p, *ET*, and *SM* forecasting hybrid DL models mostly using satellite data in Queensland, Australia. The objectives expected in this complete research are: (1) developing a deep NCA-LSTM model to predict daily E_P and performance evaluation against other benchmark models: LSTM, DNN, RF, ANN, and DT models, (2) developing three-phase multivariate sequential hybrid MEMD-Boruta-LSTM, model to forecast daily *ET* and performance evaluation against MEMD-Boruta-DNN, MEMD-Boruta-DT, and standalone LSTM, DNN and DT models, (3) developing a hybrid multi-step moDWT-Lasso-LSTM model to predict *SM* in the 0-10 cm depth and performance evaluation against eight benchmark models i.e., three standalone models (*e.g.*, LSTM), three 2-phase hybrid models with feature selection (*e.g.*, Lasso-LSTM) and two 3-phase hybrid models with feature selection and decomposition (*e.g.*, moDWT-Lasso-DNN).

The E_P provides a very accurate estimate of the height at which water is lost due to evaporation from water storage. The volume of water loss owning to evaporation which is one of the main causes of water loss from water resources can be determined by multiplying the E_P value by the surface area of water storage. Thus, the entire amount of water loss from water storage used for irrigation purposes, drinking, bathing, hydropower generation, and other recreational activities can be estimated, and appropriate water resource planning and irrigation schedules may then be established. Also, evaporation progression can quicken the drying of natural water bodies and consequently deprive the drinking water for wildlife while excessive evaporation conditions particularly in dry spells develop drought conditions and natural disasters like bushfires. Therefore, predicting E_P is a crucial factor to be considered in the current situation in the world and it is a vital solution for early water resource planning especially in arid and semi-arid regions. The first goal of this research was to develop a novel, precise deep learning hybrid LSTM model incorporating Neighbourhood Component Analysis (NCA) feature selection method to identify the most effective predictor variables as a useful tool to predict

E_p. The daily input data, which covered the period from August 31, 2002, to September 22, 2020, were extracted from GIOVANNI-AIRS satellites and trustworthy SILO data produced by the Queensland government. The model test location within the drought-prone region of Queensland, Australia were Amberley, Gatton, and Oakey, and Townsville. The target deep learning NCA-LSTM model was developed by integrating LSTM and NCA; its performance was assessed using statistical score measures and compared to that of existing benchmark models, including LSTM, DNN, RF, ANN, and DT. Comparing the NCA-LSTM hybrid model to other benchmark models, it performed better at predicting daily E_p, which was particularly noticeable for the study locations in Amberley, Gatton, and Oakey. However, the statistical metrics for the Townsville research site indicated that the proposed model performed less effectively than at the other study sites. Despite this performance gap that has a site-specific signature, the suggested NCA-LSTM hybrid model nevertheless outperformed the other benchmark models by a wide margin for this study site. This reinforced the ability of the NCA-LSTM hybrid model to forecast daily E_p effectively.

Additionally, *ET* provides an estimate of the amount of water lost by crops via transpiration and from the soil surface by evaporation. Thus, the second objective is to create a brand-new deep learning multi-stage hybrid MEMD-Boruta-LSTM model that can be used as a useful tool to anticipate daily *ET* utilizing satellite-based and ground-based data. Input data has been decomposed into IMFs using MEMD and the most correlated IMFs were screened by the Boruta feature selection algorithm by incorporating with the LSTM network. The GIOVANNI-AIRS, GLDAS model satellites, and the SILO ground daily-based input data were extracted for the Gatton, Fordsdale, and Cairns' test sites located in the drought-prone region of Queensland, Australia for the period from 01 February 2003 to 19 April 2011. The novel multistage deep learning MEMD-Boruta-LSTM hybrid model was created by integrating LSTM with MEMD and Boruta and its performance was assessed using statistical score metrics and compared to that of other hybrid and standalone models, including MEMD-Boruta-DNN, MEMD-Boruta-DT, LSTM, DNN, and DT. In terms of normalized performance measures, the target MEMD-Boruta-LSTM model produced the highest values for *r, NS, WI,* and *LM* and the lowest values for *RMSE, MAE, RRMSE,* and *APB* across all locations. All these results offered compelling proof of the proposed MEMD-Boruta-LSTM model performed better in forecasting *ET* at a daily forecasting horizon than the comparable hybrid and standalone models.

Furthermore, Soil moisture (*SM*) refers to the water availability in the soil and is crucial for sustaining plant growth. Forecasting *SM* gives the knowledge to develop adaption and management strategies to protect natural ecosystems from the threat of climate change owing to precipitation deficiencies while geoscientists and the appropriate authorities can prioritize the areas needed for water allocations. *SM* forecasting is useful in scheduling irrigation programs, drought monitoring, and early identification of bushfire and flood threats. Therefore, the third objective of this study focuses to design a precise and effective data-driven AI model for *SM* forecasting. In this study, a multi-step hybrid deep learning moDWT-Lasso-LSTM model was developed to forecast *SM* (up to 10 cm depth on topsoil) for 1, 14, and 30 days in advance. The daily satellite data from the Global Land Data Assimilation System (GLDAS) and Land Data Assimilation System (FLDAS) and ground data from SILO were extracted from January 1, 2005, to December 31, 2020, for Bundaberg region in Queensland, Australia. To create a robust and accurate model, retrieved data was decomposed by the moDWT method and then selected best features by the Lasso algorithm before incorporating it with the LSTM network. The performance of this target moDWT-Lasso-LSTM model was assessed using statistical score measures and compared to the eight comparator models separately for each 1, 14, and 30 days, including moDWT-Lasso-DNN, moDWT-Lasso-ANN, Lasso-LSTM, Lasso-DNN, Lasso-ANN, LSTM, DNN, and ANN. In comparison to the benchmark models, the target moDWT-Lasso-LSTM model produced overall better results when forecasting *SM* for 1, 14, and 30 days ahead. This confirmed the developed moDWT-Lasso-LSTM model's efficacy in forecasting *SM* for 1, 14, and 30 days in advance.

In addition to the above socioeconomic benefits anticipated, the outcome of this PhD study covered a significant research gap in science and technology as all models suggested here to forecast *Ep*, *ET*, and *SM* in Queensland are hybridized DL networks. Furthermore, most of the data utilized in this study are taken from satellite and ground sources, and there is no evidence in the literature to support the use of the methods suggested in this study to forecast *Ep*, *ET*, and *SM* in Queensland, Australia.

7.2 Novel contributions of the study

The development of hybridized deep learning models for hydrological forecasting is one of the advanced contributions made by this PhD thesis. In addition to creating novel deep hydrological predictive models, further unique methodological advancements include as follows:

7.2.1 Two-phase deep and machine learning models

One of the major contributions of this PhD study is designing two-phase models i.e., original standalone models integrated with the feature selection methods. For instance, the LSTM network was combined with the feature selection method, Neighbourhood Component Analysis (NCA) to create NCA- LSTM model for predicting evaporation that was compared with the standalone LSTM, DNN, ANN, RF, and DT. Furthermore, the Least Absolute Shrinkage and Selection Operator (Lasso) feature selection method coupled with LSTM, DNN, and ANN models to build up novel Lasso-LSTM, Lasso-ANN, and Lasso-DT models used as comparative models in soil moisture forecasting scenarios.

7.2.2 Three-phase deep and machine learning models

A major contribution of this PhD thesis is the design of three-phase hybrid models coupled with feature selection and data decomposition techniques. When designing *ET* forecasting models, three-phase deep learning hybrid models with Multivariate Empirical Mode Decomposition (MEMD) and Boruta-Random Forest (Boruta) algorithms were developed as MEMD-Boruta-LSTM, MEMD-Boruta-DNN, and MEMD-Boruta-DT. Furthermore, another three-phase model with Maximum Overlap Discrete Wavelet Transform (moDWT) decomposition and the Lasso feature selection techniques integrated with LSTM, DNN, and ANN denoted as moDWT-Lasso-LSTM, moDWT-Lasso-DNN, and moDWT-Lasso-ANN were created in multistep soil moisture forecasting scenario. The moDWT-Lasso-LSTM model was the highest performed hybridized approach over the moDWT-Lasso-DNN and moDWT-Lasso-ANN models in soil moisture forecasting for 1 day, 14 days, and 30 days in advance.

7.3 Limitations of the current study and Recommendations for future research

Although this study made foremost contributions to a PhD on research, it had some limitations and suggestions for future research, and are discussed in this section as follows:

- Only seven study sites in Queensland (used as a case study) were selected to develop models in this study. Future research can include more locations that represent the entire drought prone regions in Australia or elsewhere.

- Incorporated with Variation Mode Decomposition (VMD) or Improved Complete Empirical Ensemble Mode Decomposition with Adaptive Noise (ICEEMDAN) techniques could also improve the efficiency of the proposed models.

- Target models could also incorporate optimizer algorithms, such as the Quantum-Behaved Particle Swarm Optimization (Q-PSO) or the Firefly Optimizer Algorithm (FFA).

- Data intelligent standard statistical tool, Bayesian Model Averaging (BMA) can be used to rank the model performance and avoid the hurdle of model uncertainties that may result in overly confident inferences and risky agricultural decisions.

- The suggested models can be experimented with to predict important drought indices such as the Palmer drought severity index (PDSI), standardized precipitation index (SPI), and standardized precipitation and evaporation index (SPEI).

- Some additional feature selection algorithms like the Rule-and-Tree-based algorithm, multivariate adaptive regression spline (MARS), iterative input selection (IIS), or joint mutual information maximization feature selection (JMIM) can be further incorporated to increase the efficacy of the models.

- Dimensionality reduction algorithms can be used as a data transform pre-processing method, such as principal component analysis (PCA), non-negative matrix factorization (NNMF), and linear discriminant analysis (LDA).

In conclusion, this PhD study has contributed in a novel way to the practical issues of hydrological forecasting by combining deep learning and optimization techniques in data science. Proposed new hybridized forecasting approaches are very computationally efficient and have low latency that could be easy to use for real-world problems with having access to upgrade the models. This could enhance hydrological forecasting, acting as a key tool for applications in water resource and agricultural management.

REFERENCES

ABADI, M., BARHAM, P., CHEN, J., CHEN, Z., DAVIS, A., DEAN, J., DEVIN, M., GHEMAWAT, S., IRVING, G. & ISARD, M. {TensorFlow}: A System for {Large-Scale} Machine Learning. 12th USENIX symposium on operating systems design and implementation (OSDI 16), 2016. 265-283.

ABDULLAH, S. S., MALEK, M. A., ABDULLAH, N. S., KISI, O. & YAP, K. S. 2015. Extreme learning machines: a new approach for prediction of reference evapotranspiration. *Journal of Hydrology,* 527**,** 184-195.

ABED, M., IMTEAZ, M. A., AHMED, A. N. & HUANG, Y. F. 2021. Application of long short-term memory neural network technique for predicting monthly pan evaporation. *Scientific Reports,* 11**,** 20742.

ABED, M., IMTEAZ, M. A., AHMED, A. N. & HUANG, Y. F. 2022. Modelling monthly pan evaporation utilising Random Forest and deep learning algorithms. *Scientific Reports,* 12**,** 13132.

ADAMOWSKI, J. & CHAN, H. F. 2011. A wavelet neural network conjunction model for groundwater level forecasting. *Journal of hydrology (Amsterdam),* 407**,** 28-40.

AGENCY, A. G. N. E. M. 2022. Available: https://knowledge.aidr.org.au/resources/2018-bushfire-qld-queensland-bushfires/ [Accessed].

AL-MUSAYLH, M. S., DEO, R. C., LI, Y. & ADAMOWSKI, J. F. 2018. Two-phase particle swarm optimized-support vector regression hybrid model integrated with improved empirical mode decomposition with adaptive noise for multiple-horizon electricity demand forecasting. *Applied energy,* 217**,** 422-439.

BANK, T. W. 2020. *Water in Agriculture* [Online]. The World Bank. Available: https://www.worldbank.org/en/topic/water-in-agriculture [Accessed 8/9/2022 2022].

BASAK, A., SCHMIDT, K. M. & MENGSHOEL, O. J. 2023. From data to interpretable models: machine learning for soil moisture forecasting. *International Journal of Data Science and Analytics,* 15**,** 9-32.

BRUTSAERT, W. 2013. *Evaporation into the atmosphere: theory, history and applications*, Springer Science & Business Media.

CAI, Y., ZHENG, W., ZHANG, X., ZHANGZHONG, L. & XUE, X. 2019. Research on soil moisture prediction model based on deep learning. *PloS one,* 14**,** e0214508.

DEO, R. C., SAMUI, P. & KIM, D. 2016. Estimation of monthly evaporative loss using relevance vector machine, extreme learning machine and multivariate adaptive regression spline models. *Stochastic Environmental Research and Risk Assessment,* 30**,** 1769-1784.

DOAWE. 2020. *About my region – Queensland* [Online]. Department of Agriculture,Water and the Environment
[Accessed].

ELSAADANI, M., HABIB, E., ABDELHAMEED, A. M. & BAYOUMI, M. 2021. Assessment of a Spatiotemporal Deep Learning Approach for Soil Moisture Prediction and Filling the Gaps in Between Soil Moisture Observations. *Frontiers in artificial intelligence,* 4.

FAN, J., YUE, W., WU, L., ZHANG, F., CAI, H., WANG, X., LU, X. & XIANG, Y. 2018. Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China. *Agricultural and forest meteorology,* 263**,** 225-241.

FERREIRA, L. B. & DA CUNHA, F. F. 2020. Multi-step ahead forecasting of daily reference evapotranspiration using deep learning. *Computers and electronics in agriculture,* 178**,** 105728.

FUND, W. W. 2022. *Water Scarcity* [Online]. World Wildlife Fund. Available: https://www.worldwildlife.org/threats/water-scarcity [Accessed].

GARCÍA, S., RAMÍREZ-GALLEGO, S., LUENGO, J., BENÍTEZ, J. M. & HERRERA, F. 2016. Big data preprocessing: methods and prospects. *Big Data Analytics,* 1**,** 1-22.
GHIMIRE, S., DEO, R. C., RAJ, N. & MI, J. 2019. Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. *Applied energy,* 253**,** 113541.
GOVERNMENT, Q. 2019. *Business Queensland* [Online]. Queensland: Queensland Government. Available: https://www.business.qld.gov.au/industries/farms-fishing-forestry/agriculture/agribusiness/agricultural-land-audit/climate-risk [Accessed 2019].
GOYAL, M. K., BHARTI, B., QUILTY, J., ADAMOWSKI, J. & PANDEY, A. 2014. Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS. *Expert systems with applications,* 41**,** 5267-5276.
JAMEI, M., ALI, M., KARBASI, M., SHARMA, E., JAMEI, M., CHU, X. & YASEEN, Z. M. 2023. A high dimensional features-based cascaded forward neural network coupled with MVMD and Boruta-GBDT for multi-step ahead forecasting of surface soil moisture. *Engineering Applications of Artificial Intelligence,* 120**,** 105895.
JAMEI, M., KARBASI, M., MALIK, A., JAMEI, M., KISI, O. & YASEEN, Z. M. 2022. Long-term multi-step ahead forecasting of root zone soil moisture in different climates: Novel ensemble-based complementary data-intelligent paradigms. *Agricultural Water Management,* 269**,** 107679.
KETKAR, N. 2017. Introduction to keras. *Deep learning with Python.* Springer.
KISI, O., GENC, O., DINC, S. & ZOUNEMAT-KERMANI, M. 2016. Daily pan evaporation modeling using chi-squared automatic interaction detector, neural networks, classification and regression tree. *Computers and Electronics in Agriculture,* 122**,** 112-117.
KISI, O., MIRBOLUKI, A., NAGANNA, S. R., MALIK, A., KURIQI, A. & MEHRAEIN, M. 2022. Comparative evaluation of deep learning and machine learning in modelling pan evaporation using limited inputs. *Hydrological Sciences Journal,* 67**,** 1309-1327.
LI, Q., LI, Z., SHANGGUAN, W., WANG, X., LI, L. & YU, F. 2022. Improving soil moisture prediction using a novel encoder-decoder model with residual learning. *Computers and Electronics in Agriculture,* 195**,** 106816.
LIAO, R., YANG, P., WANG, Z., WU, W. & REN, S. 2018. Development of a soil water movement model for the superabsorbent polymer application. *Soil Science Society of America Journal,* 82**,** 436-446.
LIU, P. 2015. A survey of remote-sensing big data. *Frontiers in environmental science,* 3.
MAJHI, B., NAIDU, D., MISHRA, A. P. & SATAPATHY, S. C. 2020. Improved prediction of daily pan evaporation using Deep-LSTM model. *Neural computing & applications,* 32**,** 7823-7838.
MALIK, A., KUMAR, A. & KISI, O. 2017. Monthly pan-evaporation estimation in Indian central Himalayas using different heuristic approaches and climate based models. *Computers and Electronics in Agriculture,* 143**,** 302-313.
MARCAR, N., BENYON, R., POLGLASE, P., PAUL, K., THEIVEYANATHAN, S. & ZHANG, L. 2006. Predicting the hydrological impacts of bushfire and climate change in forested catchments of the River Murray Uplands: A review.
MCNAIRN, H., MERZOUKI, A., PACHECO, A. & FITZMAURICE, J. 2012. Monitoring soil moisture to support risk reduction for the agriculture sector using RADARSAT-2. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 5**,** 824-834.
MPELASOKA, F., HENNESSY, K., JONES, R. & BATES, B. 2008. Comparison of suitable drought indices for climate change impacts assessment over Australia towards resource management. *International journal of climatology,* 28**,** 1283-1292.
NASA 2022. GIOVANNI. NASA.
NOURANI, V., ELKIRAN, G. & ABDULLAHI, J. 2020. Multi-step ahead modeling of reference evapotranspiration using a multi-model approach. *Journal of hydrology (Amsterdam),* 581**,** 124434.
PRASAD, R., DEO, R. C., LI, Y. & MARASENI, T. 2019. Weekly soil moisture forecasting with

multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach. *Catena,* 177**,** 149-166.

QUEENSLAND. 2020. *Drought Declarations* [Online]. Available: https://www.longpaddock.qld.gov.au [Accessed].

REN, H., WANG, Y.-L., HUANG, M.-Y., CHANG, Y.-L. & KAO, H.-M. 2014. Ensemble Empirical Mode Decomposition Parameters Optimization for Spectral Distance Measurement in Hyperspectral Remote Sensing Data. *Remote sensing (Basel, Switzerland),* 6**,** 2069-2083.

RIEBSAME, W. E., CHANGNON, S. A. & KARL, T. R. 2019. *Drought and natural resources management in the United States: impacts and implications of the 1987-89 drought*, Routledge.

SAGGI, M. K. & JAIN, S. 2019. Reference evapotranspiration estimation and modeling of the Punjab Northern India using deep learning. *Computers and Electronics in Agriculture,* 156**,** 387-398.

SALAM, R. & ISLAM, A. R. M. T. 2020. Potential of RT, Bagging and RS ensemble learning algorithms for reference evapotranspiration prediction using climatic data-limited humid region in Bangladesh. *Journal of Hydrology,* 590**,** 125241.

SANNER, M. F. 1999. Python: a programming language for software integration and development. *J Mol Graph Model,* 17**,** 57-61.

SILO-QUEENSLAND 2022. SILO. Queensland Government.

SOBRINO, J., GÓMEZ, M., JIMÉNEZ-MUÑOZ, J., OLIOSO, A. & CHEHBOUNI, G. 2005. A simple algorithm to estimate evapotranspiration from DAIS data: Application to the DAISEX campaigns. *Journal of hydrology,* 315**,** 117-125.

SOCIETY, W. 2022. Available: https://www.wilderness.org.au/10-endangered-australian-animals?gclid=CjwKCAjwtp2bBhAGEiwAOZZTuPVlpoEMlltqK-xOrCCLdEdqXt_-8WDCiPkd3s8RT81R-kG2VcUsdBoCPpkQAvD_BwE [Accessed].

SUEBSOMBUT, P., SEKHARI, A., SUREEPHONG, P., BELHI, A. & BOURAS, A. 2021. Field data forecasting using LSTM and bi-LSTM approaches. *Applied Sciences,* 11**,** 11820.

TERN, E. R. I. 2022. Available: https://www.tern.org.au/news-smips-soil-moisture/#:~:text=Accurate%20soil%20moisture%20data%20can,ecological%20function%20and%20environmental%20condition. [Accessed].

TIKHAMARINE, Y., MALIK, A., KUMAR, A., SOUAG-GAMANE, D. & KISI, O. 2019. Estimation of monthly reference evapotranspiration using novel hybrid machine learning approaches. *Hydrological sciences journal,* 64**,** 1824-1842.

UR REHMAN, N. & AFTAB, H. 2019. Multivariate variational mode decomposition. *IEEE Transactions on signal processing,* 67**,** 6039-6052.

WALKER, J. P., WILLGOOSE, G. R. & KALMA, J. D. 2001. One-dimensional soil moisture profile retrieval by assimilation of near-surface measurements: A simplified soil moisture model and field application. *Journal of Hydrometeorology,* 2**,** 356-373.

WANG, L., KISI, O., ZOUNEMAT-KERMANI, M. & LI, H. 2017. Pan evaporation modeling using six different heuristic computing methods in different climates of China. *Journal of Hydrology,* 544**,** 407-427.

WEI YANG, K. W. A. W. Z. 2012. Neighborhood Component Feature Selection for High-Dimensional Data. *JOURNAL OF COMPUTERS,,* 7.

WU, M., FENG, Q., WEN, X., DEO, R. C., YIN, Z., YANG, L. & SHENG, D. 2020. Random forest predictive model development with uncertainty analysis capability for the estimation of evapotranspiration in an arid oasis region. *Hydrology Research,* 51**,** 648-665.

YIN, J., DENG, Z., INES, A. V. M., WU, J. & RASU, E. 2020. Forecast of short-term daily reference evapotranspiration under limited meteorological variables using a hybrid bi-directional long short-term memory model (Bi-LSTM). *Agricultural water management,* 242**,** 106386.

ZEYNODDIN, M. & BONAKDARI, H. 2022. Structural-optimized sequential deep learning methods for surface soil moisture forecasting, case study Quebec, Canada. *Neural Computing and Applications,* 34**,** 19895-19921.

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.

APPENDIX A: RESEARCH HIGHLIGHTS AND GRAPHICAL ABSTRACT

A.1 Paper 1

Research Highlights

- Research aims to design a deep learning hybrid model for Pan Evaporation prediction.
- Neighbourhood Component Analysis is used for feature selection.
- Long Short-Term Memory network is used as the prediction algorithm.
- Target deep learning hybrid model outperforms competing benchmark models.
- The outcomes are useful for the accurate estimation of evaporative water loss.

Graphical Abstract

Outline of the study areas, data sources, and model development methodology and procedures used in research work based on the first objective explained in the journal paper forwarded in this chapter.

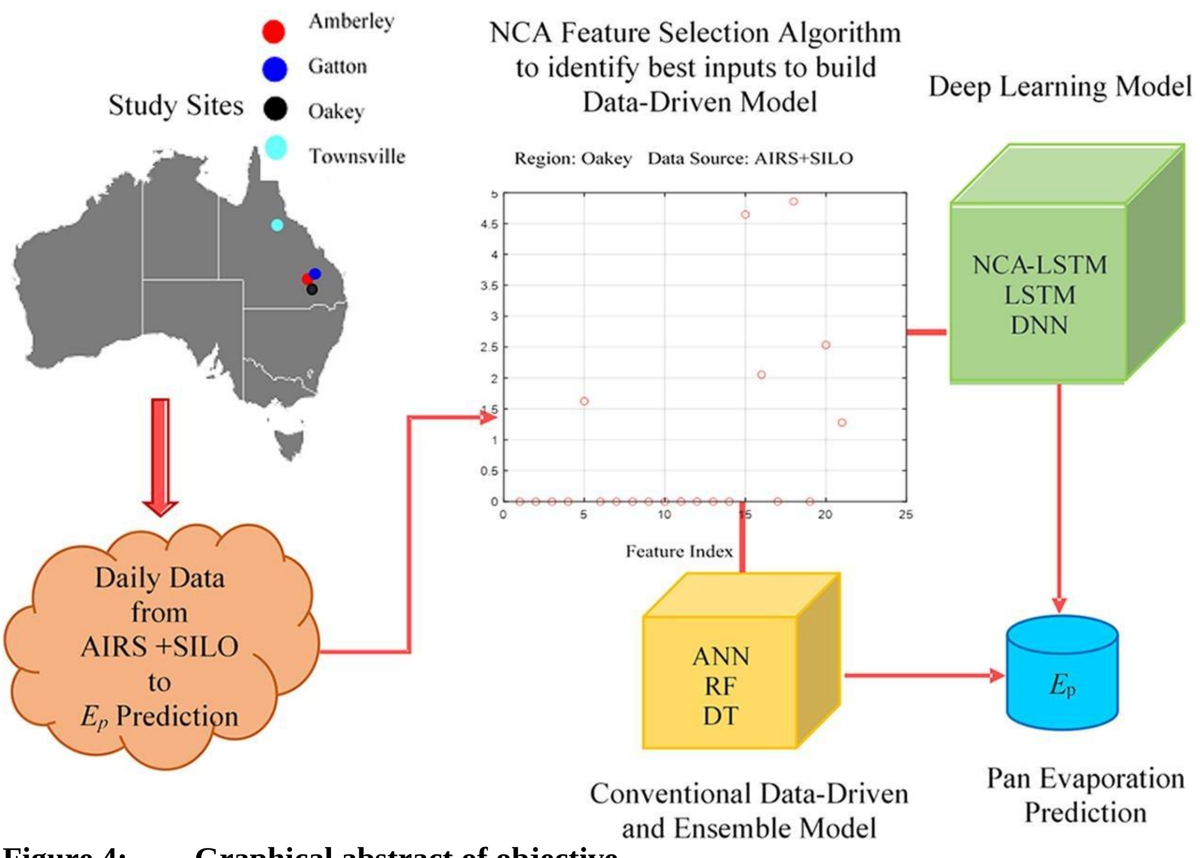


Figure 4: Graphical abstract of objective

A.2 Paper 2

Research Highlights

- This research aims to design a multi-stage deep learning hybrid model to forecast Soil Moisture.
- Maximum Overlap Discrete Wavelet Transform is used to decompose data.
- Lasso algorithm is used for feature selection.
- Long Short-Term Memory network is used as the prediction algorithm.
- Target multi-stage deep learning hybrid model beats competing benchmark models.

Graphical Abstract

Outline of the data sources and model development methodology and procedures used in research work based on the second objective explained in the journal paper forwarded in this chapter.

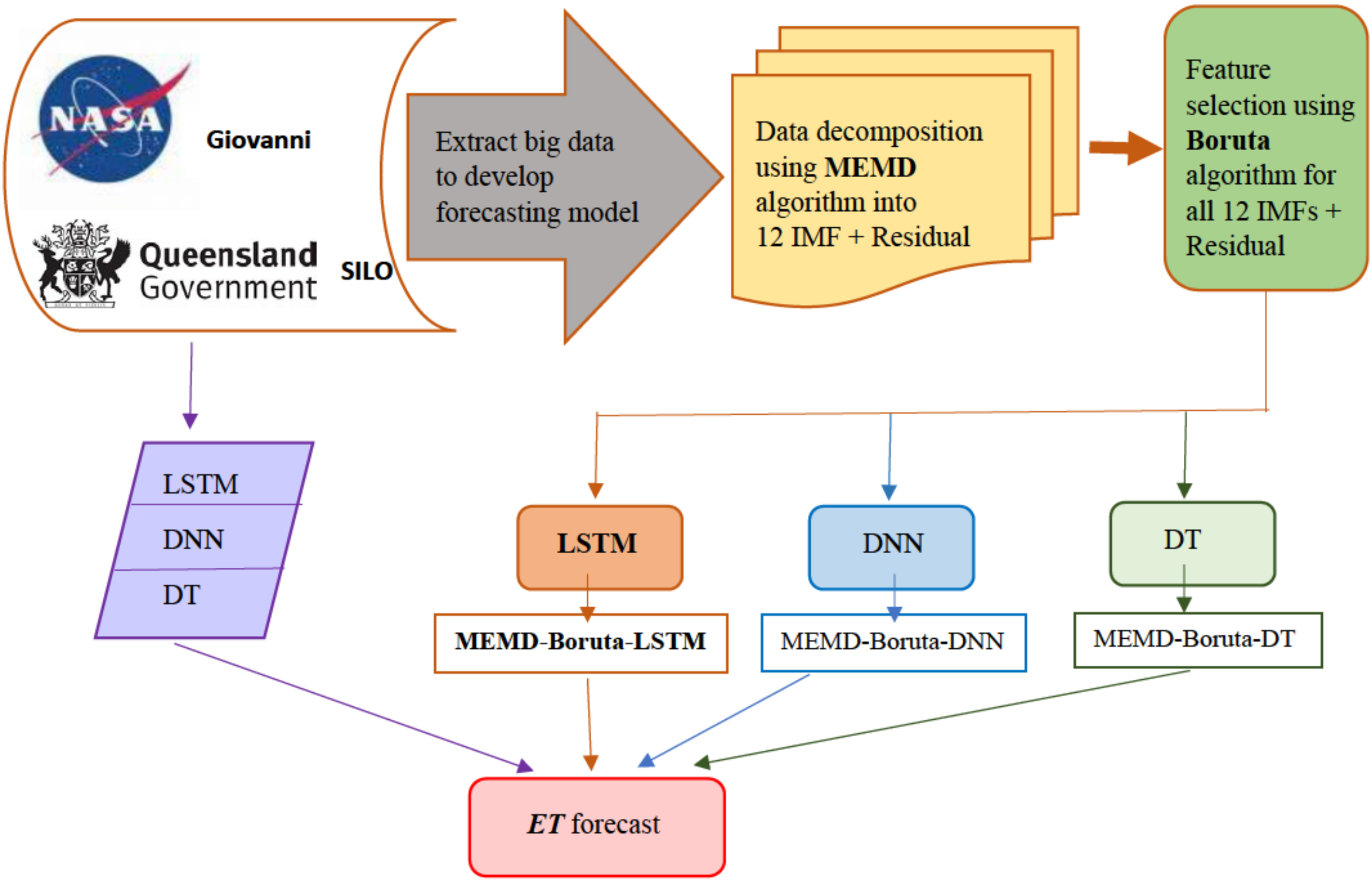


Figure 5: Graphical abstract of objective 2

A.3 Paper 3

Research Highlights

- This research aims to design a multi-stage deep learning hybrid model to forecast Soil Moisture.
- Maximum Overlap Discrete Wavelet Transform is used to decompose data.
- Lasso algorithm is used for feature selection.
- Long Short-Term Memory network is used as the prediction algorithm.
- Target multi-stage deep learning hybrid model beats competing benchmark models.
- The outcomes are useful to forecast Soil Moisture in the topsoil layer.

Graphical Abstract

Outline of the data sources and model development methodology and procedures used in research work based on the third objective explained in the journal paper forwarded in chapter 5

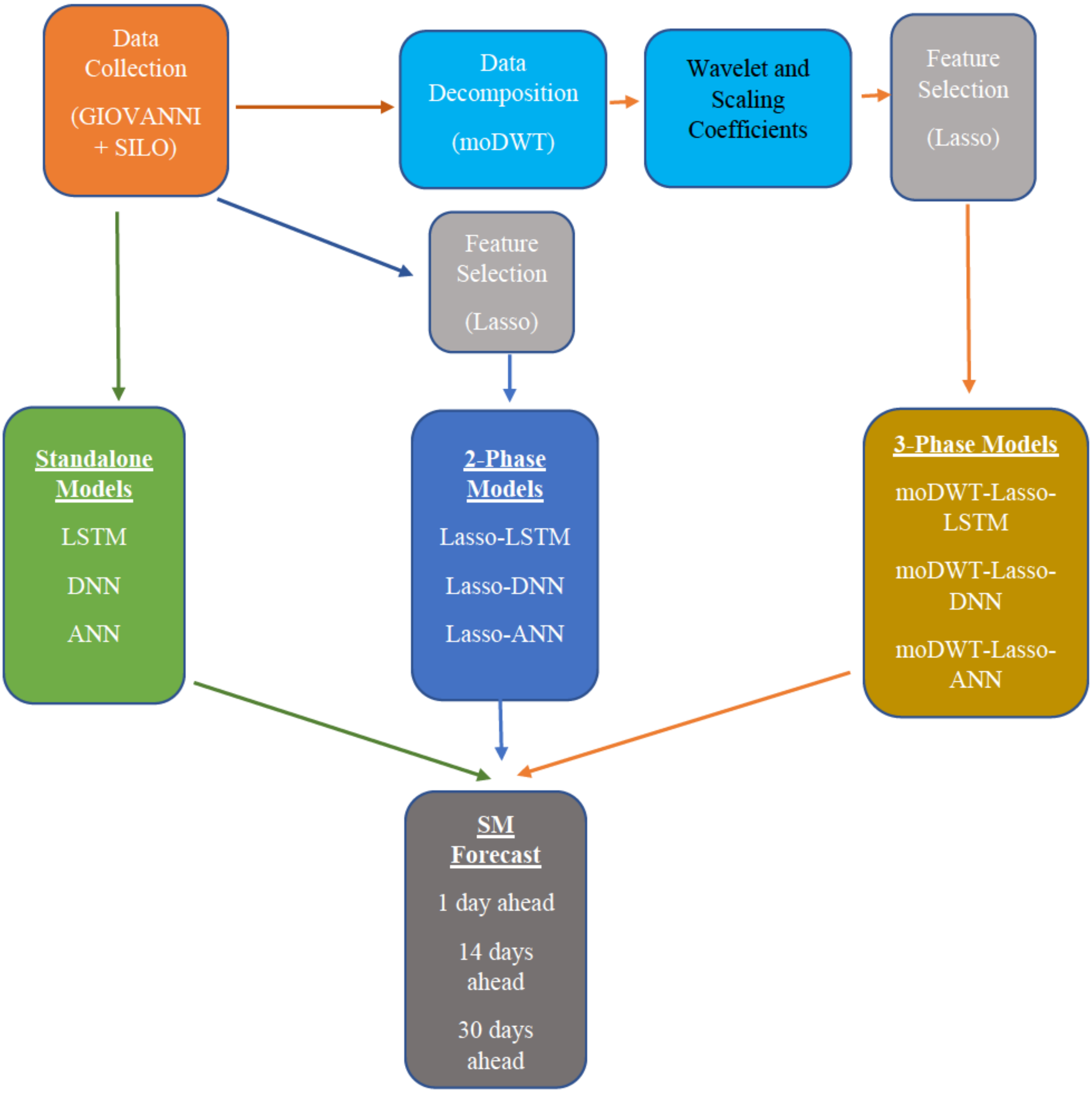
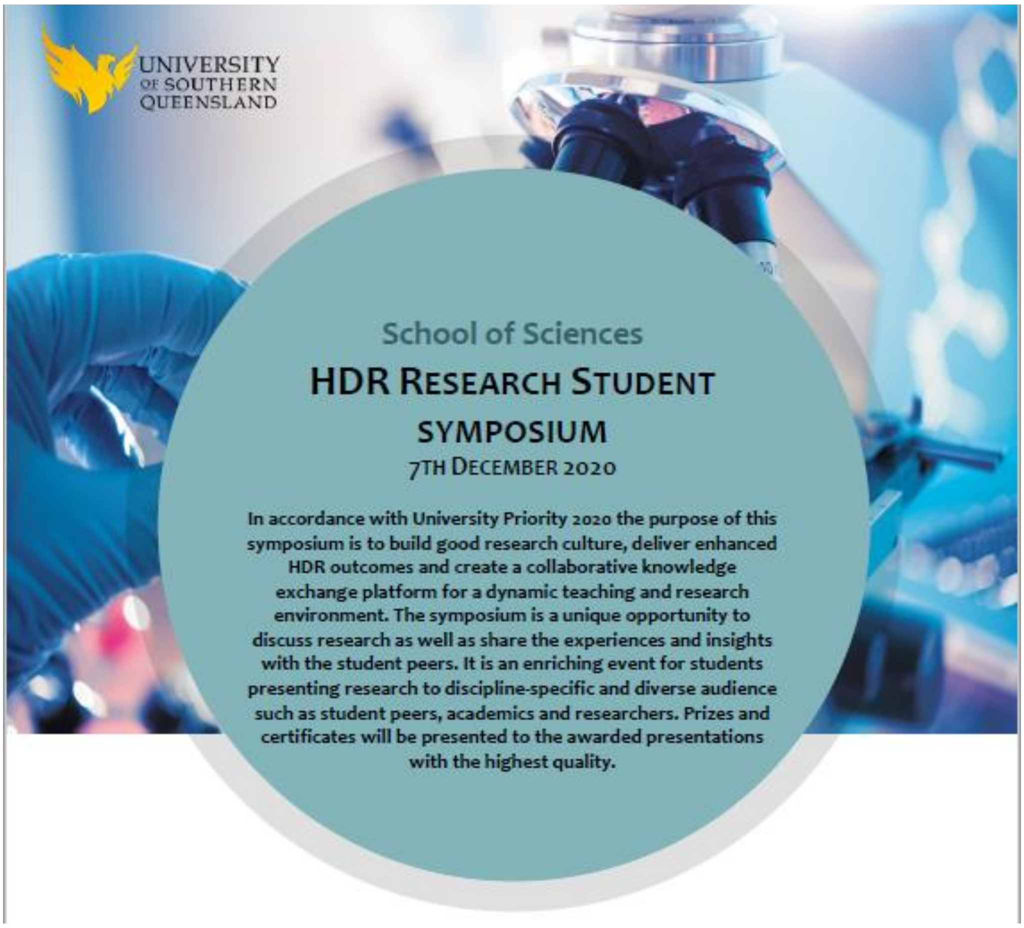


Figure 6: Graphical abstract of objective 3

APPENDIX B: PRESENTATION IN HDR SYMPOSIUM

B.1 Presentation in HDR Symposium 2020



Technical Program

School of Sciences HDR Symposium, 2020		
7th December 2020		
Zoom ID: 191 009 428 (Password 011847)		
Time slot	**Presentation**	**HDR Student**
0900-0910	Symposium Opening	
0910-0930	Defining the Australian monsoon	Joel Lisonbee
0930-0950	The East Australian Current and its role within the Climate System	Toby Pickering
0950-1010	Decadal Pacific sea surface temperature impacts on Australian monsoon rainfall variability	Hanna Heidemann
1010-1030	Drawdown and Drawup of Bi-Directional Grid Constrained Stochastic Processes	Aldo Taranto
1030-1040	Coffee break	
1040-1100	Keratinocyte skin cancer risks for working school teachers: Scenarios and implications of the timing of scheduled duty periods in Queensland, Australia	Benjamin Dexter
1100-1120	Critical analysis of past assessment within an introductory tertiary statistical course to facilitate the mastery of fundamental concepts	Taryn Axelsen
1120-1140	QLD Electricity consumption prediction	Tobias Kumie
1140-1200	Development of Flood Monitoring Index for daily flood risk evaluation: case studies in Fiji	Mohammed Moishin
1200-1300	Lunch break	
1300-1320	Deep hybrid long short-term memory network algorithm for pan evaporation prediction with neighbourhood component analysis	W.J.M. Lakmini P. Jayasinghe
[illegible]	[illegible]	

B.2 Presentation in HDR Symposium 2021

UNIVERSITY OF SOUTHERN QUEENSLAND
AUSTRALIA

School of Sciences

Certificate of Participation

Presented to

Lakmini Mudiyanselage

For abstract paper 'Daily deep multi-stage reference evapotranspiration forecasting model' presented at the

2021 School of Sciences Higher Degree Research Student Symposium

Held online on 6 December 2021 by the School of Sciences, University of Southern Queensland.

[redacted]

Associate Professor Linda Galligan
Head of School, School of Sciences

B.3 Presentation in HDR Symposium 2022

University of Southern Queensland

School of Mathematics, Physics and Computing

Certificate of Participation

Presented to

Lakmini Prarthana Jayasinghe

For 'Development of Novel Hybridized Three Phase Deep Soil Moisture Forecasting'

2022 School of Mathematics, Physics and Computing
Higher Degree Research Student Symposium

at the University of Southern Queensland

[redacted]

Professor Linda Galligan
Head of School, Mathematics, Physics and Computing

unisq.edu.au
CRICOS QLD 00244B NSW 02225M TEQSA PRV 12081

APPENDIX C: RESEARCH CONTRIBUTIONS

C.1 Journal paper reviewer

Stochastic Environmental Research and Risk Assessment (SERR) <em@editorialmanager.com>
Reply-To: "Stochastic Environmental Research and Risk Assessment (SERR)" [redacted]
To: Lakmini Prarthana Jayasinghe Warnakulasooriya Jayasinghe Mudiyanselage <lakmini.jayasinghe@gmail.com>

Wed, Apr 7, 2021 at 6:24 AM

Dear Ms Warnakulasooriya Jayasinghe Mudiyanselage,

Thank you very much for your review of manuscript
SERR-D-21-00144, "Deep Learning-based assessment of flood severity using social media streams".
We greatly appreciate your assistance.

With kind regards,
Journals Editorial Office
Springer

Animal Production Science <onbehalfof@manuscriptcentral.com>
Reply-To: editorial.an@csiro.au
To: [redacted]

Sun, Jul 3, 2022 at 4:54 PM

03-Jul-2022

Dear Dr Jayasinghe:

Thank you for reviewing manuscript # AN22172 entitled "A Multi-factor based Grading Evaluations for Pasture: A Fuzzy Data Fusion Approach" for Animal Production Science.

On behalf of the Editors of Animal Production Science, we appreciate the voluntary contribution that each reviewer gives to the Journal. We thank you for your participation in the online review process. We would also be delighted if you would consider submitting your own papers to the journal.

As a token of our appreciation of your help, CSIRO Publishing would like to offer you free online access to the journal until 31 December 2022. If you would like to take up this offer, please reply to this email with the word SUBSCRIPTION in the title. You will then be sent log-in details.

CSIRO Publishing has partnered with Publons to make it easier for you to automatically receive recognition for your review. If you have opted in to this service the review record will be added to your profile. To manually add a review record please forward this email to reviews@publons.com.

Subscribing to APS's (free) Early Alert Service at: http://www.publish.csiro.au/?nid=24 is also highly recommended.

For more information on the journal and to view the latest issue please visit the journal's website at:
www.publish.csiro.au/journals/an

Sincerely,

Dr Keith Pembleton
Associate Editor, Animal Production Science
editorial.an@csiro.au

C.2 Technical session chair

කර්මාන්ත කළමනාකරණ අධ්‍යයනාංශය
கைத்தொழில் முகாமைத்துவத் துறை
DEPARTMENT OF INDUSTRIAL MANAGEMENT

Faculty of Applied Sciences, Wayamba University of Sri Lanka, Kuliyapitiya, Sri Lanka

Your Ref: | My Ref: | Date: 27.07.2016

Mrs. WJMLF Jayasinghe.
Department of Mathematical Sciences
Faculty of Applied Sciences
Wayamba University of Sri Lanka
Kuliyapitiya.

Dear Mrs. Jayasinghe,

Letter of Appreciation –Technical Session Chair of Applied Science, Business & Industrial Research Symposium 2016

We thankfully send our sincere appreciation to you for the service rendered as the Chair of a Session in the 8th Applied Science, Business & Industrial Research Symposium (ASBIRES).

Your presence as the Chair of the Technical Session in "Mathematical Sciences" held on 27th July, 2016 at Wayamba University of Sri Lanka (WUSL), Kuliyapitiya, is a great honour for us and helped us conduct a fruitful academic session in the symposium.

We herewith acknowledge and express our gratitude and appreciation for your immense contribution towards the success of ASBIRES 2016, and we do hope your kind corporation in our future events as well.

Thank You.

Yours Sincerely,

[redacted]

CHAIR PERSON Executive Committee ASBIRES
Faculty of Applied Sciences
Wayamba University of Sri Lanka, Kuliyapitiya

Dr. MMDR Deegahawathure
Chair Person - Executive Committee (ASBIRES)

APPENDIX D: OTHER RESEARCH TALKS IN THE FIELD OF MATHEMATICAL SCIENCES

Sri Lanka Association for the Advancement of Science
Proceedings of the 65th Annual Sessions – 2009, Part I - Abstracts

SECTION E_1

501/E1

Exact formula for the sum of the squares of the Bessel and the Neumann function of the half-odd integer order

W.J.M.L.P. Jayasinghe*
Department of Mathematics, University of Wayamba, Kuliyapitiya

Sum of the squares of the spherical Bessel function and the Neumann function of the same order of an integer has been found to be very useful in theoretical nuclear physics. This sum can be obtained from the corresponding sum of the half-odd integer Bessel and Neumann functions. To our surprise, there is no exact formula for the afore mentioned sum but an approximate formula is available, which has been obtained by G.N. Watson, and is valid for the complex argument whose real part is greater than zero, and the absolute value of the upper bound of the error term is undefined in case of half-odd integers. The same formula has been obtained by G.N. Watson, which is valid for all complex arguments, using the sophisticated mathematical method called Barnes' method. However, the error term in this formula is very difficult to estimate. We have shown that the Watson formula is exact, in the important case of positive half-odd integers, using elementary mathematics and the Nicholson formula. Watson formula can be written as

Proceedings of the Annual Research Symposium 2008 – Faculty of Graduate Studies University of Kelaniya

4.18 Exact formula for the sum of the squares of spherical Bessel and Neumann function of the same order

W.J.M.L.P. Jayasinghe and R.A.D. Piyadasa
Department of Mathematics, University of Kelaniya

ABSTRACT

The sum of the squares of the spherical Bessel and Neumann function of the same order (SSSBN)is the square of the modulus of the Hankel function when the argument of all function are real, and is very important in theoretical physics. However, there is no exact formula for SSSBN.Corresponding formula, which has been derived by G.N.Watson[1] is an approximate formula[1], [2] valid for $\mathrm{Re}(z) > 0$,and it can be

canadian journal|journal|engineering|computer science|medical http://ampublisher.com/Mar 2011/CMNSEM Mar 2011.ht

Title: Simple analytical proofs of three Fermat's theorems

Authors: R.A.D.Piyadasa, A.M.D.M.Shadini, W.J.M.L.P.Jayasinghe

Pages: 50-56

Abstract—Two theorems, Fermat's last theorem for and his theorem on the Pythagorean triangles, are proved using two simple independent algorithms. A short and simple proof of Fermat's last theorems for is also discussed to point out that the method of infinite descent may be a tailor-made method by Fermat for the proof of above two theorems.

Full Text: PDF

Proceedings of the Annual Research Symposium 2008 - Faculty of Graduate Studies, University of Kelaniya

4.16 Structure of primitive Pythagorean triples and the proof of a Fermat's theorem

W.J.M.L.P. Jayasinghe, R.A.D. Piyadasa
Department of mathematics, University of Kelaniya

ABSTRACT

In a short survey of survey of primitive Pythagorean triples (x, y, z) $0 < x < y < z$, we have found that one of x, y, z is divisible by 5 and z is not divisible by 3, there are Pythagorean triples whose corresponding element are equal , but there cannot be two Pythagorean triples such that $(x_1, y_1, z_1), (x_1, z_1, z_2)$, where z_1 and z_2 hypotenuses of the corresponding Pythagorean triples. This is due to a Fermat's theorem [1] that the area of a Pythagorean triangle cannot be a perfect square of an integer, which can directly be used to prove Fermat's last theorem for $n = 4$. Therefore the preceding theorem is proved using elementary mathematics, which is the one of the main objectives of this contribution. All results in this contribution are summarized as a theorem.

Elastic scattering based on integral equation theory for potentials including the Coulomb potential

Jayasinghe WJMLP [1]

ABSTRACT

The upper bounds for the regular Coulomb wave function and the Green function are involved with the integral equation. Using the uniform convergence of the integral equation for the wave function, it is found that the wave function is an analytic function of k except at $k = 0$. In this respect it is found that S-matrix element is an analytic function of k except at $k = 0$ and at it poles.

KEYWORDS: Analytic, Elastic scattering, Poles, S-matrix, Wave function.

Proceedings of the Annual Research Symposium 2008 – Faculty of Graduate Studies University of Kelaniya

4.17 Singularities of the elastic S-matrix element

W.J.M.L.P.Jayasinghe, R.A.D.Piyadasa
Department of Mathematics, University of Kelaniya

ABSTRACT

It is well known that the standard conventional method of integral equations is not able to explain the analyticity of the elastic S-matrix element for the nuclear optical potential including the Coulomb potential. It has been shown[1],[2] that the cutting down of the potential at a large distance is essential to get rid of the redundant poles of the S-matrix element in case of an attractive exponentially decaying potential. This method has been found [3] to be quite general and it does not change the physics of the problem. Using this method , analiticity and the singularities of the S-matrix element is discussed.