



DEVELOPMENT OF DEEP LEARNING HYBRID
MODELS FOR HYDROLOGICAL PREDICTIONS

A Thesis Submitted by

ABUL ABRAR MASRUR AHMED

B. Sc., M.Sc.

For the award of

Doctor of Philosophy (PhD)

2022

ABSTRACT

Forecasting hydrologic phenomena are critical for strategic environmental planning, designing the hydrologic structures, and managing agricultural practices and water resources. Physical models are the mainstream method that helps understand the physical mechanisms and dynamics used in hydrological predictions. They are used for addressing the characteristics of hydrological phenomena while considering the initial conditions and spatial-temporal resolution of the model inputs. Data-driven models, on the other hand, are based on artificial intelligence and are designed as alternatives to discover the relationships between a set of predictors and a target variable without considering any of the initial conditions or underlying assumptions. These methods are relatively new and are becoming state-of-the-art to address different prediction problems.

This doctoral thesis, with its five primary objectives, aims to build a set of deep learning hybrid models and evaluate for their predictive skills in forecasting hydrological variables such as soil moisture (SM), evapotranspiration (ET_o), and streamflow water levels (SWL) within Australian Murray-Darling Basin. The first objective establishes the significance of feature selection to predict the monthly SWL at six study sites. The BRF-LSTM hybrid method integrated with the long-short term memory (LSTM) model with a Boruta-Random forest optimizer (BRF) is used to demonstrate the importance of feature selection for SWL forecasting problems.

The second objective is to develop a CNN-GRU hybrid model using the ant colony optimization to screen the most correlated features from a diversified set of inputs using convolutional neural network (CNN) and gated recurrent unit (GRU) networks for evapotranspiration (ET_o) forecasting. The results show that the CNN-GRU model integrated with the ACO method has outperformed the benchmark models over multi-step forecast horizons, and it has also captured the complex and non-linear relationships between predictors and daily ET_o . The third objective employs the BRF-feature selection method to identify the global climate model (GCM)-simulated variables for an LSTM model, aiming to estimate upper-layer surface soil moisture (SM) under RCP4.5 and 8.5 warming scenarios. The results demonstrate that the proposed BRF-LSTM model is more accurate than benchmark models, and this objective has established a new approach that can deal with GCM-simulated variables.

The fourth objective develops a CEEMDAN-CNN-GRU hybrid model to forecast daily surface soil moisture (SSM) by using neighbourhood component analysis (NCA), complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN), convolutional neural networks (CNN), and gated recurrent units (GRU). The CEEMDAN-CNN-GRU hybrid model outperforms all benchmark and standalone models in simulating surface soil moisture. The fifth objective is to develop the CBiLSTM hybrid model, coupled with CEEMDAN and a variational mode decomposition (VMD) to build the CVMD-CBiLSTM hybrid model for streamflow water level forecasting. This proposed model reveals that the CVMD-CBiLSTM hybrid model had outperformed the benchmark models.

The artificial intelligence (AI) methodologies developed in this PhD project are expected to be a significant step forward in developing AI-based data-driven decision support systems that will enable hydrologists and climate specialists to design water resource management strategies. Though this work focuses on soil moisture, evapotranspiration, and streamflow water level forecasting, the developed methodologies can also contribute significantly to other areas, such as flood forecasting, irrigation scheduling, and sustainable management of water resources. Overall, the doctoral study establishes significant scientific pathways for water resources management and smart farming using AI-based decision support systems.

CERTIFICATION OF THESIS

This thesis is the work of **Abul Abrar Masrur Ahmed**, except where otherwise acknowledged. The majority of the authorship in the research papers presented as a Thesis by Publication has been undertaken by the Student. The work is original and has not previously been submitted for any other award, except where acknowledged.

Signed: Abul Abrar Masrur Ahmed

Date: 25/01/2022

Endorsed by:

Professor Ravinesh C. Deo

Principal Supervisor

Dr Afshin Ghahramani

Associate Supervisor

Dr Nawin Raj

Associate Supervisor

Professor Feng Qi

External Supervisor

Professor Zhenliang Yin

External Supervisor

Dr Linshan Yang

External Supervisor

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

STATEMENT OF CONTRIBUTION

This doctoral Thesis by Publication has produced seven quartile 1 (*Q1*) publications (five articles are presented as primary contributions, and two articles are presented as supplementary contributions. One article has resulted from an Australian Postgraduate Research (APR.intern) PhD industry internship with the supervisors from CS Energy, a Queensland electricity company).

Field of Research (FOR): The focus of this doctoral thesis is in the national priority area of: ‘Artificial Intelligence and Image Processing FOR-08’, ‘Agriculture, Land and Farm Management FOR 0701, and Environmental Science and Management FOR 0502’.

Articles 1, 2, 3, 4, and 5 are primary (core) parts of this thesis, and Articles 6 and 7 are the secondary contributions placed in the Appendix section as additional research output completed during the PhD candidature. The following presents the student contributions and the contributions of the co-authors of the publications.

Article 1: Chapter 3

Ahmed A. A. M., Deo R.C., Feng Q., Ghahramani A., Raj N., Yin Z., Yang L. Deep learning hybrid model with Boruta-Random Forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. *Journal of Hydrology* 2021, 126350. (**Scopus Ranked *Q1*; Impact Factor: 5.72, SNIP: 1.87; 95th percentile in water science and technology**).

The percentage contributions of the paper are: **A. A. Masrur Ahmed** (PhD candidate) for 70% (conceptualization, data analysis, preparation of tables and figures, methodology, software, model development, application, compilation, and writing of the manuscript.). The contributions of the supervisors are Ravinesh C Deo 11% (development, and application, compilation, and writing of the manuscript), Afshin Ghahramani 5% (supervision, proofreading of the manuscript), Nawin Raj 5% (supervision, proofreading of the manuscript), Feng Qi 3% (funding, supervision & proofreading of the manuscript), Zhenliang Yin 3% (funding, supervision & proofreading of the manuscript), Linshan Yang 3% (funding, supervision & proofreading of the manuscript).

Article 2: Chapter 4

Ahmed A. A. M., Deo R.C., Feng Q., Ghahramani A., Raj N., Yin Z., Yang L. Hybrid deep learning method for a week-ahead evapotranspiration forecasting. *Stochastic Environmental Research and Risk Assessment* 2021, 1-19. (**Scopus Ranked Q1; Impact Factor: 3.38 and SNIP: 1.15; 86th percentile in water science and technology**).

The percentage contributions of the paper are: **A. A. Masrur Ahmed** (PhD candidate) for 70% (conceptualization, data analysis, preparation of tables and figures, methodology, software, model development, application, compilation, and writing of the manuscript.). The contributions of the supervisors are Ravinesh C Deo 11% (development, and application, compilation, and writing of the manuscript), Afshin Ghahramani 5% (supervision, proofreading of the manuscript), Nawin Raj 5% (supervision, proofreading of the manuscript), Feng Qi 3% (funding, supervision & proofreading of the manuscript), Zhenliang Yin 3% (funding, supervision & proofreading of the manuscript), Linshan Yang 3% (funding, supervision & proofreading of the manuscript).

Article 3: Chapter 5

Ahmed A. A. M., Deo R.C., Ghahramani A., Raj N., Feng Q., Yin Z., Yang L. 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* 2021, 35, 1851-1881. (**Scopus Ranked Q1; Impact Factor: 3.38 and SNIP: 1.15; 86th percentile in water science and technology**).

The percentage contributions of the paper are: **A. A. Masrur Ahmed** (PhD candidate) for 70% (conceptualization, data analysis, preparation of tables and figures, methodology, software, model development, application, compilation, and writing of the manuscript.). The contributions of the supervisors are Ravinesh C Deo 11% (development, and application, compilation, and writing of the manuscript), Afshin Ghahramani 5% (supervision, proofreading of the manuscript), Nawin Raj 5% (supervision, proofreading of the manuscript), Feng Qi 3% (funding, supervision & proofreading of the manuscript), Zhenliang Yin 3% (funding, supervision &

proofreading of the manuscript), Linshan Yang 3% (funding, supervision & proofreading of the manuscript).

Article 4: Chapter 6

Ahmed A. A. M., Deo R.C., Raj N., Ghahramani A., Feng Q., Yin Z., Yang L. 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* 2021, 13, 554. (*Scopus Ranked Q1; Impact Factor: 4.85 and SNIP: 1.71; 90th percentile in General Earth and Planetary Sciences*).

The percentage contributions of the paper are: **A. A. Masrur Ahmed** (PhD candidate) for 70% (conceptualization, data analysis, preparation of tables and figures, methodology, software, model development, application, compilation, and writing of the manuscript.). The contributions of the supervisors are Ravinesh C Deo 11% (development, and application, compilation, and writing of the manuscript), Afshin Ghahramani 5% (supervision, proofreading of the manuscript), Nawin Raj 5% (supervision, proofreading of the manuscript), Feng Qi 3% (funding, supervision & proofreading of the manuscript), Zhenliang Yin 3% (funding, supervision & proofreading of the manuscript), Linshan Yang 3% (funding, supervision & proofreading of the manuscript).

Article 5: Chapter 7

Ahmed A. A. M., Deo R.C., Feng Q., Ghahramani A., Raj N., Yin Z., Yang L. New double decomposition deep learning methods for river water level forecasting. *Science of Total Environment* 2022, 831, 154722 (*Scopus Ranked Q1; Impact Factor: 7.96 and SNIP: 2.015; 96th percentile in environmental engineering*).

The percentage contributions of the paper are: **A. A. Masrur Ahmed** (PhD candidate) for 70% (conceptualization, data analysis, preparation of tables and figures, methodology, software, model development, application, compilation, and writing of the manuscript). The contributions of the supervisors are Ravinesh C Deo 11%

(development, and application, compilation, and writing of the manuscript), Afshin Ghahramani 5% (supervision, proofreading of the manuscript), Nawin Raj 5% (supervision, proofreading of the manuscript), Feng Qi 3% (funding, supervision & proofreading of the manuscript), Zhenliang Yin 3% (funding, supervision & proofreading of the manuscript), Linshan Yang 3% (funding, supervision & proofreading of the manuscript).

Article 6: Appendix A

Ahmed A. A. M., Deo R.C., Segal G., Y Yu. (2022, under review). Development of Kernel Ridge Regression for bias correction of Total Cloud Cover forecast generated by Global Forecast System weather model. *Applied Energy (Scopus Ranked Q1; Impact Factor: 9.75 and SNIP: 2.696; 97th percentile in general energy)*. Manuscript No: APEN-D-21-11403. Submission Date: 26 Nov 2021. [This article was financially supported under the Australian Postgraduate Research APR.intern PhD internship conceptualised with CS Energy].

The overall contributions of **A. A. Masrur Ahmed** (PhD candidate) for this article were 80% (conceptualization, data analysis, preparation of tables and figures, methodology, software, model development, application, compilation, and manuscript writing). The contributions of the supervisors were: RCD 10% (development, application, compilation, and writing of the manuscript), GS 5% (funding, supervision, conceptualization), and YY 5% (funding, supervision, conceptualization).

Article 7: Appendix B

Ahmed A. A. M., Sharma E., Deo R.C., Nguyen T-H., Ali M., Jui S. J. J. (2022, in review). Kernel Ridge Regression hybrid method for wheat yield prediction using satellite-derived predictors. *Remote Sensing*, 2022, 14, 1136 (**Scopus Ranked Q1; Impact Factor: 4.85 and SNIP: 1.71; 90th percentile in General Earth and Planetary Sciences**).

The overall contributions of **A. A. Masrur Ahmed** (PhD candidate) for this article were 65% (conceptualization, data analysis, preparation of tables and figures, methodology, software, model development, and application, compilation, and writing of the manuscript). The contributions of the collaborators were ES 10% (writing of the manuscript), RCD 10% (development, and application, compilation, and writing of the manuscript), MA 5% (proofreading of the manuscript), and TN-H 5% (conceptualization, proofreading of the manuscript), and SJJJ 5% (proofreading of the manuscript).

LIST OF PUBLICATIONS

1. **Ahmed A. A. M.**, Deo R.C., Feng Q., Ghahramani A., Raj N., Yin Z., Yang L. Deep learning hybrid model with Boruta-Random Forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. *Journal of Hydrology* 2021, 126350. (**Q1; Impact Factor: 5.72 and SNIP: 1.87; 95th percentile**).
2. **Ahmed A. A. M.**, Deo R.C., Feng Q., Ghahramani A., Raj N., Yin Z., Yang L. Hybrid deep learning method for a week-ahead evapotranspiration forecasting. *Stochastic Environmental Research and Risk Assessment* **2021**, 1-19. (**Q1; Impact Factor: 3.38 and SNIP: 1.15; 86th percentile**).
3. **Ahmed A. A. M.**, Deo R.C., Ghahramani A., Raj N., Feng Q., Yin Z., Yang L. 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* **2021**, 35, 1851-1881. (**Q1; Impact Factor: 3.38 and SNIP: 1.15; 86th percentile**).
4. **Ahmed A. A. M.**, Deo R.C., Raj N., Ghahramani A., Feng Q., Yin Z., Yang L. 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* 2021, 13, 554. (**Q1; Impact Factor: 4.85 and SNIP: 1.71; 90th percentile**).
5. **Ahmed A. A. M.**, Deo R.C., Feng Q., Ghahramani A., Raj N., Yin Z., Yang L. New double decomposition deep learning methods for river water level forecasting. *Science of Total Environment* 2022, 831, 154722 [**Under review**] (**Q1; Impact Factor: 7.96 and SNIP: 2.015; 96th percentile**).

6. **Ahmed A. A. M.**, Deo R.C., Segal G., Yu Y. (2022, in review). Development of Kernel Ridge Regression for bias correction of Total Cloud Cover forecast generated by Global Forecast System weather model. *Applied Energy* (***Q1; Impact Factor: 9.75 and SNIP: 2.696; 97th percentile***). *Manuscript No: APEN-D-21-11403*. Submission Date: 26 November 2021. [This article was supported under the Australian Postgraduate Research APR.intern PhD internship with CS Energy].
7. **Ahmed A. A. M.**, Sharma E., Deo R.C., Nguyen T-H., Ali M., Jui S. J. J. (2022, in review). Kernel Ridge Regression hybrid method for wheat yield prediction using satellite-derived predictors. *Remote Sensing*, 14, 1136 (***Q1; Impact Factor: 4.85 and SNIP: 1.71; 90th percentile***).

ACKNOWLEDGEMENTS

First and foremost, I want to convey my gratitude and admiration for Prof Ravinesh C Deo, PhD, my Principal Research Supervisor and the Leader of USQ's Advanced Data Analytics Group. He provided me with regular, much-needed help creating the project concept and kept me motivated throughout the journey. His constant assistance and supervision allowed me to finish this thesis. His quick responses providing essential advice and feedback, assisted me widely to conceptualise the thesis that helped me publish in high-impact factored journals.

A special thanks to my associate supervisors, Dr Afshin Ghahramani and Dr Nawin Raj, who provided invaluable support and editorial help throughout the process. I would also like to extend my gratitude to Professor Qi Feng, Zhenliang Yin, and Linshan Yang, external supervisors, who gave essential guidance and editing of the manuscript. I would like to thank the entire supervisory team for their guidance in publishing high-quality research papers.

A special thank you to the University of Southern Queensland (USQ) and the Chinese Academy of Science (CAS) (Northwest Institute of Eco-Environment and Resources and its Deputy Director Professor Feng Qi) for providing a USQ-CAS Postgraduate Research Scholarship (2019-2021) to continue the study. This study would not have taken place if it hadn't been for the funding. I would like to extend my heartfelt gratitude to the Bureau of Meteorology, Australia, the New South Wales Department of Primary Industries-Office of Water, GIOVANNI that provided free-to-access data, including the Scientific Information for Landowners (SILO) and National Centre for Earth Observation.

Following that, I'd like to express my heartfelt gratitude to my mother, Mrs Nazma Khatun, for her love, support, and encouragement. Thank you to my siblings, Abul Mabrur and Sakia Naja, for their support throughout the years. I'd also like to thank my friend, Kazi Nazmul Haque, for supporting me throughout the journey. I highly appreciate the support of my friends, Manrose Mannan, Maruf Hasan, and Bristy Siddiky. I am thankful to all the Advanced Data Analytics: Environmental Modelling & Simulation Research Group members, including Md Moishin, Tobius Kumie, Aditi Bose and Ekta Sharma, for their encouragement and insightful discussions throughout this process.

A particular word of thanks goes out to Mrs. Fatimatuj Zahera, my wife, for her patience and unwavering support throughout this ordeal. I cannot thank you enough for taking care of my parents and being confident about my research journey. I would like to extend my love to Adiyah Hamid and Affan Hamid, two of my beloved sons; you have been the light of my life and have given me extra strength and motivation.

I am thankful to Dr Barbara Harmes of the English Angels Program, University of Southern Queensland, Australia, for providing help in proofreading the work. The suggestions provided by Dr Douglas Eacersall, HDR Learning Advisor, University of Southern Queensland, are highly acknowledged.

Final thoughts: I am grateful for the opportunity to attend the University of Southern Queensland, a distinguished institution. This experience has allowed me to collaborate with some of the most professional individuals globally and the resources necessary to achieve tremendous success. Thanks for this fantastic opportunity.

DEDICATION

To the memory of my beloved father, **A J Kawsar Ahmed**, who never doubted my potential to succeed, you've left, yet your faith in me has allowed me to make this voyage. May Allah bless you.

TABLE OF CONTENTS

ABSTRACT	II
CERTIFICATION OF THESIS	IV
STATEMENT OF CONTRIBUTION	V
LIST OF PUBLICATIONS	X
ACKNOWLEDGEMENTS	XII
DEDICATION	I
TABLE OF CONTENTS	I
LIST OF TABLES	III
LIST OF FIGURES	I
LIST OF ACRONYMS	XIV
HYBRID MODELS NOTATIONS	XVII
CHAPTER 1: INTRODUCTION	1
1.1 BACKGROUND	1
1.2 STATEMENT OF THE PROBLEM	3
1.3 OBJECTIVES	6
1.4 SIGNIFICANCE OF THE RESEARCH	9
1.3 THESIS LAYOUT	10
CHAPTER 2: DATA AND METHODOLOGY	14
2.1 STUDY AREA: THE MURRAY-DARLING BASIN	14
2.2 DATA DESCRIPTION	16
2.2.1 Streamflow water level (SWL) - NSW Department of Primary Industries	18
2.2.2 Meteorological data - Scientific Information for Landowners	18
2.2.3 Atmospheric Parameters - The Moderate Resolution Imaging Spectroradiometer (MODIS)	19
2.2.4 GCM/CMIP5 Simulated Variables	23
2.2.5 Synoptic Scale Climate Mode Indices	25
2.3 GENERAL METHODOLOGY	26
2.4 MODEL EVALUATION	32
CHAPTER 3: STREAMFLOW WATER LEVEL FORECASTING USING CLIMATE INDICES, RAINFALL, AND PERIODICITY	33
3.1 FOREWORD	33
3.2 RESEARCH HIGHLIGHTS	33
3.3 ARTICLE 1	34
CHAPTER 4: EVAPOTRANSPIRATION FORECASTING MODEL AT MULTI-STEP HORIZON	58

4.1 FOREWORD.....	58
4.2 RESEARCH HIGHLIGHTS	58
4.3 ARTICLE 2.....	59
CHAPTER 5: SOIL MOISTURE ESTIMATION UNDER RCP4.5 AND RCP8.5 GLOBAL WARMING SCENARIOS	79
5.1 FOREWORD.....	79
5.2 RESEARCH HIGHLIGHTS	79
5.3 ARTICLE 3.....	80
CHAPTER 6: SURFACE SOIL MOISTURE FORECASTING AT MULTI-STEP HORIZON	112
6.1 FOREWORD.....	112
6.2 RESEARCH HIGHLIGHTS	112
6.3 ARTICLE 4.....	113
CHAPTER 7: MULTI-STEP AHEAD STREAMFLOW WATER LEVEL FORECASTING USING DOUBLE DECOMPOSITION AND DEEP LEARNING METHODS	144
7.1 FOREWORD.....	144
7.2 RESEARCH HIGHLIGHTS	144
7.3 ARTICLE 5.....	145
CHAPTER 8: CONCLUSIONS AND FUTURE SCOPE	167
8.1 SYNTHESIS AND CONCLUSIONS.....	167
8.2 LIMITATIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH	170
REFERENCES	172
APPENDIX A: BIAS CORRECTION OF TOTAL CLOUD COVER FORECAST FROM GLOBAL FORECAST SYSTEM MODEL.....	185
A1.1 FOREWORD.....	185
A1.2 RESEARCH HIGHLIGHTS	185
A1.3 ARTICLE 6	186
APPENDIX B: WHEAT YIELD PREDICTION USING SATELLITE-DERIVED INFORMATION	222
B1.1 FOREWORD.....	222
B1.2 RESEARCH HIGHLIGHTS	222
B1.3 ARTICLE 7.....	223

LIST OF TABLES

Chapter 2

- Table 2.1** The datasets used in the study
- Table 2.2** Description of predictor variables used to design and evaluate hybrid CEEMDAN-CNN-GRU predictive model for daily surface soil moisture forecasting.
- Table 2.3** Twelve climate model indices were used as predictor variables to forecast the SWL using the hybrid deep learning BRF-LSTM predictive model. Source of data: monthly sea surface temperature (SST) in different oceanic regions derived from the Optimum Interpolation SST, v2 (OISST v2)
- Table 2.4** Summary of global climate models from CMIP5 simulation sets
- Table 2.5** Summary of CMIP5 simulated variables used in this study of Objective 3

Chapter 3

- Table 1** Geographic location of the study sites within the Australian Murray Darling Basin (MDB) with drainage area and mean streamflow water level (SWL) for selected study stations where the hybrid is deep learning BRF-LSTM model was developed and implemented.
- Table 2** Twelve climate model indices were used as predictor variables to forecast the SWL using the hybrid deep learning BRF-LSTM predictive model. Source of data: monthly sea surface temperature in different oceanic regions derived from the Optimum Interpolation SST, v2 (OISST v2) downloaded from Climate Prediction Center (CPC, NOAA).
- Table 3** (a) Architecture of the hybrid LSTM, GRU, and RNN-based predictive model developed through a grid search procedure. (b) The optimum architecture used to design the LSTM, GRU, and RNN-based

predictive models. (c) The hyper-parameters for the grid search process and optimum architecture of the SVR-based predictive model.

Table 4 An evaluation of the performance of the hybrid deep learning BRF-LSTM *vs.* the standalone LSTM and other predictive models (BRF-GRU, BRF-RNN, BRF-SVR) was developed and tested at six study stations within the Murray Darling Basin (MDB). The performance metrics were deduced from observed and forecasted datasets, assessed in terms of the correlation coefficient (r), root mean square error ($RMSE$; m), mean absolute error (MAE ; m), and the Nash-Sutcliffe coefficient, NS) in the testing phase.

Table 5 The forecasted error ($|FE|$, m) in the maximum and minimum values, including the standard deviations obtained in the testing phase, is calculated from the observed and forecasted SWL.

Table 6 Hybrid deep learning BRF-LSTM model performance for all stations in the training and validation set

Chapter 4

Table 1 Experimental Datasets Description. Note: The ID is the Station's ID of BOM-Australia-owned station.

Table 2 The RRMSE (%) error of CNN-GRU and other baseline models using single data sets

Table 3 Evaluation of the performance of CNN-GRU *vs.* the Benchmark models at Menindee stations. The performance metrics were deduced from observed and forecasted datasets, assessed in the correlation coefficient (r) and root mean square error ($RMSE$; mm) in the testing phase. Note: ET_7 , ET_{14} , ET_{21} , and ET_{28} are referred to as ET_o for week-1, week-2, week-3, and week-4.

Table 4 Evaluation of the performance of CNN-GRU *vs.* the Benchmark models at Fairfield stations. The performance metrics were deduced from observed and forecasted datasets assessed in the correlation coefficient (r), and root mean square error ($RMSE$; mm) in the testing

phase. Note: ET₇, ET₁₄, ET₂₁, and ET₂₈ are referred to as reference evapotranspiration for week-1, week-2, week-3, and week-4, respectively.

Table 5 Evaluation of the performance of CNN-GRU vs the Benchmark models at Gabo Island stations. The performance metrics were deduced from observed and forecasted datasets assessed in the correlation coefficient (r), and root mean square error (RMSE; mm) in the testing phase. Note: ET₇, ET₁₄, ET₂₁, and ET₂₈ are referred to as reference evapotranspiration for week-1, week-2, week-3, and week-4, respectively.

Table 6 Evaluation of the performance of CNN-GRU vs. the Benchmark models at Gatton at USQ stations. The performance metrics were deduced from observed and forecasted datasets, assessed in the correlation coefficient (r) and root mean square error (RMSE; mm) in the testing phase. Note: ET₇, ET₁₄, ET₂₁, and ET₂₈ are referred to as reference evapotranspiration for week-1, week-2, week-3, and week-4, respectively.

Chapter 5

Table 1 Summary of global climate models from CMIP5 simulation sets

Table 2 Summary of climate variables used in this study

Table 3 a) Architecture hybrid LSTM model and b) the optimum architecture used in designing the LSTM model. Hyper-parameters are obtained through a grid search procedure (Table 3a). Note: ReLU stands for Rectified Linear Units, SGD = Stochastic gradient descent optimizer

Table 4 Evaluation of the performance of BRF-LSTM vs. the BRF-SVR, BRF-MARS, SVR, and MARS models at five stations of ACCES1.1 with the correlation coefficient (r), root mean square error (RMSE; mm), mean absolute error (MAE; mm), and the standardized performance metrics (Willmott's Index, WI & Nash-Sutcliffe

coefficient, *NS*) between the predicted and observed soil moisture data in the testing phase.

Table 5 Evaluation of the performance of BRF-LSTM *vs.* the BRF-SVR, BRF-MARS, SVR, and MARS models at five stations of ACCESS1.3 with the correlation coefficient (*r*), root mean square error (*RMSE*; mm), mean absolute error (*MAE*; mm), and the standardized performance metrics (Willmott's Index, *WI* & Nash-Sutcliffe coefficient, *NS*) between the predicted and observed soil moisture data in the testing phase.

Table 6 Evaluation of the performance of BRF-LSTM *vs.* the BRF-SVR, BRF-MARS, SVR, and MARS models at five stations of HadGEM2-CC with the correlation coefficient (*r*), root mean square error (*RMSE*; mm), mean absolute error (*MAE*; mm), and the standardized performance metrics (Willmott's Index, *WI* & Nash-Sutcliffe coefficient, *NS*) between the predicted and observed soil moisture data in the testing phase

Table 7 Evaluation of the performance of BRF-LSTM *vs.* the BRF-SVR, BRF-MARS, SVR, and MARS models at five stations of HadGEM2-ES with the correlation coefficient (*r*), root mean square error (*RMSE*; mm), mean absolute error (*MAE*; mm), and the standardized performance metrics (Willmott's Index, *WI* & Nash-Sutcliffe coefficient, *NS*) between the predicted and observed soil moisture data in the testing phase

Table 8 The training performances of the proposed hybrid deep learning model (i.e., BRF-LSTM) with selected stations of respective GCMs for RCP4.5 and RCP8.5 global warming scenarios.

Table 9 Diebold–Mariano (DM) test was adopted to compare the predictive accuracy of any two forecasting Models (i.e., BRF-LSTM *vs.* SVR) for selected GCMs with RCP4.5 and RCP8.5 global warming scenarios.

Table 10 Kolmogorov-Smirnov (KS) test for normality of the estimated SM under RCP4.5 and RCP8.5 global warming scenarios.

Chapter 6

Table 1 Geographic locations and physical characteristics of selected sites in Murray Darling Basin

Table 2 Description of the global pool of 52 predictor variables used to design and evaluate hybrid CEEMDAN-CNN-GRU predictive model for daily surface soil moisture forecasting.

Table 3 (a) Range of tested hyperparameters in designing hybrid CNN-GRU and GRU predictive models through grid search. (b) Optimally selected hyperparameters. ReLU stands for Rectified Linear Units, SGD stands for stochastic gradient descent optimiser

Table 4 Evaluation of hybrid CEEMDAN-CNN-GRU *vs.* benchmark (CNN-GRU, CEEMDAN-GRU, GRU) models for the specific case of Menindee study site. The correlation coefficient (r), root mean square error ($RMSE$; Kg m^{-2}), mean absolute error (MAE ; Kg m^{-2}), and Nash-Sutcliffe coefficient, NS) is computed between forecasted and observed surface soil moisture for the 1st Day, 5th Day, 7th Day, 14thDay, 21stDay, and 30thDay ahead periods in the testing phase. The optimal model is boldfaced (blue).

Chapter 7

Table 1 Geographic locations and physical characteristics of selected sites in the Murray River System.

Table 2 Description of predictor variables used to design and evaluate hybrid CVMD-CBiLSTM predictive model for daily *SWL* forecasting.

Table 3 Optimally selected hyperparameters of deep learning models. ReLU stands for Rectified Linear Units, SGD stands for stochastic gradient descent optimiser

- Table 4** Evaluation of hybrid CVMD-CBiLSTM *vs.* benchmark (CBiLSTM, BiLSTM, SVR) models for the Murray River System study sites. The correlation coefficient (r) and root mean square error (RMSE; m) are computed between forecasted and observed stream water levels in the testing phase for the 7-Day ahead periods.
- Table 5** Promoting Percentage of Legates and McCabe’s (LM) Index (δ_{LM}), Mean Absolute Percentage Error (δ_{MAPE}), and the Relative Root Mean Square Error (δ_{RRMSE}) to compare the various models used in SWL forecasting.

Appendix A

- Table 1** List of Global Forecast System (GFS)-forecast variables (*i.e.*, 2-metre temperature, 10-metre wind speed, total cloud cover, and downward short-wave radiation flux) used as KRR model inputs, and GFS analysis variable (*i.e.*, total cloud cover used as proxy observed) in the proposed KRR model used in bias correction problem.
- Table 2** Descriptive statistics of GFS forecast and GFS analysis data were used to develop the proposed KRR model. Data were acquired from the GFS model over January 1, 2019, and April 30, 2020, used for training 70% and testing (30%), where the remaining 15% of the training set is used for model validation.
- Table 3** Mean Absolute Error (MAE, %) between ‘proxy observed’ (TCDC_{GFS-Analysis}) and ML-bias corrected TCDC_{BC} using our proposed KRR model. Our conventional bias correction MRNBC method, whereas benchmark methods include BNR, DTR, GBR, HGBR, KNN, MARS, MLR, and RF model. Approach 1 used T2m_{GFS-Forecast}, V_{GFS-Forecast}, U_{GFS-Forecast}, TCDC_{GFS-Forecast}, and DSWRF_{GFS-Forecast}. In contrast, in Approach 2, we used TCDC_{GFS-Forecast} as a predictor (or input) variable against TCDC_{GFS-Analysis} as a target variable.

Table 4	The optimal hyperparameter of the proposed KRR model, including that of the other benchmark models methods, include machine learning (i.e., BNR, DTR, GBR, HGBR, KNN, MARS, MLR, & RF)
Table 5	List of Global Forecast System (GFS)-forecast variables (<i>i.e.</i> , 2-metre temperature, 10-metre wind speed, total cloud cover, and downward short-wave radiation flux) used as KRR model inputs, and GFS analysis variable in the proposed KRR model used in bias correction problem.
Table 6	Descriptive statistics of GFS forecast and GFS analysis data were used to develop the proposed KRR model. Data were acquired from GFS model over January 1, 2019, and April 30, 2020, used for training 70% and testing (30%), where the remaining 15% of the training set is used for model validation.
Table 7	Mean Absolute Error (MAE, %) between ‘proxy observed’ ($TCDC_{GFS-Analysis}$) and ML-bias corrected $TCDC_{BC}$ using our proposed KRR model. Our conventional bias correction is an MRNBC method, whereas benchmark methods include BNR, DTR, GBR, HGBR, KNN, MARS, MLR, and RF model. Approach 1 used $T2m_{GFS-Forecast}$, $V_{GFS-Forecast}$, $U_{GFS-Forecast}$, $TCDC_{GFS-Forecast}$, and $DSWRF_{GFS-Forecast}$. In contrast, in Approach 2, we used $TCDC_{GFS-Forecast}$ as a predictor (or input) variable against $TCDC_{GFS-Analysis}$ as a target variable.
Table 8	The optimal hyperparameter of the proposed KRR model, including that of the other benchmark models methods, include machine learning (i.e., BNR, DTR, GBR, HGBR, KNN, MARS, MLR, and RF)

Appendix B

Table 1	(a) A description of the 32 predictors from the MERRA-2 satellite system used to design the hybrid GWO-CEEMDAN-KRR model for wheat yield prediction (tonnes) in South Australia. (b) Feature selections were undertaken using Grey Wolf Optimization (GWO), Ant Colony Optimization (ACO), Particle Swarm Optimization (PSO), and Atom Search Optimization (ASO) and a “√” shows the selected feature whereas a “×” shows the rejected feature.
Table 2	Evaluation of the hybrid CEEMDAN-KRR <i>vs.</i> the benchmark (i.e., CEEMDAN-MLR, CEEMDAN-RF, CEEMDAN-SVR) models and their standalone counterparts (i.e., KRR, MLR, RF, and SVR) models. The <i>r</i> and normalized root mean square error (<i>NRMSE</i>) is computed between predicted and observed Wheat Yield (Y, tones) South Australia. The optimal model, GWO-CEEMDAN-KRR, is boldfaced (blue).
Table 3	Table 3 The Optimal parameter values for the optimization algorithms (i.e., GWO, ACO, ASO, and PSO)

LIST OF FIGURES

Chapter 1

- Fig. 1.1** The layout of the thesis
Fig. 1.2 Schematic view of thesis

Chapter 2

- Fig. 2.1** Map of the study region (a) the location of Murray Darling Basin (MDB), with close-ups (b-f) illustrating the selected sites for objectives 1 to 5, respectively.
Fig. 2.2 Brief overview of artificial intelligence (AI) based-deep learning predictive models used in this doctoral research thesis.

Chapter 3

- Fig. 3.1** Graphical Abstract of Objective 1
Fig. 1 Map of the study region with oceanic representation used to calculate the climate mode indices. Notations and equations of the climate indices are provided in Table 2.
Fig. 2 The Australian Murray Darling Basin shows the selected river stations.
Fig. 3 The monthly mean values of streamflow water level (SWL) for six selected stations within the Australian Murray Darling Basin
Fig. 4 Topographical structure of the predictive models designed for the prediction of SWL (a) 4-layered Long short-term memory (LSTM) and (b) 3-layered Gated Recurrent Unit Network (GRU). Note: x_t is the new input, h_t is the hidden state, h_{t-1} is the last hidden state, \tilde{C}_t is the cell state, \tilde{C}_{t-1} is the previous cell state, \tanh is the hyperbolic tangent function, O_t is the output gate and σ is the logistic sigmoid function]
Fig. 5 Workflow detailing the steps in the model designing, as for the proposed hybrid BRF-LSTM and BRF-GRU predictive models

- Fig. 6** Box plots of the Z-scores by the Boruta for the Brewarrina (ST6) as an example used in determining significant features from a) the climate indices and b) the considerable lags of climate indices, rainfall, periodicity, and SWL. Blue resembles the shadow inputs, while green represents the Z-score distributions of confirmed inputs with notably considerable importance.
- Fig. 7** Partial autocorrelation function (PACF) plot of the SWL time series exploring the antecedent behaviour in terms of the lagged values of SWL every month. The red line in the figures indicates the $\pm 95\%$ confidence level.
- Fig. 8** Correlogram showing the covariance between the objective variable (SWL) and the predictor variable (climate indices, rainfall, and periodicity) in terms of the Cross-correlation coefficient (r_{cross}) for North Cuerindi stations of MDB. Blue lines indicate the significance of r_{cross} at the 95% confidence interval for each panel.
- Fig. 9** Comparison of the forecasting skill for all of the proposed models in terms of the relative error: RRMSE (%) and RMAE (%) within the testing period.
- Fig. 10** Cumulative frequency of the SWL generated by the objective model (BRF-LSTM) *vs.* BRF-GRU and the other comparing models in terms of absolute forecasting error ($|FE|$) within the Murray Darling Basin's six study sites.
- Fig. 11** Tylor diagram representing the correlation coefficient together with the standard deviation difference for proposed hybrid BRF-LSTM and BRF-GRU *vs.* benchmark models for (a) Coggan (b) Barham (c) Wee Jasper (d) Cowra (e) North Cuerindi and (f) Brewarrina stations.
- Fig. 12** Comparison between forecasted SWL and observed SWL during the model's testing using the objective model (BRF-LSTM), *vs.* BRF-GRU, LSTM, and GRU-based predictive models.

- Fig. 13** Monthly relative forecasting error (%) generated by the proposed hybrid BRF-LSTM and BRF-GRU model vs. the comparative counterpart models (*i.e.*, LSTM and GRU) For six stations of MDB.
- Fig. 14** Comparison of the Legates and McCabe's Index (LM) for the proposed hybrid deep learning approach (BRF-LSTM and BRF-GRU) against the other standalone models.
- Fig. 15** Bar graphs showing the mean absolute percentage error (MAPE) for the proposed hybrid deep learning approach (BRF-LSTM and BRF-GRU) in comparison with the other standalone models.

Chapter 4

- Fig 4.1** Graphical Abstract of Objective 2
- Fig. 1** Reference Evapotranspiration (ET_o) value from Menindee station January 2019 to March 2020.
- Fig. 2** Selected time-series predictors from January 2019 to March 2020 (include avg surface temp, evaporation, SOI, Rainfall, and minimum temperature).
- Fig. 3** The architecture of CNN model with 2-layered Gated Recurrent Unit for a hybrid CNN-GRU model at week 1–4 ahead ET_o forecasting with Ant Colony Optimization.
- Fig. 4** Partial autocorrelation function (PACF) plot of the ET_o time series exploring the antecedent behaviour in terms of the lag of ET_o . The red line in the figures indicates the $\pm 95\%$ confidence level.
- Fig. 5** Correlogram shows the covariance between the objective variable (ET_o) and the predictor variable (plant canopy surface water and average surface skin temperature) in terms of the cross-correlation coefficient for Menindee stations MDB.
- Fig. 6** Time series of daily evapotranspiration (ET_o , mm) for the observed and forecasted ET_o for the objective model CNN-GRU for (a) MODIS

Satellite, (b) SILO data, and (c) Climate mode indices at week 1 lead forecasting.

- Fig. 7** An empirical cumulative distribution function (CDF) plot of $|FE|$ of the CNN-GRU, CNN-LSTM the standalone models (*i.e.*, GRU, LSTM, RNN, MLR, RF, etc.) for Menindee station at Week 1, 2, 3, and 4 horizons with Ant Colony Optimization in forecasting ET_o (mm) at 95 percentiles on ECDF. Note: ET_7 , ET_{14} , ET_{21} , and ET_{28} are referred to as reference evapotranspiration for week-1, week-2, week-3, and week-4, respectively.
- Fig. 8** Comparing the forecasting skill of the proposed models in RRMSE (%) and NS in the Menindee station's testing period. Note: ET_7 , ET_{14} , ET_{21} , and ET_{28} are referred to as reference evapotranspiration for week-1, week-2, week-3, and week-4, respectively.
- Fig. 9** Scatter plot of forecasted (ET_{for}) with observed ET_o (ET_{obs}) of Menindee station at Week 1, 2, 3, and 4 horizons) with the CNN-GRU model. A least square regression line and coefficient of determination (R^2) with a linear fit equation are shown in each sub-panel. Note: ET_7 , ET_{14} , ET_{21} , and ET_{28} are referred to as reference evapotranspiration for week-1, week-2, week-3, and week-4, respectively.
- Fig. 10** Time series of daily evapotranspiration (ET_o , mm) for observed and forecasted ET_o for the (a) objective model (CNN-GRU), (b) standalone GRU, and (c) classical MLP at week 1 ahead ET_o forecast.
- Fig. 11** Wavelet coherency spectrums between week-1 observed ET_o (ET_7) and forecasted ET_o using CNN-GRU and GRU model with ACO. The arrows indicate the relative phase relationship within the significant zones of higher correlation.
- Fig. 12** The percentage change in $RMAE$ generated by the objective and benchmark models using the ACO algorithm adopted in forecasting ET_o at three sites of Murray Darling Basin. (a) Menindee, (b) Fairfield, (c) Gabo Island, and (d) Gattton at different n^{th} ($n = 1, 2, 3$, and 4) week ahead horizon.

Fig. 5.1 Graphical Abstract of Objective 3

Fig. 1 Topographical structure of LSTM memory cell following and (Olah 2015). Note: [x_t is the inputs, h_t is the next hidden state, h_{t-1} is the last hidden state, \tanh is the hyperbolic tangent function, C_t is the next cell state, and σ is the logistic sigmoid function]

Fig. 2 A detailed workflow is outlining the necessary steps taken in the design of the proposed hybrid BRF-LSTM predictive model for soil moisture (SM) estimation for two future global warming scenarios (*i.e.*, RCP 4.5 & 8.5).

Fig. 3 Map of the present study region showing the selected stations and their geographical location where the proposed hybrid deep learning BRF-LSTM model was validated for monthly SM estimation.

Fig. 4 Box plot of the Z-scores attained by the Boruta feature selection algorithm exploring the relative strength of the predictor variables used for SM estimation, provided for the study site 2 (ST4) as an example. This method was used to determine the most significant input features for the ACCESS 1.3 (RCP 8.5) global climate model.

Fig. 5 The relative performance of soil moisture estimation model at the five study sites for the four global climate model, evaluated according to the normalized Kling–Gupta efficiency for (a) RCP 4.5 and (b) RCP 8.5. Note that KGE = 0 and 1, respectively, for the worst and best performance of the proposed model.

Fig. 6 The caption is identical to Figure 5, except showing the model performance in terms of the normalized mean absolute prediction error (MAPE).

Fig. 7 Comprehensive assessment of the performance of the proposed hybrid deep learning (*i.e.*, BRF-LSTM) against the counterpart models, based on the relative root means square error for the five stations of four GCMs for RCP 4.5 and RCP 8.5 warming scenarios

- Fig. 8** Evaluation of the performance of the proposed hybrid deep learning, BRF-LSTM model with the comparative benchmark models based on the relative mean absolute error for the five stations of four GCMs. The prediction was performed for the testing period for RCP 4.5 and RCP 8.5 warming scenarios.
- Fig. 9** Histogram illustrating the frequency of the absolute estimation errors ($|FE|$) of the proposed hybrid deep learning BRF-LSTM model for RCP 4.5 warming scenarios for five stations.
- Fig. 10** Illustration of the frequency of absolute value of estimation errors ($|EE|$) of the proposed hybrid deep learning BRF-LSTM model for the case of RCP 8.5 warming scenarios at all five study stations.
- Fig. 11** Box plots were constructed to evaluate the discrepancy ratio (i.e., the estimated SM/observed SM) generated by the proposed hybrid deep learning model BRF-LSTM model relative to the benchmark models for RCP 4.5 and RCP 8.5 scenarios.
- Fig. 12** Taylor diagram demonstrating the correlation coefficient, together with the standard deviation difference for the proposed hybrid deep learning BRF-LSTM vs. benchmark models for the case of the RCP4.5 scenario.
- Fig. 13** Caption identical to Figure 12 but for the RCP 8.5 scenario.
- Fig. 14** Scatter plot of the monthly soil moisture of the estimated ('est') vs. observed ('obs') values in the testing phase for the proposed hybrid deep learning BRF-LSTM model for the selected 'best stations' for each GCMs for both global warming scenarios (2005 to 2099).
- Fig. 15** A comparison of the proposed deep learning BRF-LSTM model for SM estimation based on the Legates and McCabe's Index in the testing phase, for the two global warming scenarios. (a) RCP 4.5 and (b) RCP 8.5
- Fig. 16** A comparison of the proposed deep learning BRF-LSTM model for SM estimation based on Wavelet Coherence Spectrums for the two global warming scenarios (i.e., RCP 4.5 and RCP 8.5)

Fig. 6.1 Graphical Abstract of Objective 4

Figure 1 (a) Schematic of the hybrid CEEMDAN-CNN-GRU model with Complete Ensemble Empirical Model Decomposition (CEEMDAN), Convolutional Neural Networks (CNN), and Gated Recurrent Unit (GRU) Neural Network arrangement. The IMF's (Intrinsic Mode Functions) and residual series are generated in the CEEMDAN process, whereas the CNN algorithm represents the feature extraction stage. (b) 2-layered GRU model.

Figure 2 The Australian Murray Darling Basin with study sites & Surface Soil Moisture (SSM kgm^{-2}) where the hybrid CEEMDAN-CNN-GRU model at multi-step daily SSM forecasting

Figure 3 Workflow with the steps in model design for hybrid CEEMDAN-CNN-GRU predictive model. SSM = Surface Soil Moisture, NCA = neighbourhood component analysis for regression, IMF = Intrinsic Mode Function, CEEMDAN = Complete Ensemble Empirical Model Decomposition with adaptive noise, GRU = Gated Recurrent Units

Figure 4 Feature weight matrix of predictor variables from a pool of 52 data sources using neighbourhood component analysis at the n^{th} ($n = 1, 5$, and 30) day lead time forecasting of surface soil moisture shown for the case of Menindee study station.

Figure 5 Stair plot showing the relative root mean squared error ($RRMSE$, %) for (a) CNN-GRU, (b) GRU applied at different input combinations for Menindee station at the 1st, 5th, 7th, 14th, 21st and 30th Day lead time.

Figure 6 Probability plot (95 percentiles) for hybrid CEEMDAN-CNN-GRU, CNN-GRU, CEEMDAN-GRU & GRU. model for Menindee at different n^{th} ($n = 1, 5, 7, 14, 21$ & 30) Day lead time.

Figure 7 Time series of daily surface soil moisture (SSM, $kg m^{-2}$) for observed SSM (Gray) and forecasted SSM for the objective model,

CEEMDAN-CNN-GRU (red) against CNN-GRU (Cyan) and standalone GRU model (Purple) for Menindee station at different n^{th} ($n = 1, 5, 7, 14, 21$ and 30) Day lead times.

Figure 8 Scatter plot of the forecasted and observed SSM. (a) Menindee station, (b) Deniliquin, (c) Fairfield, and (d) Gabo Island at different n^{th} ($n = 1$ and 7) Day ahead. A least square regression line, $y = mx + C$, and coefficient of determination (R^2) are shown in each sub-panel.

Figure 9 Polar plot showing the Legates & McCabe's Index in the testing period computed for the hybrid CEEMDAN-CNN-GRU against comparative models at different n^{th} Day-ahead forecasting of SSM.

Figure 10 Contour plot of (a) KGE, (b) MAPE for hybrid CEEMDAN-CNN-GRU model against comparative models for different n^{th} ($n = 1, 5, 7, 14, 21$ & 30) Day-ahead forecasting of SSM.

Figure 11 Box plot of errors in the testing phase for hybrid CEEMDAN-CNN-GRU against comparative models at different n^{th} ($n = 1, 7$, and 30) Day-ahead lead time forecasting SSM.

Figure 12 The percentage change in *RMAE* generated by the objective and benchmark models using CEEMDAN and CNN methods was adopted in forecasting SSM at four study sites: Murray Darling Basin. (a) Menindee, (b) Deniliquin, (c) Fairfield, (d) Gabo Island at different n^{th} ($n = 1, 5, 7, 14, 21$ & 30) Day-ahead forecasting SSM.

Figure 13 The average forecasted SSM vs. observed SSM on a seasonal basis using hybrid CEEMDAN-CNN-GRU and CNN-GRU models for Menindee at different n^{th} ($n = 1, 5, 7, 14, 21$ & 30) Day-ahead periods. The forecast error ($|FE|$) in each model is plotted on a secondary axis as a line chart.

Chapter 7

Fig. 7.1 Graphical Abstract of Objective 5

Figure 1 (a) The selected river stations of the Australian Murray River System including the flood inundation area, (b) 1-Monthly rainfall totals for MDB (01/12/2020 – 31/12/2020), (c) 6-Monthly rainfall totals for

MDB, (d) 12-hour total rainfall for MDB, (e) 48-hour total rainfall for MDB, (f) The land use in the Murray River basin and (g) Irrigated area within Murray River system.

Figure 2 (a) Structure of bi-directional LSTM Network, (b) Topological structure of feature extraction algorithm (Convolutional Neural Network, CNN) that has been integrated with the objective predictive algorithm (Bi-directional Long Short Term Memory Networks (BiLSTM)) in this study used to construct CBiLSTM hybrid model in the SWL forecasting problem. The forecasting horizon was up to 7-days, 14-days, and 28-days lead time-step.

Figure 3 Schematic structure of the two-phase CVMD-CBiLSTM hybrid model integrating complete ensemble empirical mode decomposition adaptive noise (CEEMDAN) and variational mode decomposition (VMD) with Convolutional Neural Network (CNN) that has been merged with the Bi-directional Long Short Term Memory Networks (BiLSTM).

Figure 4 (a) Correlogram showing the covariance between the objective variable (SWL) and the predictor variables (i.e., $T2X$, $T2M$, sam , $ccot$, $RHmaxT$, GBI) in terms of the Cross-correlation coefficient (r_{cross}) for Howlong stations of Murray River System (b) Cross-correlation coefficient (r_{cross}) for $ccot$ of the Howlong stations decomposed by CEEMDAN (i.e., IMF_n and Residuals) and VMD (i.e., $VMIF_n$ to Res-VMD), (c) Partial autocorrelation function (PACF) plot of the SWL time series and decomposed $ccot$ using CEEMDAN and VMD. The red line in the figures indicates the $\pm 95\%$ confidence level.

Figure 5 Model performance for forecasting daily SWL (m) in terms of RRMSE (%) for Lake Albert, Wakool, and Bringenbrong Bridge with input data from SILO, MODIS and CI.

Figure 6 Box plots of proposed hybrid models (i.e., CVMD-CBiLSTM) along with their respective standalone counterparts (i.e., CBiLSTM, BiLSTM, and SVR) in forecasting SWL in terms of Correlation

Efficient (r) and Root Means Squared Error (RMSE, m) for 19 selected stations at Murray River System.

- Figure 7** Spatial plots of proposed hybrid models (i.e., CVMD-CBiLSTM) along with their respective standalone counterparts (i.e., CBiLSTM, BiLSTM, and SVR) in forecasting SWL in terms of Mean Absolute Percentage Error (MAPE, %) for 19 selected stations at Murray River System
- Figure 8** Comparison of the forecasting skill of proposed models in NSE for the testing period
- Figure 9** Scatter plot of forecasted vs. observed SWL of a) Lake Albert and b) Tocumwal sites using the proposed hybrid model and comparing models. A least square regression line and coefficient of determination (R^2) with a linear fit equation are shown in each sub-panel.
- Figure 10** Tylor diagram representing correlation coefficient together with the standard deviation difference for proposed hybrid CVMD-CBiLSTM vs. benchmark models for (a) Lake Albert (b) Wakool (c) Tocumwal, and (d) Howlong.
- Figure 11** Empirical Cumulative Distribution function (CDF) of forecasted error $|FE|$ of SWL generated by the proposed CVMD-CBiLSTM vs. benchmark models for (a) Lake Albert (b) Wakool (c) Tocumwal and (d) Howlong
- Figure 12** Comparison plots of proposed hybrid models (i.e., CVMD-CBiLSTM) vs. standalone machine learning model (i.e., SVR) in forecasting SWL in terms of RRMSE (%) for 7-Days, 14-Days and 28-Days ahead SWL forecasting
- Figure 13** Heat map showing the normalized LM index with the proposed hybrid deep learning approach (i.e., CVMD-CBiLSTM) compared to standalone models for 19 selected stations of Murray River Basin

Figure 14 Comparison between Forecasted and Observed SWL during model testing using CVMD-CBiLSTM model for 7-Day, 21-Day, and 28-Day ahead forecasting

Appendix A

Fig. A1 Graphical Abstract of Article 6

Fig. 1 Geographic location of our study site: *Columboola solar energy farm in Queensland Australia*, where the proposed kernel ridge regression (KRR)-based machine learning model (ML) model for bias correction of total cloud cover (TCDC) was developed utilizing Global Forecast System (GFS) analysis (*i.e.*, proxy observed) and forecasted variables.

Fig. 2 Schematic of the proposed KRR-based bias correction method that is benchmarked with the conventional (*i.e.*, multivariate recursive nesting bias correction, MRNBC) and nine ML (*i.e.*, Bayesian ridge regression (BNR), Decision Tree (DTR), Gradient Boosting Regressor (GBR), Hist Gradient Boosting Regressor (HGBR), *k*-nearest regression (KNN), multivariate adaptive regression splines (MARS), extreme gradient boosting (XGB), and random forest (RF) methods adopted to correct the bias in total cloud cover

Fig. 3 Schematic illustration of the 3-h GFS forecasts initialized at 0000 UTC compared with Australian Eastern Standard Time used to develop KRR bias correction method.

Fig. 4 Schematic of the traditional method, *i.e.*, multivariate recursive nested bias correction (MRNBC) presented in this study as a comparison method against the proposed KRR bias correction method used to correct bias in total cloud cover (TCDC)

Fig. 5 Box plots of Willmott's Index of Agreement (*d*) were calculated for all nine ML-bias corrections models (*i.e.*, KRR, BNR, DTR, GBR, HGBR, KNN, MARS, RF, XGB) pooled together, including conventional bias correction (*i.e.*, MRNBC) and their respective

reference values (d calculated between $TCDC_{GFS-Forecast}$ and $TCDC_{GFS-Analysis}$) for (a) Approach 1, & (b) Approach 2

Fig. 6 Box plots of bias-corrected root mean square error (RMSE) calculated between data for all the nine ML-based bias correction methods pooled together (i.e., KRR, BNR, DTR, GBR, HGBR, KNN, MARS, RF, XGB), conventional bias correction method (i.e., MRNBC) & along with their respective reference values ($RMSE$ calculated between $TCDC_{GFS-Forecast}$ and $TCDC_{GFS-Analysis}$). (a) Approach 1 and (b) Approach 2.

Fig. 7 Comparative analysis of *four* selected ML-based bias correction methods (i.e., KRR, MARS, KNN, RF) by means of correlation coefficient (r) between the **corrected** $TCDC_{GFS-Forecasts}$ and the reference $TCDC_{GFS-Analysis}$. Included is a respective reference r -value computed using ‘non-corrected’ $TCDC_{GFS-forecasts}$ and bias-corrected $TCDC_{GFS-Forecasts}$ but using a traditional method (i.e., MRNBC). (a) Approach 1, and (b) Approach 2.

Fig. 8 Change (∇) in mean absolute percentage error, $MAPD$ (%) generated by proposed KRR bias correction method with respect to a reference value of $MAPD$ deducted from $TCDC_{GFS-Forecast}$ and $TCDC_{GFS-Analysis}$. (a) Approach 1, and (b) Approach 2.

Fig. 9 The percentage change in Legates & McCabe’s Index (LM) was deduced by comparing the LM values obtained using the proposed KRR-bias correction model in respect to the LM values generated by KNN, MARS, and RF Models. (a) Approach-1, (b) Approach-2.

Fig. 10 Taylor diagram showing the correlation coefficient, standard deviation, and root mean square centered difference (RMSD). (a) The objective model (KRR) compared with (b) KNN, (c) MARS, and (d) RF) for the most accurate approach (i.e., Approach-2).

- Fig. B1** Graphical Abstract of Article 7
- Figure 1** Integrated workflow showing the study area and atmospheric domain of South Australia with a schematic structure of KRR model integrating with GWO and CEEMDAN methods for the proposed GWO-CEEMDAN-KRR model for wheat yield prediction
- Figure 2** Flowchart of the grey wolf optimization (GWO) algorithm
- Figure 3** Comparison of the predictive skill of the proposed wheat yield prediction models in terms of the relative error: RMAE (%) and the correlation of determination (R^2) within the testing period.
- Figure 4** An assessment of four distinct feature selection methods regarding the percentage change in relative error (i.e., RMAE) and relative index of agreement (d_{rel}) with all methods using a CEEMDAN data decomposition approach the model's testing phase
- Figure 5** The prompting percentage (Δ) for correlation coefficient (ΔR), RMAE ($\Delta RMAE$), and NRMSE ($\Delta NRMSE$) between the proposed GWO-CEEMDAN-KRR model, other ACO, ASO, PSO used models, and the standalone models.
- Figure 6** Scatter plot of the predicted and observed Y generated by proposed GWO-CEEMDAN-KRR model vs. the other models. A least square regression line, $Y = mX + C$, and the coefficient of determination (R^2) are shown in each sub-panel.
- Figure 7** The discrepancy ratio (i.e., the predicted Y/ observed Y) generated by the proposed hybrid CEEMDAN-KRR model using the four optimization algorithms and their respective standalone counterparts.
- Figure 8** (a) An empirical cumulative distribution function (ECDF) plot of $|FE|$ and (b) Taylor diagram demonstrating the correlation coefficient, together with the standard deviation difference of the hybrid KRR model and standalone KRR with four optimization algorithms (i.e., GWO, ASO, ACO, and PSO)

LIST OF ACRONYMS

ACF	Autocorrelation Function
ACO	Ant Colony Optimization
ADF	Augmented Dickey-Fuller
AEZ	Australian Agro-ecological Zones
ANN	Artificial Neural Network
AO	Arctic Oscillation
AR5	Fifth Assessment Report
BiLSTM	Bi-directional Long- short term memory
BOM	Australian Bureau of Meteorology
BRF-LSTM	Hybrid Model integrating the Boruta feature selection with LSTM
BRF-GRU	Hybrid Model integrating the Boruta feature selection with GRU
BRF-RNN	Hybrid Model integrating the Boruta feature selection with RNN
BRF-SVR	Hybrid Model integrating the Boruta feature selection with SVR
BRF-MARS	Two-stage Hybrid Model integrating the Boruta feature selection algorithm and significant lagged memory with MARS
CVMD-CBiLSTM	Hybrid Model integrating the CEEMDAN, VMD, CNN with BiLSTM
CBiLSTM	Hybrid Model integrating the CNN with BiLSTM
CCF	Cross Correction Function
CEEMDAN	Complete Empirical Ensemble Mode Decomposition with Adaptive Noise
CEDA	Centre for Environmental Data Analysis
CI	Climate Mode Indices
CMIP5	Coupled Model Inter-comparison Project Phase 5
CMIP3	Coupled Model Inter-comparison Project Phase 3
CNN	Convolutional Neural Network
CNN-LSTM	Hybrid Model integrating the CNN with LSTM
CNN-GRU	Hybrid Model integrating the CNN with GRU

CSIRO	Commonwealth Scientific and Industrial Research Organization
DBN	Deep Belief Network
DL	Deep Learning
DMI	Dipole Model Index
DTR	Decision Tree
ECDF	Empirical Cumulative Distribution Function
ELM	Extreme Learning Machine
EMI	El-Nino southern oscillation Modoki index
ENSO	El Niño Southern Oscillation
EPI	East Pole Index
ET₀	Evapotranspiration
FCM	Fuzzy C-means
FE	Forecasting Error
FFNN	Feed Forward Neural Networks
GCM	Global Climate Models
GIOVANNI	Geospatial Online Interactive Visualization & Analysis Infrastructure
GLDAS	Global Land Data Assimilation System
GRU	Gated Recurrent Unit
IOD	Indian Ocean Dipole
IPO	Interdecadal Pacific Oscillation
IPCC	Intergovernmental Panel on the Climate Change
KNMI	Royal Netherlands Meteorological Institute
LM	Legates-McCabe's Index
LSTM	Long- short term memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MARS	Multivariate Adaptive Regression Splines
MDB	Murray-Darling Basin
MODIS	Moderate Resolution Imaging Spectroradiometer
MOHC	Met Office Hadley Centre
MRA	Multi-Resolution Analysis

MSE	Mean Squared Error
NAO	North Atlantic Oscillation
NOAA	National Oceanic and Atmospheric Administration
NSE	Nash–Sutcliffe Efficiency
NSW	New South Wales
PACF	Partial Autocorrelation Function
PDO	Pacific Decadal Oscillation
QLD	Queensland
r	Correlation Coefficient
RMSE	Root-Mean-Square-Error
RNN	Recurrent Neural Network
RRMSE	Relative Root-Mean-Square Error
SD	Standard Deviation
SAM	Southern Annular Mode
SILO	Scientific Information for Landowners
SM	Soil Moisture
SOI	Southern Oscillation Index
SSM	Surface soil moisture
SST	Sea Surface Temperature
STR	Subtropical Ridge
SVR	Support Vector Regression
SWL	Stream Water Level
TPI	Tri-pole Index
UK	United Kingdom
VMD	Variation Mode Decomposition
WI	Willmott’s Index of Agreement
WPI	West Pole Index

HYBRID MODELS NOTATIONS

Chapter 3 (Published Article 1)

BRF-LSTM	Two-stage Hybrid Model integrating the Boruta feature selection algorithm and significant lagged memory with LSTM
BRF-GRU	Two-stage Hybrid Model integrating the Boruta feature selection algorithm and significant lagged memory with GRU
BRF-SVR	Two-stage Hybrid Model integrating the Boruta feature selection algorithm and significant lagged memory with SVR
BRF-MARS	Two-stage Hybrid Model integrating the Boruta feature selection algorithm and significant lagged memory with MARS

Chapter 4 (Published Article 2)

CNN-GRU	Two-stage Hybrid Model integrating the ACO and significant lagged memory with CNN and GRU
CNN-LSTM	Two-stage Hybrid Model integrating the ACO and significant lagged memory with CNN and GRU

Chapter 5 (Published Article 3)

BRF-LSTM	Two-stage Hybrid Model integrating the Boruta feature selection algorithm and significant lagged memory with LSTM
BRF-SVR	Two-stage Hybrid Model integrating the Boruta feature selection algorithm and significant lagged memory with SVR
BRF-MARS	Two-stage Hybrid Model integrating the Boruta feature selection algorithm and significant lagged memory with MARS

Chapter 6 (Published Article 4)

CEEMDAN-CNN-GRU	Hybrid Model integrating the NCA, CEEMDAN, and CNN algorithm with GRU
CEEMDAN-GRU	Hybrid Model integrating the NCA, CEEMDAN algorithm with GRU
CNN-GRU	Hybrid Model integrating the NCA and CNN algorithm with GRU

Chapter 7 (Article Under Review 5)

CVMD-CBiLSTM	Hybrid Model integrating the ACO, CEEMDAN, VMD, CNN with BiLSTM
CBiLSTM	Hybrid Model integrating the CNN with BiLSTM

CHAPTER 1: INTRODUCTION

1.1 Background

Increased agricultural and industrial activity and water-based recreation increase demand for water resources. As a result of anthropogenic influences, the distribution and accessibility of this valuable and rare resource adversely impacts the environment. In order to avoid any potential catastrophes, water resource management strategies that are prudent and effective are required to meet the expanding demand and inconsistent supply. Additionally, changing weather patterns and climate due to anthropogenic influences harm the distribution and accessibility of this scarce and essential resource. Due to the rising demand and inconsistent supply, a sustainable water resource management system must be developed to avoid potential calamities.

Increased susceptibility to limited water resources has resulted from recent variability in long-term climate (e.g., seasonal variations) and short-term (weather) patterns due to natural and manufactured influences. The chaotic behaviour of climatological occurrences results in non-linearity and non-stationarity in hydrological phenomena. The impacts of extreme weather experiences such as excessive rainfall, droughts, and hail, as well as heatwaves and extreme temperatures, have frequently had a substantial impact on agricultural productivity and water resource management. As a result, improvements in the understanding of climate risk on water resources are required urgently to minimise the climate-related impact and assist agricultural and water resource managers in developing and implementing strategies to avoid any possible catastrophe.

The most significant changes in the climatic conditions of water-dependent ecosystems due to climate change are likely to involve hydrological regimes (Barron et al. 2012). Climate change also impacts drivers such as land-use change, dams and other hydrological changes, and water extraction for consumptive use, all of which affect the aquatic environment. Rather than direct precipitation, terrestrial water reservoirs primarily control agricultural, hydrological, ecological, and interconnected socio-economic systems (Van Loon & Laaha 2015). In particular, streamflow water level (SWL) and soil moisture (SM) are two prime components of terrestrial water

reservoirs. The functioning of ecological and hydrological systems depends on soil moisture (Prasad et al. 2019), which plays an essential role in plant growth and maintenance and is linked with the water cycle of soil-plant-atmosphere systems (Cai et al. 2019). A continued lack of soil moisture, combined with a lack of adequate planning strategies, may significantly impact agricultural and hydro-meteorological processes (Zaman & McKee 2014; Prasad et al. 2018a, 2019). In addition, streamflow is a collection of surface runoff from a catchment or basin, a crucial driver of soil water retention, infiltration, and evaporation. Projected climate changes significantly impacts streamflow, a vital element of the hydrological cycle (Alaoui et al. 2014; Abera et al. 2019). Changes in rainfall patterns, temperature, and evaporation indirectly affect streamflow (Guo et al. 2020). Watershed management is crucial to managing the consequences of climate change and creating effective adaptation measures. Interannual fluctuations in streamflow (McMahon et al. 1992; Deo & Sahin 2016) pose irrigation, marine life, and ecosystem management (Verdon & Franks 2005).

Hydrologic drought studies rely heavily on streamflow data, whereas agricultural drought depends on SM levels. Hence, SWL and SM are essential to managing this limited resource properly. An early warning system for drought can be developed by forecasting SWL and SM to understand future water resource availability. Moreover, reference crop evapotranspiration (ET_o) plays a vital role in agriculture, ecosystems, and ecological modelling (Feng et al. 2017). It connects the atmospheric and surface water flows by transporting water vapour to the atmosphere. A thorough understanding of the hydrological cycle dynamics is necessary to monitor ET_o stochasticity and improve sustainable freshwater use (Zeng et al. 2019). Knowledge-based intelligent systems for monitoring soil moisture, evapotranspiration, and streamflow water level could benefit from optimal water distribution and exploitation for domestic, industrial, agricultural, hydroelectricity generating, and recreation applications.

Recent developments in processing capability have made it possible to apply machine learning-based predictive models in a wide range of fields, including energy (Ghimire et al. 2019a; Kong et al. 2020; Rajagukguk et al. 2020), medicine (Yuan et al. 2020), and hydrology (Deo et al. 2015; Deo & Sahin 2016; Ali et al. 2018). While data-driven models capture relevant predicting elements from past data sets,

traditional models do not (Ali et al. 2019). However, much of the research employed standalone data-driven models with specific simplification limitations due to complicated inputs with stochastic features generated by highly interrelated meteorological and hydrological parameters (Adamowski et al. 2012). The intricate and non-linear interactions between the predictors cause overfitting for big datasets (Zhang, W. et al. 2017). To overcome the limitations of standalone models, deep learning (DL) was employed (Li et al. 2007), which efficiently extracted compound relationships from data (Ghimire et al. 2019a). Given the importance of predicting climatological, hydrological, and agricultural sustainability, this is an area of investigation that is still being explored. Consequently, in this study, new and advanced deep learning-based predictive models, hybridised with different feature selection and feature decomposition approaches, are being investigated for the purpose of forecasting soil moisture, streamflow water level, and evapotranspiration in the Murray-Darling Basin, the Australian agricultural hub. The result of the study brings significant contributions to flood forecasting and irrigation scheduling for the sustainable management of water resources in Australia.

1.2 Statement of the problem

Australia is the driest inhabited continent on the planet, with one of the most inhospitable climates on the Earth (Ummenhofer et al. 2009). Rainfall records show regular drought cycles that can last years or even decades, and these cycles are alternated by years of above-average rainfall. Climate change has significantly impacted Australia's regional water availability and ecosystem health (CSIRO 2016). Hydrological abnormalities, such as frequent and long-lasting droughts, are a common feature of the Murray-Darling Basin (MDB), the focused area of this study (Deo et al. 2009; McAlpine et al. 2009). Three major drought events, such as the Federation drought (1895–1902), World War II (1937– 1945), and the Millennium drought event (1997–2009), were all reported in the MDB region of Australia (Ummenhofer et al. 2009; Deo & Şahin 2015). The most recent occurrence, the Millennium drought, together with the ever-changing weather patterns, impacts water resources, water management, and distribution policy. As a result, in the future high-emission scenario, seasonal changes of hydrological variables (i.e., SM) show a significant decrease,

primarily in the winter and spring seasons, while the annual-mean decrease by 10% in the MDB regions (Timbal et al. 2015). Moreover, the droughts and reduced rainfall have also resulted in a drop in agricultural yields (van Dijk et al. 2013).

Even though climate change estimates imply that the MDB will become drier in the future, there is a huge transaction of uncertainty in the future rainfall projections, which makes the future runoff projections complex. Using the medium warming scenario, recent hydrological modelling studies reveal that the southern MDB will see a median predicted drop in mean annual runoff of 14% by 2046–75 (10–90 percentile range: -38 % to +8 %) (Whetton & Chiew 2021). The current era of a changing and highly unpredictable climate necessitates proactive and wise planning of sustainable water management procedures to meet the increasing need for water for agricultural and residential purposes, given limited water supplies and increasing demand.

Experts use various methods to anticipate hydrological variables, including soil moisture, evapotranspiration, and streamflow, by utilising multiple models, namely physical-dynamical, statistical, and artificial intelligence approaches. Physical models are the mainstream model to quantify the hydrological variables, but they require many variables to validate the model (Li et al. 2017; Prasad et al. 2019). However, the non-linearities that arise between the hydrological variables resulting from hydro-physical interactions occurring at large scales can further conceal forecasting capability, making the model more laborious to use for forecasting purposes (Deo & Sahin 2016; Yaseen et al. 2016). As an alternative, data-driven models are intended to find the association between the predictors and target variables without considering the fundamental operations of hydrological systems (Kisi & Parmar 2016; Yaseen et al. 2016).

The use of data-driven models in predicting hydrological variables (Deo & Sahin 2016; Prasad et al. 2019) is becoming increasingly promising. The application of multiple data-driven models to forecast stream water level, evapotranspiration, and soil moisture significantly have been investigated in the previous decade, including adaptive neuro-fuzzy inference system (Ehteram et al. 2019), support vector machine (Bafitlhile & Li 2019), and extreme learning machine (Deo & Sahin 2016; Yaseen et al. 2019). However, many studies have used standalone machine learning models, which have specific limitations in simplification competencies due to complex inputs

with deterministic characteristics induced by densely interrelated climatic and hydrological factors to address the vital temporal and seasonal patterns (Adamowski et al. 2012). Moreover, many standalone machine learning techniques often overfit when used on large datasets (Zhang, Y. et al. 2017) because of the intricate and non-linear relationships between the predictors and the target variable. Deep learning (DL) models, on the other hand, have the potential to overcome the constraints of standalone models (Li et al. 2017). It has been proven to be more accurate than other methods. As a result, deep learning algorithms can potentially be incorporated in predicting stream water levels.

Several multiple feature extraction layers are employed by deep learning (DL) models, which allow them to efficiently capture compound relationships within the predictor variables. These deep learning algorithms have been effectively applied in a variety of applications. The long short-term memory (LSTM) model was successful in hydrology and water resources (Zhang et al. 2018) as a proven and feasible forecasting strategy. The LSTM can extract the relative extrapolative features from the historical data. Ahmed et al. (2021) used a long short term memory (LSTM) and gated recurrent unit (GRU) model to forecast monthly stream water levels to avoid the issues associated with overfitting the inputs. According to the authors, the target variables (i.e., SWL) and model achieved better generalisation than the standalone machine learning models (e.g., RNN and SVR). Thus, DL approaches outperform the conventional machine learning model (Li et al. 2021), and need to be implemented in forecasting hydrological variables in Australia's Murray Darling Basin for better management of water resources.

Moreover, the hydrological variables are complex time series superimposed by multi-scale regulations, making them difficult to understand. It does not consider the trend of actual hydrological data and its periodicity and randomness, putting classical artificial intelligence prediction algorithms at a competitive disadvantage. Considering the complex and stochastic behaviour of the hydrological variables, a hybrid deep learning approach incorporating feature extraction (i.e., convolutional neural network, CNN) and multi-resolution analysis (MRA) is recommended to enhance the forecasting skill. A complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) is an improved version of ensemble empirical mode

decomposition (EEMD), and empirical mode decomposition (EMD) showed improved performance in forecasting problems. Previous investigations employed CEEMDAN to predict soil moisture (Prasad et al. 2018b), and an earlier version of the model (i.e., EEMD) was used to predict stream-flow (Seo & Kim 2016), wave height (Raj & Brown 2021) and rainfall (Ouyang et al. 2016) with improved performance. Moreover, the complex relations between the input features can also be extracted by combining two decomposition methods, eventually increasing the forecasting ability (Peng et al. 2017; Prasad et al. 2020; Sibtain et al. 2020). Thus, decomposition techniques can increase the forecasting performance of hydrological variables (i.e., SM, ET_o , and SWL) in Australia's Murray-Darling Basin.

Overall, the purpose of this doctoral study is to address the concerns of appropriate input selection, non-linearity, and non-stationarity of the predictor variables in the context of forecasting soil moisture, evapotranspiration, and streamflow water level in the Murray-Darling Basin of Australia. In addition, the multivariate sequential CEEMDAN and VMD approaches and feature selection methods (i.e., Boruta feature selection, neighbourhood component analysis (NCA) for regression and ant colony optimization) is used to investigate a unique ensemble forecasting technique incorporated with deep learning predictive models (i.e., LSTM, BiLSTM, GRU, and CNN-LSTM).

1.3 Objectives

The primary aim of this doctoral work, presented as a collection of peer-reviewed papers, was to develop hybrid deep learning predictive models for hydrological applications, focusing on predicting soil moisture, evapotranspiration, and streamflow water levels within Australia's Murray-Darling Basin. Therefore, this doctoral thesis, presented as a collection of Quartile 1 (Q1) papers, has adopted an artificial intelligence modelling strategy to achieve five distinct objectives. The research has developed, applied, and assessed high accuracy forecast models for hydrological variables.

The objectives of this study were designed to enable governments, water resource managers, and farmers to utilise these tools for future streamflow water levels and soil moisture predictions that can help make strategic decisions for agriculture and

water management in a highly variable and changing climate. The study aimed to build forecast models over medium (monthly) to short-term (daily) frames. The research is expected to provide accurate and reliable forecasting models to explore the hydrological response to climate change and variability at multiple forecast horizons. Overall, the modelling framework and the relevant data intelligent models aim to achieve the following specific objectives.

Objective 1: Develop Hybrid Deep Learning Models for Feature Selection at Monthly Streamflow Forecasting

1. To develop a hybrid version of the long short-term memory network (LSTM) and gated recurrent unit (GRU) models that adopt the Boruta-Random forest feature selection algorithm and forecast streamflow water levels at monthly horizons. The research has incorporated the synoptic-scale climate mode indices as predictor variables to demonstrate the importance of atmospheric teleconnections in streamflow predictions. The preciseness of the developed hybrid model has been validated using the respective standalone counterpart models.

The outcomes of this objective have been published in the *Journal of Hydrology* ranked Q1 (Vol. 599, 2021, 126350).

Objective 2: Develop Hybrid Deep Learning Models for Weekly Evapotranspiration Forecasting

2. To develop a CNN-GRU hybrid model to forecast reference evapotranspiration (ET_o) at four weekly steps. This main objective is to build a hybrid predictive model with ant colony optimisation (ACO) in forecasting ET_o . The ant colony optimization, a feature selection incorporated with deep learning hybrid model (CNN-GRU), was tested for its precision in simulating reference evapotranspiration employing fifty-two potential inputs. The novel CNN-GRU model with ACO outperforms other benchmark models over various time horizons apprehends the complex and non-linear relationships between predictor variables and the daily ET_o .

The outcomes of this objective have been published in the *Stochastic Environmental Research and Risk Assessment* ranked **Q1** (2021, 1-19).

**Objective 3: Develop Hybrid Deep Learning Model for Soil Moisture
Forecasting Under Global Warming Scenarios**

3. To formulate a new hybrid long short-term memory (LSTM) predictive framework that can emulate the monthly moisture in an upper portion of the soil column (SM) under global warming scenarios based on representative concentration pathways RCP4.5 and RCP8.5 CO₂ emissions. This objective has integrated Boruta-Random forest (BRF) feature selection to capture the significant antecedent memory of SM behaviour to estimate future SM between 2006 and 2100 using the Coupled Model Intercomparison Phase-5 (CMIP5) repository of four global climate models (GCM).

The outcomes of this objective have been published in *Stochastic Environmental Research and Risk Assessment* ranked **Q1** (Vol: **35**, pages: 1851–1881).

**Objective 4: Develop Hybrid Deep Learning Models for Daily Soil
Moisture Forecasting Over Multi Horizons**

4. To develop a deep learning hybrid strategy for daily surface soil moisture forecasting over multi horizons. In this objective, the model has combined a feature selection algorithm using neighbourhood component analysis for regression and a feature decomposition CEEMDAN approach to generate the CNN-GRU predictive model. The CEEMDAN-CNN-GRU hybrid model is tested over the 1st, 5th, 7th, 14th, 21st, and 30th day ahead period by assimilating a large pool of 52 predictors obtained from three distinct data sources (satellite-derived data, ground-based variables from Scientific Information Landowners SILO, and synoptic-scale climate indices) to establish the model's viability for forecasting at multi-step daily horizons

The outcomes of this objective have been published in *Remote Sensing* ranked **Q1** (Vol. 13, Pages 554).

Objective 5: Develop Hybrid Deep Learning Model for Streamflow Water Level Forecasting at 7-day, 14-day, and 28-day Horizons for Medium-Term Decisions

5. To develop a deep learning hybrid model for streamflow water level (SWL) forecasting by convolutional neural networks (CNN), bi-directional long-short term memory (BiLSTM), and ant colony optimization (ACO) methods along with a two-phase decomposition technique. The proposed CVMD-CBiLSTM model combines the CBiLSTM model with a complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and variational mode decomposition (VMD) to extract the significant features of the predictor variables. The model is tested at three forecasting horizons of 7-days, 14-days, and 28-days with potential applications in medium-term decision making. This goal discovered that the CVMD-CBiLSTM model outperformed the CBiLSTM and standalone BiLSTM and SVR, allowing better water resource management.

The outcomes of this objective have been submitted to *Science of Total Environment* ranked Q1 2022.

1.4 Significance of the Research

The outcomes of this research are highly significant as the Murray-Darling Basin is the most productive agricultural area in Australia. The basin is considered the agricultural hub of Australia, which covers almost 67% of agricultural land (Australian Bureau of Statistics 2010), and is 14% of the mainland of the country. The drainage area of the basin is one of the world's largest and the largest on the continent (AIDR 2021). The region is semi-arid, and the variability in the quantity and timing of streamflow is extreme and unpredictable. The temperature varies from 15°C to 28°C from winter to summer. The significant connotation of the Inter-decadal Pacific Oscillation through the Pacific Ocean influences the ENSO phenomena, which impact the drought events and can eventually affect the streamflow and precipitation of the region.

The long-term variations in the hydrological scenarios affect the frequency of flooding events, and changes in the temporal pattern of hydrological variables in the

region of streamflow can hamper irrigation scheduling and drought management. Understanding the features of the hydrological phenomena of the Murray-Darling Basin would help plan the water resources effectively. The developed hybrid artificial intelligence models are vital for policymakers and governments for better future planning concerning trade, development policies, and mitigating climate extreme events. The incorporation of feature selection, decomposition, and extraction methods reduce the irrelevant features from the predictors, producing a precise forecast of the hydrological phenomena in the Murray-Darling Basin. Moreover, these modelling strategies can provide timely information for rapid decision-making during agricultural production. The outcomes of this study will help farmers and decision-makers optimise the hydrological parameters for better flood forecasting and irrigation scheduling for sustainable management of water resources in Australia, which would subsequently impact socio-economic, environmental, and agricultural development.

1.3 Thesis layout

The layout of the thesis is depicted schematically in Fig 1.1, and the overview of the thesis is illustrated in Fig 1.2. It clearly outlines the graphical abstract for comprehension, and the need for a dependable and precise forecasting tool for soil moisture and stream flows level and reference evapotranspiration. This thesis is divided into eight chapters, which are as follows:

- Chapter 1** Provides the introductory background, the problem statement about the research, and the objectives of this study
- Chapter 2** Introduces the study area, data, and general methodology of this study. This chapter provides general viewpoints while the specific study area, data, and methods are presented in the respective chapters.
- Chapter 3** This chapter is provided as a published journal article in the **Journal of Hydrology**. This study is devoted to incorporating a feature selection (i.e., BRF) based LSTM modelling approach to forecasting the streamflow water level at the monthly forecast horizon. Chapter 3 addresses the first objective of this study.
- Chapter 4** This chapter is provided as a published journal article in the *Stochastic Environmental Research and Risk Assessment* journal. It describes a

hybrid CNN-GRU application with ant colony optimisation (ACO) for early warning evapotranspiration, forecasting four weekly steps using deep hybrid learning. The third research objective of this study is described in detail in Chapter 4.

Chapter 5 This chapter is provided as a published journal article in the *Stochastic Environmental Research and Risk Assessment* journal. It addresses a hybrid LSTM predictive framework coupled with Boruta-random forest (BRF) feature selection to emulate SM under global warming scenarios. The Coupled Model Intercomparison Phase-5 (CMIP5) repository estimated future SM. The fourth research objective of this study is described in detail in Chapter 5.

Chapter 6 This chapter is provided as a published journal article in the Journal of *Remote Sensing*. This chapter describes the application of the updated CEEMDAN model with an NCA-based feature selection approach in the CNN-GRU model in forecasting soil moisture. It describes the model development process, and the outcomes are compared to comparative models (CEEMDAN-GRU, CNN-GRU, and GRU).

Chapter 7 This chapter is provided as a published journal article in the *Science of Total Environment*. A newly proposed double decomposed model, CVMD-CBiLSTM, coupled with a complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and variational mode decomposition (VMD) to extract the significant features of the predictor variables. This chapter describes the novel techniques to mimic the future scenarios of streamflow water level

Chapter 8 This chapter provides the synthesis of the study with concluding remarks, limitations, and recommendations for future works.

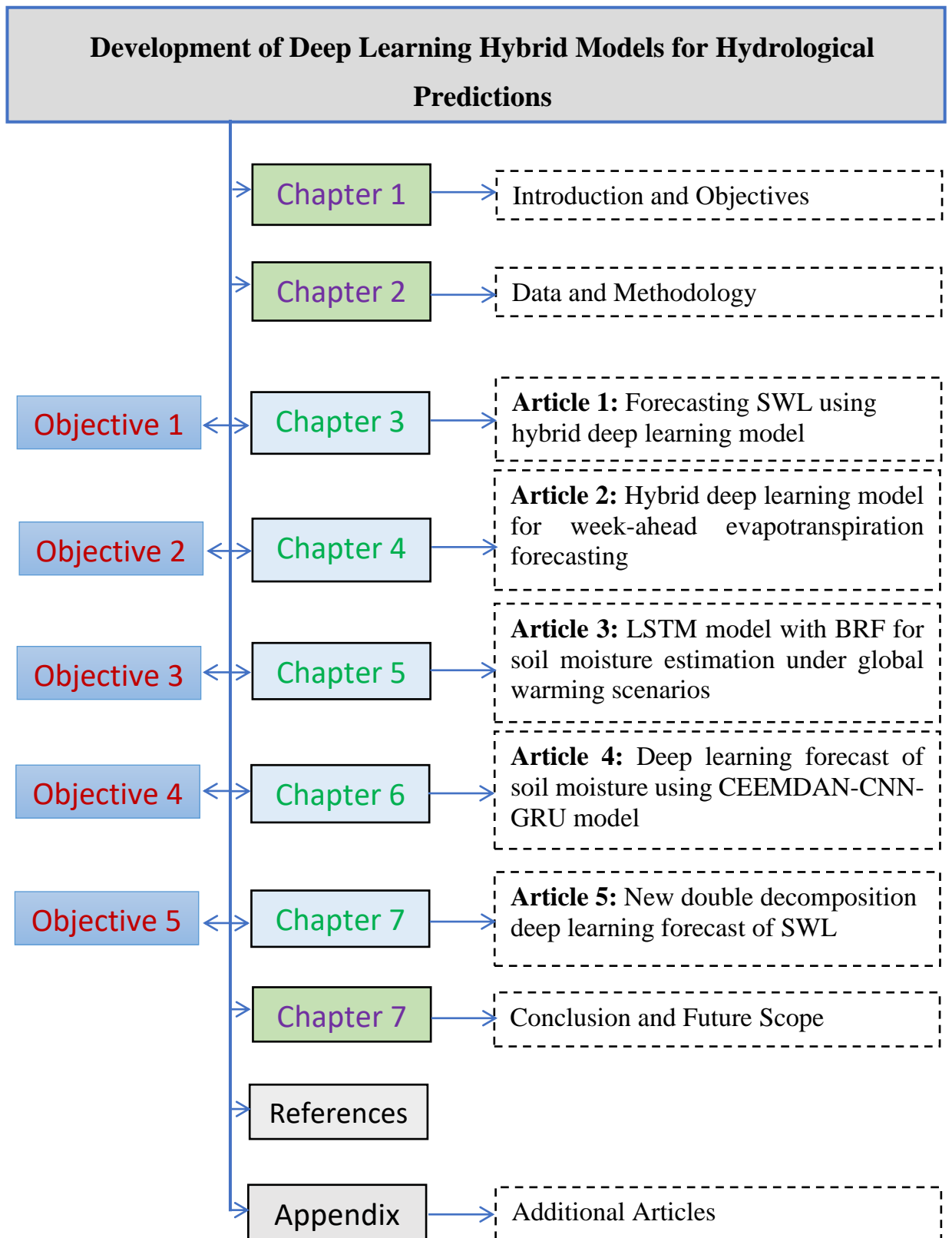


Fig 1.1: Layout of the thesis

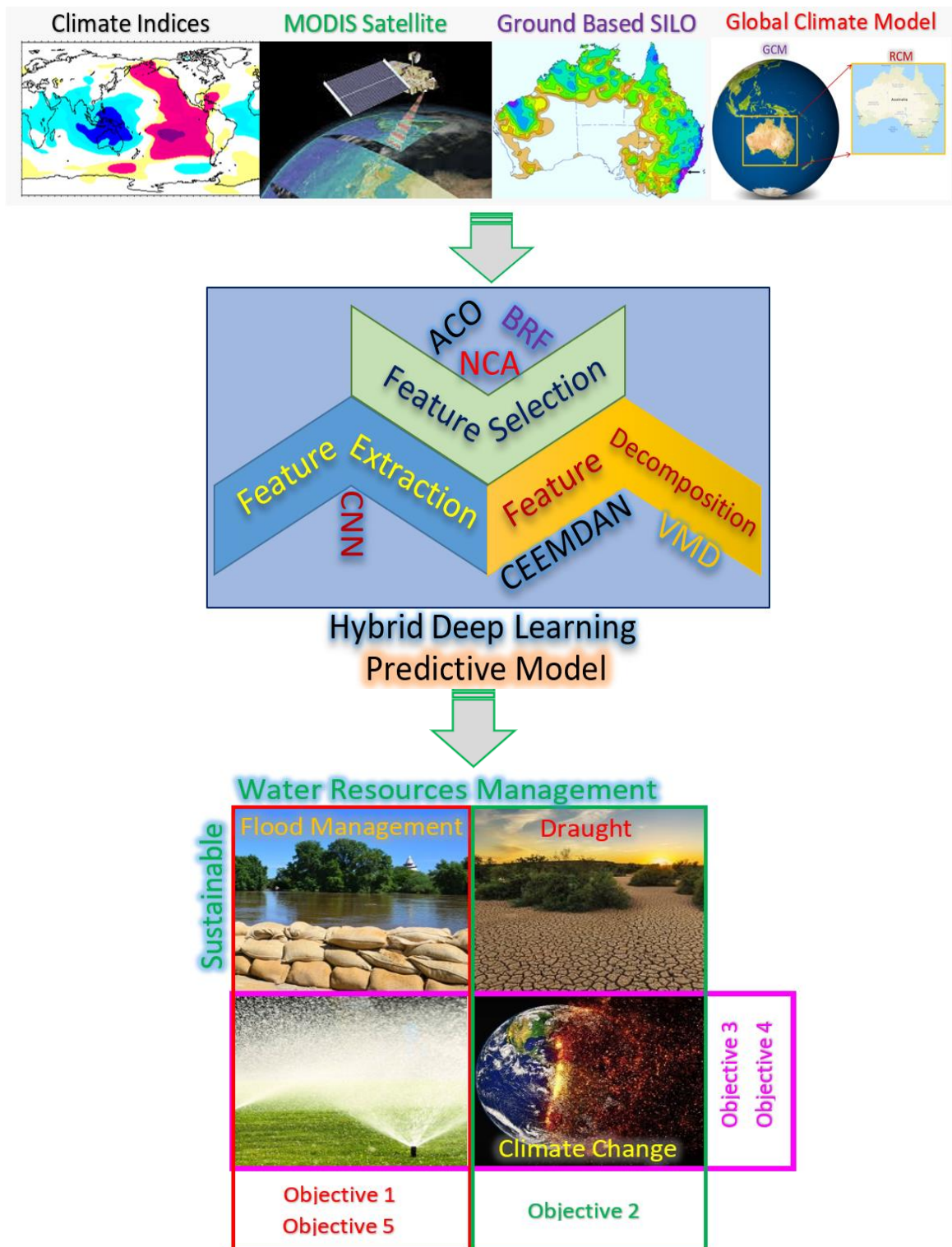


Fig 1.2 Schematic view of thesis

CHAPTER 2: DATA AND METHODOLOGY

This chapter provides an overview of the location of the study sites in developing the deep learning hybrid forecasting models. Different study sites within the study region were selected to achieve each objective, described in detail in each chapter. The description of data used, length of data, and limitations, if any, are also presented. This chapter also introduces a brief account of the methodology, while specific model development techniques have been described in respective chapters.

2.1 Study Area: the Murray-Darling Basin

The study focused on the rich agricultural zones of Australia, the Murray Darling Basin (MDB). The Murray-Darling drainage area is one of the world's largest and the largest on the continent (AIDR 2021). In Australia, the MDB comprises the Murray River and Darling River catchments and is referred to as the country's agricultural heartland (Prasad et al. 2019). Approximately 67% of Australia's agricultural land is covered by this MDB basin agricultural land (Australian Bureau of Statistics 2010), the most productive agricultural region in the country.

Located in southeast Australia, the Murray-Darling Basin encompasses more than one million km² (Beare & Heaney 2002) and accounts for approximately 14 % of the country's total landmass (Leblanc et al. 2012). With a combined length of nearly 7000 km, Australia's largest rivers - the Darling, Murray, and Murrumbidgee - are all situated within the basin catchment (Beare & Heaney 2002). As a result of the Murray Darling Basin's vast size, it has a highly diverse range of climatic conditions and natural environments. Although the MDB is dominated by extensive dryland agriculture, irrigated agriculture accounts for nearly 75% of all agricultural production in the MDB, where water demand and levels of water extraction from rivers and groundwater have reached at unsustainable levels (Chartres & Williams 2006).

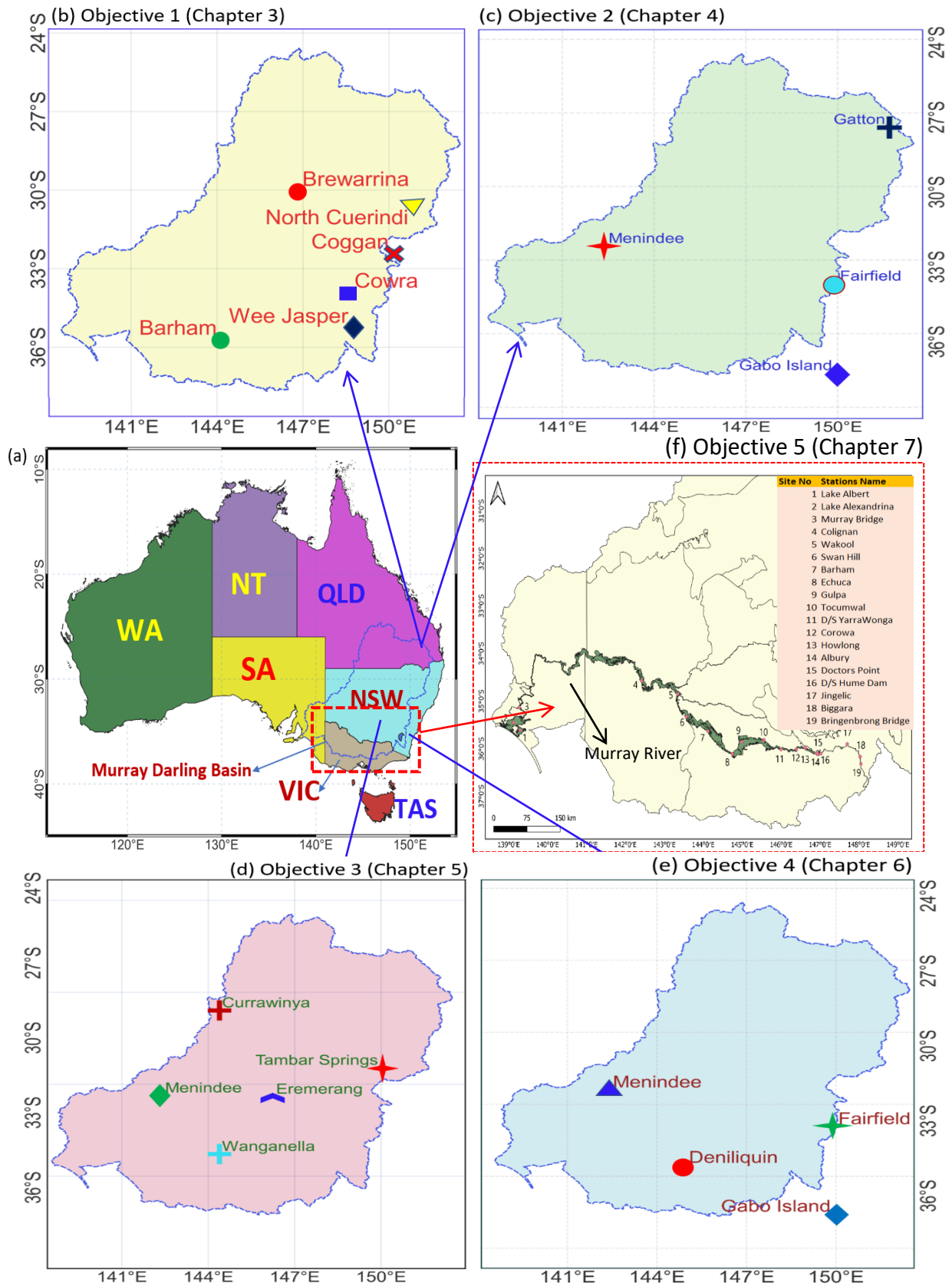


Fig 2.1 Map of the study region (a) the location of Murray-Darling Basin (MDB), with close-ups (b-f) illustrating the selected sites for objectives 1 to 5, respectively.

The basin's mean annual rainfall ranges from 200 mm in the southwest to 1800 mm in the southeast, while altitude ranges from 0 m at the Murray River mouth to 2225 m in the Upper Murray catchment (Austin et al. 2010). The distribution of rainfall is also anticipated to shift due to climate change. In both scenarios, summer rainfall is expected to decrease, most notably in the basin's southern parts. Climate change affects precipitation, temperature, humidity, and wind speed, affecting potential evaporation. By 2100, both global warming scenarios predict a rise in annual average potential evaporation across the basin, particularly in the Murray River tributary catchments (Austin et al. 2010). On average, the Murray-Darling Basin's estimates indicate a small to moderate decline in water availability for dryland agriculture and a medium to a significant decrease in surface water flows. Fig 2.1 illustrates the geographical locations of the study area

2.2 Data Description

To construct high-precision deep learning hybrid predictive models, a range of data from different sources were used in conjunction with each other. Concisely, Table 2.1 contains the data used, the sources of the data, and additional essential information for attaining each objective.

The development of a streamflow water level forecasting model (Objective 1), in particular, used multiple synoptic mode climate indices (CI) from various sources where the target variable was streamflow water level (SWL) obtained from the NSW Department of Primary Industries of Water. Moreover, Objectives 2 and 4 were formulated by incorporating three data sources: remotely-sensed MODIS data, ground-based SILO database, and synoptic mode climate indices to develop a hybrid deep learning model to forecast soil moisture and reference evapotranspiration. Objective 3 is prepared by utilising CMIP5-derived variables obtained from 4 global climate models (GCM).

Table 2.1 The datasets used in the study

Objective	Data	Source	Study Period	Data Predict	Study Area
1 (Chapter 3)	Predictors: Synoptic climate mode indices	Refer to Table 2.3	1915-2019	Monthly	Fig 2.1(b)
	Target: Streamflow water level	NSW Department of Primary Industries (NSW-Water, 2021)			
2 (Chapter 4)	Predictors: Synoptic climate mode indices	Refer to Table 2.3	2003-2020	Daily	Fig 2.1(c)
	MODIS Satellite	Refer to Table 2.2			
	Ground-Based SILO				
	Target: Reference Evapotranspiration				
3 (Chapter 5)	Predictors: CMIP5-derived Variables	Refer to Table 2.5	1960-2100	Monthly	Fig2.1(d)
	Target: Moisture in an upper portion of the soil column				
4 (Chapter 6)	Predictors: Synoptic climate mode indices	Refer to Table 2.3	2003-2020	Daily	Fig 2.1(e)
	MODIS Satellite				
	Ground-Based SILO				
	Target: Surface Soil Moisture				
5 (Chapter 7)	Predictors: Synoptic climate mode indices	Refer to Table 2.3	2000-2020	Daily	Fig 2.1(f)
	MODIS Satellite	Refer to Table 2.2			
	Ground-Based SILO				
	Target: Streamflow water level				

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

The streamflow water level (SWL) for each month (in meters) was collected from the NSW Department of Primary Industries (DPI). Surface and groundwater resources in New South Wales are monitored and managed by the Department of Primary Industry—Office of Water, responsible for creating water policy. On specific sites in river basins, it monitors daily staff gauge readings gas purge pressure. It floats well-water level recording systems, electronic pressure sensors, telemetered digital logging systems, and telemetered pressure sensors. Generally, the data were from direct gauging, while some were adjusted during processing due to anomalies. The monthly (Objective 1) and daily (Objective 5) were utilised in this study were obtained from the NSW Department of Primary Industries.

2.2.2 Meteorological data - Scientific Information for Landowners

The target variables for objective 2 (i.e., reference evapotranspiration, ET_0) and predictor variables of objectives 2 (Chapter 4), 4 (Chapter 6) objective 5 (Chapter 7) were collected from the repository of Scientific Information for Land Owners (SILO) database: <https://www.longpaddock.qld.gov.au/silo/ppd/index.php> developed by Queensland Department of Environment and Resource Management (Jeffrey et al. 2001; Beesley et al. 2009). A list of ground-based SILO variables is tabulated in Table 2.2(a).

SILO is a database system that provides ready-to-use climate data to biological and hydrological models users. Missing values in the SILO database were interpolated using robust statistical tools in the quality control stages implemented by the Australian Bureau of Meteorology (BOM) (Zajackowski et al. 2013). This has enabled previous hydrological studies to use SILO-based meteorological data (Deo et al. 2016; Salcedo-Sanz et al. 2018; Ahmed, M. et al. 2021). Furthermore, in this doctoral research thesis, the meteorological data (maximum temperature, minimum temperature, relative humidity, rainfall, evaporation, evapotranspiration, and vapour pressure) from SILO-database are integrated with synoptic mode climate indices and Remotely sensed Moderate Resolution Imaging Spectroradiometer (MODIS) data (Chapter 4 and Chapter 7).

2.2.3 Atmospheric Parameters - The Moderate Resolution Imaging Spectroradiometer (MODIS)

The hybrid deep learning models dealing with objectives 2, 4, and 5 were developed using remotely sensed Moderate Resolution Imaging Spectroradiometer (MODIS) datasets. The datasets were collected from NASA's Geospatial Online Interactive Visualization and Analysis Infrastructure (GIOVANNI) repository. GIOVANNI, which captures 2000 satellite variables, is a powerful online visualisation and analysis tool for geoscience datasets (Berrick et al. 2008; Chen et al. 2010). As shown in Table 2.2, MODIS-based predictor variables are used in this study to design and evaluate the hybrid model for hydrological forecasting.

Satellite-based atmospheric parameters (MODIS) are used to create the predictive model for Objectives 2 and 4. MODIS is a satellite instrument found on the Terra (EOS AM) and Aqua (EOS PM) satellites. Terra's orbit around the Earth is timed to cross the equator from north to south in the morning, while aqua crosses the equator from south to north in the afternoon. Terra and aqua orbit the Earth every 1-2 days, providing data with a moderate spatial resolution (250 m at nadir), a large swath (2330 km), and a wide spectral range (36 channels between 0.412 and 14.2 m) (López & Batlles 2014). The MODIS observations yield forty-four data products. Among these products, the MOD08-M3 contains approximately 800 sub-datasets that describe atmospheric features such as cloud fraction, cloud optical thickness, precipitable water vapour amount, and aerosol optical thickness (Kim & Liang 2010).

The predictor variables used to develop and execute the predictive model for objective 4 were obtained from the Global Land Data Assimilation System (GLDAS) system, which represents a high-temporal resolution terrestrial modelling system comprised of the land surface state and several flux parameters with three temporal resolution products: hourly, daily, and monthly. GLDAS 2.0 datasets extracted at daily temporal resolutions were used in our study made publicly available. The study used MODIS-based surface soil moisture (SSM) data obtained from the GLDAS 2.0 model as a target variable (Objective 4), as tabulated in Table 2.2(b).

Table 2.2 Description of predictor variables used to design and evaluate hybrid CEEMDAN-CNN-GRU predictive model for daily surface soil moisture forecasting.

Acronyms	Description	Unit
(a) SILO (Ground-Based Observations)		
Tx	Maximum Temperature	°C
Tn	Minimum Temperature	°C
r	Rainfall	mm
Ep	Evaporation	mm
Rd	Radiation	MJ m ⁻²
VP	Vapour Pressure	hPa
Rx	Relative Humidity at Temperature T.Max	%
Rn	Relative Humidity at Temperature T.Min	%
Mp	Morton potential evapotranspiration overland	mm
(b) GLDAS 2.0: MODIS Satellite Data from GIOVANNI Repository		
St	Average Surface Skin temperature	K
CW	Plant canopy surface water	Kg m ⁻²
CE	Canopy water evaporation	kg m ⁻² s ⁻¹
Es	Direct Evaporation from Bare Soil	kg m ⁻² s ⁻¹
ET	Evapotranspiration	kg m ⁻² s ⁻¹
Es	Snow Evaporation	kg m ⁻² s ⁻¹
GW	Groundwater storage	mm
LW	Net longwave radiation flux	W m ⁻²
Qg	Ground heat flux	W m ⁻²
Qh	Sensible heat net flux	W m ⁻²
Qle	Latent heat net flux	W m ⁻²
Qs	Storm surface runoff	kg m ⁻² s ⁻¹
Qb	Baseflow-groundwater runoff	kg m ⁻² s ⁻¹
Qm	Snow melt	kg m ⁻² s ⁻¹
Sn	Snow depth	kg m ⁻² s ⁻¹
Snt	Snow Surface temperature	m
Sp	Profile Soil moisture	kg m ⁻²
Sz	Root Zone Soil moisture	kg m ⁻²
Ssur	Surface Soil moisture	kg m ⁻²
SW	Snow depth water equivalent	kg m ⁻²

Acronyms	Description	Unit
SR	Net short-wave radiation flux	kg m ⁻²
Tr	Transpiration	kg m ⁻² s ⁻¹
TW	Terrestrial water storage	mm

(c) MERRA-2 Model based Satellite Data

T2X	2-meter air temperature - daily max	K
T2A	2-meter air temperature - daily mean	K
T2M	2-meter air temperature - daily min	K

(d)MODIS-Aqua Model based Satellite Data

AOD	Aerosol Optical Depth 550 nm (Dark Target)	-
asam	Aerosol Scattering Angle: Mean of Daily	Degrees
cdt	Combined Dark Target and Deep Blue AOD at 0.55 micron	-
pwvm	Precipitable Water Vapor 440 to 10mb: Mean	cm
pwvs	Precipitable Water Vapor Surface to 680mb	cm
pwvt	Precipitable Water Vapor Total Column: Mean of Level-3 QA Weighted	cm
crm	Cirrus Reflectance: Mean	-
icm	Ice Cloud Effective Particle Radius: Mean	microns
lwce	Liquid Water Cloud Effective Particle Radius: Mean	microns
cfcm	Cloud Fraction from Cloud Mask	-
ccot	Combined Cloud Optical Thickness: Mean	-
icot	Ice Cloud Optical Thickness: Mean	-
lwco	Liquid Water Cloud Optical Thickness: Mean	-
cpdm	Cloud Top Pressure (Day): Mean	hPa
cpav	Cloud Top Pressure: Mean	hPa
cpnm	Cloud Top Pressure (Night): Mean	hPa
ctav	Cloud Top Temperature: Mean	K
ctdm	Cloud Top Temperature (Day): Mean	K
ctnm	Cloud Top Temperature (Night): Mean	K
icwm	Ice Cloud Water Path: Mean	gm ⁻²
lwcm	Liquid Water Cloud Water Path: Mean	gm ⁻²
AODl	Aerosol Optical Depth 550 nm (Deep Blue, Land-only)	-

Table 2.3 Twelve climate model indices were used as predictor variables to forecast the SWL using the hybrid deep learning BRF-LSTM predictive model. Source of data: monthly sea surface temperature (SST) in different oceanic regions derived from the Optimum Interpolation SST, version 2 (OISST v2)

Variable	Notation	Name and Description	Region	Data Source
Nino3.0	N3	Average SST over 150°–90°W & 5°N–5°S	Pacific	OISST v2,
Nino3.4	N34	Average SST over 170°E–120°W & 5°N–5°S	Pacific	OISST v2,
Nino4.0	N4	Average SST over 160 °E–150°W & 5°N–5°S	Pacific	OISST v2,
Nino1+2	N12	Average SST over 90°W–80°W & 0°–10°S	Pacific	OISST v2,
AO	AO	Arctic Oscillation	ET	OISST v2,
DMI	DMI	DMI = WPI – EPI WPI = Mean SST over 50°–70°E & 10°N–10°S EPI = Mean SST over 90°–110°E & 0°N–10°S	Indian	OISST v2, NOAA
EMI	EMI	EMI = C – 0.5 x (E+W) Where the components are average SSTA over C: 165 °E–140 °W and 10°N–10°S E: 110°–70°W and 5°N–15°S W: 125°–145°E and 20°N–10°S	Pacific	ERRSST.v.3b
NAO	NAO	North Atlantic Oscillation		OISST v2,
PDO	PDO	Pacific Decadal Oscillation	Pacific	OISST v2,
SAM	SAM	Southern Annular Mode index	Pacific	OISST v2,
SOI	SOI	Southern Oscillation Index The pressure difference between Tahiti and Darwin as defined by Troup (1965)	Pacific	BOM
TPI	TPI	Tripole Index for the Interdecadal Pacific Oscillation	Tropical	OISST v2, NOAA
MJO1	MJ1	Madden Julian Oscillation-1		
MJO2	MJ2	Madden Julian Oscillation-2		
MJO4	MJ4	Madden Julian Oscillation-4		
MJO5	MJ5	Madden Julian Oscillation-5		
MJO6	MJ6	Madden Julian Oscillation-6		
MJO7	MJ7	Madden Julian Oscillation-7		
MJO8	MJ8	Madden Julian Oscillation-8		

Note: EMI = ENSO Modoki Index; WPI= West Pole Index; EPI = East Pole Index;
DMI = Dipole Model Index, ET= Extra-Tropical

The study also used Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2) data set spans 1980 to the present. Along with the improvements to meteorological assimilation, MERRA-2 makes significant progress towards the Earth System. MERRA-2 is the first long-term global reanalysis that incorporates space-based aerosol observations and their interactions with other physical processes in the climate system. It was developed to replace the original MERRA dataset due to advancements in assimilation technology that allows for the assimilation of contemporary hyperspectral and microwave observations and GPS-Radio Occultation datasets (Gelaro et al. 2017). The spatial resolution is about 50 kilometres latitudinal direction. MERRA-2 is the first long-term global reanalysis that incorporates space-based aerosol measurements and their interactions with other physical processes in the climate system (Draper et al. 2018).

In Objective 5, twenty-two predictor variables from the MODIS-aqua model were also incorporated with the MERRA-2 model. MODIS-aqua views the entire Earth's surface every 2 days, acquiring data in 36 spectral bands. This data helps us understand global dynamics and processes on land, oceans, and the lower atmosphere. They play an important role in developing validated, global, interactive Earth system models capable of accurately predicting global change enough to assist policymakers in making sound decisions about environmental protection. A list of MODIS-aqua-derived predictor variables is provided in Table 2.2 (d).

2.2.4 GCM/CMIP5 Simulated Variables

Global climate models (GCMs) are recognised as full gears for studying the present, past, and future (Xu et al. 2009; Ramesh & Goswami 2014). The 5th phase of the Coupled Model Intercomparison Project (CMIP5) of the World Climate Research Programme (WCRP) demonstrates numbers of GCMs compared to CMIP3 GCMs for the understanding of the mechanisms of climate system change and to improve the capability to simulate climate change (Sillmann et al. 2013; Sun et al. 2015; Meher et al. 2017). CMIP5 GCMs had improved skills in describing the seasonality of precipitation regimes compared to their predecessors over the Asian monsoon region (2016). Still, the performance of the GCMs varied for different river basins and catchments. The CMIP5 simulated variables were utilised throughout the study

periods for the RCP4.5 and RCP8.5 global warming scenarios to estimate the SM under Objective 3. The data sets of GCMs were obtained from the archive of the Centre for Environmental Data Analysis (CEDA): <http://data.ceda.ac.uk/badc/cmip5/data/cmip5/output1>. Accordingly, the list of GCM models and the predictors used in Objective 3 are tabulated in Table 2.4 and Table 2.5.

Table 2.4 Summary of global climate models from CMIP5 simulation sets

Model	Centre/ Country	Spatial Resolution	Data Length		References
			Historical	RCPs	
ACCESS 1.0	CSIRO- BOM/Australia	1.875°×1.25 °	1960-2005	2006-2100	(Marsland et al. 2013)
ACCESS 1.3	CSIRO- BOM/Australia	1.875°×1.25 °	1960-2005	2006-2100	(Marsland et al. 2013)
HadGEM2-CC	MOHC/UK	1.875°×1.25 °	1960-2005	2006-2100	(Martin et al. 2011)
HadGEM2-ES	MOHC/UK	1.875°×1.25 °	1960-2005	2006-2100	(Martin et al. 2011)

Note: **CSIRO**: Commonwealth Scientific and Industrial Research Organization

MOHC: Met Office Hadley Centre

Table 2.5 Summary of CMIP5 simulated variables used in this study of Objective 3

Acronyms	Variables	Units
clt	Total Cloud Fraction	%
evp	Evaporation	mm
hfls	Surface Upward Latent Heat Flux	$\text{MJm}^{-2}\text{d}^{-1}$
hfss	Surface Upward Sensible Heat Flux	$\text{MJm}^{-2}\text{d}^{-1}$
huss	Near Surface Specific Humidity	$\text{MJm}^{-2}\text{d}^{-1}$
mrso	Total soil moisture content	mm
pr	Precipitation	mm
prc	Convective Precipitation	mm
prw	Atmospheric water vapour content	kgm^{-2}
ps	Surface air pressure	$^{\circ}\text{C}$
psl	Sea Level Pressure	Pa
rlds	Surface Downwelling Longwave Radiation	$\text{MJm}^{-2}\text{d}^{-1}$
rlus	Surface Upwelling Longwave Radiation	$\text{MJm}^{-2}\text{d}^{-1}$
rlut	TOA Outgoing Longwave Radiation	$\text{MJm}^{-2}\text{d}^{-1}$
rsds	Surface Downwelling Shortwave Radiation	$\text{MJm}^{-2}\text{d}^{-1}$
rsus	Surface Upwelling Shortwave Radiation	$\text{MJm}^{-2}\text{d}^{-1}$
sfcwind	Monthly mean Near-Surface Wind Speed	ms^{-1}
tas	Near-Surface Air Temperature	$^{\circ}\text{C}$
tasmax	Daily Maximum Near-Surface Air Temperature	$^{\circ}\text{C}$
tasmin	Daily Minimum Near-Surface Air Temperature	$^{\circ}\text{C}$
tsl	Soil Temperature	$^{\circ}\text{C}$
mrsos (SM)	Moisture in an upper portion of the soil column	mm

2.2.5 Synoptic Scale Climate Mode Indices

In Australia, various synoptic-scale climate indices have been acknowledged as a suitable approach to climate variability based on locations and seasons (Schepen et al. 2012; Nguyen-Huy et al. 2017). The impacts of IOD and the subtropical ridge on Australian rainfall variability are well recognised (Murphy & Timbal 2008; Cai & Cowan 2009; Kirono et al. 2010). The La Niña events are responsible for significant precipitation (in eastern Australia), whereas the El Niño events are related to high-scale drought situations (Yuan & Yamagata 2015). In the north and east of Australia, the El Niño Southern Oscillation (ENSO) phenomenon greatly influences precipitation, with regional differences (Risbey et al. 2009). However, IOD shows a superior impact than ENSO during winter and spring rainfall for southern Australia

due to the robust covariation between ENSO and the IOD (Cai et al. 2011). The ENSO hydroclimate relates to below-normal streamflow for El Niño and above regular for La Niño events (Deo & Sahin 2016), which indicates that SST-associated climate indices have robust potential in forecasting streamflow water level in Australia.

Consequently, twenty-one climate indices were acquired from a variety of sources, including the National Climate Prediction Centre (BOM 2020), the Australian Bureau of Meteorology (BOM 2020), and the National Oceanic and Atmospheric Administration (NOAA) with daily sea soil (McKenzie et al. 2005). The monthly synoptic-scale climate indices from various credible and trustworthy databases are shown in Table 2.3. The sea surface temperatures (SSTs) are the most critical indices because they indicate climate variability. In contrast, the other indices (Pacific Decadal Oscillation (PDO), Indian Ocean Dipole (IOD), and El Niño Modoki Index) are dependent on the SSTs. As a result, this study uses the most recent version of SST; Extended Reconstructed Sea Surface Temperature Version 4 (ERSST.v4). The daily sea surface temperature (Nino1 + 2SST, Nino4SST, Nino3SST, and Nino3.4SST) was collected from KNMI- Climate Explorer (Trouet & Van Oldenborgh 2013). The monthly (Objective 1) and daily (Objective 2, 4, and 5) climate mode indices were utilised in this study.

2.3 General Methodology

A number of basic tasks were applied to the predictor variables prior to the model development. Converting the inputs and the target data into their required forecast horizons and using the partial autocorrelation function (PACF) to select the best statistically significant lags from the target variables were the first two steps used in all objectives except the fourth objective of this thesis. Secondly, the datasets were decomposed using a complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) for Objective 4 and a combination of CEEMDAN and variational mode decomposition (VMD) for Objective 5 to address the non-stationarity issues associated with the data. The cross-correlation functions (CCF) were also adopted between the target and the inputs to select the best input variables in Objective 1, Objective 2, and Objective 5.

The selection of features within the inputs used to forecast target variables (i.e., SWL, SM, and ET_o) is a vital stage in the practical application of a predictive model.

This is implemented to reduce the dimensionality of model inputs and computational cost, including the desired improvements in forecasting accuracy and interpretation of predictors' predictive model characteristics and nature (Bowden et al. 2005; Maier et al. 2010; Yang et al. 2012; Prasad et al. 2018b). The study incorporated three potential feature selection algorithms, namely, Boruta-random forest (Objective 1 and 3), ant colony optimization (Objective 2 and 5), and neighbourhood component analysis (Objective 4), to select appropriate input variables for the model.

Furthermore, the higher frequency data from inputs and target datasets of all study objectives were normalized between zero and one using Eq. (1) to avoid large numeric ranges from the values of the predictor variables. Finally, the best parameters and boundary conditions of the models developed in the respective objectives of this study were selected using optimisation and trial-and-error methods. The tasks above were described in detail in their respective chapters.

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (1)$$

In Eq. (1), x is the respective variable, x_{min} is the minimum value, x_{max} is the maximum and x_{norm} is the normalised value.

This thesis considers various data-driven forecasting models and pre-processing methods to evaluate their ability to forecast over different horizons. The models were multivariate adaptive regression splines (MARS), support vector regression (SVR), multilayer perceptron (MLP), decision tree (DTR), random forest (RF), multilinear regression (MLR), recurrent neural regression (RNN), long short-term memory (LSTM), a convolutional neural network (CNN), gated recurrent unit (GRU) network and bi-directional LSTM (BiLSTM). Pre-processing approaches were incorporated to increase the forecast accuracy by decomposing the data into low and high pass filters using CEEMDAN and VMD and selecting the best input parameters using ACO, BRF, and NCA tools. A trial-and-error method was also used in this study to select the appropriate hyperparameter of the models.

The proposed data-driven models were considered to evaluate their precision in forecasting hydrological variables. The deep learning hybrid models range from the double-phase (Chapter 3 to 5) to the more advanced multi-phase hybrid approaches (Chapter 6 to 7). A multilayer perceptron (MLP) is an artificial neural network (ANN)

that uses feedforward learning. MLP is used ambiguously; it is sometimes used broadly to refer to feedforward ANN. Based on the neurological structure of the human brain, it is possible to define ANN as a simplified mathematical model of the brain. Its capability to determine complex correlations among variables makes artificial neural networks (ANN) one of the most powerful tools available in the field of data modelling (Akbari et al., 2014). Compared to the ANN, SVR is significantly faster and more computationally convenient (Akbari et al. 2014; Song & Dai 2017; Blanchard et al. 2018). An implicit feature space mapping from the dimension of the data to a potentially infinite feature space is used by SVR, in the same way as ANNs do, to provide a non-linear representation of the modelled data; this is accomplished by the use of the ‘kernel trick’ (Akbari et al. 2014; Dhiman et al. 2019). The SVR model has been widely regarded as a universal method for handling multidimensional function estimation problems. Figure 2.2 shows a brief overview of artificial intelligence (AI) based on the deep learning predictive models used in this doctoral research thesis.

Deep learning (DL) models use multiple feature extraction layers and efficiently acquire compound associations within the data (Ghimire et al. 2019b). These DL methods have been successfully utilised in medical imaging, natural language processing, and computer vision. Nevertheless, only a few prior studies have explicitly employed a DL model for hydrological forecasting. The deep learning algorithm exhibits excellent quality to extract data characteristics when processing large amounts of complex data with substantial computing power and complex mapping ability (Gong et al. 2019). The convolutional neural network (CNN) is a variant of a traditional neural network consisting of one or more convolution, pooling, and fully connected layers (Wang et al. 2019). Each convolutional layer consists of several convolutional units, and a backpropagation algorithm optimizes the parameters of every unit. The purpose of a convolutional manipulation in CNN is to extract unique features of the input layer. As a distinctive class of recurrent neural networks, LSTMs utilize special units named memory blocks to take the place of the traditional neurons in the hidden layers (Sainath et al. 2015). Moreover, there exist three gates units called input gates, output gates, and forget gates in memory blocks, and hence LSTMs can update and control the information flow in the block through these gates (Chen et al. 2018).

In Chapter 4 and Chapter 6 of this research, the CNN-GRU model is proposed where the CNN algorithm is incorporated to extract intrinsic features of the target time series. At the same time, in the second phase, GRU is connected to CNN to utilize all relevant features for prediction. GRU is a distinct type of long short-term memory (LSTM) network presented by Cho et al. (2014). GRU can achieve long-short reliance on declining ignition gradients. Along with similarities, GRU possesses different characteristics from the LSTM. For instance, the GRU owns two gates: the update gate and reset gate, whereas the LSTM has three gates (i.e., the input gate, forget gate, and output gate).

Chapter 7 of this doctoral research thesis proposes the CBiLSTM model; in this CBiLSTM model, the CNN algorithm is incorporated to extract intrinsic features of the target time series, while in the second phase, LSTM is connected to CNN to utilize all relevant features for prediction. The BiLSTM is an LSTM deformation structure with forwards and backward LSTM layers (Peng et al. 2021), which uses past and future information (Kulshrestha et al. 2020).

In order to handle the non-stationarity features within the inputs, data pre-processing via a proper multi-resolution analysis tool is necessary. Hence, hybridised models with complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) for Objective 4 and a combination of CEEMDAN and variational mode decomposition (VMD) is acquired for Objective 5. In addition, appropriate input selection is imperative not only for input dimension reduction but also to improve the model performances. The optimization by input selection approaches also has its advantages and disadvantages. Therefore, many algorithms were explored, including the neighbourhood component analysis (NCA) for regression, ant colony optimization (ACO), and Boruta-Random forest (BRF) algorithm. Table 2.6 summarizes the details of the methodology and tools used to develop artificial intelligence-based predictive models. The specific models developed in this study include:

- Two DL models (i.e., LSTM and GRU) and two ML models (RNN and SVR) were designed for monthly SWL forecasting. PACF, CCF, and BRF were utilised for input selection (Chapter 3).

- Hybrid DL models (i.e., CNN-GRU and CNN-LSTM) and 7 standalone ML and DL models (i.e., GRU, LSTM, RNN, DTR, MLP, MLR, RF, and MARS) were designed for forecasting ETo. The proposed model is designed by integrating PACF and CCF, ACO, and CNN to optimize the models (Chapter 4).
- Two-phase DL model BRF-LSTM for monthly SM estimation was developed where SVR and MARS are used as comparison models. BRF was utilised to select appropriate model inputs (Chapter 5).
- A three-phase hybrid model (i.e., CEEMDAN-CNN-GRU) is designed for daily SSM forecasting. The proposed hybrid model is designed by integrating NCA, CEEMDAN, and CNN to optimize the models (Chapter 6).
- A new hybrid DL model (i.e., CVMD-CBiLSTM) and SVR model were designed for weekly SWL forecasting. The model is designed by integrating double decomposition methods (i.e., CEEMDAN and VMD) and PACF, CCF, ACO, CNN, CEEMDAN, and VMD for the optimization (Chapter 7).

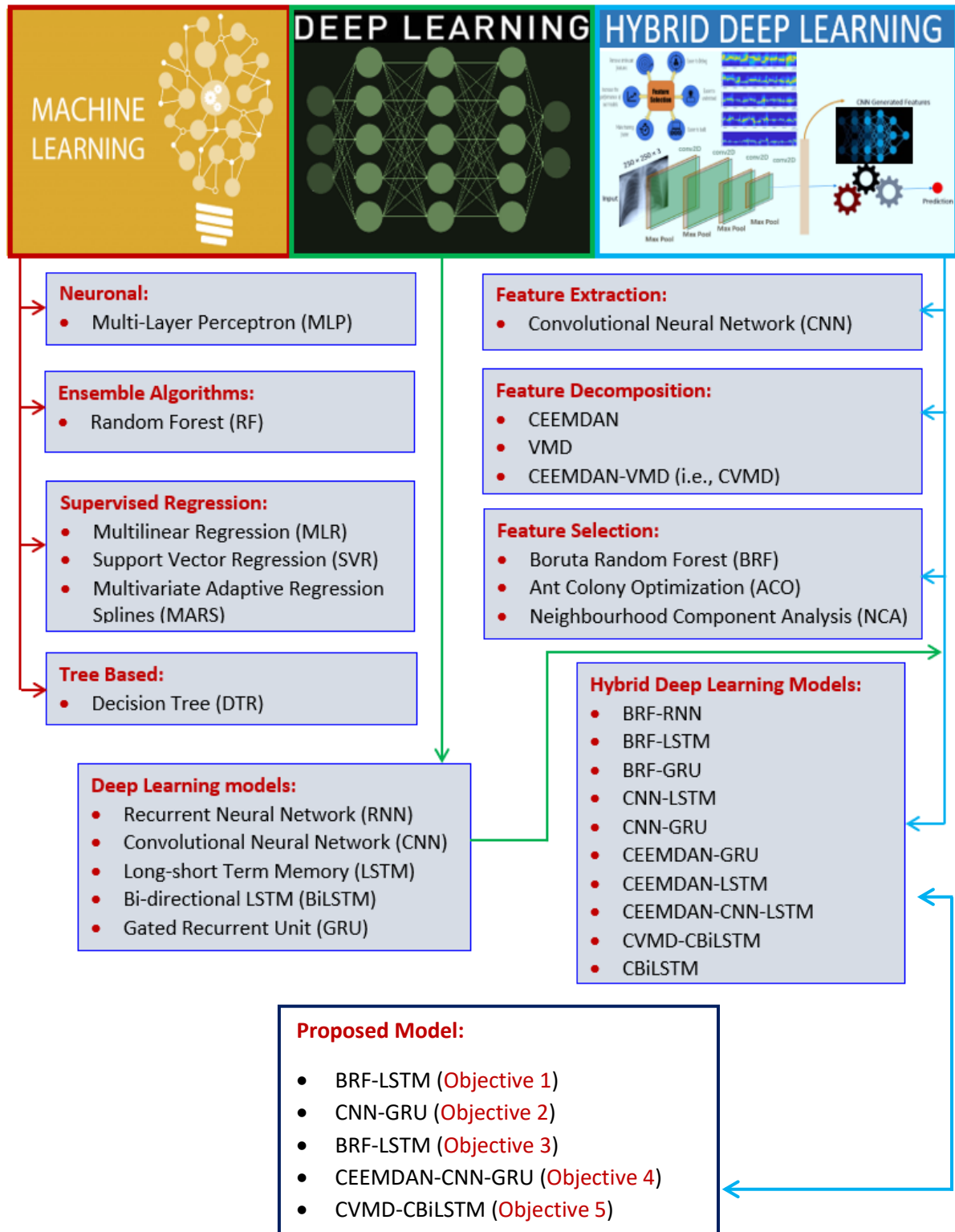


Fig 2.2 Brief overview of artificial intelligence (AI) based-deep learning predictive models used in this doctoral research thesis.

Table 2.6: Summary of the methodology and tools used to develop the predictive models.

Objective	Main Model	Benchmark Model	Hybrid Approaches	Modelling Platform	Target
1 (Chapter 3)	LSTM	GRU, RNN, SVR	<ul style="list-style-type: none"> • PACF and CCF • BRF 	Python R	SWL
2 (Chapter 4)	CNN-GRU	CNN-LSTM, LSTM, RNN, DTR, MLP, MLR, RF, MARS	<ul style="list-style-type: none"> • PACF and CCF • ACO • CNN 	Python Matlab Minitab	ET _o
3 (Chapter 5)	LSTM	SVR, MARS	<ul style="list-style-type: none"> • PACF • BRF 	Python R	SM
4 (Chapter 6)	CNN-GRU	GRU	<ul style="list-style-type: none"> • NCA • CNN • CEEMDAN 	Python Matlab Minitab R	SSM
5 (Chapter 7)	CNN-BiLSTM	BiLSTM, SVR	<ul style="list-style-type: none"> • PACF and CCF • ACO • CNN • CEEMDAN • VMD 	Python Matlab Minitab R QGIS	SWL

2.4 Model Evaluation

Several statistical metrics were employed to evaluate the performance of artificial intelligence-based predictive models. They were based on root mean square error (RMSE), mean absolute error (MAE), correlation coefficient (r), Willmott's index (WI), Nash–Sutcliffe coefficient (NSE), and the Legates and McCabe index (LM). Furthermore, relative (%) error values based on the RMSE (RRMSE) and MAE (RMAE) are also used for model comparison at geographically distinct sites. Besides these statistical metrics, the data-driven predictive models are also analysed with diagnostic plots, including box plots, scatter diagrams, frequency histograms, time series plots, spider plots, and Taylor plots.

CHAPTER 3: STREAMFLOW WATER LEVEL FORECASTING USING CLIMATE INDICES, RAINFALL, AND PERIODICITY

3.1 Foreword

This Chapter is an exact copy of the published manuscript to the *Journal of Hydrology* 599 (2021), 126350 (Scopus Impact Factor 5.72). The title of the manuscript is:

“Deep learning hybrid model with Boruta-Random Forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity.”

This chapter is based on real-time forecasting using a deep learning (DL) hybrid model. The long-short term memory (LSTM) model has hybridized with the Boruta-Random Forest (BRF) algorithm for feature selection. The BRF-LSTM hybrid model exhibits a significant advantage in SWL forecasting compared to conventional approaches. 98% prediction errors in a testing dataset were within 0.015 (m), with a low relative error of 1.30%, outperforming all benchmark models. Using the BRF-based feature selection technique based on maximum-optimal feature selection, the BRF-LSTM model builds a significant predictor set using a random forest model as its underlying learning process. As a result, reducing inputs was complemented properly in addressing the forecasting issues in the study areas, which was accomplished using the proposed model.

3.2 Research Highlights

- A deep learning hybrid model (BRF-LSTM) is built for streamflow forecasts.
- Climate mode indices, rainfall, and periodicity are incorporated for accurate forecasts.
- BRF was used as feature selection with LSTM and GRU models.
- Six gauging sites were tested: BRF-LSTM yields >98% errors in ± 0.015 m, ~1.30% relative error.
- The proposed model is useful for hydrological and strategic water resources planning.

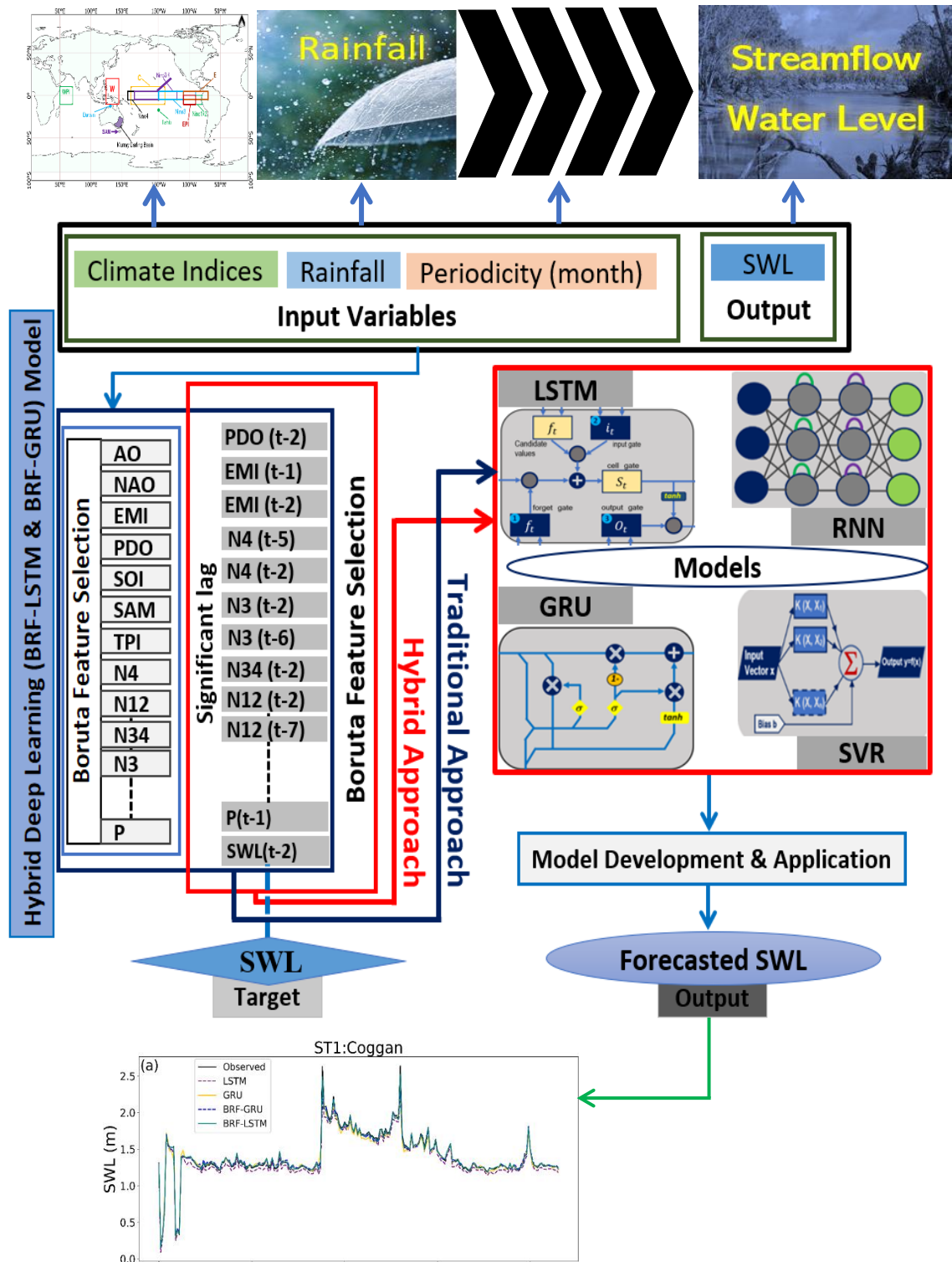


Fig. 3.1 Graphical abstract of Objective 1

3.3 Article 1

This article cannot be displayed due to copyright restrictions. See the article link in the Related Outputs field on the item record for possible access.

CHAPTER 4: EVAPOTRANSPIRATION FORECASTING MODEL AT MULTI-STEP HORIZON

4.1 Foreword

This Chapter is an exact copy of the published manuscript to the *Stochastic Environmental Research and Risk Assessment* 241(2021) (Scopus Impact Factor 3.38). The title of the manuscript is:

“Hybrid deep learning method for a week-ahead evapotranspiration forecasting.”

This chapter comprises a combination of convolutional neural network (CNN) and gated recurrent unit (GRU) network coupled with ant colony optimization (ACO); this research provides a new hybrid-deep learning strategy for forecasting multi-step (weeks 1 to 4) daily-ET_o. The findings demonstrate an excellent forecasting capacity, which shows that the hybrid CNN-GRU model is superior to the other benchmark models in terms of mean absolute error and efficiency. Ultimately, the results of this study reveal that the proposed hybrid CNN-GRU model can capture the complicated and non-linear interactions between predictor variables and the daily ET_o.

4.2 Research Highlights

- A hybrid predictive model (i.e., CNN-GRU) with ant colony optimisation is implemented in forecasting reference evapotranspiration at a multi-step horizon.
- Data from MODIS satellite, ground-based SILO, and synoptic-scale climate mode indices are incorporated into the technique.
- The ACO was found as a realistic approach to obtain the best features from an optimal set of predictor variables.
- The hybrid CNN-GRU model significantly improved the forecasting performance of evapotranspiration against standalone models.
- In general, the results showed ACO estimates would benefit future studies to characterise the regional scenarios of water resources.

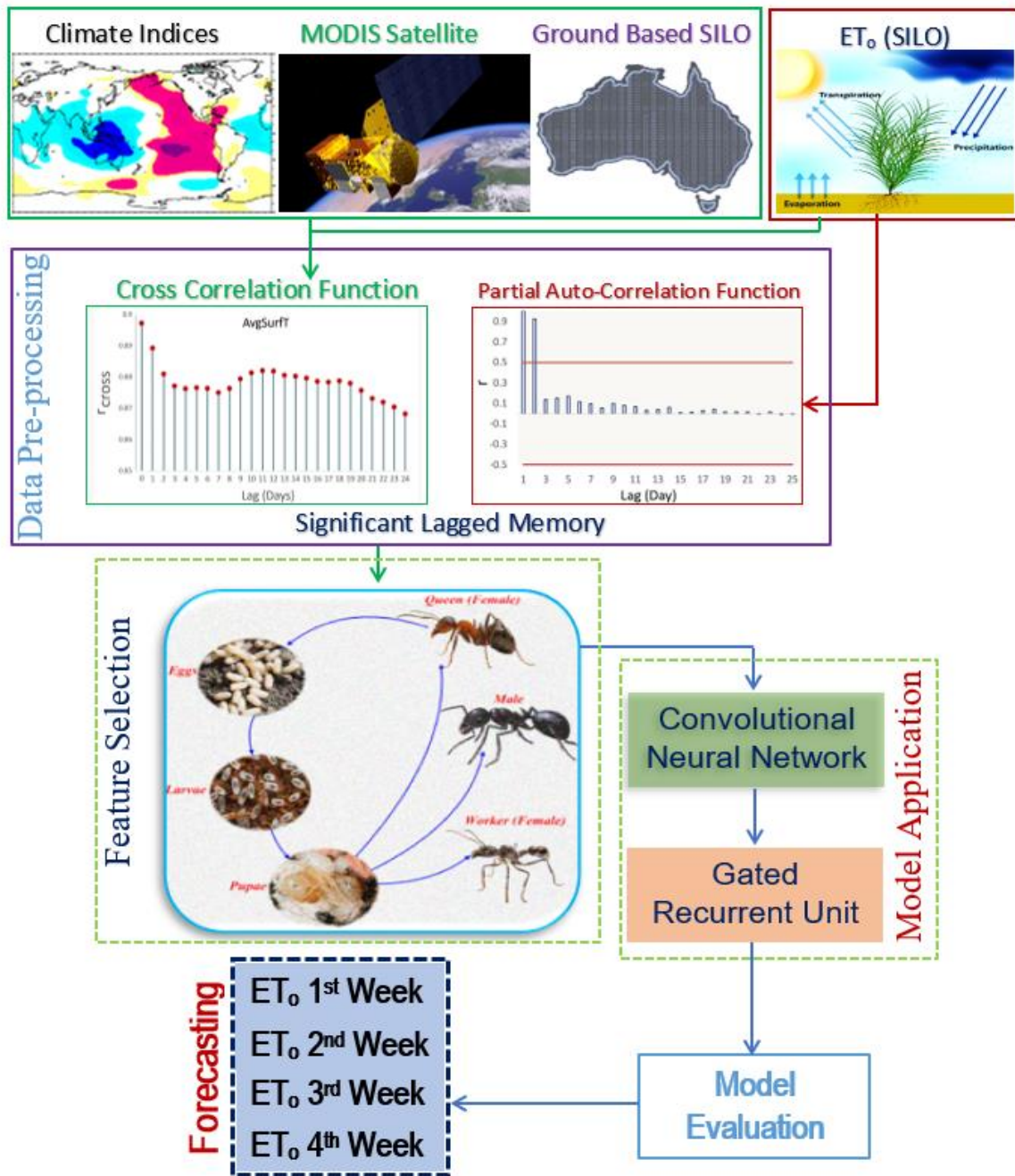


Fig. 4.1 Graphical abstract of Objective 2

4.3 Article 2

This article cannot be displayed due to copyright restrictions. See the article link in the Related Outputs field on the item record for possible access.

CHAPTER 5: SOIL MOISTURE ESTIMATION UNDER RCP4.5 AND RCP8.5 GLOBAL WARMING SCENARIOS

5.1 Foreword

This Chapter is an exact copy of the published manuscript to the *Stochastic Environmental Research and Risk Assessment* **35**, pages 1851–1881 (2021) (Scopus Impact Factor 3.38). The title of the manuscript is:

“LSTM integrated with Boruta-Random Forest optimiser for soil moisture estimation under RCP4.5 and RCP8.5 global warming scenarios.”

This chapter provides a hybridised LSTM framework to simulate moisture in an upper portion of the soil column (SM) in RCP4.5 and RCP8.5 global warming scenarios. The proposed model incorporates Boruta-Random Forest (BRF) feature selection and significant antecedent memory of predictor variables to estimate future SM using the CMIP5 repository. Five study sites in Australia's Murray-Darling Basin were chosen to test the deep learning model's viability for SM estimation till 2100. The BRF-LSTM model is compared to standalone models (i.e., LSTM, SVR, & MARS). The results showed that the BRF-LSTM hybrid model outperformed the standalone models in both warming scenarios.

5.2 Research Highlights

- A hybrid predictive model (i.e., BRF-LSTM) coupled with Boruta-Random Forest was used in estimating the CMIP5 simulated SM under RCP4.5 and RCP8.5.
- It was discovered that the BRF was a viable way to acquire the best features from a collection of predictor variables.
- The suggested hybrid model outperformed all benchmark models in SM estimation, with over 95% of all predicted errors below 0.02 mm and low relative root mean square error (1.06% for RCP4.5 and 1.88% for RCP8.5).
- This study shows that using an LSTM algorithm with BRF feature selection can simulate future SM under climate change.

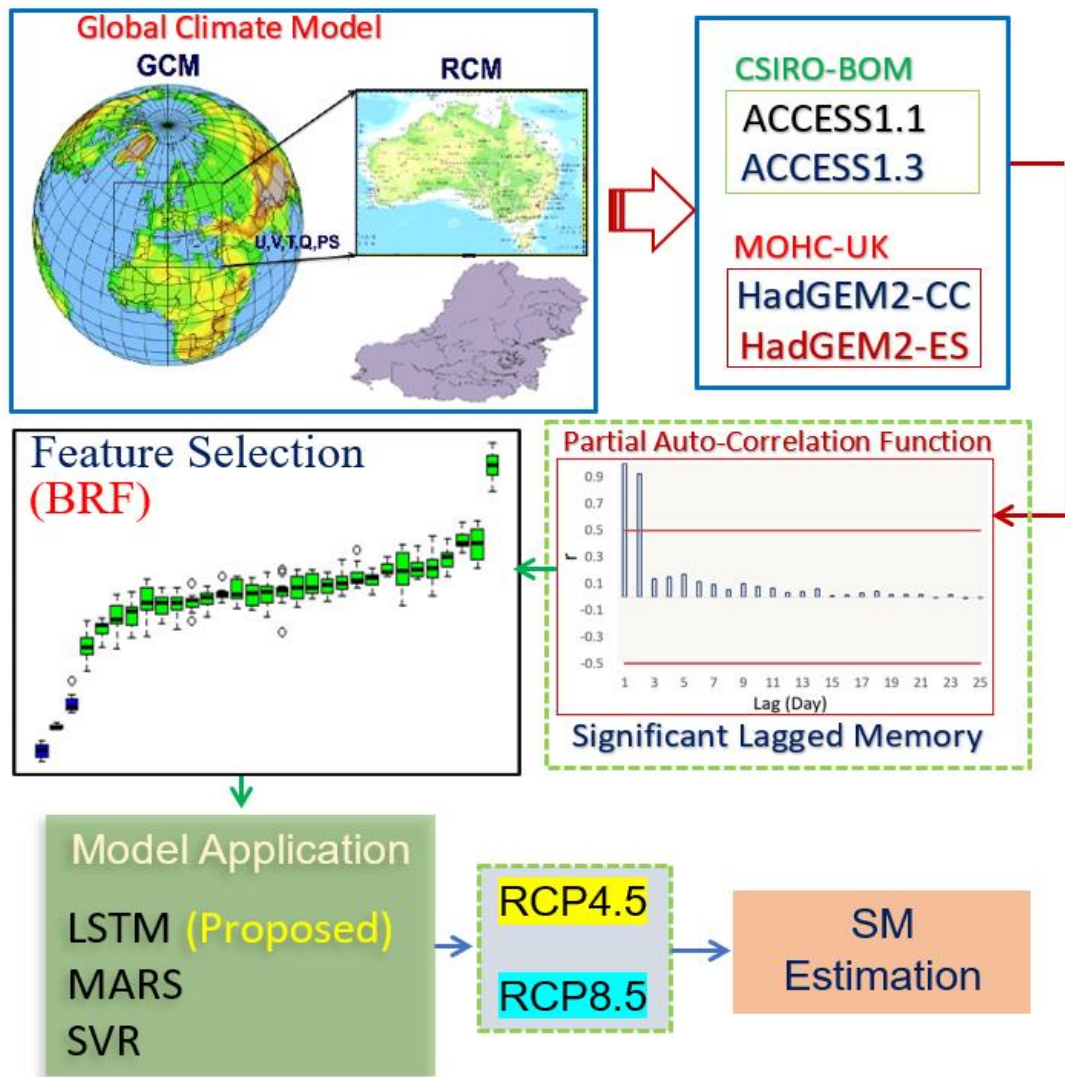


Fig. 5.1 Graphical abstract of Objective 3

5.3 Article 3

This article cannot be displayed due to copyright restrictions. See the article link in the Related Outputs field on the item record for possible access.

CHAPTER 6: SURFACE SOIL MOISTURE FORECASTING AT MULTI-STEP HORIZON

6.1 Foreword

This Chapter is an exact copy of the published manuscript to the *Remote Sensing* **2021**, 13(4), 554 (Scopus Impact Factor 4.85).

“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”

This chapter proposed a deep learning hybrid model for daily time-step surface soil moisture (SSM) forecasts, incorporating the gated recurrent unit (GRU), complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN), and a convolutional neural network (CNN). The CEEMDAN-CNN-GRU hybrid model is tested for SSM forecasting on the 1st, 5th, 7th, 14th, 21st, and 30th-day ahead. The results demonstrate that the proposed model can successfully forecast the surface soil moisture when compared to benchmark models. Therefore, it can be said that the proposed CEEMDAN-CNN-GRU model can be successfully implemented in hydrology and farm management.

6.2 Research Highlights

- A hybrid deep learning CEEMDAN-CNN-GRU model coupled with neighbourhood component analysis was used to forecast the surface soil moisture is incorporated.
- A pool of 52 predictor datasets obtained from three distinct data sources is used for SSM forecasting at multi-step daily horizons.
- The proposed CEEMDAN-CNN-GRU models perform all benchmark models.
- Forecasting surface soil moisture using satellite-based sensors at multi-step horizons can be beneficial for planning and managing environmentally friendly agricultural methods.

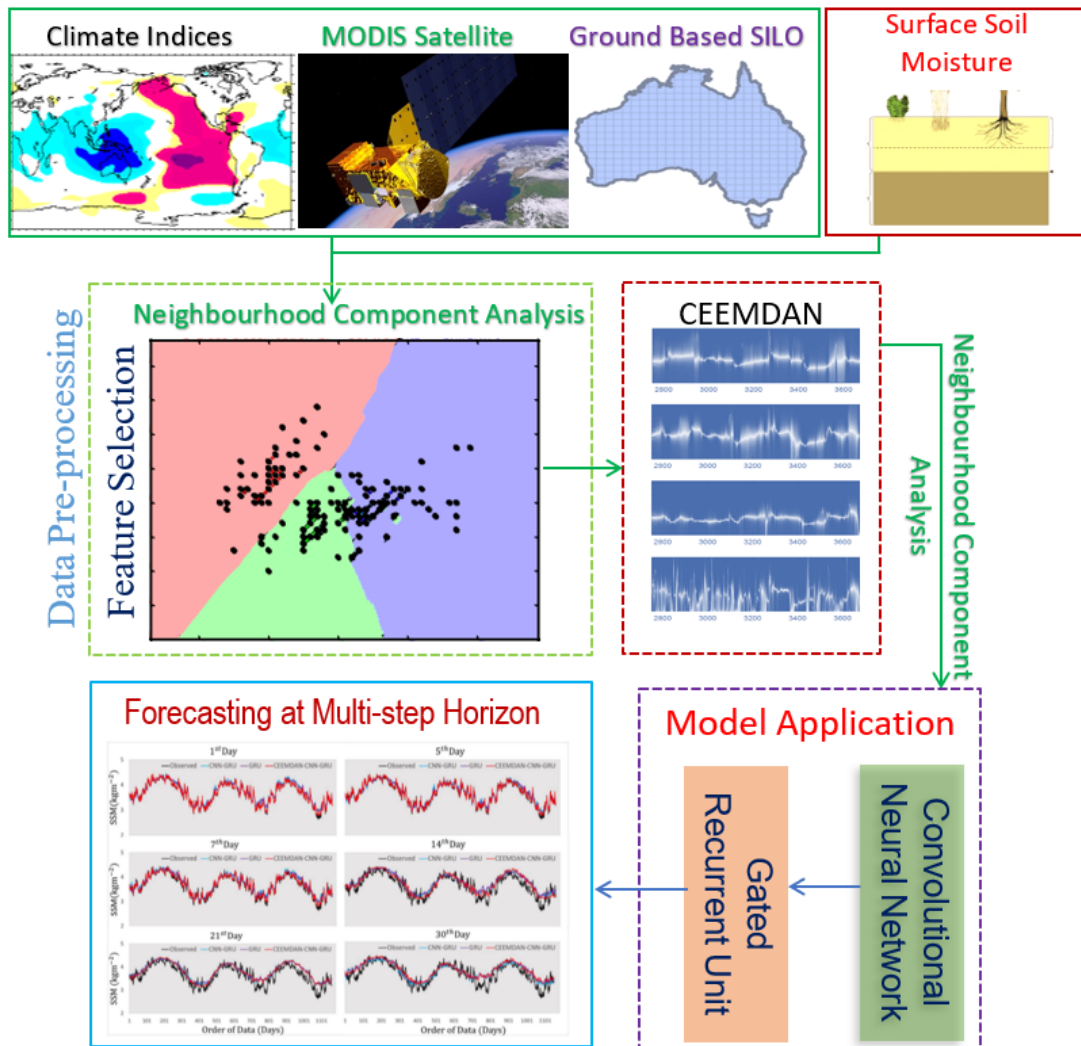







Fig. 6.1 Graphical abstract of Objective 4

6.3 Article 4

Article

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

A. A. Masrur Ahmed ¹, Ravinesh C Deo ^{1,*}, Nawin Raj ¹, Afshin Ghahramani ¹, Qi Feng ², Zhenliang Yin ² and Linshan Yang ²

¹ School of Science, University of Southern Queensland, Springfield, QLD 4300, Australia; abulabramasrur.ahmed@usq.edu.au (A.A.M.A.); Nawin.raj@usq.edu.au (N.R.); afshin.ghahramani@usq.edu.au (A.G.)

² Key Laboratory of Ecohydrology of Inland River Basin and Northwest, Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China; qifeng@lzb.ac.cn (Q.F.); yinzhengliang@lzb.ac.cn (Z.Y.); yanglsh08@lzb.ac.cn (L.Y.)

* Correspondence: ravinesh.deo@usq.edu.au; Tel.: +61-734704430



Citation: Ahmed, A.A.M.; Deo, R.C.; Raj, N.; Ghahramani, A.; Feng, Q.; Yin, Z.; Yang, L. 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 Sens.* **2021**, *13*, 554. <https://doi.org/10.3390/rs13040554>

Academic Editor:

Nemesio Rodriguez-Fernandez

Received: 16 December 2020

Accepted: 27 January 2021

Published: 4 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Abstract: Remotely sensed soil moisture forecasting through satellite-based sensors to estimate the future state of the underlying soils plays a critical role in planning and managing water resources and sustainable agricultural practices. In this paper, Deep Learning (DL) hybrid models (i.e., CEEMDAN-CNN-GRU) are designed for daily time-step surface soil moisture (SSM) forecasts, employing the gated recurrent unit (GRU), complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN), and convolutional neural network (CNN). To establish the objective model's viability for SSM forecasting at multi-step daily horizons, the hybrid CEEMDAN-CNN-GRU model is tested at 1st, 5th, 7th, 14th, 21st, and 30th day ahead period by assimilating a comprehensive pool of 52 predictor dataset obtained from three distinct data sources. Data comprise satellite-derived Global Land Data Assimilation System (GLDAS) repository a global, high-temporal resolution, unique terrestrial modelling system, and ground-based variables from Scientific Information Landowners (SILO) and synoptic-scale climate indices. The results demonstrate the forecasting capability of the hybrid CEEMDAN-CNN-GRU model with respect to the counterpart comparative models. This is supported by a relatively lower value of the mean absolute percentage and root mean square error. In terms of the statistical score metrics and infographics employed to test the final model's utility, the proposed CEEMDAN-CNN-GRU models are considerably superior compared to a standalone and other hybrid method tested on independent SSM data developed through feature selection approaches. Thus, the proposed approach can be successfully implemented in hydrology and agriculture management.

Keywords: deep learning algorithm; MODIS; gated recurrent unit; satellite models of soil moisture

1. Introduction

The precise requirements for water resource supply, constant monitoring, and forecasting are changing continuously with population growth, agricultural and human activities. Any variations in weather and perturbations in climate patterns due to anthropogenically-induced factors affect usable water distribution and accessibility. Instead of precipitation playing a paramount role, the terrestrial water basin tends to dominate the actual functioning of the hydrological, ecological, and inter-coupled socio-economic systems [1]. Notably, the knowledge of fundamental components of water reservoirs, e.g., soil moisture (SM) and streamflow, is essential for an effective water resources management strategy. SM also governs the physical interactions between land and the atmosphere [2,3] and acts as

a driver to feed irrigation systems [4], grazing and crop yield predictions [5]. A decline in groundwater reduces soil water content and the storage volume in underlying soils. A lack of soil moisture can affect agricultural and hydro-meteorological processes. Therefore, predictive models providing prior information on monitoring and forecasting water, such as in this study, are critical to soil moisture forecasts as a principal regulating factor in groundwater hydrology to understand the soil's future state.

With increasing computer power, researchers are developing intelligent models to extract features in historical data (e.g., SM). Such models demonstrate acceptable skills in forecasting hydro-metrological variables, e.g., precipitation [6–9], drought [10], stream-flow [11,12], runoff [13,14], floods [15,16], soil moisture [17], water demand and water quality [18–21]. However, very few studies have focused on the prediction of soil moisture, with most examples being the artificial neural networks (ANN) [22] and the extreme learning machines (ELM) [23]. Irrespective of the model type and domain of applications, accurately forecasted soil moisture presents a greater understanding of water resources and agricultural management, leading to more sustainable decisions. Intelligent systems based on deep learning utilise feature extraction and reveal the compounded association between predictors and targets [24]. Hence, soil moisture prediction with advanced algorithms is a highly practical tool for agricultural water management. DL methods, however, are yet to be explored in the present study region (i.e., Australian Murray Darling Basin). In this study, we adopt a gated recurrent unit (GRU) neural networks as a modified long-short term memory (LSTM) that has attracted good research attention [25]. There appear to be only a few studies on GRU-based models, especially in hydrology [26,27]. Convolutional Neural Networks (CNNs) is a useful feature extraction method to improve the overall predictive process [28]. Therefore, an integration of CNN and GRU can, in foreseeable possibilities, lead to a robust pre-processing of data providing a viable option to improve the model's forecasting skill. This has been evident in some studies that integrated CNN with LSTM for improved performance, with Ghimire et al. [28] showing the superior skill of the CNN-LSTM model in the problem of solar radiation. Integration of deep learning (i.e., CNN-GRU) for soil moisture forecasting is yet to be tested explicitly, with no studies previously using this method, the focus of this study.

Given the stochastic nature of hydrological variables, multi-resolution analysis (MRA) can enhance any model's performance as a tool to reveal the data features. Conventional MRA, for example, discrete wavelet transforms (DWT), have long been implemented [29–32]. However, DWT appears to have drawbacks, and this critical issue is resolved by the maximum-overlap discrete wavelet transform (MODWT), an advanced DWT method [11,33,34]. In this study, we adopt an improved version of EMD, i.e., complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) to implement a self-adaptive decomposition of the predictor variables [23]. In CEEMDAN-based decomposition, a coefficient representing Gaussian white noise with a unit variance is added consecutively at each stage to reduce the forecasting procedure's complexity, avoiding the time series' intricacy [35]. Previous studies used CEEMDAN in forecasting soil moisture [23,36] with an earlier version (i.e., EEMD) used in forecasting streamflow [37] and rainfall [38–40]. Moreover, The multivariate empirical mode decomposition (MEMD) is a self-adaptive algorithm that establishes multivariate inputs to perform a proper investigation [41]. The MEMD method has been successfully applied in time series forecasting [42,43]. The study incorporates the CEEMDAN method as neither the EEMD nor the CEEMDAN decomposition approach has been assimilated with any deep learning approach (i.e., GRU) to produce a soil moisture forecast system, as attempted in the present study.

Climate indices have long been recognised as a useful synoptic-scale indicator of teleconnections representing climate variability [9,44]. La Niña, represented by climate indices, is accountable for substantial rainfall in eastern Australia, whereas the El Niño phenomenon is related to drought [45]. However, El Niño Southern Oscillation (ENSO) has a potential impact on precipitation in northern and eastern Australia [46]. Considering the substantial effects of ENSO phenomena on Australia's climate variability, some

studies [9,47,48] have correlated ENSO effects with hydrological variables (e.g., streamflow, rainfall, and droughts). Rashid et al. [49] aimed to predict a drought index in Australian catchments by aggregating synoptic-scale climate mode indices. Considering these studies, the design of an artificial intelligence model utilising synoptic-scale climate indices, as done in this paper, can be of great practical value in developing sustainable river systems and drought management strategies.

In our paper, we rely on satellite (i.e., MODIS) sensors providing a flexible remote system to explore the nexus between physical, chemical, and biological parameters related to ground variables (i.e., observations) and how these affect future changes in daily soil moisture. However, the inclusion of three distinct datasets has a high potential to address the uncertainties in the predictor variables, especially the remote sensing data's errors. The variables from satellite sensors are associated with errors that propagate to the prediction of hydrologic variables [50–52]. To address this issue, it is preferable to integrate satellite and ground-based variables. Ghimire et al. [53] integrated GIOVANNI data with ECMWF Reanalysis to predict long-term solar radiation. However, the integration of satellite-based, ground-based SILO data, and climate indices for soil moisture forecasts, particularly with deep learning methods (e.g., LSTM), is yet to be implemented.

The objectives are, therefore, fourfold. (1) To build deep learning approaches to forecast surface soil moisture (SSM) at 2 cm depth, incorporating CEEMDAN (i.e., data splitting method) with CNN (i.e., feature extraction method) to generate a GRU-based predictive model. This predictive system, denoted as the CEEMDAN-CNN-GRU hybrid model, is improved with neighbourhood component analysis as a feature selection tenet on diverse predictors obtained from MODIS data, climate mode indices, and ground-based SILO product. (2) To adopt the hybrid CEEMDAN-CNN-GRU model for daily SSM forecasts at a multi-step horizon (i.e., 1st, 5th, 7th, 14th, 21st, and 30th day lead time). (3) To explore the contributory influence of climate indices on the accuracy of the CEEMDAN-CNN-GRU model. (4) To comprehensively benchmark the objective model against alternative tools such as the GRU standalone algorithm, CEEMDAN-GRU, and CNN-GRU hybrid model. This study's primary contribution is to generate a skilful deep learning method for soil moisture prediction, capitalising on remote sensing and ground data while capturing pertinent relationships between soil moisture and synoptic-scale drivers of climate variability in the Australian Murray Darling Basin.

2. Materials and Methods

2.1. Theoretical Frameworks

2.1.1. Convolutional Neural Network

To build the CEEMDAN-CNN-GRU hybrid model trained for daily SSM forecasts, this study purposely employs the Convolutional Neural Networks (CNN) for optimal feature extraction from the input dataset. CNN's have some similarities with conventional neural networks. They are, however, different in their connectivity between and within neuronal layers. In conventional neural networks, every neuron is wholly connected to all neurons in prior layers, whereas single layer neurons do not contribute to the model's network. CNN's are similar to Feed Forward Neural Networks [54], with its model architecture having three layers based on pooling, convolutional, and fully connected layer settings.

The connected layer is employed to estimate objective variables depending on the predictor variable's input features. CNN has proven to be a reliable modelling tool to extract hidden features in inputs and generating filters capturing data features in predictors [55]. To extract the pattern in an objective variable (i.e., SSM) and associated predictor variables, each convolutional layer is established as follows [56]:

$$h_{ij}^k = f \left(\left(W^k \times x \right)_{ij} + b_k \right) \quad (1)$$

Here, W^k is referred to as the weight of the kernel associated with k th feature map, f is the activation function, and the operator of the convolutional procedure is denoted by

multiplication sign (\times). The rectified linear unit (ReLU) is used as an activation function and the adaptive moment estimation (Adam) is selected as an optimisation algorithm using the grid search approach. The ReLU is described as:

$$f(x) = \max(0, x) \quad (2)$$

A one-dimensional convolutional operative was adopted to directly forecast the 1-Dimensional dataset, which eventually simplifies the modelling procedures for real-time forecasting execution.

2.1.2. Gated Recurrent Unit Network

The hybrid CEEMDAN-CNN-GRU model utilises Gated Recurrent Unit (GRU) neural network as the predictive tool after extracting features based on the CNN algorithm (Section 2.1.1). GRU is a distinct type of long short-term memory (LSTM) network presented by Cho et al. [57]. Along with similarities, GRU possesses different characteristics from the LSTM. For instance, the GRU owns two gates, namely the update gate and reset gate, whereas the LSTM has three gates (i.e., the input gate, forget gate, and output gate). Figure 1 provides a schematic of the hybrid CEEMDAN-CNN-GRU model with CEEMDAN data decomposition and model architecture. Moreover, Figure 1b shows the structure of the gated recurrent unit network.

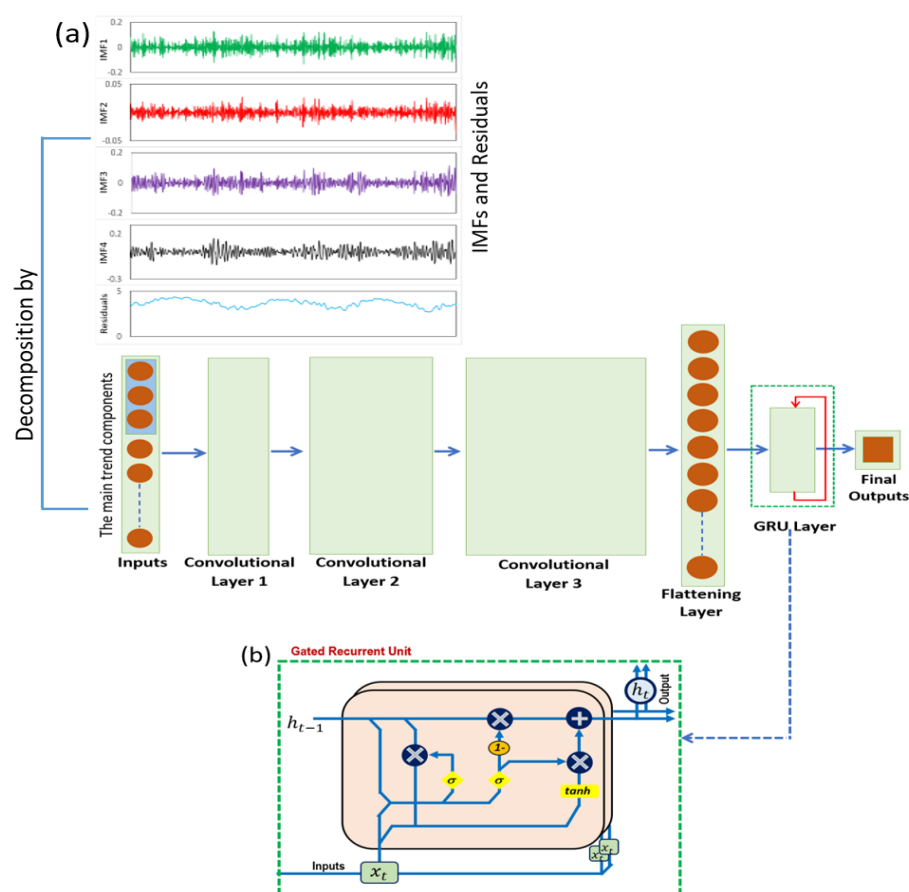


Figure 1. (a) Schematic of the hybrid CEEMDAN-CNN-GRU model with Complete Ensemble Empirical Model Decomposition (CEEMDAN), Convolutional Neural Networks (CNN), and Gated Recurrent Unit (GRU) Neural Network arrangement. The IMFs (Intrinsic Mode Functions) and residual series are generated in the CEEMDAN process, whereas the CNN algorithm represents the feature extraction stage. (b) 2-layered GRU model.

In a GRU Network, two input features, including the input vector $x(t)$ and output vector $h(t - 1)$, are present in each layer. The yield of each gate is achieved by logical operation and non-linear transformation of predictors. Moreover, the association between predictors and predictand can be defined as follows:

$$r(t) = \sigma_g(W_r x(t) + U_r h(t - 1) + b_r) \quad (3)$$

$$z(t) = \sigma_g(W_z x(t) + U_z h(t - 1) + b_z) \quad (4)$$

$$h(t) = (1 - z(t))o(t - 1) + z(t)o\hat{h}(t) \quad (5)$$

$$\hat{h}(t) = \sigma_h(W_h x(t) + U_h(r(t))o h(t - 1)) + b_h \quad (6)$$

where $r(t)$ is the reset gate vector, $z(t)$ is defined as the update gate vector, W and U are parameter metrics and vector. σ_h is referred to as a hyperbolic tangent, and σ_g is defined as a sigmoid function. Finally, given the architecture of GRU, a training approach is chosen, which includes backpropagation through time. Based on previous studies, *Adam* optimiser was implemented as it has enhanced expertise.

2.1.3. Hybrid CNN–GRU. Neural Network

In this paper, the hybrid modelling approach utilises a deep learning method built upon a feature extraction procedure under a forecast model framework. This research demonstrates how the CNN–GRU model comprised of three-layered CNN is used for feature extraction to generate future changes in the objective variable (i.e., SSM). In particular, the GRU layer is employed to integrate input features extracted by the CNN algorithm to finally forecast the target variable (i.e., SSM) with minimal training and testing error.

2.1.4. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN)

As elucidated in Section 1, CEEMDAN is adopted as an improved version of EMD and EEMD to perform a self-adaptive decomposition of model input signals [23] prior to modelling the target variable. The CEEMDAN decomposition process commences by discretising the n -length inputs of any model $\chi(t)$ into intrinsic mode functions (IMFs) and residues to comply with tolerability provision. Nevertheless, to ensure no leakage of information in the IMFs and residues from the training series into the future (i.e., testing and validation subset), the decomposition is performed separately for training, validation, and testing. The actual IMF is produced by taking the mean of the EMD-grounded I.M.F.s across a trial and the combination of white noise to model the predictor–target variables.

Assume that we have D -dimensional set, with n -length X_i matrix (i.e., inputs selected by two-phase decomposed sub-series achieved during the decomposition) and the 1-dimensional surface soil moisture as the target variable. The difference between CEEMDAN and EEMD is that in the CEEMDAN case, a restricted noise (ε_i) across $[0, 1]$ is included at every single decomposition stage, calculated to induce the IMF to take the lead to insignificant error. Considering $E_j(\cdot)$ as an operator producing J th modes obtained from EMD, we follow Torres et al. [58] to implement the CEEMDAN process as follows:

Step 1: The decomposition of p -realizations of $\chi[n] = \varepsilon_1 \omega^p[n]$ using EMD to develop their first intrinsic approach, as explained according to the equation:

$$IMF_1[n] = \frac{1}{p} \sum_{p=1}^P IMF_1^p[n] = \overline{IMF}_1[n] \quad (7)$$

Step 2: Putting $k = 1$, the 1st residue is computed following Equation (7).

$$Res_1[n] = \chi[n] - IMF_1[n] \quad (8)$$

Step 3: Putting $k = 2$, the 2nd residual is obtained as:

$$IMF_2[n] = \frac{1}{p} \sum_{p=1}^P E_1(r_1[n] + \varepsilon_1 E_1(\omega^p[n])) \quad (9)$$

Step 4: Setting $k = 2 \dots K$ calculates the k th residue as:

$$Res_k[n] = Res_{k-1}[n] - IMF_k[n] \quad (10)$$

Step 5: Now we decompose the realisations $Res_k[n] + \varepsilon_1 E_1(\omega^p[n])$, Here, $k = 1, \dots, K$ until their first model of EMD is reached; here, the $(k + 1)$ is:

$$IMF_{(k+1)}[n] = \frac{1}{p} \sum_{p=1}^P E_1(r_k[n] + \varepsilon_k E_k(\omega^p[n])) \quad (11)$$

Step 6: Now the k value is incremented, and steps 4–6 are repeated. Consequently, the final residue is achieved:

$$RES_k[n] = \chi[n] - \sum_{k=1}^K IMF_k \quad (12)$$

Here, K is defined as the limiting case (i.e., the highest number of modes). To comply with the replicability of the earliest input, $\chi[n]$, the following is performed for the CEEMDAN approach.

$$\chi[n] = \sum_{k=1}^K IMF_k + RES_k[n] \quad (13)$$

The additive noise demonstrates that signal-to-noise ratio (ε) is operated at every phase [59,60] and must connect the low magnitude with high-frequency signals in the data [61,62]. Figure 1a provides the CEEMDAN decomposed IMFs and residuals and CNN architecture.

2.1.5. Feature Selection: Neighbourhood Component Analysis

The selection of features within the inputs used to forecast soil moisture is vital in applying a predictive model. This is implemented to reduce the dimensionality of model inputs and computational cost, including the desired improvements in the forecasting accuracy and interpretation of the predictive model characteristics and nature of its predictors [59,63–65]. This study has adopted Neighbourhood Component Analysis (NCA) based on regressions applied to segregate the potential input variables from 52 predictor variables. Introduced by Yang et al., this method uses a competent, non-rectilinear, and non-parametric implanted approach. The MATLAB function called “*fsrnca*” performs NCA feature selection with regularisation to learn feature weights for the minimisation of an objective function that measures the average ‘leave-one-out’ regression loss over the training data. The NCA process’s *fsrnca* approach is adopted to train a variable set to better understand the importance of features through weight by minimising the objective function and calculating the regression loss of predictive model for soil moisture forecasts.

Consider training a dataset $T = \{(x_i, y_i) : i = 1, 2, 3, \dots, N\}$ where $x_i \in R^P$ is the feature vectors (i.e., predictor variables), $y_i \in R$ is the target (i.e., SSM), and N is the sample number for the training set. A function $g(x) : R^P \rightarrow R$ is absorbed by *fsrnca* algorithm to forecast the response y from several input variables, optimising their nearest spaces. The weighted distance (D_w) amongst any two samples is calculated as:

$$D_w(x_a, x_b) = \sum_{j=1}^J w_j^2 |x_a, x_b| \quad (14)$$

where x_a and x_b are the two samples used during training, and w_j is defined as the weight-related to the j th feature. Furthermore, the probability distribution ($p_{\alpha\beta}$) is employed to increase its leave-one-out forecasting correctness in the training phase. By contrast, the probability is that x_a chooses x_b as its reference argument. The algorithm acquires a weighting vector 'w' for gradient the ascent method to determine the feature subset with a regularisation factor to prevent overfitting.

3. Study Area and Data

3.1. Study Area and Description of Predictive Model Development Dataset

For the first time, this study aims to build a new forecast for daily surface soil moisture (SSM) with convolutional-gated recurrent unit neural networks within the Australian Murray Darling Basin (MDB). The MDB covers $\sim 1,042,730 \text{ km}^2$ (or 14%) of mainland Australia [24,66] and $\sim 67\%$ of agricultural lands [67]. As illustrated (Figure 2), the sites are selected based on climate class and soil type diversity, namely Menindee, Deniliquin, Fairfield, and Gabo Island.

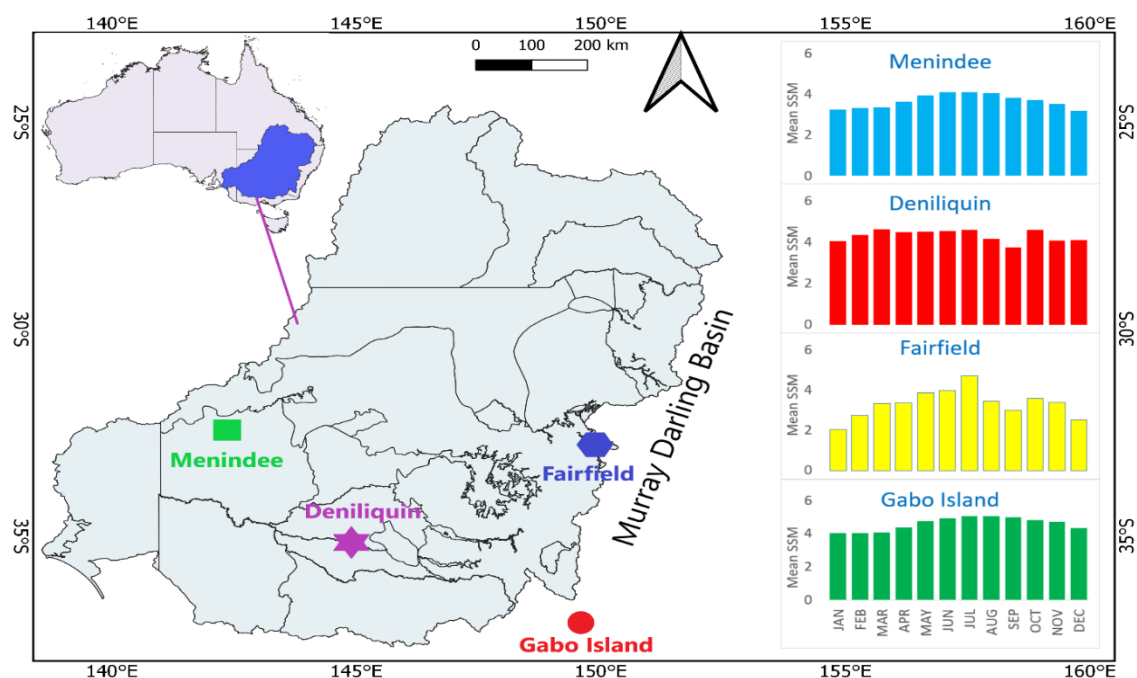


Figure 2. The Australian Murray Darling Basin with study sites and Surface Soil Moisture (SSM, kgm^{-2}) where the hybrid CEEMDAN-CNN-GRU model at multi-step daily SSM forecasting.

The geographical locations and physical characteristics of the sites in Murray Darling Basin are tabulated in Table 1. It should be noted that the site Gabo Island is located at the border of the MDB region for comparison purposes with the other study stations, whereas ~ 20 lakes surround Menindee in a harsh desert environment. The site Fairfield lies within the savannah climate class with land-use patterns of dryland cropping [23]. Figure 2 also shows a histogram of monthly surface soil moisture patterns for the candidate sites.

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

Station Name	BOM Station ID	SILO Position (MODIS Grid Area)	Major Climate Class [68]	Soil Type [69]	Elevation [70]
Menindee	047019	32.39°S, 142.42°E (142.5°E, 32.5°S, 142.25°E, 32.25°S)	Desert	Calcarosol	61
Deniliquin	074128	35.53°S, 144.97° (145°E, 35.25°S, 144.75°E, 35°S)	Savannah	Calcarosol	94
Fairfield	066137	33.92°S, 150.98°E (149.75°E, 37.75°S, 150.0°E, 37.5°S)	Savannah	Vertosol	15
Gabo Island	084016	37.57°S, 149.92°E (150°E, 37.75°S, 149.75°E, 37.5°S)	Sub-Tropical	Sodosol	15

BOM = Bureau of Meteorology, Australia.

The appropriate selection of predictors related to the objective variable has a crucial role in predictive model design. To build a robust model, we adopt remotely sensed MODIS satellite-derived data identified as potential predictor variables in other studies, e.g., solar radiation prediction [24,71,72]. We consider different studies that demonstrate the potential utility of synoptic-scale climate indices that modulate Australian rainfall and crops [41,73,74]. This study integrates three unique data (i.e., satellite-derived data, climate indices, and ground-based variables) to capture a diverse suite of predictive features to forecast SSM, enabling the deep-learning approach a significant edge over the solely station-based models.

3.1.1. MODIS Satellite Dataset

Our hybrid deep learning model (i.e., CEEMDAN-CNN-GRU) is built upon NASA's Geospatial Online Interactive Visualization and Analysis Infrastructure (GIOVANNI) repository (1 February 2003 to 31 March 2020). GIOVANNI represents a powerful online visualisation and analysis tool for geoscience datasets, capturing 2000 satellite variables [75,76]. In this study, MODIS-based predictor variables, presented in Table 2, are utilised to design and evaluate the hybrid CEEMDAN-CNN-GRU model for SSM forecasting. These are extracted from the GLDAS system representing the high-temporal resolution terrestrial modelling system consisting of the land surface state and several flux parameters with three temporal resolution products: hourly, daily, and monthly. Our study has used GLDAS 2.0 datasets extracted in daily temporal resolutions available publicly. The study utilised MODIS-based surface soil moisture (SSM) data as a target variable obtained from the GLDAS 2.0 model.

3.1.2. Scientific Information for Landowners (SILO) Dataset

To increase the pool of predictors, enabling effective feature engineering and increased performance of the DL model, this study selects nine meteorological variables from Scientific Information for Landowners (SILO): <https://www.longpaddock.qld.gov.au/silo/ppd/index.php> (accessed on 31 December 2020). SILO, managed by the Department of Environment and Science, Queensland Government [77], is popular for studying the Australian climate. Table 2 provides a list of SILO data.

3.1.3. Climate Indices

In previous studies, e.g., [9,29,59,74] on modelling precipitation, streamflow, and soil moisture, the role of synoptic-scale and climate indices were found significant in improving the overall model. In this study, twenty-one climate indices are thus obtained from many sources: National Climate Prediction Centre, Australian Bureau of Meteorology [70], and National Oceanic and Atmospheric Administration (NOAA) with daily sea

surface temperature (Nino1 + 2SST, Nino3SST, Nino3.4SST, Nino4SST) over 1 March 2003 to 31 March 2020 from KNMI Climate Explorer [78]. As the positive SOI is related to La-Nina and negative SOI concurs with El-Nino events [79,80], this study has used all of these indices due to strongly correlated rainfall with lagged SOI showing high predictability of rainfall from August–November [44,81]. To further enhance the predictive skill of the deep learning model, we consider Madden-Julian Oscillation (MJO) known to produce a substantial effect on tropical weather [70], which indeed entails a change in rainfall, wind, sea surface temperature (SST), and cloudiness [82]. Hence, eight daily MJO indices were adopted from KNMI Climate Explorer [78], together with Interdecadal Pacific Oscillation (IPO), was introduced by Henley et al. [83], collected from NOAA National Climate Prediction Centre. Detailed information on climate indices and SSTs are in Table 2.

Table 2. Description of the global pool of 52 predictor variables used to design and evaluate hybrid CEEMDAN-CNN-GRU predictive model for daily surface soil moisture forecasting.

GLDAS 2.0: Modis Satellite Data from Giovanni Repository			
Predictor Variable	Notation	Description	Units
SurT	St	Average Surface Skin temperature	K
CSW	CW	Plant canopy surface water	Kg m ^{−2}
CWE	CE	Canopy water evaporation	kg m ^{−2} s ^{−1}
Esoil	Es	Direct Evaporation from Bare Soil	kg m ^{−2} s ^{−1}
ET	ET	Evapotranspiration	kg m ^{−2} s ^{−1}
Esnow	Es	Snow Evaporation	kg m ^{−2} s ^{−1}
GWS	GW	Groundwater storage	mm
LWR.	LW	Net longwave radiation flux	W m ^{−2}
Qg	Qg	Ground heat flux	W m ^{−2}
Qh	Qh	Sensible heat net flux	W m ^{−2}
Qle	Qle	Latent heat net flux	W m ^{−2}
Qs	Qs	Storm surface runoff	Kg m ^{−2} s ^{−1}
Qsb	Qb	Baseflow-groundwater runoff	Kg m ^{−2} s ^{−1}
Qsm	Qm	Snow-melt	Kg m ^{−2} s ^{−1}
Snd	Sn	Snow depth	m
Snt	Snt	Snow Surface temperature	m
SMp	Sp	Profile Soil moisture	Kg m ^{−2}
SMrz	Sz	Root Zone Soil moisture	Kg m ^{−2}
SSM	SSM	Surface Soil moisture	Kg m ^{−2}
SWE	SW	Snow depth water equivalent	Kg m ^{−2}
SWR	SR	Net short-wave radiation flux	W m ^{−2}
Tra	Tr	Transpiration	Kg m ^{−2} s ^{−1}
TWS	TW	Terrestrial water storage	mm
SILO (Ground-Based Observations)			
T.Max	Tx	Maximum Temperature	°C
T.Min	Tn	Minimum Temperature	°C
Rain	r	Rainfall	mm
Evap	Ep	Evaporation	mm
Radn	Rd	Radiation	MJ m ^{−2}

Table 2. Cont.

GLDAS 2.0: Modis Satellite Data from Giovanni Repository			
Predictor Variable	Notation	Description	Units
VP	VP	Vapour Pressure	hPa
RHmaxT	Rx	Relative Humidity at Temperature T.Max	%
RHminT	Rn	Relative Humidity at Temperature T.Min	%
Mpot	Mp	Morton potential evapotranspiration overland	mm
SYNOPTIC-SCALE (Climate Mode Indices)			
Nino3.0	N3	Average SSTA over 150°–90°W and 5°N–5°S	NONE
Nino3.4	N34	Average SSTA over 170°E–120°W and 5°N–5°S	
Nino4.0	N4	Average SSTA over 160°E–150°W and 5°N–5°S	
Nino1+2	N12	Average SSTA over 90°W–80°W and 0°–10°S	
AO	A	Arctic Oscillation	
AAO	AO	Antarctic Oscillation	
MJO1	MJ1	Madden Julian Oscillation-1	
MJO2	MJ2	Madden Julian Oscillation-2	
MJO4	MJ4	Madden Julian Oscillation-4	
MJO5	MJ5	Madden Julian Oscillation-5	
MJO6	MJ6	Madden Julian Oscillation-6	
MJO7	MJ7	Madden Julian Oscillation-7	
MJO8	MJ8	Madden Julian Oscillation-8	
MJO10	MJ10	Madden Julian Oscillation-10	
EPO	EP	East Pacific Oscillation	
GBI	G	Greenland Blocking Index (GBI)	
WPO	WP	Western Pacific Oscillation (WPO.)	
PNA	PN	Pacific North American Index	
NAO	N	North Atlantic Oscillation	
SAM	SM	Southern Annular Mode index	
SOI	SOI	Southern Oscillation Index, as per Troup [84]	

SSTA = Sea Surface Temperature anomalies (°C).

3.2. Predictive Model Development

To design a forecast model for SSM over multi-step periods of 1st, 5th, 7th, 14th, 21st, and 30th day lead time, three distinct datasets from satellites (i.e., GIOVANNI), climate indices, and ground source (SILO) for 17 years, 1 February 2003 to 31 March 2020 are used. Hybrid DL is implemented under Intel i7 @ 1.5 GHz and 16 GB memory. The proposed model algorithms were demonstrated using freely available DL libraries, namely the Keras [85,86] and TensorFlow [87] libraries. MATLAB 2020 software is used for Neighbourhood Component Analysis feature selection with packages matplotlib, and Minitab is used to visualise the forecasted SSM in the testing phase.

Data-driven models were built by normalising the input variables, transforming these predictors into a more consistent form [88]. To ensure the variable features were given proportional attention in network training, all were normalised [89] between (0, 1) [41,53,90].

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (15)$$

In Equation (15), x is the respective variable, x_{min} is the minimum value, x_{max} is the maximum and x_{norm} is the normalised value. After normalising the variables, the datasets are partitioned into training (February 2003–December 2013), validation (January 2014–December 2016), and testing (January 2017–March 2020) subsets. Figure 3 shows the methodological steps of the proposed CEEMDAN-CNN-GRU model. CEEMDAN is implemented in four stages.

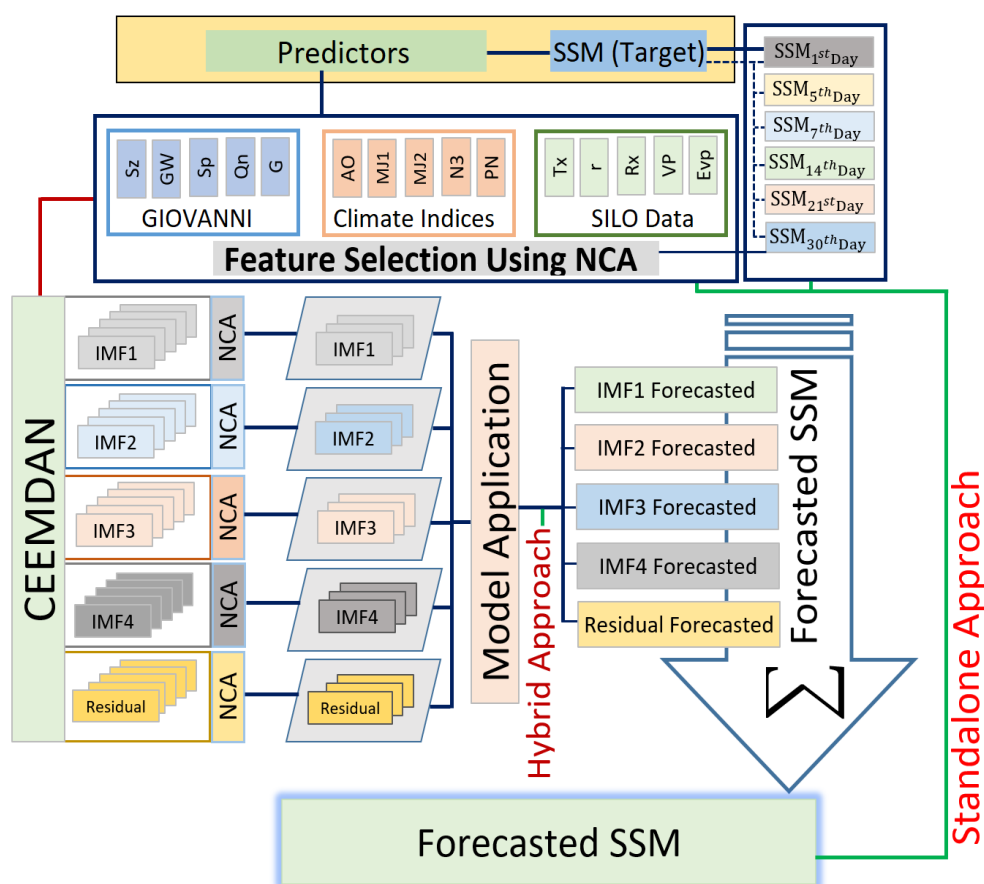


Figure 3. Workflow with the steps in model design for hybrid CEEMDAN-CNN-GRU predictive model. SSM = Surface Soil Moisture, NCA = neighbourhood component analysis for regression, IMF = Intrinsic Mode Function, CEEMDAN = Complete Ensemble Empirical Model Decomposition with adaptive noise, GRU = Gated Recurrent Units.

3.2.1. Feature Selection

By incorporating the MODIS satellite and ground and climate indices, this study has utilised 52 different predictors for SSM forecasting; hence, feature selection was crucial for data pre-processing. This is because irrelevant and redundant features increase the network size, congestion and cause a reduction in the algorithm's speed, reducing the efficiency of the predictive model [91]. Therefore, our study has used the NCA algorithm to screen an optimal set of predictor variables out of the 52-variable set. In general, f_{srnca} calculates every predictor's relative weight against a target (SSM), illustrated in Figure 4. Following this, the standalone GRU and hybrid CNN-GRU models were executed with

predictors added one by one from the highest feature to the lowest feature weight until an optimal performance was achieved. Figure 5 illustrates the the relative root mean squared error (RRMSE) value of different combinations prepared based on NCA. Tables A1–A6 shows the GRU and CNN-GRU model's performance accordingly.

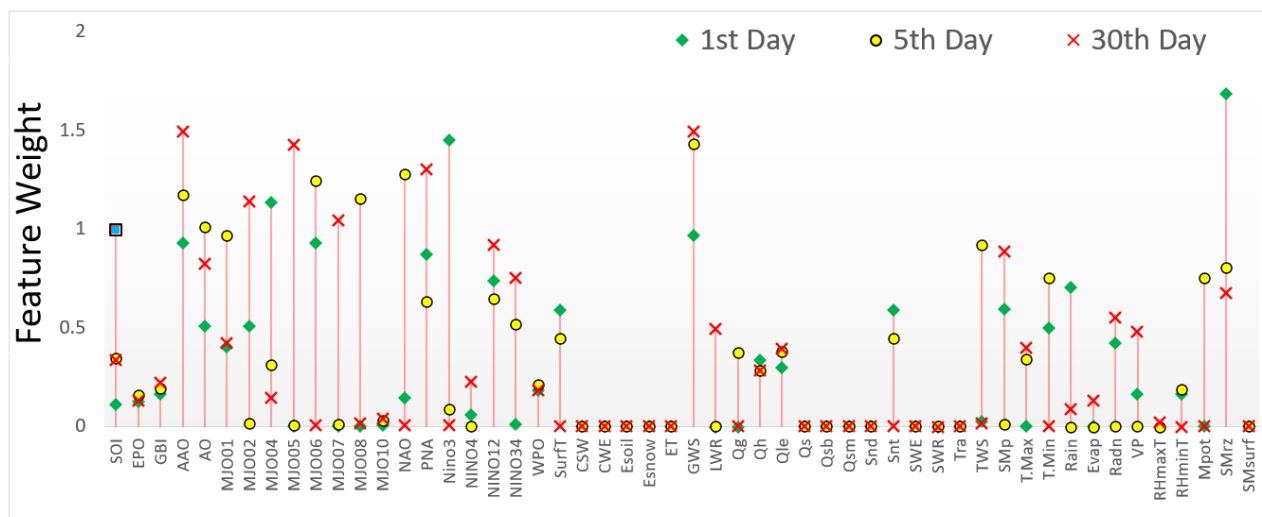


Figure 4. Feature weight matrix of predictor variables from a pool of 52 data sources using neighbourhood component analysis at the n th ($n = 1, 5$, and 30) day lead time forecasting of surface soil moisture shown for the case of Menindee study station. Details of the variables are mentioned in Table 2.

Figure 4 illustrates the respective feature weights of predictor variables, using the Menindee station as an example. For the 1st day of SSM forecasting, the root zone soil moisture (kg m^{-2}) is found to generate the highest feature weight, whereas, for the 5th day, groundwater storage (mm) is found to be the most significant feature weight. Notably, the groundwater storage contributed to the largest feature weighted for the 7th, 14th, 21st, and 30th day SSM forecasting. This evaluation indicates that groundwater has a strong influence on SSM over inter-daily scales. Tables S1–S6 illustrates the input combination for SSM forecasting in the n th day lead period with their respective forecasting performance with CNN-GRU and GRU model. It is imperative to note that *fsrnca* algorithm is used in two distinct phases before applying the hybrid-deep learning (i.e., CEEMDAN-CNN-GRU) model. In the first phase, *fsrnca* attains the feature weights and acquires the optimal predictor variable list required for SSM forecasts. Subsequently, the second phase incorporates the data decomposition process utilising CEEMDAN to each variable selected from the feature weights. Finally, the feature weight is calculated for IMF (t) deduced for each predictor variable against the objective variable (i.e., SSM). Here, the term t refers to the number of IMFs for each variable, removing four to five least significant features from the hybrid CEEMDAN-CNN-GRU model.

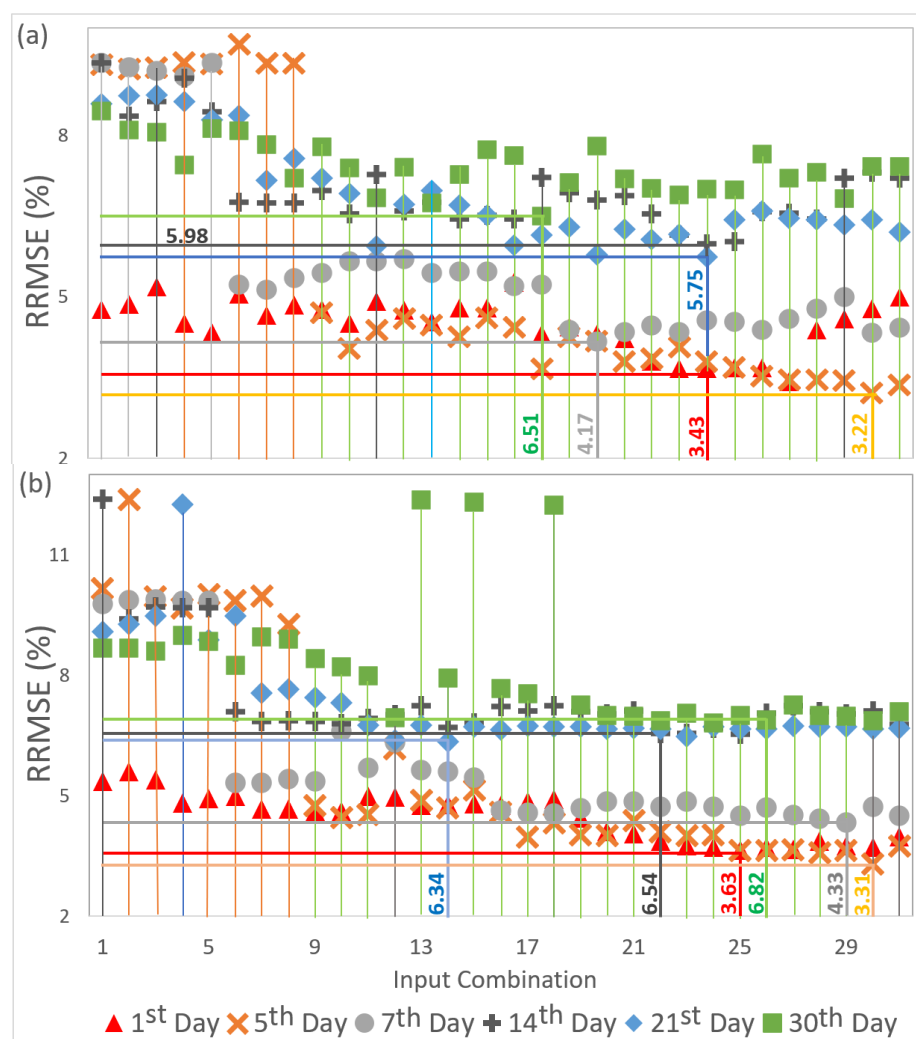


Figure 5. Stair plot showing the relative root mean squared error (RRMSE, %) for (a) CNN-GRU and (b) GRU applied at different input combinations for the Menindee station at the 1st, 5th, 7th, 14th, 21st and 30th day lead time.

3.2.2. Hybrid Deep Learning Algorithm Implementation

Before applying the CEEMDAN-CNN-GRU model in the 1st, 5th, 7th, 14th, 21st, and 30th day SSM forecasts, hyperparameter selection is undertaken through a grid search procedure whose theoretical descriptions are provided in Section 2. Table 3 shows the hyperparameters, optimal GRU architecture, and CNN-GRU with input combinations deduced from the feature weight matrix. Finally, the deep learning forecast model combining a data decomposition (i.e., CEEMDAN) stage with a three-layered feature extraction stage (i.e., CNN) and feature selection stage (i.e., *fsrnca*) is implemented to forecast SSM.

The proposed CEEMDAN-CNN-GRU model is implemented in four stages, as shown in Figure 3. Firstly, CEEMDAN is applied to decompose historical training data into IMFs and residual signals (Figure 1a) followed by segregation of each IMFs and residual, such as collecting all the IMF1 for predictor variables. The relative feature weights of respective IMFs related to IMF of the target variable (i.e., SSM) are determined. The optimal signal selection enables the algorithm to remove the least important feature-weighted IMFs, allowing the predictive model network to be noise-free. Finally, the forecasted SSM utilising the CEEMDAN-based model (i.e., the hybrid CEEMDAN-CNN-GRU) is obtained by aggregating the IMFs of the predictor variables. The robustness of the model is investigated by several evaluation criteria (Section 3.2.3).

Table 3. (a) Range of tested hyperparameters in designing hybrid CNN-GRU and GRU predictive models through trial and error method. (b) Optimally selected hyperparameters. ReLU stands for Rectified Linear Units, SGD stands for stochastic gradient descent optimiser.

(a) Tested Range of Model Hyper-Parameters		
Model	Model Hyper-parameter Names	Search Space for Optimal Hyper-Parameters
CNN-GRU	Filter 1	(70, 80, 100, 150)
	Filter 2	(70, 80, 100,150)
	Filter 3	(70, 80, 100, 150)
	GRU Cell Units	(40, 50, 70, 80, 100, 150)
	Epochs	(500, 800, 1000)
	Activation function	(ReLU)
	Optimiser	(Adam, SGD)
	Batch Size	(5, 10, 20, 50, 100)
GRU	GRU Cell 1	(70, 80, 100, 110)
	GRU Cell 2	(70, 80, 100,150, 200, 210)
	Epochs	(500, 800, 1000)
	Activation function	(ReLU)
	Optimiser	(Adam, SGD)
	Batch Size	(5, 10, 20, 50, 100)
(b) Optimally Selected Hyper-Parameters		
CNN-GRU	Convolution Layer 1 (C1)	80
	C1-Activation function	ReLU
	C1-Pooling Size	1
	Convolution Layer 2 (C2)	70
	C2-Activation function	ReLU
	C2-Pooling Size	1
	Convolution Layer 3 (C3)	80
	C3-Activation function	ReLU
	C3-Pooling Size	1
	GRU Layer 1 (L1)	200
	L1-Activation function	ReLU
	GRU Layer 2 (L2)	60
	L2-Activation function	ReLU
	Drop-out rate	0.2
	Optimiser	Adam
	Padding	Same
	Batch Size	5
	Epochs	400
GRU	GRU Cell 1 (G1)	110
	G1-Activation function	ReLU
	GRU Cell 2 (G2)	250
	G2-Activation function	ReLU
	Epochs	300

Table 3. Cont.

(a) Tested Range of Model Hyper-Parameters	
Optimiser	SGD
Drop-out rate	0.2
Batch Size	15
Epochs	1000

It is worth noting that climate indices (CIs) have a notable signature of climate variability in Australia, leading to substantial influence on rainfall and a potential effect on future surface soil moisture patterns. In the final task, climate indices' relative contribution to building the CEEMDAN-CNN-GRU model is assessed by Multivariate Adaptive Regression Splines (MARS) utilising the ARESLab toolbox. Following Friedman [86], MARS can determine each predictor variable's significance by evaluating its complex and non-linear interaction with the target (i.e., SSM) based on best regressors and provide the importance of each variable. The relative importance of any predictor variable is the square root of GCV (Generalised Cross-Validation) with all basic functions involving the respective variable minus the root square of the GCV score of that full model. However, this process is scaled in such a way that the relative importance has a value of 100, expressed:

$$GCV = \frac{MSE}{\left(1 - \frac{enp}{N}\right)^2} \quad (16)$$

Here, enp is the significant number of model parameters, $p = k + c(k - 1)/2$; k = basis function in MARS model; c = penalty (set to 2 or 3). However, if enp is greater or equal to N , GCV is an Inf, which indicates the model is flawed [92].

3.2.3. Predictive Model Evaluation

The efficacy of deep learning hybrid model is evaluated using different performance evaluation criteria e.g., Pearson's Correlation Coefficient (r), root mean square error (RMSE), Nash-Sutcliffe efficiency (NSE) [93], mean absolute error (MAE), and Kling-Gupta efficiency [94]. Due to geographic differences between the study stations, we employ relative error-based metrics: i.e., relative RMSE (denoted as RRMSE) and relative MAE (denoted as RMAE). The appraisal of a predictive model's efficacy depends on the exactness between the predicted and observed values. RMSE is an appropriate measure of model performance compared to MAE when the error distribution in the tested data is Gaussian [95] but for an improved model evaluation, the Willmott's Index (WI) and Legates-McCabe's (LM) Index are used as more sophisticated and compelling measures [96,97]. Mathematically, the metrics are as follows:

Correlation coefficient (r):

$$r = \left\{ \frac{\sum_{i=1}^N (SSM_{obs} - \overline{SSM}_{obs}) (SSM_{for} - \overline{SSM}_{for})}{\sqrt{\sum_{i=1}^N (SSM_{obs} - \overline{SSM}_{obs})^2 \sum_{i=1}^N (SSM_{for} - \overline{SSM}_{for})^2}} \right\}^2 \quad (17)$$

Mean absolute error (MAE: kg m^{-2}):

$$MAE = \frac{1}{N} \sum_{i=1}^N |SSM_{for} - SSM_{obs}| \quad (18)$$

Root mean squared error (RMSE: kg m^{-2}):

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (SSM_{for} - SSM_{obs})^2} \quad (19)$$

Nash-Sutcliffe Efficiency (NSE):

$$NSE = 1 - \left[1 - \frac{\sum_{i=1}^N (SSM_{for})^2}{\sum_{i=1}^N (SSM_{obs} - \overline{SSM}_{for})^2} \right] \quad (20)$$

Kling-Gupta efficiency (KGE):

$$KGE = 1 - \sqrt{(r-1)^2 + \left(\frac{\overline{SSM}_{for}}{\overline{SSM}_{obs}} - 1 \right)^2 + \left(\frac{CV_p}{CV_s} \right)^2} \quad (21)$$

Mean Absolute Percentage Error (MAPE, %):

$$MAPE = \frac{1}{N} \left(\sum_{i=1}^N \left| \frac{SSM_{for} - SSM_{obs}}{SSM_{obs}} \right| \right) \times 100, (0\% \leq MAPE \leq 100\%) \quad (22)$$

Willmott's Index (WI):

$$WI = 1 - \left[\frac{\sum_{i=1}^N (SSM_{for} - SSM_{obs})^2}{\sum_{i=1}^N (|SSM_{for} - \overline{SSM}_{obs}| + |SSM_{obs} - \overline{SSM}_{obs}|)^2} \right] \quad (23)$$

Legates-McCabe's Index (LM):

$$LM = 1 - \left[\frac{\sum_{i=1}^N |SSM_{for} - SSM_{obs}|}{\sum_{i=1}^N |SSM_{obs} - \overline{SSM}_{obs}|} \right] \quad (24)$$

Relative Root Mean Squared Error (RRMSE, %):

$$RRMSE(\%) = \frac{\sqrt{\frac{1}{N} \sum_{i=1}^N (SSM_{for} - SSM_{obs})^2}}{\frac{1}{N} \sum_{i=1}^N (SSM_{obs})} \times 100 \quad (25)$$

Relative Mean Absolute Error (RMAE, %):

$$RMAE(\%) = \frac{\frac{1}{N} \sum_{i=1}^N |SSM_{for} - SSM_{obs}|}{\frac{1}{N} \sum_{i=1}^N (SSM_{obs})} \times 100 \quad (26)$$

Absolute percentage bias (APB, %):

$$APB = \left[\frac{\sum_{i=1}^N |SSM_{obs} - SSM_{for}| \times 100}{\sum_{i=1}^N |SSM_{obs}|} \right] \quad (27)$$

In Equations (17)–(27), SSM_{obs} and SSM_{for} represents the observed and forecasted values for i th test value; \overline{SSM}_{obs} and \overline{SSM}_{for} refer to their averages, accordingly, and N is defined as the number of observations, while CV stands for the coefficient of variation. CV is a standardised measure of the dispersion of the frequency distribution.

4. Results

The practical utility of the hybrid DL (i.e., CEEMDAN-CNN-GRU) model is established by integrating diverse data in its training and model testing phase. Significant features from predictor variables are used by incorporating NCA, and the predictive model is evaluated using statistical metrics (Equations (17)–(27)), infographics, and visualisations to appraise the degree of agreements between simulated and observed soil moisture. By

several measures, the CEEMDAN-CNN-GRU model appears to outperform all the comparative models with superior r and NSE and low RMSE, MAE, and APB in the testing phase. An extensive analysis of tabulated results (Table 4) provides convincing arguments that the hybrid deep learning method is effective for surface soil moisture forecasts and can perhaps be a potential tool in agriculture water management. However, among all study sites, the CEEMDAN-CNN-GRU model for the Menindee station showed the best performance, considering r (0.996), NSE (0.995), and lowest RMSE (0.021), MAE (0.013), and APB (0.359) values for the 1st day of SSM forecasting. The performance of this model is followed by the CEEMDAN-GRU and CNN-GRU model.

For the 5th day of SSM forecasting, the results of the objective model for Menindee had the best performance ($r = 0.993$; $NSE = 0.991$; $RMSE = 0.040 \text{ kg m}^{-2}$) followed by Deniliquin ($r = 0.989$; $NSE = 0.975$; $RMSE = 0.091 \text{ kg m}^{-2}$). Likewise, for the 7th, 14th, 21st, and 30th days of SSM forecasting, the CEEMDAN-CNN-GRU model outperformed the other models by a notable margin for all the respective periods of SSM forecasting. However, a site-specific signature in the model accuracy was also evident, with the results for Menindee registering the lowest value of RMSE generated by the CEEMDAN-CNN-GRU model. In terms of MAE, the CEEMDAN-CNN-GRU model returned the lowest value for Menindee, suggesting that the CEEMDAN-CNN-GRU model was a potential forecasting tool SSM at the 1st, 5th, and 7th day ahead periods. Not surprisingly, in accordance with other studies, e.g., the present study indicates that as the length of the forecasting period was increased, the model's performance appear to reduce at a significant rate in such a way that the r -values reduced by 0.30%, 1.10%, 9.15%, 11% and 15% for the 1st to 5th, 7th, 14th, 21st and 30th day of SSM forecasting. The change of the performance metrics (i.e., NSE, MAE, and APB) for longer-term horizons relative to the shorter-term horizons also concurred with the respective changes in the r -values and is consistent with earlier studies [60,98]. For a longer-term horizon, the present r value was lower, and the MAE increased, suggesting that for the longer forecast horizon, the model appeared to lose the relevant data features in the predictor variables required to maintain precise SSM forecasting performance. The hybrid CEEMDAN-CNN-GRU model is further evaluated using a probability plot of errors at the 95th percentile, including those of the benchmark model (i.e., CNN-GRU, CEEMDAN-GRU) and the standalone model (i.e., GRU) with an illustration for Menindee at the different n th ($n = 1, 5, 7, 14, 21$ and 30) days (Figure 6). The CEEMDAN-CNN-GRU model results show that ~95% of SSM forecasting had the lowest error (<0.1) for the 1st and 5th days of SSM forecasting. Among all the predictive models and the forecast periods over n th days, the GRU-based model showed a more significant proportion of $|FE|$ values at a 95% confidence level. Notably, consistently good results were also achieved for the other stations (i.e., Deniliquin, Fairfield, and Gabo Island), which are shown in supplementary materials (Figure S1a–c). The lowest value of $|FE|$, with <0.063 with a 95% percentile, was evident for Fairfield compared to the other two study stations. The correlation between observed and forecasted daily surface soil moisture datasets generated by the proposed CEEMDAN-CNN-GRU model vs. the corresponding benchmark models (i.e., CNN-GRU and GRU), for the case of Menindee station, is illustrated in Figure 7. The correlations for the hybrid GRU model are positioned close to the observed SSM values up to the 7th day, revealing a high degree of forecasting accuracy. An improvement in the model's forecasting performance was attained by applying the CNN algorithm (i.e., soil moisture generated by the CNN-GRU model) and data decomposition (i.e., CEEMDAN-CNN-GRU) method on standalone GRU model. The disparity between the forecasted SSM and the reference SSM values was significantly higher for the 14th, 21st, and 30th days of SSM forecasting, which concurs with earlier metrics suggesting a potential inadequacy of the data features long time ahead periods [60].

Table 4. Evaluation of hybrid CEEMDAN-CNN-GRU vs. benchmark (CNN-GRU, CEEMDAN-GRU, GRU) models for the specific case of Menindee study site. The correlation coefficient (r), root mean square error (RMSE; Kg m^{-2}), mean absolute error (MAE; Kg m^{-2}), and Nash-Sutcliffe coefficient, (NS) is computed between forecasted and observed surface soil moisture for the 1st day, 5th day, 7th day, 14th day, 21st day, and 30th day ahead periods in the testing phase. The optimal model is boldfaced.

Soil Moisture Forecasting Horizon, n th Day Lead Time																														
1st Day					5th Day					7th Day					14th Day					21st Day					30th Day					
r	NSE	RMSE	MAE	APB	r	NSE	RMSE	MAE	APB	r	NSE	RMSE	MAE	APB	r	NSE	RMSE	MAE	APB	r	NSE	RMSE	MAE	APB	r	NSE	RMSE	MAE	APB	
Study Station 1: Menindee																														
CEEMDAN-CNN-GRU	0.996	0.995	0.021	0.013	0.359	0.993	0.991	0.040	0.030	0.823	0.985	0.967	0.075	0.057	1.559	0.906	0.896	0.226	0.185	5.079	0.895	0.787	0.230	0.186	5.098	0.869	0.714	0.255	0.201	5.493
CNN-GRU	0.967	0.892	0.135	0.112	3.061	0.966	0.918	0.117	0.094	2.569	0.945	0.861	0.152	0.121	3.330	0.892	0.770	0.235	0.193	5.285	0.899	0.788	0.210	0.168	4.594	0.851	0.765	0.238	0.181	4.945
CEEMDAN-GRU	0.976	0.937	0.116	0.094	2.234	0.970	0.933	0.120	0.095	2.265	0.957	0.909	0.140	0.110	2.613	0.882	0.738	0.237	0.186	4.424	0.864	0.781	0.262	0.206	4.918	0.866	0.742	0.275	0.217	5.163
GRU	0.962	0.893	0.134	0.110	3.020	0.962	0.933	0.121	0.094	2.589	0.940	0.851	0.158	0.126	3.452	0.882	0.745	0.244	0.197	5.390	0.887	0.748	0.243	0.196	5.360	0.863	0.726	0.251	0.197	5.386
Study Station 2: Deniliquin																														
CEEMDAN-CNN-GRU	0.990	0.899	0.048	0.034	0.778	0.989	0.975	0.091	0.065	1.489	0.959	0.917	0.165	0.113	2.611	0.801	0.607	0.355	0.247	5.716	0.768	0.573	0.374	0.266	6.130	0.703	0.465	0.415	0.295	6.807
CNN-GRU	0.979	0.955	0.098	0.075	1.799	0.945	0.866	0.169	0.137	3.270	0.929	0.846	0.181	0.143	3.405	0.866	0.624	0.283	0.224	5.333	0.873	0.749	0.231	0.181	4.298	0.848	0.687	0.258	0.202	4.806
CEEMDAN-GRU	0.987	0.958	0.106	0.081	1.930	0.968	0.929	0.123	0.096	2.279	0.969	0.920	0.131	0.106	2.524	0.872	0.730	0.240	0.189	4.505	0.859	0.712	0.249	0.197	4.701	0.869	0.671	0.264	0.207	4.926
GRU	0.967	0.927	0.125	0.099	2.350	0.947	0.889	0.154	0.121	2.874	0.918	0.822	0.195	0.153	3.655	0.867	0.722	0.244	0.191	4.560	0.868	0.695	0.256	0.201	4.787	0.850	0.659	0.269	0.217	5.152
Study Station 3: Fairfield																														
CEEMDAN-CNN-GRU	0.975	0.976	0.035	0.024	0.554	0.972	0.975	0.069	0.052	1.189	0.959	0.920	0.162	0.110	2.524	0.842	0.628	0.349	0.238	5.493	0.762	0.573	0.374	0.264	6.088	0.746	0.523	0.374	0.261	6.078
CNN-GRU	0.945	0.935	0.061	0.048	1.099	0.962	0.943	0.135	0.091	2.107	0.907	0.821	0.240	0.156	3.612	0.764	0.560	0.376	0.264	6.109	0.759	0.554	0.379	0.259	5.988	0.708	0.477	0.410	0.289	6.671
CEEMDAN-GRU	0.947	0.943	0.048	0.034	0.778	0.939	0.935	0.091	0.065	1.489	0.929	0.917	0.165	0.113	2.611	0.801	0.607	0.355	0.247	5.716	0.768	0.573	0.374	0.266	6.130	0.703	0.465	0.415	0.295	6.807
GRU	0.925	0.919	0.153	0.096	2.205	0.913	0.905	0.177	0.115	2.659	0.904	0.809	0.250	0.168	3.864	0.778	0.585	0.369	0.254	5.850	0.775	0.568	0.376	0.267	6.165	0.666	0.411	0.435	0.314	7.267
Study Station 4: Gabo Island																														
CEEMDAN-CNN-GRU	0.988	0.966	0.085	0.067	1.455	0.987	0.971	0.079	0.062	1.346	0.978	0.944	0.109	0.086	1.887	0.931	0.899	0.188	0.147	3.206	0.909	0.764	0.224	0.175	3.829	0.913	0.807	0.202	0.158	3.456
CNN-GRU	0.979	0.951	0.101	0.078	1.707	0.973	0.944	0.109	0.084	1.826	0.948	0.897	0.147	0.113	2.457	0.921	0.843	0.182	0.141	3.087	0.911	0.803	0.204	0.160	3.493	0.879	0.862	0.193	0.151	3.284
CEEMDAN-GRU	0.986	0.966	0.085	0.067	1.472	0.983	0.964	0.087	0.069	1.508	0.974	0.945	0.107	0.085	1.844	0.924	0.821	0.194	0.153	3.340	0.913	0.814	0.198	0.156	3.394	0.912	0.798	0.206	0.161	3.520
GRU	0.977	0.950	0.102	0.081	1.773	0.970	0.940	0.113	0.086	1.868	0.951	0.902	0.144	0.111	2.423	0.919	0.825	0.192	0.150	3.283	0.912	0.813	0.199	0.156	3.411	0.815	0.743	0.203	0.160	3.499

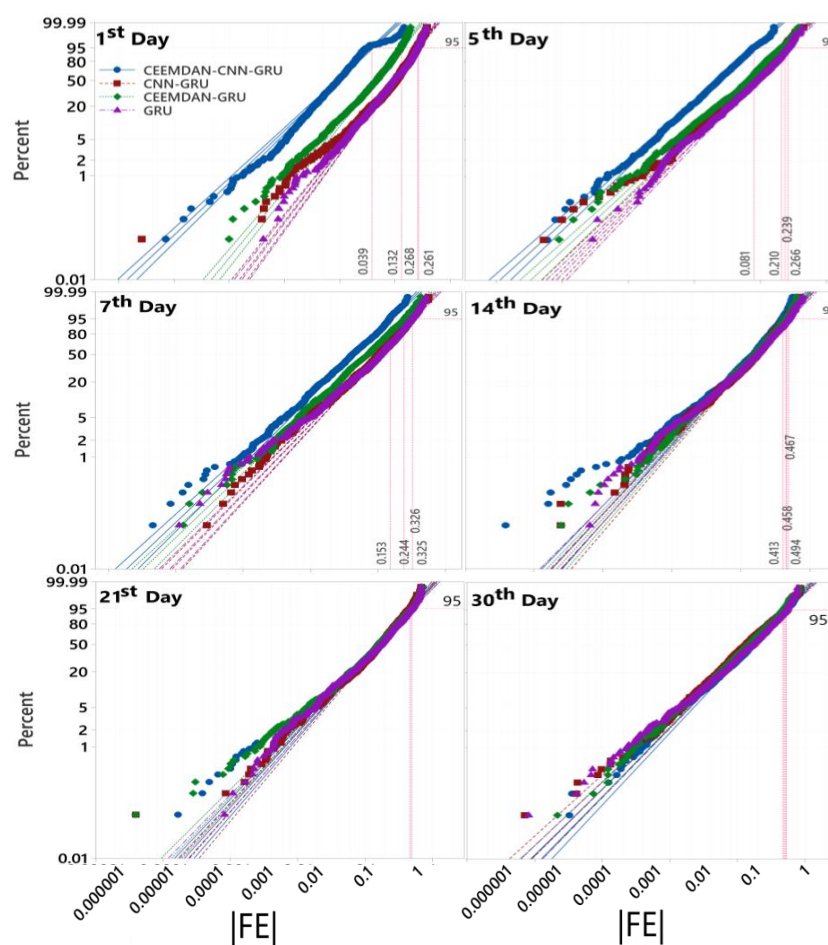


Figure 6. Probability plot (95 percentiles) for hybrid CEEMDAN-CNN-GRU, CNN-GRU, CEEMDAN-GRU, and GRU model for Menindee at different n th ($n = 1, 5, 7, 14, 21$ and 30) day lead time.

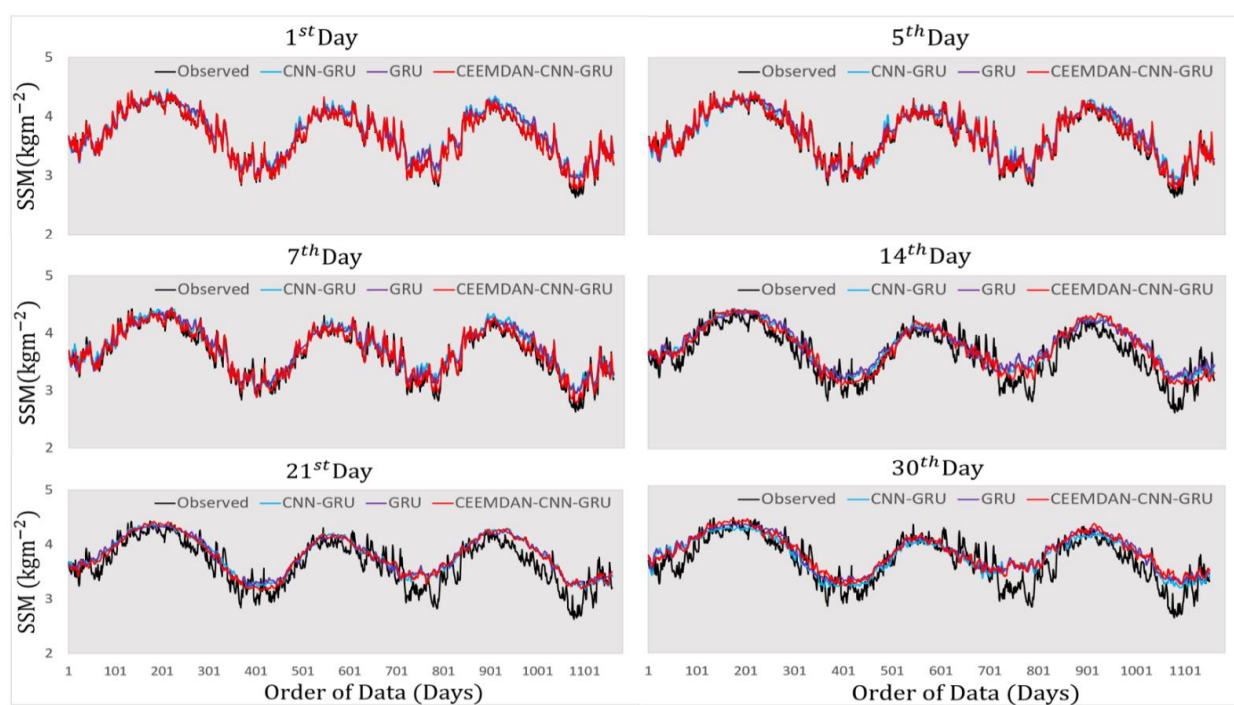


Figure 7. Time series of daily surface soil moisture (SSM, kg m^{-2}) for observed SSM (Gray) and Figure 1. 5, 7, 14, 21, and 30) day lead times.

Figure 8 shows a scatter plot of forecasted and observed SSM for the 1st and 7th days across the Murray Darling Basin with a least square regression line, $y = mx + C$, and the coefficient of determination in each sub-panel. Notably, the objective model (i.e., CEEMDAN-CNN-GRU) is seen to attain more accurate results with considerably larger r^2 values. The SSM forecast with a hybrid deep learning model for Menindee station performed significantly better than the comparative model (i.e., CNN-GRU). In the case of Menindee, for example, the values for m and r^2 are in reasonably good agreement against the 1:1 line representing the forecasted and observed SSM values in such a way that $(m | r^2)$ is $0.994 | 0.995$ for the hybrid CEEMDAN-CNN-GRU model relative to $(0.931 | 0.933)$ for CNN-GRU for the 1st day ahead of SSM forecasting. Moreover, for the 1st day of SSM forecasting, the CEEMDAN-CNN-GRU model provided results in significant proximity to the other three stations, such as Deniliquin: $0.962 | 0.966$, Fairfield: $0.928 | 0.964$, and the Gabo Island: $0.958 | 0.976$). Alternatively, the y-intercept of the regression line was close to trivial, i.e., 0.002 (Menindee: 1st day), 0.193 (Deniliquin: 1st day), 0.05 (Fairfield: 1st day), and 0.303 (Gabo Island: 1st day), revealing the efficacy of the deep learning hybrid method for surface soil moisture forecasting. For the 14th, 21st, and 30th day ahead of SSM forecasting, the y-intercept, as expected, deviated slightly from the ideal value of 0, caused by more outliers between simulated and reference values in the testing phase.

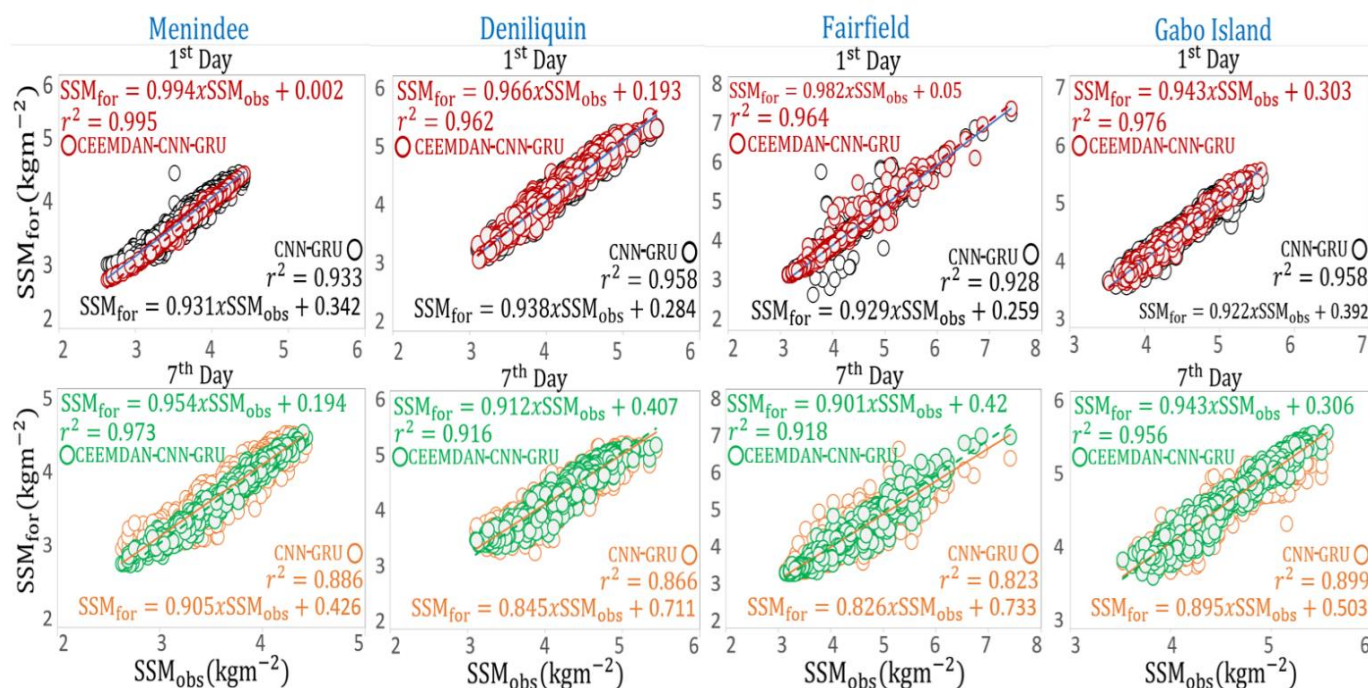


Figure 8. Scatter plot of the forecasted and observed SSM for Menindee, Deniliquin, Fairfield, and Gabo Island stations at different n th ($n = 1$ and 7) day ahead. A least square regression line, $y = mx + C$, and coefficient of determination (R^2) is shown in each sub-panel.

To further analyse the tested models' performances, we adopt the Legates and McCabe's Index [99] as a cross-validation metric for simulated data. This metric has a better model penalisation skill when high SSM values are expected in the testing set [41]. This is illustrated in Figure 9 in terms of a polar plot of the LM values for the hybrid deep learning approach (i.e., CEEMDAN-CNN-GRU) and other models for the different day ahead forecasting. The LM values accumulated across all stations in the case of CEEMDAN-CNN-GRU have a superior result with the highest LM ≈ 0.962 for Menindee and the lowest LM for the case of Gabo Island (LM ≈ 0.846) in the 1st Day ahead SSM forecasting. In agreement with earlier results, the LM values for the 14th, 21st, and 30th day ahead for other models were comparatively smaller. Figure 10a,b is a contour plot of KGE and MAPE for the hybrid DL approach (i.e., CEEMDAN-CNN-GRU) along with its benchmark (i.e.,

CNN-GRU) and standalone (i.e., GRU) methods for all four stations in MDB at different n th ($n = 1, 5, 7, 14, 21$ and 30) days in forecasting SSM. This infographic verifies the robustness of the proposed objective model that attains the highest KGE values and the lowest MAPE values for 1st and 5th day of SSM forecasting.

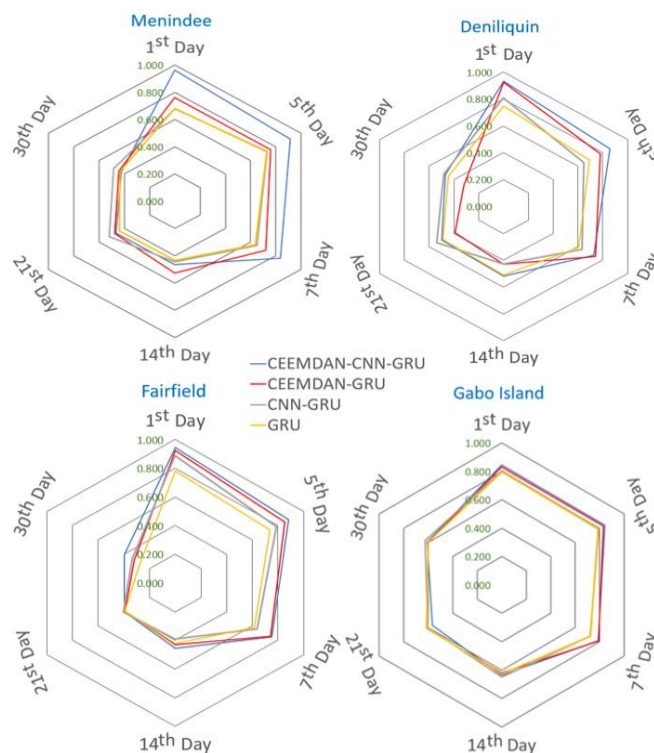


Figure 9. Polar plot showing the Legates and McCabe's Index (LM) in the testing period computed for the hybrid CEEMDAN-CNN-GRU against comparative models at different n th ($n = 1, 5, 7, 14, 21$, and 30) day ahead forecasting of SSM.

However, for the 14th, 21st, and 30th day of SSM forecasting, the KGE values range between 0.40 and 0.80, and the MAPE values range from 4–11%, demonstrating a slightly lower forecast accuracy relative to the 1st and 5th day of SSM forecasting. Figure 11 illustrates the absolute forecasted error ($|FE|$) using all the four candidate study sites' implemented models. The box plot demonstrates the data dispersal in terms of the forecasted (SSM_{for}) SSM. Figure 11 provides a clear visualisation of the closed distribution of error values for Menindee and Fairfield stations in the hybrid CEEMDAN-CNN-GRU model for 1st day ahead SSM forecasting. The lower end of the plot for $|FE|$ is situated within the lower quartile (25th) and upper quartile (75th). Moreover, the GRU and CNN-GRU models for these stations show an increased distribution of $|FE|$, except for the Fairfield station. Moreover, the forecasting of SSM for the 14th, 21st, and 30th day periods have a comparatively higher value of the absolute forecasting error for all tested models. A more comprehensive inspection of the absolute forecasting error ($|FE|$) in the case of the hybrid GRU model for the four study stations further cements the suitability of the CEEMDAN-CNN-GRU model in forecasting SSM for the 1st, 5th, and 7th day ahead periods in Australian Murray Darling Basin, evidenced by the narrowest error distribution in comparison with the other models.

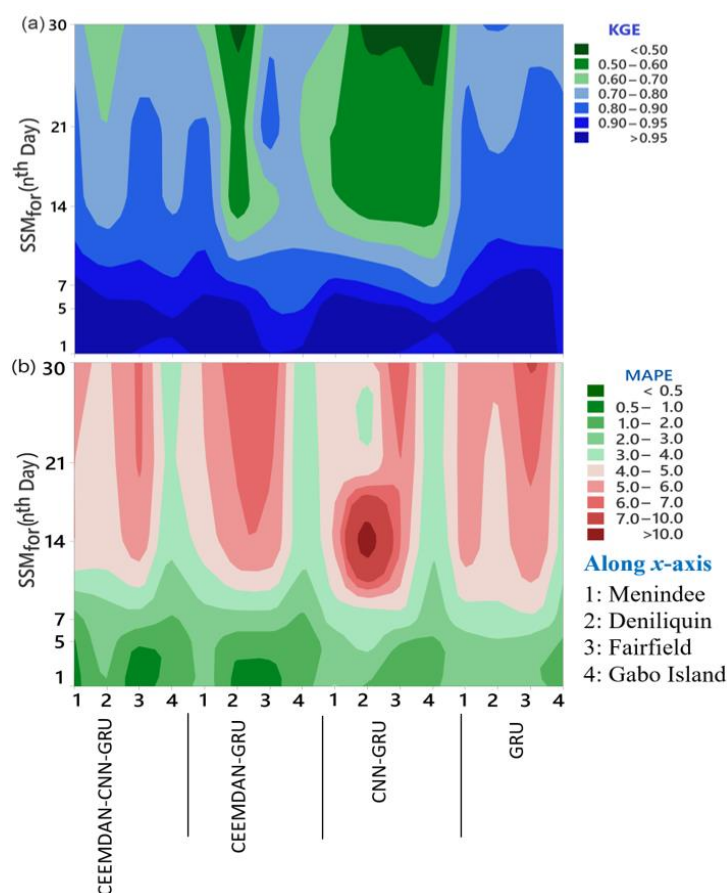


Figure 10. Contour plot of (a) KGE, (b) MAPE for hybrid CEEMDAN-CNN-GRU model against comparative models for different n th ($n = 1, 5, 7, 14, 21$, and 30) day ahead forecasting of SSM.

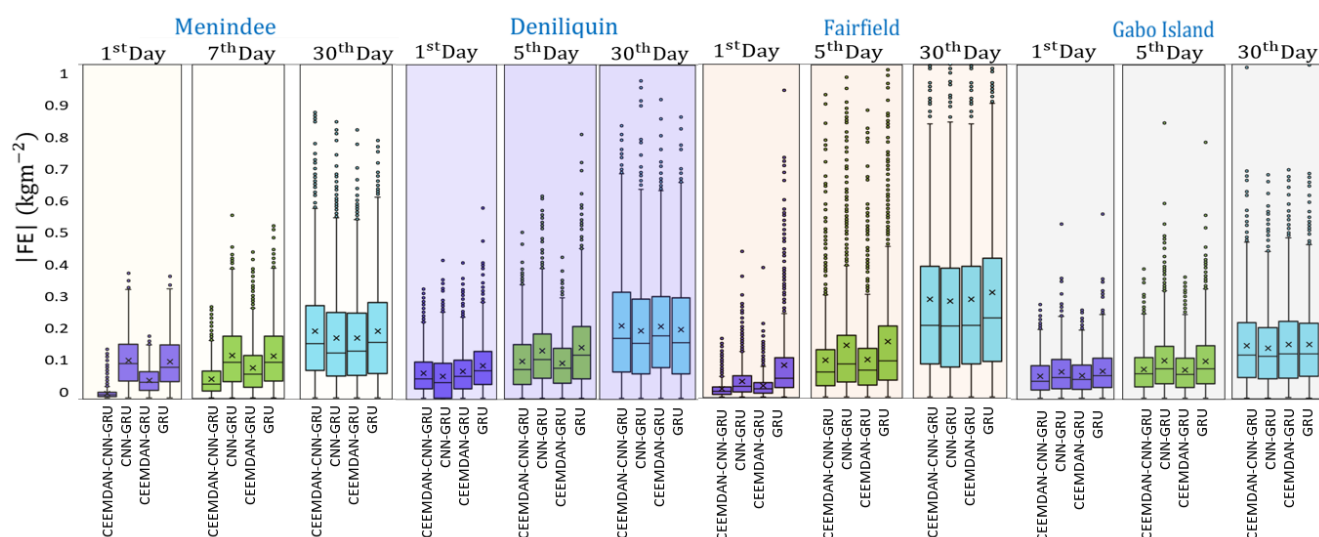


Figure 11. Box plot of errors in the testing phase for hybrid CEEMDAN-CNN-GRU against comparative models at different n th ($n = 1, 7$, and 30) day ahead lead time forecasting SSM. (Note: CEEMDAN-CNN-GRU = Hybrid Model integrating the CEEMDAN and CNN algorithm with GRU; CEEMDAN-GRU = Hybrid Model integrating the CEEMDAN algorithm with GRU; CNN-GRU = Hybrid Model integrating the CNN algorithm with GRU).

It is noteworthy that in this study, two distinct algorithms, namely the CEEMDAN and CNN, are used to improve the GRU-based predictive model. Therefore Figure 12

shows the effect of applying CEEMDAN and CNN as data pre-processing and feature extraction methods incrementally, respectively, on the per cent change in RMAE values within the testing SSM values. In terms of 1st, 5th, and 7th day of Menindee station, the RMAE (%) values of CEEMDAN-CNN-GRU model (where both CEEMDAN and CNN are integrated) appeared to decrease by ~87%, 68%, and 54%, respectively. Similarly, for the 1st-day forecasting taking the example of Fairfield station, the CNN feature-extraction skill reduced the error of ~55%, whereas an additional decrease in RMAE of ~18% was noted integration of the CEEMDAN selected variables (CEEMDAN-CNN-GRU). Additionally, for Deniliquin and Gabo Island study sites, the SSM forecasting for the 1st day ahead evaluated through RMAE values decreased by slightly less than 20%. It is worth mentioning that the per cent increase in RMAE was ~5% for Menindee for the 30th day ahead SSM forecasting with similar deductions for the other stations.

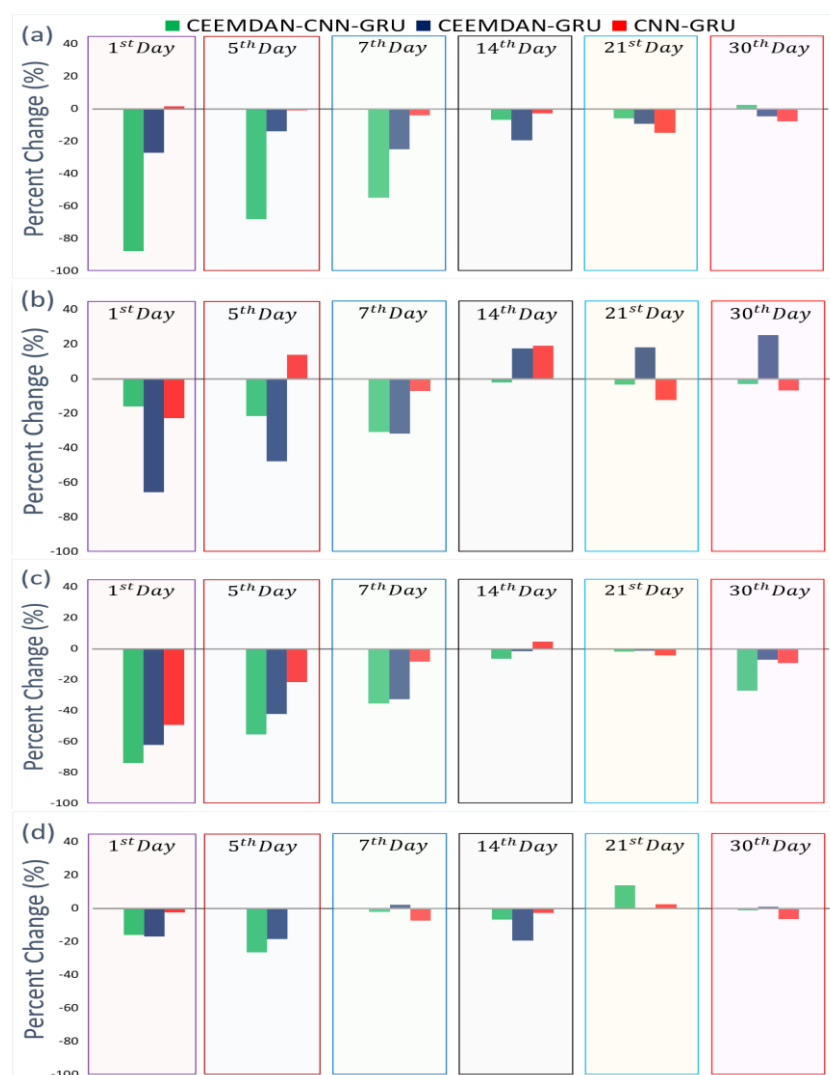


Figure 12. The percentage change in RMAE generated by the objective, and benchmark models using CEEMDAN and CNN methods (as data decomposition and feature extraction methods) adopted in forecasting SSM at four study sites: Murray Darling Basin. (a) Menindee, (b) Deniliquin, (c) Fairfield, (d) Gabo Island at different n th ($n = 1, 5, 7, 14, 21$, and 30) day ahead forecasting SSM.

We further show the CEEMDAN-CNN-GRU hybrid model's skill for seasonal forecasting for the different day ahead periods to better understand the seasonal effects of models used in SSM prediction. Figure 13 displays the average observed vs forecasted SSM on a seasonal basis (i.e., austral summer, autumn, winter, and spring) generated by CEEMDAN-

CNN-GRU model in case of Menindee study site. The forecast error across these seasons is relatively insignificant, occupying values of $(0, 0.16) \text{ kg m}^{-2}$ to demonstrate the exceptional skill of the objective model. Notably, the 1st and 5th day ahead of observed and forecasted SSM for austral summer, spring, winter, and autumn appear to match with the forecast error $(|FE|) < 0.04 \text{ kg m}^{-2}$, whereas, for winter, the $|FE|$ values are slightly higher for the 5th day ahead SSM forecasting. Not surprisingly, the CNN-GRU model possesses a larger error, ranging from 0.04 to 0.18 kg m^{-2} , establishing the CNN-GRU model's relatively poor performance compared with the hybrid CEEMDAN-CNN-GRU model. For the case of the 30th day ahead SSM forecasting, the study site Menindee registered a higher uncertainty for austral summer ($0.18 < |FE| < 0.18 \text{ kg m}^{-2}$) compared with winter and spring ($0.14 < |FE| < 0.15 \text{ kg m}^{-2}$). This indicates that the hybrid CEEMDAN-CNN-GRU model developed with NCA and CEEMDAN algorithms employing MODIS-derived satellite data, ground-based observations, and climate indices can be considered ideal in multi-step SSM forecasting.

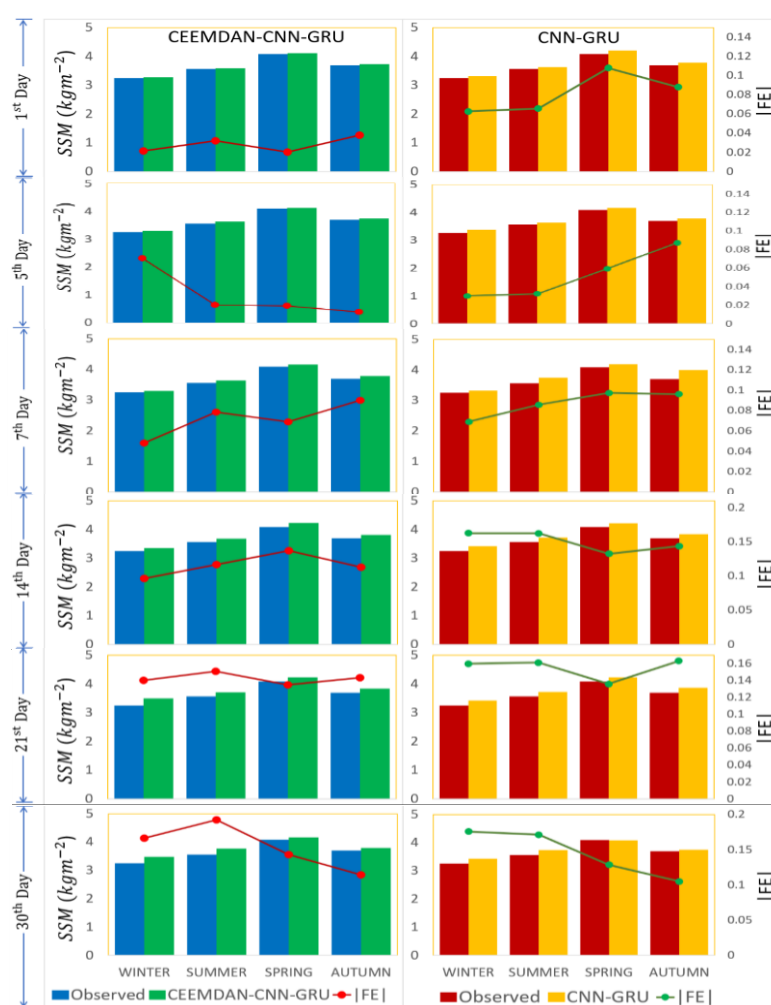


Figure 13. The average forecasted SSM vs. observed SSM on a seasonal basis using hybrid CEEMDAN-CNN-GRU and CNN-GRU models for Menindee at different n th ($n = 1, 5, 7, 14, 21$, and 30) day ahead periods. The forecast error ($|FE|$) in each model is plotted on a secondary axis as a line chart.

5. Discussions

Based on the results, we note the effects of climate indices on surface soil moisture as non-negligible. In this paper, analysing this impact is undertaken using two ways. Firstly, the NCA algorithm provides key information about how climate indices affect SSM. For

example, for SSM forecasting, climate indices based on SOI, EPO, MJOs and SST were found to significantly affect the SSM. Secondly, GCV values based on a MARS model were calculated following Friedman [100] approach to deduce the importance of input features. The contributory influence appeared to be between 12% and 53% according to GCV for Menindee station, and similarly notable effect for the other study sites. Specifically, the lowest percentage of ~12% of GCV was found for the 14th, and the highest percentage (~53%) was found for the 28th Day ahead SSM forecasting. In a nutshell, we note that climate indices make a moderate to high contribution in forecasting surface soil moisture within the Murray Darling Basin.

Neighbourhood Component Analysis (NCA) was utilised to examine significant features from a relatively large pool (or 52 different) data related to soil moisture. In data-driven modelling, selecting predictor variables is crucial, as improper variables with weak relationships against SSM can lead to undesirable uncertainties in the model. As per evaluations in Tables S1–S6, a combination of predictor variables deduced by NCA at six different lead time SSM forecasting was significant, and this result concurred with previous studies [59,101].

The objective approach based on NCA yielded good accuracy (i.e., CEEMDAN-CNN-GRU), demonstrating that best predictors were attained through a careful variable selection stage (by NCA) and feature extraction stage (by CNN and CEEMDAN methods). Accordingly, the proposed forecast model for SSM was sufficiently robust in daily and seasonal tests, as well as through the inclusion of synoptic-scale features, i.e., those captured from patterns in the SST and MJO series. The probability of absolute error placing within the 95th percentile and the substantial seasonal forecasting of SSM indicates that the model can handle satellite-derived variables' error. Our study also suggests that groundwater recharge, deep percolation, and plant uptake, which are essential factors to concentrate soil moisture in different layers [57], can be ideal variables to better understand SSM characteristics while also assisting in the prediction of future changes.

The present model's performance revealed that a shorter period forecast (i.e., 1st, 5th, or 7th) was more precise, whereas a longer forecast horizon (i.e., 14th, 21st, and 30th) registered a lower accuracy than that of the shorter span of SSM forecasting. One plausible reason for this is that our predictive model appeared to struggle to capture enough input features from the dataset for a more extended time-step forecast (i.e., 30th day against 7th day). Considering the reduction in feature capturing capability of the model, we can say that as the time series data approached close to the 7th-day boundary, the model would capture it with good forecast accuracy. Undoubtedly, this occurs due to a loss of data features in the predictor-target matrix. This indeed concurs with earlier studies (e.g., [60,98], where models for the 1- and 2-day ahead modelling horizon was more accurate than the 30-day horizon for river flow forecasting, and the 1- and 3-month runoff model was more accurate than the 6-month runoff model predicting 1-, 3-, and 6-month ahead runoff in the Yingluoxia watershed, Northwestern China. The hybrid deep learning approach (i.e., CEEMDAN-CNN-GRU) incorporated with MODIS satellite-derived data, ground-based SILO data, and climate mode indices (representing synoptic-scale climate features) can be a good modelling tool to predict soil moisture or other hydrological variables at multi-step lead times, including its future use in water resource management and sustainable agriculture.

6. Conclusions

This study reports the performance efficacy of a DL data-driven (CEEMDAN-CNN-GRU) model based on the Gated Recurrent Unit (GRU) for daily surface soil moisture forecasting at multi-step horizons. The hybrid CEEMDAN-CNN-GRU model was built by integrating MODIS sensors (satellite-derived data), ground-based observations, and climate indices tested at important stations in the Australian Murray Darling Basin. To attain an accurate and reliable model for soil moisture, a feature extraction (i.e., CNN) and feature (or variable) selection algorithm (i.e., NCA) was used, with tests at 1st, 5th, 7th, 14th,

21st, and 30th day ahead period. The input variables, comprised initially of 52 different predictors, were extracted from March 2003 to March 2020 and screened accordingly, using the NCA algorithm through a feature selection stage, to select the most relevant input variables required to forecast daily-scale soil moisture. Three other benchmarking models (i.e., CEEMDAN-GRU, CNN-GRU, and GRU) were built and evaluated against statistical score metrics and visual analysis to ascertain the predictive skill of the objective model of observed and forecasted datasets in the testing phase. The results revealed that NCA was a practical approach to acquire the best features from an optimal set of predictor variables. The hybrid CEEMDAN-CNN-GRU model has significantly improved the decomposition of input variables to provide more defined soil moisture prediction features. Thus, the proposed CEEMDAN-CNN-GRU model yielded an acceptable level of accuracy when applied at the 1st, 5th, and 7th day ahead SSM forecasting against standalone GRU model registering a comparatively higher forecast error at all these periods. This superior performance was also endorsed with low MAE values, ranging from 0.013 kg m^{-2} to 0.067 kg m^{-2} , 0.030 kg m^{-2} to 0.075 kg m^{-2} , and 0.057 kg m^{-2} to 0.113 kg m^{-2} for the 1st, 5th, and 7th day ahead period. Other results also supported the practical utility of the CEEMDAN-CNN-GRU model. For example, the probability plot of absolute error for Menindee station has 95% of SSM forecasting with the lowest error bracket (<0.1) at the 1st, and 5th day SSM prediction, and these results were better than earlier studies on forecasting soil moisture prediction, e.g., [23,36,59,102]. As the present study has focused on daily scale prediction, in a future study, researchers may also adopt the CEEMDAN-CNN-GRU model to utilise the global climate model (GCM) model-simulated variables to estimate future SSM under global warming scenarios.

Supplementary Materials: The following are available online at <https://www.mdpi.com/2072-4292/13/4/554/s1>, Table S1: Performance of CNN-GRU and GRU model to forecast the 1st day Surface Soil moisture of Minendee Station with the optimum forecasting results based on the Nash-Sutcliffe coefficient (NS) and mean absolute error (MAE; Kg m^{-2}) for the testing phase, Table S2: Caption identical to Table S1, except for the 5th day. Table S3: Caption identical to Table S1, except for the 7th day, Table S4: Caption identical to Table S1, except for the 14th day, Table S5: Caption identical to Table S1, except for the 21st day, Table S6: Caption identical to Table S1, except for the 30th day, Figure S1: Probability plot for the objective model (i.e., CEEMDAN-CNN-GRU), benchmark model (i.e., CNN-GRU, CEEMDAN-GRU), and the standalone model (i.e., GRU) for (a) Deniliquin, (b) Fairfield and (c) Gabo Island stations at different nth ($n = 1, 5, 7, 14, 21$, and 30) days. The reference points are provided at 95 percentiles on the probability of comparing the model.

Author Contributions: Conceptualisation, A.A.M.A. and R.C.D.; methodology, A.A.M.A. and R.C.D.; software, A.A.M.A.; model development, A.A.M.A.; validation, A.A.M.A.; formal analysis, A.A.M.A.; investigation, A.A.M.A.; resources, A.A.M.A.; data curation, A.A.M.A.; writing—original draft preparation, A.A.M.A.; writing—review and editing, A.A.M.A., R.C.D., N.R., A.G., Q.F., Z.Y. and L.Y.; visualisation, A.A.M.A.; supervision, R.C.D.; funding acquisition, R.C.D. All authors have read and agreed to the published version of the manuscript.

Funding: The study was supported by the Chinese Academy of Science (CAS), and University of Southern Queensland (USQ) under the USQ-CAS Postgraduate Research Scholarship (2019–2021) awarded to the first author, managed by Graduate Research School under the leadership of Professor Feng Qi (CAS) and Associate Professor Ravinesh Deo (USQ).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analysed in this study. The data can be found here: [<https://giovanni.gsfc.nasa.gov/giovanni/>; <https://www.longpaddock.qld.gov.au/silo/>].

Acknowledgments: We thank the Editor and Reviewers for their insightful comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Van Loon, A.F.; Laaha, G. Hydrological drought severity explained by climate and catchment characteristics. *J. Hydrol.* **2015**, *526*, 3–14. [\[CrossRef\]](#)
2. Brocca, L.; Melone, F.; Moramarco, T.; Morbidelli, R. Spatial—temporal variability of soil moisture and its estimation across scales. *Water Resour. Res.* **2010**, *46*, W02516. [\[CrossRef\]](#)
3. Brocca, L.; Ciabatta, L.; Massari, C.; Camici, S.; Tarpanelli, A. Soil Moisture for Hydrological Applications: Open Questions and New Opportunities. *Water* **2017**, *9*, 140. [\[CrossRef\]](#)
4. Chang, X.; Zhao, W.; Zeng, F. Crop evapotranspiration-based irrigation management during the growing season in the arid region of northwestern China. *Environ. Monit. Assess.* **2015**, *187*, 699. [\[CrossRef\]](#)
5. Gill, M.K.; Asefa, T.; Kemblowski, M.W.; McKee, M. Soil moisture prediction using support vector machines 1. *JAWRA J. Am. Water Resour. Assoc.* **2006**, *42*, 1033–1046. [\[CrossRef\]](#)
6. Akbari Asanjan, A.; Yang, T.; Hsu, K.; Sorooshian, S.; Lin, J.; Peng, Q. Short-Term Precipitation Forecast Based on the PERSIANN System and LSTM Recurrent Neural Networks. *J. Geophys. Res. Atmos.* **2018**, *123*. [\[CrossRef\]](#)
7. Tripathi, S.; Srinivas, V.V.; Nanjundiah, R.S. Downscaling of precipitation for climate change scenarios: A support vector machine approach. *J. Hydrol.* **2006**, *330*, 621–640. [\[CrossRef\]](#)
8. Yang, L.; Feng, Q.; Yin, Z.; Wen, X.; Deo, R.C.; Si, J.; Li, C. Application of multivariate recursive nesting bias correction, multiscale wavelet entropy and AI-based models to improve future precipitation projection in upstream of the Heihe River, Northwest China. *Theor. Appl. Climatol.* **2018**, *137*, 323–339. [\[CrossRef\]](#)
9. Nguyen-Huy, T.; Deo, R.C.; An-Vo, D.-A.; Mushtaq, S.; Khan, S. Copula-statistical precipitation forecasting model in Australia's agro-ecological zones. *Agric. Water Manag.* **2017**, *191*, 153–172. [\[CrossRef\]](#)
10. Deo, R.C.; Şahin, M. Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia. *Atmos. Res.* **2015**, *153*, 512–525. [\[CrossRef\]](#)
11. Prasad, R.; Deo, R.C.; Li, Y.; Maraseni, T. Input selection and performance optimisation of ANN-based streamflow forecasts in the drought-prone Murray Darling Basin region using IIS and MODWT algorithm. *Atmos. Res.* **2017**, *197*, 42–63. [\[CrossRef\]](#)
12. Ahmed, A.M.; Shah, S.M.A. Application of artificial neural networks to predict peak flow of Surma River in Sylhet Zone of Bangladesh. *Int. J. Water* **2017**, *11*, 363–375. [\[CrossRef\]](#)
13. Hu, C.; Wu, Q.; Li, H.; Jian, S.; Li, N.; Lou, Z. Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation. *Water* **2018**, *10*, 1543. [\[CrossRef\]](#)
14. Kratzert, F.; Klotz, D.; Brenner, C.; Schulz, K.; Herrnegger, M. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 6005–6022. [\[CrossRef\]](#)
15. Arto, I.; Garcia-Muros, X.; Cazcarro, I.; Gonzalez-Eguino, M.; Markandya, A.; Hazra, S. The socioeconomic future of deltas in a changing environment. *Sci. Total Environ.* **2019**, *648*, 1284–1296. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Le, H.; Lee, J. Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. *Water* **2019**, *11*, 1387. [\[CrossRef\]](#)
17. Ahmed, A.M.; Deo, R.C.; Ghahramani, A.; Raj, N.; Feng, Q.; Yin, Z.; Yang, L. LSTM integrated with Boruta-random forest optimiser for soil moisture estimation under RCP4. 5 and RCP8. 5 global warming scenarios. *Stoch. Environ. Res. Risk Assess.* **2021**, 1–31. [\[CrossRef\]](#)
18. Gedefaw, M.; Hao, W.; Denghua, Y.; Girma, A. Variable selection methods for water demand forecasting in Ethiopia: Case study Gondar town. *Cogent Environ. Sci.* **2018**, *4*, 1537067. [\[CrossRef\]](#)
19. Mouatadid, S.; Adamowski, J. Using extreme learning machines for short-term urban water demand forecasting. *Urban Water J.* **2017**, *14*, 630–638. [\[CrossRef\]](#)
20. Ahmed, A.A.M. Prediction of dissolved oxygen in Surma River by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (ANNs). *J. King Saud Univ. Eng. Sci.* **2017**, *29*, 151–158. [\[CrossRef\]](#)
21. Ahmed, A.A.M.; Shah, S.M.A. Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical oxygen demand (BOD) of Surma River. *J. King Saud Univ. Eng. Sci.* **2017**, *29*, 237–243. [\[CrossRef\]](#)
22. Huang, C.; Li, L.; Ren, S.; Zhou, Z. Research of soil moisture content forecast model based on genetic algorithm BP neural network. In Proceedings of the International Conference on Computer and Computing Technologies in Agriculture, Nanchang, China, 22–25 October 2010; pp. 309–316.
23. Prasad, R.; Deo, R.C.; Li, Y.; Maraseni, T. Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. *Geoderma* **2018**, *330*, 136–161. [\[CrossRef\]](#)
24. Ghimire, S.; Deo, R.C.; Raj, N.; Mi, J. Deep Learning Neural Networks Trained with MODIS Satellite-Derived Predictors for Long-Term Global Solar Radiation Prediction. *Energies* **2019**, *12*, 2407. [\[CrossRef\]](#)
25. Zhang, J.; Zhu, Y.; Zhang, X.; Ye, M.; Yang, J. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. *J. Hydrol.* **2018**, *561*, 918–929. [\[CrossRef\]](#)
26. Le, X.-H.; Ho, H.V.; Lee, G. Application of gated recurrent unit (GRU) network for forecasting river water levels affected by tides. In Proceedings of the International Conference on Asian and Pacific Coasts, Hanoi, Vietnam, 25–28 September 2019; pp. 673–680.
27. Gao, S.; Huang, Y.; Zhang, S.; Han, J.; Wang, G.; Zhang, M.; Lin, Q. Short-term runoff prediction with GRU and LSTM networks without requiring time step optimisation during sample generation. *J. Hydrol.* **2020**, *589*, 125188. [\[CrossRef\]](#)

28. Ghimire, S.; Deo, R.C.; Raj, N.; Mi, J. Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms. *Appl. Energy* **2019**, *253*, 113541. [\[CrossRef\]](#)
29. Deo, R.C.; Sahin, M. An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland. *Environ. Monit. Assess.* **2016**, *188*, 90. [\[CrossRef\]](#) [\[PubMed\]](#)
30. Nourani, V.; Komasi, M.; Mano, A. A multivariate ANN-wavelet approach for rainfall-runoff modeling. *Water. Resour. Manag.* **2009**, *23*, 2877–2894. [\[CrossRef\]](#)
31. Nourani, V.; Baghanam, A.H.; Adamowski, J.; Kisi, O. Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review. *J. Hydrol.* **2014**, *514*, 358–377. [\[CrossRef\]](#)
32. Deo, R.C.; Wen, X.; Qi, F. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. *Appl. Energy* **2016**, *168*, 568–593. [\[CrossRef\]](#)
33. Cornish, C.R.; Bretherton, C.S.; Percival, D.B. Maximal overlap wavelet statistical analysis with application to atmospheric turbulence. *Bound.-Layer Meteorol.* **2006**, *119*, 339–374. [\[CrossRef\]](#)
34. Rathinasamy, M.; Khosa, R.; Adamowski, J.; Ch, S.; Partheepan, G.; Anand, J.; Narsimlu, B. Wavelet-based multiscale performance analysis: An approach to assess and improve hydrological models. *Water Resour. Res.* **2014**, *50*, 9721–9737. [\[CrossRef\]](#)
35. Di, C.; Yang, X.; Wang, X. A four-stage hybrid model for hydrological time series forecasting. *PLoS ONE* **2014**, *9*, e104663. [\[CrossRef\]](#) [\[PubMed\]](#)
36. Prasad, R.; Deo, R.C.; Li, Y.; Maraseni, T. Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridiser algorithm approach. *Catena* **2019**, *177*, 149–166. [\[CrossRef\]](#)
37. Seo, Y.; Kim, S. Hydrological Forecasting Using Hybrid Data-Driven Approach. *Am. J. Appl. Sci.* **2016**, *13*, 891–899. [\[CrossRef\]](#)
38. Beltrán-Castro, J.; Valencia-Aguirre, J.; Orozco-Alzate, M.; Castellanos-Domínguez, G.; Travieso-González, C.M. Rainfall forecasting based on ensemble empirical mode decomposition and neural networks. In Proceedings of the International Work-Conference on Artificial Neural Networks, Tenerife, Spain, 12–14 June 2013; pp. 471–480.
39. Jiao, G.; Guo, T.; Ding, Y. A new hybrid forecasting approach applied to hydrological data: A case study on precipitation in Northwestern China. *Water* **2016**, *8*, 367. [\[CrossRef\]](#)
40. Ouyang, Q.; Lu, W.; Xin, X.; Zhang, Y.; Cheng, W.; Yu, T. Monthly rainfall forecasting using EEMD-SVR based on phase-space reconstruction. *Water Resour. Manag.* **2016**, *30*, 2311–2325. [\[CrossRef\]](#)
41. Ali, M.; Deo, R.C.; Maraseni, T.; Downs, N.J. 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.* **2019**, *576*, 164–184. [\[CrossRef\]](#)
42. Adarsh, S.; Sanah, S.; Murshida, K.; Nooramol, P. Scale dependent prediction of reference evapotranspiration based on Multi-Variate Empirical mode decomposition. *Ain Shams Eng. J.* **2018**, *9*, 1839–1848. [\[CrossRef\]](#)
43. Hu, W.; Si, B.C. Soil water prediction based on its scale-specific control using multivariate empirical mode decomposition. *Geoderma* **2013**, *193*, 180–188. [\[CrossRef\]](#)
44. Schepen, A.; Wang, Q.J.; Robertson, D. Evidence for Using Lagged Climate Indices to Forecast Australian Seasonal Rainfall. *J. Clim.* **2012**, *25*, 1230–1246. [\[CrossRef\]](#)
45. Yuan, C.; Yamagata, T. Impacts of IOD, ENSO and ENSO Modoki on the Australian winter wheat yields in recent decades. *Sci. Rep.* **2015**, *5*, 1–8. [\[CrossRef\]](#)
46. Risbey, J.S.; Pook, M.J.; McIntosh, P.C.; Wheeler, M.C.; Hendon, H.H. On the remote drivers of rainfall variability in Australia. *Mon. Weather Rev.* **2009**, *137*, 3233–3253. [\[CrossRef\]](#)
47. Royce, F.S.; Fraisse, C.W.; Baigorria, G.A. ENSO classification indices and summer crop yields in the Southeastern USA. *Agric. For. Meteorol.* **2011**, *151*, 817–826. [\[CrossRef\]](#)
48. Shuai, J.; Zhang, Z.; Sun, D.-Z.; Tao, F.; Shi, P. ENSO, climate variability and crop yields in China. *Clim. Res.* **2013**, *58*, 133–148. [\[CrossRef\]](#)
49. Rashid, M.M.; Sharma, A.; Johnson, F. Multi-model drought predictions using temporally aggregated climate indicators. *J. Hydrol.* **2020**, *581*. [\[CrossRef\]](#)
50. Nikolopoulos, E.I.; Anagnostou, E.N.; Borga, M. Using high-resolution satellite rainfall products to simulate a major flash flood event in northern Italy. *J. Hydrometeorol.* **2013**, *14*, 171–185. [\[CrossRef\]](#)
51. Nikolopoulos, E.I.; Anagnostou, E.N.; Hossain, F.; Gebremichael, M.; Borga, M. Understanding the scale relationships of uncertainty propagation of satellite rainfall through a distributed hydrologic model. *J. Hydrometeorol.* **2010**, *11*, 520–532. [\[CrossRef\]](#)
52. Yong, B.; Hong, Y.; Ren, L.L.; Gourley, J.J.; Huffman, G.J.; Chen, X.; Wang, W.; Khan, S.I. Assessment of evolving TRMM-based multisatellite real-time precipitation estimation methods and their impacts on hydrologic prediction in a high latitude basin. *J. Geophys. Res. Atmos.* **2012**, *117*. [\[CrossRef\]](#)
53. Ghimire, S.; Deo, R.C.; Downs, N.J.; Raj, N. 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 Sens. Environ.* **2018**, *212*, 176–198. [\[CrossRef\]](#)
54. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Oehmcke, S.; Zielinski, O.; Kramer, O. Input quality aware convolutional LSTM networks for virtual marine sensors. *Neurocomputing* **2018**, *275*, 2603–2615. [\[CrossRef\]](#)

56. Nunez, J.C.; Cabido, R.; Pantrigo, J.J.; Montemayor, A.S.; Velez, J.F. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognit.* **2018**, *76*, 80–94. [\[CrossRef\]](#)
57. Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv* **2014**, arXiv:1406.1078.
58. Torres, M.E.; Colominas, M.A.; Schlotthauer, G.; Flandrin, P. A complete ensemble empirical mode decomposition with adaptive noise. In Proceedings of the 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 4144–4147.
59. Prasad, R.; Deo, R.C.; Li, Y.; Maraseni, T. Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors. *Soil Tillage Res.* **2018**, *181*, 63–81. [\[CrossRef\]](#)
60. Wen, X.; Feng, Q.; Deo, R.C.; Wu, M.; Yin, Z.; Yang, L.; Singh, V.P. Two-phase extreme learning machines integrated with the complete ensemble empirical mode decomposition with adaptive noise algorithm for multi-scale runoff prediction problems. *J. Hydrol.* **2019**, *570*, 167–184. [\[CrossRef\]](#)
61. Wu, Z.; Huang, N.E. Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Adv. Adapt. Data Anal.* **2009**, *1*, 1–41. [\[CrossRef\]](#)
62. Wu, Z.; Huang, N.E.; Chen, X. The multi-dimensional ensemble empirical mode decomposition method. *Adv. Adapt. Data Anal.* **2009**, *1*, 339–372. [\[CrossRef\]](#)
63. Bowden, G.J.; Dandy, G.C.; Maier, H.R. Input determination for neural network models in water resources applications. Part 1—background and methodology. *J. Hydrol.* **2005**, *301*, 75–92. [\[CrossRef\]](#)
64. Maier, H.R.; Jain, A.; Dandy, G.C.; Sudheer, K.P. Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environ. Model. Softw.* **2010**, *25*, 891–909. [\[CrossRef\]](#)
65. Yang, W.; Wang, K.; Zuo, W. Neighborhood Component Feature Selection for High-Dimensional Data. *JCP* **2012**, *7*, 161–168. [\[CrossRef\]](#)
66. Murray–Darling Basin Authority. *Guide to the Proposed Basin Plan*; Murray–Darling Basin Auth.: Canberra, Australia, 2010.
67. Australian Bureau of Statistics. *Household Use of Information Technology*; Australia Bureau of Statistics: Canberra, Australia, 2010.
68. Hijmans, R.J.; Cameron, S.E.; Parra, J.L.; Jones, P.G.; Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol. J. R. Meteorol. Soc.* **2005**, *25*, 1965–1978. [\[CrossRef\]](#)
69. ASRIS. *The Australian Soil Resource Information System*; Department of Agriculture, Fisheries and Forestry: Canberra, Australia, 2014. Available online: <https://www.asris.csiro.au/> (accessed on 12 December 2020).
70. BOM. Bureau of Meteorology. 2020. Available online: <http://www.bom.gov.au/> (accessed on 31 December 2020).
71. Deo, R.C.; Sahin, M. Forecasting long-term global solar radiation with an ANN algorithm coupled with satellite-derived (MODIS) land surface temperature (LST) for regional locations in Queensland. *Renew. Sustain. Energy Rev.* **2017**, *72*, 828–848. [\[CrossRef\]](#)
72. Deo, R.C.; Şahin, M.; Adamowski, J.F.; Mi, J. Universally deployable extreme learning machines integrated with remotely sensed MODIS satellite predictors over Australia to forecast global solar radiation: A new approach. *Renew. Sustain. Energy Rev.* **2019**, *104*, 235–261. [\[CrossRef\]](#)
73. Deo, R.C.; Syktus, J.I.; McAlpine, C.A.; Lawrence, P.J.; McGowan, H.A.; Phinn, S.R. Impact of historical land cover change on daily indices of climate extremes including droughts in eastern Australia. *Geophys. Res. Lett.* **2009**, *36*. [\[CrossRef\]](#)
74. Nguyen-Huy, T.; Deo, R.C.; Mushtaq, S.; An-Vo, D.-A.; Khan, S. Modeling the joint influence of multiple synoptic-scale, climate mode indices on Australian wheat yield using a vine copula-based approach. *Eur. J. Agron.* **2018**, *98*, 65–81. [\[CrossRef\]](#)
75. Berrick, S.W.; Leptoukh, G.; Farley, J.D.; Rui, H. Giovanni: A web service workflow-based data visualization and analysis system. *IEEE Trans. Geosci. Remote Sens.* **2008**, *47*, 106–113. [\[CrossRef\]](#)
76. Chen, C.; Jiang, H.; Zhang, Y.; Wang, Y. Investigating spatial and temporal characteristics of harmful Algal Bloom areas in the East China Sea using a fast and flexible method. In Proceedings of the 2010 18th International Conference on Geoinformatics, Beijing, China, 18–20 June 2010; pp. 1–4.
77. Morshed, A.; Aryal, J.; Dutta, R. Environmental spatio-temporal ontology for the Linked open data cloud. In Proceedings of the 2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, Melbourne, VIC, Australia, 16–18 July 2013; pp. 1907–1912.
78. Trouet, V.; Van Oldenborgh, G.J. KNMI Climate Explorer: A web-based research tool for high-resolution paleoclimatology. *Tree-Ring Res.* **2013**, *69*, 3–13. [\[CrossRef\]](#)
79. Adnan, M.; Rehman, N.; Sheikh, M.; Khan, A.; Mir, K.; Khan, M. Influence of natural forcing phenomena on precipitation of Pakistan. *Pak. J. Meteorol.* **2016**, *12*, 23–35.
80. Philander, S.G.H. El Nino southern oscillation phenomena. *Nature* **1983**, *302*, 295–301. [\[CrossRef\]](#)
81. Chiew, F.H.; Piechota, T.C.; Dracup, J.A.; McMahon, T.A. El Nino/Southern Oscillation and Australian rainfall, streamflow and drought: Links and potential for forecasting. *J. Hydrol.* **1998**, *204*, 138–149. [\[CrossRef\]](#)
82. Madden, R.A.; Julian, P.R. Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *J. Atmos. Sci.* **1971**, *28*, 702–708. [\[CrossRef\]](#)
83. Henley, B.J.; Gergis, J.; Karoly, D.J.; Power, S.; Kennedy, J.; Folland, C.K. A tripole index for the interdecadal Pacific oscillation. *Clim. Dyn.* **2015**, *45*, 3077–3090. [\[CrossRef\]](#)
84. Troup, A. The ‘southern oscillation’. *Q. J. R. Meteorol. Soc.* **1965**, *91*, 490–506. [\[CrossRef\]](#)
85. Ketkar, N. Introduction to keras. In *Deep Learning with Python*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 97–111.

86. Brownlee, J. *Deep Learning with Python: Develop Deep Learning Models on Theano and TensorFlow Using Keras*; Machine Learning Mastery: Vermont, VIC, Australia, 2016.
87. 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. In Proceedings of the 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
88. Jayalakshmi, T.; Santhakumaran, A. Statistical normalization and back propagation for classification. *Int. J. Comput. Theory Eng.* **2011**, *3*, 1793–8201.
89. Maier, H.R.; Dandy, G.C. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. *Environ. Model. Softw.* **2000**, *15*, 101–124. [[CrossRef](#)]
90. Deo, R.C.; Downs, N.; Parisi, A.V.; Adamowski, J.F.; Quilty, J.M. Very short-term reactive forecasting of the solar ultraviolet index using an extreme learning machine integrated with the solar zenith angle. *Environ. Res.* **2017**, *155*, 141–166. [[CrossRef](#)]
91. Arhami, M.; Kamali, N.; Rajabi, M.M. Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations. *Environ. Sci. Pollut. Res.* **2013**, *20*, 4777–4789. [[CrossRef](#)]
92. Jekabsons, G. ARESLab: Adaptive Regression Splines Toolbox for Matlab/Octave. 2011. Available online: <http://www.cs.rtu.lv/jekabsons> (accessed on 18 January 2021).
93. Nash, J.E.; Sutcliffe, J.V. River flow forecasting through conceptual models part I—A discussion of principles. *J. Hydrol.* **1970**, *10*, 282–290. [[CrossRef](#)]
94. Gupta, H.V.; Kling, H.; Yilmaz, K.K.; Martinez, G.F. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.* **2009**, *377*, 80–91. [[CrossRef](#)]
95. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)?—Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* **2014**, *7*, 1247–1250. [[CrossRef](#)]
96. Willmott, C.J.; Robeson, S.M.; Matsuura, K. A refined index of model performance. *Int. J. Climatol.* **2012**, *32*, 2088–2094. [[CrossRef](#)]
97. Legates, D.R.; McCabe, G.J. A refined index of model performance: A rejoinder. *Int. J. Climatol.* **2013**, *33*, 1053–1056. [[CrossRef](#)]
98. Yin, Z.; Feng, Q.; Wen, X.; Deo, R.C.; Yang, L.; Si, J.; He, Z. Design and evaluation of SVR, MARS and M5Tree models for 1, 2 and 3-day lead time forecasting of river flow data in a semiarid mountainous catchment. *Stoch. Environ. Res. Risk Assess.* **2018**, *32*, 2457–2476. [[CrossRef](#)]
99. Legates, D.R.; McCabe Jr, G.J. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **1999**, *35*, 233–241. [[CrossRef](#)]
100. Friedman, J.H. *Estimating Functions of Mixed Ordinal and Categorical Variables Using Adaptive Splines*; Stanford University CA Lab for Computational Statistics: Stanford, CA, USA, 1991.
101. Ghimire, S.; Deo, R.C.; Downs, N.J.; Raj, N. Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia. *J. Clean. Prod.* **2019**, *216*, 288–310. [[CrossRef](#)]
102. Cai, Y.; Zheng, W.; Zhang, X.; Zhangzhong, L.; Xue, X. Research on soil moisture prediction model based on deep learning. *PLoS ONE* **2019**, *14*, e0214508. [[CrossRef](#)]

CHAPTER 7: MULTI-STEP AHEAD STREAMFLOW WATER LEVEL FORECASTING USING DOUBLE DECOMPOSITION AND DEEP LEARNING METHODS

7.1 Foreword

This Chapter is an exact copy of the published manuscript [under Review] to *Science of the Total Environment*, (Scopus Impact Factor 7.96). The title of the manuscript is:

“New double decomposition deep learning methods for river water level forecasting.”

This chapter discusses a newly developed double decomposed (CEEMDAN and VMD) deep learning hybrid model, CVMD-CBiLSTM comprising CNN and BiLSTM model coupled with ant colony optimization to forecast the streamflow water level at the multi-step horizon. At 19 gauging stations on the Murray River, Australia, the streamflow water levels were forecasted using satellite-derived data, climate mode indices, and ground-based meteorological data. For SWL forecasting, the CVMD-CBiLSTM hybrid model performed better than the standalone model. Outperforming the benchmark models, almost 98% of the prediction errors were less than 0.020 meters and had a low relative error (RRMSE of 0.08%). The study concluded that integrating deep learning algorithms with ACO feature selection can improve water resource management decisions.

7.2 Research Highlights

- Deep learning CVMD-CBiLSTM model is proposed for streamflow water level forecasts
- Satellite predictors are incorporated with climate indices and ground-based data
- Two phases of feature decomposition (CEEMDAN, VMD) integrated with CNN and BiLSTM
- CVMD-CBiLSTM with ACO feature selection has a distinct advantage in forecasting.
- Our advanced AI model can empower strategic water management decisions.

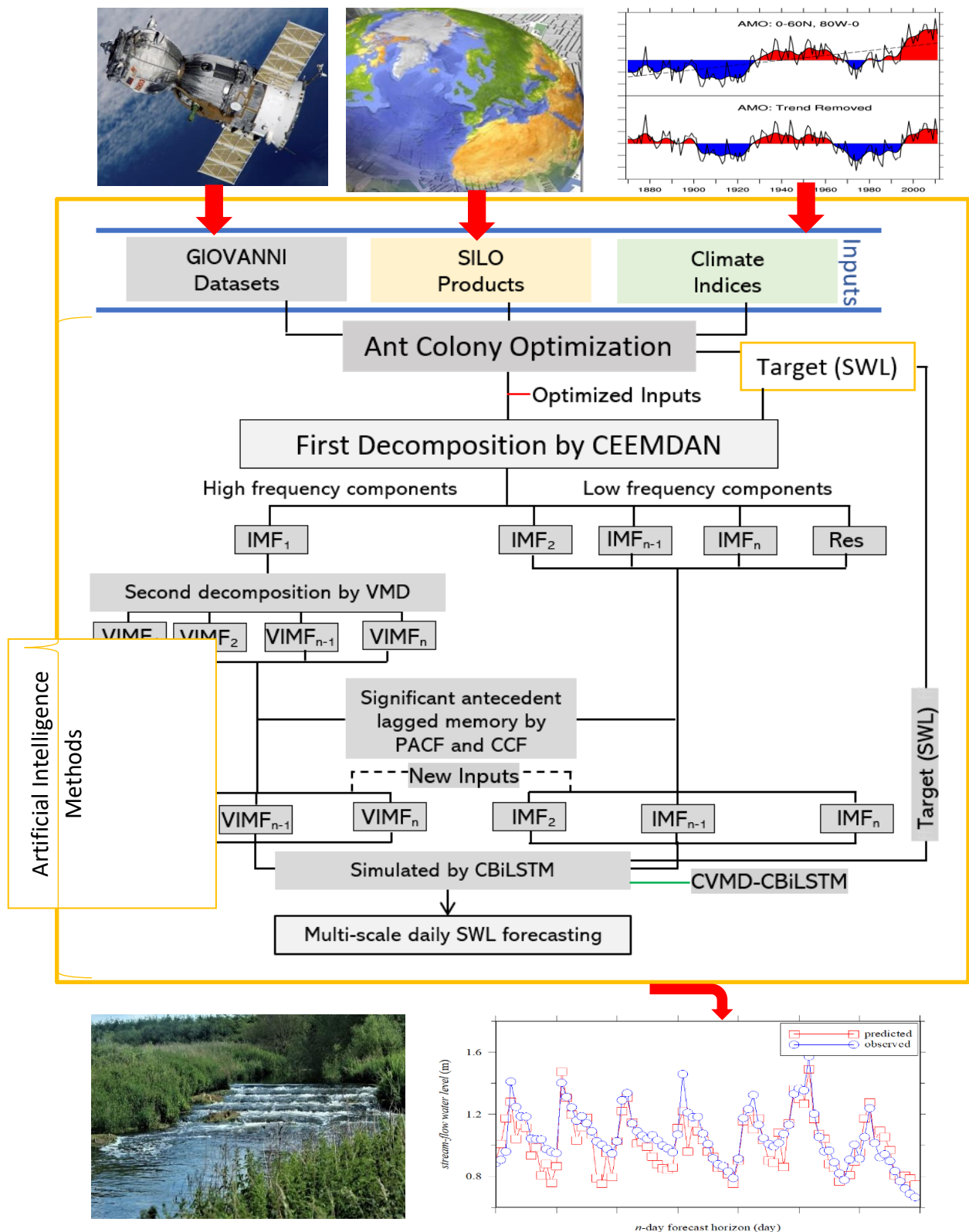


Fig. 7.1 Graphical abstract of Objective 5

7.3 Article 5

This article cannot be displayed due to copyright restrictions. See the article link in the Related Outputs field on the item record for possible access.

CHAPTER 8: CONCLUSIONS AND FUTURE SCOPE

8.1 Synthesis and Conclusions

This study developed advanced deep learning hybrid forecasting models in Australia's Murray-Darling Basin. These are accurate and reliable data intelligence models for soil moisture, evapotranspiration, and streamflow water level forecasting. Several standalone, hybrid and high precision models were also presented at multi-forecasting horizons, including the short-term (daily) and medium-term (monthly) periods. The streamflow water level and soil moisture were forecasted monthly and daily time periods. The reference evapotranspiration was forecasted at multi-step ahead horizon (week-1 to week-4 ahead forecasting) of daily data. In subsequent studies, the complex relationships between hydrological variables associated with extreme weather and climatic events are implemented. Therefore, the findings of this study can provide valuable information and powerful tools to hydrologists and climate scientists. The significant contributions of the study are summarised in this section. Finally, potentially interesting paths for future research are suggested at the end of this chapter.

The study addressed three distinct issues: i) the problem of selecting appropriate predictor variables from sets of multivariate inputs in hydrological forecasting; ii) the overfitting of large datasets due to the complex and non-linear relationships between predictors, and iii) addressing non-stationarity and non-linearity issue of the hydrological predictor variables. The issue of selecting appropriate predictor variables was resolved by incorporating different feature selection algorithms, namely ant colony optimization, Boruta-random forest optimiser, and neighbourhood component analysis. In contrast, the latter was resolved by deep learning predictive models. In contrast, multi-resolution analysis, namely complete ensemble empirical mode decomposition with adaptive noise and variational mode decomposition, was used to address the non-stationarity and non-linearity issue.

The findings showed improved performances of hybridised models against the standalone counterparts. The outcomes from the **Objective 1** served as the milestone of the future state of any river system through accurate predictive modelling considering the climate indicators. The results have demonstrated that the Boruta-

Random forest feature selection algorithm is an important tool that can successfully identify the significant features within the predictor variable(s), as required to model the hydrological state of a river system. **Objective 2** provided relevant information from a set of predictors, including climate mode indices, ground-based meteorological inputs, and satellite-derived data, to develop an early warning decision support system for reference crop evapotranspiration. The ant colony optimisation successfully selected appropriate inputs from three distinct datasets, and a convolutional neural network extracted the feature of the target in forecasting reference evapotranspiration. The ET_o estimates are beneficial for future studies to characterize the water availability of terrestrial ecosystems, assess climate change impacts, and provide guidance to agriculture in the Murray-Darling Basin.

The study from **Objective 3** demonstrates the capability of a hybridised long short-term memory predictive framework to emulate soil moisture under global warming scenarios. The proposed model is developed by integrating the Boruta-Random Forest feature selection. Significant antecedent memory of SM behaviour was successfully applied to estimate the future SM using a coupled model intercomparison phase-5 (CMIP5) repository. This study demonstrates the capability of the proposed algorithm to simulate future soil moisture under climate change, which can be implemented in hydrology, agriculture, soil use management, and environmental management. **Objective 4** provides a deep learning hybrid model to forecast the surface soil moisture at a multi-step ahead horizon. The incorporation of neighbourhood component analysis with the model revealed that a shorter period forecast (i.e., 1st, 5th, and 7th) was more precise, whereas a longer forecast horizon (i.e., 14th, 21st, and 30th) registered a lower accuracy than that of the shorter span. Lastly, **Objective 5** used a new double decomposition approach for forecasting stream-flow water levels using deep learning algorithms with remotely sensed MODIS satellites, climate mode indices, and ground-based atmospheric products. The results are critical to managing water quality and the availability of palatable water for many practical applications.

Constructing parsimonious yet high-performing data-intelligent models requires feature selection. This study found the successful application of Boruta-random forest, ant colony optimisation, and neighbourhood component analysis feature selection to identify the potential predictor variables for forecasting the hydrological

variables. Moreover, the CEEMDAN is a self-adaptive multi-resolution method; hence, the number of IMFs and residual components (*i.e.*, resolved frequencies) is contingent upon the embedded features within the data. The CEEMDAN improved the model performance of the standalone models. The CEEMDAN based proposed model outperformed all other models in forecasting soil moisture. In addition, a multi-phase multi-resolution technique coupled with CEEMDAN and variational mode decomposition (VMD), referred to as CVMD, substantially improved streamflow water level forecasting. The CEEMDAN-decomposed high-frequency component is further decomposed into several components by VMD. The hybrid model requires minimal human interaction, which is a fundamental benefit of the self-adaptive MRA tool, CEEMDAN, coupled with CNN-LSTM. This might be integrated into advanced forecasting software for mobile devices like tablets and phones and provide hydrological forecasts at the farm level.

In summary, various novel contributions were provided by this study in the development of data-intelligent predictive models for hydrologic forecasting. According to the results, the study found that the performances of the suggested models were relatively better than standalone models. Hence, new innovative approaches were explored, and the main contributions of the research could be summarized as follows:

- The initial contribution was to investigate previously unexplored forecasting methodologies in hydrologic forecasting in the Murray-Darling Basin using deep learning hybrid models.
- The current research enhances this general goal by combining a deep learning predictive method (*i.e.*, LSTM and GRU) with the three feature selection algorithms for increased performance accuracy. According to the findings, the feature selection methods used in this study are promising tools for identifying relevant characteristics within the predictor variable(s), which is required for modelling the hydrological phenomena.
- The developed hybrid two-phase model utilised the CEEMDAN method to address the data's non-stationarity and the NCA algorithm to choose the optimal parameters for the CNN-GRU model.

- Another contribution was a new double decomposition approach using CEEMDAN and VMD methods using deep learning algorithms.
- The extreme values are represented by high peaks of a dominant wave moving along the field. The data pre-processing approaches such as feature selection (i.e., BRF, ACO and NCA), feature extraction (i.e., CNN) and feature decomposition (i.e., CEEMDAN, VMD and EEMD) successfully applied in gripping the extreme events of different hydrological phenomenon.
- Moreover, our study, which has not been previously investigated, was the incorporation of predictor variables from three distinct datasets: climate mode indices, satellite-derived variables, and ground-based hydro-climatological variables. The inclusion of ground-based measurements, climate indicators, and atmospheric satellite data increased the diversity of the input signals, improving forecasting skills.
- The study improved the hydrological forecast of the Murray-Darling Basin using antecedent memory of long and short-term climate indices, satellite-derived variables, and ground-based hydro-climatological variables.
- Finally, this study's novel and robust data-driven deep learning framework have developed a novel approach to assessing climate risk on agricultural production and water resource management that can be applied to other sectors.

8.2 Limitations and Recommendations for Future Research

The explored innovative approaches showed promising outcomes and could provide the scientific pathway for integrated on-farm decision-support systems for hydrological and precision agricultural purposes.

- In the streamflow water level forecasting study, the critical limitation was the unavailability of concurrently recorded streamflow water level and hydro-meteorological data at the same hydrological stations. In future studies, the use of concurrently observed data is recommended to improve the accuracy of the respective models.
- Individual forecasts of high, moderate, and low streamflow events and soil moisture level events could also be explored independently.

- Studies with maximum overlap discrete wavelet transform (MODWT) and multivariate ensemble mode decomposition (MEMD) could also provide greater insight into the performance of these predictive models.
- Integration of add-on optimizer algorithms (e.g., firefly optimizer algorithm (FFA) or quantum-behaved particle swarm optimization (Q-PSO)) could also be applied in these hydrological models.
- Since the standard statistical approaches tend to avoid the hurdle of model uncertainty that potentially leads to over-confident inferences and risky agricultural decisions, alternative feature selection algorithms like iterative input selection (IIS), modified minimum redundancy maximum relevance (mMRMR) algorithm, or joint mutual information maximisation feature selection (JMIM) can be further explored.
- Uncertainty is crucial to machine learning, although it is one of the factors that causes the most difficulties for adopting the machine learning problems. It is increasingly important to evaluate the reliability and efficacy of artificial intelligence (AI) systems before they could be applied in practise, since the predictions made by such models are subject to noise and model inference errors. It is therefore critically desirable for any AI-based system to express uncertainty in a reliable manner. The sources of uncertainty begin when the test and training data are mismatched, whereas data uncertainty arises from class overlap or noise in the data; therefore, estimating uncertainty is recommended for the future study.
- The Spatio-temporal dependency of the stations and variables could be addressed in future studies. The contribution of neighbouring stations can provide important information in extracting the feature of the target variables.

Finally, using new hybridised deep learning techniques, this study has made novel contributions to the practical problem of hydrological forecasting. The hybridised machine learning data-intelligent forecasting models are simple to implement with high computational efficiency and low latency. This approach can serve as an essential tool for water resource management applications for urban and agricultural management.

REFERENCES

Note that the references presented here do not include the references from the published articles (Chapters 3,4, 5, and 6) and the submitted manuscript (Chapter 7). These references are provided in the reference sections of the respective articles.

Abera, W, Tamene, L, Abegaz, A & Solomon, D 2019, 'Understanding climate and land surface changes impact on water resources using Budyko framework and remote sensing data in Ethiopia', *Journal of arid environments*, vol. 167, pp. 56-64.

Adamowski, J, Fung, C, Prasher, S, Ozga-Zielinski, B & Sliusarieva, A 2012, 'Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada', *Water Resources Research*, vol. 48, no. 1.

Ahmed, AM, Deo, RC, Feng, Q, Ghahramani, A, Raj, N, Yin, Z & Yang, L 2021, 'Deep learning hybrid model with Boruta-Random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity', *Journal of Hydrology*, vol. 599, p. 126350.

Ahmed, M, Deo, R, Feng, Q, Ghahramani, A, Raj, N, Yin, Z & Yang, L 2021, 'Hybrid deep learning method for a week-ahead evapotranspiration forecasting', *Stochastic Environmental Research and Risk Assessment*, pp. 1-19.

AIDR 2021, 'Flood - Murray River, South Australia | Australian Disaster Resilience Knowledge Hub'.

Akbari, E, Buntat, Z, Enzevae, A, Ebrahimi, M, Yazdavar, AH & Yusof, R 2014, 'Analytical modeling and simulation of I–V characteristics in carbon nanotube based gas sensors using ANN and SVR methods', *Chemometrics and Intelligent Laboratory Systems*, vol. 137, pp. 173-80.

Alaoui, A, Willmann, E, Jasper, K, Felder, G, Herger, F, Magnusson, J & Weingartner, R 2014, 'Modelling the effects of land use and climate changes on hydrology in the Ursern Valley, Switzerland', *Hydrological Processes*, vol. 28, no. 10, pp. 3602-14.

Ali, M, Deo, RC, Downs, NJ & Maraseni, T 2018, 'An ensemble-ANFIS based uncertainty assessment model for forecasting multi-scalar standardized precipitation index', *Atmospheric Research*, vol. 207, pp. 155-80.

Ali, M, Deo, RC, Maraseni, T & Downs, NJ 2019, 'Improving SPI-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridized with simulated annealing and kernel ridge regression algorithms', *Journal of Hydrology*, vol. 576, pp. 164-84.

Austin, J, Zhang, L, Jones, RN, Durack, P, Dawes, W & Hairsine, P 2010, 'Climate change impact on water and salt balances: an assessment of the impact of climate change on catchment salt and water balances in the Murray-Darling Basin, Australia', *Climatic Change*, vol. 100, no. 3, pp. 607-31.

Australian Bureau of Statistics 2010, *Household use of information technology*, Australia.

Bafitlhile, TM & Li, Z 2019, 'Applicability of ϵ -Support Vector Machine and Artificial Neural Network for Flood Forecasting in Humid, Semi-Humid and Semi-Arid Basins in China', *Water*, vol. 11, no. 1.

Barron, O, Silberstein, R, Ali, R, Donohue, R, McFarlane, D, Davies, P, Hodgson, G, Smart, N & Donn, M 2012, 'Climate change effects on water-dependent ecosystems in south-western Australia', *Journal of Hydrology*, vol. 434, pp. 95-109.

Beare, S & Heaney, A 2002, 'Climate change and water resources in the Murray Darling Basin, Australia', *Conference paper*, Citeseer.

Beesley, CA, Frost, AJ & Zajackowski, J 2009, 'A comparison of the BAWAP and SILO spatially interpolated daily rainfall datasets', *18th World IMACS / MODSIM Congress*, Cairns, Australia.

Berrick, SW, Leptoukh, G, Farley, JD & Rui, H 2008, 'Giovanni: a web service workflow-based data visualization and analysis system', *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 1, pp. 106-13.

Blanchard, J, Martin, C & Liu, W 2018, 'Effect of ELMS and disruptions on FNSF plasma-facing components', *Fusion Engineering and Design*, vol. 135, pp. 337-45.
BOM 2020, 'Bureau of Meteorology'.

Bowden, GJ, Dandy, GC & Maier, HR 2005, 'Input determination for neural network models in water resources applications. Part 1—background and methodology', *Journal of Hydrology*, vol. 301, no. 1-4, pp. 75-92.

Cai, W & Cowan, T 2009, 'La Niña Modoki impacts Australia autumn rainfall variability', *Geophysical Research Letters*, vol. 36, no. 12.

Cai, W, Van Rensch, P, Cowan, T & Hendon, HH 2011, 'Teleconnection pathways of ENSO and the IOD and the mechanisms for impacts on Australian rainfall', *Journal of Climate*, vol. 24, no. 15, pp. 3910-23.

Cai, Y, Zheng, W, Zhang, X, Zhangzhong, L & Xue, X 2019, 'Research on soil moisture prediction model based on deep learning', *PLoS One*, vol. 14, no. 4, p. e0214508.

Chartres, C & Williams, J 2006, 'Can Australia overcome its water scarcity problems?', *Journal of Developments in Sustainable Agriculture*, vol. 1, no. 1, pp. 17-24.

Chen, C, Jiang, H, Zhang, Y & Wang, Y 2010, 'Investigating spatial and temporal characteristics of harmful Algal Bloom areas in the East China Sea using a fast and flexible method', *2010 18th International Conference on Geoinformatics*, IEEE, pp.

1-4.

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*, vol. 165, pp. 681-95.

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*.

CSIRO 2016, *Climate change in Australia information for Australia's natural resource management regions: Technical report*, CSIRO and Bureau of Meteorology Australia.

Deo, RC & Şahin, M 2015, 'Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia', *Atmospheric Research*, vol. 153, pp. 512-25.

Deo, RC & Sahin, M 2016, 'An extreme learning machine model for the simulation of monthly mean streamflow water level in eastern Queensland', *Environ Monit Assess*, vol. 188, no. 2, p. 90.

Deo, RC, Wen, X & Qi, F 2016, 'A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset', *Applied Energy*, vol. 168, pp. 568-93.

Deo, RC, Byun, H-R, Adamowski, JF & Kim, D-W 2015, 'A real-time flood monitoring index based on daily effective precipitation and its application to Brisbane and Lockyer Valley flood events', *Water Resources Management*, vol. 29, no. 11, pp. 4075-93.

Deo, RC, Syktus, JI, McAlpine, CA, Lawrence, PJ, McGowan, HA & Phinn, SR 2009, 'Impact of historical land cover change on daily indices of climate extremes

including droughts in eastern Australia', *Geophysical Research Letters*, vol. 36, no. 8.

Dhiman, HS, Deb, D & Guerrero, JM 2019, 'Hybrid machine intelligent SVR variants for wind forecasting and ramp events', *Renewable and Sustainable Energy Reviews*, vol. 108, pp. 369-79.

Draper, CS, Reichle, RH & Koster, RD 2018, 'Assessment of MERRA-2 land surface energy flux estimates', *Journal of Climate*, vol. 31, no. 2, pp. 671-91.

Ehteram, M, Afan, HA, Dianatikhah, M, Ahmed, AN, Ming Fai, C, Hossain, MS, Allawi, MF & Elshafie, A 2019, 'Assessing the Predictability of an Improved ANFIS Model for Monthly Streamflow Using Lagged Climate Indices as Predictors', *Water*, vol. 11, no. 6.

Feng, Y, Cui, N, Gong, D, Zhang, Q & Zhao, L 2017, 'Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling', *Agricultural Water Management*, vol. 193, pp. 163-73.

Gelaro, R, McCarty, W, Suárez, MJ, Todling, R, Molod, A, Takacs, L, Randles, CA, Darmenov, A, Bosilovich, MG & Reichle, R 2017, 'The modern-era retrospective analysis for research and applications, version 2 (MERRA-2)', *Journal of Climate*, vol. 30, no. 14, pp. 5419-54.

Ghimire, S, Deo, RC, Downs, NJ & Raj, N 2019a, 'Global solar radiation prediction by ANN integrated with European Centre for medium range weather forecast fields in solar rich cities of Queensland Australia', *Journal of Cleaner Production*, vol. 216, pp. 288-310.

Ghimire, S, Deo, RC, Downs, NJ & Raj, N 2019b, 'Deep Learning Neural Networks Trained with MODIS Satellite-Derived Predictors for Long-Term Global Solar Radiation Prediction', *Energies*, vol. 12, no. 12.

Gong, G, An, X, Mahato, NK, Sun, S, Chen, S & Wen, Y 2019, 'Research on Short-Term Load Prediction Based on Seq2seq Model', *Energies*, vol. 12, no. 16.

Guo, Y, Fang, G, Xu, YP, Tian, X & Xie, J 2020, 'Identifying how future climate and land use/cover changes impact streamflow in Xinanjiang Basin, East China', *Sci Total Environ*, vol. 710, p. 136275.

Jeffrey, SJ, Carter, JO, Moodie, KB & Beswick, AR 2001, 'Using spatial interpolation to construct a comprehensive archive of Australian climate data', *Environmental Modelling & Software*, vol. 16, no. 4, pp. 309-30.

Kim, H-Y & Liang, S 2010, 'Development of a hybrid method for estimating land surface shortwave net radiation from MODIS data', *Remote Sensing of Environment*, vol. 114, no. 11, pp. 2393-402.

Kirono, DG, Chiew, FH & Kent, DM 2010, 'Identification of best predictors for forecasting seasonal rainfall and runoff in Australia', *Hydrological Processes: An International Journal*, vol. 24, no. 10, pp. 1237-47.

Kisi, O & Parmar, KS 2016, 'Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution', *Journal of Hydrology*, vol. 534, pp. 104-12.

Kong, W, Jia, Y, Dong, ZY, Meng, K & Chai, S 2020, 'Hybrid approaches based on deep whole-sky-image learning to photovoltaic generation forecasting', *Applied Energy*, vol. 280, p. 115875.

Kulshrestha, A, Krishnaswamy, V & Sharma, M 2020, 'Bayesian BILSTM approach for tourism demand forecasting', *Annals of tourism research*, vol. 83, p. 102925.

Leblanc, M, Tweed, S, Van Dijk, A & Timbal, B 2012, 'A review of historic and future hydrological changes in the Murray-Darling Basin', *Global and Planetary Change*, vol. 80, pp. 226-46.

Li, F, Ma, G, Chen, S & Huang, W 2021, 'An Ensemble Modeling Approach to Forecast Daily Reservoir Inflow Using Bidirectional Long-and Short-Term Memory (Bi-LSTM), Variational Mode Decomposition (VMD), and Energy Entropy Method', *Water Resources Management*, vol. 35, no. 9, pp. 2941-63.

Li, H, Hou, G, Dakui, F, Xiao, B, Song, L & Liu, Y 2007, 'Prediction and elucidation of the population dynamics of *Microcystis* spp. in Lake Dianchi (China) by means of artificial neural networks', *Ecological Informatics*, vol. 2, no. 2, pp. 184-92.

Li, J, Johnson, F, Evans, J & Sharma, A 2017, 'A comparison of methods to estimate future sub-daily design rainfall', *Advances in Water Resources*, vol. 110, pp. 215-27.

López, G & Batlles, F 2014, *Estimating solar radiation from MODIS data*, *Energy Proced.*, 49, 2362–2369.

Maier, HR, Jain, A, Dandy, GC & Sudheer, KP 2010, 'Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions', *Environmental Modelling & Software*, vol. 25, no. 8, pp. 891-909.

Marsland, S, Bi, D, Uotila, P, Fiedler, R, Griffies, S, Lorbacher, K, O'Farrell, S, Sullivan, A, Uhe, P & Zhou, X 2013, 'Evaluation of ACCESS climate model ocean diagnostics in CMIP5 simulations', *Aust. Meteorol. Oceanogr. J.*, vol. 63, pp. 101-19.

Martin, G, Bellouin, N, Collins, W, Culverwell, I, Halloran, P, Hardiman, S, Hinton, T, Jones, C, McDonald, R & McLaren, A 2011, *The HadGEM2 family of Met Office Unified Model climate configurations*, *Geosci. Model Dev.*, 4, 723–757, doi: 10.5194, gmd-4-723-2011.

McAlpine, C, Syktus, J, Ryan, J, Deo, R, McKeon, G, McGowan, H & Phinn, S 2009, 'A continent under stress: interactions, feedbacks and risks associated with impact of modified land cover on Australia's climate', *Global Change Biology*, vol. 15, no. 9, pp. 2206-23.

McKenzie, NJ, Jacquier, DW & Gregory, LJ 2005, 'The Australian Soil Resource Information System', *Technical specifications. Version*, vol. 1, p. 93.

McMahon, TA, Finlayson, B, Haines, A & Srikanthan, R 1992, *Global runoff: continental comparisons of annual flows and peak discharges*, Catena Verlag.

Meher, JK, Das, L, Akhter, J, Benestad, RE & Mezghani, A 2017, 'Performance of CMIP3 and CMIP5 GCMs to Simulate Observed Rainfall Characteristics over the Western Himalayan Region', *Journal of Climate*, vol. 30, no. 19, pp. 7777-99.

Murphy, BF & Timbal, B 2008, 'A review of recent climate variability and climate change in southeastern Australia', *International Journal of Climatology*, vol. 28, no. 7, pp. 859-79.

Nguyen-Huy, T, Deo, RC, An-Vo, D-A, Mushtaq, S & Khan, S 2017, 'Copula-statistical precipitation forecasting model in Australia's agro-ecological zones', *Agricultural Water Management*, vol. 191, pp. 153-72.

Olah, C 2015, 'Understanding lstm networks, 2015', URL <http://colah.github.io/posts/2015-08-Understanding-LSTMs>.

Ouyang, Q, Lu, W, Xin, X, Zhang, Y, Cheng, W & Yu, T 2016, 'Monthly rainfall forecasting using EEMD-SVR based on phase-space reconstruction', *Water Resources Management*, vol. 30, no. 7, pp. 2311-25.

Peng, T, Zhou, J, Zhang, C & Zheng, Y 2017, 'Multi-step ahead wind speed forecasting using a hybrid model based on two-stage decomposition technique and AdaBoost-extreme learning machine', *Energy Conversion and Management*, vol. 153, pp. 589-602.

Peng, T, Zhang, C, Zhou, J & Nazir, MS 2021, 'An integrated framework of Bi-directional Long-Short Term Memory (BiLSTM) based on sine cosine algorithm for

hourly solar radiation forecasting', *Energy*, vol. 221, p. 119887.

Prasad, R, Deo, RC, Li, Y & Maraseni, T 2018a, 'Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition', *Geoderma*, vol. 330, pp. 136-61.

Prasad, R, Deo, RC, Li, Y & Maraseni, T 2018b, 'Ensemble committee-based data intelligent approach for generating soil moisture forecasts with multivariate hydro-meteorological predictors', *Soil and Tillage Research*, vol. 181, pp. 63-81.

Prasad, R, Deo, RC, Li, Y & Maraseni, T 2019, 'Weekly soil moisture forecasting with multivariate sequential, ensemble empirical mode decomposition and Boruta-random forest hybridizer algorithm approach', *Catena*, vol. 177, pp. 149-66.

Prasad, R, Ali, M, Xiang, Y & Khan, H 2020, 'A double decomposition-based modelling approach to forecast weekly solar radiation', *Renewable Energy*, vol. 152, pp. 9-22.

Raj, N & Brown, J 2021, 'An EEMD-BiLSTM Algorithm Integrated with Boruta Random Forest Optimiser for Significant Wave Height Forecasting along Coastal Areas of Queensland, Australia', *Remote Sensing*, vol. 13, no. 8, p. 1456.

Rajagukguk, RA, Ramadhan, RA & Lee, H-J 2020, 'A Review on Deep Learning Models for Forecasting Time Series Data of Solar Irradiance and Photovoltaic Power', *Energies*, vol. 13, no. 24, p. 6623.

Ramesh, KV & Goswami, P 2014, 'Assessing reliability of regional climate projections: the case of Indian monsoon', *Scientific Reports*, vol. 4, p. 4071.

Risbey, JS, Pook, MJ, McIntosh, PC, Wheeler, MC & Hendon, HH 2009, 'On the remote drivers of rainfall variability in Australia', *Monthly Weather Review*, vol. 137, no. 10, pp. 3233-53.

Sainath, TN, Vinyals, O, Senior, A & Sak, H 2015, 'Convolutional, long short-term memory, fully connected deep neural networks', *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, pp. 4580-4.

Salcedo-Sanz, S, Deo, RC, Cornejo-Bueno, L, Camacho-Gómez, C & Ghimire, S 2018, 'An efficient neuro-evolutionary hybrid modelling mechanism for the estimation of daily global solar radiation in the Sunshine State of Australia', *Applied Energy*, vol. 209, pp. 79-94.

Schepen, A, Wang, QJ & Robertson, D 2012, 'Evidence for Using Lagged Climate Indices to Forecast Australian Seasonal Rainfall', *Journal of Climate*, vol. 25, no. 4, pp. 1230-46.

Seo, Y & Kim, S 2016, 'Hydrological Forecasting Using Hybrid Data-Driven Approach', *American Journal of Applied Sciences*, vol. 13, no. 8, pp. 891-9.

Sibtain, M, Li, X, Nabi, G, Azam, MI & Bashir, H 2020, 'Development of a three-stage hybrid model by utilizing a two-stage signal decomposition methodology and machine learning approach to predict monthly runoff at Swat river basin, Pakistan', *Discrete Dynamics in Nature and Society*, vol. 2020.

Sillmann, J, Kharin, VV, Zhang, X, Zwiers, FW & Bronaugh, D 2013, 'Climate extremes indices in the CMIP5 multimodel ensemble: Part 1. Model evaluation in the present climate', *Journal of Geophysical Research: Atmospheres*, vol. 118, no. 4, pp. 1716-33.

Song, G & Dai, Q 2017, 'A novel double deep ELMs ensemble system for time series forecasting', *Knowledge-Based Systems*, vol. 134, pp. 31-49.

Sun, Q, Miao, C & Duan, Q 2015, 'Comparative analysis of CMIP3 and CMIP5 global climate models for simulating the daily mean, maximum, and minimum temperatures and daily precipitation over China', *Journal of Geophysical Research:*

Atmospheres, vol. 120, no. 10, pp. 4806-24.

Timbal, B, Abbs, D, Bhend, J, Chiew, F, Church, J, Ekström, M, Kirono, D, Lenton, A, Lucas, C & McInnes, K 2015, 'Murray basin cluster report', *Climate Change in Australia Projections for Australia's Natural Resource Management Regions: Cluster Reports*. CSIRO and Bureau of Meteorology, Australia, Australia.

Trouet, V & Van Oldenborgh, GJ 2013, 'KNMI Climate Explorer: a web-based research tool for high-resolution paleoclimatology', *Tree-Ring Research*, vol. 69, no. 1, pp. 3-13.

Ummenhofer, CC, England, MH, McIntosh, PC, Meyers, GA, Pook, MJ, Risbey, JS, Gupta, AS & Taschetto, AS 2009, 'What causes southeast Australia's worst droughts?', *Geophysical Research Letters*, vol. 36, no. 4.

van Dijk, AIJM, Beck, HE, Crosbie, RS, de Jeu, RAM, Liu, YY, Podger, GM, Timbal, B & Viney, NR 2013, 'The Millennium Drought in southeast Australia (2001-2009): Natural and human causes and implications for water resources, ecosystems, economy, and society', *Water Resources Research*, vol. 49, no. 2, pp. 1040-57.

Van Loon, A & Laaha, G 2015, 'Hydrological drought severity explained by climate and catchment characteristics', *Journal of Hydrology*, vol. 526, pp. 3-14.

Verdon, DC & Franks, SW 2005, 'Indian Ocean sea surface temperature variability and winter rainfall: Eastern Australia', *Water Resources Research*, vol. 41, no. 9.

Wang, Y, Wang, W, Wang, S, Chang, J & Bai, D 2019, 'A model for soil moisture dynamics estimation based on artificial neural network', *E3S Web of Conferences*, vol. 81.

Whetton, P & Chiew, F 2021, 'Climate change in the Murray–Darling Basin', in *Murray-Darling Basin, Australia*, Elsevier, pp. 253-74.

Xu, Y, Xu, C, Gao, X & Luo, Y 2009, 'Projected changes in temperature and precipitation extremes over the Yangtze River Basin of China in the 21st century', *Quaternary International*, vol. 208, no. 1-2, pp. 44-52.

Yang, W, Wang, K & Zuo, W 2012, 'Neighborhood Component Feature Selection for High-Dimensional Data', *JCP*, vol. 7, no. 1, pp. 161-8.

Yaseen, ZM, Sulaiman, SO, Deo, RC & Chau, K-W 2019, 'An enhanced extreme learning machine model for river flow forecasting: State-of-the-art, practical applications in water resource engineering area and future research direction', *Journal of Hydrology*, vol. 569, pp. 387-408.

Yaseen, ZM, Jaafar, O, Deo, RC, Kisi, O, Adamowski, J, Quilty, J & El-Shafie, A 2016, 'Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq', *Journal of Hydrology*, vol. 542, pp. 603-14.

Yuan, AY, Gao, Y, Peng, L, Zhou, L, Liu, J, Zhu, S & Song, W 2020, 'Hybrid deep learning network for vascular segmentation in photoacoustic imaging', *Biomedical Optics Express*, vol. 11, no. 11, pp. 6445-57.

Yuan, C & Yamagata, T 2015, 'Impacts of IOD, ENSO and ENSO Modoki on the Australian winter wheat yields in recent decades', *Scientific Reports*, vol. 5, no. 1, pp. 1-8.

Zajaczkowski, J, Wong, K & Carter, J 2013, 'Improved historical solar radiation gridded data for Australia', *Environmental Modelling & Software*, vol. 49, pp. 64-77.

Zaman, B & McKee, M 2014, 'Spatio-Temporal Prediction of Root Zone Soil Moisture Using Multivariate Relevance Vector Machines', *Open Journal of Modern Hydrology*, vol. 04, no. 03, pp. 80-90.

Zeng, Z, Wu, W, Zhou, Y, Li, Z, Hou, M & Huang, H 2019, 'Changes in Reference Evapotranspiration over Southwest China during 1960–2018: Attributions and

Implications for Drought', *Atmosphere*, vol. 10, no. 11.

Zhang, J, Zhu, Y, Zhang, X, Ye, M & Yang, J 2018, 'Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas', *Journal of Hydrology*, vol. 561, pp. 918-29.

Zhang, W, Qu, Z, Zhang, K, Mao, W, Ma, Y & Fan, X 2017, 'A combined model based on CEEMDAN and modified flower pollination algorithm for wind speed forecasting', *Energy Conversion and Management*, vol. 136, pp. 439-51.

Zhang, Y, Yamaguchi, R, Imoto, S & Miyano, S 2017, 'Sequence-specific bias correction for RNA-seq data using recurrent neural networks', *BMC Genomics*, vol. 18, no. Suppl 1, p. 1044.

APPENDIX A: BIAS CORRECTION OF TOTAL CLOUD COVER FORECAST FROM GLOBAL FORECAST SYSTEM MODEL

A1.1 Foreword

This Chapter is an exact copy of the submitted manuscript to the *Applied Energy* Manuscript Number: APEN-S-21-14660 (Scopus Impact Factor 9.75). The title of the manuscript is:

Development of Kernel Ridge Regression for bias correction of Total Cloud Cover forecast generated by Global Forecast System weather model

In this Chapter, a new kernel ridge regression (KRR) approach is used to reduce bias in total cloud cover (TCDC) for inter-daily scales (i.e., 2–8 days ahead) of the GFS forecast datasets. The KRR model is compared against a multivariate recursive nesting bias correction (MRNBC) and classical machine learning (ML) approaches to determine its performance. A significant reduction in mean bias error (20–50%) relative to MRNBC and reference value during the model's testing phase indicates the objective model's efficacy. Because of this, it is concluded that the proposed KRR method should be investigated further to reduce the uncertainties in weather simulations, which has positive contributory implications and practicality in solar energy generation systems and energy conversion and monitoring systems.

A1.2 Research Highlights

- Kernel ridge regression (KRR) is constructed for bias correction of total cloud cover
- KRR is evaluated using Global Forecast System 0 UTC, 3 UTC & 6 UTC simulation data
- We see a reduction in the bias of GFS cloud cover for two-to-six-day forecasts
- KRR is validated against multivariate recursive nested bias correction and ML model
- KRR is a viable tool to correct cloud cover bias in solar energy monitoring systems

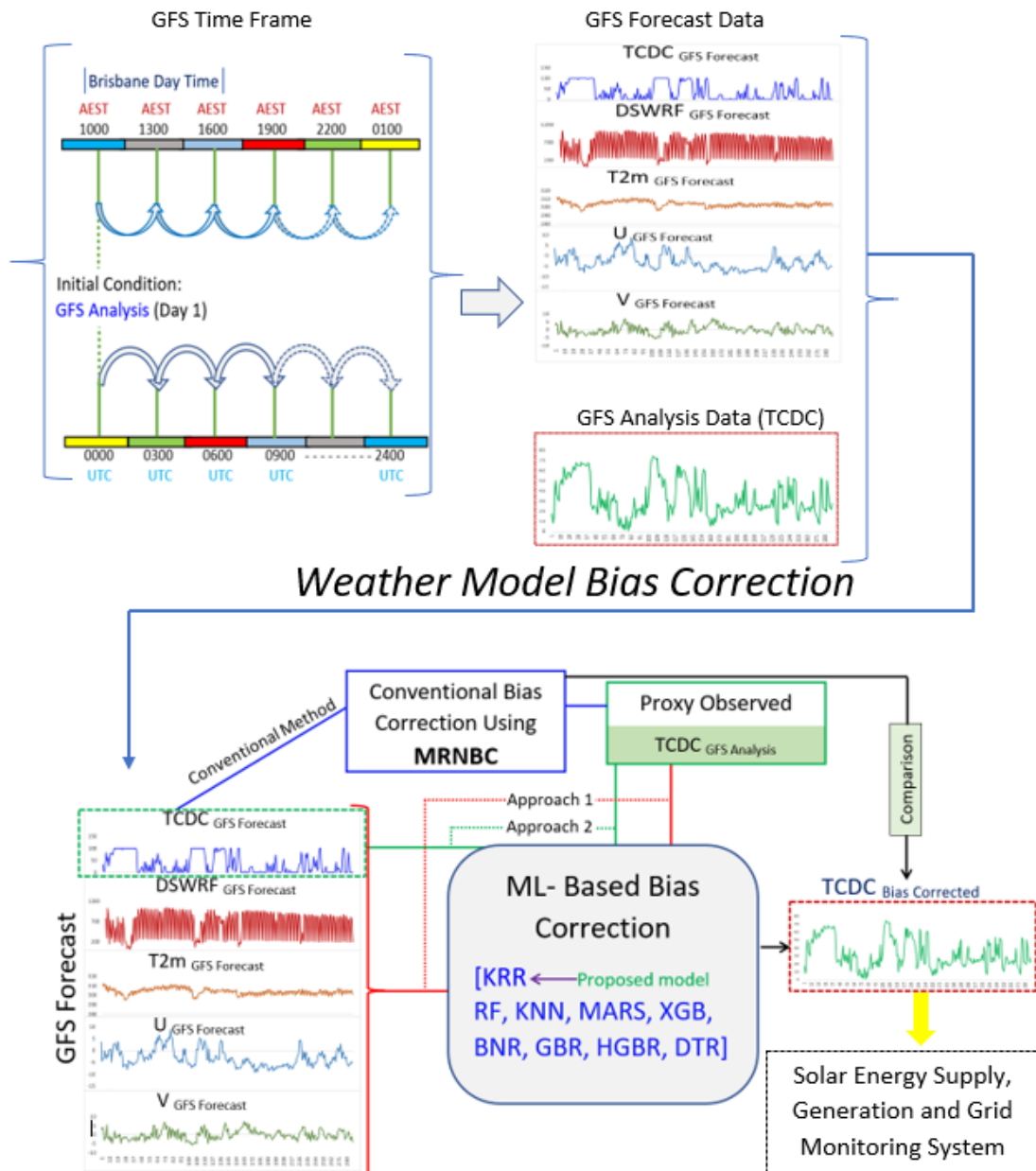


Fig A1 Graphical Abstract of Article 6

A1.3 Article 6

Development of Kernel Ridge Regression for bias correction of Total Cloud Cover forecast generated by Global Forecast System weather model

A. A. Masrur Ahmed^{1*}, Ravinesh C Deo^{1*}, Gary Segal², and Yanshan Yu²

¹ School of Sciences, University of Southern Queensland, Springfield, QLD 4300, Australia

² CS Energy, Level 2, HQ North Tower, 540 Wickham St Fortitude Valley QLD 4006, Australia

Corresponding Author: ravinesh.deo@usq.edu.au (Prof Ravinesh Deo)

Abstract

Accurate forecasting of total cloud cover (TCDC) in Global Forecast System (GFS) model with 3-h time horizon can monitor photovoltaic power production in solar energy generation systems. The bias in a GFS-based variable can hinder the monitoring and integration of energy into real electricity grids. This study proposes a new kernel ridge regression (KRR) strategy to reduce bias in TCDC for inter-daily (*i.e.*, 2–8 Day ahead) scales. The KRR model is evaluated against a multivariate recursive nesting bias correction (MRNBC) and competing machine learning (ML) methods. The testing phase showed that in terms of mean absolute error, the KRR model trained with cloud cover inputs (*i.e.*, TCDC_{GFS-Forecast}) has outperformed MRNBC and all other ML models for Day 2-7 forecast (MAE \approx 20.20–27.47%). The appraisal of objective model's effectiveness is ascertained by a notable reduction in mean bias error (20–50%) against MRNBC and reference accuracy values generated using the proxy-observed and the non-corrected GFS-forecasted TCDC data in model's testing phase. The study, therefore, ascertains that the proposed KRR method could be explored to reduce the uncertainties in weather simulation that have positive contributory implication and practicality in solar energy generation, energy conversion, and monitoring systems.

Keywords solar energy generation; bias correction in weather models; global forecast system, cloud cover study; solar radiation; energy conversion and management

1.0 Introduction

It is a relatively challenging yet an essential task to address biases in forecasted data generated by numerical weather prediction and physical models [1, 2]. These models have practical applications in solar energy integration and power monitoring in household rooftop systems and solar energy farms. The fidelity of any physical model or their predictive uncertainties can be associated with many factors, *e.g.*, unrealistic estimates of greenhouse gases within the physical model, the model equation's uncertainty, and incorrectly parametrized model's internal climate variability effects [3]. Solar photovoltaic (PV) power forecasting, which uses weather simulations, is necessary to ensure supply of solar energy, economic viability, and stability of electricity grid. While it is attractive from an

environmental and economic viewpoint to harness this freely available solar energy, uncertainties from weather captured in solar simulation models can cause significant problems, particularly due to their strong dependence on uncontrolled weather elements such as the cloud cover, the positioning of the sun relative to a solar PV panel, bushfire dust, aerosols, ozone, and other conditions, which can change dramatically over an hourly and inter-hourly interval [4]. Therefore, greater understanding any weather model's uncertainty, capability to generate accurate cloud cover patterns, and the use of a novel learning method to correct the bias in forecasted cloud cover is a crucial tool for management of solar energy farms, loop virtual power plant (VPP) as a new, technology-cent energy grid, rooftop solar energy production, and energy uptake by the consumers attained through automated intelligent solar monitoring systems.

Total cloud cover, denoted in this paper as TCDC, is considered to be a chief cause of the intermittency in solar energy supply given that the performance of any solar photovoltaic (PV) system is likely to drop by as much as 60% within a few seconds of the passing clouds over a solar PV panel [5]. When the sun travels across the sky and is obscured by a cloud cover band, the intensity of solar radiation reaching the solar PV panel may also fluctuate, which may cause a significant drop in the solar power production. Consequently, this can reduce the quantity of solar power produced by solar energy companies. A cloudy day can impact the solar PV output much differently as the clouds affecting solar energy production quite diversely affects solar power from on a rooftop [5]. Therefore, accurate cloud forecasts, both at short-term (*i.e.*, sub-hourly, hourly, or inter-hourly) and medium term (*i.e.*, daily, or inter-daily) period has particular industry implications for solar energy monitoring with broader application in agriculture, natural disaster (e.g., cyclone) prediction, air quality and environmental monitoring tasks.

To support growing renewable energy industries in decisions regarding the sustainability of solar supply and integration of power into national grids, reliable forecasts of cloud cover are crucial for studying the variations in ground-based solar and the underlying intermittencies in the supply of energy [6, 7]. Typically, cloud cover is defined as a “fraction of the sky covered by all the visible clouds” [8], so unlike weather variables, *e.g.*, temperature or precipitation, TCDC observation data are somewhat different in terms of their characteristics [6]. For example, the movement of clouds over a solar PV panel can be relatively stochastic (*i.e.*, rapidly changing, unpredictable, intermittent). These uncertain features of clouds no doubt hamper energy production rate, so it is highly desirable to create a better understanding of features present in total clouds that affect the overall solar energy generation system.

This research aims to build machine learning approaches for bias correction of TCDC data derived from a Global Forecast System (GFS). Maintained by the National Centre for Environment Prediction, the GFS physics-based weather model has a 0.25° grid resolution and a 3h temporal

resolution with forecasts initialized six times per day for cloud cover, including the 2-meter temperature, zonal and meridional windspeed, downward shortwave radiation flux, and other atmospheric variables. These variables are widely employed in solar PV system prediction modules such as the *pvl* package that is used to monitor solar energy generation by industries. The bias correction of these variables has traditionally focused on correcting individual variable representations across a single time (*e.g.*, daily, monthly). However, these corrections aim to determine the bias in a statistical or a quantile sense and utilise the corrected data for future scenarios of solar energy production. Daily and monthly standardization can address systematic biases in mean and variance of simulated variables [9] to handle energy generation applications. Bias correction with non-parametric approaches, *e.g.*, quantile matching [10-13] and equidistant quantile [14], were found to be successful methods. Still, a shortcoming of these techniques is that they tend to examine only the bias in distribution of GFS and GCM simulation and not an effect of their persistence [15]. The study of Johnson and Sharma [16] suggested a nested bias correction (NBC) to reduce variability and persistence at different time scales. In contrast, techniques like multivariate bias correction (MBC) [17], copula-based bias correction [18], empirical copula bias correction (EC-BC) [19], distribution transfer [13], power transformation [20-22] and local intensity scaling [22, 23] have also been utilized for numerous locations and a plethora of weather variables. To the best of the authors' knowledge, no method has successfully eliminated the bias given that the relationships between simulated and observed variables are complex [24]. Machine learning (ML) approaches have thus been demonstrated as alternative methods to model highly non-linear features [25-27], appearing to offer a strong potential to correct bias in numerical weather variables such as those produced by GFS or another weather model.

In general, ML discovers the associations between predictors and a predictand without considering the system's operation [28-30]. The mathematical complexity of a physical model can also be reduced using ML algorithms, or the data features be better understood with a physical model employing partial differential equations with initial conditions [31, 32]. Such conditions are somewhat difficult to estimate over spatial and temporal domains. Artificial neural networks (ANN) are a multivariate non-parametric ML algorithm used to correct inter-instrument biases [33, 34] but as a "black box" model, ANN can only identify potential causal relationships [35]. However, support vector machines (SVM) have long been recognized as a sophisticated model with a sound theoretical foundation in statistical learning [36-38]. The use of SVMs has explored a kernel-based ANN to address the drawbacks of conventional ANNs [39]. SVM can be very resilient and efficient for non-linear modelling of noisy mixed data [25, 27, 40].

The kernel ridge regression (KRR) [41] advocated in this study is based on the integration of the kernels and the ridge regression approach to capture non-linear correlative features and to address with regression-based over-fitting issues found in some of other ML models [42]. To its primary advantage,

the proposed KRR method can also use a regularized variant of least-squares to learn a global function, and therefore, predict any target variable. Although ML had been used for general bias correction, the proposed KRR technique has remained somewhat underexplored. It should also be noted that the proposed KRR model in its generic sense has been used in many research including the forecasting of precipitation [45], drought [46], wind speed [47-51] and solar power [52]. It has also been demonstrated that the KRR method offers significant benefits in terms of computational simplicity relative to an SVM model, thus limiting its practical applicability in real-time solar energy monitoring systems.

This study, therefore, aims to build a new KRR-based bias correction method for GFS derived total cloud cover (TCDC) for Columboola Solar Farm in Queensland, Australia where solar energy projects are strategic to the cleaner energy utilisation efforts of the Queensland Government. To fulfil this zero-carbon vision envisaged by United Nations Sustainable Development Goal # 7 (*i.e.*, cleaner & affordable energy), we have adopted two distinct modelling strategies: *First*, the KRR model is trained using the 2-m temperature, 10-m zonal (*U*)-wind, 10-m meridional (*V*)-wind, downward shortwave radiation flux, and the total cloud cover regressed against the proxy-observed (*i.e.*, GFS-Analysis) dataset. *Second*, only the cloud cover data series (*i.e.*, TCDC_{GFS-Forecast}) are used as a single input with TCDC_{GFS-Analysis} as a target to test this alternative method's viability. The proposed KRR model is then evaluated extensively using conventional methods (*i.e.*, MRNBC) and nine other ML methods. The proposed KRR and its counterpart models are tested over inter-daily horizons utilizing Day-2 to Day-8 forecasts. Based on the potential computational capabilities, it is envisaged that the biases of GFS-derived TCDC at multi-step horizons can be reduced, and that the new predictive system may be useful for solar generation monitoring to make important industry decisions for solar power utilisation in national electricity markets.

2.0 Materials and Method

2.1 Study Area

We implement newly developed KRR model for cloud cover bias correction in Queensland, which is referred to as Australia's "Sunshine State," with enormous solar energy potential [53, 54]. Under the United Nations Sustainable Development Goal #7 (SDG7) [55], the State government is committed towards increasing renewable energy uptake to account for up to 50% of the overall future energy supply by 2030. These projects represent an investment of \$8.5 billion, the creation of 7000 jobs, the generation of 4600 MW of renewable electricity, and the reduction of more than 11 million tonnes of air pollutants. As of January 2021, Queensland has 6200 megawatts (MW) of renewable energies including rooftop solar systems. According to Queensland Government, renewable energy accounts for 20% of electricity consumed [56] so they have set a 50% renewable energy generating capacity by 2030. To add value to research methodologies that assist solar energy producers, we aim to correct bias in TCDC_{GFS-Forecast}

obtained from the Columboola Solar Farm in Queensland, Australia which is expected to support 100,000 households. This Solar Farm is also expected to create ~440 GWh of renewable energy annually when it gets complete in 2022, which is enough to power 75,000 households over a 35-year period. Figure 1 shows the geographic location of the present study site where the KRR model for cloud cover bias correction was implemented.

[FIGURE 1]

2.2 Global Forecasting System

We develop KRR model using the Global Forecasting System (GFS) managed by National Oceanic and Atmospheric Administration (NOAA) that aims to deliver operational set of global weather predictions [58]. The GFS data repository aims to produce forecast variables up to 16 days in advance with a temporal resolution of 3h and 6h, and a spatial resolution of $0.25^\circ \times 0.25^\circ$ [59]. The GFS is not a frozen system, so its dynamic core and physical package is modified regularly [60]. For example, after a single-member prediction was replaced by GFS ensemble mean forecast in late 2001, this method was modified again in late 2003 to properly incorporate the bias-corrected GFS ensemble mean forecast [61, 62].

As this physics-based model is initialised every three hours, newly predicted variables are generated eight times a day at 0 UTC, 3 UTC, 6 UTC, 9 UTC, 12 UTC, 15 UTC, 18 UTC, 21 UTC, and 24 UTC. The GFS utilises Global Data Assimilation System (GDAS) [63] that augments a gridded three-dimensional model space with surface observations, balloon data, wind profiler data, buoy observations, radar observations, or satellite observations. The GDAS Model output is emulated four times daily and includes projections for the next three hours, six hours, and nine hours.

The present study attempts to build a new modelling strategy to correct inherent bias in GFS-derived total cloud cover forecasts (*i.e.*, $\text{TCDC}_{\text{GFS-Forecast}}$) for 3 distinct forecast horizons, which according to Brisbane daytime zones, are: at 0 UTC (10 AEST), 3 UTC (13 AEST), and 6 UTC (16 AEST). The 3-h GFS forecast experiments that are initialized from 0000 UTC compared to AEST (Australian Eastern Standard Time) as illustrated schematically in Figure 3. For comparison, the GFS-analysis total cloud cover ($\text{TCDC}_{\text{GFS-Analysis}}$) is used as a proxy for observed variables. We also utilised temperature ($\text{T2m}_{\text{GFS-Forecast}}$), downward shortwave radiation flux ($\text{DSWRF}_{\text{GFS-Forecast}}$), windspeeds (UGFS-Forecast , and $\text{V}_{\text{GFS-Forecast}}$) to further improve the bias through our newly proposed KRR modelling strategies.

2.3 Theoretical Overview

This section summarises the objective model (*i.e.*, KRR) and the conventional bias correction model (*i.e.*, MRNBC). Technical details of the other ML methods such as the decision tree [64], random forest [65, 66], multivariate adaptive ridge regression [67], Bayesian ridge regression [68], *k*-nearest

neighborhood [27, 69], gradient boosting regression [70] and the histogram-based gradient boosting regression [71] are explained elsewhere.

2.3.1 Kernel ridge regression

Kernel ridge regression (KRR) is a novel algorithm with an unlimited number of non-linear transformations of the independent variables used as regressors [72]. KRR model utilises ML strategy based on kernel and ridge regressions [41] to avoid issues of overfitting found in other regression methods. It therefore utilizes regularizations and a kernel technique to capture non-linear connections viz [46].

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

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

The Hilbert normed space of equation (1) is defined as $\| \cdot \|_H$. For a given $m \times m$ kernel matrix, K is developed by $\omega(x_p, x_o)$ from some fixed predictor variables where y is the input $q \times 1$ regression vector and is the $q \times 1$ unknown situation vector that reduces as follows:

$$y = (K + \lambda qI) \quad 3$$

$$\tilde{y} = \sum_{p=1}^q \alpha_o \omega(x_o, \tilde{x}) \quad 4$$

In model training stage, KRR technique is applied by solving Eq. (3) but utilised to predict the regression of an unknown sample x in Eq (4) in a testing step. To achieve highest accuracy possible, linear, polynomial, and Gaussian kernels are employed [42, 43, 73].

2.3.2 Multivariate recursive nesting bias correction (MRNBC):

As a traditional method, the multivariate recursive nesting bias correction (MRNBC) approach aims to correct the seasonal and non-seasonal time series based on multivariate autoregressive modelling. First introduced by Mehrotra et al. (2018), the MRNBC aims to incorporate the Recursive Nested Bias Correction (RNBC) so in this approach, the TCDC_{GFS-Forecast} simulations are nested into the observed data for all timescales of interest. Before applying the nesting, both timeseries are standardised to a mean of zero and a standard deviation of 1.

With m predictor variables at an i time step for a $Z (m \times t)$ matrix, the lag-one autocorrelation and the lag-one and lag-zero cross-correlation in TCDC_{GFS-Forecast} simulations can be modified to match the observed correlations in the time and space [74]. The multivariate autoregressive order 1 (MAR1) model for TCDC_{GFS-forecast} data and observed variables is therefore expressed as follows [75]:

$$\hat{Z}_i^h = C \hat{Z}_{i-1}^h + D_{\epsilon i} \quad 5$$

$$\hat{Z}_i^g = E \hat{Z}_{i-1}^g + F_{\epsilon i} \quad 6$$

where Z^h represents the observations and Z^g is the TCDC_{GFS-forecast} data. Data are standardised to construct a periodic time series \hat{Z}_i^g to be modified to match the observation \hat{Z}_i^h where εi is a mutually independent vector with random variation having zero mean value and an identity covariance matrix. C and D are lag-zero and lag-one cross-correlation coefficient matrices for observation \hat{Z}_i^h and the coefficients E and F are calculated for the standardised TCDC_{GFS-forecast} output.

Equations (5) and (6) are rearranged and modified \hat{Z}_i^g along with lag-zero and lag-one correlation matrices such as C and D to \hat{Z}_i^g have the desired dependence properties [75].

$$\hat{Z}_i^h = C Z_{i-1}'^g + D F^{-1} \hat{Z}_i^g - D F^{-1} E \hat{Z}_{i-1}^g \quad 7$$

For correction of periodic parameters, let vectors and $Z_{t,i}^h$ and $Z_{t,i}^g$ represent the observations and the TCDC_{GFS forecast} outputs, respectively, with m variables for month i and year t . The standardised periodic time series with a mean of zero and a unit variance is denoted as $\hat{Z}_{t,i}$. Following Eq. (7), the series $\hat{Z}_{t,i}'^g$ which maintains the observed lag-one serial and cross dependence as follows [75]:

$$\hat{Z}_{t,i}'^g = C_i Z_{t,i-1}'^g + D_i F_i^{-1} \hat{Z}_{t,i}^g - D_i F_i^{-1} E_i \hat{Z}_{t,i-1}^g \quad 8$$

Here $Z_{t,i-1}'^g$ = corrected time series from a previous month in year t . After corrections, the resulting time series Z'^g is rescaled by observed mean and standard deviation to yield the final corrected time series \bar{Z}^g whose are found elsewhere [74, 76, 77].

After correcting monthly timeseries, Z is combined to produce seasonal sequence and the periodic correction. This timeseries is connected to an annual time series and the correlation, standard deviation, and mean are corrected to form Ag (A = matrix of yearly data, $p \times n/12$). Subsequently, each time, aggregation corrections can be applied to daily time series to create a simple correction step [78]:

$$\bar{Z}_{i,j,s,t}^g = \left(\frac{\bar{Y}_{j,s,t}^g}{Y_{j,s,t}^g} \right) \times \left(\frac{\bar{S}_{s,t}^g}{S_{s,t}^g} \right) \times \left(\frac{\bar{A}_t^g}{A_t^g} \right) \times Z_{i,j,s,t}^g \quad 9$$

Here $\bar{Y}_{j,s,t}^g$, $\bar{S}_{s,t}^g$ and \bar{A}_t^g indicate the monthly, seasonally, and annually corrected values, respectively, and $Y_{j,s,t}^g$, $S_{s,t}^g$ and A_t^g represent the accumulated monthly, seasonal, and annual values accordingly. The subscript i stands for day, j for the day, s for the season, and t for the year. The three-step bias correction technique confirms that future variation is not influenced by bias correction procedure utilised to correct TCDC_{GFS-Forecast} [76].

2.4 Implementation of the Machine Learning (ML)-based Bias Correction

The fundamental idea behind bias correction is to identify a sufficiently adaptable and flexible approach that is capable of learning from an available data and then constructing a prediction function that performs well across the projection period (*i.e.*, forecast horizon). To develop a robust bias correction (*i.e.*, one that can precisely reduce bias in TCDC_{GFS-Forecast} data produced by GFS weather model at the Columboola solar farm), it was critical to firstly optimise the architecture of the proposed KRR model,

and then to take advantage of the associative links between the bias-corrected TCDC and the learned ML model. An ML-based Python package [79], *scikit-learn* [80, 81], was thus employed to develop the objective (*i.e.*, KRR) and other benchmark (*i.e.*, BNR, DTR, GBR, HGBR, KNN, MLR, XGB, and RF) models. For the case of MARS model, we have used *py-earth* package, and programming software R for traditional bias correction (*i.e.*, MRNBC) prescribed by Mehrotra et al. [77]. As defined in Section 2.5, six statistical measures are used to evaluate experimental outcome of the bias-corrected model, created using Intel i7 processor running at 3.6GHz and 16 GB RAM. Visualisation of bias-corrected TCDC dataset were made through *matplotlib* [82], *seaborn* [83] and Microsoft Excel.

[FIGURE 2]

[FIGURE 3]

Figure 2 is a schematic representation of KRR-based bias correction approach including the conventional (*i.e.*, multivariate recursive nested bias correction, MRNBC) methods. In summary we developed the proposed KRR method as follows:

(a) **Data:** GFS-forecast and GFS-analysis data were downloaded from NCEP repository: <https://www.ncei.noaa.gov/products/weather-climate-models/global-forecast>. As this repository provides 384-hours ahead data at a 3-hr interval, this study has only measured three time periods within the Brisbane daytime zone considering the relevance to solar PV power production at 0 UTC, 3 UTC, and 6 UTC. Figure 3 shows a schematic illustration of 3-h GFS forecast experiments initialized at 0000 UTC, compared with the Australian Eastern Standard Time (AEST). We adopted the *pygrib* python package to extract five selected variables and the datasets were sorted for Day-2 to Day-8 forecast. To apply the bias correction method, we adopted the TCDC_{GFS-Analysis} dataset as a proxy for the observation and used these to correct the systemic biases that were present the TCDC_{GFS-Forecast} dataset.

(b) **Pre-possessing and post-processing:** Missing values were replaced using a preceding seven data point and all data normalised to be bounded by [0, 1] [84]. As the TCDC dataset had significant zero values which is normal for cloud cover properties (*i.e.*, the presence of no clouds) and that this can affect an ML model's performance, we used four normalization techniques and the best normalization technique was selected based on minimum mean absolute error (MAE). These techniques were: max-min normalization (T_{MinMax}), maximum absolute normalization (T_{MaxAbs}), z-score normalization (T_{Std}), and robust scaler normalization (T_{Robust}) and their mathematical formulations are:

$$(a) \text{ Max-min normalization } (T_{MinMax}) = \frac{T_i - T_{min}}{T_{max} - T_{min}} \quad 10$$

$$(b) \text{ z-score normalization } (T_{Std}) = \frac{T_i - \bar{T}_i}{Std} \quad 11$$

$$(c) \text{ Maximum Absolute normalization } (T_{MaxAbs}) = \frac{T_i}{Max(Abs(x))} \quad 12$$

$$(d) \text{ Robust scaler normalization } (T_{Robust}) = \frac{T_i - T_\omega}{(Q_3 - Q_1)} \quad 13$$

In Eq. (10 - 13), T_i = respective predictors, \bar{T}_i = average of T_i , T_{min} = minimum value for predictors, T_{max} = maximum value and Std = standard deviation, T_ω = median of T_i and $(Q_3 - Q_1)$ = interquartile range between 1st quartile (25th) and 3rd quartile (75th) quantile. As there is no specific rule for data partitioning [84, 85] we used 70% training, 15% testing with a validation set as the last 15% of training set for all data collected between 1 January 2019 and 30 April 2020.

(c) *Implementation of ML-based Bias Correction*

This study has developed a total of 10 different ML models (*i.e.*, KRR with nine other benchmark models) to correct the bias in TCDC_{GFS-Forecast} for Day-2 to Day-8 forecasts. Our multivariate adaptive regression splines (MARS) model considers multivariate data with basic functions to investigate the predictor variable and identifies the predictor and target features [87]. The decision tree (DTR) was our non-parametric, supervised system to approximate a sine curve using ‘if-then-else’ decision where generally, the deeper the tree the more complicated a rule could be to fit a model. A prime task of ML is to set hyper-parameters for optimal bias correction method so an optimum architecture of the KRR model was created using GridSearchCV (regularization strength, $\alpha = 1.5$; gamma parameter = *None*, with a degree of the polynomial kernel = 3 and kernel = *linear*; see Table 4). The performance of ML bias correction was compared with traditional bias corrections (*i.e.*, MRNBC), and the reference value usually calculated between the TCDC_{GFS-Forecast} and the TCDC_{GFS-Analysis} was used with TCDC_{GFS-Analysis} considered as the proxy of the observed cloud cover dataset.

(d) *Implementation of MRNBC Bias Correction Method*

In this section, we detail the procedure developed to correct bias using MRNBC methods, a traditional non-ML approach used previously. We made univariate adjustments followed by multivariate corrections using a time series with appropriate bias correction statistics generated for all variables and location. Therefore, the MRNBC method corrected the bias in TCDC_{GFS-Forecast} by removing the current GFS mean and adding the observed mean. The time series adjusted in Step-2 are standardised, and this residual time series is adapted for bias using auto and cross-correlations for day lag-1 and lag-0. To summarise the corrections necessary at each time scale, a weighting factor may also be computed. The TCDC_{GFS-Forecast} daily time series is multiplied by the weighting factor from each time scale to produce the final bias-corrected time series. The MRNBC bias correction procedure is schematized in Figure 4.

[FIGURE 4]

2.5 Evaluation of ML-Based Bias Correction Method

The effectiveness of an ML-based bias correction method in comparison with the traditional bias correction (*i.e.*, MRNBC) and reference value (calculated between TCDC_{GFS-Analysis} and TCDC_{GFS-Forecast}) method was evaluated using performance metrics such as the Pearson's Correlation Coefficient (r), root means square error (RMSE, %), and mean absolute error (MAE, %) in the testing phase. In its most general sense, the effectiveness of any model is determined by the corrected (*i.e.*, TCDC_{BC}) and the proxy of the observed (TCDC_{GFS-analysis}) datasets. While RMSE is a more appropriate measure of performance than MAE when error distribution is Gaussian [88], for a more persuasive model, the Willmott's Index (WI) [89-91] and Legates –McCabe's Index (LM) [92-94] are employed in this research. Mathematically, these are expressed as follows:

Correlation coefficient (r):

$$r = \left\{ \frac{\sum_{i=1}^N (TCDC_{BC} - \overline{TCDC}_{ANL})(TCDC_{BC} - \overline{TCDC}_{BC})}{\sqrt{\sum_{i=1}^N (TCDC_{ANL} - \overline{TCDC}_{ANL})^2 \sum_{i=1}^N (TCDC_{BC} - \overline{TCDC}_{BC})^2}} \right\}^2 \quad 14$$

Mean absolute error (MAE: %):

$$MAE = \frac{1}{N} \sum_{i=1}^N |TCDC_{BC} - TCDC_{ANL}| \quad 15$$

Root mean squared error (RMSE: %):

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (TCDC_{BC} - TCDC_{ANL})^2} \quad 16$$

Willmott's Index of Agreement (d):

$$d = 1 - \left[\frac{\sum_{i=1}^N (TCDC_{BC} - TCDC_{ANL})^2}{\sum_{i=1}^N (|TCDC_{BC} - \overline{TCDC}_{ANL}| + |TCDC_{ANL} - \overline{TCDC}_{ANL}|)^2} \right] \quad 17$$

Legates –McCabe's Index (LM):

$$LM = 1 - \left[\frac{\sum_{i=1}^N |TCDC_{BC} - TCDC_{ANL}|}{\sum_{i=1}^N |TCDC_{ANL} - \overline{TCDC}_{ANL}|} \right] \quad 18$$

Mean Absolute Percentage Deviation (MAPD: %):

$$MAPD (\%) = \frac{1}{N} \left(\sum_{i=1}^N \left| \frac{TCDC_{BC} - TCDC_{ANL}}{TCDC_{ANL}} \right| \right) * 100 \quad 19$$

In Eq. (14–19) we note that the $TCDC_{ANL}$ and $TCDC_{BC}$, respectively, represents the proxy of the observed (TCDC_{GFS-Analysis}) and bias-corrected data series for i^{th} tested value, and \overline{TCDC}_{ANL} and \overline{TCDC}_{BC} refer to their average values, accordingly, and the number of observations is denoted by N , while the coefficient of variation is denoted by CV.

In comparing the different models adopted to this bias correction problem, this study has used the promoting percentage of Legate McCabe's Index ($\Delta_{LM}(\%)$) as a complementary measure of the model efficiency. The $\Delta_{LM}(\%)$ was thus calculated comparing the actual LM obtained using the proposed KRR and LM values generated by the KNN, MARS, and RF model. Mathematically, the $\Delta_{LM}(\%)$ is computed as follows:

$$\Delta_{LM}(\%) = \left(\frac{LM_{KRR} - LM_{COM}}{LM_{KRR}} \right) \times 100 \quad 20$$

In Eq. (20) the LM_{COM} represents the LM value of the benchmark (*e.g.*, KNN, MARS, or RF) model.

3.0 Results and Discussion

The practicality of ML-based bias correction method developed using KRR model is established using two distinct approaches where the bias in TCDC_{GFS-Forecast} data is reduced relative to the observed (TCDC_{GFS-Analysis}) data. The first approach integrates five GFS forecast data series (*i.e.*, TCDC_{GFS-Forecast}, T2m_{GFS-Forecast}, DSWRF_{GFS-Forecast}, U_{GFS-Forecast}, V_{GFS-Forecast}) as inputs in the model's training phase while the 2nd approach uses a single TCDC_{GFS-Forecast} data in the model's training phase. In the initial phase, we examined 10 ML-based bias correction methods pooled together to broadly identify the bias correction performance in comparison with conventional bias correction method (*i.e.*, MRNBC) and the respective reference values calculated between TCDC_{GFS-Forecast} and TCDC_{GFS-Analysis} for our study site. The ML models (*i.e.*, BNR, DTR, GBR, HGBR, KNN, KRR, MARS, MLR, XGB, and RF) were assessed using statistical metrics (Eq. 15–19), infographics, and visualisations to determine the degree of agreement between TCDC_{GFS-Forecast} and TCDC_{GFS-Analysis}. Performance metrics indicate that the proposed KRR model outperforms all comparative counterpart models in the testing phase and that these models attain a superior value of r and d , with a low value of $RMSE$ and MAE in the independent testing phase.

According to results presented in Figures 5 and 6, an in-depth examination of the Willmott's Index (d) and root mean squared error ($RMSE$) accordingly provides a persuasive evidence that ML approach has substantial benefits in reducing the bias compared with the traditional MRNBC method and the respective reference values tested for all the days over which GFS total cloud cover forecast is considered. This figure clearly shows the closer distribution of $RMSE$ and d values for the case of ML models using Approach 2 (see Figs. 5b & 6b) compared with Approach 1 (Figs. 5a & 6a). The lower end of the plot for the value of d is relatively situated within the lower quartile (25th) and the upper quartile (75th) range for the Day-2 GFS forecast data series.

There appears to also be a single outlier found further than the 75th percentile. However, for the Day-3 to Day-8 forecasts, the bias correction of TCDC_{GFS-Forecast} time series are found to result in a lesser improvement, except for Day-6 forecasts, which is reasonable as the uncertainties in TCDC are

likely to increase with an increment in forecast horizon. Noticeably, as the forecasting period changes from Day-2 to Day-8, the performance for our bias correction model decreases significantly. Despite this, we can note through the overall findings in the visualized result that the ML model can be considered the most potent strategy for bias correction at solar farms, at least for the present study site and the suite of models considered.

Further analysis is performed through a boxplot of errors (*i.e.*, RMSE) for results obtained through Approach-2. This shows the bias-corrected total cloud cover ($TCDC_{BC}$) *vs.* $TCDC_{GFS-Analysis}$ of all the ML models as illustrated in Figure 5b. For Day-2 $TCDC_{GFS-Forecast}$ data series, it is noticeable that the dispersion of RMSE for bias correction methods concerning the quartile values has distinct outliers. The lower end of the boxplot seems to precisely lie between lower quartile (25th percentile) and upper quartile (75th percentile). Likewise, the correlation coefficient (d) and RMSE are higher for the other days (Day-2 to Day-8) forecast except for Day-6. Therefore, the improvement of bias using ML methods signifies potentially improved performance compared with the MRNBC and the respective reference values of the $TCDC_{GFS-Forecast}$ and $TCDC_{GFS-Analysis}$. When data from the other models were compared, the accuracy of KRR-based bias correction outweighed those of the other ML models (see Figure 5).

[FIGURE 5]

[FIGURE 6]

To investigate the performance of ML-based bias correction, the mean absolute error (MAE) for ten machine learning models is listed (Table 2), along with traditional bias correction method (MRNBC) and the reference value method. The boxplots of bias-corrected root mean square error (RMSE) calculated between data for all the nine ML-based bias correction methods pooled together (*i.e.*, KRR, BNR, DTR, GBR, HGBR, KNN, MARS, RF, XGB), conventional bias correction method (*i.e.*, MRNBC) and along with their respective reference values (*RMSE* calculated between $TCDC_{GFS-Forecast}$ and $TCDC_{GFS-Analysis}$) are also shown (Figure 6). When used to correct TCDC, it appears that the KRR model with Approach-2 produces the highest MAE compared to other machine learning models and reference value method. For Approach-2, the MAE for Day 2 forecast is bounded by [20.20, 26.75] %, with a best value obtained for the proposed KRR, which also indicates a modest 14% improvement over the reference MAE value. A similar type of improvement in the cloud cover bias is also seen for Day-3 to Day-7. The KRR model's performance is also compared to the KNN model's performance for Day 2 forecasting. However, it is imperative to note that Approach 1, which employs a MARS model, was more effective in correcting the TCDC bias for the Day 8 cloud cover forecasts relative to Approach 1. Consequently, the ML-based KRR model is seen to outperform the classic bias correction strategy in correcting the GFS-derived TCDC. In accordance with this result, the four best methods (*i.e.*, KNN,

KRR, MARS, and RF) were then chosen to conduct an in-depth examination of the bias correction approaches utilizing these machine learning models.

[FIGURE 7]

An evaluation of the robustness of the four selected ML models (*i.e.*, KNN, including the KRR, MARS, and RF), with the correlation coefficient (r) for the study site for Approach-1 and Approach-2 are plotted in Figure 7. For the proposed model (*i.e.*, KRR), the bias correction performance in terms of the r value shows substantial performance for Approach-1 ($r \approx 0.69$) and Approach-2 ($r \approx 0.76$) for Day-2 TCDC_{GFS-Forecast}. The KRR model accuracy is then followed by the MARS model ($r \approx 0.731$). For the Day-3 to Day-8 forecast data, the bias correction performance in terms of r ranged from 0.21 to 0.60 for Approach-2; Day-6 forecast data has the highest ($r \approx 0.61$). Reasonably, the proposed KRR model demonstrates the highest r values for Day-3 to Day-8 forecast data of TCDC bias correction. Considering the traditional bias correction method (*i.e.*, MRNBC) and reference r -value, for Day-2 of GFS forecast TCDC for Approach-2, the bias correction performance in terms of r is increased by 22% and 53% accordingly. Similarly, for Day-3 to Day-8 of TCDC bias correction, the improvement is 70% to 75% which is significant. Because the benchmark models performed poorly, as demonstrated in Figure 7, the newly proposed KRR model was superior for the research study site.

[FIGURE 8]

The change (∇) in mean absolute percentage error $MAPD$ (%) generated by the proposed KRR method in respect to reference value deduced from TCDC_{GFS-Forecast} and TCDC_{GFS-Analysis} is presented in Figure 8. A positive change is expected to show the objective model (*i.e.*, KRR) outperforming the benchmark model. For both approaches, $\nabla MAPD$ (%) is significant for Day-2 GFS forecast whereas Approach 2 with KRR shows the lowest value at ~48%. For Approach 1, the MAE value from an SVR model is ~17.5% higher whereas for Day-3 to Day-8 forecasts, $\nabla MAPD$ range from [5, 35] % for Approach 2 with some deviation noted for KNN model. In a rational sense, the proposed KRR model is seen to demonstrate the most significant improvement in $MAPD$ ($\nabla MAPD$; %) ranging from 15% to 14% for Day-2 to Day-8 in respect to a reduction in bias for the TCDC dataset. Accordingly, we can ascertain that our newly developed KRR model appears to fall within the criterion of an acceptable predictive model that can correct the bias in GFS-derived total cloud cover forecasts, and therefore, may be a useful tool for solar energy monitoring and forecasting systems.

[FIGURE 9]

To further demonstrate the KRR model's particular skills to correct TCDC_{GFS-Forecast} data biases for Day 2-8 forward periods, we show the LM values that aim to compare the promoting percentages referred to as an incremental performance in LM (Δ_{LM} , %) of the comparative model against our

objective model (*i.e.*, KRR model). Figure 9 shows the result of KRR model against those of KNN, MARS, and RF applied to correct bias in TCDC data for Day-2 to Day-8 forecast horizons. The bias correction, evaluated in respect to the KRR model, is relatively diverse. Notwithstanding this, Figure 9 shows that the bias correction using the proposed KRR method is more notable by 20% to 65% for all the forecasted days. Overall, the highest gain appears to have reached ~70% for the KNN model for the case of Day-4 forecasts of the total cloud cover conditions.

[FIGURE 10]

Figure 10 is an alternative representation of KRR model's performance in respect to benchmark models using a Taylor diagram [95]. In this case, a significant correlation seems to exist between bias-corrected TCDC (*i.e.*, TCDC_{BC}) and the proxy observed variable (TCDC_{GFS-Analysis}) for the case of the proposed KRR model. It is therefore clear that the bias corrected TCDC data produced from KRR model is close match to the proxy of the observed TCDC data with the other ML models. Therefore, in a nutshell, based on the statistical performance measures, we can ascertain that the newly developed KRR model has the predictive skills to reduce the overall bias in total cloud cover generated by a weather simulation model used in this study.

4.0 Conclusions and Future Research Insights

This paper has utilised an ML-based bias correction (*i.e.*, KRR) method to reduce bias in total cloud (TCDC_{GFS-Forecast}) variable at solar energy farm. To demonstrate the feasibility of the developed KRR model, data from Columboola solar energy farm located in Queensland, Australia, were used where the findings indicated a superior performance of this objective model in respect to an ensemble of machine learning and conventional bias correction methods. We learned that the ML-based bias correction approach had a solid potential to significantly reduce, if not eradicate, the bias in TCDC_{GFS-Forecast} by utilising cloud cover, temperature, windspeed and downward solar radiation flux that provides adequate predictive features and relationships in observed cloud cover variable. Precisely, the KRR model's capability to correct the bias in TCDC_{GFS-Forecast} dataset was established in terms of the percentage improvement in mean bias error that for this study site has ranged from ~20% to ~50% using traditional MRNBC method for Day-2 to Day-8 forecast. The study showed that the integration of multiple predictor variables such as the TCDC_{GFS-Forecast}, T2m_{GFS-Forecast}, DSWRF_{GFS-Forecast}, U_{GFS-Forecast}, and V_{GFS-Forecast} into the model's input matrix was able to successfully correct the bias in cloud cover as it provided historical information on cloud evolution and its lagged stochastic behaviour. Nonetheless, we contend that the biases in all these individual forecasted variables that are produced by the numerical weather model may also affect the accuracy of the cloud cover bias correction task. Therefore, using a single set of model input variable (*i.e.*, TCDC_{GFS-Forecast}) was somewhat better suited compared to the multi-

variable approach such that the results have established high predictive potency of employing a single variable to resolve the bias-related problem for this solar energy site.

These results have shown that the performance of ML-based bias correction for longer-term forecast horizon (*i.e.*, Day-8) was much better with Approach-1 where multiple predictor variables: TCDC_{GFS-Forecast}, T2m_{GFS-Forecast}, DSWRF_{GFS-Forecast}, U_{GFS-Forecast}, and V_{GFS-Forecast} were incorporated in the KRR model's input matrix. This outcome appears to reveal the interactions of these variables with proxy-observed cloud cover over the passage of time leading to improved overall performance *i.e.*, for a longer-term Day-8 bias correction result although this multi-variable approach (*i.e.*, Approach-1) registered comparatively large bias. While the results of this pilot study may not be explicitly conclusive and may require further investigations, one possible explanation for comparatively large bias could be the interference of disproportionately embedded biases within each of these forecast variables that could hinder the correlation among such bias to further affect TCDC produced by the GFS model. We therefore conclude that in a future study, the development of deep learning algorithms that have exceptional skills in terms of extracting the more complex data features may hold a greater promise to correct bias in real-time weather model data used for solar energy monitoring. Some relevance may be drawn from recent studies where deep learning was broadly implemented, for example, in hydrology [29, 31] and solar energy studies [97, 98]. Therefore, a deep learning hybrid approach could be adopted as a future bias correction method both for solar power production monitoring and power failure risk analysis when solar energy is integrated into real energy grids.

Credit authorship contribution statement

A. A. Masrur Ahmed: Writing - original draft, Conceptualization, Methodology, Software, Data Curation, Formal analysis, Investigation, Model development and application. **Ravinesh C. Deo:** Conceptualization, Investigation, Project administration, Writing - review & editing, Investigation, Supervision. **Gary Segal:** Conceptualization. **Yanshan Yu:** Conceptualization.

Acknowledgment

The project was supported by the Australian Postgraduate Research Program (APR.intern). It was partly funded by the CS Energy, Queensland under an APR.intern program. We also thank the Australian Mathematical Sciences Institute (AMSI) for the development and administration of a research partnership between USQ and CS Energy.

References:

1. Christensen, J.H., et al., *On the need for bias correction of regional climate change projections of temperature and precipitation*. Geophysical Research Letters, 2008. **35**(20).
2. Ayar, PV, M. Vrac, and A. Mailhot, *Ensemble bias correction of climate simulations: preserving internal variability*. Scientific Reports, 2021. **11**(1): p. 1-9.

3. Evin, G., et al., *Partitioning uncertainty components of an incomplete ensemble of climate projections using data augmentation*. Journal of Climate, 2019. **32**(8): p. 2423-2440.
4. Zhang, J., et al., *Deep photovoltaic nowcasting*. Solar Energy, 2018. **176**: p. 267-276.
5. Mills, A., *Understanding variability and uncertainty of photovoltaics for integration with the electric power system*. 2009.
6. Baran, A., et al., *Machine learning for total cloud cover prediction*. Neural Computing and Applications, 2021. **33**(7): p. 2605-2620.
7. Matuszko, D., *Influence of the extent and genera of cloud cover on solar radiation intensity*. International Journal of climatology, 2012. **32**(15): p. 2403-2414.
8. World-Meteorological-Organization, *International Cloud Atlas: Manual on the observation of clouds and other meteors*. WMO-No. 407. 2017, World Meteorological Organization Geneva.
9. Wilby, R.L., et al., *Guidelines for use of climate scenarios developed from statistical downscaling methods*. Supporting material of the Intergovernmental Panel on Climate Change, available from the DDC of IPCC TGCIA, 2004. **27**: p. -.
10. Chen, J., et al., *Finding appropriate bias correction methods in downscaling precipitation for hydrologic impact studies over North America*. Water Resources Research, 2013. **49**(7): p. 4187-4205.
11. Piani, C., J. Haerter, and E. Coppola, *Statistical bias correction for daily precipitation in regional climate models over Europe*. Theoretical and Applied Climatology, 2010. **99**(1): p. 187-192.
12. Wood, AW, et al., *Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs*. Climatic change, 2004. **62**(1): p. 189-216.
13. Teutschbein, C. and J. Seibert, *Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods*. Journal of Hydrology, 2012. **456-457**: p. 12-29.
14. Li, H., J. Sheffield, and EF Wood, *Bias correction of monthly precipitation and temperature fields from Intergovernmental Panel on Climate Change AR4 models using equidistant quantile matching*. Journal of Geophysical Research, 2010. **115**(D10).
15. Johnson, F. and A. Sharma, *What are the impacts of bias correction on future drought projections?* Journal of Hydrology, 2015. **525**: p. 472-485.
16. Johnson, F. and A. Sharma, *A nesting model for bias correction of variability at multiple time scales in general circulation model precipitation simulations*. Water Resources Research, 2012. **48**(1).
17. Cannon, A.J., *Multivariate bias correction of climate model output: Matching marginal distributions and intervariable dependence structure*. Journal of Climate, 2016. **29**(19): p. 7045-7064.
18. Mao, G., et al., *Stochastic bias correction of dynamically downscaled precipitation fields for Germany through Copula-based integration of gridded observation data*. Hydrology and Earth System Sciences, 2015. **19**(4): p. 1787-1806.
19. Vrac, M. and P. Friederichs, *Multivariate—intervariable, spatial, and temporal—bias correction*. Journal of Climate, 2015. **28**(1): p. 218-237.
20. Leander, R. and T.A. Buishand, *Resampling of regional climate model output for the simulation of extreme river flows*. Journal of Hydrology, 2007. **332**(3-4): p. 487-496.
21. Leander, R., et al., *Estimated changes in flood quantiles of the river Meuse from resampling of regional climate model output*. Journal of Hydrology, 2008. **351**(3-4): p. 331-343.
22. Smitha, P.S., et al., *An improved bias correction method of daily rainfall data using a sliding window technique for climate change impact assessment*. Journal of Hydrology, 2018. **556**: p. 100-118.
23. Schmidli, J., C. Frei, and P.L. Vidale, *Downscaling from GCM precipitation: a benchmark for dynamical and statistical downscaling methods*. International Journal of Climatology: A Journal of the Royal Meteorological Society, 2006. **26**(5): p. 679-689.
24. Pour, SH, et al., *Model output statistics downscaling using support vector machine for the projection of spatial and temporal changes in rainfall of Bangladesh*. Atmospheric Research, 2018. **213**: p. 149-162.
25. Shi, Y., et al., *Mapping annual precipitation across mainland China in the period 2001–2010 from TRMM3B43 product using spatial downscaling approach*. Remote Sensing, 2015. **7**(5): p. 5849-5878.
26. Sa'adi, Z., et al., *Projection of spatial and temporal changes of rainfall in Sarawak of Borneo Island using statistical downscaling of CMIP5 models*. Atmospheric Research, 2017. **197**: p. 446-460.
27. Devak, M., C. Dhanya, and A. Gosain, *Dynamic coupling of support vector machine and K-nearest neighbour for downscaling daily rainfall*. Journal of Hydrology, 2015. **525**: p. 286-301.

28. Ahmed, A.M. and S.M.A. Shah, *Application of artificial neural networks to predict peak flow of Surma River in Sylhet Zone of Bangladesh*. International Journal of Water, 2017. **11**(4): p. 363-375.
29. Ahmed, A.M., et al., *Deep learning hybrid model with Boruta-Random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity*. Journal of Hydrology, 2021. **599**: p. 126350.
30. Yaseen, ZM, et al., *Complementary data-intelligence model for river flow simulation*. Journal of Hydrology, 2018. **567**: p. 180-190.
31. Ahmed, A., et al., *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, 2021. **13**(4): p. 554.
32. Ahmed, A.M., et al., *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, 2021: p. 1-31.
33. Haykin, S., *Kalman filtering and neural networks*. Vol. 47. 2004: John Wiley & Sons.
34. Lary, D.J., et al., *Machine learning and bias correction of MODIS aerosol optical depth*. IEEE Geoscience and Remote Sensing Letters, 2009. **6**(4): p. 694-698.
35. Tu, J.V., *Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes*. Journal of clinical epidemiology, 1996. **49**(11): p. 1225-1231.
36. Deo, R.C., O. Kisi, and V.P. Singh, *Drought forecasting in eastern Australia using multivariate adaptive regression spline, least square support vector machine and M5Tree model*. Atmospheric Research, 2017. **184**: p. 149-175.
37. Lazri, M. and S. Ameer, *Combination of support vector machine, artificial neural network and random forest for improving the classification of convective and stratiform rain using spectral features of SEVIRI data*. Atmospheric research, 2018. **203**: p. 118-129.
38. Roodposhti, M.S., T. Safarrad, and H. Shahabi, *Drought sensitivity mapping using two one-class support vector machine algorithms*. Atmospheric Research, 2017. **193**: p. 73-82.
39. Tripathi, S., V.V. Srinivas, and R.S. Nanjundiah, *Downscaling of precipitation for climate change scenarios: A support vector machine approach*. Journal of Hydrology, 2006. **330**(3-4): p. 621-640.
40. Li, W., et al., *Assessment for surface water quality in Lake Taihu Tiaoxi River Basin China based on support vector machine*. Stochastic Environmental Research and Risk Assessment, 2013. **27**(8): p. 1861-1870.
41. Zhang, Y., J. Duchi, and M. Wainwright. *Divide and conquer kernel ridge regression*. in *Conference on learning theory*. 2013. PMLR.
42. You, Y., et al. *Accurate, fast and scalable kernel ridge regression on parallel and distributed systems*. in *Proceedings of the 2018 International Conference on Supercomputing*. 2018.
43. Saunders, C., A. Gammerman, and V. Vovk, *Ridge regression learning algorithm in dual variables*. 1998.
44. Orsenigo, C. and C. Vercellis, *Kernel ridge regression for out-of-sample mapping in supervised manifold learning*. Expert Systems with Applications, 2012. **39**(9): p. 7757-7762.
45. Ali, M., et al., *Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts*. Journal of Hydrology, 2020. **584**: p. 124647.
46. Ali, M., et al., *Improving SPI-derived drought forecasts incorporating synoptic-scale climate indices in multi-phase multivariate empirical mode decomposition model hybridized with simulated annealing and kernel ridge regression algorithms*. Journal of Hydrology, 2019. **576**: p. 164-184.
47. Douak, F., F. Melgani, and N. Benoudjit, *Kernel ridge regression with active learning for wind speed prediction*. Applied energy, 2013. **103**: p. 328-340.
48. Mishra, S., S. Dhar, and P. Dash, *An effective battery management scheme for wind energy systems using multi Kernel Ridge regression algorithm*. Journal of Energy Storage, 2019. **21**: p. 418-434.
49. Naik, J., P. Satapathy, and P. Dash, *Short-term wind speed and wind power prediction using hybrid empirical mode decomposition and kernel ridge regression*. Applied Soft Computing, 2018. **70**: p. 1167-1188.
50. Alalami, M.A., M. Maalouf, and T.H. EL-Fouly. *Wind Speed Forecasting Using Kernel Ridge Regression with Different Time Horizons*. in *International Conference on Time Series and Forecasting*. 2019. Springer.
51. Zhang, S., et al., *Kernel Ridge Regression Model Based on Beta-Noise and Its Application in Short-Term Wind Speed Forecasting*. Symmetry, 2019. **11**(2): p. 282.

52. Dash, P., et al., *Point and interval solar power forecasting using hybrid empirical wavelet transform and robust wavelet kernel ridge regression*. Natural Resources Research, 2020: p. 1-29.
53. Salcedo-Sanz, S., et al., *An efficient neuro-evolutionary hybrid modelling mechanism for the estimation of daily global solar radiation in the Sunshine State of Australia*. Applied Energy, 2018. **209**: p. 79-94.
54. Zahedi, A., *Australian renewable energy progress*. Renewable and Sustainable Energy Reviews, 2010. **14**(8): p. 2208-2213.
55. Martin, D.A., *Linking fire and the United Nations sustainable development goals*. Science of the Total Environment, 2019. **662**: p. 547-558.
56. Works-DoEaP, *Achieving our renewable energy targets*. Queensland Government. 2021.
57. Wang, X., et al., *GSI 3DVar-based ensemble-variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments*. Monthly Weather Review, 2013. **141**(11): p. 4098-4117.
58. Ward, J., *Australian Solar Energy Forecasting System Final report: project results and lessons learnt 2016*.
59. Kistler, R., et al., *The NCEP–NCAR 50-year reanalysis: monthly means CD-ROM and documentation*. Bulletin of the American Meteorological society, 2001. **82**(2): p. 247-268.
60. Fan, Y. and H. van den Dool, *Bias correction and forecast skill of NCEP GFS ensemble week-1 and week-2 precipitation, 2-m surface air temperature, and soil moisture forecasts*. Weather and forecasting, 2011. **26**(3): p. 355-370.
61. Van den Dool, H., J. Huang, and Y. Fan, *Performance and analysis of the constructed analogue method applied to US soil moisture over 1981–2001*. Journal of Geophysical Research: Atmospheres, 2003. **108**(D16).
62. Huang, J., HM van den Dool, and K.P. Georgarakos, *Analysis of model-calculated soil moisture over the United States (1931–1993) and applications to long-range temperature forecasts*. Journal of Climate, 1996. **9**(6): p. 1350-1362.
63. Blankenau, P.A., *Bias and Other Error in Gridded Weather Data Sets and Their Impacts on Estimating Reference Evapotranspiration*. 2017.
64. Senthil Kumar, A.R., et al., *Modeling of Suspended Sediment Concentration at Kasol in India Using ANN, Fuzzy Logic, and Decision Tree Algorithms*. Journal of Hydrologic Engineering, 2012. **17**(3): p. 394-404.
65. Feng, Y., et al., *Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration modelling*. Agricultural Water Management, 2017. **193**: p. 163-173.
66. Breiman, L., *Random Forests*. Machine Learning, 2001. **45**: p. 5–32.
67. Al-Musaylh, M.S., et al., *Short-term electricity demand forecasting using machine learning methods enriched with ground-based climate and ECMWF Reanalysis atmospheric predictors in southeast Queensland, Australia*. Renewable and Sustainable Energy Reviews, 2019. **113**.
68. Xu, W., et al., *Blood-based multi-tissue gene expression inference with Bayesian ridge regression*. Bioinformatics, 2020. **36**(12): p. 3788-3794.
69. Shabani, S., et al., *Modeling pan evaporation using gaussian process regression k-nearest neighbors random forest and support vector machines; Comparative analysis*. Atmosphere, 2020. **11**(1): p. 66.
70. Cai, J., et al., *prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest*. Applied energy, 2020. **262**: p. 114566.
71. Guryanov, A. *Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees*. in *International Conference on Analysis of Images, Social Networks and Texts*. 2019. Springer.
72. Exterkate, P., *Model selection in kernel ridge regression*. Computational Statistics & Data Analysis, 2013. **68**: p. 1-16.
73. Alaoui, A., et al., *Modelling the effects of land use and climate changes on hydrology in the Ursern Valley, Switzerland*. Hydrological processes, 2014. **28**(10): p. 3602-3614.
74. Sarhadi, A., et al., *Water resources climate change projections using supervised non-linear and multivariate soft computing techniques*. Journal of hydrology, 2016. **536**: p. 119-132.
75. Salas, J.D., GQ Tabios III, and P. Bartolini, *Approaches to multivariate modeling of water resources time series 1*. JAWRA Journal of the American Water Resources Association, 1985. **21**(4): p. 683-708.
76. Mehrotra, R. and A. Sharma, *Correcting for systematic biases in multiple raw GCM variables across a range of timescales*. Journal of Hydrology, 2015. **520**: p. 214-223.

77. Mehrotra, R., F. Johnson, and A. Sharma, *A software toolkit for correcting systematic biases in climate model simulations*. Environmental Modelling & Software, 2018. **104**: p. 130-152.
78. Srikanthan, R. and G.G.S. Pegram, *A nested multisite daily rainfall stochastic generation model*. Journal of Hydrology, 2009. **371**(1-4): p. 142-153.
79. Sanner, M.F., *Python: a programming language for software integration and development*. J Mol Graph Model, 1999. **17**(1): p. 57-61.
80. Kramer, O., *Scikit-learn*, in *Machine learning for evolution strategies*. 2016, Springer. p. 45-53.
81. Pedregosa, F., et al., *Scikit-learn: Machine learning in Python*. Journal of machine learning research, 2011. **12**(Oct): p. 2825-2830.
82. Barrett, P., et al. *matplotlib--A Portable Python Plotting Package*. in *Astronomical data analysis software and systems XIV*. 2005.
83. Waskom, M., et al., *Seaborn: statistical data visualization*. Astrophysics Source Code Library, 2020: p. ascl: 2012.015.
84. Ahmed, A.A.M., *Prediction of dissolved oxygen in Surma River by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (ANNs)*. Journal of King Saud University - Engineering Sciences, 2017. **29**(2): p. 151-158.
85. Deo, R.C., X. Wen, and F. Qi, *A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset*. Applied Energy, 2016. **168**: p. 568-593.
86. Prasad, R., et al., *Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition*. Geoderma, 2018. **330**: p. 136-161.
87. Deo, R.C., et al., *Very short-term reactive forecasting of the solar ultraviolet index using an extreme learning machine integrated with the solar zenith angle*. Environ Res, 2017. **155**: p. 141-166.
88. Chai, T. and R.R. Draxler, *Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature*. Geoscientific Model Development, 2014. **7**(3): p. 1247-1250.
89. Willmott, C.J. and K. Matsuura, *Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance*. Climate research, 2005. **30**(1): p. 79-82.
90. Willmott, C.J., S.M. Robeson, and K. Matsuura, *A refined index of model performance*. International Journal of Climatology, 2012. **32**(13): p. 2088-2094.
91. Willmott, C.J., et al., *Statistics for the evaluation and comparison of models*. Journal of Geophysical Research: Oceans, 1985. **90**(C5): p. 8995-9005.
92. Legates, D.R. and R.E. Davis, *The continuing search for an anthropogenic climate change signal: Limitations of correlation-based approaches*. Geophysical Research Letters, 1997. **24**(18): p. 2319-2322.
93. Legates, D.R. and G.J. McCabe, *Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation*. Water Resources Research, 1999. **35**(1): p. 233-241.
94. Legates, D.R. and G.J. McCabe, *A refined index of model performance: a rejoinder*. International Journal of Climatology, 2013. **33**(4): p. 1053-1056.
95. Taylor, K.E., *Summarizing multiple aspects of model performance in a single diagram*. Journal of Geophysical Research: Atmospheres, 2001. **106**(D7): p. 7183-7192.
96. Jovanovic, B., et al., *A high-quality monthly total cloud amount dataset for Australia*. Climatic change, 2011. **108**(3): p. 485-517.
97. Ghimire, et al., *Deep Learning Neural Networks Trained with MODIS Satellite-Derived Predictors for Long-Term Global Solar Radiation Prediction*. Energies, 2019. **12**(12).
98. Ghimire, S., et al., *Deep solar radiation forecasting with convolutional neural network and long short-term memory network algorithms*. Applied Energy, 2019. **253**.

Credit authorship contribution statement

A. A. Masrur Ahmed: Writing - original draft, Conceptualization, Methodology, Software, Data Curation, Formal analysis, Investigation, Model development and application. **Ravinesh C. Deo:** Conceptualization, Investigation, Project administration, Writing - review & editing, Investigation, Supervision. **Gary Segal:** Conceptualization. **Yanshan Yu:** Conceptualization.

Declaration of Interest

The authors declare no conflict of interest.

List of Figures

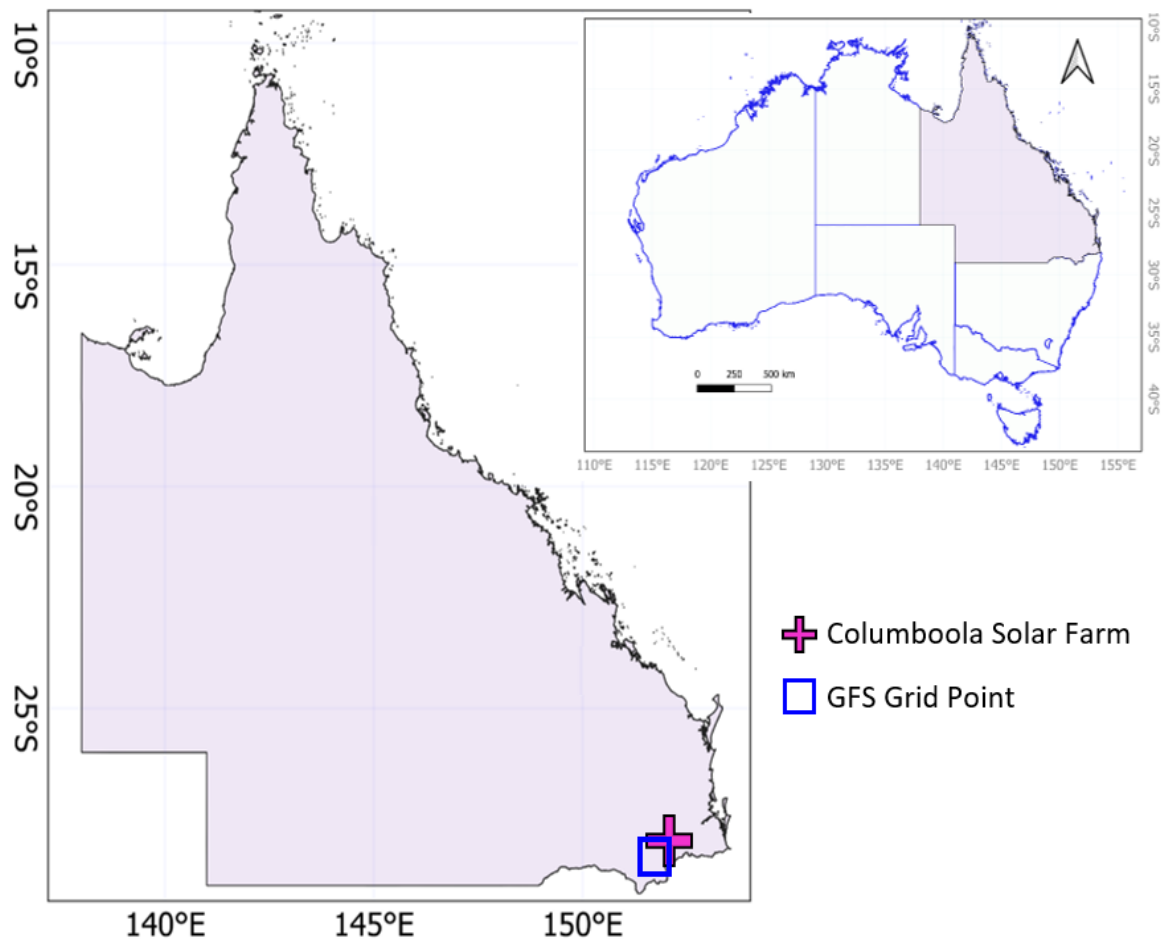


Fig. 1. Geographic location of our study site: *Columboola solar energy farm in Queensland Australia* where the proposed kernel ridge regression (KRR)-based machine learning model (ML) model for bias correction of total cloud cover (TCDC) was developed utilizing Global Forecast System (GFS) analysis (*i.e.*, proxy observed) and forecasted variables.

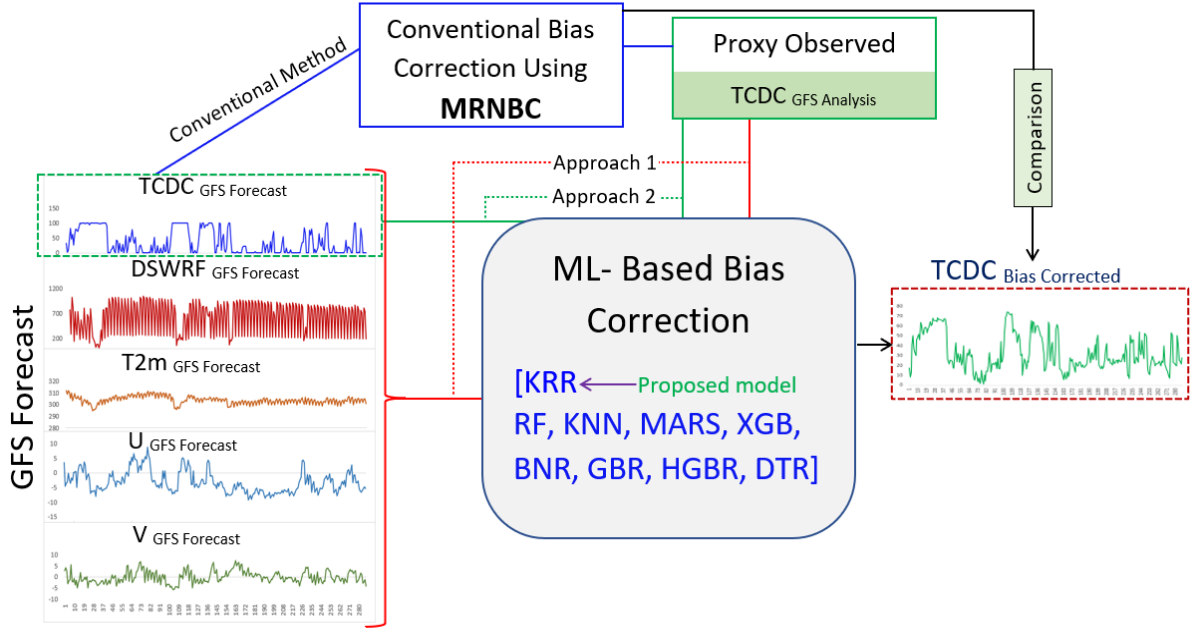


Fig. 2. Schematic of the proposed KRR-based bias correction method that is benchmarked with the conventional (*i.e.*, multivariate recursive nesting bias correction, MRNBC) and nine ML (*i.e.*, Bayesian ridge regression (BNR), Decision Tree (DTR), Gradient Boosting Regressor (GBR), Hist Gradient Boosting Regressor (HGBR), k - nearest regression (KNN), multivariate adaptive regression splines (MARS), extreme gradient boosting (XGB), and random forest (RF) methods adopted to correct the bias in total cloud cover (TCDC).

Interpretive Statement: The proposed KRR bias correction method uses: (i) **Approach 1** taking in five GFS inputs: *i.e.*, $TCDC_{GFS-Forecast}$, downward short-wave radiation flux $DSWRF_{GFS-Forecast}$, 2-meter temperature ($T2m_{GFS-forecast}$), zonal $U_{GFS-Forecast}$ and meridional $V_{GFS-Forecast}$ against the total cloud cover $TCDC_{GFS-Analysis}$ (reference or proxy observed) target, and (ii) **Approach 2** taking in $TCDC_{GFS-Forecast}$ as an input with $TCDC_{GFS-Analysis}$ target for which bias is corrected.

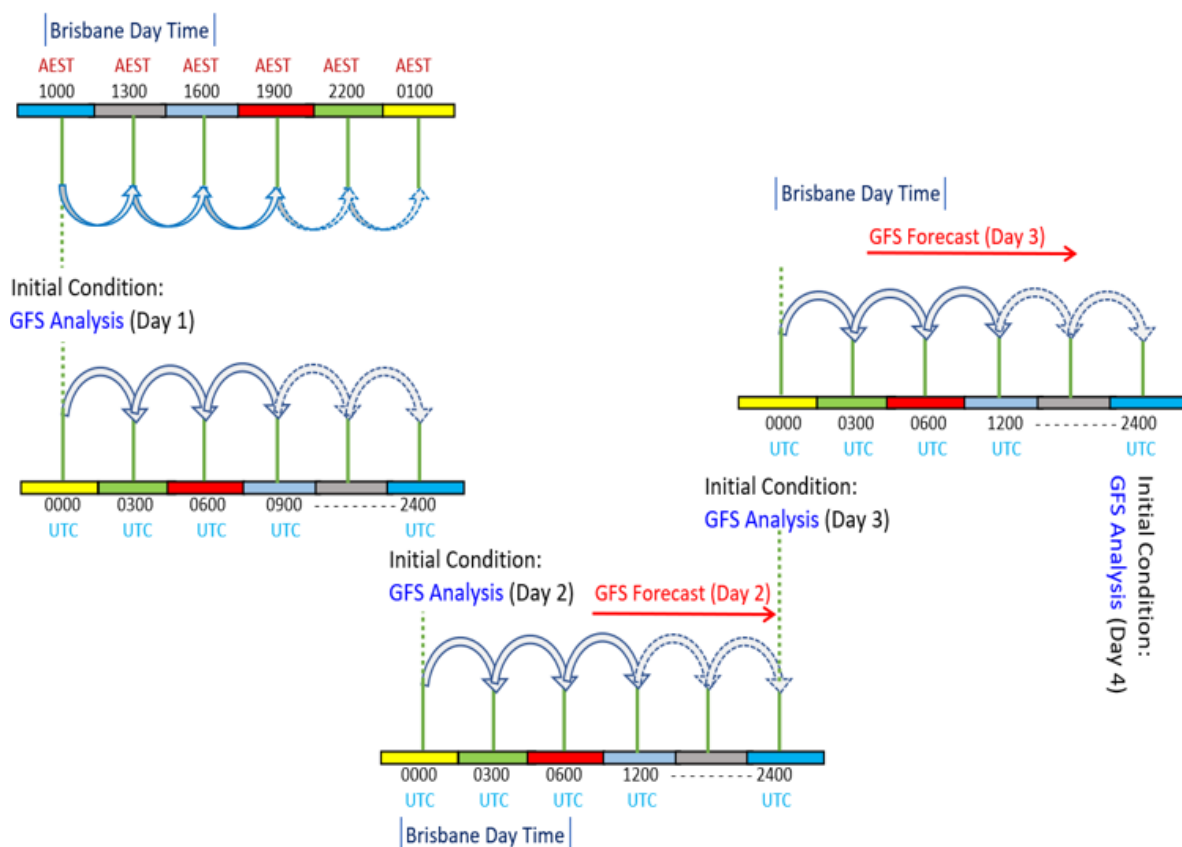


Fig 3. Schematic illustration of the 3-h GFS forecasts initialized at 0000 UTC compared with Australian Eastern Standard Time (AEST) used to develop KRR bias correction method.

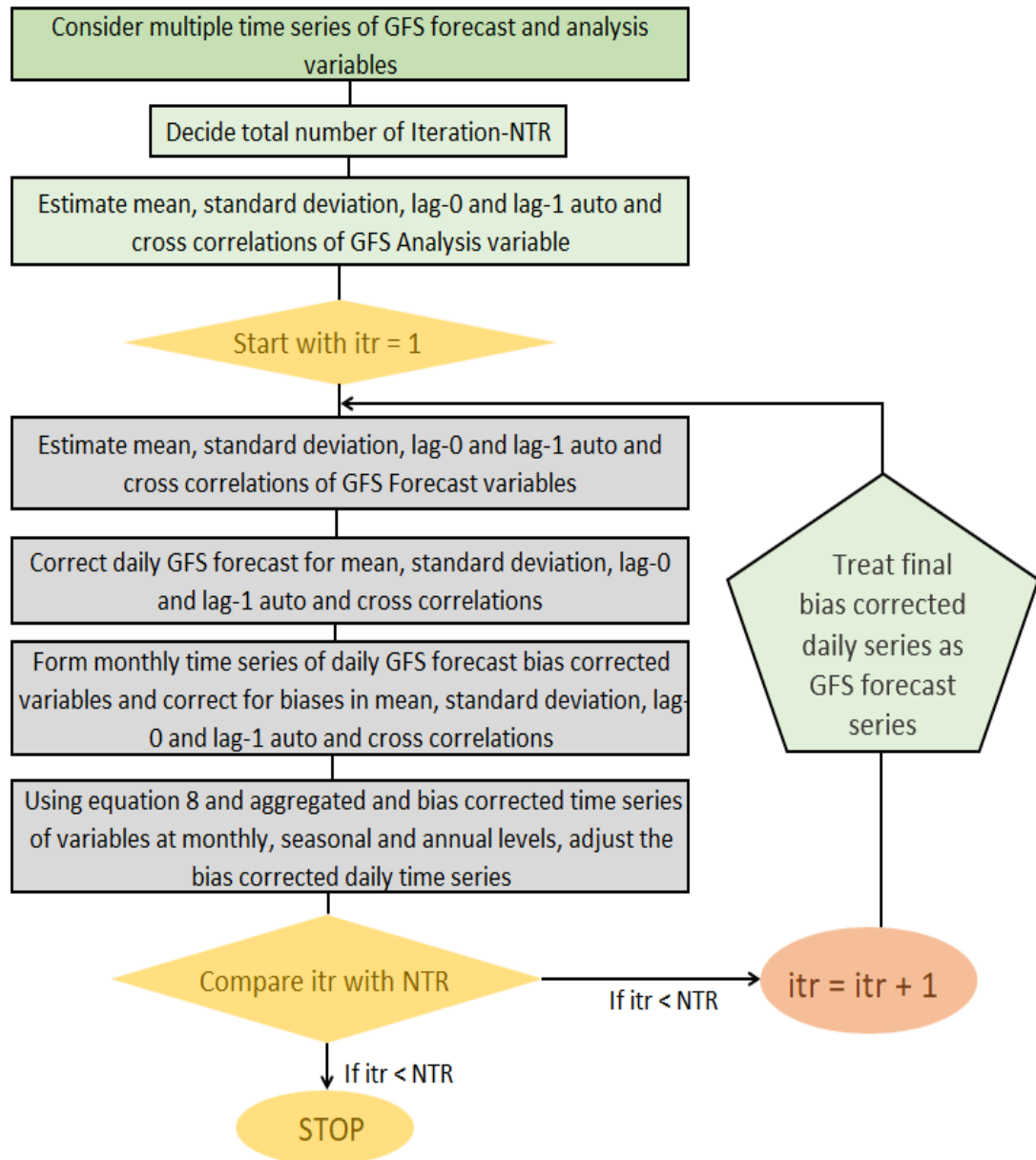


Fig. 4. Schematic of the traditional method, *i.e.*, multivariate recursive nested bias correction (MRNBC) presented in this study as a comparison method against the proposed KRR bias correction method used to correct bias in total cloud cover (TCDC).

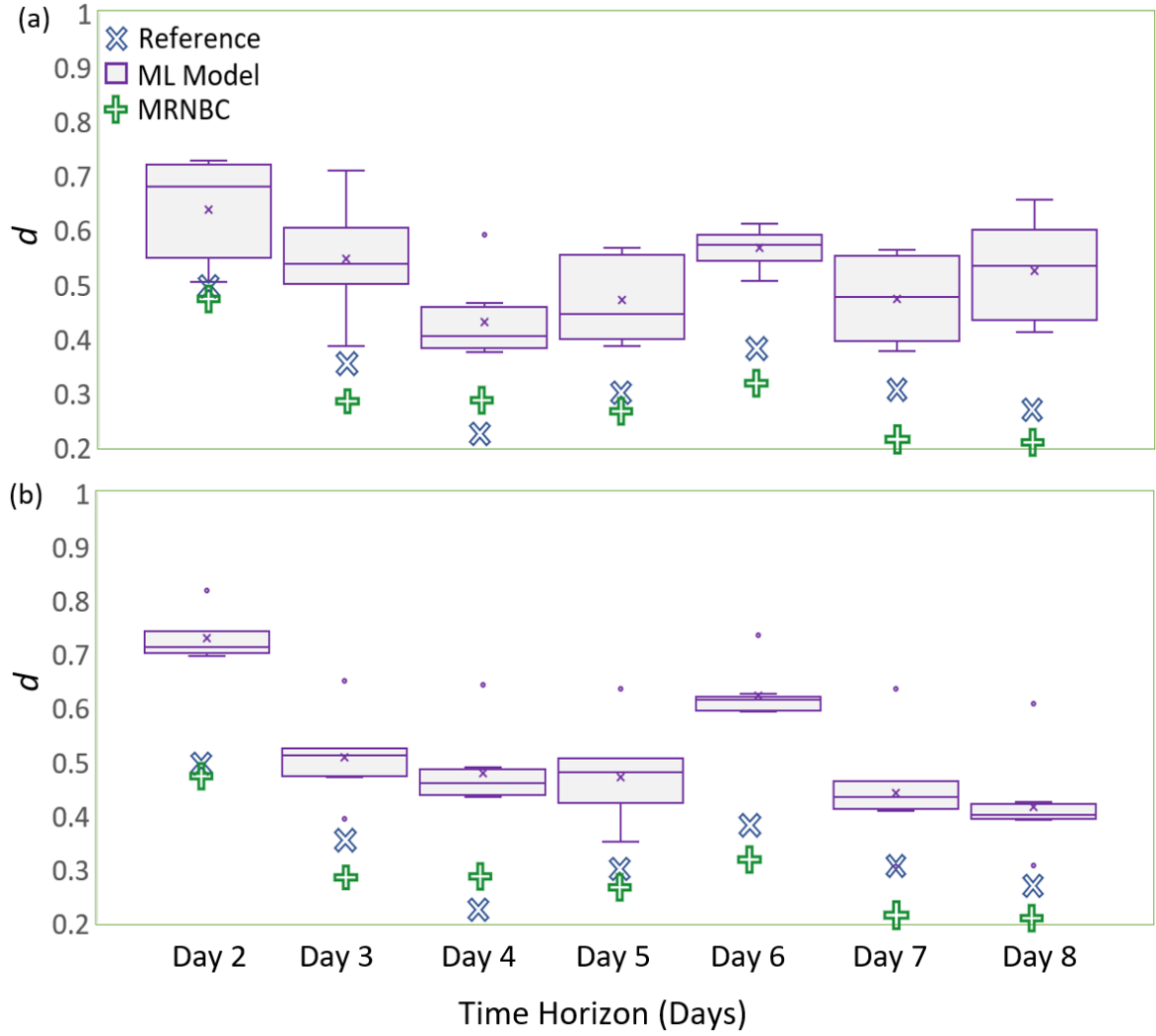


Fig. 5. Box plots of Willmott's Index of Agreement (d) calculated for all nine ML-bias corrections models (*i.e.*, KRR, BNR, DTR, GBR, HGBR, KNN, MARS, RF, XGB) pooled together including conventional bias correction (*i.e.*, MRNBC) and their respective reference values (d calculated between $\text{TCDC}_{\text{GFS-Forecast}}$ and $\text{TCDC}_{\text{GFS-Analysis}}$) for (a) Approach 1, and (b) Approach 2. [For more details on each approach, see Fig. 2]

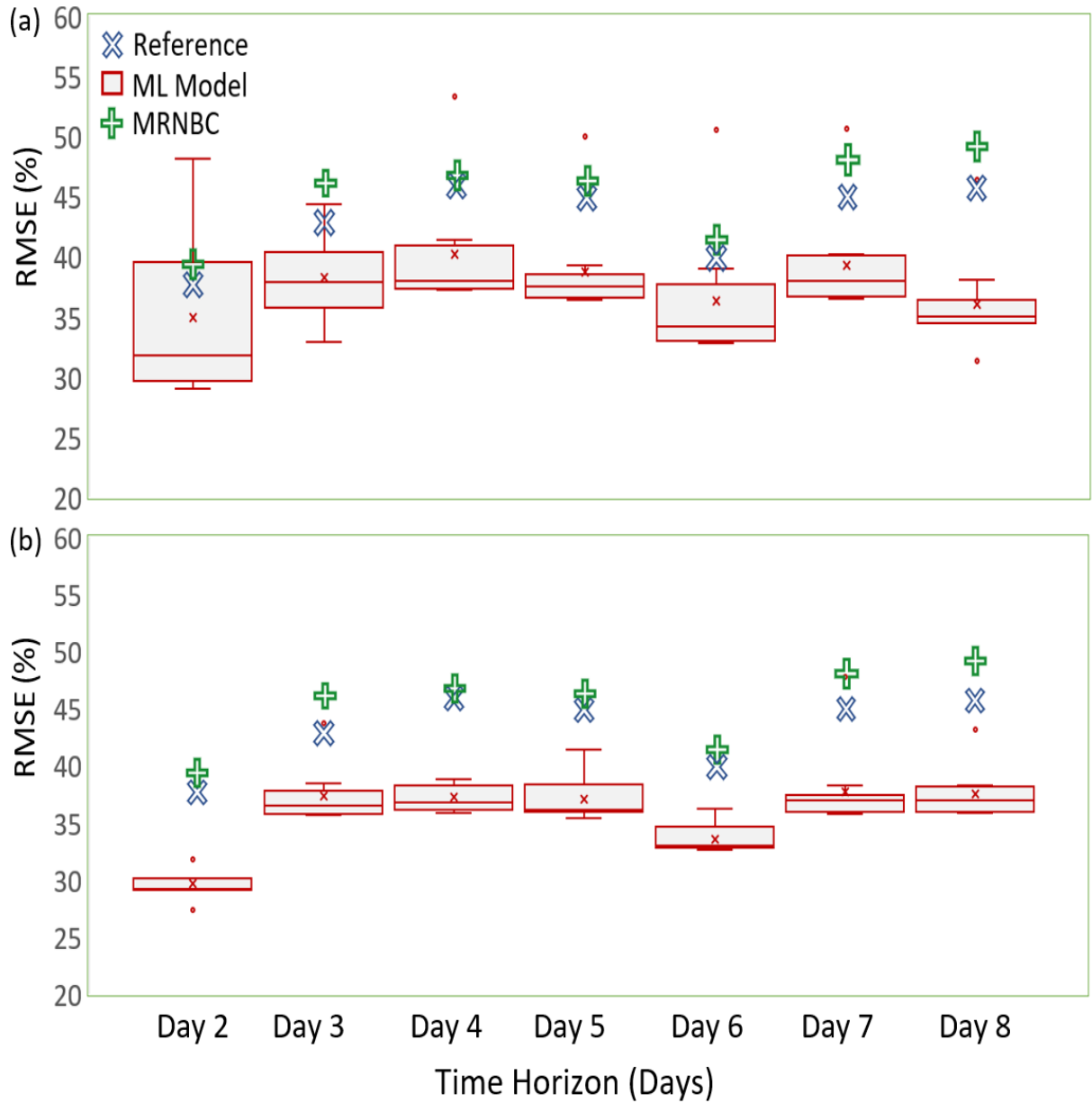


Fig. 6 Box plots of bias-corrected root mean square error (RMSE) calculated between data for all the nine ML-based bias correction methods pooled together (i.e., KRR, BNR, DTR, GBR, HGBR, KNN, MARS, RF, XGB), conventional bias correction method (i.e., MRNBC) and along with their respective reference values ($RMSE$ calculated between $TCDC_{GFS-Forecast}$ and $TCDC_{GFS-Analysis}$). (a) Approach 1 and (b) Approach 2. [For more details on each approach, see Fig. 2].

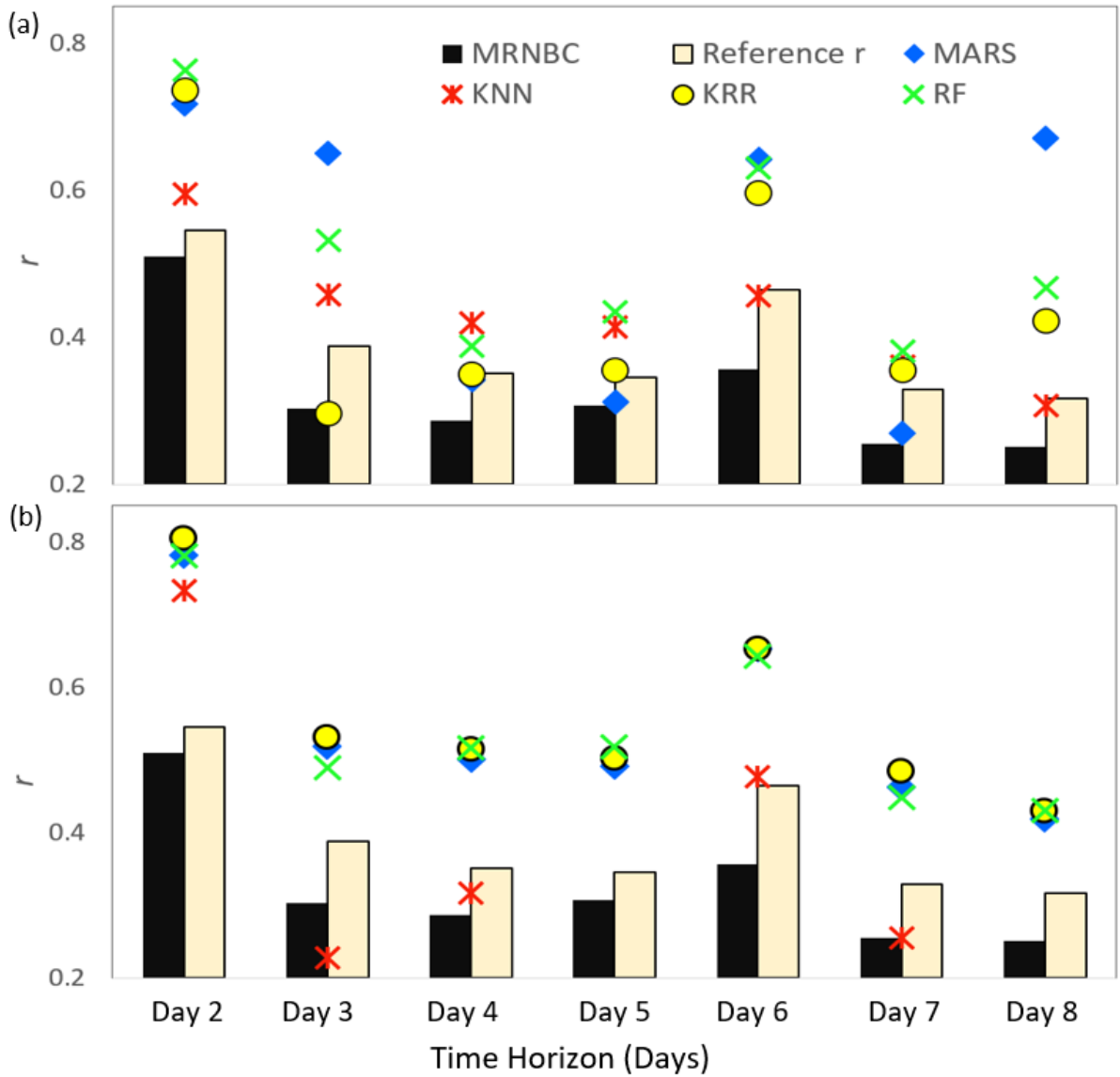


Fig. 7 Comparative analysis of *four* selected ML-based bias correction methods (*i.e.*, KRR, MARS, KNN, RF) by means of correlation coefficient (r) between the **corrected** TCDC_{GFS-Forecasts} and the reference TCDC_{GFS-Analysis}. Included is a respective reference r -value computed using ‘non-corrected’ TCDC_{GFS-forecasts} and bias-corrected TCDC_{GFS-Forecasts} but using a traditional method (*i.e.*, MRNBC). (a) Approach 1, and (b) Approach 2. [For more details on each approach, see Fig. 2].

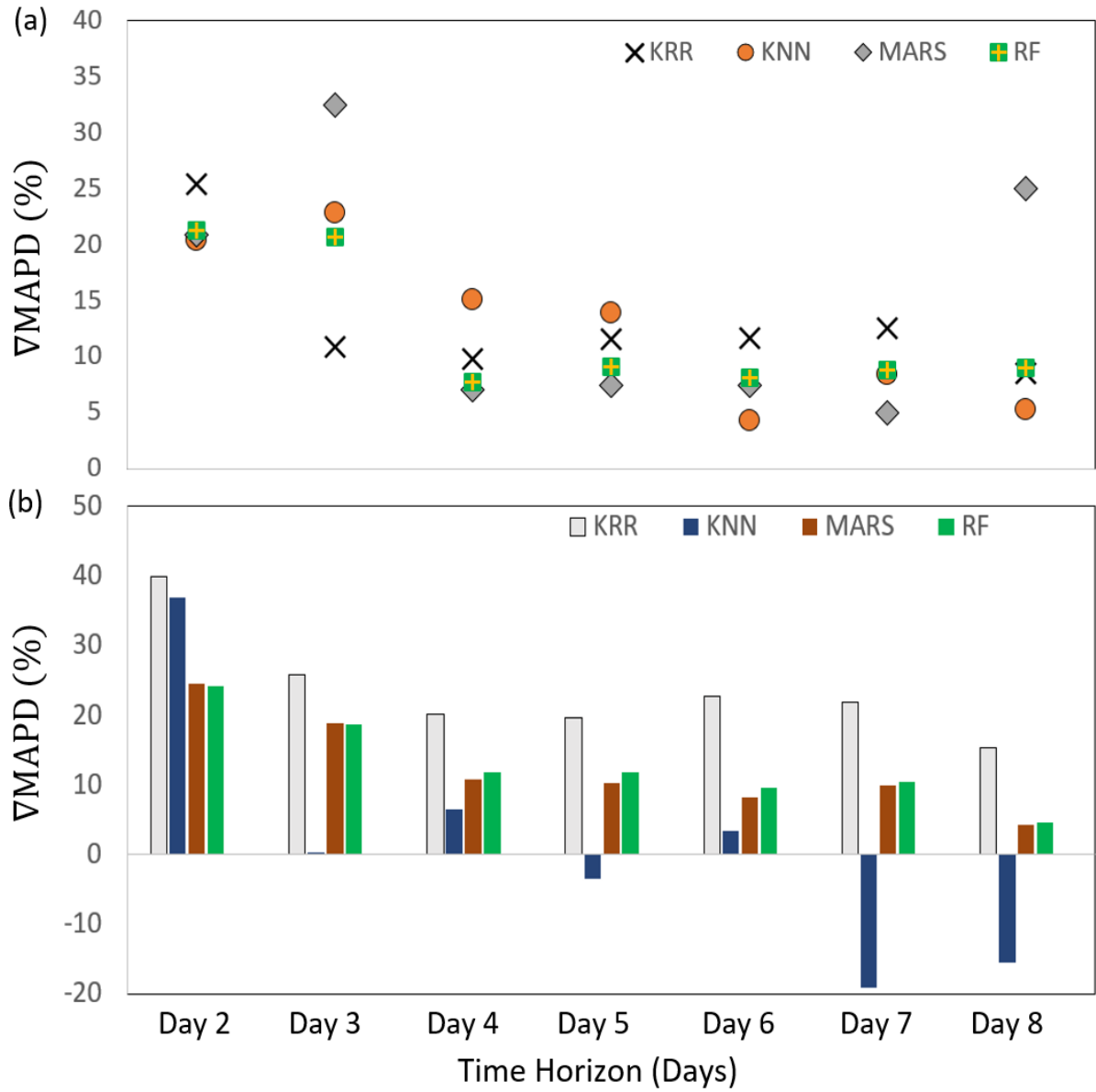


Fig. 8 Change (∇) in mean absolute percentage error, $MAPD$ (%) generated by proposed KRR bias correction method with respect to a reference value of $MAPD$ deducted from $TCDC_{GFS-Forecast}$ and $TCDC_{GFS-Analysis}$. (a) Approach 1, and (b) Approach 2. [For more details on each approach, see Fig. 2].

Interpretive statement: a positive change is used to show the objective model (*i.e.*, KRR) has outperformed the other benchmark models.

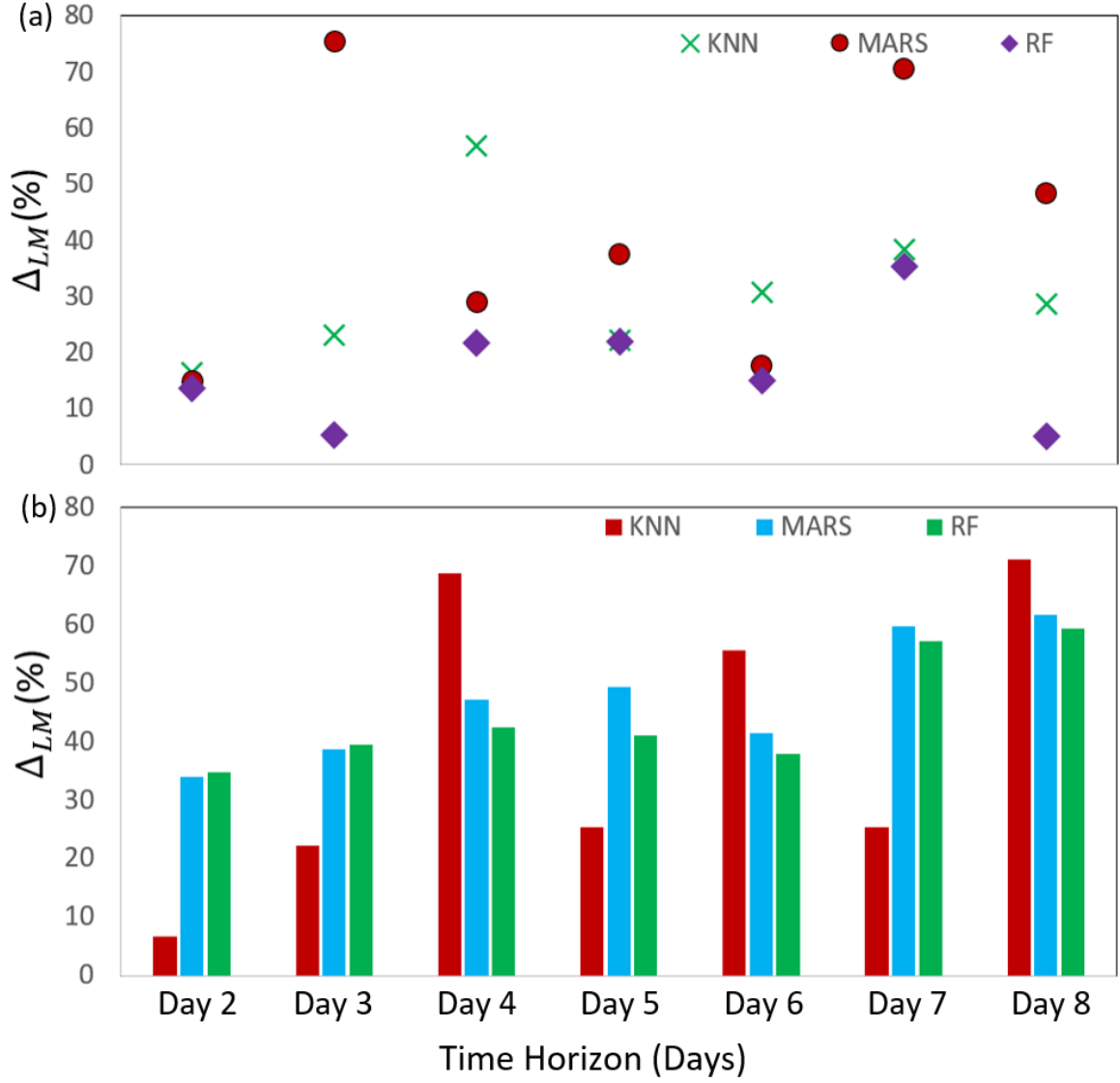


Fig. 9 The percentage change in Legates & McCabe's Index (LM) was deduced by comparing the LM values obtained using the proposed KRR-bias correction model in respect to the LM values generated by KNN, MARS and RF Models. (a) Approach-1, (b) Approach-2.

Note that: $\left(\Delta_{LM}(\%) = \left(LM_{KRR} - LM_{COM} / LM_{KRR}\right) \times 100\right)$

Note: LM_{COM} represents the LM value of the benchmark (KNN, MARS or RF) model.
 [For more details on each approach, see Fig. 2].

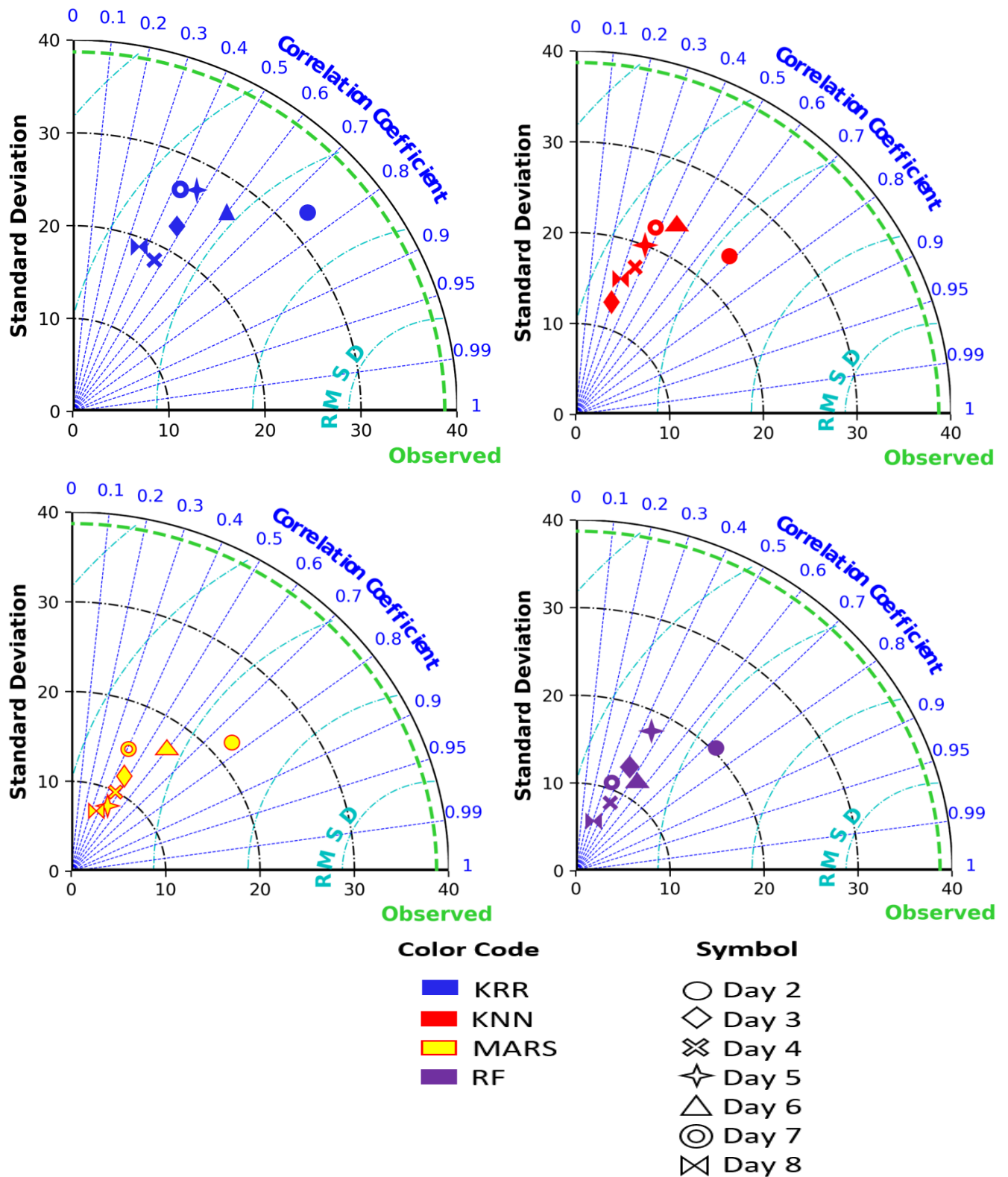


Fig. 10 Taylor diagram showing the correlation coefficient, standard deviation, and root mean square centered difference (RMSD). (a) The objective model (KRR) compared with (b) KNN, (c) MARS, and (d) RF) for the most accurate approach (i.e., Approach-2). [For more details on each approach, see Fig. 2].

List of Tables

Table 1 List of Global Forecast System (GFS)-forecast variables (*i.e.*, 2-metre temperature, 10-metre wind speed, total cloud cover, and downward short-wave radiation flux) used as KRR model inputs, and GFS analysis variable (*i.e.*, total cloud cover used as proxy observed) in the proposed KRR model used in bias correction problem.

Variable Short Name	Variable Description	Level	Units
KRR Model Inputs: GFS Forecast (Inputs)			
T2m _{GFS-Forecast}	2-metre temperature	Height Above Ground	K
U _{GFS-Forecast}	10-metre <i>U</i> wind component	Height Above Ground	m s ⁻¹
V _{GFS-Forecast}	10-metre <i>V</i> wind component	Height Above Ground	m s ⁻¹
TCDC _{GFS-Forecast}	Total Cloud Cover	Atmosphere	%
DSWRF _{GFS-Forecast}	Downward short-wave radiation flux	Surface	W m ⁻²
KRR Model Target: GFS Analysis (proxy observed)			
TCDC _{GFS-Analysis}	Total Cloud Cover	Atmosphere	%

Table 2

Descriptive statistics of GFS forecast and GFS analysis data used to develop the proposed KRR model. Data were acquired from GFS model over January 1, 2019 and April 30, 2020 used for training 70% and testing (30%) where the remaining 15% of training set is used for model validation.

Variable	Forecast	Max	Min	Mean	Skewness	Kurtosis
DSWRF _{GFS-Forecast}	Day 2	1100	0.00	601.07	-0.22	-1.38
	Day 3	1100	0.00	605.30	-0.23	-1.46
	Day 4	1100	0.00	595.55	-0.20	-1.47
	Day 5	1100	0.00	595.71	-0.20	-1.46
	Day 6	1100	0.00	599.78	-0.20	-1.39
	Day 7	1090	0.00	604.91	-0.24	-1.44
	Day 8	1100	0.00	605.01	-0.27	-1.42
TCDC _{GFS-Forecast}	Day 2	100	0.00	27.82	1.01	-0.56
	Day 3	100	0.00	29.38	0.91	-0.74
	Day 4	100	0.00	32.80	0.73	-1.04
	Day 5	100	0.00	32.95	0.73	-1.05
	Day 6	100	0.00	32.62	0.70	-1.12
	Day 7	100	0.00	31.88	0.77	-0.96
	Day 8	100	0.00	33.87	0.66	-1.11
T2m _{GFS-Forecast}	Day 2	314.55	285.38	301.64	-0.31	-0.62
	Day 3	314.76	285.36	301.57	-0.35	-0.59
	Day 4	313.59	285.24	301.49	-0.33	-0.67
	Day 5	314.74	284.35	301.45	-0.34	-0.61
	Day 6	314.65	284.76	301.53	-0.33	-0.54
	Day 7	315.22	285.20	301.45	-0.34	-0.55
	Day 8	313.45	285.54	301.70	-0.45	-0.42
U _{GFS-Forecast}	Day 2	10.49	-12.23	-4.25	0.99	0.94
	Day 3	7.38	-13.03	-3.50	0.49	-0.37
	Day 4	8.56	-11.41	-4.37	1.08	1.09
	Day 5	8.80	-12.24	-4.37	1.02	0.95
	Day 6	8.83	-10.67	-4.46	1.13	1.25
	Day 7	10.93	-11.93	-4.52	1.19	1.74
	Day 8	8.85	-13.19	-4.05	0.66	0.01
V _{GFS-Forecast}	Day 2	10.29	-7.74	0.14	0.22	-0.08
	Day 3	10.06	-9.55	-0.70	-0.03	-0.34
	Day 4	8.53	-7.08	0.09	0.25	-0.10
	Day 5	8.65	-7.22	0.12	0.31	-0.03
	Day 6	9.57	-6.64	0.03	0.30	-0.10
	Day 7	8.58	-10.66	-0.07	0.22	0.10
	Day 8	13.70	-7.37	-0.22	0.21	0.35
TCDC _{GFS-Analysis}	Day 2	100	0.00	31.70	0.78	-1.01
	Day 3	100	-5.83	31.82	0.78	-1.02
	Day 4	100	-5.83	31.89	0.77	-1.03
	Day 5	100	-5.83	31.95	0.77	-1.03
	Day 6	100	-5.83	31.95	0.77	-1.03
	Day 7	100	-5.83	31.92	0.77	-1.03
	Day 8	100	-5.83	32.02	0.76	-1.04

Table 3 Mean Absolute Error (MAE, %) between ‘proxy observed’ (TCDC_{GFS-Analysis}) and ML-bias corrected TCDC_{BC} using our proposed KRR model. Our conventional bias correction is a multivariate recursive nesting bias correction (MRNBC) method, whereas benchmark methods include BNR, DTR, GBR, HGBR, KNN, MARS, MLR, and RF model. In **Approach 1**, we used T2m_{GFS-Forecast}, V_{GFS-Forecast}, U_{GFS-Forecast}, TCDC_{GFS-Forecast}, and DSWRF_{GFS-Forecast}. In contrast, in **Approach 2**, we used TCDC_{GFS-Forecast} as a predictor (or input) variable against TCDC_{GFS-Analysis} as a target variable. *The reference MAE is computed between TCDC_{GFS-Forecast} and TCDC_{GFS-Analysis} data to provide additional benchmarks for the proposed KRR bias correction method.*

Model and Method		Inter-daily Forecast Horizon						
		Day 2	Day 3	Day 4	Day 5	Day 6	Day 7	Day 8
Based on TCDC _{GFS-Forecast} and TCDC _{GFS-Analysis} datasets	Reference	23.45	29.36	32.93	31.49	27.59	31.68	32.36
Conventional Bias Correction	MRNBC	25.90	32.05	32.65	32.76	30.28	33.57	34.50
Approach 1								
Objective Model	KRR	25.07	34.56	32.23	31.33	27.68	30.76	30.26
Benchmark Machine Learning Models	BNR	25.35	31.90	32.93	32.63	29.08	32.41	31.31
	DTR	35.65	30.47	41.35	37.00	38.24	37.98	34.46
	GBR	32.52	31.68	34.32	32.38	29.85	31.73	28.77
	HGBR	32.45	32.39	34.15	30.95	30.73	33.18	28.77
	KNN	26.76	29.90	30.32	30.48	29.98	32.20	31.31
	MARS	26.60	26.18	33.21	32.77	28.99	33.40	24.80
	RF	25.19	32.14	32.84	32.52	28.94	32.27	31.16
	XGB	26.47	30.74	32.96	32.17	28.80	32.08	30.08
Approach 2								
Objective Model	KRR	20.20	28.75	28.52	28.44	24.20	27.47	27.99
Benchmark Machine Learning Models	BNR	25.32	31.63	31.89	31.78	28.77	31.57	31.69
	DTR	26.75	32.22	33.19	31.82	29.23	31.55	32.74
	GBR	25.81	31.73	32.36	31.27	28.52	31.36	31.82
	HGBR	25.91	31.70	32.24	31.55	28.37	31.46	32.19
	KNN	21.22	38.64	33.39	36.67	30.29	41.85	38.18
	MARS	25.36	31.46	31.85	31.75	28.74	31.67	31.66
	RF	25.28	31.60	31.85	31.75	28.74	31.54	31.66
	XGB	25.48	31.50	31.52	31.20	28.36	31.49	31.52

Table 4

The optimal hyperparameter of the proposed KRR model, including that of the other benchmark models methods include machine learning (i.e., BNR, DTR, GBR, HGBR, KNN, MARS, MLR, and RF)

Model	Name	Hyper-parameters	Acronym	Optimum
Objective Model	KRR	Regularization strength	<i>alpha</i>	<i>1.5</i>
		Kernel mapping	<i>kernel</i>	<i>linear</i>
		Gamma parameter	<i>gamma</i>	<i>None</i>
		Degree of the polynomial kernel	<i>degree</i>	<i>3</i>
		Zero coefficient for polynomial and sigmoid	<i>coef0</i>	<i>1.2</i>
Benchmark Machine Learning Models	BNR	Maximum number of iterations	<i>n_iter</i>	<i>200</i>
		Stop the algorithm if w has converged	<i>tol</i>	<i>0.0001</i>
		Shape parameter for Gamma distribution over	<i>alpha_1</i>	<i>1e-05</i>
		Inverse scale parameter over alpha	<i>alpha_2</i>	<i>1e-05</i>
		Shape parameter for Gamma distribution over lambda	<i>lambda_1</i>	<i>1e-06</i>
		Inverse scale parameter for Gamma distribution over lambda	<i>lambda_2</i>	<i>1e-04</i>
		The initial value for alpha	<i>alpha_init</i>	<i>None</i>
	DTR	Maximum depth of the tree	<i>max_depth</i>	<i>None</i>
		Minimum number of samples for an internal node	<i>min_sample_split</i>	<i>2</i>
		Number of features for the best split	<i>max_features</i>	<i>Auto</i>
	GBR	Number of boosting stages	<i>n_estimators</i>	<i>102</i>
		Minimum number of samples for an internal node	<i>min_sample_split</i>	<i>2</i>
		Learning rate	<i>learning_rate</i>	<i>0.1</i>
		Maximum depth of individual regression estimators' estimators	<i>max_depth</i>	<i>3</i>
		Number of features to consider for the best split	<i>max_feature</i>	<i>None</i>
	HGBR	Learning rate	<i>learning_rate</i>	<i>0.1</i>
		Maximum number of iterations	<i>max_iter</i>	<i>120</i>
		maximum number of leaves for each tree	<i>max_leaf_nodes</i>	<i>31</i>
		Maximum number of bins	<i>max_bins</i>	<i>260</i>
	KNN	Number of neighbours	<i>n_neighbors</i>	<i>5</i>
		Weights	<i>Weights</i>	<i>uniform</i>
		The algorithm used to compute the nearest	<i>algorithm</i>	<i>auto</i>
		Leaf-size passed	<i>leaf_size</i>	<i>30</i>
		Power parameter for the Minkowski metric	<i>p</i>	<i>2</i>
		The distance metric to use for the tree.	<i>metric</i>	<i>minkowski</i>
		Additional keyword arguments for the metric	<i>metric_params</i>	<i>none</i>
		The number of parallel jobs	<i>n_jobs</i>	<i>int</i>
	MARS	maximum degree of terms	<i>max_degree</i>	<i>1</i>
		Smoothing parameter used to calculate GCV	<i>penalty</i>	<i>3.0</i>
	RF	Number of trees in the forest	<i>n_estimators</i>	<i>120</i>
		Maximum depth of the tree	<i>max_depth</i>	<i>2</i>
		Minimum number of samples for an internal node	<i>min_sample_split</i>	<i>2</i>
		Number of features for the best split	<i>max_features</i>	<i>auto</i>

APPENDIX B: WHEAT YIELD PREDICTION USING SATELLITE-DERIVED INFORMATION

B1.1 Foreword

This Chapter is an exact copy of the published manuscript to the *Remote Sensing* Manuscript Number: APEN-S-21-14660 (Scopus Impact Factor 4.85). The title of the manuscript is:

“Kernel Ridge Regression hybrid method for wheat yield prediction using satellite-derived predictors”

In this Chapter, Wheat, which controls the Australian grain market and accounts for 10-15% of the world's annual 100 million tonnes of wheat trade, dominates the Australian grain market, is predicted using the kernel ridge regression (KRR) method in conjunction with complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and the grey wolf optimization (GWO). Remote satellite-based information is used in this paper to estimate yield in a wheat-growing region in South Australia. The GWO-CEEMDAN-KRR hybrid model outperforms all benchmark models and feature selection (ant colony, atom search, and particle swarm optimization). The GWO-CEEMDAN-KRR model, with this improved methodology, may be used in agricultural yield simulations that require remote sensing data to establish relationships between crop health, yield, and other productivity features to support precision agriculture, such as crop rotation.

B1.2 Research Highlights

- a hybrid kernel ridge regression (KRR) method that is developed to predict the wheat yield of South Australia.
- KRR is coupled with CEEMDAN and GWO, referred to as GWO-CEEMDAN-KRR.
- A pool of 23 different satellite-based predictors is used for wheat yield prediction.
- The predicted results show that prediction error can be reduced by ~20% by employing the proposed GWO-CEEMDAN-KRR model.

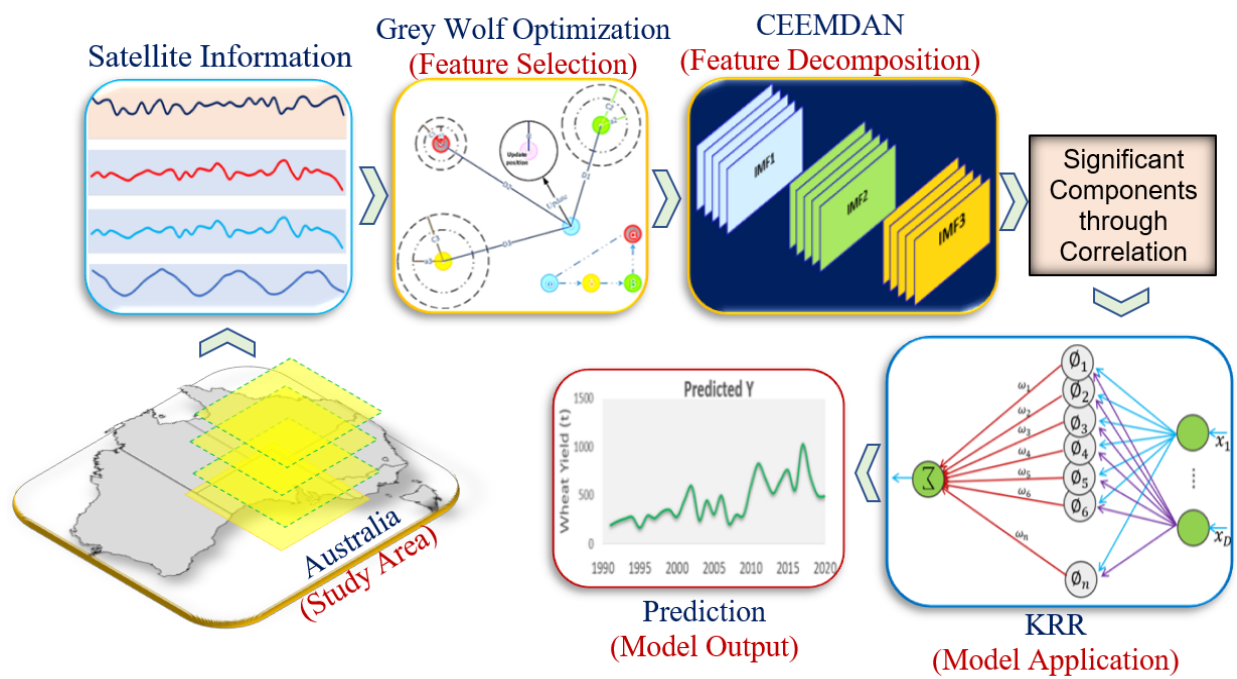


Fig B1 Graphical Abstract of Article 7

B1.3 Article 7

Article

Kernel Ridge Regression Hybrid Method for Wheat Yield Prediction with Satellite-Derived Predictors

A. A. Masrur Ahmed ¹, Ekta Sharma ¹, S. Janifer Jabin Jui ², Ravinesh C. Deo ^{1,*}, Thong Nguyen-Huy ^{3,4,5} and Mumtaz Ali ⁶

¹ School of Mathematics, Physics and Computing, University of Southern Queensland, Springfield, QLD 4300, Australia; abulabramasrur.ahmed@usq.edu.au (A.A.M.A.); ekta.sharma@usq.edu.au (E.S.)

² Global Project Management (Advanced), Torrens University, Sydney, NSW 2000, Australia; s.jui@business.torrens.edu.au

³ SQNNNSW Drought Resilience Adoption and Innovation Hub, University of Southern Queensland, Toowoomba, QLD 4350, Australia; thong.nguyen-huy@usq.edu.au

⁴ Centre for Applied Climate Sciences, University of Southern Queensland, Toowoomba, QLD 4350, Australia

⁵ Ho Chi Minh City Space Technology Application Center, Vietnam National Space Center, VAST, Ho Chi Minh 700000, Vietnam

⁶ Deakin-SWU Joint Research Centre on Big Data, School of Information Technology, Deakin University, Deakin, VIC 3125, Australia; mumtaz.ali@deakin.edu.au

* Correspondence: ravinesh.deo@usq.edu.au; Tel.: +61-7-34704430

Abstract: Wheat dominates the Australian grain production market and accounts for 10–15% of the world's 100 million tonnes annual global wheat trade. Accurate wheat yield prediction is critical to satisfying local consumption and increasing exports regionally and globally to meet human food security. This paper incorporates remote satellite-based information in a wheat-growing region in South Australia to estimate the yield by integrating the kernel ridge regression (KRR) method coupled with complete ensemble empirical mode decomposition with adaptive noise (CEEMDAN) and the grey wolf optimisation (GWO). The hybrid model, 'GWO-CEEMDAN-KRR,' employing an initial pool of 23 different satellite-based predictors, is seen to outperform all the benchmark models and all the feature selection (ant colony, atom search, and particle swarm optimisation) methods that are implemented using a set of carefully screened satellite variables and a feature decomposition or CEEMDAN approach. A suite of statistical metrics and infographics comparing the predicted and measured yield shows a model prediction error that can be reduced by ~20% by employing the proposed GWO-CEEMDAN-KRR model. With the metrics verifying the accuracy of simulations, we also show that it is possible to optimise the wheat yield to achieve agricultural profits by quantifying and including the effects of satellite variables on potential yield. With further improvements in the proposed methodology, the GWO-CEEMDAN-KRR model can be adopted in agricultural yield simulation that requires remote sensing data to establish the relationships between crop health, yield, and other productivity features to support precision agriculture.

Keywords: wheat yield; satellite data; machine learning; kernel ridge regression; South Australia

Citation: Ahmed, A.A.M.; Sharma, E.; Jui, S.J.J.; Deo, R.C.; Nguyen-Huy, T.; Ali, M. Kernel Ridge Regression Hybrid Method for Wheat Yield Prediction with Satellite-Derived Predictors. *Remote Sens.* **2022**, *14*, 1136. <https://doi.org/10.3390/rs14051136>

Academic Editors: Nathaniel K. Newlands and Jianxi Huang

Received: 30 December 2021

Accepted: 23 February 2022

Published: 25 February 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Agriculture and climate change are interrelated sciences [1], with adverse climate variability being a fundamental factor disrupting agricultural production. This may correlate to food availability, decreased food access, and even food quality [2]. Such an effect is likely to happen with subsequent changes in temperature, rainfall, and extreme climatic conditions such as heatwaves, diseases, pest invasions, and varying nutritional quality of some foods, to name a few [3]. Quantifying and modelling the impacts of these factors on crop yield is vital for improving the resilience of our agricultural system in a highly

variable environment [4]. The authors in [5] state that three variables, such as crop yield, cropping area, and crop frequency, are fundamental to the crop production equation. Modelling crop yield has often been estimated based on the sensitivity of agricultural outputs to climate variability under global warming scenarios. It has also been estimated that changes in frequency and/or cropping can cause roughly 70% of the change in agricultural output driven by climate variability [6].

Several types of research have assessed climate change impacts on crop yields at local and global scales. Some examples stated in [7–10] use either deterministic or artificial intelligence methods for modelling. Romeijn et al. [11] have evaluated deterministic and complex analytical hierarchy process methods for agricultural land suitability analysis in a changing climate. Aschonitis et al. [12] assessed the intrinsic vulnerability of agricultural land to water and nitrogen losses via a deterministic approach and regression analysis. Several studies, such as [13–15], have used deterministic or probabilistic approaches for modelling. However, these methods lack automation and can be time-consuming, complex, and resource-intensive [16,17]. Machine learning (ML) methods have gained significant attention from researchers keen to develop yield prediction models. One such study is the work of Kouadio et al. [18], which used soil fertility properties as fertiliser constituents (i.e., soil organic matter (SOM), available potassium, boron, sulfur, zinc, phosphorus, nitrogen, exchangeable calcium, magnesium, and pH) to predict Robusta coffee yield in Vietnam.

Wheat yield predictions based on multi-source data from climate, satellite, soil, and historical yield records have developed rapidly using linear regression [19,20], machine learning [21,22], and deep learning algorithms [23,24]. The research interest has focused on identifying the most important predictors and developing robust prediction models. Kolotii et al. [25] applied single-factor linear regression to forecast winter wheat crop yield in Ukraine using normalised difference vegetation index (NDVI), leaf area index (LAI), and a fraction of absorbed photosynthetically active radiation (fAPAR) derived from satellite data and crop growth model. The author indicated that the satellite-based biophysical parameter predictor, LAI, yielded the most accurate result at each scale. Cai et al. [26] combined climate and satellite data to achieve the best performance for wheat yield prediction in Australia. The findings also indicated that the yield prediction models based on machine learning methods outperformed the regression methods used by earlier researchers, such as [27–29]. Among satellite-based inputs, using the enhanced vegetation index (EVI) provided better performance in yield prediction than the solar-induced chlorophyll fluorescence (SIF). Kamir et al. [30] integrated the benefits of machine learning and regression methods, climate records, and satellite image time series to estimate wheat yields across the Australian wheat belt. The results show that the combination of support vector regression (SVR) and radial basis function is the best model while the additional information from climate (temperatures and rainfall) significantly improved yield predictions compared to the pure NDVI-based model. Moreover, the author suggested that the resulting yield estimates meet the accuracy requirements for mapping the yield gap and identifying yield gap hotspots that could be targeted for further work. Bali and Singla [31] demonstrated that deep learning-based Recurrent Neural Network (RNN)-long short-term memory (LSTM) outperformed machine learning models, Artificial Neural Network (ANN), Random Forest (RF), and Multivariate Linear Regression (MLR) model in predicting wheat yield in the northern region of India using climate variables. The results also show that machine and deep learning models outperformed the two linear regression methods in predicting wheat yield; however, the LSTM did not perform better than SVR. Overall, it is clear that studies focused on the importance of incorporating satellite data for modelling purposes to capture spatially relevant information for yield prediction, the performance of different predictors and models requires further investigation.

This paper contributes to the development of the robust method for predictor selection and accuracy of wheat yield prediction using large datasets derived from satellites. The study also aims to report on the modelling impacts of climate variability on

agricultural crop yields in South Australia using satellite-derived information. This is why this study is necessary and has several advantages [32], such as eliminating the provision to collect unobstructed spatial data physically without a piece of measuring equipment and considering satellite sensors that can passively record electromagnetic energy reflected from or emitted by the phenomena of interest [33]. In other words, using the satellite method to collect data means that passive remote sensing does not disturb the object or the area of interest and can help collect the data over relatively large spatial areas. The use of remote sensing methods in satellite datasets can also help to characterise the natural weather or climatic features without being affected by the physical objects on the ground surface. Using satellites to monitor the ground, the surface areas can be observed systematically, and the changes in soil or other properties affecting crop yield can also be monitored systematically and regularly over time. Remote sensing methods also enable us to obtain repetitive coverage, which becomes quite handy when collecting data on dynamic themes such as soil moisture, water, agricultural fields, etc. Australian farmers are vulnerable to climate variability and change [34]. As for South Australia as a specific choice of an area of interest, it is observed that the region has varying rainfall patterns, droughts, and higher temperatures that pose significant risks to the state's urban water supplies and agricultural areas [35,36]. This affects wheat production, a prime employment source in South Australia, and its export. Therefore, climate change variability in the region, especially during the austral winter, is a potential threat to production [35]. Lastly, this study maps ground conditions at small-to-medium scales, making the data acquisition methods cheaper and faster.

2. Materials and Methods

2.1. Theoretical Frameworks

This section summarises the proposed objective model (i.e., KRR) and related algorithms (i.e., CEEMDAN and GWO) used in this study. The use of hybrid models in the study can amplify the strengths of the individual techniques to provide a more robust approach to the modelling process and make the model more accurate and efficient [37–39]. This paper aims to use the predictive merits of the CEEMDAN (a data decomposition method) combined with the KRR algorithm to achieve what has never been done in crop yield modelling before, especially in South Australia. To improve the CEEMDAN-KRR and other comparative models by selecting the most relevant satellite variables, we optimise the overall predictive system using feature selections based on grey wolf optimisation (GWO), ant colony optimisation (ACO, [40]), atom search optimisation (ASO, [41]), and particle swarm optimisation (PSO) [42]. It is imperative to note that the CEEMDAN method is a variation of the Ensemble Empirical Mode Decomposition (EEMD) algorithm that provides a near-exact reconstruction of the original signal and a better spectral separation of the Intrinsic Mode Functions (IMFs) [43]. Several other comparison approaches include CEEMDAN-MLR or Multiple Linear Regression, CEEMDAN-RF or Random Forest, and CEEMDAN-SVR or Support Vector Regression, and their respective standalone counterparts such as KRR, MLR, RF, and SVR models are also used in this study. Technical details of multi-linear regression (MLR) [44], random forest (RF) [45], and support vector regression (SVR) [46,47], and the feature optimization methods ACO [40], ASO [41], and PSO [42] are explained elsewhere.

2.1.1. Kernel Ridge Regression (KRR)

Ridge Regression (RR) is a simple yet powerful non-linear regression for forecasting, especially when the kernel is introduced into RR (KRR) as it maps out the time-series non-linearly transformed [48] input data to high dimensional space from low dimension [49] and the kernel function is a feature map from d dimensional Hilbert Space $\mathcal{H}_k, \Psi : \mathcal{X} \rightarrow \mathcal{H}_k$ such that $k(x_i, x_j) = \{\Psi(x_i), \Psi(x_j)\}_{\mathcal{H}_k}$. In this study, we follow Li et al. [49] to

implement KRR. With kernel functions and n data samples $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n) \in X * Y$ (y_i is the target value of corresponding $x_i, i = 1, 2, \dots, n$), the kernel matrix equation is:

$$K = \begin{bmatrix} k(x_1, x_1) & k(x_1, x_2) & \dots & k(x_1, x_n) \\ k(x_2, x_1) & k(x_2, x_2) & \dots & k(x_2, x_n) \\ \dots & \dots & \dots & \dots \\ k(x_n, x_1) & k(x_n, x_2) & \dots & k(x_n, x_n) \end{bmatrix} \quad (1)$$

The KRR problem can be formulated as

$$\min_w \|Y - Kw\|^2 + \lambda \|w\|^2 \quad (2)$$

Here Y is the target vector of all n data samples, w is the unknown vector, I_n is an $n * n$ identity matrix and regularisation item $\lambda \geq 0$ to avoid a large range of w .

$$w = (K + \lambda I_n)^{-1} Y \quad (3)$$

2.1.2. Complete Ensemble Empirical Mode Decomposition with Adaptive Noise

Complete Ensemble Empirical Mode Decomposition with Adaptive Noise (CEEMDAN) is an improved version of ensemble mode decomposition (EMD) and empirical ensemble mode decomposition (EEMD). The EMD is an adaptive time-space analysis used to process non-linear and non-stationary time series of data. Due to the nonlinearity of the data, EMD uses the divide and conquer strategy to decompose and ensemble complex signals into simple components and extract those components as intrinsic mode functions (IMFs) and residues [49]. To avoid EMD's mode mixing problem, EEMD decomposes signals by adding white Gaussian noise, but they cannot be offset after multiple averaging [50,51]. Moreover, this method's intricacy and computational complexity are significantly raised when white noise is expanded numerous times [52]. The CEEMDAN overcomes this problem with adaptive noise by reconstructing the original input/output variables. Compared to EEMD, the reconstruction of CEEMDAN is comprehensive and noise-free, and it requires fewer trials [53]. This study follows Torres et al. [54] and Ahmed et al. [55] to implement CEEMDAN using the following steps.

Step 1: Decompose by EMD P realisation $x[n] + \varepsilon_0 \omega^i[n]$ to receive the first modal component

$$\widehat{IMF}_1[n] = \frac{1}{P} \sum_{p=1}^P IMF_1^p[n] = \overline{IMF}_1[n] \quad (4)$$

Step 2: The first residual component is calculated by putting $k = 1$ in Equation (1),

$$Res_1[n] = \chi[n] - \widehat{IMF}_1[n] \quad (5)$$

Step 3: Putting $k = 2$, the second residual component is obtained as

$$\widehat{IMF}_2[n] = \frac{1}{P} \sum_{p=1}^P E_1(r_1[n] + \varepsilon_1 E_1(\omega^p[n])) \quad (6)$$

Step 4: Similarly calculating k th residue as

$$Res_k[n] = Res_{k-1}[n] - \widehat{IMF}_k[n] \quad (7)$$

Step 5: Decomposing the realizations $Res_k[n] + \varepsilon_1 E_1(\omega^p[n])$. Here, $k = 1, \dots, K$ until their first model of EMD reached and the $(k + 1)$ is

$$\widehat{IMF}_{(k+1)}[n] = \frac{1}{P} \sum_{p=1}^P E_1(r_k[n] + \varepsilon_k E_k(\omega^p[n])) \quad (8)$$

Step 6: Here, the k value is incremented and steps 4–6 are repeated, and the final residue is achieved

$$\text{RES}_k[n] = \chi[n] - \sum_{k=1}^K \text{IMF}_k \quad (9)$$

Here, k is the highest number of modes.

Therefore, the signal $\chi[n]$ can be expressed as

$$\chi[n] = \sum_{k=1}^K \text{IMF}_k + \text{RES}_k[n] \quad (10)$$

2.1.3. Grey Wolf Optimizer (GWO)

Grey wolf optimiser proposed by Mirjalili et al. (2014) depicts the interesting and systematic lifestyle of grey wolves belonging to the Canidae family. Grey wolves lie at the top of the food chain living in a pack of 5 to 12 with a social hierarchy naming alpha (α -leaders), beta (β -advisors of alpha and commands δ and ω), gamma (δ -commands ω), and omega (ω -follow every other wolf's command). During hunting, α , β , and δ work as guides and ω follow them. During encircling of prey for hunting, it is as described by Al-Tashi et al. [56]

$$\vec{X}(t+1) = \vec{X}_p(t) + \vec{A} \cdot \vec{D} \quad (11)$$

$$\vec{D} = \left| \vec{C} \cdot \vec{X}_p(t) - \vec{X}(t) \right| \quad (12)$$

where t indicates the current iteration, \vec{A} and \vec{C} are coefficient vectors, \vec{X}_p is the prey's positions vector, \vec{X} is the position of the wolves in d dimensional space, as d is the variable number. \vec{A} and \vec{C} can be calculated as the following:

$$\vec{A} = 2\vec{a} \cdot \vec{r}_1 - \vec{a} \quad (13)$$

$$\vec{C} = 2 \cdot \vec{r}_2 \quad (14)$$

where \vec{r}_1 and \vec{r}_2 are vectors randomly in $[0, 1]$ and \vec{a} is a set vector that linearly decreases from 2 to 0 over iterations. In the hunting process, α , β , and δ command and ω follow them modifying their positions as required by the pack until a suitable position or, in this case, a solution is achieved. The position selection can be calculated as

$$\vec{X}(t+1) = \frac{\vec{x}_1 + \vec{x}_2 + \vec{x}_3}{3} \quad (15)$$

where \vec{x}_1, \vec{x}_2 , and \vec{x}_3 can be defined as:

$$\vec{x}_1 = \vec{X}_\alpha - A_1 \cdot (\vec{D}_\alpha) \quad (16)$$

$$\vec{x}_2 = \vec{X}_\beta - A_2 \cdot (\vec{D}_\beta) \quad (17)$$

$$\vec{x}_3 = \vec{X}_\delta - A_3 \cdot (\vec{D}_\delta) \quad (18)$$

where \vec{x}_1, \vec{x}_2 , and \vec{x}_3 are the best solutions at iteration t , A_1, A_2 , and A_3 can be calculated using Equation (13) and $\vec{D}_\alpha, \vec{D}_\beta$ and \vec{D}_δ calculated from Equation (19) and \vec{C}_1, \vec{C}_2 and \vec{C}_3 from Equation (14)

$$\begin{aligned} \vec{D}_\alpha &= |\vec{C}_1 \cdot \vec{X}_\alpha - \vec{X}| \\ \vec{D}_\beta &= |\vec{C}_2 \cdot \vec{X}_\beta - \vec{X}| \\ \vec{D}_\delta &= |\vec{C}_3 \cdot \vec{X}_\delta - \vec{X}| \end{aligned} \quad (19)$$

In the main paper, to tune exploration and exploitation \vec{a} vector is suggested to decrease for each dimension linearly proportional to the number of iterations from 2 to 0. The equation is as follows, and ter is the optimisation total iterations number:

$$\vec{a} = 2 - t \cdot \frac{2}{\max_t ter} \quad (20)$$

Figure 1 illustrates the flowchart of the grey wolf optimisation algorithm. The figure shows that only one wolf can conduct a mating action in a wolf pack. It is not required for the alpha (α) wolf to be the strongest wolf in the pack, but the wolf must have the finest management skills. The beta (β) wolf possesses the group's second-best command. The wolf supports each other and serve as a liaison with all other wolves in the pack. The second is the delta (Δ) and omega (ω) wolves, respectively, maintaining the group's diminishing authority level. The wolf is the group's lowest level of the hierarchy, and it obeys the orders and instructions of a wolf. The GWO method uses four sorts of grey wolves for the simulation, representing the four fitness functions.

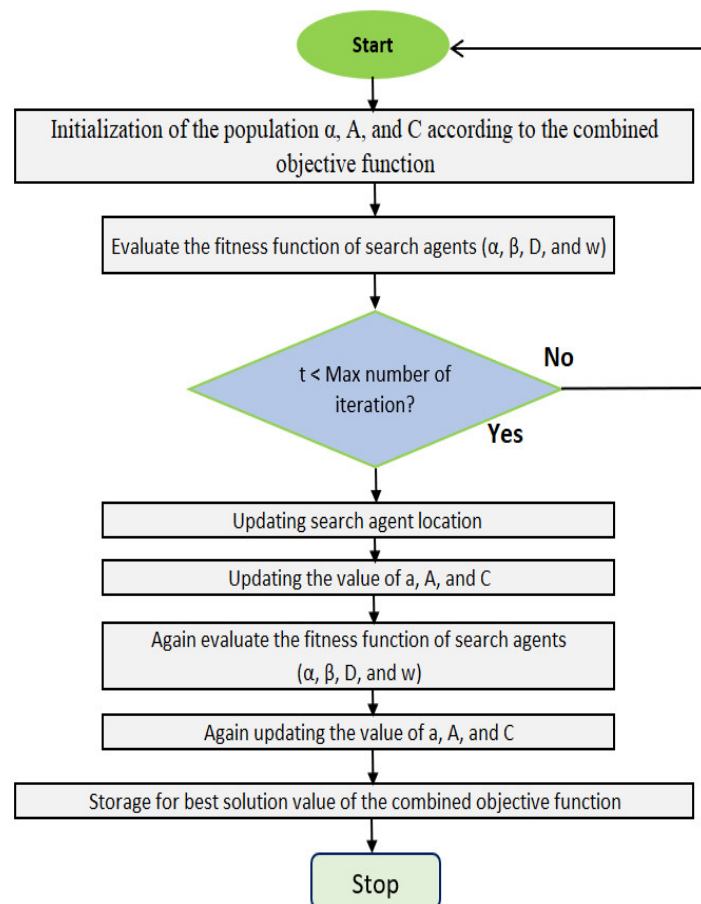


Figure 1. Flowchart of the grey wolf optimisation (GWO) algorithm.

2.1.4. Particle Swarm Optimiser (PSO)

The particle swarm optimiser algorithm is a population-based stochastic optimisation inspired by social and psychological considerations [57]. The PSO relates to swarm intelligence principles, which imitate the social behaviour of flocking birds or schooling fish. The algorithm has gained popularity due to its several favourable properties, including its basic structure, resilient mobility, and ease of implementation [58,59], which enable the training of various intelligent models. Each particle is considered a plausible solution in the search space of an optimisation problem. The control parameters determine the convergence of particle trajectories, keeping track of each particle's unique best fitness

value, locating the global best particle, and updating each particle's location and velocity. If convergence is not achieved, the iterative process is repeated until either the optimisation problem converges to an optimal solution, or the maximum number of iterations is reached. The number of particles is 10, with a maximum number of iterations of 150. The maximum and minimum bond inertia weights were 0.9 and 0.4, respectively.

2.1.5. Atom Search Optimiser (ASO)

Zhao et al. [60] introduced Atom Search Optimisation (ASO) as a new metaheuristic algorithm in 2019. The ASO simulates the fundamental concepts of molecular dynamics and atom movement principles, such as potential function characteristics, contact force, and geometric constraint force. In ASO, each atom keeps track of two vectors: position and velocity. When it comes to binary optimisation, the atoms only have to deal with two numbers ("1" or "0"). As a result, a means to leverage the atom's velocity to alter the position from "0" to "1" or vice versa should be discovered. Previous research has shown that the transfer function helps convert a continuous optimisation algorithm to a binary one [61]. During the initial iterations of ASO, each atom interacts with others via attraction or repulsion. Repulsion can help avoid over-concentration of atoms and premature algorithm convergence, improving exploration capability across the search space. As iterations progress, the repulsion becomes weaker, and the attraction becomes more robust, indicating that exploration diminishes, and exploitation grows. Finally, each atom interacts with other atoms by attraction, ensuring that the algorithm has a lot of power to use.

2.1.6. Ant Colony Optimiser (ASO)

Dorigo and Caro [62] proposed Ant Colony Optimization (ACO), which is technically motivated by the behaviours of ant colonies. We used an ACO algorithm to identify features as a comparing approach in this work. According to the ACO algorithm's theory, when ants discover a sign of food, they leave a fragrant chemical known as a pheromone to mark the trail [63]. When an ant seeks food, it follows the pheromone trail. Additionally, this ant deposits pheromones along the path, allowing other ants to follow suit. When an ant must choose between two roads, it chooses the one with a high pheromone level, indicating that more ants have travelled the path. It is a question of convenience for the ants; shorter trails become more fragrant than longer paths. If an ant does not follow a trail, the pheromone degrades over time. As a result, the intensity of the pheromone is diminished [64], and over time, all ants will take the shorter route to food. Finally, "pheromone evaporation" and "probabilistic path selection" supply information to ants for them to identify the shortest food path. The notions enable elasticity in the solution of optimisation problems. In a nutshell, an ant can use the information contained in the bodies of other ants to select a more practical choice.

2.1.7. Comparing Predictive Models

Three machine learning models were also included in determining a viable approach to machine learning and a feature selection approach. Multiple Linear Regression (MLR) seeks to model the relationship between two or more explanatory variables and a target variable. It aids in determining the extent to which variables vary [65]. Support Vector Regression (SVR) is a machine learning kernel approach is used for various tasks, including forecasting time series. SVRs that employ kernels can also learn the training data's non-linear trend. Three SVR models are available, each with a unique kernel (RBF, poly, and linear) [66]. Additionally, the SVR model has been used in a variety of research applications, including precipitation [67], solar radiation [68], wind energy [69], flood forecasting [68,70], evaporation [71], and crop yield [72,73] prediction.

Breiman [74] developed the random forest (RF) model, and it contains regression and classification methods. The RF model assembles tree predictors linked to distinct values of randomly sampled random vectors. The model creates decor-related decision trees

during the training phase, and the overall model output is derived by averaging the output values of all the individual trees [75]. The bootstrap resampling process generates a new set of training data from the initial training sample set N , and then K decision trees are used to construct bootstrap-set random forests. The complete specifications for the RF model may be found here [45]. RF is a collection (ensemble) of fundamental tree predictors. Each tree can generate a response given a set of predictor values [75]. The random forest method has been successfully implemented in predicting different crop yields worldwide [75,76].

3. Study Area and Data

3.1. Study Area and Wheat Yield Data

This study focused on a wheat yield prediction problem in South Australia, the fourth largest state in southern central Australia. Australia has a Mediterranean climate with an abundance of rain suitable for rainfed agriculture and crops like wheat. Wheat is the largest broadacre winter crop typically sown from May to June and harvested by November and December [26,77]. In the 2019–2020 financial year, 14 million tonnes of wheat were harvested in Australia, 18% less than the previous year, and 2.689 million tonnes of wheat were produced in South Australia (SA) [78]. Several climatic conditions, such as rainfall, soil moisture, temperature, solar radiation, humidity, etc., determine wheat production and are essential inputs to empirical and process-based models [26]. Delayed harvest due to a series of heavy rainfall in November and flooding in some regions is likely to cause a fall in wheat production in 2021–2022, which leads to poor grain quality [79]. Wang et al. [80] showed that climate variability could impact wheat production by 31% to 47%.

For this study, the average yearly wheat yield data for South Australia (SA) from 1990 to 2020 was downloaded from the Australian Bureau of Agricultural and Resource Economics and Sciences (ABARES) (<http://apps.agriculture.gov.au/agsurf/>, 29 December 2021). The dataset was acquired through farm surveys where the farm population ranged from 1967 to 9018 farms, and the sample was between 73 and 206 farms. The farm population was stratified based on operation size using the estimated value of the agricultural operation. The size of each stratum was decided using the Dalenius–Hodges method, while the sample was assigned to each stratum using a mixture of the Neyman allocation [81]. This dataset has been a prime source of information on the current and historical economic performance of Australian farm business units and has been used to undertake research and analysis on a range of industry issues and government policy areas.

It is worth to mention that South Australia's cropping zones are of three types, namely pastoral (411: SA North Pastoral), wheat-sheep (421: SA Eyre Peninsula; 422: SA Murray Lands and the Yorke Peninsula), and high rainfall (431: SA South East) zone [82]. Except for the Murray lands, where rainfall was generally average, most farming districts in South Australia experienced below-average rainfall in September. Most crops' yield potential was increased by adequate rainfall and mild temperatures in October, especially those sown later. However, the recovery in growth conditions in October came too late for crops in the upper regions of the Eyre Peninsula and the Yorke Peninsula, which had been harmed by dry conditions in early spring [79].

3.2. Predictor Variables

The monitoring of crop conditions using remote sensing is being used extensively to assess crop conditions, soil moisture, and the probability of natural disasters such as pest infestation, drought, and precipitation [83]. Ahmed et al. [55,63,84] has discussed the importance of satellite-based remote sensing to forecasting and constant monitoring of soil moisture and its importance in agriculture and human activities. Several other studies have also projected the correlation between weather conditions and remote sensing information to address the situation [26,85]. This study collected satellite data from NASA's

GES-DISC Interactive Online Visualization and Analysis Infrastructure (GIOVANNI) repository from 1991 to 2020. Specifically, the study used Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2), whose data set spans 1980 to the present.

Along with the improvements to meteorological assimilation, MERRA-2 makes significant progress towards the Earth System. MERRA-2 is the first long-term global reanalysis that incorporates space-based aerosol observations and their interactions with other physical processes in the climate system. The MERRA-2 model has a native spatial resolution of $0.5^\circ \text{ lat} \times 0.625^\circ \text{ long}$ and four temporal resolutions: daily, hourly, 3-hourly, and monthly. The study used 32 monthly hydro-climatic variables converted to the yearly data to be correlated with wheat yield, as tabulated in Table 1. The predictor variables were extracted as area-averaged of the time series data, as the target data (i.e., wheat yield) was provided for the whole of South Australia. Figure 1 depicts the atmospheric domain of South Australia between 127.44°E , 38.79°S , and 141.77°E , 23.76°S to extract the area-averaged wheat yield data. Satellite data collection has significant advantages overground stations regarding costs and coverage range. Local factors significantly impact ground stations and do not typically have a logical distribution system [86]. On the other hand, satellite remote sensing is not impacted by local conditions and captures data with a uniform cell size throughout the world. Interestingly, satellite data tracks crop growth conditions and gradually captures the variability in yield as the growing season progresses, and their contribution to yield prediction peaks during the growing season's peak [26,87].

Table 1. A description of the 32 predictors from the MERRA-2 satellite system used to design the hybrid GWO-CEEMDAN-KRR model for wheat yield prediction (tonnes) in South Australia. The feature selections were undertaken using GWO, ACO, PSO, and ASO, and a \checkmark shows the selected feature, whereas a \times shows the rejected feature.

Information of Satellite Derived Variables			Results of Feature Selection			
Notation	Description	Units	GWO	ACO	ASO	PSO
Q	Specific humidity @1000 hPa	kg/kg	\times	\checkmark	\checkmark	\checkmark
TA	Air temperature monthly @1000 hPa	K	\checkmark	\checkmark	\checkmark	\times
Q10	10-m specific humidity	kg/kg	\times	\checkmark	\checkmark	\checkmark
TO3	Total column ozone	Dobsons	\checkmark	\times	\times	\checkmark
T2X	2-m air temperature-daily max	K	\checkmark	\times	\checkmark	\checkmark
T2A	2-m air temperature-daily mean	K	\checkmark	\checkmark	\checkmark	\checkmark
T2M	2-m air temperature-daily min	K	\checkmark	\times	\checkmark	\checkmark
LE	Total latent energy flux	W/m^2	\times	\times	\times	\times
PR	Total precipitation	Kg/m^2	\times	\checkmark	\checkmark	\times
TA	Surface air temperature monthly	K	\times	\checkmark	\times	\checkmark
GRN	Greenness fraction	-	\checkmark	\times	\checkmark	\times
SW	Surface soil wetness	-	\times	\checkmark	\times	\checkmark
LAI	Leaf area index	-	\checkmark	\times	\times	\checkmark
ALB	Surface albedo	-	\checkmark	\times	\times	\checkmark
CL	Total cloud area fraction	-	\checkmark	\times	\checkmark	\times
SSF	Surface incoming shortwave flux	W/m^2	\times	\times	\times	\times
Q250	Specific humidity at 250 hPa	kg/kg	\checkmark	\times	\checkmark	\times
Q500	Specific humidity at 500 hPa	kg/kg	\times	\times	\checkmark	\checkmark
Q850	Specific humidity at 850 hPa	kg/kg	\times	\checkmark	\checkmark	\checkmark
Q10	10-m specific humidity	kg/kg	\times	\times	\checkmark	\checkmark
Q2	2-m specific humidity	kg/kg	\checkmark	\times	\checkmark	\times
SLP	Sea level pressure	hPa	\checkmark	\checkmark	\checkmark	\times
T10	Temperature at 10 m above surface	K	\times	\times	\checkmark	\times

T2	2-m air temperature	K	×	√	√	√
TS	Surface skin temperature	K	√	√	×	×
U10	10-m eastward wind	m/s	×	√	√	×
U2	2-m eastward wind	m/s	√	×	√	×
U50	Eastward wind at 50-m	m/s	√	×	√	√
V10	10-m northward wind	m/s	√	√	√	√
V2	2-m northward wind	m/s	√	√	√	√
V50	Northward wind at 50-m	m/s	√	√	√	×
A	Area	Ha	×	×	√	×
Total Number of Selected Features			18	15	24	17

3.3. Development of GWO-CEEMDAN-KRR Model

The proposed GWO-CEEMDAN-KRR model was developed on a personal computer (PC) equipped with an Intel i7 processor running at 3.6 GHz and 16 GB of RAM. A publicly available machine learning library, *scikit-learn* [88,89] using Python, was employed to execute the KRR model for the proposed framework. An implementation of the feature optimisation (i.e., GWO, ACO, ASO, and PSO) has been developed using MATLAB R2020b. The CEEMDAN method is executed with the programming language software R. To visualise further the anticipated wheat yield, tools such as *matplotlib* [90] and *seaborn* [91] are used, in addition to standalone methods. The following steps were carried out to develop the proposed GWO-CEEMDAN-KRR model.

Step 1: The 31 predictor variables obtained from the MERRA-2 satellite model were combined to screen the best-correlated input predictors using grey wolf optimisation (GWO) techniques. The use of GWO resulted in the best-selected predictors being used for feature decomposition. The optimal values of four selected feature selection algorithms are tabulated in Table 2. For the GWO, the optimal number of wolves is fixed at 10 with 100 iterations. Similarly, ACO, PSO, and ASO algorithms provide essential information on selecting significant predictor variables.

Table 1 shows the optimal set of satellite-derived features selected by GWO, ACO, ASO, and PSO methods.

Table 1 demonstrates that the GWO optimised diversified hydro-climatological variables for the predictive model.

Table 2 provides the optimum parameters of the GWO, ACO, ASO, and PSO algorithms

Step 2: In this step, each of the GWO optimised predictor variables was resolved into 4-IMFs (i.e., IMF1, IMF2, IMF3, and IMF4) and 1-residual (RES) using the CEEMDAN method ($18 \times 5 = 90$ IMFs in total). Gaussian Noise realisations ($N = 500$) and the provided amplitude in terms of added white noise ($=0.2$). The implementation of the CEEMDAN process is in Figure 1. The decomposed component was then correlated with the variable (i.e., wheat yield) by Pearson's correlation, and the most highly correlated components were chosen as the target components for the KRR model.

Table 2. The optimal parameter for the optimization algorithms such as grey wolf optimization (GWO), ant colony optimization (ACO), atom search optimization (ASO), and particle swarm optimization (PSO).

Characteristics	Optimal Value
Grey Wolf Optimization (GWO)	
Number of wolves	10
Maximum number of iterates	100
Curve	Convergence
Ant Colony Optimization (ACO)	
Number of ants	10

Maximum number of iterations	100
Coefficient control tau	1
Coefficient control eta	2
Initial tau	1
Initial beta	1
Pheromone	0.2
Coefficient	0.5
Atom Search Optimization (ASO)	
Number of particles	10
Maximum number of iterations	100
Depth weight	50
Multiplier weight	0.2
Particle Swarm Optimization (PSO)	
Number of particles	10
Maximum number of iterations	150
Cognitive factor	2
Social factor	2
Maximum velocity	6
Maximum bound on inertia weight	0.9
Minimum bound on inertia weight	0.4

Step 3: The selected predictor IMFs are normalised to minimise the overinfluence of one input to another. Using the following equation, all variable features were normalised to ensure that they received proportional attention in network training [0, 1] [26–28].

$$\delta_{norm} = \frac{\delta - \delta_{min}}{\delta_{max} - \delta_{min}} \quad (21)$$

In Equation (21), δ is the respective variable, δ_{min} is the minimum variable, δ_{max} is the maximum, and δ_{norm} is the normalised variable. After normalising the variables, the datasets are partitioned into training (1991–2010), validation (2011–2016), and testing (2017–2020) subsets. The data partitioning is done by the trial-and-error method.

Figure 2 shows the methodological steps of the proposed GWO-CEEMDAN-KRR model.

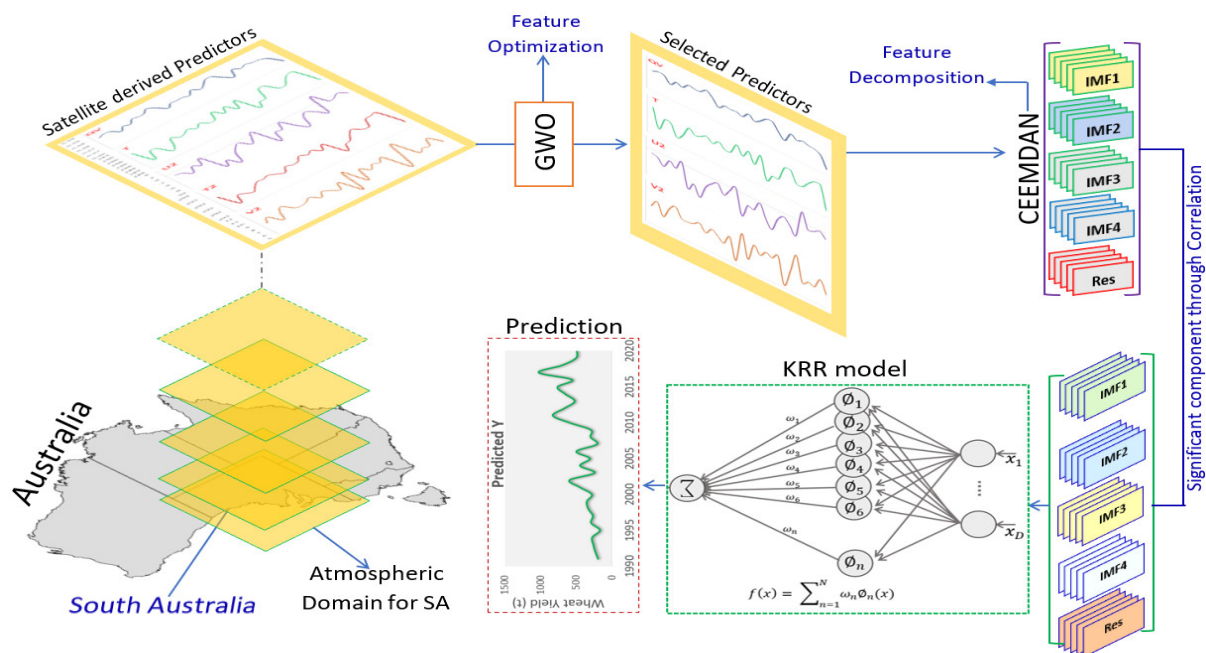


Figure 2. Integrated workflow showing the study area and atmospheric domain of South Australia with a schematic structure of KRR model integrating with GWO and CEEMDAN methods for the proposed GWO-CEEMDAN-KRR model for wheat yield prediction.

Step 4: To predict the wheat yield of South Australia, this study developed the KRR model to use the predictors' data in step 3. GridSearchCV was used to create an optimal architecture of the KRR model (regularisation strength = 1.5; gamma parameter = None, with a degree of the polynomial kernel = 3 and kernel = *rbf*). The performance of the proposed model was compared to that of standalone machine learning models.

3.4. Predictive Model Evaluation

The robustness of the proposed machine learning model (i.e., GWO-CEEMDAN-KRR) and the benchmark model is assessed using numerous performance metrics, e.g., Pearson's Correlation Coefficient (*r*), root mean square error (RMSE), and normalised root means square error (RMSE). Due to geographic differences between the study stations, we also employ the relative error-based metrics: i.e., relative MAE (denoted as RMAE), to compare geographically and climatologically diverse wheat yield sites. The accuracy of any predictive model is evaluated by comparing its predicted test values to the actual test results. The relative index of agreement (*d_{rel}*) can be a more sophisticated and compelling measure form than the RMSE when the error distribution in the tested data is Gaussian [92]. A sensitivity analysis was undertaken to evaluate the contributing response to the anticipated *Y* of the provided set of predictor variables to verify the prediction models created in our study. The goal was to see which predictor variables contributed the most to modelling the monthly evaporative loss value. Following previous research [93–95], we calculated the sensitivity % of the output (*E*) to each predictor (*x*) variable as follows:

$$Z_i = f_{\max}(x_i) - f_{\min}(x_i) \quad (22)$$

$$S_i = \frac{Z_i}{\sum_{i=1}^n Z_i} \times 100 \quad (23)$$

where $f_{\max}(x_i)$ and $f_{\min}(x_i)$ are, respectively, the maximum and the minimum predicted *Y* over the *i*th domain, where other variables are equal to their mean values. Z_i is the predicted value

4. Results

In this study, a hybrid KRR predictive model denoted as GWO-CEEMDAN-KRR is developed and evaluated for its capability to predict wheat yield (Y) in South Australia. The performance accuracy of wheat yield prediction is evaluated in comparison with several comparing models (i.e., CEEMDAN-MLR, CEEMDAN-RF, and CEEMDAN-SVR) and standalone methods (e.g., KRR, RF, MLR, and SVR) with all models employing four competitive feature optimisation algorithms (i.e., GWO, ACO, ASO, and PSO). The outcomes of the newly designed hybrid KRR predictive models were evaluated using statistical score metrics in conjunction with the diagnostic plots of both the observed and the predicted Y for the testing datasets.

Comparing the observed and predicted Y test data, we note that the newly developed CEEMDAN-KRR model can generate the highest value of R (showing a good degree of agreement between observed and predicted Y) while also generating the lowest values of NRMSE using the grey wolf optimisation method, according to the findings in Table 3. The GWO-CEEMDAN-KRR model with GWO produced $R \approx 0.998$, $NRMSE \approx 0.437\%$, followed by ACO-CEEMDAN-KRR ($R \approx 0.990$ and $NRMSE \approx 0.452\%$), PSO-CEEMDAN-KRR ($R \approx 0.980$ and $NRMSE \approx 0.477\%$) model that also produced substantially good, yet a lower performance relative to the GWO-CEEMDAN-KRR model. We discovered that an MLR model could produce better performance with a high R-value (0.963) in the standalone model. However, this model still underperforms the objective GWO-CEEMDAN-KRR model. Therefore, we note that the proposed CEEMDAN-KRR model using the grey wolf optimisation feature selection with an appropriate feature decomposition using the CEEMDAN) method provided the most satisfactory performance. Regarding the benchmark models' poor performance (as shown in Table 3), the newly proposed hybrid KRR (i.e., GWO-CEEMDAN-KRR) predictive model has proven to be a superior tool for predicting the wheat yield in South Australia using a carefully selected set of satellite-based predictor variables.

Table 3. Evaluation of the hybrid CEEMDAN-KRR vs. the benchmark (i.e., CEEMDAN-MLR, CEEMDAN-RF, CEEMDAN-SVR) models and their respective standalone counterpart (i.e., KRR, MLR, RF, and SVR) models. The r and normalized root mean square error (NRMSE) is computed between predicted and observed Wheat Yield (Y, tonnes) South Australia.

Predictive Model	R	NRMSE
GWO–Objective Feature Selection Method		
CEEMDAN-KRR	0.998	0.437
CEEMDAN-MLR	0.896	1.144
CEEMDAN-RF	0.751	0.589
CEEMDAN-SVR	0.840	0.614
ACO–benchmark method		
CEEMDAN-KRR	0.990	0.452
CEEMDAN-MLR	0.860	1.122
CEEMDAN-RF	0.847	0.531
CEEMDAN-SVR	0.681	0.743
ASO–benchmark method		
CEEMDAN-KRR	0.974	0.738
CEEMDAN-MLR	0.866	0.659
CEEMDAN-RF	0.768	0.601
CEEMDAN-SVR	0.849	0.523
PSO–benchmark method		
CEEMDAN-KRR	0.980	0.475
CEEMDAN-MLR	0.973	0.655
CEEMDAN-RF	0.784	0.689

CEEMDAN-SVR	0.929	0.525
Standalone		
KRR	0.738	0.761
MLR	0.963	2.992
RF	0.882	0.653
SVR	0.758	0.710

The predictive performance of the proposed hybrid GWO-CEEMDAN-KRR model is further evaluated by examining the relative error values (i.e., RMAE) and coefficient of determination (R^2) between the observed and predicted wheat yield in the testing phase as shown in Figure 3. According to Figure 3, the newly developed CEEMDAN-KRR model, with GWO algorithm, has the lowest percentage of RMAE ($\approx 32\%$) and the highest R^2 value (≈ 0.997), which is an impressive result compared with the same model with other optimisation techniques. We also noted that the GWO method could produce the best performance when integrated with the KRR-based predictive model compared with the different feature selection techniques.

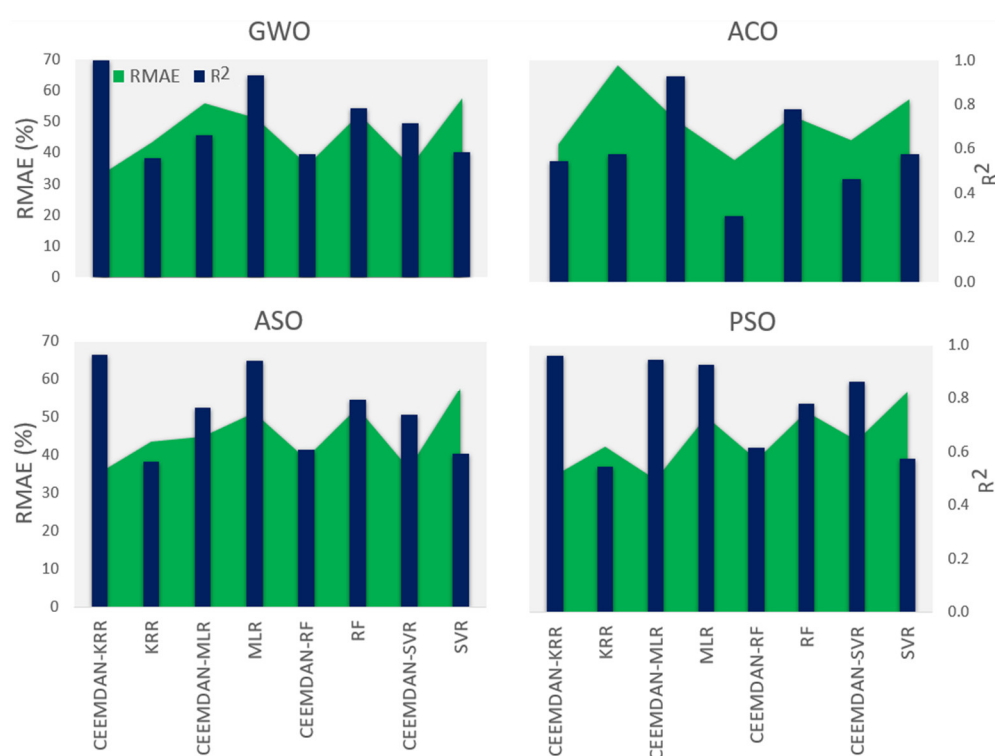


Figure 3. Comparison of the predictive skill of the proposed wheat yield prediction models in terms of the relative error: RMAE (%) and the correlation of determination (R^2) within the testing period.

Interestingly, the KRR model performs best when the predictor variables are decomposed with the CEEMDAN method using all optimisation techniques. Our results range between 32% and 36% regarding the RMAE value. The improvement in the prediction performance is more evident after applying the feature decomposition (i.e., CEEMDAN) and the feature optimisation (i.e., GWO) techniques. By these results, the hybrid CEEMDAN-KRR model seems to outperform the comparison benchmark models and the standalone machine learning models, demonstrating superior performance.

It is worth noting that this study has employed two distinct algorithms (one for feature selection, namely the GWO, and the other for decomposition of selected features, namely the CEEMDAN) to improve the overall performance of the hybrid KRR-based predictive model. As a result, in Figure 4, we illustrate the effects of incrementally

applying the CEEMDAN and different optimisation methods such as GWO, ACO, ASO, and PSO as the data pre-processing and the feature selection methods on the percentage change in error (i.e., RMAE) and percentage change in Willmott's Index (i.e., d_{rel}) from their respective standalone models. The RMAE (%) values of the CEEMDAN-KRR model, which incorporates a GWO method for satellite predictor variable feature selection, appear to decrease by $\approx 20\%$. For d_{rel} , this is a 35% increment from the standalone KRR model. Moreover, for the case of the ACO feature selection method, the change of RMAE and d_{rel} is 16 and 34%, for ASO, this change is 18 and 31%, and for PSO, the change is 10 and 30%, accordingly relative to the standalone KRR model. The other models, such as the SVR, RF, and MLR, showed a minimum improvement in utilising the four optimisation techniques and the CEEMDAN data decomposition technique. This indicates that incorporating the CEEMDAN and the GWO methods can improve the model's predictive capability in simulating the wheat yield tested data values. This is notable by these values outperforming the indices generated for the comparison model by a significant margin. Therefore, this exemplifies that the proposed hybrid predictive model is more accurate than competing methods used to predict wheat yield.

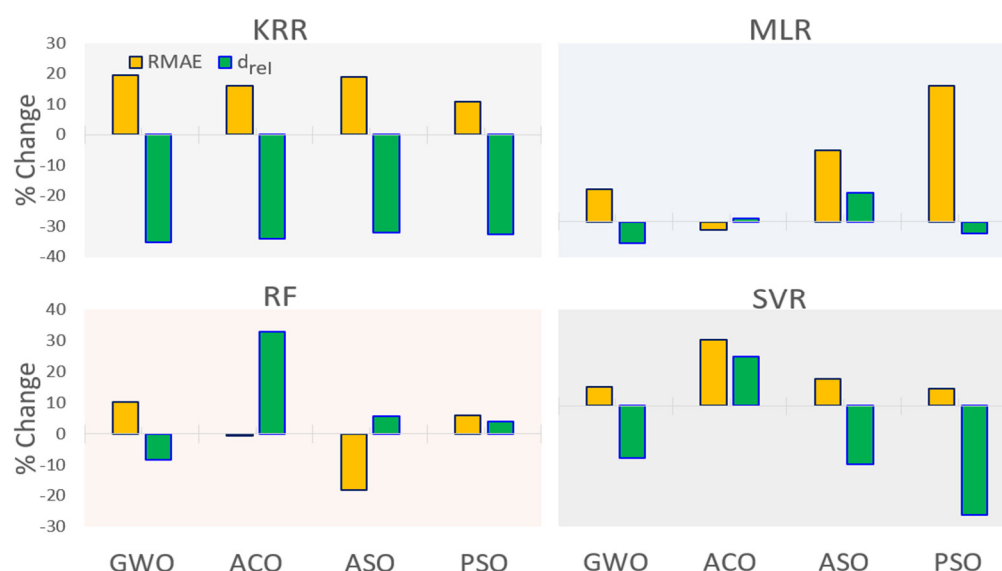


Figure 4. An assessment of four distinct feature selection methods regarding the percentage change in relative error (i.e., RMAE) and relative index of agreement (d_{rel}) with all methods using a CEEMDAN data decomposition approach in the model's testing phase.

To demonstrate a superior performance of the proposed GWO-CEEMDAN-KRR and its standalone counterpart models, we have also examined the prompting percentage of the correlation coefficient (ΔR), RMAE ($\Delta RMAE$), and NRMSE ($\Delta NRMSE$) for wheat yield prediction, as illustrated in Figure 5. Note that the prompting percentage, presented as the incremental performance (Δ) of the objective model over the competing approaches, aims to evaluate the difference in the R, RMAE, and NRMSE of the GWO-CEEMDAN-KRR against the other models. In general, the metrics ΔR , $\Delta RMAE$, and $\Delta NRMSE$ are used to demonstrate a performance edge of the preferred (i.e., GWO-CEEMDAN-KRR) model over the comparative counterparts. Figure 5 shows the results as for the case of ΔR , the improvement is found to be $\approx 1\%$ to 25% ; for the case of $\Delta RMAE$, the improvement is ≈ 2 to 60% . Likewise, improving prediction performance in terms of $\Delta NRMSE$ also demonstrates significant improvements. This demonstrates that our proposed model (i.e., GWO-CEEMDAN-KRR) was the most responsive in the prediction process.

The discrepancy ratio (Dr) is used to further investigate the proposed model's robustness. In general, the discrepancy ratio (Dr) measures whether a model overestimates or underestimates a simulated wheat yield value. The Dr value that begins with a "one"

indicates that an exact prediction can be made for a specific observation. According to Figure 6, the GWO-CEEMDAN-KRR model shows that the distribution of Dr is within a $\pm 30\%$ band error for observation of the testing phase. As determined by the discrepancy ratio, hybrid machine learning approaches were the most accurate predictive models compared to other models on the same basis. As shown in Figure 7, a scatter plot is used to perform an additional evaluation of the hybrid predictive model (i.e., CEEMDAN-KRR) where the GWO algorithm and the previous evaluation. The scatter plot is plotted with the goodness-of-fit between the predicted and observed Y, and a least-square fitting line to represent the relationship between the two variables. The suggested model outperforms the standalone model with an R^2 value significantly higher than the baseline model.

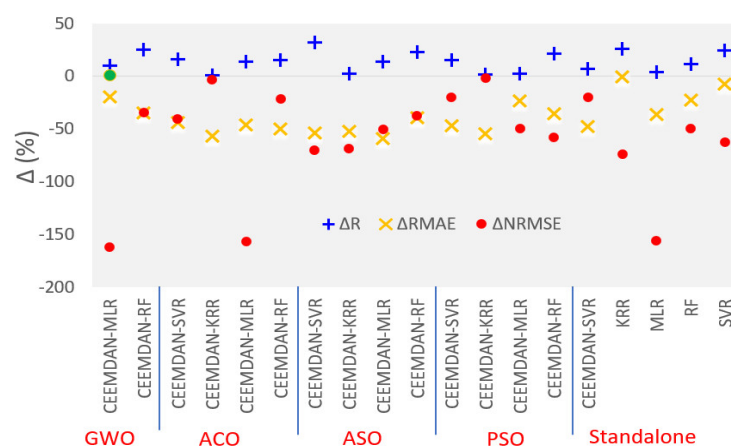


Figure 5. The prompting percentage (Δ) for correlation coefficient (ΔR), RMAE ($\Delta RMAE$), and NRMSE ($\Delta NRMSE$) between the proposed GWO-CEEMDAN-KRR model, other ACO, ASO, PSO used models, and the standalone models.

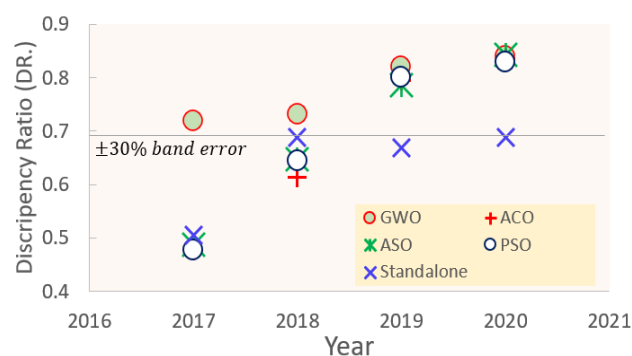


Figure 6. The discrepancy ratio (i.e., the predicted Y/observed Y) generated by the proposed hybrid CEEMDAN-KRR model using the four optimization algorithms and their respective standalone counterparts.

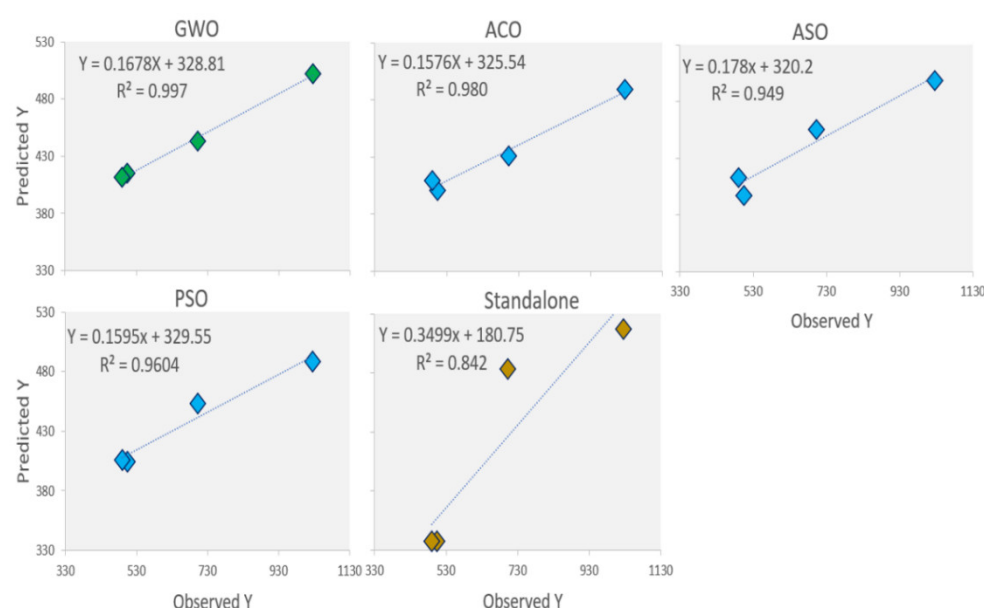


Figure 7. Scatter plot of the predicted and observed Y generated by proposed GWO-CEEMDAN-KRR model vs. the other models. A least square regression line, $Y = mX + C$, and the coefficient of determination (R^2) are shown in each sub-panel.

Concerning the proposed model with GWO algorithm, it performed significantly better than the other feature optimisation algorithms (i.e., ACO, ASO, and PSO), registering magnitudes that were the closest to unity ($m \mid R^2 \mid 0.167$), followed by the CEEMDAN-KRR model with ACO ($0.980 \mid 0.157$). For the case of standalone KRR, the unity has far deviated from the proposed model's exhibits statistically significant performance with the proposed model. Therefore, the learning hybrid CEEMDAN-KRR model with the GWO algorithm is exceptionally well suited for predicting wheat yield for South Australia.

The performance of Wheat prediction using GWO-CEEMDAN-KRR that is shown in Figure 8a (ECDF) examines the plots of various prediction skills using an empirical cumulative distribution function (ECDF). Comparing the performance of the proposed hybrid KRR model to the benchmark models, the generated error ranged from 50 to 300 within the 95 per cent percentile, demonstrating that the CEEMDAN-KRR model with the GWO model was the most accurate and responsive wheat yield prediction model. A Taylor diagram provides a more specific and conclusive argument about how strongly the predicted and observed Y are correlated than a simple correlation coefficient. As illustrated in Figure 8b, the output of the GWO-CEEMDAN-KRR model is significantly closer to the observation than the output of other comparing models, as indicated by the Taylor diagram. The GWO outperformed other models' optimised CEEMDAN-KRR model to achieve the observed values' closest match; however, the proposed model outperformed against other counterpart models. The study site had a higher R-value than the observed Y for the proposed CEEMDAN-KRR model, further supporting the findings of improved performance by this model, which was previously reported.

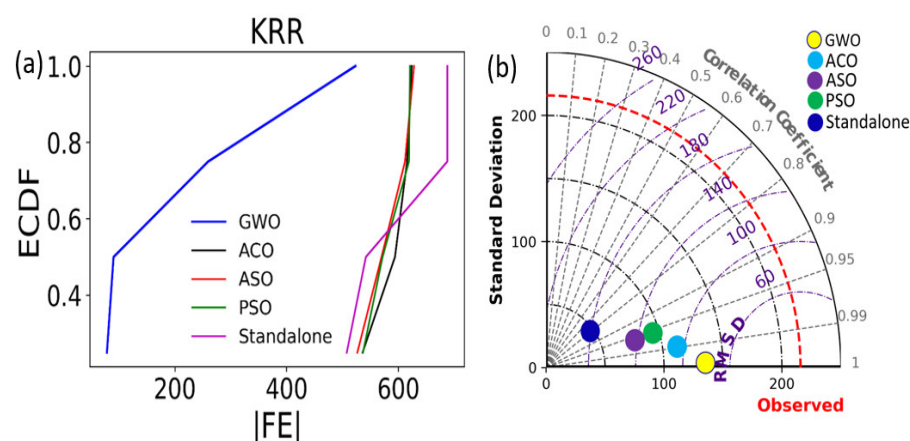


Figure 8. (a) An empirical cumulative distribution function (ECDF) plot of $|FE|$ and (b) Taylor diagram demonstrating the correlation coefficient, together with the standard deviation difference of the hybrid KRR model and standalone KRR with four optimisation algorithms (i.e., GWO, ASO, ACO, and PSO).

In addition to understanding the contribution of the input variables to the yield prediction, a sensitivity analysis of individual variables was performed. Figure 9 shows the results of sensitivity analysis for the proposed GWO-CEEMDAN-KRR model. It can be observed that almost all the parameters selected by GWO were significant, ranging from 20% to 33%. Specifically, the leaf area index (LAI) had the highest sensitivity, which is endorsed by other researchers [96,97]. However, inputs like V50, V10, V2, T2A, TS, and Q2 show a similar sensitivity percentage, ranging from 28% to 31%. The high sensitivity of the northward wind values is substantial, which is needed to be explored in further study. Moreover, surface albedo and other meteorological variables were also found to be significant in predicting wheat yield in South Australia.

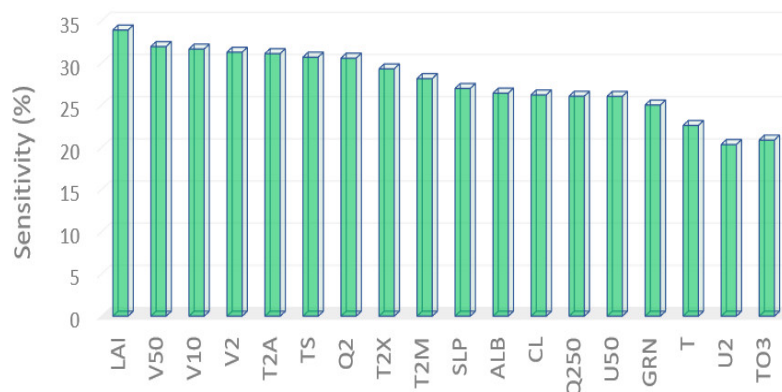


Figure 9. Sensitivity (%) analysis of predictor variables for the prediction of wheat yield (Y).

5. Discussion

The useful information derived from space combined with advanced machine algorithms enabled the development of more accurate near-real-time forecasts for different crops at different scales [98,99]. The findings from this study clearly showed that spatial information derived from MERRA-2 combined with the hybrid CEEMDAN-KRR model could provide an accurate forecast tool for wheat yield in Australia. The high accuracy of the proposed model has been proven through the reported model performance using different evaluation criteria and benchmark models. In this study, the CEEMDAN-KRR model with GWO generated R (0.998), and NRMSE (0.437%) outperformed other hybrid and standalone models. Furthermore, the integration of the GWO technique has indicated the most important predictors among 32 variables for wheat yield forecast. While the

present study contributes to the current research avenue, several limitations, challenges, and suggestions for further research are discussed.

This study used the space-based MERRA-2 dataset to exploit many variables related to atmospheric, weather, and canopy conditions. However, this dataset's coarse resolution ($0.5^\circ \times 0.625^\circ$) might affect the forecast accuracy of wheat yield. The predictor variables obtained as area-averaged of the time series data for the whole of South Australia's atmospheric domain would minimise the effect. Moreover, integrating vegetation indices (VIs), land surface temperature (LTS), and weather variables acquired from higher spatial resolution satellite data is highly recommended to overcome this issue. Multi-temporal VIs such as Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI) derived from MODIS (250 m), Landsat (30 m), and Sentinel (10 m) data have been successfully explored in predicting crop yield [100,101]. Furthermore, the composited products (e.g., from MODIS) on a near real-time basis of 8 days, 16 days, and months can overcome the cloud cover problem and, thus, improve the model performance. In addition, gridded precipitation retrieved from the Climate Hazards Group Infrared Precipitation with Stations (CHIRPS) dataset (~5.5 km) can provide helpful information to develop crop yield forecast models [102].

One of the main challenges of using satellite-based data to predict crop yield at a regional level is the lack of cropland cover masks. Zhang et al. [103] reported a consistent improvement in yield prediction using crop-specific masks at all regions and scales. In contrast, Shao et al. [104] claimed that using available cropland masks (e.g., summer crop or cultivated crops) generated similar results to using an annual corn-specific mask. It is also worth noting that the MERRA-2 dataset used in this study was extracted in the atmospheric domain of South Australia between 127.44°E , 38.79°S and 141.77°E , 23.76°S to extract the area-averaged wheat yield data, which potentially affects the forecast results. Therefore, it is interesting to explore whether the wheat crop's growing boundary could enhance the forecast accuracy.

Another factor that may affect the model performance is the algorithms used for modelling relationships between crop yield and predictors. Methods such as RF or SVM might not perform well with time-ordered data such as multi-temporal VIs and weather variables [102]. The authors demonstrated that the LSTM neural network model outperformed the multivariate OLS regression and random forest in soybean yield prediction. Our results also showed that the CEEMDAN-KRR model is superior to MLR, RF, and SVR-based models. In addition, deep learning methods are up-and-coming for the crop yield prediction problem [105,106]. Therefore, future research can consider using space-based datasets and deep learning approaches combined with automatic feature extraction to improve yield forecasts. This study has established an essential framework for building smart farming services. The high accuracy of crop yield prediction information in different climatic conditions using the proposed model is an essential element that helps agricultural producers and other stakeholders improve decision-making. In addition, this research helps rural areas where gauge-based observations are not always available. This is because satellite data can be used to help this research.

6. Conclusions

The prediction of wheat, subsistence, or commercial agricultural commodities using freely available satellite data and remote sensing methods can add value to new initiatives in precision agriculture. The active promotion of Agriculture 4.0, an Austrade strategy, showcases our competitive advantage in agtech and foodtech to a global audience through digital practices such as modelling crop yields through machine learning methods. This paper has developed and implemented a hybrid machine learning algorithm with an artificial intelligence methodology for wheat yield prediction in South Australia. The new approach uses a feature selection strategy and the subsequent decomposition of the selected features as an optimisation algorithm to improve the proposed Kernel Ridge Regression (KRR) and a set of competitive compression models. To train the prescribed

models, we have used thirty-two predictors derived from the MERRA-2 satellite datasets to encapsulate the features to model wheat yield and quantify the relationships between satellite-derived information and ground-based wheat yield. Our novel method combined the CEEMDAN, a feature decomposition method, and the grey wolf optimisation, a feature selection method, to improve kernel ridge regression prediction accuracy. The proposed hybrid GWO-CEEMDAN-KRR model, composed of five distinct modules for optimal accuracy, was tested on area-aggregated wheat yield data in South Australia. A common problem in data-driven modelling was solved when the GWO algorithm was used in the machine learning model. It reduced the number of predictor variables to solve this problem.

According to the results of this study, the proposed hybrid CEEMDAN-KRR model demonstrated the best performance in predicting wheat yield when it was optimised by the GWO method. The high R-value of the CEEMDAN-KRR predictive model, which ranged from 0.0980 to 0.998, and the low NRMSE value, which ranged from 0.437 to 0.475, supported the different feature selection techniques of the model's superior testing performance. More precisely, the CEEMDAN-KRR model improved with the GWO feature selection algorithm and registered the best performance. The scatterplot revealed that the merits of the CEEMDAN-KRR model with GWO are the closest to unity, supporting the applicability of the newly designed hybrid CEEMDAN-KRR model in real-time applications. Therefore, we ascertain that the proposed model can address a wide range of complex or challenging prediction tasks in agriculture and can be a helpful method for predicting other variables such as rainfall, wind speed, flood, or drought index. Global climate model (GCM) datasets could be used in the future to predict crop yields under different global warming scenarios, assess CO₂ emissions, and measure agricultural sustainability to figure out how future climate change and climate variability will affect farming.

Author Contributions: Conceptualisation, A.A.M.A.; methodology, A.A.M.A.; software, A.A.M.A.; model development, A.A.M.A.; validation, A.A.M.A.; formal analysis, A.A.M.A.; investigation, A.A.M.A.; resources, A.A.M.A. and T.N.-H.; data curation, A.A.M.A. and T.N.-H.; writing—original draft preparation, A.A.M.A., S.J.J.J. and E.S.; writing—review and editing, A.A.M.A., R.C.D., E.S., M.A., T.N.-H. and S.J.J.J.; visualisation, A.A.M.A.; funding acquisition, R.C.D. All authors have read and agreed to the published version of the manuscript.

Funding: The study was supported by the Chinese Academy of Science (CAS) and the University of Southern Queensland (USQ) USQ-CAS Postgraduate Research Scholarship (2019–2021).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Data were obtained from the GIOVANNI database, duly acknowledged. We thank the Editors and Reviewers for their time and insightful comments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Pathak, H.; Aggarwal, P.K.; Singh, S. *Climate Change Impact, Adaptation and Mitigation in Agriculture: Methodology for Assessment and Applications*; Indian Agricultural Research Institute, New Delhi, India, 2012; Volume 302.
2. Rosenberg, N.J. Adaptation of agriculture to climate change. *Clim. Chang.* **1992**, *21*, 385–405.
3. Rickards, L.; Howden, S.M. Transformational adaptation: Agriculture and climate change. *Crop Pasture Sci.* **2012**, *63*, 240–250.
4. Leng, G.; Hall, J.W. Predicting spatial and temporal variability in crop yields: An inter-comparison of machine learning, regression and process-based models. *Environ. Res. Lett.* **2020**, *15*, 044027.
5. Iizumi, T.; Ramankutty, N. How do weather and climate influence cropping area and intensity? *Glob. Food Secur.* **2015**, *4*, 46–50.
6. Ruane, A.C.; Major, D.C.; Winston, H.Y.; Alam, M.; Hussain, S.G.; Khan, A.S.; Hassan, A.; Al Hossain, B.M.T.; Goldberg, R.; Horton, R.M. Multi-factor impact analysis of agricultural production in Bangladesh with climate change. *Glob. Environ. Chang.* **2013**, *23*, 338–350.

7. Challinor, A.J.; Watson, J.; Lobell, D.B.; Howden, S.; Smith, D.; Chhetri, N. A meta-analysis of crop yield under climate change and adaptation. *Nat. Clim. Chang.* **2014**, *4*, 287–291.
8. Olesen, J.E.; Bindi, M. Consequences of climate change for European agricultural productivity, land use and policy. *Eur. J. Agron.* **2002**, *16*, 239–262.
9. Thornton, P.K.; Jones, P.G.; Alagarswamy, G.; Andresen, J. Spatial variation of crop yield response to climate change in East Africa. *Glob. Environ. Chang.* **2009**, *19*, 54–65.
10. Alexandrov, V.; Hoogenboom, G. The impact of climate variability and change on crop yield in Bulgaria. *Agric. For. Meteorol.* **2000**, *104*, 315–327.
11. Romeijn, H.; Faggian, R.; Diogo, V.; Sposito, V. Evaluation of deterministic and complex analytical hierarchy process methods for agricultural land suitability analysis in a changing climate. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 99.
12. Aschonitis, V.; Mastrocicco, M.; Colombani, N.; Salemi, E.; Kazakis, N.; Voudouris, K.; Castaldelli, G. Assessment of the intrinsic vulnerability of agricultural land to water and nitrogen losses via deterministic approach and regression analysis. *Water Air Soil Pollut.* **2012**, *223*, 1605–1614.
13. Meenken, E.; Wheeler, D.; Brown, H.; Teixeira, E.; Espig, M.; Bryant, J.; Triggs, C. Framework for uncertainty evaluation and estimation in deterministic agricultural models. *Nutr. Manag. Farmed Landsc. Occas. Rep.* **2020**, *33*, 1–11.
14. Kingsley, J.; Afu, S.M.; Isong, I.A.; Chapman, P.A.; Kebonye, N.M.; Ayito, E.O. Estimation of soil organic carbon distribution by geostatistical and deterministic interpolation methods: A case study of the southeastern soils of Nigeria. *Environ. Eng. Manag. J. EEMJ* **2021**, *20*, 1077–1085.
15. Holman, I.; Tascone, D.; Hess, T. A comparison of stochastic and deterministic downscaling methods for modelling potential groundwater recharge under climate change in East Anglia, UK: Implications for groundwater resource management. *Hydrogeol. J.* **2009**, *17*, 1629–1641.
16. Sharma, E.; Deo, R.C.; Prasad, R.; Parisi, A.V. A hybrid air quality early-warning framework: An hourly forecasting model with online sequential extreme learning machines and empirical mode decomposition algorithms. *Sci. Total Environ.* **2020**, *709*, 135934.
17. Sharma, E.; Deo, R.C.; Prasad, R.; Parisi, A.V.; Raj, N. Deep Air Quality Forecasts: Suspended Particulate Matter Modeling With Convolutional Neural and Long Short-Term Memory Networks. *IEEE Access* **2020**, *8*, 209503–209516.
18. Kouadio, L.; Deo, R.C.; Byrareddy, V.; Adamowski, J.F.; Mushtaq, S.; Nguyen, V.P. Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. *Comput. Electron. Agric.* **2018**, *155*, 324–338.
19. Ren, J.; Chen, Z.; Zhou, Q.; Tang, H. Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong, China. *Int. J. Appl. Earth Obs. Geoinf.* **2008**, *10*, 403–413.
20. Franch, B.; Vermote, E.; Becker-Reshef, I.; Claverie, M.; Huang, J.; Zhang, J.; Justice, C.; Sobrino, J.A. Improving the timeliness of winter wheat production forecast in the United States of America, Ukraine and China using MODIS data and NCAR Growing Degree Day information. *Remote Sens. Environ.* **2015**, *161*, 131–148.
21. Han, J.; Zhang, Z.; Cao, J.; Luo, Y.; Zhang, L.; Li, Z.; Zhang, J. Prediction of winter wheat yield based on multi-source data and machine learning in China. *Remote Sens.* **2020**, *12*, 236.
22. Wang, Y.; Zhang, Z.; Feng, L.; Du, Q.; Runge, T. Combining multi-source data and machine learning approaches to predict winter wheat yield in the conterminous United States. *Remote Sens.* **2020**, *12*, 1232.
23. Wang, X.; Huang, J.; Feng, Q.; Yin, D. Winter wheat yield prediction at county level and uncertainty analysis in main wheat-producing regions of China with deep learning approaches. *Remote Sens.* **2020**, *12*, 1744.
24. Haider, S.A.; Naqvi, S.R.; Akram, T.; Umar, G.A.; Shahzad, A.; Sial, M.R.; Khaliq, S.; Kamran, M. LSTM neural network based forecasting model for wheat production in Pakistan. *Agronomy* **2019**, *9*, 72.
25. Kolotii, A.; Kussul, N.; Shelestov, A.; Skakun, S.; Yailymov, B.; Basarab, R.; Lavreniuk, M.; Oliinyk, T.; Ostapenko, V. Comparison of biophysical and satellite predictors for wheat yield forecasting in Ukraine. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2015**, *XL-7/W3*, 39–44.
26. Cai, Y.; Guan, K.; Lobell, D.; Potgieter, A.B.; Wang, S.; Peng, J.; Xu, T.; Asseng, S.; Zhang, Y.; You, L. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agric. For. Meteorol.* **2019**, *274*, 144–159.
27. Landau, S.; Mitchell, R.; Barnett, V.; Colls, J.; Craigon, J.; Payne, R. A parsimonious, multiple-regression model of wheat yield response to environment. *Agric. For. Meteorol.* **2000**, *101*, 151–166.
28. Kumar, S.; Attri, S.; Singh, K. Comparison of Lasso and stepwise regression technique for wheat yield prediction. *J. Agrometeorol.* **2019**, *21*, 188–192.
29. Kogan, F.; Kussul, N.N.; Adamenko, T.I.; Skakun, S.V.; Kravchenko, A.N.; Krivobok, A.A.; Shelestov, A.Y.; Kolotii, A.V.; Kussul, O.M.; Lavrenyuk, A.N. Winter wheat yield forecasting: A comparative analysis of results of regression and biophysical models. *J. Autom. Inf. Sci.* **2013**, *45*, 68–81.
30. Kamir, E.; Waldner, F.; Hochman, Z. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. *ISPRS J. Photogramm. Remote Sens.* **2020**, *160*, 124–135.
31. Bali, N.; Singla, A. Deep Learning Based Wheat Crop Yield Prediction Model in Punjab Region of North India. *Appl. Artif. Intell.* **2021**, 1–25, doi: 10.1080/08839514.2021.1976091.
32. Liaghat, S.; Balasundram, S.K. A review: The role of remote sensing in precision agriculture. *Am. J. Agric. Biol. Sci.* **2010**, *5*, 50–55.
33. Ozdogan, M.; Yang, Y.; Allez, G.; Cervantes, C. Remote sensing of irrigated agriculture: Opportunities and challenges. *Remote Sens.* **2010**, *2*, 2274–2304.

34. Nelson, R.; Kokic, P.; Crimp, S.; Meinke, H.; Howden, S. The vulnerability of Australian rural communities to climate variability and change: Part I—Conceptualising and measuring vulnerability. *Environ. Sci. Policy* **2010**, *13*, 8–17.
35. Luo, Q.; Bellotti, W.; Williams, M.; Wang, E. Adaptation to climate change of wheat growing in South Australia: Analysis of management and breeding strategies. *Agric. Ecosyst. Environ.* **2009**, *129*, 261–267.
36. Luo, Q.; Bellotti, W.; Williams, M.; Bryan, B. Potential impact of climate change on wheat yield in South Australia. *Agric. For. Meteorol.* **2005**, *132*, 273–285.
37. Tikhamarine, Y.; Malik, A.; Kumar, A.; Souag-Gamane, D.; Kisi, O. Estimation of monthly reference evapotranspiration using novel hybrid machine learning approaches. *Hydrol. Sci. J.* **2019**, *64*, 1824–1842.
38. Gundoshmian, T.M.; Ardabili, S.; Mosavi, A.; Várkonyi-Kóczy, A.R. Prediction of combine harvester performance using hybrid machine learning modeling and response surface methodology. In Proceedings of the 18th International Conference on Global Research and Education, Inter-Academia 2019, Budapest, Hungary, 4–7 September 2019; pp. 345–360.
39. Shin, J.-Y.; Kim, K.R.; Ha, J.-C. Seasonal forecasting of daily mean air temperatures using a coupled global climate model and machine learning algorithm for field-scale agricultural management. *Agric. For. Meteorol.* **2020**, *281*, 107858.
40. Kabir, M.M.; Shahjahan, M.; Murase, K. A new hybrid ant colony optimization algorithm for feature selection. *Expert Syst. Appl.* **2012**, *39*, 3747–3763.
41. Too, J.; Abdullah, A.R. Chaotic atom search optimization for feature selection. *Arab. J. Sci. Eng.* **2020**, *45*, 6063–6079.
42. Abualigah, L.M.; Khader, A.T.; Hanandeh, E.S. A new feature selection method to improve the document clustering using particle swarm optimization algorithm. *J. Comput. Sci.* **2018**, *25*, 456–466.
43. Wang, Y.; Yuan, Z.; Liu, H.; Xing, Z.; Ji, Y.; Li, H.; Fu, Q.; Mo, C. A new scheme for probabilistic forecasting with an ensemble model based on CEEMDAN and AM-MCMC and its application in precipitation forecasting. *Expert Syst. Appl.* **2022**, *187*, 115872.
44. Ghali, U.M.; Usman, A.; Degm, M.A.A.; Alsharksi, A.N.; Naibi, A.M.; Abba, S. Applications of artificial intelligence-based models and multi-linear regression for the prediction of thyroid stimulating hormone level in the human body. *Int. J. Adv. Sci. Technol.* **2020**, *29*, 3690–3699.
45. Ali, M.; Prasad, R.; Xiang, Y.; Yaseen, Z.M. Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. *J. Hydrol.* **2020**, *584*, 124647.
46. Kisi, O.; Parmar, K.S. Application of least square support vector machine and multivariate adaptive regression spline models in long term prediction of river water pollution. *J. Hydrol.* **2016**, *534*, 104–112.
47. Zhao, P.; Xia, J.; Dai, Y.; He, J. Wind speed prediction using support vector regression. In Proceedings of the 2010 5th IEEE Conference on Industrial Electronics and Applications, Auckland, New Zealand, 15–17 June 2015; pp. 882–886.
48. Naik, J.; Satapathy, P.; Dash, P. Short-term wind speed and wind power prediction using hybrid empirical mode decomposition and kernel ridge regression. *Appl. Soft Comput.* **2018**, *70*, 1167–1188.
49. Li, T.; Zhou, Y.; Li, X.; Wu, J.; He, T. Forecasting daily crude oil prices using improved CEEMDAN and ridge regression-based predictors. *Energies* **2019**, *12*, 3603.
50. Santhosh, M.; Venkaiah, C.; Kumar, D.V. Ensemble empirical mode decomposition based adaptive wavelet neural network method for wind speed prediction. *Energy Convers. Manag.* **2018**, *168*, 482–493.
51. Liang, T.; Xie, G.; Fan, S.; Meng, Z. A Combined Model Based on CEEMDAN, Permutation Entropy, Gated Recurrent Unit Network, and an Improved Bat Algorithm for Wind Speed Forecasting. *IEEE Access* **2020**, *8*, 165612–165630.
52. Jin, T.; Li, Q.; Mohamed, M.A. A novel adaptive EEMD method for switchgear partial discharge signal denoising. *IEEE Access* **2019**, *7*, 58139–58147.
53. Zhang, W.; Qu, Z.; Zhang, K.; Mao, W.; Ma, Y.; Fan, X. A combined model based on CEEMDAN and modified flower pollination algorithm for wind speed forecasting. *Energy Convers. Manag.* **2017**, *136*, 439–451. <https://doi.org/10.1016/j.enconman.2017.01.022>.
54. Torres, M.E.; Colominas, M.A.; Schlotthauer, G.; Flandrin, P. A complete ensemble empirical mode decomposition with adaptive noise. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 4144–4147.
55. Ahmed, M.; Deo, R.C.; Raj, N.; Ghahramani, A.; Feng, Q.; Yin, Z.; Yang, L. 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 Sens.* **2021**, *13*, 554.
56. Al-Tashi, Q.; Kadir, S.J.A.; Rais, H.M.; Mirjalili, S.; Alhussian, H. Binary optimization using hybrid grey wolf optimization for feature selection. *IEEE Access* **2019**, *7*, 39496–39508.
57. Kennedy, J.; Eberhart, R. Particle swarm optimization. In Proceedings of the ICNN'95—International Conference on Neural Networks, Perth, WA, Australia, 27 November–1 December 1995.
58. Roy, D.K.; Lal, A.; Sarker, K.K.; Saha, K.K.; Datta, B. Optimization algorithms as training approaches for prediction of reference evapotranspiration using adaptive neuro fuzzy inference system. *Agric. Water Manag.* **2021**, *255*, 107003.
59. Sun, L.; Song, X.; Chen, T. An improved convergence particle swarm optimization algorithm with random sampling of control parameters. *J. Control. Sci. Eng.* **2019**, *2019*, 7478498.
60. Zhao, W.; Wang, L.; Zhang, Z. Atom search optimization and its application to solve a hydrogeologic parameter estimation problem. *Knowl. Based Syst.* **2019**, *163*, 283–304.
61. Mirjalili, S.; Lewis, A. S-shaped versus V-shaped transfer functions for binary particle swarm optimization. *Swarm Evol. Comput.* **2013**, *9*, 1–14.

62. Dorigo, M.; Di Caro, G. Ant colony optimization: A new meta-heuristic. In Proceedings of the 1999 Congress on Evolutionary Computation-CEC99 (Cat. No. 99TH8406), Washington, DC, USA, 6–9 July 1999; pp. 1470–1477.
63. Ahmed, M.; Deo, R.; Feng, Q.; Ghahramani, A.; Raj, N.; Yin, Z.; Yang, L. Hybrid deep learning method for a week-ahead evapotranspiration forecasting. *Stoch. Environ. Res. Risk Assess.* **2022**, *36*, 831–849.
64. Sweetlin, J.D.; Nehemiah, H.K.; Kannan, A. Feature selection using ant colony optimization with tandem-run recruitment to diagnose bronchitis from CT scan images. *Comput. Methods Programs Biomed.* **2017**, *145*, 115–125.
65. Abba, S.; Hadi, S.J.; Abdullahi, J. River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques. *Procedia Comput. Sci.* **2017**, *120*, 75–82.
66. Yang, P.; Xia, J.; Zhang, Y.; Hong, S. Temporal and spatial variations of precipitation in Northwest China during 1960–2013. *Atmos. Res.* **2017**, *183*, 283–295. <https://doi.org/10.1016/j.atmosres.2016.09.014>.
67. Belayneh, A.; Adamowski, J. Standard precipitation index drought forecasting using neural networks, wavelet neural networks, and support vector regression. *Appl. Comput. Intell. Soft Comput.* **2012**, *2012*, 6.
68. Deo, R.C.; Wen, X.; Qi, F. A wavelet-coupled support vector machine model for forecasting global incident solar radiation using limited meteorological dataset. *Appl. Energy* **2016**, *168*, 568–593. <https://doi.org/10.1016/j.apenergy.2016.01.130>.
69. Dhiman, H.S.; Deb, D.; Guerrero, J.M. Hybrid machine intelligent SVR variants for wind forecasting and ramp events. *Renew. Sustain. Energy Rev.* **2019**, *108*, 369–379.
70. Dodangeh, E.; Panahi, M.; Rezaie, F.; Lee, S.; Bui, D.T.; Lee, C.-W.; Pradhan, B. Novel hybrid intelligence models for flood-susceptibility prediction: Meta optimization of the GMDH and SVR models with the genetic algorithm and harmony search. *J. Hydrol.* **2020**, *590*, 125423.
71. Baydaroglu, Ö.; Koçak, K. SVR-based prediction of evaporation combined with chaotic approach. *J. Hydrol.* **2014**, *508*, 356–363. <https://doi.org/10.1016/j.jhydrol.2013.11.008>.
72. Khosla, E.; Dharavath, R.; Priya, R. Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression. *Environ. Dev. Sustain.* **2020**, *22*, 5687–5708.
73. Jaikla, R.; Auephanwiriyakul, S.; Jintrawet, A. Rice yield prediction using a support vector regression method. In Proceedings of the 2008 5th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, Chiang Rai, Thailand, 14–17 May 2008; pp. 29–32.
74. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32.
75. Jui, S.J.J.; Ahmed, A.A.M.; Bose, A.; Raj, N.; Sharma, E.; Soar, J.; Chowdhury, M.W.I. Spatiotemporal Hybrid Random Forest Model for Tea Yield Prediction Using Satellite-Derived Variables. *Remote Sens.* **2022**, *14*, 805. <https://doi.org/10.3390/rs14030805>.
76. Prasad, N.; Patel, N.; Danodia, A. Crop yield prediction in cotton for regional level using random forest approach. *Spat. Inf. Res.* **2021**, *29*, 195–206.
77. Zhao, Y.; Potgieter, A.B.; Zhang, M.; Wu, B.; Hammer, G.L. Predicting wheat yield at the field scale by combining high-resolution Sentinel-2 satellite imagery and crop modelling. *Remote Sens.* **2020**, *12*, 1024.
78. ABS. Agricultural Commodities, Australia, 2019–2020 Financial Year. 2020. Available online: <https://www.abs.gov.au/statistics/industry/agriculture/agricultural-commodities-australia/latest-release> (accessed on 25 December 2021).
79. AWE. Australian Government Department of Agriculture, Water and the Environment. National Overview—DAWE. 2021. Available online: <https://www.awe.gov.au/abares/research-topics/agricultural-outlook/australian-crop-report/overview> (accessed on 25 December 2021).
80. Wang, B.; Chen, C.; Li Liu, D.; Asseng, S.; Yu, Q.; Yang, X. Effects of climate trends and variability on wheat yield variability in eastern Australia. *Clim. Res.* **2015**, *64*, 173–186.
81. Lehtonen, R.; Pahkinen, E. *Practical Methods for Design and Analysis of Complex Surveys*; John Wiley & Sons: Hoboken, NJ, USA, 2004.
82. ABARES. Department of Agriculture, Water and the Environment-ABARES. 2022. Available online: <https://www.awe.gov.au/abares> (accessed on 25 December 2021).
83. Doraiswamy, P.C.; Moulin, S.; Cook, P.W.; Stern, A. Crop yield assessment from remote sensing. *Photogramm. Eng. Remote Sens.* **2003**, *69*, 665–674.
84. Ahmed, A.A.M.; Ahmed, M.H.; Saha, S.K.; Ahmed, O.; Sutradhar, A. Optimization Algorithms as Training Approach with Deep Learning Methods to Develop an Ultraviolet Index Forecasting Model. 2021. Available online: https://www.researchgate.net/publication/354741827_Optimization_Algorithms_As_Training_Approach_With_Deep_Learning_Methods_To_Develop_An_Ultraviolet_Index_Forecasting_Model (accessed on 20 December 2021).
85. Teng, W.; de Jeu, R.; Doraiswamy, P.; Kempfer, S.; Mladenova, I.; Shannon, H. Improving world agricultural supply and demand estimates by integrating NASA remote sensing soil moisture data into USDA world agricultural outlook board decision making environment. In Proceedings of the American Society of Photogrammetry and Remote Sensing 2010 Annual Conference, San Diego, CA, USA, 26–30 April 2010.
86. Sohrabinia, M.; Khorshiddoust, A.M. Application of satellite data and GIS in studying air pollutants in Tehran. *Habitat Int.* **2007**, *31*, 268–275. <https://doi.org/10.1016/j.habitatint.2007.02.003>.
87. Guan, K.; Berry, J.A.; Zhang, Y.; Joiner, J.; Guanter, L.; Badgley, G.; Lobell, D.B. Improving the monitoring of crop productivity using spaceborne solar-induced fluorescence. *Glob. Chang. Biol.* **2016**, *22*, 716–726.
88. Kramer, O. Scikit-learn. In *Machine Learning for Evolution Strategies*, Springer: Berlin/Heidelberg, Germany, 2016; pp. 45–53.

89. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
90. Barrett, P.; Hunter, J.; Miller, J.T.; Hsu, J.-C.; Greenfield, P. matplotlib--A Portable Python Plotting Package. In Proceedings of the Astronomical Data Analysis Software and Systems XIV, Pasadena, CA, USA, 24–27 October 2004; p. 91.
91. Waskom, M.; Botvinnik, O.; Ostblom, J.; Gelbart, M.; Lukauskas, S.; Hobson, P.; Gemperline, D.C.; Augspurger, T.; Halchenko, Y.; Cole, J.B. Mwaskom/Seaborn: v0. 10.1 (April 2020). Zenodo. 2020. Available online: <https://ui.adsabs.harvard.edu/abs/2020zndo...3767070W%2F/abstract> (accessed on 25 December 2021).
92. Krause, P.; Boyle, D.; Bäse, F. Comparison of different efficiency criteria for hydrological model assessment. *Adv. Geosci.* **2005**, *5*, 89–97.
93. Gandomi, A.H.; Yun, G.J.; Alavi, A.H. An evolutionary approach for modeling of shear strength of RC deep beams. *Mater. Struct.* **2013**, *46*, 2109–2119. <https://doi.org/10.1617/s11527-013-0039-z>.
94. Samui, P.; Dixon, B. Application of support vector machine and relevance vector machine to determine evaporative losses in reservoirs. *Hydrol. Processes* **2012**, *26*, 1361–1369.
95. Deo, R.C.; Samui, P.; Kim, D. Estimation of monthly evaporative loss using relevance vector machine, extreme learning machine and multivariate adaptive regression spline models. *Stoch. Environ. Res. Risk Assess.* **2015**, *30*, 1769–1784. <https://doi.org/10.1007/s00477-015-1153-y>.
96. Baez-Gonzalez, A.D.; Kiniry, J.R.; Maas, S.J.; Tiscareno, M.L.; Macias, C.J.; Mendoza, J.L.; Richardson, C.W.; Salinas, G.J.; Manjarrez, J.R. Large-area maize yield forecasting using leaf area index based yield model. *Agron. J.* **2005**, *97*, 418–425.
97. Huang, J.; Tian, L.; Liang, S.; Ma, H.; Becker-Reshef, I.; Huang, Y.; Su, W.; Zhang, X.; Zhu, D.; Wu, W. Improving winter wheat yield estimation by assimilation of the leaf area index from Landsat TM and MODIS data into the WOFOST model. *Agric. For. Meteorol.* **2015**, *204*, 106–121.
98. Sagan, V.; Maimaitijiang, M.; Bhadra, S.; Maimaitiyiming, M.; Brown, D.R.; Sidike, P.; Fritsch, F.B. Field-scale crop yield prediction using multi-temporal WorldView-3 and PlanetScope satellite data and deep learning. *ISPRS J. Photogramm. Remote Sens.* **2021**, *174*, 265–281.
99. Shetty, S.A.; Padmashree, T.; Sagar, B.; Cauvery, N. Performance analysis on machine learning algorithms with deep learning model for crop yield prediction. In *Data Intelligence and Cognitive Informatics*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 739–750.
100. Son, N.; Chen, C.; Chen, C.; Minh, V.; Trung, N. A comparative analysis of multitemporal MODIS EVI and NDVI data for large-scale rice yield estimation. *Agric. For. Meteorol.* **2014**, *197*, 52–64.
101. Satir, O.; Berberoglu, S. Crop yield prediction under soil salinity using satellite derived vegetation indices. *Field Crops Res.* **2016**, *192*, 134–143.
102. Schwalbert, R.A.; Amado, T.; Corassa, G.; Pott, L.P.; Prasad, P.V.; Ciampitti, I.A. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agric. For. Meteorol.* **2020**, *284*, 107886.
103. Zhang, Y.; Chipanshi, A.; Daneshfar, B.; Koiter, L.; Champagne, C.; Davidson, A.; Reichert, G.; Bédard, F. Effect of using crop specific masks on earth observation based crop yield forecasting across Canada. *Remote Sens. Appl. Soc. Environ.* **2019**, *13*, 121–137.
104. Shao, Y.; Campbell, J.B.; Taff, G.N.; Zheng, B. An analysis of cropland mask choice and ancillary data for annual corn yield forecasting using MODIS data. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *38*, 78–87.
105. Nevavuori, P.; Narra, N.; Lipping, T. Crop yield prediction with deep convolutional neural networks. *Comput. Electron. Agric.* **2019**, *163*, 104859.
106. Van Klompenburg, T.; Kassahun, A.; Catal, C. Crop yield prediction using machine learning: A systematic literature review. *Comput. Electron. Agric.* **2020**, *177*, 105709.