

UNIVERSITY OF THE WESTERN CAPE

Geo-physical parameter forecasting on
imagery-based data sets using machine
learning techniques



A thesis submitted in fulfilment for the
degree of Master of Science

in the
Faculty of Natural Sciences
Department of Computer Science

Supervisor: Dr Mehrdad Ghaziasgar
Co-supervisor: Prof Christopher Thron and Prof Mattia Vaccari

March 2022

<http://etd.uwc.ac.za/>

Declaration of Authorship



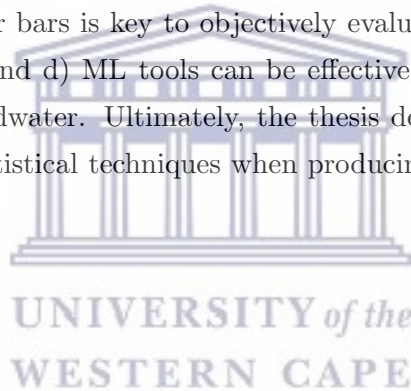
I, ESLAM HUSSEIN, declare that this thesis “*Geo-physical parameter forecasting on imagery-based data sets using machine learning techniques*” is my own work, that it has not been submitted before for any degree or assessment at any other university, and that all the sources I have used or quoted have been indicated and acknowledged by means of complete references.

Signed: _____

Date: _____

Abstract

This research objectively investigates the effectiveness of machine learning (ML) tools towards predicting several geo-physical parameters. This is based on a large number of studies that have reported high levels of prediction success using ML in the field. Therefore, several widely used ML tools coupled with a number of different feature sets are used to predict six geophysical parameters namely rainfall, groundwater, evaporation, humidity, temperature, and wind. The results of the research indicate that: a) a large number of related studies in the field are prone to specific pitfalls that lead to over-estimated results in favour of ML tools; b) the use of gaussian mixture models as global features can provide a higher accuracy compared to other local feature sets; c) ML never outperform simple statistically-based estimators on highly-seasonal parameters, and providing error bars is key to objectively evaluating the relative performance of the ML tools used; and d) ML tools can be effective for parameters that are slow-changing such as groundwater. Ultimately, the thesis demonstrates the importance of using well-grounded statistical techniques when producing and analyzing the results of ML predictive models.



Acknowledgements

First and foremost, all praises and thanks to the Almighty for granting me the guidance and strength to complete this thesis.

Secondly, my sincere gratitude to my research supervisors for their patience and invaluable guidance throughout the research.

Thirdly, to my family, this thesis is a product of almost 30 years of sacrifice, love, support and patience. Thank you all for the selfless support.


Finally, I would like to thank the South African National Research Foundation (NRF CSUR Grant Number 121291 for the HIPPO project) and to Telkom/Openserve/Aria Technologies Centre-of-Excellence at the University of the Western Cape for their essential financial support.



Publications

- **Title: Regional Rainfall Prediction Using Support Vector Machine Classification of Large-Scale Precipitation Maps**
Authors: Eslam A. Hussein, Mehrdad Ghaziasgar, Christopher Thron
Published [18] in the proceedings of the 23rd International IEEE conference on information fusion 2020 (FUSION 2020).
- **Title: Groundwater Prediction Using Machine-Learning Tools**
Authors: Eslam A. Hussein, Christopher Thron, Mehrdad Ghaziasgar, Antoine Bagula, Mattia Vaccari
Published [21] in the MDPI Algorithms journal (2020).
- **Title: Basic Statistical Estimation Outperforms Machine Learning in Monthly Prediction of Seasonal Climatic Parameters**
Authors: Eslam A. Hussein, Mehrdad Ghaziasgar, Christopher Thron, Mattia Vaccari, Antoine Bagula
Published [19] in the MDPI Atmosphere journal (2021).
- **Title: Rainfall Prediction Using Machine Learning Models: Literature Survey**
Authors: Eslam A. Hussein, Mehrdad Ghaziasgar, Christopher Thron, Mattia Vaccari, Yahlieel Jafta
Accepted for publication [20] as a book chapter in the book entitled "Empowering Artificial Intelligence in Data Science" published by Springer (2021).

Contents

Declaration of Authorship	i
Abstract	ii
Acknowledgements	iii
Publications	iv
List of Figures	vii
Abbreviations	viii
 <p>UNIVERSITY of the WESTERN CAPE</p>	
1 Introduction	1
1.1 Background	1
1.2 Research Question	5
1.3 Research Objectives	5
1.4 Thesis Structure and Outline	6
1.4.1 Chapter 2	8
1.4.2 Chapter 3	9
1.4.3 Chapter 4	9
1.4.4 Chapter 5	10
1.4.5 Chapter 6	12
2 Literature review	13
3 Manuscript “Regional Rainfall Prediction Using Support Vector Machine Classification of Large-Scale Precipitation Maps Groundwater Prediction Using Machine-Learning Tools”	44
4 Manuscript “Groundwater Prediction Using Machine-Learning Tools”	54
5 Manuscript “Basic Statistical Estimation Outperforms Machine Learning in Monthly Prediction of Seasonal Climatic Parameters”	72
6 Conclusion	93
6.1 Future Work	97

6.2 Concluding Comments 97

Bibliography **98**



List of Figures

1.1	Satellite precipitation images an image from the NCEP dataset [15]. . . .	3
1.2	The growth of machine learning for modeling geophysical parameters [2]. .	4
1.3	Summary of the mapping between the research objectives (labelled Obj1–Obj5) and the thesis chapters (labelled CH2–CH5), showing the chapter(s) that address each objective.	6
1.4	Summary of the mapping between the research sub-questions (labelled Q1–Q4) and the thesis chapters (labelled CH2–CH5), showing the chapter(s) that address each sub-question.	7



Abbreviations

ML	Machine learning
GMMs	Gaussian Mixture Models
ANNs	Artificial neural networks
CNNs	Convolution neural networks
LSTMs	Long short term memory
ConvLSTMs	Convolutions layers with Long short term memory
MLP	Multilayer perceptron
RF	Random forest
SVMs	Support vector machines
SVR	Support vector regression
XGB	eXtreme gradient boosting
MLR	Multi linear regression
KNN	K-nearest neighbour
RMSE	Root mean square error
MAE	Mean absolute error
Rain	Rainfall
Temp	Temperature
Evap	Evaporation
Humid	Humidity

Chapter 1

Introduction

1.1 Background

The study of physical geography is about understanding and representing changes on the Earth's surface. It focuses on factors that affect the geographic nature of the Earth. The functioning of the Earth can be classified into the following processes: hydrological processes like storm waves and groundwater; biological processes like forest growth; atmospheric processes like thunderstorms and rainfall; human processes like urban development; and geological processes like earthquakes. So the field of physical geography seeks to investigate the distribution of the different features/parameters that describe the landscape and functioning of the planet by analyzing the processes that shape it. These features/parameters have been referred to as geophysical parameters in the literature [24], and the overall field that involves the study of various geophysical parameters that describe the functioning of the planet is called physical geography.

Modelling geophysical parameters to forecast their future behavior is essential for long-term planning in a number of different contexts. The following are some important examples of uses of accurate geophysical parameter forecasting:

- Accurate geophysical parameter forecasting can help to take necessary steps and make preparations for extreme weather such as droughts and floods etc.
- The forecasting of rainfall and groundwater can be immensely helpful in decision-making for the agricultural sector.
- The forecasting of food and energy demand and consumption can be very instrumental in urban planning and development.

- The ability to predict potential disasters such as earthquakes, tsunamis and volcanic eruptions can help mitigate their effects on the population.

Many other example use-cases exist for geophysical parameter forecasting, and it is clear that it is a very important area of research.

Historically, in order to be able to understand and predict geophysical parameters, researchers began to gather raw data through geographic enquiries. The collection of some datasets started as early as centuries ago [22, 24]. In the past century, the advancement of computer data processing and remote sensing has helped geophysical parameter researchers to record a large number of varied geophysical parameter data, resulting in petabytes of data recorded per day [24]. This made a shift in the geophysical parameter research community to move from a data-scarce to a data-rich environment [29].

The datasets used in geophysical parameters can be characterized according to two properties:

1. **The time interval of the data set:** this specifies the interval between data points during data collection. It can vary from very short intervals of a few seconds or minutes, to medium intervals of one or more hours or days, to long and very long intervals of one or more months or years. Shorter time intervals provide more fine-grained information of variations in a given parameter, whereas longer time intervals provide a better overview of long-term trends in the data.
2. **The space dimensionality of the data set:** This specifies the number of dimensions of each time point where:
 - (a) 1D data means that all data points in the data set were collected at a static known location, which often comes from e.g. meteorological stations.
 - (b) 2D data means that the data point associated with each time point is a 2D structure of data values, where the two dimensions implicitly specify the location on a 2D grid of a given data value recording, such as longitude and latitude. This is common with satellite images and radar.
 - (c) 3D data is similar to 2D data, with a third spatial dimension that specifies e.g. height/depth. This is common with datasets collected by means of LiDAR.

As an example, Figure 1.1 represents a map of 2D precipitation data on a daily time interval from the National Center for Environmental Prediction (NCEP) dataset [15].

The two properties mentioned above i.e. time interval and space dimensionality directly shape and influence the methodology used in generating a forecasting model, right from

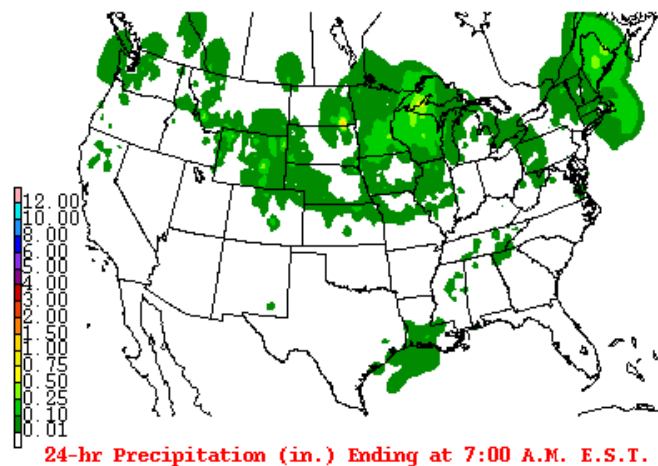


FIGURE 1.1: Satellite precipitation images an image from the NCEP dataset [15].

the pre-processing phase up to creation of the predictive model [24]. A wide variety of features can be used in geophysical parameter forecasting and prediction such as previous lags, and climate indices. [4, 9, 16, 17, 25, 26, 30, 41, 43].

In general, the use of machine learning (ML) towards the creation of predictive models for geophysical parameter forecasting has received increased attention in research. Figure 1.3 shows the progress in the usage of ML towards geophysical parameter forecasting [2]. The graph clearly demonstrates a consistent increase in the number of research papers produced, year-on-year from 2010 onwards. This demonstrates the importance and relevance of this field.

The exponential increase in the use of ML towards geophysical parameter forecasting can be attributed to the availability of data, the abundance and increase in computational resources and significant advances in ML frameworks. ML is one of the approaches that scientific communities have utilized extensively to model large complex datasets in a variety of fields, which includes physical geography.

The ML tools used in geophysical parameters forecasting can be categorized into two broad types:

1. **Deep learning tools:** Deep learning tools have consistently been shown to perform well on very large data sets typical of datasets with short and very short time intervals. Furthermore, the ability of deep learning tools to perform automatic feature selection and reduction on high-dimensional data with a large number of features is another advantage [18]. The tools used include Recurrent Neural Networks

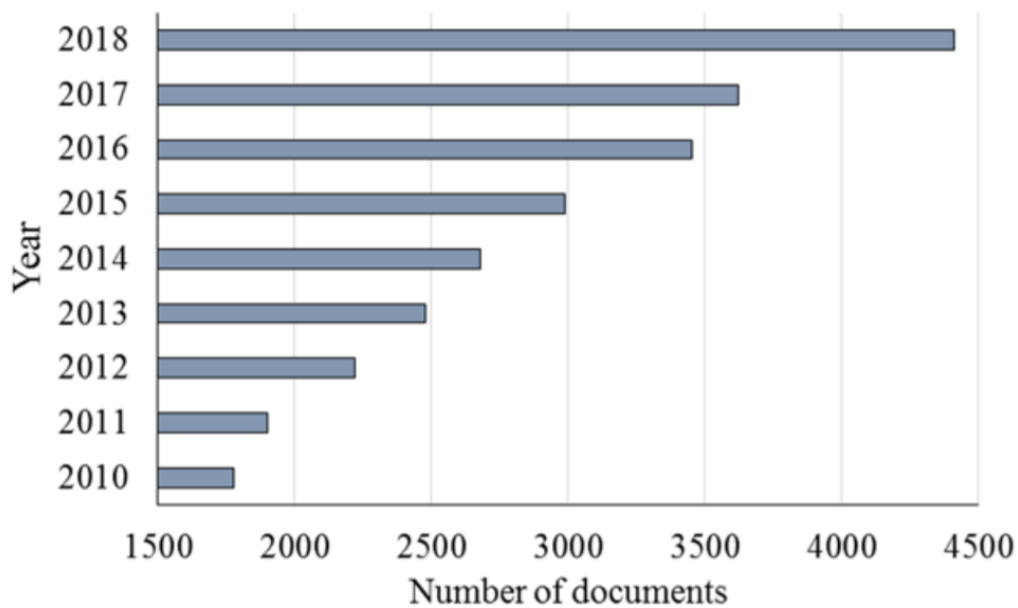


FIGURE 1.2: The growth of machine learning for modeling geophysical parameters [2].

(RNNs) in [32, 38], convolutional neural networks (CNNs) in [5, 6, 31, 32, 48], and various combinations between the two techniques in [3, 7, 23, 36–40, 42, 44].

2. **Classical ML tools:** These tools are more often used with small-to-medium sized data sets, typical of data sets with medium or long time intervals, accompanied with manual or auto feature selection using external techniques such as Principal Components Analysis (PCA) [12–14, 31, 33, 47]. The ML tools used in the literature vary from simple models like multivariate linear regression (MLR) [8, 9, 14, 26], to more advanced models like support vector machines (SVMs) and random forest (RF) in [1, 11, 14, 28, 34, 35, 46].

The accuracy and effectiveness of ML tools is still under debate in the literature [10]. To elaborate on this point: there is a tendency in the literature to make use of more sophisticated methods such as neural networks, without first investigating simpler methods. This is problematic as neglecting the use of simpler baselines makes it difficult to objectively determine the true effectiveness of ML tools relative to simpler baselines. A recent study carried out an investigation of several papers that used RNNs for top-n-recommendation tasks, and the study showed that a simple model using k -Nearest Neighbours (k NNs) outperforms several of the more sophisticated models [10]. Another study discussed two methods based on neural networks in the field of information retrieval. The results showed that the proposed methods do not significantly outperform simpler baseline methods [27].

Therefore, this research closely investigates the ability of several ML tools such as SVMs, extreme gradient boosting (XGB) and multilayer perceptron (MLP), to predict several geo-physical parameters relative to simple statistical estimators. Several features are also investigated in terms of their effectiveness on prediction success. Features investigated include global features such as the use of gaussian mixture models (GMMs), and local features such as previous lags of geophysical parameters used. The features are detailed later in this thesis.

1.2 Research Question

The main research question posed in this research is phrased as follows: “How effective is ML at predicting geophysical parameter data, given the spatio-temporal aspect of this kind of data?”

This research question can be broken down into the following sub-questions:

1. What are the evaluation and pre-processing pitfalls when it comes to spatio-temporal parameter forecasting?
2. What features/parameters affect the prediction success of the geophysical parameter prediction models?
3. How do various ML techniques compare in terms of geophysical parameter forecasting success?
4. How do the various ML techniques compare to appropriate simple baselines?

1.3 Research Objectives

The following research objectives will be used to obtain answers to the research sub-questions mentioned in the previous section, the answers to which will culminate in an answer to the main research question posed in the previous section:

1. Conduct a detailed literature survey on geophysical parameter forecasting, to show how related and relevant research studies approach the problem of geophysical parameter forecasting, with a focus on data sets, data pre-processing, evaluation and baselining strategies.

2. Investigate, use and compare various appropriate features of varying complexity for geophysical forecasting. The features that can be considered include global features such as the use of GMMs, and local features such as pixel values.
3. Select and train several ML models on the features selected in the previous objective to forecast geophysical parameters.
4. Compare the prediction success of the trained ML models.
5. Develop appropriate baselines and compare the trained models to the baselines, in order to put the success of the more sophisticated trained models into perspective.

1.4 Thesis Structure and Outline

Chapters 2–5 of this thesis takes the form of a series of peer-reviewed double-blind publications with the researcher as first author in all of these publications. The publications include two journal papers, one International conference paper and one book chapter that have either been published or have been accepted for publication at the time of writing, through which the objectives set out in this chapter were progressively met and the research sub-questions set out in this chapter, answered.

Each paper is self-contained and provides a complete set of information. This includes introduction and background, related work, implementation, results and conclusions, as relevant in each case. The formatting of each paper is in line with the requirements of the publishing outlet to which it was submitted for publication.

	CH2	CH3	CH4	CH5
Obj 1	○			
Obj 2		○	○	○
Obj 3		○	○	○
Obj 4			○	○
Obj 5		○	○	○

FIGURE 1.3: Summary of the mapping between the research objectives (labelled Obj1–Obj5) and the thesis chapters (labelled CH2–CH5), showing the chapter(s) that address each objective.

.Further details on each of the papers which form Chapters 2–5 are provided below, specifically pointing out which research objective(s) each paper addresses, and this information is visually summarized in Figure 1.3, as well as pointing out which research

	CH2	CH3	CH4	CH5
Q1	○			
Q2		○	○	○
Q3			○	○
Q4		○	○	○

FIGURE 1.4: Summary of the mapping between the research sub-questions (labelled Q1–Q4) and the thesis chapters (labelled CH2–CH5), showing the chapter(s) that address each sub-question.

sub-question(s) each paper addresses, and this information is visually summarized in Figure 1.4. This will form the basis on which to conclude the thesis in Chapter 6 by answering each of the research sub-questions, culminating in an answer to the main research question set out in Section 1.2.

The following sub-sections 1.4.1–1.4.5 describe each of the papers, providing overviews of each study and pointing out which research objectives/sub-questions are addressed, and the contributions that are made in each case.

It should be noted that the studies in Chapters 3–5 are similar in methodology whereby in each case several ML techniques coupled with several features were trained to predict one or more geophysical parameters, and the final models were compared to appropriate simple baselines. Therefore, in general, the following contributions are common for these papers:

- In line with research objective 2 and providing insight into research sub-question 2, the use and comparison of several features of varying complexity for geophysical parameter prediction.
- In line with research objectives 3 and 4 and providing insight into research sub-question 3, the use and comparison of several commonly used ML techniques for geophysical parameter prediction.
- In line with research objective 5 and providing insight into research sub-question 4, the development of appropriate simple baselines to which the ML models were compared.

Having noted these contributions, they will be omitted in the subsections below for the sake of brevity.

1.4.1 Chapter 2

The study in this chapter has been accepted for inclusion in the book titled "Empowering Artificial Intelligence in Data Science" that will be published by Springer. The study in this chapter looks at related work in the literature, focusing on geophysical parameter prediction implementations, with a focus on rainfall which is one of the most prominent geophysical parameters. It demonstrates the methods used in the literature, including data sets, pre-processing techniques, right through to ML models used, and evaluation and baselining strategies. This chapter aims to meet research objective 1 in order to provide an answer to research sub-question 1.

The contributions of the study in this chapter can be summarised as follows:

- A systematic review of 66 papers, which focuses the use of machine learning for rainfall forecasting.
- The survey provides researchers with crucial awareness of the different pitfalls that can lead to unrealistic and over-estimated model performance. These pitfalls apply equally to rainfall as to any other geophysical parameter, and in fact time series data prediction in general. Some of the key findings in this paper that can be useful to other researchers in this field include the following:
 - Many studies fail to utilize practices that prevent data leakage, leading to overestimates of predictive accuracy. These practices are clearly highlighted in the study.
 - Many studies do not provide error bars for prediction errors, so that the significance of differences between prediction methods cannot be determined.
 - Many studies in the literature do not use simple baselines for comparison, but rather compare several variations or architectures of more advanced ML methods such as SVR or MLP. This makes it very difficult to objectively gauge the performance of the proposed advanced techniques in real terms.

These contributions serve to meet research objective 1 and provide an answer to research sub-question 1, whereby a large number of relevant papers were reviewed and examined, and potential errors during the pre-processing and evaluation process that are made by authors were pointed out, which this research tries to consider in subsequent chapters.

1.4.2 Chapter 3

The study in this chapter has been published in the proceedings of the 23rd International IEEE conference on information fusion (Fusion 2020) [18]. The study proposes, implements and evaluates a class-based approach to rainfall prediction, with predictions ranging from 1–30 days ahead. The study made regional predictions based on sequences of daily rainfall maps of the continental US, with rainfall quantized at 3 levels: light or no rain; moderate rain; and heavy rain.

In this initial study, one ML technique was used for prediction, namely, support vector machine (SVM), in partial satisfaction of research objective 3. Furthermore, the trained model was compared to a simple baseline developed in the study. The results showed that predictions by the SVM in edge regions were less accurate than predictions obtained by a simple untrained classifier. However, in central regions, the SVM outperformed the untrained classifier. That paper provides initial evidence that SVMs applied to large-scale precipitation maps can under some conditions, but not all, provide useful information for predicting regional rainfall, but care must be taken to avoid pitfalls pointed out in the study in Chapter 2.

In addition to the common contributions of Chapters 3–5 mentioned previously, the study in this chapter additionally makes the following contributions:

- In line with research objective 2 and providing insight into research sub-question 2, the chapter shows that increasing the complexity of the features does not necessarily yield better prediction success, and in fact may contribute towards worsening the prediction success.
- In line with research objective 5 and providing insight into research sub-question 4, the chapter shows that a very simple baseline in most cases outperforms SVMs.

Accordingly, the chapter helps to build towards research objectives 2, 3 and 5 towards developing answers to research sub-questions 2 and 4.

1.4.3 Chapter 4

This chapter has been published as a paper in the MDPI Algorithms journal in 2020 [21]. This chapter proposes, implements and evaluates a machine-learning approach to groundwater prediction with the following characteristics:

1. the use of a regression-based approach to predict full groundwater images based on sequences of monthly groundwater maps.
2. the use and comparison of a range of local and global features, coupled with strategic automatic feature selection using extreme gradient boosting.
3. the use of a multiplicity of ML techniques, namely, extreme gradient boosting, multivariate linear regression, random forests, multilayer perceptron and support vector regression, in line with research objective 3 and 4.
4. the development of a simple baseline to which the ML techniques were compared, in line with research objective 5

In addition to the common contributions of Chapters 3–5 mentioned previously, the study in this chapter additionally makes the following contributions:

- In line with research objective 2 and providing insight into research sub-question 2, the paper proposes a novel global feature obtained from a GMM which ultimately produced ensemble based models with lower error than the best models which could be obtained with local geographical features.
- As a further confirmation of the findings in Chapter 3 in line with research objective 2 and providing insight into research sub-question 2, the study further confirms that more complicated features do not necessarily contribute towards better prediction success.
- In line with research objectives 3 and 4 and providing insight into research sub-question 3, while all techniques were able to successfully predict groundwater one month ahead, support vector regression consistently performed best in terms of minimizing root mean square error and mean absolute error.
- In line with research objective 5 and providing insight into research sub-question 4, the study successfully showed that all ML techniques were able to out-perform the simple baseline in this study for this geophysical parameter.

Accordingly, the study in the chapter helps to build towards research objectives 2–5 towards developing answers to research sub-questions 2–4.

1.4.4 Chapter 5

This chapter has been accepted for publication as a paper in the MDPI Atmosphere journal in 2021 [19]. In line with research objectives 2–4 to provide insight into research

sub-questions 2–4, this study scrutinizes the effectiveness of five widely used ML algorithms using several combinations of features towards the monthly prediction of several geophysical parameters, namely, rainfall, humidity, evaporation, temperature and wind, using monthly image data.

Furthermore, in line with research objective 5 the study compares the predictive accuracy of the resulting trained ML models to that of simple baseline statistical estimators that are computed directly from the training data.

Unlike other related studies, the study provides error bars for the relative performance of different predictors based on jackknife estimates applied to differences in predictive error magnitudes. The study also shows that the practice of shuffling data sequences which was employed in some previous related studies leads to data leakage, resulting in over-estimated performance.

Ultimately, the paper demonstrates the importance of using well-grounded statistical techniques when producing and analyzing the results of ML predictive models.

In addition to the common contributions of Chapters 3–5 mentioned previously, the study in this chapter additionally makes the following contributions:

- As a further confirmation of findings in previous chapters and in line with research objective 2 and providing further insight into research sub-question 2, the study confirms yet again that more complicated features do not necessarily contribute towards better prediction success.
- As a further confirmation of the findings in Chapter 2, and in line with research objective 5 and providing further insight into research sub-question 4, the study shows that ML never significantly outperforms the statistical baseline for these geophysical parameters, and under-performs for most feature sets, ranging from simple to more sophisticated feature sets.
- Further to the previous contribution, the study provides error bars to measure the relative performance of different predictors based on jackknife estimates applied to differences in predictive error magnitudes, which has not been done in any related studies previously. This tool can help objectively and realistically determine the effectiveness of ML models compared to the baselines, in line with research objective 5 and providing further insight into research sub-question 4.
- Linked to the previous contribution, the study also clearly shows that data leakage, which is regularly to be found in related studies, can lead to over estimation in model performance.

Accordingly, this culminating study in this chapter helps to finalize research objectives 2–5 towards developing final answers to research sub-questions 2–4.

1.4.5 Chapter 6

This chapter concludes the thesis by providing clear answers to the research sub-questions, thereby providing an answer to the main research question posed in this chapter. The chapter closes with several directions for future work.



Chapter 2

Literature review



UNIVERSITY *of the*
WESTERN CAPE

Rainfall Prediction Using Machine Learning Models: Literature Survey

Eslam A. Hussein^{1*}, Mehrdad Ghaziasgar¹, Christopher Thron², Mattia Vaccari³, and Yahleel Jafta¹

¹ Department of Computer Science, University of the Western Cape, Cape Town, 7535, South Africa

ehussein@uwc.ac.za (E.A.H.), mghaziasgar@uwc.ac.za (M.G.),
2858132@myuwc.ac.za (Y.J.)

² Department of Science and Mathematics, Texas A&M University-Central, Texas, Killeen, TX 76549, USA

thron@tamuct.edu

³ Department of Physics and Astronomy, University of the Western Cape, Cape Town, 7535, South Africa;

mvaccari@uwc.ac.za

Abstract. Research on rainfall prediction contributes to different fields that have a huge impact on our daily life. With the advancement of computer technology, machine learning has been extensively used in the area of rainfall prediction. However, some papers suggest that applications of machine learning in different fields are deficient in some respects. This chapter performs a review on 66 research papers that use machine learning tools to predict rainfall. The papers are examined in terms of the source of the data, output objective, input features, pre-processing, model used, and the results. The review shows questionable aspects present in many studies. In particular, many studies lack a baseline predictor for comparison. Also, many references do not provide error bars for prediction errors, so that the significance of differences between prediction methods cannot be determined. In addition, some references utilize practices that permit data leakage, leading to overestimates of predictive accuracy.

Keywords: forecasting, short and long term data, geophysical, deep learning, sequence prediction, data leakage, baselining, error bars, shuffling, seasonality.

1 Introduction

Natural processes on Earth can be classified into several categories, including hydrological processes like storm waves and groundwater; biological processes

* E.A.H. acknowledges financial support from the South African National Research Foundation (NRF CSUR Grant Number 121291 for the HIPPO project) and from the Telkom-Openseve-Aria Technologies Center of Excellence at the Department of Computer Science of the University of the Western Cape.

like forest growth; atmospheric processes like thunderstorms and rainfall; human processes like urban development; and geological processes like earthquakes. The field of physical geography seeks to investigate the distribution of the different features/parameters that describe the landscape and functioning of the Earth by analyzing the processes that shape it. These features/parameters have been referred to as geophysical parameters in the literature [37].

Rainfall is a key geophysical parameter that is essential for many applications in water resource management, especially in the agriculture sector. Predicting rainfall can help managers in various sectors to make decisions regarding a range of important activities such as crop planting, traffic control, the operation of sewer systems, and managing disasters like droughts and floods [32]. A number of countries such as Malaysia and India depend on the agriculture sector as a major contributor to the economy [32, 58] and as a source of food security. Hence, an accurate prediction of rainfall is needed to make better future decisions to help manage activities such as the ones mentioned above.

Rainfall is considered to be one of the most complicated parameters to forecast in the hydrological cycle [32, 34, 52]. This is due to the dynamic nature of environmental factors and random variations, both spatially and temporally, in these factors [32]. Therefore, to address random variations in rainfall, several machine learning (ML) tools including artificial neural networks (ANN), k -nearest neighbours (KNNs), decision trees (DT), etc. are used in the literature to learn patterns in the data to forecast rainfall. In this chapter, a review of past work in the area of rainfall prediction using ML models is carried out.

A number of related review papers exist as follows. The authors in [51] focused on reviewing studies that use ML for flood prediction, which closely resembles rainfall prediction. The authors in [70] focused on the use of ML for generic spatiotemporal sequence forecasting. Finally, the authors in [58] conducted a survey on the use of ML for rainfall prediction: however the study was limited to rainfall prediction in India.

This chapter serves as an addition to the field by surveying recent relevant studies focusing on the use of ML in rainfall prediction in a variety of geographic locations from 2016–2020. After detailing the methods used to forecast rainfall, one of the important contributions of this chapter is to demonstrate various pitfalls that lead to an overestimation in model performance of the ML models in various papers. This in turn leads to unrealistic hype and expectations surrounding ML in the current literature. It also leads to an unrealistic understanding of the advancements in, and gains by, ML research in this field. It is therefore important to clearly state and demonstrate these pitfalls in order to help researchers avoid them.

The rest of this review is organized as follows: Section 2 discusses the methodology used to survey and review the literature which defines the discussion framework used in all subsequent sections; Section 3 describes the data sets used; Section 4 provides a description of the output objective in the various papers; Sections 5 – 7 describe the input features used, common methods of pre-processing and the ML models used; Section 8 summarizes the results obtained

in various studies; and Section 9 then provides a discussion of the procedures used, specifically pointing out the pitfalls mentioned before towards obtaining over-estimated and unrealistic results. The section that follows concludes the paper.

2 Methodology

This chapter carries out an in-depth review of relevant literature to reveal the different practices authors take to predict rainfall. The review covers several aspects which relate to the input into, output from, and methods used in the various systems devised in the literature for this purpose. The review specifically focuses on studies that use supervised learning for both regression and classification problems.

Google scholar was used to collect papers from 2016 to 2020, with the following key words: ("machine learning" OR "deep learning") AND ("precipitation prediction" OR "rainfall prediction" OR "precipitation nowcasting"). Almost 1240 results were obtained, and of these only supervised rainfall prediction papers that used meteorological data from e.g. radar, satellites and stations were selected, while papers that used data from normal cameras e.g. photographs were excluded. Even though this review focuses on the prediction of rainfall, the methods used to achieve this can be extended and applied to other geophysical parameters like temperature and wind. Hence, the conclusions and discussions of this chapter can be adapted to other parameters.

The total number of reviewed papers are 66, which are a combination of conferences and journal papers published from 2016–2020, except for one paper [68] which was published in 2015 and is a seminal work in this field. Figure 1 shows the reviewed studies per year. Tables which summaries the reviewed paper can be found in Appendices A and B.

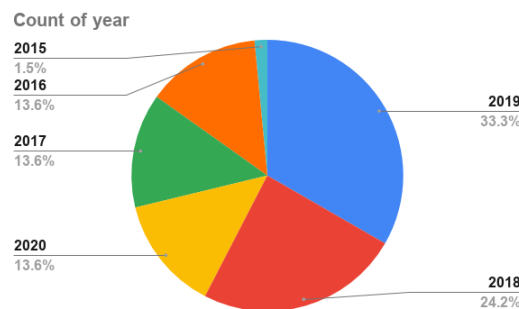


Fig. 1. Pie chart showing proportions by publication year for papers in this review .

Figure 2 shows the generic structure of supervised ML models. This structure was used as a guideline to construct a set of questions used to systematically categorize and analyze the 66 papers. The questions are as follows:

1. What data sets are used and where are they sourced?
2. What is the output objective in the various papers in terms of what the goal of prediction/forecasting?
3. What input features are extracted from the data set(s) to be used to achieve the output objective?
4. What pre-processing methods are used prior to classification/regression?
5. What ML models are used to achieve classification/regression towards the output objective?
6. What results were obtained from the above-mentioned steps, and how were they reported?

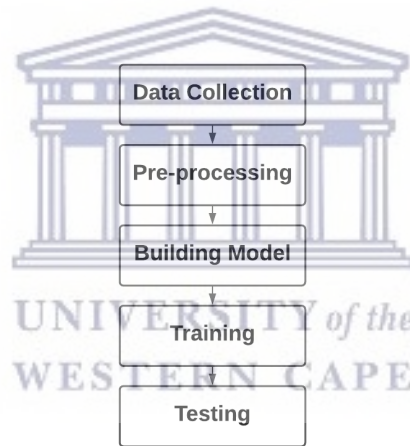


Fig. 2. Basic flow for building machine learning (ML) models [51]

These questions provide the framework for the rest of this paper. Sections 3–8 address questions 1–6 in sequence. Section 9 discusses the findings in the previous six sections, and Section 10 provides conclusions.

3 Data Sets

This section provides a breakdown of the data sets used in the 66 studies surveyed, based on the sources of the data sets, availability, and geographical locations where the data sets were collected.

Figure 3 (left) provides a breakdown of the studies based on the sources/availability of the data sets used in those studies. About 75% of the studies used private data,

sourced from meteorological stations of their prospective countries [60, 61, 84, 85, 82, 71, 55, 7, 27, 46, 38, 68, 72, 36, 6, 18, 75, 67, 78, 76, 69, 26, 16, 86, 28, 23, 20, 40, 62, 73, 4, 1, 63, 66, 8, 48, 10, 54, 2, 39, 31, 29, 80, 5, 19, 11, 14, 47, 79, 30, 49]. Most of these data sets are not readily available for use. Only 10% of the studies use data sourced from freely available sources such as Kaggle (www.kaggle.com), and the National Oceanic and Atmospheric Administration (NOAA) [64, 15, 56, 83, 59, 22, 5]. The remaining 13% of studies in this review use data from both private and publicly available sources [77, 81, 25, 17, 33, 12, 13, 41, 3].

Figure 3 (right) summarizes the geographical regions included in this review. The continent of Asia accounts for around 68% of all studies [61, 81, 82, 46, 38, 68, 64, 69, 12, 13, 73, 4, 8, 48, 22, 41, 10, 29, 19, 11, 47, 79, 85, 17, 27, 33, 36, 18, 75, 67, 15, 78, 76, 26, 28, 80, 30, 7, 23, 40, 62, 63, 39, 49]. Of these, studies that focus on China and India make up almost one quarter and one tenth respectively of all studies in this review. The remaining Asian studies focus on countries such as Iran, South Korea and Japan.

The rest of the chart is distributed as follows: the Americas make up 12.1% of studies [77, 71, 72, 56, 83, 16, 86, 14]; Europe accounts for 9.1% [25, 6, 20, 66, 54, 3]; Australia comprises 6.1% [55, 1, 2, 31], and the remaining 4.5% either involve multiple regions, or involve the use of the whole global map [60, 59, 5].

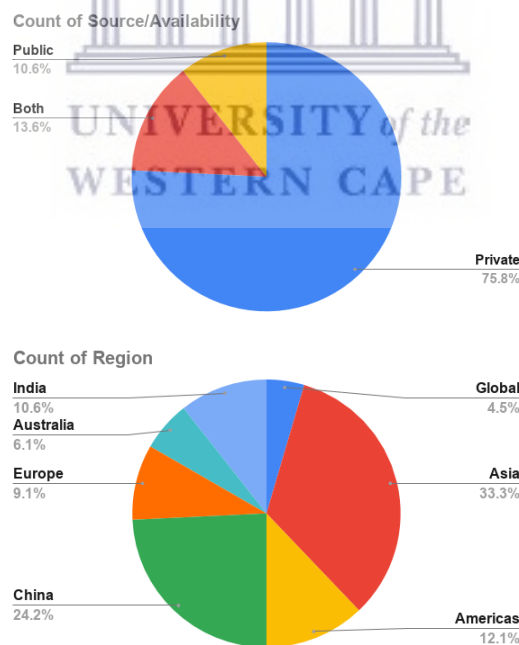


Fig. 3. Pie chart of the percentage of data sets in this survey in terms of source/availability (top) and geographical region (bottom).

4 Output Objectives

The output objectives of rainfall forecasting studies can be analyzed in terms of three factors: the forecasting time frame of the output; whether the output is continuous or discrete; and the dimensionality of the output. The forecasting time frame of the output specifies the time span of the forecast made, i.e. hourly, daily, monthly etc. The output can also be discrete (e.g. classification into “Rain”/“No Rain” classes), or continuous (e.g. predicting the quantity of rain), or both. Finally, the output can be 1-dimensional (1D) in the form of a single number or label representing a rainfall measure or category, or 2-dimensional (2D) in the form of a geospatial map of rainfall measures or categories on a grid of the geographical location under study.

In terms of the forecasting time frame, the studies can be broken down into those that make long-term predictions and those that focus on making short-term predictions. In this review, long-term prediction is defined as predictions of one months up to a year ahead, while short-term prediction can be a few minutes ahead (e.g. 5–15 minutes), up to one or more days ahead. Figure 4 (left), shows the distribution of papers’ forecasting time frames. Of the 66 reviewed papers, 30 papers (45%) make long-term predictions, the majority of which focus on monthly forecasting [20, 40, 62, 73, 4, 1, 63, 66, 8, 48, 22, 41, 10, 54, 2, 39, 31, 29, 80, 5, 14, 11, 19, 47, 79, 49, 3]. Only two studies focus on seasonal forecasting [28, 23], while a single study aims towards yearly forecasting [30]. As for studies that focus on short-term prediction, these are broken down nearly evenly between daily [60, 61, 82, 25, 71, 55, 7, 33, 15, 56, 86, 12, 13], hourly [77, 84, 85, 81, 17, 27, 26, 83, 16, 59], and one or more minutes ahead [46, 38, 68, 72, 36, 64, 6, 18, 75, 67, 78, 76, 69].

In terms of the type of output i.e. discrete (classification) or continuous (regression), Figure 4 (right) shows the distribution between the different output types. The majority carried out regressions to obtain continuous output [28, 23, 20, 40, 62, 73, 4, 1, 63, 66, 8, 48, 22, 41, 10, 54, 2, 39, 31, 29, 80, 5, 14, 11, 19, 47, 79, 60, 61, 77, 84, 85, 81, 75, 78, 69, 26], while slightly more than one third carried out classification into discrete classes [82, 25, 71, 55, 7, 17, 27, 46, 38, 33, 68, 72, 36, 64, 18, 67, 16, 59, 86, 12, 13, 30, 49, 3]. Only 3 studies applied both classification and regression [6, 76, 83].

For studies that applied classification, mostly carried out binary classification [82, 25, 71, 55, 7, 17, 27, 46, 38, 33, 68, 72, 36, 64, 18, 67, 59, 86, 12], with the majority of these classified into “Rain”/“No Rain” classes. Relatively fewer studies aim towards carrying out classification into multiple classes [33, 16, 13, 30, 49], varying from three to five classes.

Finally, for the dimensionality of the output, 54 out of 66 studies produce 1D output [28, 23, 8, 48, 22, 41, 10, 54, 2, 39, 31, 30, 49, 84, 85, 81, 82, 25, 71, 55, 3, 60, 61, 77, 7, 17, 27, 46, 38, 33, 56, 26, 83, 16, 59, 86, 12, 13], with the remaining 12 studies producing a series 2D images as output [68, 72, 36, 64, 6, 18, 75, 67, 15, 78, 76, 69]. Of the studies with 2D output, all except one [15] involve short-term prediction intervals of 10 minutes or less.

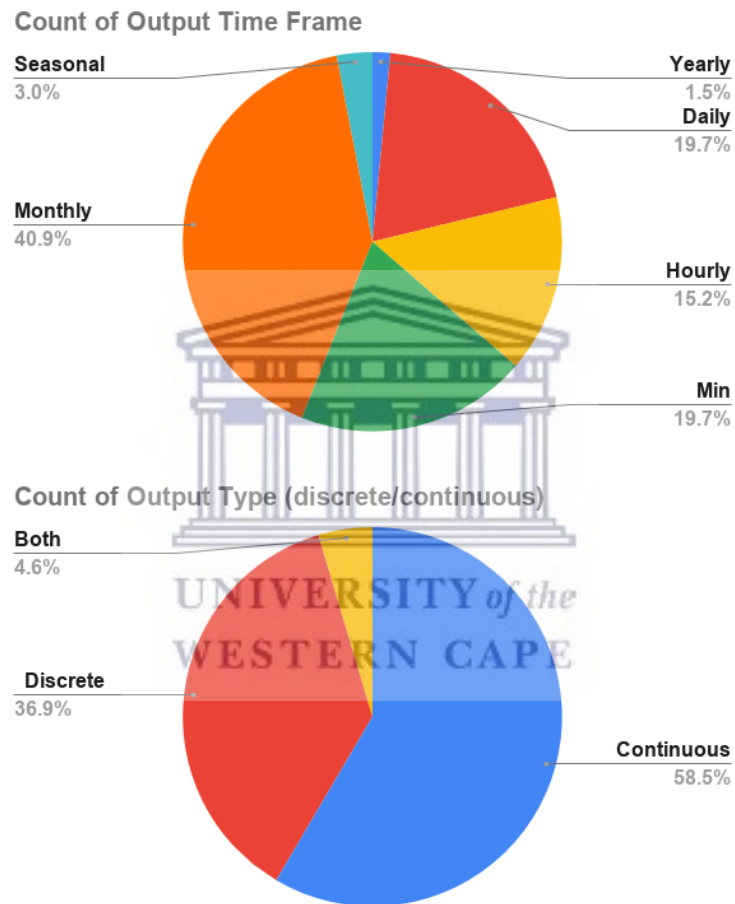


Fig. 4. Pie chart of the percentage of data sets in this survey in terms of forecasting time frame (top) and the discrete (classification)/continuous (regression) nature of the prediction output (bottom).

A connection between forecasting time frame and the discrete/continuous nature of the output can be observed. In general, studies involving longer-term predictions tend to make use of regression which produces continuous output, whereas on short-term time frames, studies tend towards using classification that gives discrete output. Specifically, 27 out of the 30 papers that focus on long-term prediction carry out regression [28, 23, 20, 40, 62, 73, 4, 1, 63, 66, 8, 48, 22, 41, 10, 54, 2, 39, 31, 29, 80, 5, 14, 11, 19, 47, 79], and 23 of the 36 papers that focus on short-term prediction carry out classification [82, 25, 71, 55, 7, 17, 27, 46, 38, 33, 68, 72, 36, 64, 6, 18, 67, 83, 16, 59, 86, 12, 13]. This relation may be explained by the fact that longer-term studies usually aim at predicting averages over several days (up to a month), while short-term studies predict instantaneous conditions. Multi-day averaged data assumes a continuous range of values, while in instantaneous rainfall datasets most values are null. It follows that classification into rain/no rain is useful for short term, but not for long-term prediction.

5 Input Features

In order to make future predictions, studies make use of data from one or more time steps (called “lags” or “time lags”) as input features to predict one or more future lags. For example, to predict rainfall at lag T , two previous time lags ($T - 1$) and ($T - 2$) may be used.

The actual input features in each lag vary across studies. In general, the input features used in the studies in this review were found to be of two types: 1D input features in which each time lag in the data set represents one or a set of geophysical parameters that have been collected at static known locations i.e. meteorological stations; and 2D input features in which each time lag in the data set is a 2D spatial map of values representing rainfall in the geographical area under review, usually collected by satellite or radar.

1D input features used include geophysical parameters such as temperature, humidity, wind speed and air pressure [62, 4, 8, 22, 3, 55, 46, 38, 43, 80]. In a smaller number of cases, climatic indices such as the Pacific Decadal Oscillation may also be used [28, 1, 41, 30, 77]. Studies that use 1D input features tend to use a relatively small number of overall input features, ranging from 2–12 features used for prediction.

With 2D input features, one or more images are taken as input features, depending on the number of time lags used as input e.g. two time lags used as input implies that two images are used as input. The number of time lags used as input is henceforth referred to as the “sequence length”.

There is no rule of thumb for how many time lags should be used as input, and this is mostly selected arbitrarily, and in fewer cases via trial and error. The vast majority of the studies under review select a fixed sequence length. The sequence length can be viewed as a hyper-parameter that affects the prediction outcome, but the optimization of this hyper-parameter is not investigated in the studies under review. The studies under review were found to be more focused on the

machine learning component, mostly at devising new deep learning architectures, than selecting and tuning other aspects of their systems.

The most common sequence lengths used are 5 frames [72, 36, 64, 78] and 10 frames [72, 36, 64, 78]. Other sequence lengths are also used, such as 2 [72], 4 [64], 7 [78] and 20 [36].

Studies that use 2D input features tend to use a relatively large number of input features. This can be attributed to the fact that the feature vectors produced are associated with one or more 2D images, resulting in vectors of size (Image width \times Image height \times Sequence length). Overall, the number of features can grow as high as several thousands.

Typically, 1D or 2D inputs are used to predict 1D or 2D outputs, respectively. As noted in the previous section, longer-term predictions tend to make 1D predictions, so it follows these studies also tend to use 1D data [28, 23, 20, 40, 62, 73, 4, 1, 63, 8, 48, 22, 41, 10, 54, 2, 39, 29, 80, 14, 11, 19, 47, 49, 3], while those that make shorter-term predictions tend towards the use of 2D data [68, 72, 36, 64, 6, 18, 75, 67, 15, 78, 76, 69, 13]

6 Input Data Pre-processing

Before ML tools are applied to make predictions on the available data, the input data is usually pre-processed to reformat the data into a form that will make training of, and prediction by, the ML tool(s) easier and faster. The pre-processing techniques usually applied in geophysical parameter forecasting can be broken down into three broad categories, namely data imputation; feature selection/reduction; and data preparation for classification. The following subsections describe these categories, as well as their application in the papers in this review.

6.1 Data Imputation

Data sets are regularly found to have missing data entries, which is caused by a range of factors such as data corruption, data sensor malfunction etc. This is a serious issue faced by researchers in data mining or analysis, and needs to be addressed as part of pre-processing before feature selection/preparation and training.

The techniques used to infer and substitute missing data are collectively referred to as data imputation techniques. Data imputation is challenging and is an on-going research area. In the papers in this review, it was found that very little focus was placed on this problem, with most of the studies making use of simple statistical techniques such as averaging to interpolate missing data entries [73, 31, 14, 11, 82, 55]. While not used in the papers in this review, more advanced data imputation techniques exist beyond the use of simple statistics, such as the use of ML to impute the data. The interested reader may refer to [74, 65, 57].

6.2 Feature Selection/Reduction

Feature selection/reduction aims to determine and use salient features in the data, and disregard irrelevant features in the data. This helps to reduce training time, decrease the model complexity and increase its performance. In the papers in this review, it is observed that feature selection is carried out either automatically or manually.

For automatic feature selection, various algorithms are used to determine the most salient features in the data. The most common method used in the papers in this review involved the use of deep learning techniques such as ANNs and convolutional neural networks (CNNs), to select/reduce features automatically, most especially when high-dimensional data such as radar and satellite images was used [68, 72, 36, 64, 18, 67, 15, 78, 76, 69, 59, 86, 12, 13, 31, 79]. The use of deep learning techniques was found to be much more common with short-term data sets which are generally much larger, therefore making it possible to achieve convergence on deep networks. Another category of ML tools used for automatic feature selection includes ensemble methods like random forests (RFs) which automatically order features in terms of importance, as used in [22, 11, 3, 77, 81, 82, 25, 71, 7]. Finally, principle components analysis (PCA) has also been used to reduce features in [28, 60, 85, 25, 27, 56].

As regards manual feature selection, researchers may either use prior experience and trial and error to manually select relevant features such as in [23, 20, 23, 20, 40, 73, 63, 66, 8, 54, 31, 29, 14], or use correlation analysis methods such as auto correlation to indirectly inform the manual feature selection process as in [28, 62, 1, 48, 22, 41, 2, 39, 30]. Where images are used, image cropping and resizing is applied to, respectively, dispose of irrelevant/static image segments and reduce the number of features [68, 72, 36, 64, 18, 67, 15, 78, 76, 69].

Manual feature selection is much more common with long-term data sets, with very few long-term prediction studies in this review making use of automatic feature selection methods. This is partly attributed to the relatively smaller amount of data available in these sets, as mentioned before, which makes it challenging, or even rules out, the application of e.g. deep learning methods for automatic feature selection.

The rotation of the earth around the sun can cause data to exhibit a seasonal behavior on an annual basis i.e. they exhibit annual periodicity [24, 9]. This is most prominent in long-term data sets and less prominent in shorter-term data sets. Addressing seasonality in long-term data sets is critical when traditional time series models are used, since these models assume stationarity [24, 53], while seasonality and trends in general makes time series non-stationary. Converting data from a non-stationary to a stationary state involves is a process of generating a time series with statistical properties that do not change over time. For further information about seasonal and non-stationary data sets and the conversion of non-stationary to stationary time series, the interested reader is referred to [53]. Another way to deal with seasonality is the inclusion of features that exhibit seasonal behavior, such as the usage of the same month previous year.

Figure 5 shows the methodologies used in the long-term prediction studies in this review. 11 of the 30 long-term papers (37%) did not address seasonality in the data [28, 23, 62, 63, 8, 22, 41, 47, 49, 3, 30], while the remaining 19 papers used some means of addressing seasonality in the data [11, 4, 66, 80, 48, 10, 14, 20, 40, 73, 1, 54, 2, 39, 31, 29, 5, 19, 79].

In the papers that addressed seasonality, four unique approaches were identified, and some were combined with others. The first approach involves including features from lag $T - 12$ (same month previous year) in the feature set used to predict rainfall at month T [48, 10, 14, 20, 40, 73, 1, 54, 2, 39, 31, 29, 5, 19, 79]. A less common approach is to use the index of the current month in the year (1=January, . . . , 12=December) as an input feature [31, 10].

Alternative approaches include performing time series decomposition, either using singular spectrum analysis as in [11] or wavelet transformation as in [4, 80, 66]. One paper [10] combined time series decomposition using singular spectrum analysis with the inclusion of features from lag $T - 12$ in the feature set. This has been included in the segment labelled “Combination” in Figure 5.

The final approach used to address seasonality takes the form of data de-seasonalization by subtracting the monthly averages from the data as in [19, 48]. All of the papers in this review that used this approach combined this subtraction with the first approach i.e. including features from lag $T - 12$ in the feature set. These two papers have also been included in the segment labelled “Combination” in Figure 5.

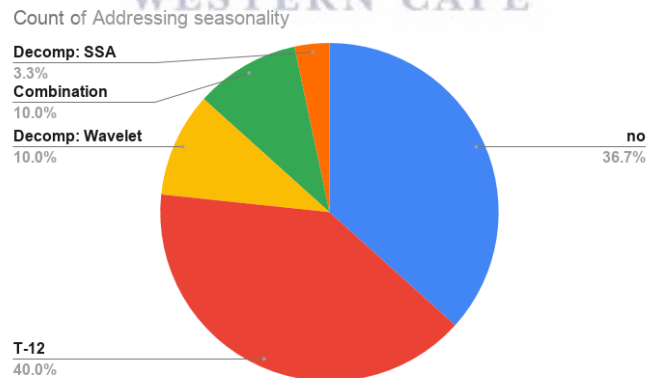


Fig. 5. Methods used to account for seasonality in studies with long-term data, by percentage.

6.3 Data Preparation for Classification

When attempting to carry out classification into discrete classes, it is either necessary to use a data set in which the desired output variable is discrete, or to convert a desired continuous-valued output variable into discrete classes. This involves setting the desired number of classes, which is usually done manually and arbitrarily, followed by determining the range of values represented by each class i.e. determining the thresholds that divide the continuous scale into the desired classes. Finally, where the number of instances across classes is imbalanced, it is necessary to balance them.

In the papers in this survey that carried out classification, most made use of data that was continuous, yet very few provide details on the process used to convert from a continuous to a discrete scale. Select studies in this survey that provide information about their data preparation process are described below.

In converting from continuous to discrete data, after manually specifying the number of classes (which has been explained in Section 4), studies automate the selection of the class thresholds using clustering tools, specifically k-means and k-medoids [16, 49, 3]. Another approach taken is to manually determine suitable thresholds, by performing a series of experiments to compare various threshold values [69]. To address any resulting class imbalances, researchers may perform random down-sampling to obtain an equal sample distribution across classes as in [3, 55, 46].

7 Machine Learning Techniques Used

The studies in this survey made use of a wide range of ML techniques which can be subdivided into two main groups: “classical” techniques such as multivariate linear regression (MLR), KNN ANNs, SVMs, and RF; and modern deep learning methods such as CNNs and Long-Short-Term-Memory (LSTM). It was observed that classical ML models tended to work with 1D data from meteorological stations, such as in [45, 28, 23, 4, 61, 81, 62] for short-term data and [28, 20, 40, 62, 1, 66, 48, 22, 47, 30, 3] for long-term data.

Some papers use hybrid models that combine two or more approaches. A popular hybrid approach is to combine ML with optimization tools such as genetics and particle swarm optimization to optimize hyper-parameters [47, 10, 48, 26, 27]. Multiple ML techniques are combined in [19, 71, 60], and ML is used with ARIMA in [61].

Deep learning models usually requires huge datasets to avoid overfitting on the data, which explains their popularity among short term data sets, especially those using 2D data [68, 72, 36, 64, 6, 18, 75, 67, 15, 78, 76, 69, 56, 26, 83, 16, 59, 86, 12, 13]. 2D data in particular has a huge feature space, which requires authors to implement automated feature reduction models like CNNs [83, 16, 59, 86, 12].

In order to accommodate the time dimension in the data, many researchers try to adapt time series models such as LSTMs for 1D data in [39, 29, 80, 5, 14, 84, 59]. For 2D data, models combining CNNs with LSTMs (designated as

ConvLSTMs models) were first used in in[68] in 2015, and subsequently several variations have been implemented [72, 36, 64, 6, 18, 75, 67, 15, 78, 76, 69].

8 Reporting of Results and Accuracy Measures

Several different metrics are used in the literature to measure the performance of the ML models according to the type of the problem. In classification problems, authors tend use metrics such as precision, recall, and accuracy [30, 49, 3, 82, 25, 71, 55, 17, 46, 38, 33, 72, 83, 16, 59, 86, 12]. If the data is not balanced then f1-score is used rather than the accuracy, since accuracy does not take the imbalance between the classes into account [82, 25, 71, 55, 72, 59]. For sequence classification prediction, other metrics are used such as the critical success (CSI) [64, 6, 18, 75, 67, 76, 69]. For continuous outputs, then the mean absolute error, and the root mean squared error are the most commonly used metrics in the literature [60, 61, 77, 84, 85, 81, 28, 23, 20, 40, 62, 73, 4, 1, 63, 66, 8, 48, 22, 41, 10, 54, 2, 39, 31, 29, 80, 5, 14, 11, 19, 47]

A direct comparison of these results across different papers is a nearly impossible task, since each paper uses its own models, pre-processing, metrics, data sets and parameters. However, individual authors frequently compare multiple algorithms, and there are a few ML algorithms that stand out as being most frequently mentioned as better performers. ANNs and deep learning are most frequently mentioned as best performing models, for both long-term prediction [28, 40, 4, 1, 63, 8, 2, 79, 5, 39, 31, 29, 14, 19, 47] and especially for short-term prediction [68, 72, 36, 64, 6, 18, 75, 67, 15, 78, 76, 69, 56, 26, 83, 16, 59, 86, 12, 13, 84, 85, 25, 38].

Other algorithms mentioned as best performers are SVMs in 6 studies [81, 46, 17, 27, 66, 48], ensemble in [77, 82, 62, 54, 80, 3], logistic regression in three studies [55, 7, 30] and KNNs in two studies [33, 20].

9 Discussion

The above sections clearly demonstrate that there is a robust, growing literature on rainfall prediction, which covers an extremely wide variety of time-scales, features used, pre-processing techniques, and ML algorithms used. From a high-level perspective, the field can be divided into short versus long time scales (time intervals of a one day or less, versus intervals of a month or more), which tend to have divergent characteristics.

Short term studies typically rely on huge datasets, and require deep learning applied to large feature sets to find hidden patterns in those datasets. On the other hand, long term studies rely more on pre-processing methods such as feature selection, data imputation, and data balancing in order to make effective predictions. ANNs and deep learning seem are becoming increasingly prevalent in long term studies as well as short term: since 2018, 7 of 23 papers on long-term prediction utilized deep learning tools.

There are reasons to regard the trend towards more complicated models with skepticism. Some recent studies have shown that much simpler models such as knns can sometimes outperform advanced ML techniques like RNNs, [42, 35, 44, 20]. Similar findings have been reported for other ML applications, such as the top n recommendation problem [21].

These results underscore the importance of providing simple but statistically well-motivated baselines to verify whether ML truly is effective in improving predictive accuracy. However, many papers do not provide simple baselines, but rather compare several variations or architectures of more advanced ML methods such as SVR or MLP [50, 11, 47, 14, 60, 17, 27, 46, 26, 83, 59, 12, 13]. Of the total reviewed papers, almost half (48.2%) of papers did not supply simple baselines. Of those papers that did supply baselines, a variety of methods was used. For short-term image data, the previous image is frequently used as an untrained predictor for the next image [75, 67, 76, 69]. For monthly data, some papers use MLR based on multiple previous lags [19, 28, 20, 40]; while same-month averages, though statistically well-motivated, are used much less frequently [80].

Besides the issue of baselining, the use of error bars is essential for comparison purposes, as it highlights whether the improvement obtained by the models are significant. Unfortunately most of the literature in ML does not provide error bars around the measured metrics. In the case of our reviewed literature shows that 88% of the papers did not give error bars.

A final issue of concern is data leakage. Data leakage refers to allowing data from the testing set to influence the training set. Data leakage occurs during the pre-processing of the data, and can take various forms as follows:

- Random shuffling, which involves choosing sequences from a common data pool for both training and testing;
- Imputation, which involves filling missing records using statistical methods on the entire data set (including both training and testing)
- De-seasonalization which utilizes the monthly averages from the entire data set.
- Using current lags, e.g. using temperature at a time T to predict rainfall at the same time T . (Depending on the application, this may or may not constitute data leakage)
- Combination: Which uses two of the above mentioned techniques.

Figure 6, shows the reviewed papers in terms of data leakage. The top chart focuses on long term data, where the bottom focuses on short term data. We mentioned previously that long term data often undergoes more pre-processing than short term data. This reflects on the graph, as leakage-producing methods are more than twice as common for long term as for short term. Random shuffling was performed in [28, 40, 41, 10, 29, 30, 49, 3] for long term data, and in [77, 27, 46, 56, 26, 16] for short term data. Data imputation was performed in [73, 31, 14, 11] for long term data, and in [55] for short term data. Faulty de-seasonalization was carried out in [48] for long term data. Using the current lags was seen only implemented in [62]. Multiple leakage issues (denoted as “combination” in the figure) were observed in [19, 82].

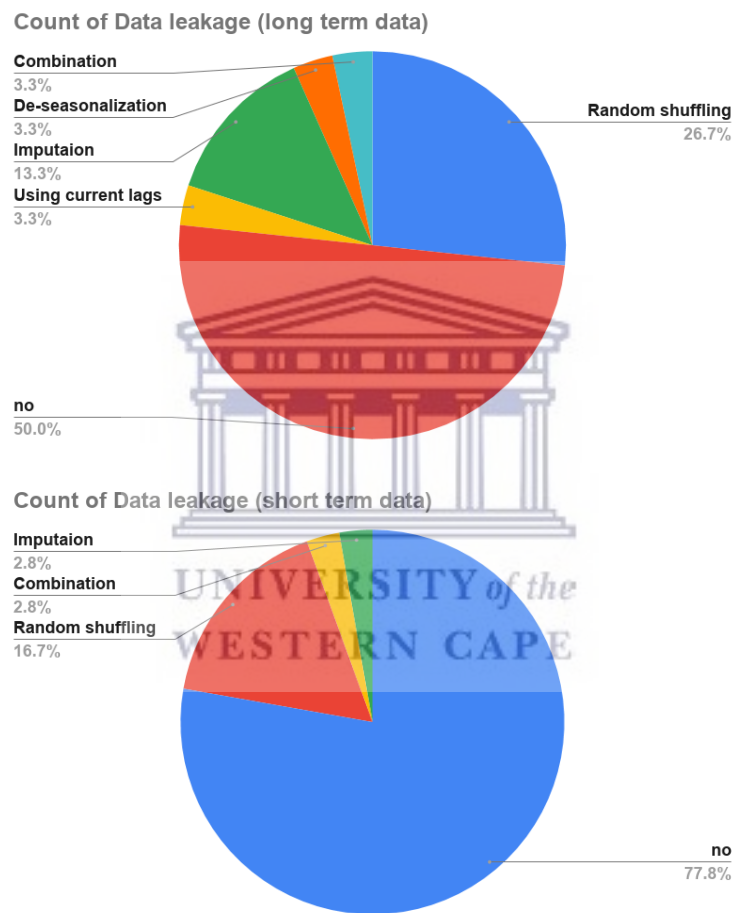


Fig. 6. Percentage of papers which introduced data leakage during pre-processing, for long term data (*top*) and short term data (*bottom*).

10 Conclusions

In the area of rainfall prediction, 66 relevant papers are reviewed, by examining the data source, output objective, input feature, pre-processing methods, models used, and finally the results. Different pre-processing like random shuffling used in the literature suggests that in some cases model performance is inaccurately represented. The aim of the survey is to make researches aware of the different pitfalls that can leads to unreal models performance, which does not only apply for rainfall, but for other time series data.

A Appendix: List of abbreviations

- ML Machine learning
- AD Author defined
- ANNs Artificial neural networks
- CNNs Convolution neural networks
- LSTMs Long short term memory
- ConvLSTMs Convolutions layers with Long short term memory
- RF Random forest
- SVMs Support vector machines
- DT Decision tress
- XGB Extreme gradient boosting
- LogReg Logistic regression
- MLR Multi linear regression
- KNNs K-nearest neighbour
- RMSE Root mean square error
- MAE mean absolute error
- CA Classification accuracy
- pre precision
- f1 f1-score
- PACF Partial autocorrelation function
- ACF Autocorrelation function
- PCA principle component analysis
- NOAA National Oceanic and Atmospheric Administration

B Appendix: Summary Tables for References

This appendix contains four tables which summarize the findings for the reviewed papers for long term data 1, 2, and short term data 3, 4. Tables 1 and 3 contain information regarding the source, period, region, input, output; while Tables 2 and 4 include information about the pre-processing tools, data leakage, and the ML used.

No.	Source	Period	Region	Input	Output	Ref
1	China Meteorological Administration (CMA)	1916-2015	China	6 climatic indices	Seasonal regression	[28]
2	Indian Institute of Tropical Meteorology (IITM)	1817-2016	India	8 past lags	Seasonal regression	[23]
3	Romanian rainfall	1991-2015	Romania	12 past lags	Monthly regression	[20]
4	Rainfall from the India Water Portal	1901-2002	India	11 climatic parameters	Monthly regression	[40]
5	Tuticorin meteorological station	1980-2002	India	Four climatic parameters	Monthly regression	[62]
6	Malaysian Department of Irrigation and Drainage	1965-2015	Malaysia	10 past lags	Monthly regression	[73]
7	National Cartographic Center of Iran (NCC)	1996-2010	Iran	Four climatic parameters	Monthly regression	[4]
8	Royal Netherlands Meteorological Institute Climate Explorer	2004-2014	Australia	Seven climatic indices	Monthly regression	[1]
9	Indian water portal	1901-2000	India	four climatic parameters	Monthly regression	[63]
10	Serbian meteorological stations	1946-2012	Serbia	past rainfall lags	Monthly regression	[66]
11	Iran meteorological department	2000-2010	Iran	Two Climatic parameters	Monthly regression	[8]
12	Iran meteorological department	1990-2014	Iran	four past lags	Monthly regression	[48]
13	CHIRPS, and NCEP-NCAR Reanalysis	1918-2001	Indus basin	5 climatic features	Monthly regression	[22]
14	World AgroMeteorological Information Service (WAMIS) and NOAA	1966-2017	South Korea	11 climatic indices	Monthly regression	[41]
15	Malaysian Department of Irrigation and Drainage	1950-2010	Malaysia	6 past lags and time stamp	Monthly regression	[10]
16	Turkish stations	2007-2016	Turkey	3 rainfall lags	Monthly regression	[54]

17	Australian stations	1885-2014	Australia	10 climatic indices and parameters	Monthly regression	[2]
18	Indian Meteorological Department	1871-2016	India	12 past lags	Monthly regression	[39]
19	Bureau of Meteorology (BOM), Royal Netherlands Meteorological Institute Climate, more	1908-2012	Australia	43 climatic indices and parameters	Monthly regression	[31]
20	Vietnam's hydrological gauging	1971-2010	Vietnam	12 features	Monthly regression	[29]
21	Global Precipitation Climatology Center (GPCC)	1901-2013	China	6-9 climatic indices and parameters	Monthly regression	[80]
22	Precipitation from NCEP	1979-2018	GLOBAL	164 past lags	Monthly regression	[5]
23	National Center of Hydrology and Meteorology Department (NCHM)	1997-2015	Bhutan	6 climates parameters	Monthly regression	[14]
24	Taiwan Water Resource Bureau	1958-2018	Taiwan	3 past lag	Monthly regression	[11]
25	Instituto de Hidrología, Meteorología y Estudios Ambientales (IDEAM) of Colombia	1983-2016	Colombia	6 past lags	Monthly regression	[19]
26	Islamic Republic of Iran Meteorological Organization (IRIMO)	1981- 2012	Iran	5 past lags	Monthly regression	[47]
27	Pluak Daeng Station in Thailand	1991-2016	Thailand	346 climatic indices and parameters	Monthly regression	[79]
28	National Climate Center of China Meteorological Administration (NCC-CMA)	1952-2012	China	84 climatic indices	Yearly Classification	[30]
29	The Department of Agricultural Meteorology Indira	2011-2013	India	five climatic parameters	Monthly Classification	[49]
30	meteorological stations of the island of Tenerife and NOAA databases	1976 - 2016	Tenerife Island	12 climatic indices and parameters	Monthly Classification	[3]

Table 1: Data sources, spatio-temporal coverage, inputs and outputs, and references for long-term predictive studies

No.	Pre-processing	Data leakage	ML used	Ref
1	Normalization, Random shuffling, feature correlation	Random shuffling	PCA-ANN, PCA-MLR	[28]
2	Normalization	no	Knns, ANNs, ELM	[23]
3	Windowing	no	Knns, ARIMA, ANNs	[20]
4	windowing, random shuffling	Random shuffling	ANN, ARMA, LR	[40]
5	Data imputation, noise removal, correlation analysis	Using current lags	DT, ANNs	[62]
6	Normalization, and data imputation	Imputation	ANNs, ARIMA	[73]
7	Normalization, Decomposition	no	WTANN, ANNs	[4]
8	features correlation	no	ANNs, POAMA	[1]
9	Normalization	no	Different ANNs	[63]
10	N/A	no	ANN, WT-SVM, GP	[66]
11	Normalization, optimization	no	AD-MLP, AD-SVM, DT	[8]
12	correlation analysis (PACF), square root transformation, standardization, de-seasonalization	De-seasonalization	SVR, AD-SVR, more	[48]
13	feature correlation, random shuffling	no	MLP, SVR, MLR, RF, Knns	[22]
14	feature correlation, random shuffling	Random shuffling	ANNs	[41]
15	Decomposition	Random shuffling	AD-MLP	[10]
16	normalization	no	Ensemble method, SVM, ANNs, more	[54]
17	Feature selection	no	ANNs, POAMA	[2]
18	Feature correlation, windowing	N/A	LSTM, RNN	[39]
19	Data imputation, normalization	Imputation	1D-CNN, MLP, baseline (ACCESS-S1)	[31]
20	random shuffling	Random shuffling	MLP, LSTM, SNN	[29]

21	Normalization, Wavelet	no	MLR, MLP, LSTM, SVMs, ConvLSTMs, ensemble methods	[80]
22	greyscale, windowing	no	LSTM, ConvNet	[5]
23	Normalization, data imputation	Imputation	MLR, AD-LSTM, LSTM, MLP	[14]
24	Decomposition	Imputation	UD-RF, RF, UD-SVR, SVR	[11]
25	Imputation, de-seasonalization	Imputation, de-seasonalization	3 AD-ANNs models	[19]
26	Normalization	no	ANNs, AD-ANNs, AD-gene expression programming	[47]
27	N/A	no	DNNs	[79]
28	feature correlation, feature reduction	Random shuffling	MLogR	[30]
29	Clustering	Random shuffling	GPR, DT, NB	[49]
30	Random Down sampling, feature correlation	Random shuffling	XGB, RF, more	[3]

Table 2: Pre-processing, data leakage characteristics, machine learning algorithms used, and reference numbers for long term predictive studies



C

No.	Source	Period	Region	Input	Output	Ref
1	Indian Statistical Institute	1989-1995	Multiple regions	10 climatic parameters	Daily Regression	[60]
2	Vietnamese stations	1978-2016	Vietnam	previous lags	Daily Regression	[61]
3	Meteoblue Data , MODIS, and more	2012-2014	Colombia	12 climatic indices and parameters	Hourly Regression	[77]
4	Central Meteorological Observatory of Shanghai	2015-2017	China	24 climatic parameters	Hourly Regression	[84]
5	China Meteorological Administration	2015-2017	China	13 climatic parameters	Hourly Regression	[85]
6	Taiwan and the National Severe Storms Laboratory and NOAA	2012-2015	Taiwan	3-4 parameters	Hourly Regression	[81]
7	Meteorological Drainage and the Irrigation departments in Malaysia	2010-2014.	Malaysia	4 parameters	Daily Classification	[82]
8	The water planing and managing ageny for Tenerife Island, and NOAA	1979-2015	Spain	1800 parameters	Daily Classification	[25]
9	U.S. Government's open data	2010-2017	US	25 parameters	Daily Classification	[71]
10	Kaggle and the australian government	2008-2017	Australia	23 parameters	Daily Classification	[55]
11	Indian Meteorological Department	2008-2017	India	8 parameters	Daily Classification	[7]
12	satellite imagery data are from FY-2G, and meteorological station located in Shenzhen	2015	China	8 parameters	Hourly Classification	[17]
13	Data from the Nanjing Station	N/A	China	6 parameters	Hourly Classification	[27]
14	Singapore related weather stations	2012-2015	Singapore	15 climatic parameters	Min Classification	[46]
15	Japan Meteorological Agency	2000-2012	Japan	8 features	Min Classification	[38]
16	NCEP-NCAR and Beijing Meteorological station	1990-2012	China	6 climatic indices and parameters	Daily Classification	[33]
17	Radar images collected in Hong Kong	2011-2013	Hong Kong	5 frames	Min Classification	[68]

18	Radar images from USA from 2008-2015	2008-2015	US	10 frames	Min Classification	[72]
19	Radar images from National Meteorological Information Center	2016-2017	China	10 frames	Min Classification	[36]
20	Radar images are retrieved using Yahoo! Static Map API	2013-2017	Japan	10 frames	Min Classification	[64]
21	Radar images from the German Weather Service (DWD)	2006-2017	Germany	2 frames	Min Both	[6]
22	Weather Surveillance Radar-1988 Doppler Radar (WSR-88D)	2015-2018	China	20 frames	Min Classification	[18]
23	CIKM AnalytiCup 2017 competition	N/A	China	5 frames	Min Regression	[75]
24	CINRAD-SA type Doppler weather radar	2016	China	4 frames	Min Classification	[67]
25	CHIRPS	1918-2019	China	5 frames	Daily Regression	[15]
26	Radar images collected in Hong Kong	2011-2013	China	10 frames	Min Regression	[78]
27	CIKM AnalytiCup 2017 competition	N/A	China	7 frames	Min Both	[76]
28	dataset from HKO-7	2009-2015	Hong Kong	5 frames	Min Regression	[69]
29	NCEP, and NOAA	1979-2017	US	A tensor of $8 \times 4 \times 25 \times 25$	Daily Regression	[56]
30	China meteorological data network	N/A	China	7 climatic parameters	Hourly Regression	[26]
31	NOAA	1800-2017	US	30 climatic parameters	Hourly Both	[83]
32	Large Ensemble (LENS) community project	1920-2005	US	$3 \times 28 \times 28 \times 3$	Hourly Classification	[16]
33	Kaggle	2012-2017	US and India	120 climatic lags	Hourly Classification	[59]
34	Iowa state	1948-2010	USA	9 climatic parameters	Daily Classification	[86]
35	Meteorological Department of Thailand and the Petroleum Authority of Thailand	2017-2017	Thailand	one image	Daily Classification	[12]
36	Meteorological Department of Thailand and the Petroleum Authority of Thailand	2017-2018	Thailand	one and batch of images	Daily Classification	[13]

Table 3: Data sources, spatio-temporal coverage, inputs and outputs, and references for short-term predictive studies

No.	Pre-processing	Data leakage	ML used	Ref
1	normalization, cross validation, feature reduction (PCA)	no	AD-ELM	[60]
2	normalization, feature correlation	no	ARIMA-MLP, ARIMA-SVM, ARIMA-HW, ARIMA-NF, more	[61]
3	data imputation, data shuffling	Random shuffling	RF, Cubist	[77]
4	feature selection, correlation analysis, Interpolation, clustering	no	LSTM, MLR, SVMs, ECM-FWF	[84]
5	normalization, feature reduction (PCA)	no	DRCF, ARIMA, more	[85]
6	N/A	no	RF, SVM	[81]
7	Normalization, data imputation, shuffling	Data imputation, Random shuffling	SVM, RF, DT, NB, ANN	[82]
8	Feature reduction (PCA)	no	ANNs, RF, Knns, LogR	[25]
9	Feature selection (RF), k-fold cross validation	no	RF, AD[ANNs, Adaboost, SVM, KNN]	[71]
10	Feature selection, Feature correlation, data imputation, over, and down sampling	Imputation	LogR, DT, Knns, more	[55]
11	N/A	no	LogReg, DT, RF, more	[7]
12	Radiometric, and geometric correction, and windowing	no	SVM	[17]
13	Normalization, random shuffling	Random shuffling	AD-SVMs	[27]
14	Down-sampling, feature correlation	Random shuffling	SVM	[46]
15	outliers removal, normalization	no	MLP, RBFN	[38]
16	Normalization	no	Knns	[33]
17	Feature reduction, noise removal, windowing	no	ConvLSTM, FC-LSTM, more	[68]
18	Resizing, windowing	no	Eulerian persistence, AD-Conv-RNN, Conv-LSTM	[72]

19	Feature reduction, windowing	no	MLC-LSTM, ConvLSTM, more	[36]
20	Feature reduction, windowing	no	SDPredNet, TrajGRU, more	[64]
21	logarithmic transformation	no	Optical flow, Dozhdy.Net	[6]
22	Noise removal, remove corrupted images, windowing, Normalization	no	COTREC, ConvLSTM, AD-ConvLSTM, more	[18]
23	Normalization, windowing	no	Last frame, TrajGRU, ConvLSTM, AD-TrajGRU, more	[75]
24	Windowing, greyscale transformation	no	Last input, COTREC, AD-CNN	[67]
25	Windowing, grey-scale, resizing	no	ConvLSTM, AD-ConvLSTMs	[15]
26	Windowing, grey-scale, resizing	no	ConvLSTM, PredRNN, VPN-baseline	[78]
27	Windowing, grey-scale, resizing, data augmentation	no	ConvLSTM, ConvGRU, TrajGRU, PredRNN, PredRNN++, last frame	[76]
28	Windowing, grey-scale, noise removal, normalization	no	2D CNN, 3D CNN, ConvGRU, TrajGRU, last frame, more	[69]
29	Normalization, random shuffling	Random shuffling	LR, CNNs, base model (NARR)	[56]
30	Random shuffling and Normalization	Random shuffling	DBN, GA-SVM, more	[26]
31	N/A	no	CNN, LPBoost, more	[83]
32	clustering, down-sampling, random shuffling	Random shuffling	CNN, LogReg	[16]
33	normalization	no	CNN, LSTM	[59]
34	cropping	no	CNN	[86]
35	cropping	N/A	CNN	[12]
36	cropping	N/A	CNN	[13]

Table 4: Pre-processing, data leakage characteristics, machine learning algorithms used, and reference numbers for short term predictive studies

References

1. John Abbot and Jennifer Marohasy. Forecasting monthly rainfall in the western australian wheat-belt up to 18-months in advance using artificial neural networks. In *Australasian Joint Conference on Artificial Intelligence*, pages 71–87. Springer, 2016.
2. John Abbot and Jennifer Marohasy. Application of artificial neural networks to forecasting monthly rainfall one year in advance for locations within the murray darling basin, australia. *International Journal of Sustainable Development and Planning*, 12(8):1282–1298, 2017.
3. Ricardo Aguasca-Colomo, Dagoberto Castellanos-Nieves, and Máximo Méndez. Comparative analysis of rainfall prediction models using machine learning in islands with complex orography: Tenerife island. *Applied Sciences*, 9(22):4931, 2019.
4. Mohammad Arab Amiri, Yazdan Amerian, and Mohammad Saadi Mesgari. Spatial and temporal monthly precipitation forecasting using wavelet transform and neural networks, qara-qum catchment, iran. *Arabian Journal of Geosciences*, 9(5):421, 2016.
5. S Aswin, P Geetha, and R Vinayakumar. Deep learning models for the prediction of rainfall. In *2018 International Conference on Communication and Signal Processing (ICCSP)*, pages 0657–0661. IEEE, 2018.
6. G Ayzel, M Heistermann, A Sorokin, O Nikitin, and O Lukyanova. All convolutional neural networks for radar-based precipitation nowcasting. *Procedia Computer Science*, 150:186–192, 2019.
7. MS Balamurugan and R Manojkumar. Study of short term rain forecasting using machine learning based approach. *Wireless Networks*, pages 1–6, 2019.
8. Fatemeh Barzegari Banadkooki, Mohammad Ehteram, Ali Najah Ahmed, Chow Ming Fai, Haitham Abdulmohsin Afan, Wani M Ridwan, Ahmed Sefelnasr, and Ahmed El-Shafie. Precipitation forecasting using multilayer neural network and support vector machine optimization based on flow regime algorithm taking into account uncertainties of soft computing models. *Sustainability*, 11(23):6681, 2019.
9. Adrian G Barnett, Peter Baker, and Annette Dobson. Analysing seasonal data. *R Journal*, 4(1):5–10, 2012.
10. Zahra Beheshti, Morteza Firouzi, Siti Mariyam Shamsuddin, Masoumeh Zibarzani, and Zulkifli Yusop. A new rainfall forecasting model using the capso algorithm and an artificial neural network. *Neural Computing and Applications*, 27(8):2551–2565, 2016.
11. Pa Ousman Bojang, Tao-Chang Yang, Quoc Bao Pham, and Pao-Shan Yu. Linking singular spectrum analysis and machine learning for monthly rainfall forecasting. *Applied Sciences*, 10(9):3224, 2020.
12. Kitinan Boonyuen, Phisan Kaewprapha, and Patchanok Srivihok. Daily rainfall forecast model from satellite image using convolution neural network. In *2018 IEEE International Conference on Information Technology*, pages 1–7, 2018.
13. Kitinan Boonyuen, Phisan Kaewprapha, Uruya Weesakul, and Patchanok Srivihok. Convolutional neural network inception-v3: A machine learning approach for leveling short-range rainfall forecast model from satellite image. In *International Conference on Swarm Intelligence*, pages 105–115. Springer, 2019.
14. Teresita Canchala, Wilfredo Alfonso-Morales, Yesid Carvajal-Escobar, Wilmar L Cerón, and Eduardo Caicedo-Bravo. Monthly rainfall anomalies forecasting for southwestern colombia using artificial neural networks approaches. *Water*, 12(9):2628, 2020.

15. Rafaela Castro, Yania M Souto, Eduardo Ogasawara, Fabio Porto, and Eduardo Bezerra. Stconvs2s: Spatiotemporal convolutional sequence to sequence network for weather forecasting. *Neurocomputing*, 2020.
16. Ashesh Chattopadhyay, Pedram Hassanzadeh, and Saba Pasha. Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data. *Scientific Reports*, 10(1):1–13, 2020.
17. Kai Chen, Jun Liu, Shanxin Guo, Jinsong Chen, Ping Liu, Jing Qian, Huijuan Chen, and Bo Sun. Short-term precipitation occurrence prediction for strong convective weather using fy2-g satellite data: a case study of shenzhen, south china. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41:215, 2016.
18. Lei Chen, Yuan Cao, Leiming Ma, and Junping Zhang. A deep learning based methodology for precipitation nowcasting with radar. *Earth and Space Science*, page e2019EA000812, 2020.
19. Manoj Chhetri, Sudhanshu Kumar, Partha Pratim Roy, and Byung-Gyu Kim. Deep blstm-gru model for monthly rainfall prediction: A case study of simtokha, bhutan. *Remote Sensing*, 12(19):3174, 2020.
20. Marinoiu Cristian et al. Average monthly rainfall forecast in romania by using k-nearest neighbors regression. *Analele Universității Constantin Brâncuși din Târgu Jiu: Seria Economie*, 1(4):5–12, 2018.
21. Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, pages 101–109, 2019.
22. Hamidreza Ghasemi Damavandi and Reepal Shah. A learning framework for an accurate prediction of rainfall rates. *arXiv preprint arXiv:1901.05885*, 2019.
23. Yajnaseni Dash, Saroj K Mishra, and Bijaya K Panigrahi. Rainfall prediction for the kerala state of india using artificial intelligence approaches. *Computers & Electrical Engineering*, 70:66–73, 2018.
24. Jacques W Delleur and M Levent Kavvas. Stochastic models for monthly rainfall forecasting and synthetic generation. *Journal of Applied Meteorology*, 17(10):1528–1536, 1978.
25. Javier Diez-Sierra and Manuel del Jesus. Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods. *Journal of Hydrology*, page 124789, 2020.
26. Jinglin Du, Yayun Liu, and Zhijun Liu. Study of precipitation forecast based on deep belief networks. *Algorithms*, 11(9):132, 2018.
27. Jinglin Du, Yayun Liu, Yanan Yu, and Weilan Yan. A prediction of precipitation data based on support vector machine and particle swarm optimization (pso-svm) algorithms. *Algorithms*, 10(2):57, 2017.
28. Yiheng Du, Ronny Berndtsson, Dong An, Linus Zhang, Feifei Yuan, Cintia Bertacchi Uvo, and Zhenchun Hao. Multi-space seasonal precipitation prediction model applied to the source region of the yangtze river, china. *Water*, 11(12):2440, 2019.
29. Tran Anh Duong, Minh Duc Bui, and Peter Rutschmann. A comparative study of three different models to predict monthly rainfall in ca mau, vietnam. In *Wasserbau-Symposium Graz 2018. Wasserwirtschaft–Innovation aus Tradition. Tagungsband. Beiträge zum 19. Gemeinschafts-Symposium der Wasserbau-Institute TU München, TU Graz und ETH Zürich*, pages Paper–G5, 2018.
30. Lihao Gao, Fengying Wei, Zhongwei Yan, Jin Ma, and Jiangjiang Xia. A study of objective prediction for summer precipitation patterns over eastern china based on a multinomial logistic regression model. *Atmosphere*, 10(4):213, 2019.

31. Ali Haidar and Brijesh Verma. Monthly rainfall forecasting using one-dimensional deep convolutional neural network. *IEEE Access*, 6:69053–69063, 2018.
32. Kyaw Kyaw Htike and Othman O Khalifa. Rainfall forecasting models using focused time-delay neural networks. In *International Conference on Computer and Communication Engineering (ICCCCE'10)*, pages 1–6. IEEE, 2010.
33. Mingming Huang, Runsheng Lin, Shuai Huang, and Tengfei Xing. A novel approach for precipitation forecast via improved k-nearest neighbor algorithm. *Advanced Engineering Informatics*, 33:89–95, 2017.
34. Nguyen Q Hung, Mukand S Babel, S Weesakul, and NK Tripathi. An artificial neural network model for rainfall forecasting in bangkok, thailand. *Hydrology and Earth System Sciences*, 13(8):1413–1425, 2009.
35. Eslam Hussein, Mehrdad Ghaziasgar, and Christopher Thron. Regional rainfall prediction using support vector machine classification of large-scale precipitation maps. In *2020 IEEE 23rd International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE, 2020.
36. Jinrui Jing, Qian Li, and Xuan Peng. Mlc-lstm: Exploiting the spatiotemporal correlation between multi-level weather radar echoes for echo sequence extrapolation. *Sensors*, 19(18):3988, 2019.
37. Hassan A Karimi. *Big Data: techniques and technologies in geoinformatics*. Crc Press, 2014.
38. Tomoaki Kashiwao, Koichi Nakayama, Shin Ando, Kenji Ikeda, Moonyong Lee, and Alireza Bahadori. A neural network-based local rainfall prediction system using meteorological data on the internet: A case study using data from the japan meteorological agency. *Applied Soft Computing*, 56:317–330, 2017.
39. Deepak Kumar, Anshuman Singh, Pijush Samui, and Rishi Kumar Jha. Forecasting monthly precipitation using sequential modelling. *Hydrological sciences journal*, 64(6):690–700, 2019.
40. K Lakshmaiah, S Murali Krishna, and B Eswara Reddy. Application of referential ensemble learning techniques to predict the density of rainfall. In *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*, pages 233–237. IEEE, 2016.
41. Jeongwoo Lee, Chul-Gyum Kim, Jeong Eun Lee, Nam Won Kim, and Hyeonjun Kim. Application of artificial neural networks to rainfall forecasting in the geum river basin, korea. *Water*, 10(10):1448, 2018.
42. Jimmy Lin. The neural hype and comparisons against weak baselines. In *ACM SIGIR Forum*, volume 52, pages 40–51. ACM New York, NY, USA, 2019.
43. Jing Lu, Wei Hu, and Xiakun Zhang. Precipitation data assimilation system based on a neural network and case-based reasoning system. *Information*, 9(5):106, 2018.
44. Malte Ludewig and Dietmar Jannach. Evaluation of session-based recommendation algorithms. *User Modeling and User-Adapted Interaction*, 28(4-5):331–390, 2018.
45. M Mallika and M Nirmala. Chennai annual rainfall prediction using k-nearest neighbour technique. *International Journal of Pure and Applied Mathematics*, 109(8):115–120, 2016.
46. Shilpa Manandhar, Soumyabrata Dev, Yee Hui Lee, Yu Song Meng, and Stefan Winkler. A data-driven approach for accurate rainfall prediction. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):9323–9331, 2019.
47. Saeid Mehdizadeh, Javad Behmanesh, and Keivan Khalili. New approaches for estimation of monthly rainfall based on gep-arch and ann-arch hybrid models. *Water resources management*, 32(2):527–545, 2018.

48. A Danandeh Mehr, Vahid Nourani, V Karimi Khosrowshahi, and Moahmmad Ali Ghorbani. A hybrid support vector regression–firefly model for monthly rainfall forecasting. *International Journal of Environmental Science and Technology*, 16(1):335–346, 2019.
49. Niharika Mishra and Ajay Kushwaha. Rainfall prediction using gaussian process regression classifier. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 8(8), 2019.
50. S Mohamadi, M Ehteram, and A El-Shafie. Accuracy enhancement for monthly evaporation predicting model utilizing evolutionary machine learning methods. *International Journal of Environmental Science and Technology*, pages 1–24, 2020.
51. Amir Mosavi, Pinar Ozturk, and Kwok-wing Chau. Flood prediction using machine learning models: Literature review. *Water*, 10(11):1536, 2018.
52. Mohsen Nasser, Keyvan Asghari, and MJ Abedini. Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network. *Expert systems with applications*, 35(3):1415–1421, 2008.
53. Aileen Nielsen. *Practical time series analysis: prediction with statistics and machine learning*. O’Reilly, 2020.
54. Vahid Nourani, Selin Uzelaltinbulat, Fahreddin Sadikoglu, and Nazanin Behfar. Artificial intelligence based ensemble modeling for multi-station prediction of precipitation. *Atmosphere*, 10(2):80, 2019.
55. Nikhil Oswal. Predicting rainfall using machine learning techniques. *arXiv preprint arXiv:1910.13827*, 2019.
56. Baoxiang Pan, Kuolin Hsu, Amir AghaKouchak, and Soroosh Sorooshian. Improving precipitation estimation using convolutional neural network. *Water Resources Research*, 55(3):2301–2321, 2019.
57. Adam Pantanowitz and Tshilidzi Marwala. Missing data imputation through the use of the random forest algorithm. In *Advances in Computational Intelligence*, pages 53–62. Springer, 2009.
58. Aakash Parmar, Kinjal Mistree, and Mithila Sompura. Machine learning techniques for rainfall prediction: A review. In *International Conference on Innovations in Information Embedded and Communication Systems*, 2017.
59. Maitreya Patel, Anery Patel, Dr Ghosh, et al. Precipitation nowcasting: Leveraging bidirectional lstm and 1d cnn. *arXiv preprint arXiv:1810.10485*, 2018.
60. Yuzhong Peng, Huasheng Zhao, Hao Zhang, Wenwei Li, Xiao Qin, Jianping Liao, Zhiping Liu, and Jie Li. An extreme learning machine and gene expression programming-based hybrid model for daily precipitation prediction. *International Journal of Computational Intelligence Systems*, 12(2):1512–1525, 2019.
61. Quoc Bao Pham, Sani Isah Abba, Abdullahi Garba Usman, Nguyen Thi Thuy Linh, Vivek Gupta, Anurag Malik, Romulus Costache, Ngoc Duong Vo, and Doan Quang Tri. Potential of hybrid data-intelligence algorithms for multi-station modelling of rainfall. *Water Resources Management*, 33(15):5067–5087, 2019.
62. N Ramsundram, S Sathya, and S Karthikeyan. Comparison of decision tree based rainfall prediction model with data driven model considering climatic variables. *Irrigation and Drainage Systems Engineering*, 2016.
63. Kaushik D Sardeshpande and Vijaya R Thool. Rainfall prediction: A comparative study of neural network architectures. In *Emerging Technologies in Data Mining and Information Security*, pages 19–28. Springer, 2019.
64. Ryoma Sato, Hisashi Kashima, and Takehiro Yamamoto. Short-term precipitation prediction with skip-connected prednet. In *International Conference on Artificial Neural Networks*, pages 373–382. Springer, 2018.

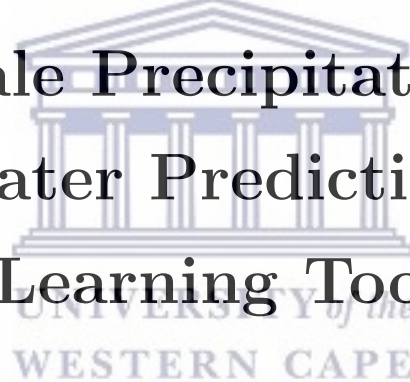
65. Anoop D Shah, Jonathan W Bartlett, James Carpenter, Owen Nicholas, and Harry Hemingway. Comparison of random forest and parametric imputation models for imputing missing data using mice: a caliber study. *American journal of epidemiology*, 179(6):764–774, 2014.
66. Mohamed Shenify, Amir Seyed Danesh, Milan Gocić, Ros Surya Taher, Ainuddin Wahid Abdul Wahab, Abdullah Gani, Shahaboddin Shamshirband, and Dalibor Petković. Precipitation estimation using support vector machine with discrete wavelet transform. *Water resources management*, 30(2):641–652, 2016.
67. En Shi, Qian Li, Daquan Gu, and Zhangming Zhao. Convolutional neural networks applied on weather radar echo extrapolation. *DEStech Transactions on Computer Science and Engineering*, (csae), 2017.
68. Xingjian Shi, Zhourong Chen, Hao Wang, D. Yeung, W. Wong, and Wang chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *ArXiv*, abs/1506.04214, 2015.
69. Xingjian Shi, Zhihan Gao, Leonard Lausen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun Woo. Deep learning for precipitation nowcasting: A benchmark and a new model. In *Advances in neural information processing systems*, pages 5617–5627, 2017.
70. Xingjian Shi and Dit-Yan Yeung. Machine learning for spatiotemporal sequence forecasting: A survey. *arXiv preprint arXiv:1808.06865*, 2018.
71. Gurpreet Singh and Deepak Kumar. Hybrid prediction models for rainfall forecasting. In *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 392–396. IEEE, 2019.
72. Sonam Singh, Sudeshna Sarkar, and Pabitra Mitra. Leveraging convolutions in recurrent neural networks for doppler weather radar echo prediction. In *International Symposium on Neural Networks*, pages 310–317. Springer, 2017.
73. Junaida Sulaiman and Siti Hajar Wahab. Heavy rainfall forecasting model using artificial neural network for flood prone area. In *IT Convergence and Security 2017*, pages 68–76. Springer, 2018.
74. Fei Tang and Hemant Ishwaran. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 10(6):363–377, 2017.
75. Quang-Khai Tran and Sa-kwang Song. Computer vision in precipitation nowcasting: Applying image quality assessment metrics for training deep neural networks. *Atmosphere*, 10(5):244, 2019.
76. Quang-Khai Tran and Sa-kwang Song. Multi-channel weather radar echo extrapolation with convolutional recurrent neural networks. *Remote Sensing*, 11(19):2303, 2019.
77. Cristian Valencia-Payan and Juan Carlos Corrales. A rainfall prediction tool for sustainable agriculture using random forest. In *Mexican International Conference on Artificial Intelligence*, pages 315–326. Springer, 2018.
78. Yunbo Wang, Mingsheng Long, Jianmin Wang, Zhifeng Gao, and S Yu Philip. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *Advances in Neural Information Processing Systems*, pages 879–888, 2017.
79. Uruya Weesakul, Pisan Kaewprapha, Kitinan Boonyuen, and Ole Mark. Deep learning neural network: A machine learning approach for monthly rainfall forecast, case study in eastern region of thailand. *Engineering and Applied Science Research*, 45(3):203–211, 2018.
80. Lei Xu, Nengcheng Chen, Xiang Zhang, and Zeqiang Chen. A data-driven multi-model ensemble for deterministic and probabilistic precipitation forecasting at seasonal scale. *Climate Dynamics*, pages 1–20, 2020.

81. Pao-Shan Yu, Tao-Chang Yang, Szu-Yin Chen, Chen-Min Kuo, and Hung-Wei Tseng. Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. *Journal of Hydrology*, 552:92–104, 2017.
82. Suhaila Zainudin, Dalia Sami Jasim, and Azuraliza Abu Bakar. Comparative analysis of data mining techniques for malaysian rainfall prediction. *International Journal on Advanced Science, Engineering and Information Technology*, 6(6):1148–1153, 2016.
83. Choujun Zhan, Fujian Wu, Zhengdong Wu, and K Tse Chi. Daily rainfall data construction and application to weather prediction. In *2019 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5. IEEE, 2019.
84. Chang-Jiang Zhang, Jing Zeng, Hui-Yuan Wang, Lei-Ming Ma, and Hai Chu. Correction model for rainfall forecasts using the lstm with multiple meteorological factors. *Meteorological Applications*, 27(1):e1852, 2020.
85. Pengcheng Zhang, Yangyang Jia, Jerry Gao, Wei Song, and Hareton KN Leung. Short-term rainfall forecasting using multi-layer perceptron. *IEEE Transactions on Big Data*, 2018.
86. WY Zhuang and Wei Ding. Long-lead prediction of extreme precipitation cluster via a spatiotemporal convolutional neural network. In *Proceedings of the 6th International Workshop on Climate Informatics: CI*, 2016.



Chapter 3

Manuscript “Regional Rainfall Prediction Using Support Vector Machine Classification of Large-Scale Precipitation Maps Groundwater Prediction Using Machine-Learning Tools”



See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/343333939>

Regional Rainfall Prediction Using Support Vector Machine Classification of Large-Scale Precipitation Maps

Preprint · July 2020

CITATIONS
0

READS
44

3 authors:



Eslam A. Hussein
University of the Western Cape
5 PUBLICATIONS 1 CITATION

SEE PROFILE



Mehrdad Ghaziasgar
University of the Western Cape
17 PUBLICATIONS 31 CITATIONS

SEE PROFILE



Christopher Thron
Texas A&M University Central Texas
83 PUBLICATIONS 353 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Transformations in plane geometry [View project](#)



Applications of wireless sensor networks [View project](#)



Regional Rainfall Prediction Using Support Vector Machine Classification of Large-Scale Precipitation Maps

Eslam Hussein

*Department of Computer Science
University of the Western Cape
Cape Town, South Africa
eslamhuss34@gmail.com*

Mehrdad Ghaziasgar

*Department of Computer Science
University of the Western Cape
Cape Town, South Africa
mghaziasgar@uwc.ac.za*

Christopher Thron

*Department of Science and Mathematics
University-Central Texas
Killeen, Texas, USA
thron@tamuct.edu*

Abstract—Rainfall prediction helps planners anticipate potential social and economic impacts produced by too much or too little rain. This research investigates a class-based approach to rainfall prediction from 1-30 days in advance. The study made regional predictions based on sequences of daily rainfall maps of the continental US, with rainfall quantized at 3 levels: light or no rain; moderate; and heavy rain. Three regions were selected, corresponding to three squares from a 5×5 grid covering the map area. Rainfall predictions up to 30 days ahead for these three regions were based on a support vector machine (SVM) applied to consecutive sequences of prior daily rainfall map images. The results show that predictions for corner squares in the grid were less accurate than predictions obtained by a simple untrained classifier. However, SVM predictions for a central region outperformed the other two regions, as well as the untrained classifier. We conclude that there is some evidence that SVMs applied to large-scale precipitation maps can under some conditions give useful information for predicting regional rainfall, but care must be taken to avoid pitfalls.

Index Terms—a comparison study, a sequence of images, SVMs

I. INTRODUCTION

A. Role of rainfall maps in water resource management

Rainfall maps provide essential information about intensity, temporal, and spatial location which are essential in water resource management. Historical rainfall maps data can help different management sectors such as agriculture to make informed decisions about water supply management strategies to better utilize the occurrence of precipitation events [1]. Historical data can be most effectively utilized by developing prediction models such as machine learning to capture historical rainfall patterns.

In previous literature, prediction models based on rainfall maps may be grouped into two main categories. The first type involves applying deep learning to a sequence of images as an input to predict future frames. Usually, the images used for this type of prediction are separated by relatively small time intervals e.g 6-10 minutes [2]–[8]. The second type consists of single output regression or classification-based models. These models predict rainfall on an hourly [9]–[11], daily [12], [13],

or monthly [14] basis using prior rainfall maps. Regression-based models may use a single frame [14] or batch of frames [10] as an input to give a numerical rainfall prediction. In contrast, classification based models categorize local rainfall into two or more discrete classes and predict the classes of future precipitation events based on a single frame or a batch of frames. [12], [13], [15]. Both regression and classification models can be used to predict entire images one pixel at a time [11], [16].

For purposes of comparison, we describe the work of Boonyuen in [12] and [13]. In [12] the authors used a single image to produce a binary classification (rain/no-rain) for three days ahead in Thailand. Using the inception-V3 based CNN model the authors had up to 54.84% classification accuracy for three days ahead prediction. The study also concluded that including neighboring countries in the images increases the efficiency of the model compared to cropping the image to focus only on Thailand. In [13] the authors developed an inception-V3 model to classify predicted rainfall into four categories (No-rain, light-rain, moderate-rain, heavy-rain). Both batches of satellite images and single images were used as input. The study demonstrated that using batches of images as input makes the model more robust at classifying upcoming rainfall. The trained model was able to predict one, two, three days ahead with an accuracy of 70.58%. Having the same accuracy up to three days ahead is an issue of concern, as we suspect that the trained model has a bias towards the majority class (no-rain), making the accuracy to be constant. Measuring the efficiency of models using classification accuracy on imbalanced data is not ideal, because the results obtained may reflect the relative frequencies of the classes more than the actual effectiveness of the model. When imbalanced classes are involved, the f1-score can be a better measure of the method's effectiveness [17], as it takes the weighted average of precision and recall and is less influenced by class imbalance.

In The literature, various image sizes and sequence lengths are used in different prediction models. It appears that these parameters are usually chosen arbitrarily, or determined by

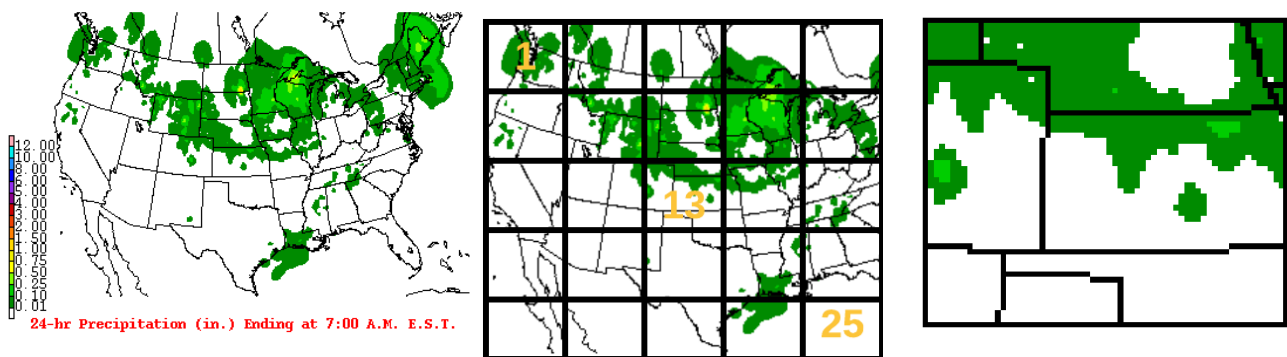


Fig. 1. Satellite precipitation images: (left) An image from the NCEP dataset; (centre) 5 grid overlay; (right) tile 13.

trail and error. For example, the authors in [10] used up to 60 images to predict rainfall on an hourly base. the authors in [3], [7], [8], [18]–[20], used 4, 4, 5, 10, 10, 20, .20 images respectively. As to image size, the authors in [10] used several image sizes between 101×101 and 10×10 , and compared the performance.

Even though most previous studies made use of deep learning related techniques to capture rainfall patterns on the historical images like Convolutional LSTM in [2], in some cases the deep learning approach has significant drawbacks. In particular, as it often overfits when the training set is relatively small [21], [22]. In addition, those models have many hyperparameters that need to be optimized.

These difficulties can be avoided by using a support vector machine (SVM) approach instead of deep learning. SVM is a powerful machine learning technique where it is often used in classification and regression problems [23]. SVM is a classifier that generates a hyperplane to classify data instances [24], [25], where optimal hyperplanes are determined by constructing the largest margin of separation between the different instances [26]. In contrast to deep learning, SVMs are suited to be trained on small and medium-sized complex datasets [23], [27]. SVM also has the capability to perform structural risk minimization (SRM), which enables SVMs to avoid overfitting by minimizing the bound on the generalization error [28].

Using SVMs can also avoid the need for extensive hyperparameter tuning. For example, the authors in [29] considered the use of linear kernel (SVMs) in case of having the number of features exceeds the size of the dataset. Linear kernels use only a single parameter, the regularization parameter C , that determines the trade-off between minimizing the training error and the model complexity [30].

Several recent papers in the literature use SVMs on rainfall prediction for different classifications and regression problems. The authors in [31] investigated the use of SVMs as well as other techniques to classify rainfall on a very small training set 10% (2245), where the output was a binary classification rain/no-rain daily. Another binary classification problem (rain/no-rain) was studied in [25], which investigated the use

of SVMs on weather stationary data to classify rainfall for the next five minutes. The data were highly imbalanced due to the rare occasion of rain, which made the researchers perform down-sampling on the dataset. As for regression problems, The authors in [32] used SVM to predict daily and the accumulated rainfall on 42 different cities from Europe and the US. The authors in [11] investigated the use of SVMs with hourly radar-derived rainfall to predict precipitation during typhoons. Another study linked the observations from satellite imagery data to predict rainfall up to 6-hours [15].

B. Scope of this research

This research aims to investigate precipitation forecasting on a dataset from the National Center for Environmental Prediction (NCEP) using SVMs. Our investigation has three aspects: i) Determine the effect of image sequence input length on class prediction accuracy, ii) Assess the effect of image size on class prediction accuracy, and iii) Compare the accuracy of rainfall class predictions for three selected squares (tiles) from a 5×5 grid covering the map area, for up to 30 days ahead.

This paper is organized as follows. Section 2 presents the methodology, including a discussion of the datasets and their preparation as well as the SVM specification and training. Section 3 presents the results in tabular and graphical form and provides analysis and discussion. Section 4 summarizes our conclusions.

II. METHODOLOGY

Following the flow chart in Figure 2, we start by discussing the data set, followed by the pre-processing of the images and the preparation of the data then explaining one of the models for this prediction.

A. Data Set

1) *24-Hour-Precipitation-Forecasting*: The data used for this study are radar images taken daily at 7 a.m. Eastern Standard Time, from Jan 2012 to Oct 2019, with a total of 2,835 images. The data comes from the NCEP, with a size of 400×320 pixels which represents the United States. Each image contains 16 different rainfall intensity level, Figure 1 (Left) shows a full image of the used dataset.

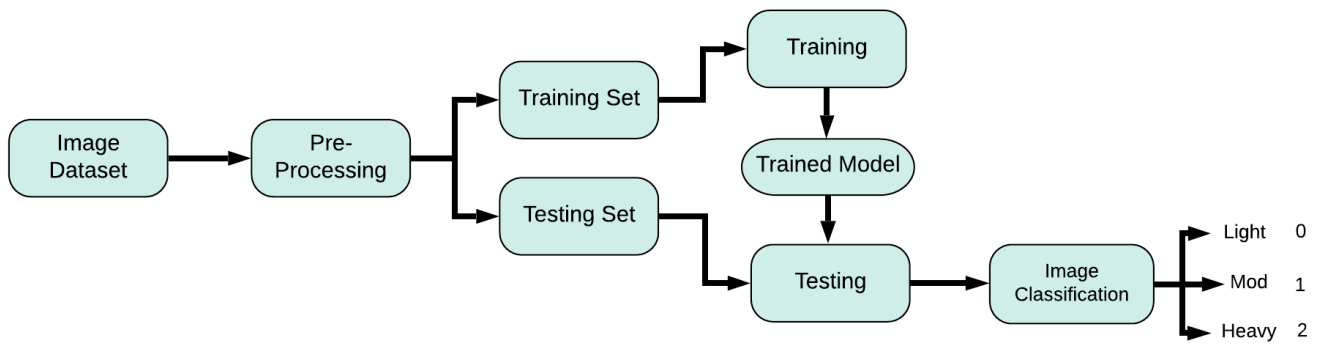


Fig. 2. Flow chart showing the implementation process.

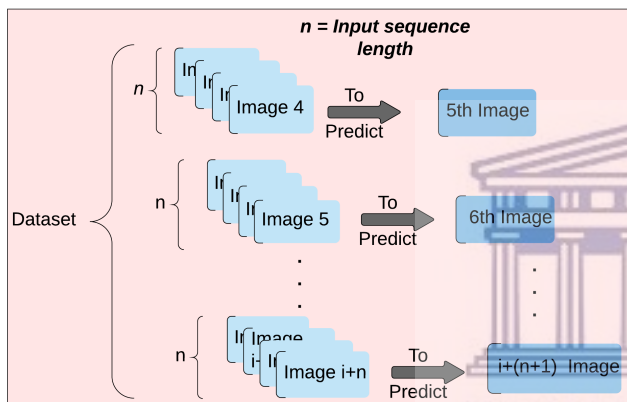


Fig. 3. Overview of dataset preparation.

B. Pre-Processing

The input data set went through a pre-processing transformation as shown in Figure 2. The first transformation included size reduction, as we reduced the size of the images from 400×320 to 172×123 (image size a), and 87×61 (image size b). For both reduced data sets, we cropped the images to remove some of the irrelevant information that exists in our data set like the color bar on the left of Figure 1 (left map). In this study, we did not consider the full image size due to memory and time consideration. After resizing we transformed the images into one channel by performing a grayscale transformation on the images to reduce the complexity of the model.

Figure 3, shows how we prepared the input dataset in a sliding window fashion to predict the next image in a sequence. Our use of sliding windows resembles the approach in several previous references [8], [20], [32]–[34]. The figure shows the case where $n = 4$, which stands for the size of the window (sequence length). In general, the number of feature (pixels) of one sample of the data set can be determined by $n \times w \times h$, where n is the size of the window, w is the

width of the image, and h is the height of the image, which depends on the size of the images. For optimization, the study tested different size window sizes n , where $n \in \{2, 4, 6, 8\}$. We divided that data into 90% training and 10% testing for the whole experiment. Moreover, we divided the US map into a 5×5 grid squares (tiles), as We trained on specific grid tiles which are 1, 13, 25.

Initially, we divided the 16 classes on the color bar by Figure 1 left image equally into three classes light, moderate, heavy. For each image, tiles were classified according to the highest level observed within the tile: for example, if the tile had one or more pixels showing heavy rainfall, the entire tile was classified as heavy rain for that image. However, this equal division produced highly unbalanced data, due to the rare occasion of very heavy rainfall. Consequently, our classification accuracy on the testing test was constant for predictions up to 30 days ahead, which is similar to what was observed in [13] as discussed in the Introduction. To circumvent this problem, we made an unequal division between the classes by designating the lowest three classes as no/light rain, the next three classes as moderate rain, and all remaining classes as heavy rain. This improved the balance between the three classes: for the three tiles, we observed the following frequencies (no/light rain, moderate rain, heavy rain): (25%, 49%, 26%) for tile 1, (38%, 30%, 32%) for tile 13, and (36%, 37%, 27%) for tile 25.

The three tiles show quite different seasonal behavior, as shown in Figure 4. For Tile 13 (central), there is a clear distinction between the light rain and heavy rain class frequencies between summer months (May-Aug) and winter months (Oct-Mar). For Tile 25 (southeast), the light/no rain class shows strong seasonality, while the other two classes less so. For Tile 1 (northwest) the seasonality for all classes is less distinct.

To summarise, the input to our model is a set of full images, with different images sizes and windows (sequence length), while the prediction of the rainfall intensity happens on three specific tiles corresponding to three local regions within the U.S.

TABLE I
TILE 1 (F1-SCORE ACCURACY ON THE TESTING SET). WITH k DAYS AHEAD (DA), WITH DIFFERENT n INPUT IMAGES AND SIZES.

Image Scale	Input Images	F1-score of Days Ahead (DA) (%)									Mean
		1DA	2DA	3DA	4DA	5DA	6DA	7DA	14DA	30DA	
Img(size a)	2	60	31	27	30	26	23	34	22	22	30.55
	4	52	37	40	26	24	32	22	22	22	30.77
	6	50	37	37	32	31	22	22	22	22	30.66
	8	44	31	30	29	27	25	22	22	22	28
Img(size b)	2	55	33	40	25	25	33	30	22	22	31.66
	4	57	43	30	35	37	25	22	22	22	32.55
	6	51	42	32	32	32	22	22	29	23	31.66
	8	53	34	36	35	24	25	23	22	22	30.44
Mean		52.75	36	34	30.5	28.25	25.875	24.625	22.875	22.125	

TABLE II
TILE 13 (F1-SCORE ACCURACY ON THE TESTING SET). WITH k DAYS AHEAD (DA), WITH DIFFERENT n INPUT IMAGES AND SIZES.

Image Scale	Input Images	F1-score of Days Ahead (DA) (%)									Mean
		1DA	2DA	3DA	4DA	5DA	6DA	7DA	14DA	30DA	
Img(size a)	2	51	49	49	42	42	42	43	38	44	44.44
	4	58	50	45	41	44	44	45	46	45	46.44
	6	51	50	42	39	44	44	40	48	41	44.33
	8	54	45	46	43	44	44	46	45	44	45.66
Img(size b)	2	55	46	48	46	41	40	48	46	37	45.22
	4	55	47	43	45	43	41	43	43	40	44.44
	6	57	48	46	43	49	43	43	42	41	45.77
	8	60	49	48	42	44	41	45	42	45	46.22
Mean		55.125	48	45.875	42.625	43.875	42.375	44.125	43.75	42.125	

TABLE III
TILE 25 (F1-SCORE ACCURACY ON THE TESTING SET). WITH k DAYS AHEAD (DA), WITH DIFFERENT n INPUT IMAGES AND SIZES.

Image Scale	Input Images	F1-score of Days Ahead (DA) (%)									Mean
		1DA	2DA	3DA	4DA	5DA	6DA	7DA	14DA	30DA	
Img(size a)	2	41	40	40	37	40	46	49	44	39	41
	4	45	35	46	36	43	44	36	40	49	41.55
	6	46	48	39	46	39	41	39	42	49	43.22
	8	53	37	41	46	42	41	42	39	42	42.55
Img(size b)	2	48	41	39	37	41	42	41	45	34	40.88
	4	42	38	42	48	41	47	41	42	42	42.55
	6	56	50	49	44	46	39	41	41	41	45.22
	8	49	42	42	46	42	43	41	43	42	43.33
Mean		47.5	41.375	42.25	42.5	41.75	42.875	40.375	42	42.25	

C. Classification

In this study, we used linear kernel SVMs, based on [29] which advised using a linear kernel when the number of features exceeds the size of the dataset. These were trained on the training set, which comprised 90% of the data. Training was accomplished using the *sklearn.svm.SVC* class in scikit-learn (www.scikit-learn.org). Since we are working with a linear kernel, we had only one parameter to optimize which is the regularization parameter C . We tried values $C = 2^k$ for $k \in -15, \dots, 6$ with 10-fold cross-validation on the training

set, and we chose the best C value separately for each input configuration and each day ahead prediction, for each of the three tiles. (It was observed that in most cases the value of C thus obtained was in the 2^{-13} – 2^{-10} range.) The models obtained were then applied to the testing set, and confusion matrices were obtained which were used to compute macro f1 scores. The macro f1-score was calculated as the average of the f1 scores of each class, where the per-class f1 score is

TABLE IV
F1-SCORE RANGES FOR 8 DIFFERENT SVM INPUTS FOR DIFFERENT DAYS AHEAD (DA), FOR THREE REGIONS (%)

Region	F1-score range of Days Ahead (DA) (%)								
	1DA	2DA	3DA	4DA	5DA	6DA	7DA	14DA	30DA
tile 1	16	12	13	10	13	11	12	7	1
tile 13	9	5	7	7	8	4	8	10	8
tile 25	15	15	10	12	7	8	6	6	15

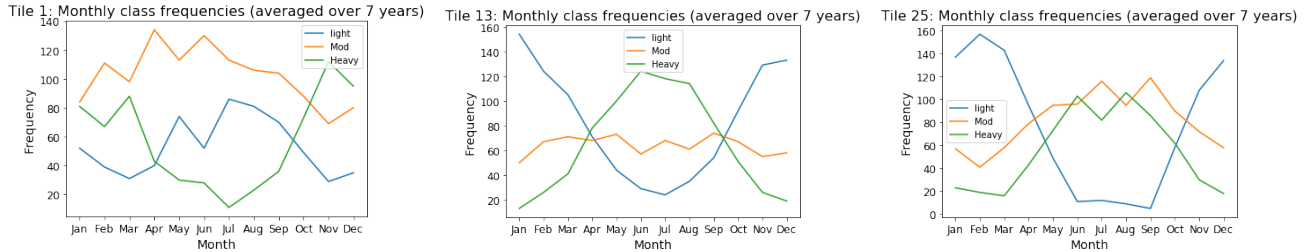


Fig. 4. Seasonality on a monthly basis for Tile 1 (left), Tile 2 (middle), and Tile 3 (right).

computed as follows:

$$f1\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (1)$$

III. RESULTS AND DISCUSSION

Tables I, II, III show the macro f1-score accuracy on the testing set for tile 1 (northwest), tile 13 (central) and tile 25 (southeast) respectively. Each table gives results for different days-ahead predictions for 8 different SVM inputs (four different input sequences and two different image sizes). The bold numbers in each column represent the highest macro f1-score among the 8 SVM inputs for that specific days-ahead prediction for the given tile.

In the following discussion, we will first compare the prediction performance for the different SVM inputs. Then we will compare the prediction performance for the three different geographical regions.

A. Comparison between different SVM inputs

From Tables I, II, and III we do not find that any one input configuration is clearly better than the others. In Table I for example, we find that 5 different input configurations attain the best accuracy for different days-ahead predictions. There is considerable variation within each column of the tables, as well as from column to column for each row. Table IV shows the f1 score range (maximum – minimum) among the 8 predictions for each days-ahead, for the three regions. From the table, it is clear that Tile 13, in general, has the lowest variability in the 1-6 day range, while the variability for Tile 1 reduces to almost 0 after 30 days.

The observed variabilities may be attributed at least partially to the relatively small size of the testing set, which consists of about 300 images. For purposes of comparison, a sequence of 300 Bernoulli trials with success probability $p = 0.5$ will have a standard deviation of ± 3 percentage points. So a 95% confidence interval of ± 2 standard deviations will have a width

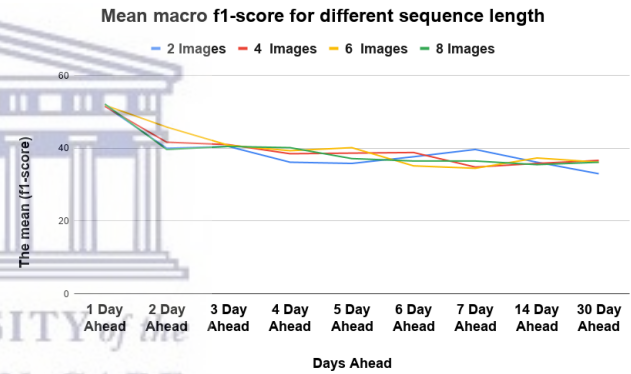


Fig. 5. Mean macro f1-scores as a function of days ahead for different input image sequence length, averaged over all image sizes and tiles.

of 12 percentage points. So in this Bernoulli trial scenario, the probability of getting 8 independent trials within a range of 12 is roughly $0.95^8 = 0.66$.

Figures 5 and 6 isolate the effect of input sequence length and image size, respectively. In Figure 5 the macro f1 scores for all predictions for all tiles for each days-ahead were averaged, and the results plotted as a function of days-ahead. The figure shows that no particular input sequence length is superior to the others. Figure 6 similarly averages macro f1 scores separately for each image size. There appears to be a slight advantage of about 1 percentage point when using the larger image size (172×123) instead of the smaller size (87×61). Both figures show a clear decrease in accuracy as days-ahead increases, in contrast to the constant classification accuracy found in [13].

B. Comparison between regional predictions

Figure 7 shows the macro f1 scores averaged over 8 inputs (4 sequence lengths \times 2 images sizes) for each day ahead,

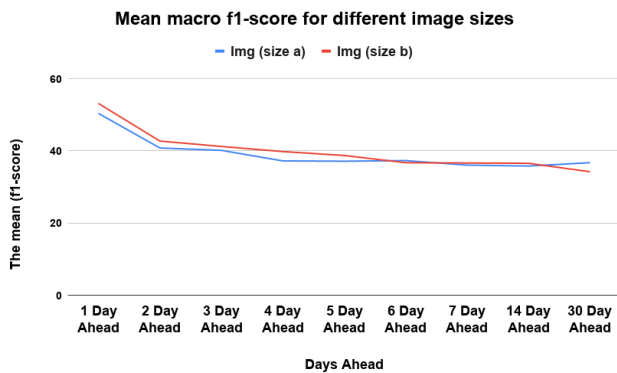


Fig. 6. Mean macro f1 score as a function of days ahead for the two different image sizes, averaged over all input sequence lengths and tiles.

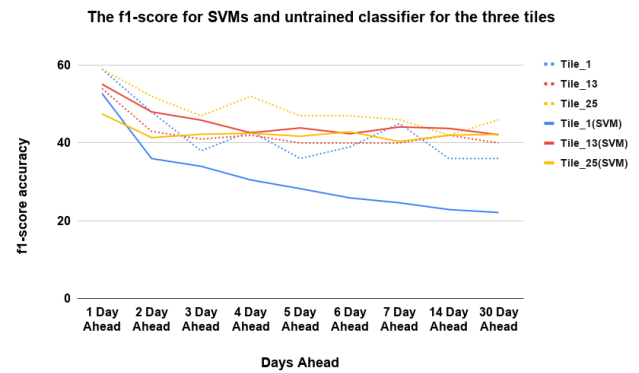


Fig. 7. Per-tile average of f1-scores for 8 SVM inputs as a function of days ahead, compared to untrained benchmark predictor.

for each image separately. These f1 scores are compared to f1 scores obtained from an untrained predictor that simply uses the class of the final image in the sequence as the predicted class for 1, 2, 3, . . . 30 days ahead. From the figure, it is clear that the SVM significantly underperforms the untrained benchmark predictor for Tiles 1 and 25. For Tile 13, the SVM outperforms the untrained predictor by about 1-3 percent.

The SVM performance for Tile 1 is particularly poor, especially for longer-range predictions. Upon closer examination of the confusion matrices produced by the simulation, we found that for predictions longer than 4 days ahead, almost all of the SVM inputs were producing a classifier that always predicted the most frequent class (moderate rain) regardless of the SVM input. An elementary calculation based only on class frequencies (25%, 49%, 26%) shows that a predictor which ignores inputs and always chooses the majority class will have a macro f1-score of 21.5%, which is consistent with our observed result.

The SVM performance for Tile 25 is also worse than the untrained predictor. The confusion matrices for Tile 25 shows that for larger-DA predictions, most predictors simply choose between light and moderate classes, and never predict the heavy class. This result is understandable based on the seasonal behavior of Tile 25, shown in Figure 4 (right). During the winter months (Nov.-Dec.), the light class is by far the most frequent. So if the input images show the tile as belonging to the light class, the chances are that the class will remain light for the next 30 days. On the other hand, in the summer months (May-Sep), the light class is virtually nonexistent, and moderate and heavy are about equal. Since moderate is a more frequent class than heavy (37% versus 27%), the classifier favors moderate over heavy. Apparently, the untrained benchmark predictor more closely matches the seasonal pattern, which is why the untrained predictor performs better. We note that for Tile 25, a predictor that always chooses the majority class will have a macro f1 score equal to 18%, so the SVM does represent a large improvement over a majority-class predictor.

Tile 13 showed the best SVM performance among the three tiles and was the only tile where the SVM outperformed the untrained benchmark classifier. We also noted from Table IV that there was a smaller variation in macro f1 scores among the 8 different SVM inputs, for predictions between 1-5 days ahead. It is reasonable that the SVM's in Tile 13 are finding useful features and converging, while the SVM's for the other tiles are not locating truly useful features, so they are overfitting the training set which means that they no longer give consistent accuracy when applied to the testing set. It may be argued that Tile 13 has better input data than the other tiles because the input images contain information from all regions surrounding Tile 13 which is not the case for the other two tiles.

IV. CONCLUSION

Our study points out some of the limitations and potential pitfalls in using SVMs with linear kernels for weather prediction up to 30 days in advance. We have shown that some classifiers used by previous researchers (e.g. [12] and [13]) which seem to show good performance may be largely due to unequal class divisions rather than the classifier itself. We have also shown that unequal classes may cause linear SVMs to converge on majority-class classifiers, or to completely neglect classes of low frequency. Among the three geographical regions predicted, only the central region had an SVM-based classifier that performed better than a simple untrained classifier that used the tile's class in the final image of the input sequence as the prediction. We conjecture that SVMs may be performing better on the central tile because the input sequence contains precipitation information for all surrounding tiles, which is not the case for tiles at the corners of the map (such as Tile 1 and Tile 25).

In this study, we propose to divide the map of the US onto 25 tiles, as for the optimization we used different input configurations. The support vector machine was used to classify the image sequences as no/ light-rain: a) Our study shows that f1-score for tile 13 has a generally better f1-score than tile 1 and tile 25, which goes back to the position of the tile. b) Taking a

bigger size scale as an input appears to provide slightly better performance than the smaller scale, but with higher time cost. c) The f1-score shows a decay while predicting days ahead, but the decay does not appear as prominent with imbalanced tiles e.g tile 1.

We have argued that the variability in the results obtained from day to day and from input method to input method was at least partially due to the insufficient amount of data for training and testing. Much of the variation in observed f1 scores is attributable to the small testing set size. Recent research has shown that using data augmentation to augment the set of training images may improve the efficiency of the trained model [5]. With an augmented training set, a larger portion of the actual data may be used for testing.

Another significant drawback of the SVM classifiers used was that they did not take seasonality into account. This is why a simple untrained classifier that took advantage of seasonality was able to substantially outperform the SVM classifiers on two out of the three regions examined. It is possible that SVMs that take seasonality into account may perform much better: this is a possible area for future research.

In the current research we did not attempt to include additional engineered features, because the number of features used was already very large. For future investigation, a possible approach would be to use PCA to reduce the number of features, and then apply feature engineering. In addition to time stamp, Gaussian Mixture Models may be used to capture spatial means and variances, which could be used as global features.

ACKNOWLEDGMENT

We acknowledge the use of the ilifu cloud computing facility- (www.ilifu.ac.za)

REFERENCES

- [1] R. Harmel, K. King, C. Richardson, and J. Williams, "Long-term precipitation analyses for the central texas blackland prairie," *Transactions of the ASAE*, vol. 46, no. 5, p. 1381, 2003.
- [2] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, 2015, pp. 802–810.
- [3] E. Shi, Q. Li, D. Gu, and Z. Zhao, "A method of weather radar echo extrapolation based on convolutional neural networks," in *International Conference on Multimedia Modeling*. Springer, 2018, pp. 16–28.
- [4] X. Shi, Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Deep learning for precipitation nowcasting: A benchmark and a new model," in *Advances in neural information processing systems*, 2017, pp. 5617–5627.
- [5] Q.-K. Tran and S.-k. Song, "Multi-channel weather radar echo extrapolation with convolutional recurrent neural networks," *Remote Sensing*, vol. 11, no. 19, p. 2303, 2019.
- [6] Y. Wang, M. Long, J. Wang, Z. Gao, and S. Y. Philip, "Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms," in *Advances in Neural Information Processing Systems*, 2017, pp. 879–888.
- [7] S. Singh, S. Sarkar, and P. Mitra, "A deep learning based approach with adversarial regularization for doppler weather radar echo prediction," in *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2017, pp. 5205–5208.
- [8] L. Chen, Y. Cao, L. Ma, and J. Zhang, "A deep learning based methodology for precipitation nowcasting with radar," *Earth and Space Science*, p. e2019EA000812, 2020.
- [9] S. Kim, S. Hong, M. Joh, and S.-k. Song, "Deeprain: ConvLstm network for precipitation prediction using multichannel radar data," *arXiv preprint arXiv:1711.02316*, 2017.
- [10] c. zhang, H. Wang, J. Zeng, L. Ma, and L. Guan, "Tiny-rainnet: A deep cnn-bilstm model for short-term rainfall prediction."
- [11] P.-S. Yu, T.-C. Yang, S.-Y. Chen, C.-M. Kuo, and H.-W. Tseng, "Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting," *Journal of Hydrology*, vol. 552, pp. 92–104, 2017.
- [12] K. Boonyuen, P. Kaewprapha, and P. Srivihok, "Daily rainfall forecast model from satellite image using convolution neural network," in *2018 IEEE International Conference on Information Technology*, 2018, pp. 1–7.
- [13] K. Boonyuen, P. Kaewprapha, U. Weesakul, and P. Srivihok, "Convolutional neural network inception-v3: A machine learning approach for leveling short-range rainfall forecast model from satellite image," in *International Conference on Swarm Intelligence*. Springer, 2019, pp. 105–115.
- [14] S. Aswin, P. Geetha, and R. Vinayakumar, "Deep learning models for the prediction of rainfall," in *2018 International Conference on Communication and Signal Processing (ICCSPP)*. IEEE, 2018, pp. 0657–0661.
- [15] K. Chen, J. Liu, S. Guo, J. Chen, P. Liu, J. Qian, H. Chen, and B. Sun, "Short-term precipitation occurrence prediction for strong convective weather using fy2-g satellite data: a case study of shenzhen, south china," *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 41, p. 215, 2016.
- [16] A. Mukhopadhyay, B. P. Shukla, D. Mukherjee, and B. Chanda, "A novel neural network based meteorological image prediction from a given sequence of images," in *2011 Second International Conference on Emerging Applications of Information Technology*. IEEE, 2011, pp. 202–205.
- [17] L. A. Jeni, J. F. Cohn, and F. De La Torre, "Facing imbalanced data—recommendations for the use of performance metrics," in *2013 Humaine association conference on affective computing and intelligent interaction*. IEEE, 2013, pp. 245–251.
- [18] E. Shi, Q. Li, D. Gu, and Z. Zhao, "Convolutional neural networks applied on weather radar echo extrapolation," *DEStech Transactions on Computer Science and Engineering*, no. csae, 2017.
- [19] L. Tian, X. Li, Y. Ye, P. Xie, and Y. Li, "A generative adversarial gated recurrent unit model for precipitation nowcasting," *IEEE Geoscience and Remote Sensing Letters*, 2019.
- [20] Y. Cao, Q. Li, H. Shan, Z. Huang, L. Chen, L. Ma, and J. Zhang, "Precipitation nowcasting with star-bridge networks," *arXiv preprint arXiv:1907.08069*, 2019.
- [21] S. Liu and W. Deng, "Very deep convolutional neural network based image classification using small training sample size," in *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*. IEEE, 2015, pp. 730–734.
- [22] P. Liu, K.-K. R. Choo, L. Wang, and F. Huang, "Svm or deep learning? a comparative study on remote sensing image classification," *Soft Computing*, vol. 21, no. 23, pp. 7053–7065, 2017.
- [23] A. Géron, *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems*. "O'Reilly Media, Inc.", 2017.
- [24] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: A review," *Sensors*, vol. 18, no. 8, p. 2674, 2018.
- [25] S. Manandhar, S. Dev, Y. H. Lee, Y. S. Meng, and S. Winkler, "A data-driven approach for accurate rainfall prediction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 9323–9331, 2019.
- [26] H.-C. Kim, S. Pang, H.-M. Je, D. Kim, and S. Y. Bang, "Constructing support vector machine ensemble," *Pattern recognition*, vol. 36, no. 12, pp. 2757–2767, 2003.
- [27] J. Du, Y. Liu, Y. Yu, and W. Yan, "A prediction of precipitation data based on support vector machine and particle swarm optimization (psvm) algorithms," *Algorithms*, vol. 10, no. 2, p. 57, 2017.
- [28] K. Chau and C. Wu, "A hybrid model coupled with singular spectrum analysis for daily rainfall prediction," *Journal of Hydroinformatics*, vol. 12, no. 4, pp. 458–473, 2010.
- [29] C.-W. Hsu, C.-C. Chang, C.-J. Lin *et al.*, "A practical guide to support vector classification," 2003.

- [30] K. Duan, S. S. Keerthi, and A. N. Poo, "Evaluation of simple performance measures for tuning svm hyperparameters," *Neurocomputing*, vol. 51, pp. 41–59, 2003.
- [31] S. Zainudin, D. S. Jasim, and A. A. Bakar, "Comparative analysis of data mining techniques for malaysian rainfall prediction," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 6, no. 6, pp. 1148–1153, 2016.
- [32] S. Cramer, M. Kampouridis, A. A. Freitas, and A. K. Alexandridis, "An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives," *Expert Systems with Applications*, vol. 85, pp. 169–181, 2017.
- [33] M. Qiu, P. Zhao, K. Zhang, J. Huang, X. Shi, X. Wang, and W. Chu, "A short-term rainfall prediction model using multi-task convolutional neural networks," in *2017 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2017, pp. 395–404.
- [34] B. Klein, L. Wolf, and Y. Afek, "A dynamic convolutional layer for short range weather prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4840–4848.



Chapter 4

Manuscript ‘Groundwater Prediction Using Machine-Learning Tools’



UNIVERSITY *of the*
WESTERN CAPE

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/346998362>

Groundwater Prediction Using Machine-Learning Tools

Article in *Algorithms* · November 2020

DOI: 10.3390/a13110300

CITATIONS

0

READS

25

5 authors, including:



Eslam A. Hussein
University of the Western Cape

5 PUBLICATIONS 1 CITATION

SEE PROFILE



Christopher Thron
Texas A&M University Central Texas

83 PUBLICATIONS 353 CITATIONS

SEE PROFILE



Mehrdad Ghaziasgar
University of the Western Cape

17 PUBLICATIONS 31 CITATIONS

SEE PROFILE



Antoine Bagula
University of the Western Cape

105 PUBLICATIONS 1,041 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



HELP: The Herschel Extragalactic Legacy Project [View project](#)



Stress, Cardiovascular Health, and Accident Risks for Commercial Drivers in Abuja, Nigeria: Causes and Correlations [View project](#)



<http://etd.uwc.ac.za/>

All content following this page was uploaded by [Eslam A. Hussein](#) on 18 December 2020.

The user has requested enhancement of the downloaded file.



Article

Groundwater Prediction Using Machine-Learning Tools

Eslam A. Hussein ^{1,*}, Christopher Thron ², Mehrdad Ghaziasgar ¹, Antoine Bagula ¹
and Mattia Vaccari ³

¹ Department of Computer Science, University of the Western Cape, Cape Town 7535, South Africa; mghaziasgar@uwc.ac.za (M.G.); abagula@uwc.ac.za (A.B.)

² Department of Science and Mathematics, University-Central Texas, Killeen, TX 76549, USA; thron@tamuct.edu

³ Department of Physics and Astronomy, University of the Western Cape, Cape Town 7535, South Africa; mattia.vaccari@gmail.com

* Correspondence: ehusein@uwc.ac.za

Received: 6 October 2020; Accepted: 2 November 2020; Published: 17 November 2020



Abstract: Predicting groundwater availability is important to water sustainability and drought mitigation. Machine-learning tools have the potential to improve groundwater prediction, thus enabling resource planners to: (1) anticipate water quality in unsampled areas or depth zones; (2) design targeted monitoring programs; (3) inform groundwater protection strategies; and (4) evaluate the sustainability of groundwater sources of drinking water. This paper proposes a machine-learning approach to groundwater prediction with the following characteristics: (i) the use of a regression-based approach to predict full groundwater images based on sequences of monthly groundwater maps; (ii) strategic automatic feature selection (both local and global features) using extreme gradient boosting; and (iii) the use of a multiplicity of machine-learning techniques (extreme gradient boosting, multivariate linear regression, random forests, multilayer perceptron and support vector regression). Of these techniques, support vector regression consistently performed best in terms of minimizing root mean square error and mean absolute error. Furthermore, including a global feature obtained from a Gaussian Mixture Model produced models with lower error than the best which could be obtained with local geographical features.

Keywords: time series data; pixel estimation; full image prediction; gaussian mixture model; global features; feature engineering; square root transformation

1. Introduction

In many countries, groundwater is one of the key natural resources that supplies a large portion of the water used by a nation. Besides its use in households and businesses, some other groundwater consumers include: (i) rural households and public water supplies that depend on wells and groundwater; (ii) farmers who use groundwater to irrigate crops and water their animals; and (iii) commercial businesses and industries that depend on groundwater for their processes and operations. Furthermore, the importance of groundwater can be revealed in its usage in supplying springs, water in ponds, marshlands, swamps, streams, rivers and bays. However, despite its unequivocal importance, groundwater levels in aquifer systems are often not constant and depend on recharge from infiltration of precipitation.

Several major acts and regulations such as the South African national water Act [1] and the 4th World Water Forum [2] recognize water as a basic human need, which is a major contributor to social development since it helps to alleviate poverty [1]. Hence, there is a growing interest towards the use of groundwater to help alleviate this crisis [2]. Groundwater is a vital freshwater resource which provides around 50% of the available drinking water according to UNESCO [3]. Also, sectors like

agriculture, and industry greatly depend on groundwater for their operations due to its widespread availability and the fact that it is not easy polluted [3,4]. Therefore, in 2015 the United Nations have reaffirmed their commitment regarding the human right to safe drinking water and sanitation by identifying it as one of the 17 Sustainable Development Goals to be pursued by 2030 [5].

Predicting groundwater availability is important to water sustainability and drought mitigation. It can provide useful insights based on real data of what happened when the flow of streams and rivers declined and/or when water supply issues developed into a drought. Machine-learning tools technologies have the potential to drive groundwater knowledge discovery and management by assisting in the prediction of groundwater availability. This can be done by enabling the collection of massive water datasets, storing these datasets into databases and processing these datasets to get useful insights which can be used by water resource managers to: (1) anticipate water quality in unsampled areas or depth zones; (2) design targeted monitoring programs; (3) inform groundwater protection strategies; and (4) evaluate the sustainability of groundwater sources of drinking water.

This paper uses a regression-based approach to predict full groundwater images based on sequences of monthly groundwater maps of the southern part of the African continent using the Gravity Recovery and Climate Experiment (GRACE) dataset [6]. Five machine-learning techniques are implemented on the GRACE dataset and compared to predict pixels in future frames of the dataset. These are extreme gradient boosting (XGB), multivariate linear regression (MLR), random forest (RF), multilayer perceptron (MLP) and support vector regression (SVR). The prediction is guided by: (i) performing feature selection based on the XGB feature importance bar on the previous lags (pixels); and (ii) investigating the effect of adding other features such as the temporal features, position indices, and global features obtained by Gaussian mixture models (GMMs) fitted to peak areas on each image.

This paper is organized as follows: Section 2 provides a background on water prediction, citing relevant literature in the field; Section 3 describes the algorithms used in this work; Section 4 discusses the methodology used for ground water prediction; Section 5 provides and discusses the results obtained; and Section 6 furnishes the conclusions.

2. Background on Groundwater Prediction

With the increase in population size coupled with urban expansion, water demand has dramatically increased, which has led to the over-exploitation of groundwater in many countries around the world [7,8]. This highlights the importance of groundwater forecasting. Accurate prediction of groundwater can help government officials determine the best approach to manage groundwater effectively [9]. The main tools for groundwater prediction are based on physical models and data-driven models [10]. Physical models require a large amount of detailed hydrological data, which suffers from a lack of accuracy during its collection and pre-processing [9]. Therefore, data-driven models tend to be more appealing, since they require less data and are more reliable [3,11].

Statistical models like multivariate linear regression (MLR), and various time series models such as autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA) and seasonal ARIMA (SARIMA), have been used to investigate patterns between the input and the output of groundwater data to make future predictions. The following studies have investigated: MLR for groundwater prediction [12–14]; and time series models for groundwater prediction [12,15–17]. Both techniques are considered to be linear fitting models [11]. Time series models have the advantage of accounting for the correlations that arise between data points [18]. In general, however, linear fitting is not ideal in describing the nonlinear behavior of groundwater. Hence, recent research has made use of MLR models more for comparative purposes [11].

In addition to these techniques, a range of machine-learning techniques have been applied to the problem, including MLP in [12,19–23], SVR in [19,24] and recently RFs in [25,26]. The use of XGB is rare in the scheme of groundwater prediction, and is found in only a few studies such as [27,28].

The above studies can be broadly divided into those that predict the groundwater level (GWL) and those that estimate the terrestrial water storage (Δ TWS). GWL provides an idea of the groundwater

level, whereas Δ TWS provides an idea about the volume of the groundwater. The GRACE database gives geographical Δ TWS levels monthly [6]. The significance of GRACE in hydrology is that it can provide an understanding of groundwater storage conditions and trends [29]. GRACE has been used as a predictor to help in the estimation of the GWL in [29,30].

In the literature, there are two main approaches to the problem of sequential image prediction. The first type involves taking a sequence of images as an input to predict future frames using deep learning techniques such as Convolutional Long-Short-Term-Memory (ConvLSTM). Usually, the images used for this type of prediction are separated by relatively short time intervals e.g., 6–10 min [31–37]. This approach normally does not involve any feature selection approach, since deep learning related techniques are known for their feature selection and reduction properties. However there are several concerns when using deep learning models: they depend heavily on a large quantity of high quality data to produce an effective model; they are very costly to train and use, in terms of time and resources; deep learning models are often viewed as black boxes [38] which means that it is very difficult to unpack and understand the automated feature selection process that eventually takes place and the predictions that arise from any given deep-learning-based model.

This leads us to the second approach in which machine-learning techniques can be used for single-output regressing problems. For GRACE Δ TWS image reconstruction, the authors in [27] used both XGB and RFs to acquire the importance of 20 features. It was shown that the precipitation of the two months prior to prediction is the most important variable for estimating the TWS dynamics. In [28], authors manually selected 11 hydrological predictors including the total precipitation and snow cover to predict Δ TWS. As for the idea of using previous pixels to predict current pixels has not been investigated enough in the literature. The authors in [39] made a comparison between Support Vector Machines (SVMs) and RF in predicting the present grid-based rainfall up to 3 h ahead, where the input involved the antecedent radar-derived rainfall. The authors in [40], used ANNs to predict full water vapor images every 30 min, where they included information from two previous images.

3. Techniques Used

In the following section we describe the tools and the technologies that has been used during the study. A total of five machine-learning techniques were used in this study for image prediction: (a) MLR; (b) MLP; (c) RF; (d) XGB; and (e) SVR.

Aside from the task of prediction, XGB was also used as a feature extraction and selection tool. As for feature engineering we used Gaussian Mixture Models (GMMs) to capture global features—mean and variance—of past images. For evaluation of the trained models, we used the RMSE and the MAE as evaluation metrics. All of the above-mentioned tools were implemented using the scikit-learn library [41] in the Anaconda python distribution (version 2020.07) with their default hyper-parameter settings.

Sections 3.1–3.6 describe each of the five machine-learning techniques listed above. Section 3.7 describes the metrics used to evaluate the trained models.

3.1. Multivariate Linear Regression

The level of correlation between the predictors and the output variables are usually estimated by regression models to determine their relationship form [42]. In linear regression, the mean square error is used to fit the models and to assess the performance of the trained models on the testing set [42,43]. In general, MLR is used to discover the hyperplane that best fits all individual data points [42]. For simplicity, in the following sections, MLR will be abbreviated as LR.

3.2. Multilayer Perceptron

MLPs are a type of artificial neural networks, which is a class of models inspired by the biological nervous system of the human brain. They can emulate complex functions like decision making, and pattern generation [44]. Like the human brain, MLPs also consist of a set of processing units called

'neurons', which are connected to each other. Each neuron is a multi-input and single-output nonlinear element [45]. Neurons mostly operate in parallel, and are arranged in multiple layers which include an input layer into which the data are fed; hidden layer where the learning takes place; and an output layer [44]. MLPs can detect complex nonlinear relationships through a learning process that involves the adjustment of the weighted connections that exists between the neurons. This gives MLPs the ability to perform two important functions: pattern classification and nonlinear adaptive filtering [46].

3.3. Random Forest

RF uses an ensemble of classification and regression trees. Each tree is built using a different bootstrap sample (with replacement) from the original data [47]. Compared to traditional trees, RF adds a randomness layer to bagging, since in traditional trees each node is split using the best split among all variables [48]. As for RF, only a random subset of the variables is used when splitting a node during the construction of a tree [47,48]. As a result of the random constructions, RF provides robustness to overfitting as compared to some other techniques [48,49].

3.4. eXtreme Gradient Boosting

Like RF, XGB is an ensemble learning technique. XGB relies on gradient boosting to form a combined prediction. In XGB, the predictors in a tree are built in a sequential manner, and are trained on the residuals of the previous learners, so that errors are reduced step by step [27].

In the scikit-learn implementation in XGB, the `plot.importance` command can be used to determine feature importance for the features in trained predictive model [50]. The `plot.importance` command computes for each separate feature the sum of estimated improvements in squared error risk for all decision nodes employing that feature, averaged over all trees used in the model. The averaging greatly reduces the masking effect which occurs when variables are correlated [51].

3.5. Support Vector Machine and Support Vector Regression

SVM is a powerful machine-learning technique that has the capability to perform structural risk minimization (SRM), which enables it to avoid overfitting by minimizing the bound on the generalization error [52]. SVMs may be extended to apply to estimation and regression problems: this extension is known as support vector regression (SVR) [53]. SVR maps the input data into a higher-dimensional feature space via nonlinear kernel functions [54]. The objective is to choose a vector of regression coefficients with a small norm, while minimizing the sum of the distances between the data points and the regression hyperplane in the higher-dimensional space [55].

3.6. Gaussian Mixture Models

Gaussian mixture models (GMMs) may be used for clustering [56] or as parametric models of the probability distribution of continuous features [57]. The user specifies the number of Gaussians in the model, and the means and covariances of the Gaussians are automatically computed using an expectation maximization (EM) algorithm [58].

3.7. Performance Metrics

The accuracies of the above machine-learning models are evaluated using the root mean square error (RMSE) and the mean absolute error (MAE). Both metrics are commonly used to measure the forecasting accuracy [59]. RMSE is more sensitive to outliers and is more appropriate for Gaussian-distributed errors, while MAE weights all errors equally [60]. The RMSE and MAE are computed as follows:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i^{\text{obs}} - y_i^{\text{pre}})^2} \quad (1)$$

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i^{\text{obs}} - y_i^{\text{pre}}| \quad (2)$$

where y_i^{obs} and y_i^{pre} refer to the observed and predicted value of the i th output, respectively.

4. Groundwater Prediction Methodology

Following the flowchart in Figure 1, we start by discussing the data set, followed by the pre-processing of the images and the preparation of the data set. Then we speak about the feature selection that was done using XGB, and feature engineering using GMM. Finally, we end up with the experiment. Our end goal is to predict groundwater on a pixel level to end up with a full image using a sequence of images as an input.

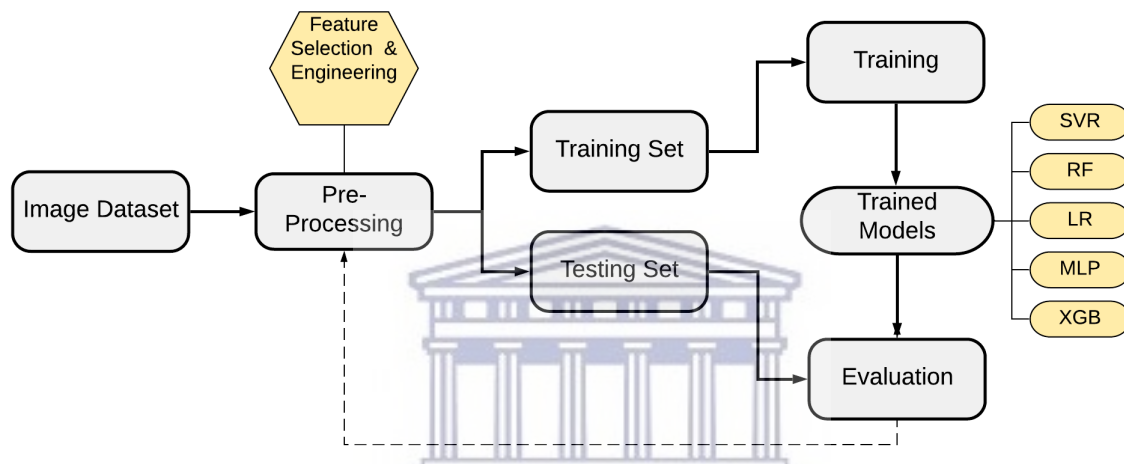


Figure 1. Flowchart showing the implementation process.

4.1. Monthly Groundwater Data Set

The dataset used in this study consists of 174 monthly groundwater satellite images between March 2002 and May 2019. Each image has a size of 360×180 pixels, and provides a color-coded representation of the ΔTWS of the earth's land surface. A sample image from the dataset is provided in Figure 2. The images were originally obtained as part of the GRACE survey conducted by the U.S. National Aeronautical and Space Administration (NASA). The actual data was obtained from the Physical Oceanography Distributed Active Archive Center (PO. DAAC) website [6].

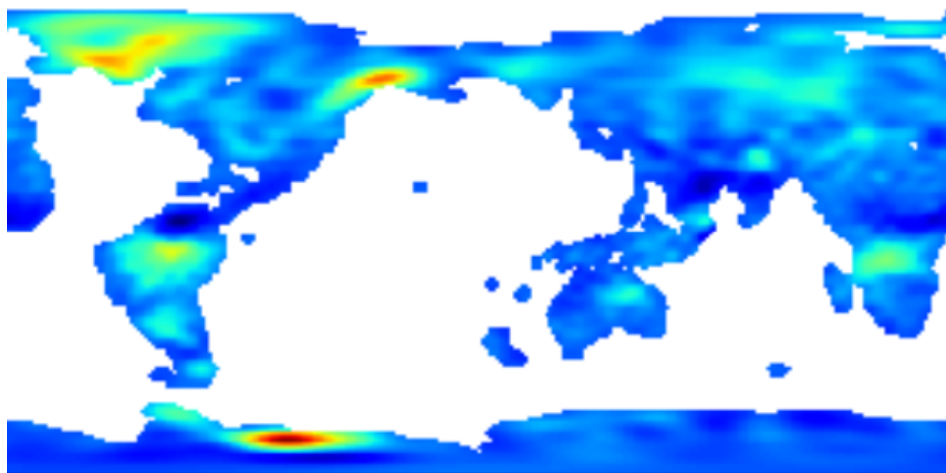


Figure 2. A sample full image of the GRACE groundwater dataset used in this research [6].

Using this dataset posed some serious challenges. Several monthly images in the dataset were missing. Neglecting these months would disrupt the periodicity/seasonality in the data, which is a key aspect of the data. Further reducing the dataset was not feasible, since the number of images is already on the small side for machine-learning applications. Finding better methods for dealing with missing images is an ongoing research topic.

4.2. Image Pre-Processing

Predicting a full color image at the pixel level would be computationally expensive, since the full image consists of R,G,B values for $360 \times 180 \times 3$ pixels. Hence, to reduce the computational cost, we transformed the images into grayscale, and cropped the images to focus on the southern area of the African continent with a size of 47×51 . An image of the pre-processed data is shown in Figure 3 (left).

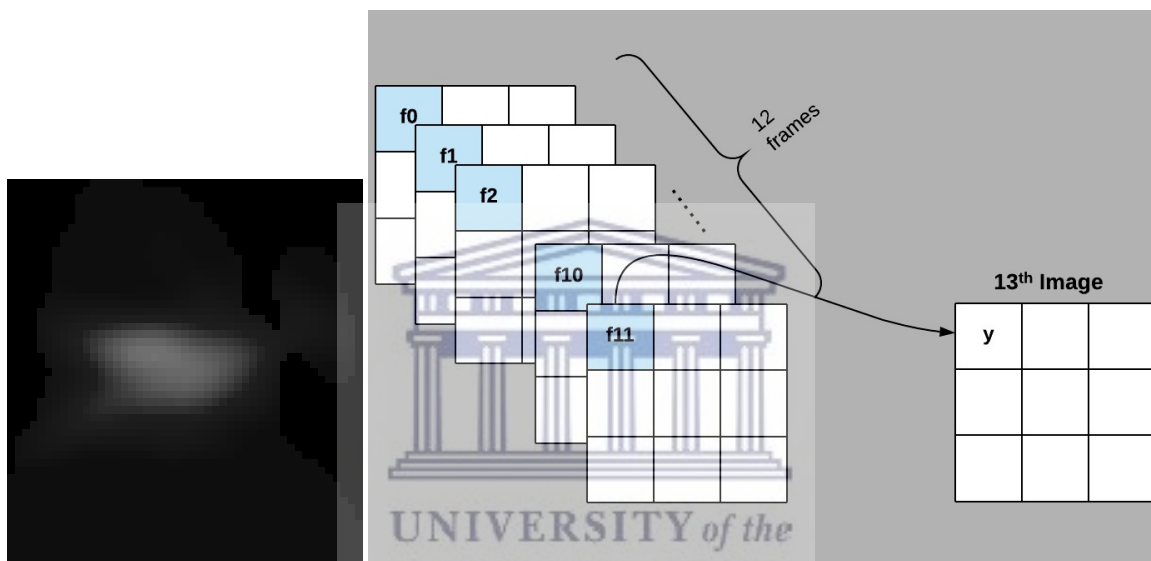


Figure 3. Overview of dataset preparation for feature selection: (left) Example of groundwater image of southern Africa before pre-processing (note image is inverted vertically); (right) Notation for same-pixel features used in image prediction.

The dataset provided from PO.DAAC had missing months. We imputed the data for the missing months by replicating the previous months' images. Out of the 174 frames, we deleted the first two frames because of a gap of about 100 days between the 2nd and the 3rd image in the data. This left 172 images, and after image imputation we ended up with a dataset of 190 images. Altogether 18 images were imputed, which amount to about 10% of the original 172 images. We then applied a sliding window to form 161 sequences of 12 consecutive images. The first 149 sequences were used for training, and the rest for testing. We did not use any of the imputed images as labels for the models to train on.

4.3. Feature Selection

Since the dataset in this research is small, it was particularly important to choose a set of features of limited size (to avoid overfitting) but which still captures essential information that can be used for prediction. In our model we used both local and global spatiotemporal features, and additionally performed a rescaling, as described in the following subsections.

4.3.1. Same-Pixel Features

Our first set of candidate features for prediction of an image pixel consists of the same pixel location for the 12 previous months. A similar choice has been made in previous studies [39,40,61]. To select the most important of these 12 features, a sliding window technique is used, as shown

in Figure 3 (right): the same pixels within a prior 12-month window were used to predict the corresponding pixel in the 13th month.

To evaluate the relative importance of these features, XGB with the gain metric was applied to the training set. Figure 4 shows the results. In the figure, $f(0)$ stands for the same month in the previous year while $f(11)$ stands for the previous month. The graph shows that $f(0)$, $f(11)$, and $f(1)$ (12, 1, and 11 months previous, respectively) have the greatest importance. This finding agrees with [62,63], which also used the previous and 12-month prior pixels to predict corresponding points in current month.

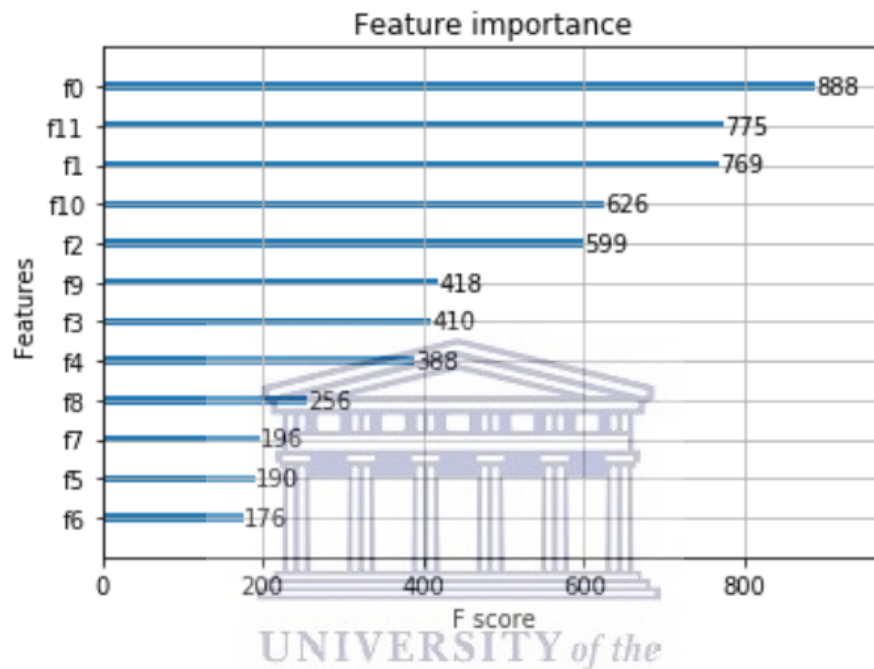


Figure 4. Feature importance of same pixel of previous months, where $f(0)$ stands for same month previous year, and $f(11)$ stands for previous month.

Based on our results, we created seven different feature sets labelled **a–g** as follows:

- **a** = $f(0, 11)$
- **b** = $f(0, 11, 1)$
- **c** = $f(0, 11, 1, 10)$
- **d** = $f(0, 11, 1, 10, 2)$
- **e** = $f(0, 11, 1, 10, 2, 9)$
- **f** = $f(0, 11, 1, 10, 2, 9, 3)$
- **g** = $f(0, 11, 1, 10, 2, 9, 3, 8, 4)$

4.3.2. Other Local Spatiotemporal Features, and Rescaling

Because of the geographical and seasonal nature of groundwater levels, the following spatiotemporal features were also deemed to be significant and were used:

- Pixel's x, y coordinate;
- Time stamp $(0, \dots, 11)$ ($0 = \text{January}, \dots, 11 = \text{December}$)

Since most of the pixel values are low and high values were relatively infrequent, the pixel values were replaced by their square roots to regularize the scale. The square root transformation was similarly applied to inputs in [64], and to outputs in [65,66].

4.3.3. Global Feature Generation Using Gaussian Mixture Models

In this subsection we describe how we used Gaussian mixture models (GMMs) to generate global features. This idea came from observing that regions of high groundwater level seemed to form shapes that could be described as Gaussian distributions, which propagated from image to image. The means and the covariances of these Gaussian-shaped features can be used as global features that describe the motion of high regions from image to image. To apply GMM to an image, we converted the image to a set of pixel locations by randomly selecting pixels with selection probability proportional to the pixel's scaled intensity (see Figure 5). These pixel locations were fed to the GMM algorithm which returned means, covariance matrices, and weights of Gaussian clusters. In this study, we set up the algorithm to have only one cluster. The pixel located at the cluster mean and the two eigenvalues of the covariance matrix were used as global features.



Figure 5. Two image representations of groundwater, where (A) represents a normal frame of groundwater; (B) represents the captured high pixels intensity.

We applied GMM to image 10 (two months previous), and image 11 (one month previous), yielding a total of 8 global features. As described above, our application of GMM involved a randomization when choosing pixel locations. To account for this randomness, when evaluating models that used GMM features we created 100 different models using different randomization. From those 100 results we took the per-pixel averages to get a single model, and took the RMSE and MAE for this model to obtain accuracy estimates.

5. Performance Results and Discussion

5.1. Performance Results

Tables 1–4 show performance accuracies for models trained using different features. Table 1 uses only same-pixel data from previous months; Table 2 adds the features (i,j) , which stands for the pixels position in a 2D array; Table 3, adds the time stamp (denoted by t); Table 4, additionally applies the square root transformation (denoted by s) to the pixel values. The code together with image data is available on GitHub at <https://github.com/EslamHussein55/Groundwater-Pixel-Prediction-using-ML-tools>.

Table 1. RMSE and MAE for the same-pixel features from previous months, using seven different configurations and five different machine-learning techniques.

Features	MAE XGB	RMSE XGB	MAE LR	RMSE LR	MAE RF	RMSE RF	MAE MLP	RMSE MLP	MAE SVR	RMSE SVR	MAE Mean	RMSE Mean
a	2.887	5.790	2.915	5.649	2.911	5.878	2.843	5.639	2.700	5.720	2.851	5.735
b	2.890	6.064	2.840	5.642	3.047	6.402	3.008	5.952	2.677	5.895	2.892	5.991
c	2.912	6.078	2.909	5.630	3.048	6.255	2.782	5.844	2.640	5.861	2.858	5.933
d	2.928	6.145	2.900	5.625	3.074	6.407	2.844	5.657	2.626	5.857	2.874	5.938
e	2.890	6.060	2.913	5.621	3.034	6.351	2.829	5.723	2.617	5.751	2.856	5.901
f	2.957	6.104	2.942	5.641	3.065	6.293	2.763	5.655	2.616	5.710	2.868	5.880
g	2.936	6.065	2.933	5.628	2.954	5.981	2.826	5.803	2.617	5.685	2.853	5.832

Table 2. RMSE and MAE using the same-pixel features from Table 1, plus an additional pixel location feature, using five different machine-learning techniques.

Features	MAE XGB	RMSE XGB	MAE LR	RMSE LR	MAE RF	RMSE RF	MAE MLP	RMSE MLP	MAE SVR	RMSE SVR	MAE Mean	RMSE Mean
a + i, j	2.655	5.571	2.655	5.571	2.996	6.358	2.73	5.540	2.436	5.413	2.694	5.690
b + i, j	2.736	5.884	2.736	5.884	2.893	6.057	2.838	5.809	2.526	5.657	2.745	5.858
c + i, j	2.716	5.763	2.716	5.763	2.781	5.736	2.838	5.908	2.493	5.625	2.708	5.750
d + i, j	2.805	5.983	2.805	5.983	2.759	5.770	2.760	5.594	2.479	5.626	2.721	5.791
e + i, j	2.753	5.904	2.753	5.904	2.714	5.668	2.838	5.809	2.481	5.565	2.707	5.770
f + i, j	2.844	5.890	2.844	5.890	2.811	5.806	2.860	5.907	2.491	5.592	2.770	5.817
g + i, j	2.887	5.996	2.887	5.996	2.804	5.679	2.813	5.742	2.529	5.607	2.784	5.804

Table 3. RMSE and MAE using the same feature sets as Table 2 plus time stamp, using five different machine-learning techniques.

Features	MAE XGB	RMSE XGB	MAE LR	RMSE LR	MAE RF	RMSE RF	MAE MLP	RMSE MLP	MAE SVR	RMSE SVR	MAE Mean	RMSE Mean
a + i, j + t	2.478	5.859	2.967	5.682	2.567	5.954	2.872	5.893	2.377	5.342	2.652	5.746
b + i, j + t	2.481	5.742	2.867	5.658	2.653	5.769	2.807	5.781	2.445	5.559	2.650	5.701
c + i, j + t	2.514	5.834	2.933	5.641	2.587	5.595	2.95	6.272	2.456	5.588	2.680	5.786
d + i, j + t	2.576	5.879	2.924	5.637	2.609	5.634	2.771	5.903	2.440	5.602	2.660	5.731
e + i, j + t	2.598	5.946	2.94	5.633	2.620	5.613	2.945	6.263	2.451	5.540	2.710	5.799
f + i, j + t	2.758	6.092	2.962	5.645	2.700	5.689	2.882	6.159	2.474	5.573	2.755	5.831
g + i, j + t	2.724	5.936	2.954	5.634	2.621	5.519	2.912	5.843	2.491	5.580	2.740	5.702

Table 4. RMSE and MAE using the same feature sets as Table 3 and square root rescaling, using five different machine-learning techniques.

Features	MAE XGB	RMSE XGB	MAE LR	RMSE LR	MAE RF	RMSE RF	MAE MLP	RMSE MLP	MAE SVR	RMSE SVR	MAE Mean	RMSE Mean
a + i, j + t + s	2.342	5.544	2.857	5.582	2.536	5.897	2.490	5.598	2.542	5.313	2.553	5.586
b + i, j + t + s	2.438	5.682	2.788	5.612	2.612	5.821	2.598	5.661	2.503	5.326	2.587	5.620
c + i, j + t + s	2.417	5.602	2.797	5.575	2.558	5.668	2.633	5.634	2.450	5.275	2.571	5.550
d + i, j + t + s	2.539	5.816	2.785	5.571	2.557	5.670	2.726	5.870	2.455	5.291	2.612	5.643
e + i, j + t + s	2.554	5.757	2.796	5.550	2.569	5.629	2.945	6.039	2.455	5.289	2.663	5.652
f + i, j + t + s	2.596	5.839	2.818	5.565	2.628	5.642	2.942	6.283	2.477	5.301	2.692	5.726
g + i, j + t + s	2.557	5.639	2.811	5.553	2.631	5.632	2.859	5.964	2.477	5.315	2.667	5.620

5.2. Performance Comparisons

The data in Tables 1–4 are summarized in Figures 6 and 7. For MAE, XGB with all features (including same-pixel, spatial-temporal, and square root rescaling) gives the overall best performance. However, SVR is clearly the best performer for most feature sets. SVR tends to work best with fewest same-pixel features (i.e., configuration a, which is the previous month + 12 months prior). SVR with configuration a + (i, j) reduces the MAE by about 7% over the best result without spatial features. Adding time stamp and square root rescaling gives little additional improvement. In general, SVR gave MAE values that were between 7 and 20 % better than the worst-performing algorithms, which were

random forest (for same-pixel and same-pixel + spatial location feature sets) or linear regression (for other feature sets). It is noteworthy that adding a time stamp brought large error reductions to random forest and XGB, while having little effect on linear regression, MLP, and SVR.

The RMSE results resemble the MAE results in that SVR consistently gives the lowest error. This time however, XGB with **a + i, j + t + s** does not outperform the SVR results. Once again, same-pixel configuration **a** tends to give the best results for SVR, and **a + i, j** has about 4% lower error than **a** only. In general, SVR gave MAE values that were 2.5–15.5% better than the worst-performing algorithms.

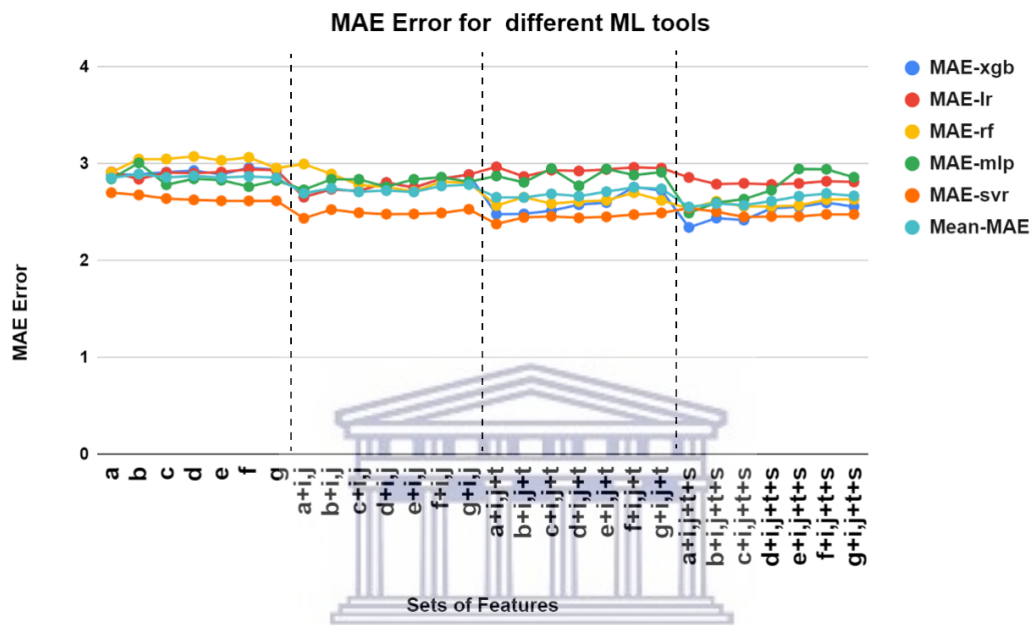


Figure 6. MAE Graph for the different set configurations.

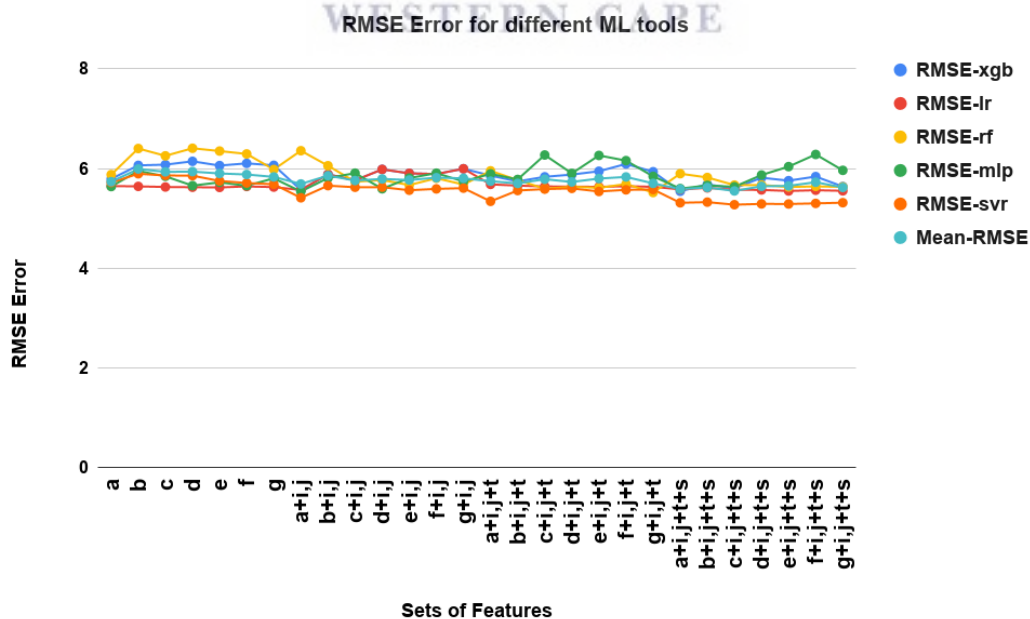


Figure 7. RMSE Graph for the different set configurations.

Figure 8, shows the percentage performance improvement (i.e., percentage error reduction) of the overall best-performing model from Tables 1–4 (SVR) compared to the untrained model based on the previous month. The MAE and RMSE values for the untrained model were 2.988 and 6.771,

respectively. When square root rescaling is used, all predictions reduced MAE and RMSE by over 15% and 20%, respectively. The overall best predictor ($a + i + j + t$) reduced both MAE and RMSE by more than 20%. This result is consistent with [39], which found that SVM outperformed RF when predicting rainfall images up to 3 h ahead on a per-pixel basis.

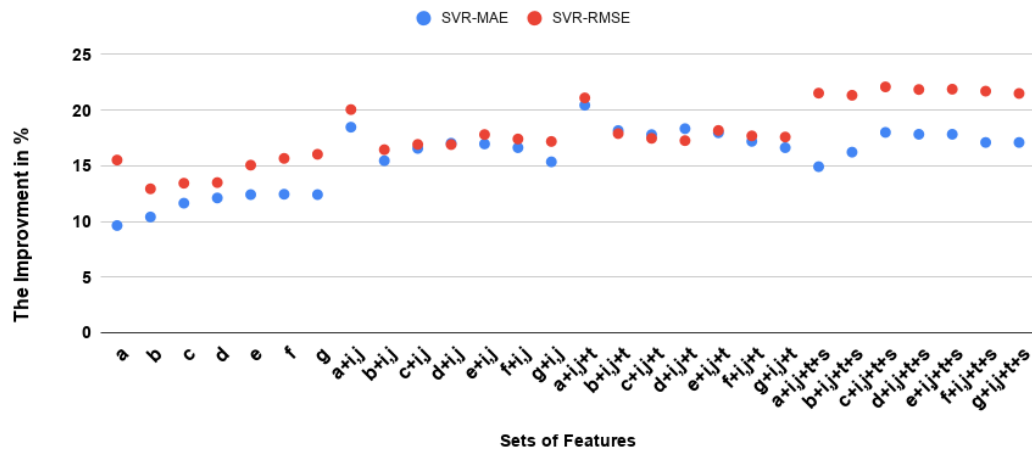


Figure 8. Performance improvement of SVR versus the untrained previous month regressor.

When GMM was added to XGB with $a + i, j + t + s$, the values of MAE and RMSE obtained were 2.258 and 4.838, respectively. These values were better than the corresponding best results without GMM by 3.6% and 9.4% respectively. Compared to the untrained predictor, this XGB+GMM model gave 25% improvement in MAE and 29% improvement in RMSE. Figure 9, shows an example of an actual image and its prediction using XGB+GMM.

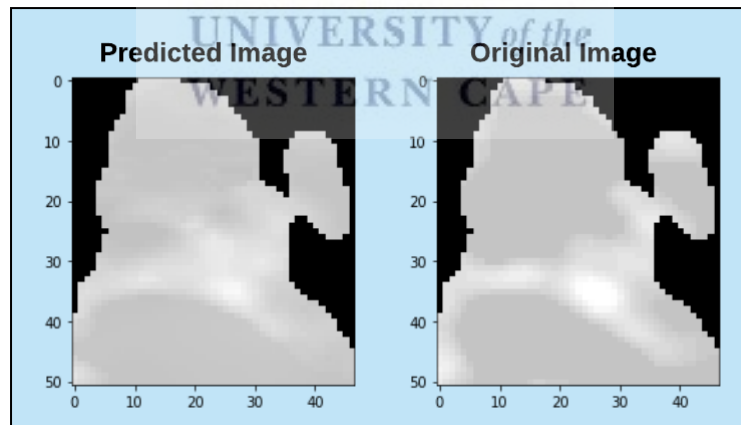


Figure 9. Example of an image prediction made with the model (XGB+GMM).

All of the above results were obtained using the default parameters in sklearn for their respective methods. Parameters were not optimized because of the large number of different methods involved. In particular, optimizing GMM is very expensive since it used 100 trained models which would all have to be optimized separately. We did conduct a preliminary investigation into parameter optimization by tuning parameters used in RF and XGB for the models in Tables 1–3. For this purpose, the scikit-optimize package was used, which employs Bayesian optimization. Improvements in MAE and RMSE were less than 6.5%, and still fell short of the performance obtained with GMMs without parameter optimization.

Figure 10 gives residual plots and R^2 values for the best XGB+GMM ($a + i, j + t + s$) model and the best SVR model ($a + i, j + t$), superimposed on the untrained model residuals. Visually, XGB+GMM and SVR are giving predictions closer to the 45° line than the untrained regressor, while the R^2 values

are more than doubled. Figure 11 presents Regression Error Characteristics (REC) curves for the same three models. For XGB+GMM, 85% of pixel predictions have a deviation of 5 or less.

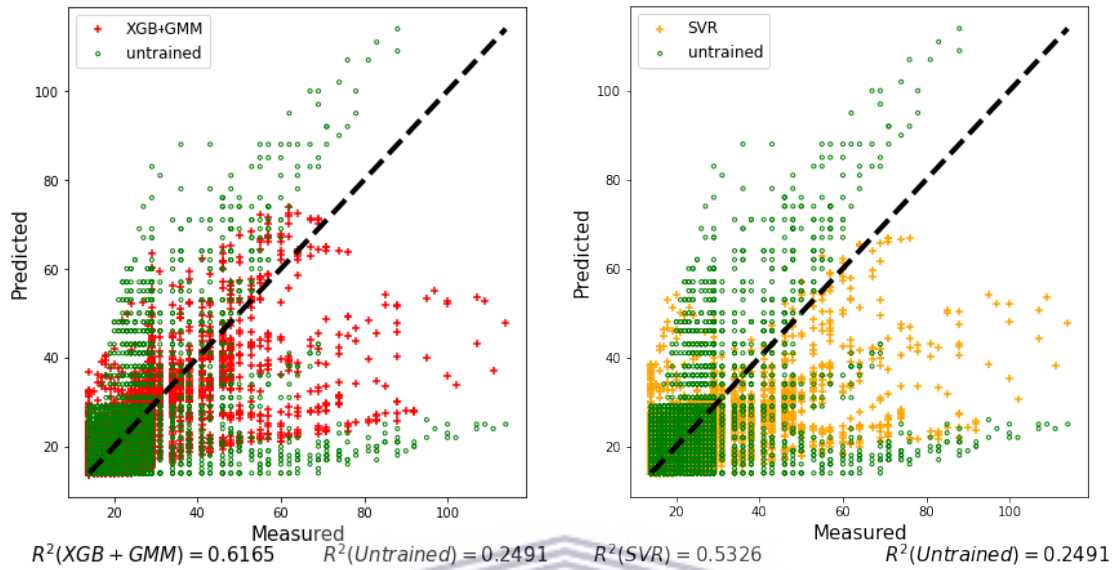


Figure 10. Residual plots and R^2 values for XGB+GMM versus untrained predictor (left), and best SVR model versus the untrained (right).

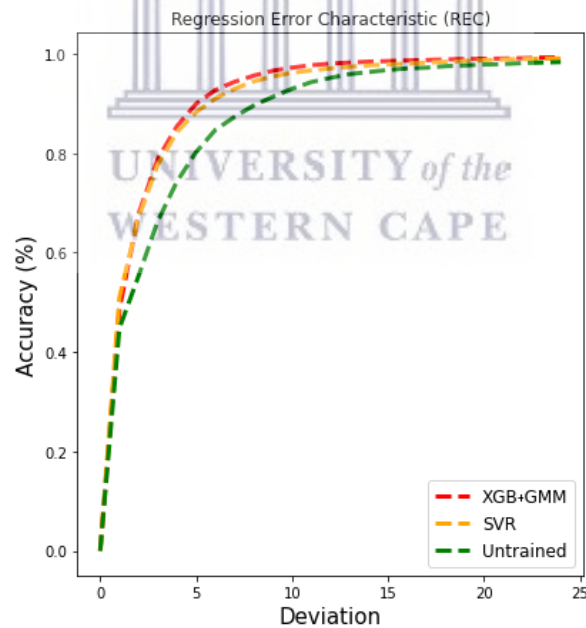


Figure 11. Regression error characteristic (REC) curves for the best XGB+GMM, and SVR models, together with the untrained regressor.

6. Conclusions

This paper investigated the automatic prediction of groundwater Δ TWS in the GRACE dataset. The proposed approach uses a regression-based approach to predict full groundwater images based on sequences of monthly groundwater maps.

Our results show that the application of appropriate machine-learning techniques can yield significantly more accurate predictions. In particular, it was shown that using SVR as a predictor, automatically selected previous same-pixel values and time stamp as features, and square root rescaling

all contributed to better overall prediction outcomes. Global features constructed from GMMs fitted to the pixel intensity distribution brought further improvements.

In future work we will apply these methods to other regions and meteorological parameters such as rainfall, temperature, air pressure, humidity etc. We shall also explore possible improvements to the method such as better imputation of missing values and the investigation of other global features. Additionally, we shall extend these methods to the joint prediction of multiple parameters.

Author Contributions: Conceptualization, M.G., A.B. and E.A.H.; methodology, M.G. and E.A.H.; software, E.A.H.; validation, M.G. and C.T.; formal analysis, C.T.; computing resources, M.G., M.V. and A.B.; writing—original draft preparation, E.A.H.; writing—review and editing, M.G., M.V., C.T. and A.B.; visualization, M.G., C.T. and E.A.H.; supervision, M.G., C.T. and A.B.; funding acquisition, M.G. and M.V. All authors have read and agreed to the published version of the manuscript.

Funding: E.A.H. acknowledges financial support from the South African National Research Foundation, and the Telkom-Openserve-Aria Technologies center of Excellence at Computer Science at UWC.

Acknowledgments: Meerkat Cluster was made use of during this study, (<http://docs.meerkat.uwc.ac.za/>) provided by the University of the Western Cape’s eRe-search Office (<https://eresearch.uwc.ac.za/>).

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

References

- Levy, J.; Xu, Y. Groundwater management and groundwater/surface-water interaction in the context of South African water policy. *Hydrogeol. J.* **2012**, *20*, 205–226. [[CrossRef](#)]
- Braune, E.; Xu, Y. Groundwater management issues in Southern Africa—An IWRM perspective. *Water SA* **2008**, *34*, 699–706. [[CrossRef](#)]
- Ghasemian, D. Groundwater Management Using Remotely Sensed Data in High Plains Aquifer. Ph.D. Thesis, The University of Arizona, Tucson, AZ, USA, 2016.
- Cao, G.; Zheng, C.; Scanlon, B.R.; Liu, J.; Li, W. Use of flow modeling to assess sustainability of groundwater resources in the North China Plain. *Water Resour. Res.* **2013**, *49*, 159–175. [[CrossRef](#)]
- Assembly, U.N.G. Transforming Our World: The 2030 Agenda for Sustainable Development. 2015. Available online: <http://www.naturalcapital.vn/wp-content/uploads/2017/02/UNDP-Viet-Nam.pdf> (accessed on 15 July 2020).
- Felix Landerer. *JPL TELLUS GRACE Level-3 Monthly Land Water-Equivalent-Thickness Surface Mass Anomaly Release 6.0 Version 03 in netCDF/ASCII/GeoTIFF Formats*; Ver. RL06 v03; PO.DAAC: Pasadena, CA, USA, 2020. [[CrossRef](#)]
- Natkhin, M.; Steidl, J.; Dietrich, O.; Dannowski, R.; Lischeid, G. Differentiating between climate effects and forest growth dynamics effects on decreasing groundwater recharge in a lowland region in Northeast Germany. *J. Hydrol.* **2012**, *448*, 245–254. [[CrossRef](#)]
- Goderniaux, P.; Brouyère, S.; Wildemeersch, S.; Therrien, R.; Dassargues, A. Uncertainty of climate change impact on groundwater reserves—Application to a chalk aquifer. *J. Hydrol.* **2015**, *528*, 108–121. [[CrossRef](#)]
- Yadav, B.; Ch, S.; Mathur, S.; Adamowski, J. Assessing the suitability of extreme learning machines (ELM) for groundwater level prediction. *J. Water Land Dev.* **2017**, *32*, 103–112. [[CrossRef](#)]
- Lo, M.H.; Famiglietti, J.S.; Yeh, P.F.; Syed, T. Improving parameter estimation and water table depth simulation in a land surface model using GRACE water storage and estimated base flow data. *Water Resour. Res.* **2010**, *46*. [[CrossRef](#)]
- Zhou, T.; Wang, F.; Yang, Z. Comparative analysis of ANN and SVM models combined with wavelet preprocess for groundwater depth prediction. *Water* **2017**, *9*, 781. [[CrossRef](#)]
- Adamowski, J.; Fung Chan, H.; Prasher, S.O.; Ozga-Zielinski, B.; Sliusarieva, A. Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resour. Res.* **2012**, *48*. [[CrossRef](#)]
- Sahoo, S.; Jha, M.K. Groundwater-level prediction using multiple linear regression and artificial neural network techniques: A comparative assessment. *Hydrogeol. J.* **2013**, *21*, 1865–1887. [[CrossRef](#)]

14. Bourennane, H.; King, D.; Couturier, A. Comparison of kriging with external drift and simple linear regression for predicting soil horizon thickness with different sample densities. *Geoderma* **2000**, *97*, 255–271. [[CrossRef](#)]
15. Tiwari, M.K.; Adamowski, J. Urban water demand forecasting and uncertainty assessment using ensemble wavelet-bootstrap-neural network models. *Water Resour. Res.* **2013**, *49*, 6486–6507. [[CrossRef](#)]
16. Arandia, E.; Ba, A.; Eck, B.; McKenna, S. Tailoring seasonal time series models to forecast short-term water demand. *J. Water Resour. Plan. Manag.* **2016**, *142*, 04015067. [[CrossRef](#)]
17. Mirzavand, M.; Ghazavi, R. A stochastic modelling technique for groundwater level forecasting in an arid environment using time series methods. *Water Resour. Manag.* **2015**, *29*, 1315–1328. [[CrossRef](#)]
18. Nielsen, A. *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*; O'Reilly: Sebastopol, CA, USA, 2020.
19. Yoon, H.; Jun, S.C.; Hyun, Y.; Bae, G.O.; Lee, K.K. A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer. *J. Hydrol.* **2011**, *396*, 128–138. [[CrossRef](#)]
20. Sun, A.Y. Predicting groundwater level changes using GRACE data. *Water Resour. Res.* **2013**, *49*, 5900–5912. [[CrossRef](#)]
21. Emamgholizadeh, S.; Moslemi, K.; Karami, G. Prediction the groundwater level of bastam plain (Iran) by artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS). *Water Resour. Manag.* **2014**, *28*, 5433–5446. [[CrossRef](#)]
22. Moosavi, V.; Vafakhah, M.; Shirmohammadi, B.; Behnia, N. A wavelet-ANFIS hybrid model for groundwater level forecasting for different prediction periods. *Water Resour. Manag.* **2013**, *27*, 1301–1321. [[CrossRef](#)]
23. Dos Santos, C.C.; Pereira Filho, A.J. Water demand forecasting model for the metropolitan area of São Paulo, Brazil. *Water Resour. Manag.* **2014**, *28*, 4401–4414. [[CrossRef](#)]
24. Huang, F.; Huang, J.; Jiang, S.H.; Zhou, C. Prediction of groundwater levels using evidence of chaos and support vector machine. *J. Hydroinform.* **2017**, *19*, 586–606. [[CrossRef](#)]
25. Rahaman, M.M.; Thakur, B.; Kalra, A.; Li, R.; Maheshwari, P. Estimating High-Resolution Groundwater Storage from GRACE: A Random Forest Approach. *Environments* **2019**, *6*, 63. [[CrossRef](#)]
26. Jing, W.; Yao, L.; Zhao, X.; Zhang, P.; Liu, Y.; Xia, X.; Song, J.; Yang, J.; Li, Y.; Zhou, C. Understanding terrestrial water storage declining trends in the Yellow River Basin. *J. Geophys. Res. Atmos.* **2019**, *124*, 12963–12984. [[CrossRef](#)]
27. Jing, W.; Zhao, X.; Yao, L.; Di, L.; Yang, J.; Li, Y.; Guo, L.; Zhou, C. Can terrestrial water storage dynamics be estimated from climate anomalies? *Earth Space Sci.* **2020**, *7*, e2019EA000959. [[CrossRef](#)]
28. Sahour, H.; Sultan, M.; Vazifedan, M.; Abdelmohsen, K.; Karki, S.; Yellich, J.A.; Gebremichael, E.; Alshehri, F.; Elbayoumi, T.M. Statistical applications to downscale GRACE-derived terrestrial water storage data and to fill temporal gaps. *Remote Sens.* **2020**, *12*, 533. [[CrossRef](#)]
29. Mukherjee, A.; Ramachandran, P. Prediction of GWL with the help of GRACE TWS for unevenly spaced time series data in India: Analysis of comparative performances of SVR, ANN and LRM. *J. Hydrol.* **2018**, *558*, 647–658. [[CrossRef](#)]
30. Seyoum, W.M.; Kwon, D.; Milewski, A.M. Downscaling GRACE TWSA data into high-resolution groundwater level anomaly using machine learning-based models in a glacial aquifer system. *Remote Sens.* **2019**, *11*, 824. [[CrossRef](#)]
31. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, Canada, 7–12 December 2015; pp. 802–810.
32. Shi, E.; Li, Q.; Gu, D.; Zhao, Z. A Method of Weather Radar Echo Extrapolation Based on Convolutional Neural Networks. In *MultiMedia Modeling (MMM 2018)*; Schoeffmann, K., Ed.; Springer: Cham, Switzerland, 2018; Volume 10704, pp. 16–28.
33. Shi, X.; Gao, Z.; Lausen, L.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Deep learning for precipitation nowcasting: A benchmark and a new model. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5617–5627.
34. Tran, Q.K.; Song, S.k. Multi-Channel Weather Radar Echo Extrapolation with Convolutional Recurrent Neural Networks. *Remote Sens.* **2019**, *11*, 2303. [[CrossRef](#)]

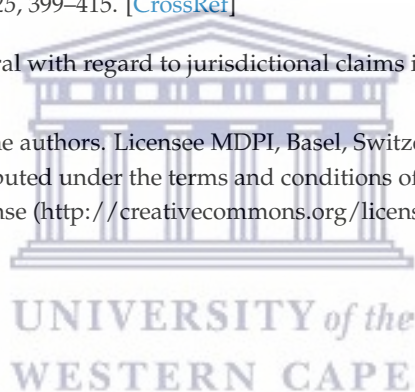
35. Wang, Y.; Long, M.; Wang, J.; Gao, Z.; Philip, S.Y. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 879–888.
36. Singh, S.; Sarkar, S.; Mitra, P. A deep learning based approach with adversarial regularization for Doppler weather radar ECHO prediction. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; pp. 5205–5208.
37. Chen, L.; Cao, Y.; Ma, L.; Zhang, J. A Deep Learning Based Methodology for Precipitation Nowcasting with Radar. *Earth Space Sci.* **2020**, *7*, e2019EA000812. [[CrossRef](#)]
38. D’Isanto, A.; Cavuoti, S.; Gieseke, F.; Polsterer, K.L. Return of the features—Efficient feature selection and interpretation for photometric redshifts. *Astron. Astrophys.* **2018**, *616*, A97. [[CrossRef](#)]
39. Yu, P.S.; Yang, T.C.; Chen, S.Y.; Kuo, C.M.; Tseng, H.W. Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting. *J. Hydrol.* **2017**, *552*, 92–104. [[CrossRef](#)]
40. Mukhopadhyay, A.; Shukla, B.P.; Mukherjee, D.; Chanda, B. A novel neural network based meteorological image prediction from a given sequence of images. In Proceedings of the 2011 Second International Conference on Emerging Applications of Information Technology, Kolkata, India, 19–20 February 2011; pp. 202–205.
41. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
42. Khademi, F.; Jamal, S.M.; Deshpande, N.; Londhe, S. Predicting strength of recycled aggregate concrete using artificial neural network, adaptive neuro-fuzzy inference system and multiple linear regression. *Int. J. Sustain. Built Environ.* **2016**, *5*, 355–369. [[CrossRef](#)]
43. Bengio, Y.; Goodfellow, I.; Courville, A. *Deep Learning*; MIT Press: New York, NY, USA, 2017; Volume 1.
44. Liakos, K.G.; Busato, P.; Moshou, D.; Pearson, S.; Bochtis, D. Machine learning in agriculture: A review. *Sensors* **2018**, *18*, 2674. [[CrossRef](#)]
45. Ding, S.; Li, H.; Su, C.; Yu, J.; Jin, F. Evolutionary artificial neural networks: a review. *Artif. Intell. Rev.* **2013**, *39*, 251–260. [[CrossRef](#)]
46. Kolluru, V.; Ussenaiah, M. A survey on classification techniques used for rainfall forecasting. *Int. J. Adv. Res. Comput. Sci.* **2017**, *8*, 226–229.
47. Voyant, C.; Notton, G.; Kalogirou, S.; Nivet, M.L.; Paoli, C.; Motte, F.; Fouilloy, A. Machine learning methods for solar radiation forecasting: A review. *Renew. Energy* **2017**, *105*, 569–582. [[CrossRef](#)]
48. Liaw, A.; Wiener, M. Classification and regression by randomForest. *R News* **2002**, *2*, 18–22.
49. Farnaaz, N.; Jabbar, M. Random forest modeling for network intrusion detection system. *Procedia Comput. Sci.* **2016**, *89*, 213–217. [[CrossRef](#)]
50. Brownlee, J. *XGBoost with Python*, 1.10 ed.; Machine Learning Mastery; Machine Learning Mastery Pty: Victoria, Australia, 2018.
51. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Science & Business Media: Berlin, Germany, 2009.
52. Chau, K.; Wu, C. A hybrid model coupled with singular spectrum analysis for daily rainfall prediction. *J. Hydroinform.* **2010**, *12*, 458–473. [[CrossRef](#)]
53. Oğcu, G.; Demirel, O.F.; Zaim, S. Forecasting electricity consumption with neural networks and support vector regression. *Procedia Soc. Behav. Sci.* **2012**, *58*, 1576–1585. [[CrossRef](#)]
54. Hsu, C.C.; Wu, C.H.; Chen, S.C.; Peng, K.L. Dynamically optimizing parameters in support vector regression: An application of electricity load forecasting. In Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS’06), Kauia, HI, USA, 4–7 January 2006; Volume 2, p. 30c.
55. Cheng, C.S.; Chen, P.W.; Huang, K.K. Estimating the shift size in the process mean with support vector regression and neural networks. *Expert Syst. Appl.* **2011**, *38*, 10624–10630. [[CrossRef](#)]
56. Zeng, J.; Xie, L.; Liu, Z.Q. Type-2 fuzzy Gaussian mixture models. *Pattern Recognit.* **2008**, *41*, 3636–3643. [[CrossRef](#)]
57. Reynolds, D.A. Gaussian Mixture Models. 2009. Available online: http://leap.ee.iisc.ac.in/sriram/teaching/MLSP_16/refs/GMM_Tutorial_Reynolds.pdf (accessed on 15 July 2020).
58. Tran, D.; Le, T.V.; Wagner, M. Fuzzy Gaussian mixture models for speaker recognition. In Proceedings of the Fifth International Conference on Spoken Language Processing, Sydney, Australia, 30 November–4 December 1998.

59. Brassington, G. Mean Absolute Error and Root Mean Square Error: Which Is the Better Metric for Assessing Model Performance? 2017. Available online: <https://meetingorganizer.copernicus.org/EGU2017/EGU2017-3574.pdf> (accessed on 15 July 2020).
60. 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](#)]
61. Mukhopadhyay, A.; Shukla, B.P.; Mukherjee, D.; Chanda, B. Prediction of meteorological images based on relaxation labeling and artificial neural network from a given sequence of images. In Proceedings of the 2012 International Conference on Computer Communication and Informatics, Coimbatore, India, 10–12 January 2012; pp. 1–5.
62. Mehr, A.D.; Nourani, V.; Khosrowshahi, V.K.; Ghorbani, M.A. A hybrid support vector regression—Firefly model for monthly rainfall forecasting. *Int. J. Environ. Sci. Technol.* **2019**, *16*, 335–346. [[CrossRef](#)]
63. Nourani, V.; Uzelaltinbulat, S.; Sadikoglu, F.; Behfar, N. Artificial intelligence based ensemble modeling for multi-station prediction of precipitation. *Atmosphere* **2019**, *10*, 80. [[CrossRef](#)]
64. Fiennen, M.N.; Nolan, B.T.; Feinstein, D.T. Evaluating the sources of water to wells: Three techniques for metamodeling of a groundwater flow model. *Environ. Model. Softw.* **2016**, *77*, 95–107. [[CrossRef](#)]
65. Abudu, S.; Cui, C.; King, J.P.; Moreno, J.; Bawazir, A.S. Modeling of daily pan evaporation using partial least squares regression. *Sci. China Technol. Sci.* **2011**, *54*, 163–174. [[CrossRef](#)]
66. Pinheiro, A.; Vidakovic, B. Estimating the square root of a density via compactly supported wavelets. *Comput. Stat. Data Anal.* **1997**, *25*, 399–415. [[CrossRef](#)]

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



© 2020 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 (<http://creativecommons.org/licenses/by/4.0/>).



Chapter 5

Manuscript “Basic Statistical Estimation Outperforms Machine Learning in Monthly Prediction of Seasonal Climatic Parameters”



Article

Basic Statistical Estimation Outperforms Machine Learning in Monthly Prediction of Seasonal Climatic Parameters

Eslam A. Hussein ^{1,*} , Mehrdad Ghaziasgar ¹ , Christopher Thron ² , Mattia Vaccari ³  and Antoine Bagula ¹ 

- ¹ Department of Computer Science, University of the Western Cape, Cape Town 7535, South Africa; mghaziasgar@uwc.ac.za (M.G.); abagula@uwc.ac.za (A.B.)
- ² Department of Science and Mathematics, Texas A&M University-Central Texas, Killeen, TX 76549, USA; thron@tamuct.edu
- ³ Department of Physics and Astronomy, University of the Western Cape, Cape Town 7535, South Africa; mvaccari@uwc.ac.za
- * Correspondence: ehusein@uwc.ac.za

Abstract: Machine learning (ML) has been utilized to predict climatic parameters, and many successes have been reported in the literature. In this paper, we scrutinize the effectiveness of five widely used ML algorithms in the monthly prediction of seasonal climatic parameters using monthly image data. Specifically, we quantify the predictive performance of these algorithms applied to five climatic parameters using various combinations of features. We compare the predictive accuracy of the resulting trained ML models to that of basic statistical estimators that are computed directly from the training data. Our results show that ML never significantly outperforms the statistical baseline, and underperforms for most feature sets. Unlike previous similar studies, we provide error bars for the relative performance of different predictors based on jackknife estimates applied to differences in predictive error magnitudes. We also show that the practice of shuffling data sequences which was employed in some previous references leads to data leakage, resulting in over-estimated performance. Ultimately, the paper demonstrates the importance of using well-grounded statistical techniques when producing and analyzing the results of ML predictive models.

Keywords: geophysical image data; high-dimensional data analysis; prediction; statistical modeling; baselining; evaluation; data leakage; seasonality; uncertainty quantification; jackknife



Citation: Hussein, E.A.; Ghaziasgar, M.; Thron, C.; Vaccari, M.; Bagula, A. Basic Statistical Estimation Outperforms Machine Learning in Monthly Prediction of Seasonal Climatic Parameters. *Atmosphere* **2021**, *12*, 539. <https://doi.org/10.3390/atmos12050539>

Academic Editor: Sándor Baran and Annette Möller

Received: 25 March 2021
Accepted: 14 April 2021
Published: 23 April 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/>).

1. Introduction

Recent advances in computing have shifted the focus of scientific communities from a data-scarce to a data-rich research environment [1]. This paradigm shift, known as the fourth paradigm of science, and often referred to as the era of “big data” [2], has emerged from the move of big data and AI into our daily lives and the pervasiveness of these two technologies, which are (i) leading to an explosion in innovation, competition, and productivity [3], (ii) causing a dramatic shift to data-driven research [4], and (iii) unleashing the benefits of data-intensive applications.

Climate science is a research field where data-driven models based on machine learning (ML) have become popular [5]. A major focus of climate science is the understanding and prediction of climate parameters such as rainfall and temperature [6] and many others. For many practical climate-influenced decisions where prediction times of months to a decade are likely to be the most important [7], providing accurate models to predict climatic parameters on these time scales is critical. The remarkable successes of ML and deep learning in a variety of fields such as computer vision and natural language processing suggests that this success may be extended to climate science as well.

However, there is a concern regarding how effective and legitimate these ML models are to address real world applications in climate science. This is reason for enthusiasm, but also for skepticism, as it is all too common to make excessive claims for new techniques,

<http://etd.uwc.ac.za/>

which turn out not to live up to their initial promise, as exemplified by Gartner's hype cycle model [8]. There are already several examples in the literature that show that ML does not always live up to its hype. A recent overview study reviewed several papers that used recurrent neural networks for top-*n* recommendation tasks, and found that a simple model using K-nearest neighbors outperformed most of the more sophisticated models [9]. One major deficiency identified by the study was the use of defective or weak baselines when quantifying the performance of newer proposed models. Other papers that also reached the conclusion that sophisticated ML models do not necessarily outperform simpler models include [10–13].

One key feature of ML methods is that they make no assumptions about the underlying distribution of inputs. This can be both an advantage and a disadvantage. The advantage is that ML methods can be applied to a wide variety of datasets without having detailed knowledge of the statistics of the individual datasets. The disadvantage is that ML may miss important characteristics of particular datasets. For this reason, if the user has some knowledge of the dataset's distribution, it is important to compare ML predictors with statistical estimates based on the presumed distribution. Such statistical estimates have the advantage that they are simple to calculate, require no training, and are easy to interpret [14].

One deep flaw in most papers in the literature is that accuracy estimates for ML methods are given (such as R^2 or root mean squared error) without providing error bars on these estimates. Hence, it is impossible to tell whether or not differences between methods are statistically significant. This may be one reason why different investigators often reach different conclusions about the relative effectiveness of different ML methods. For example, Armstrong et al. [15] concluded from an analysis in the context of ad hoc retrieval tasks that numerous published papers report mutually contradictory conclusions concerning ML model performance.

Another concern is that some common pre-processing practices produce data leakage, so that ML algorithm accuracies are over-reported. Some examples of such practices are: data shuffling, whereby researchers randomly shuffle the data [11,16–23]; data imputation methods that use statistics (such as averaging) calculated on the entire data set, including both training and testing [24–26]; and data transformations such as de-seasonalization that also use statistics calculated on the whole dataset [27,28].

It is necessary to investigate the robustness of ML models in different fields of application. The current study is aimed at investigating the above mentioned deficiencies in the area of climatic seasonal parameters. This paper is organized as follows: Section 3 describes the data used; Section 4 discusses the methodology used for climatic parameter prediction; Section 5 shows the results obtained; Section 6 discusses the results; and Section 7 furnishes the conclusions.

2. Literature Review and Scope of the Research

ML is widely used in climatology to construct predictive models based on sequential data [11]. A variety of types of input data are used, including satellite images or periodic samples from gauges or weather stations.

The studies in the literature can be largely divided into two categories in terms of the predicted output: those that predict one or more entire images which provide a visual representation of a given predicted climate parameter on spatial maps of a specific geographical area under review ("whole-image prediction"); and those that predict only a single output representing a given predicted climate parameter at a fixed location ("single-output prediction").

For whole-image prediction based on sequential images, convolutional neural networks (CNNs) and convolutional long-short term models (ConvLSTMs) are often used due to their ability to perform feature reduction on spatial information. However these models require very large datasets with tens of thousands of images, due the data-intensive training process. For this reason, CNNs and ConvLSTMs are mainly applied to data sets

with short time intervals of no more than a few minutes between data points, which are typically much larger than data sets with longer time intervals [29–40]. For single-output prediction, a wider range of ML tools and time frames have been used, from linear methods in [17,21,41,42], to ensemble methods in [43–45], to hybrid methods in [28,46–48], to deep models in [49–56] covering time scales from minutes to years.

From a practical point of view, usually the most important policy decisions involving climate require monthly predictions [7]. Relatively few studies exist which use image data to make monthly predictions [57,58]. When time scales on the order of months or longer are involved, datasets are typically much smaller than those involving shorter time scales. A broad range of ML methods are applied, from simple methods like multilinear regression (MLR) up to advanced neural networks models [13,16–18,20,21,24,25,46,47,49,59–62]. Because of the small data sets used, researchers often perform feature selection/reduction to avoid overfitting. Most often, the selected features in the literature are combinations of features derived from previous time steps in the data, for example, a parameter at month n may be predicted based one or more parameter values taken from months previous to n [25–27,46,63].

Because of the rotation of the earth around the sun, monthly time series data like rainfall exhibit a seasonal behavior on a yearly basis (exhibit a yearly periodicity) [64,65]. This is critical to address because traditional time series models tends to rely on the time series being stationary [64,66]. Hence, the authors in [64] saw it as necessary to remove the periodicity in a monthly time series data. They described three ways of going about this: (a) previous lag differencing, (b) seasonal referencing; and (c) monthly mean subtraction, where (c) was identified as the most suitable method for monthly time series data. However, we found that many papers dealing with monthly prediction of climate parameters did not transform the input data to remove seasonality. Some papers accommodate seasonality by including data from month $n - 12$ to predict parameters at month n [13,17,19,20,24,25,27,28,44,49,51,58,60,62,67]. Month n 's time stamp (defined as $n \bmod 12$) was used as a feature in [19,49], but is not common in the literature.

In a few papers, the authors subtracted the monthly mean averages computed from the whole data set [25,28], with the inclusion of data from month $n - 12$. This procedure disrupts the integrity of the data by causing data leakage, whereby information from the testing set is introduced into the training set. Other papers make no attempt to account for seasonality [18,22,23,46,48,61,68,69]. Evidently, there is no consistent procedure for dealing with the seasonality aspect of the data; this is one point that we address in this paper.

In the previous section we emphasized the importance of using simple baselines to provide benchmarks to compare with more complicated methods. According to [66], the simplest baseline for predicting time series is to use the previous lag. For short-term image data, the previous image is used as a naive predictor for the next image [36,37,40,70]. As for monthly data, using previous lags as a baseline is not a common practice. Instead, a variety of baselines are used. Some papers use MLR based on previous lags [13,16,17,25], while the authors in [45] used same-month averages. Some papers do not use simple baselines, but rather compare several variations or architectures of more advanced ML methods such as SVR or MLP [26,27,46,63]. In summary, simple baselines are not consistently used in the literature.

The Objectives of the paper are as follows: (a) perform seasonal grid prediction on multiple climatic parameters; (b) investigate multiple untrained baselines, and in particular using a statistical estimator derived from a simple statistical model of the image pixel distributions; (c) analyze the effectiveness of subtracting the seasonality using the monthly average calculated only on the training data; (d) investigate the common feature sets used in the literature; (e) calculate error bars on the relative prediction accuracies of different methods using jackknife estimation applied to pairwise differences between prediction errors; and (f) demonstrating the effect of data leakage on the reported performance.

Our results show that across all climatic parameters studied, a very limited feature set (time stamp with spatial information) without seasonal subtraction outperforms feature sets

that use previous lags, with or without seasonal subtraction. Furthermore, an untrained baseline based on a simple statistical model can outperform more sophisticated ML tools. Furthermore, handling data inappropriately so that data leakage occurs (as has been done in some previous papers) can lead to significant overestimation of predictive performance.

3. Data and Area of Interest

The climatic data were obtained from the NASA GESDISC data archive, which is accessible to users registered with NASA Earthdata [71]. The dataset used is obtained from the Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS). FLDAS contains monthly image data for 28 fields such as rainfall flux, evaporation, and temperature [72] with a spatial resolution of $0.1^\circ \times 0.1^\circ$. The data are archived in netCDF format, where it can be manipulated and displayed using freely available software packages within python and R. NASA also supplies a cross-platform application called Panoply that can be used to plot the data [71].

The downloaded data set for each parameter used contains 228 satellite frames on a monthly basis, between January 1982 and December 2000. Images depict the entire globe at a resolution of 1500×3600 pixels. Figure 1 shows a sample image of rainfall. In general, the images are color coded to provide information about the relevant parameter. In the current study, the climatic parameters used are rainfall, evaporation, humidity, temperature, and wind speed.

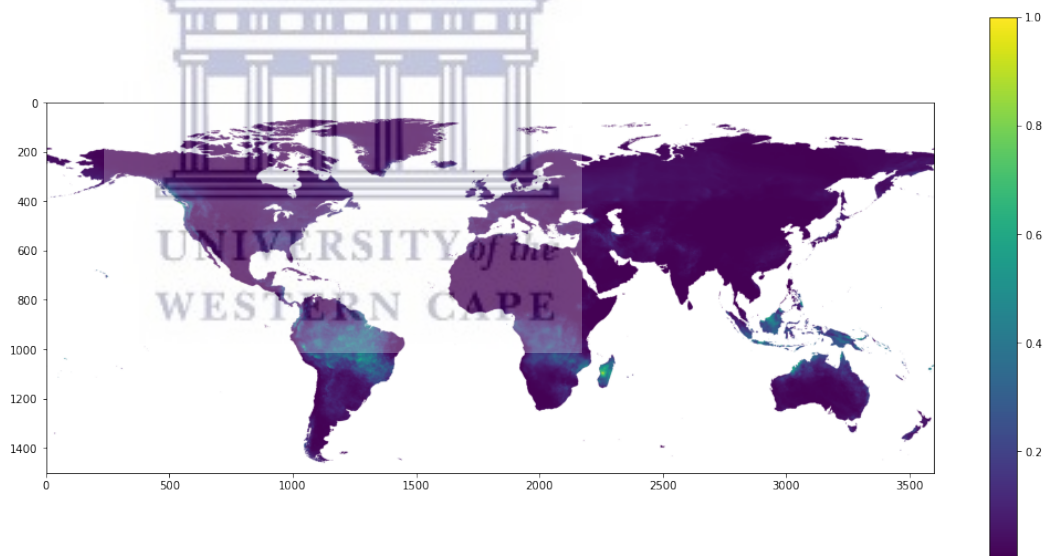


Figure 1. A sample full image of the rainfall dataset used in this research [72]. Color scale indicates normalized rainfall intensities.

To limit the computational load, we focused our prediction on Madagascar. Madagascar is the world's fourth largest island with an area of about $592,000 \text{ km}^2$ [73], and is separated from Mozambique on the main African continent by about 400 km [73]. The climate on the island is subtropical and is characterized by a dry season from May to October and rainy season from November to April [74,75]. Table 1 summarizes the characteristics of the Madagascar image data used in our study, which was extracted from the original FLDAS data.

Table 1. Properties of Madagascar image data (extracted from FLDAS dataset).

Property	Value
Latitude Extent	12°–26° S
Longitude Extent	43°–51° E
Spatial Resolution	0.1° × 0.1°
Temporal Resolution	Monthly
Temporal Coverage	January 1982 to December 2000
Dimension (lat × lon)	140 × 80

Madagascar is currently facing several challenges due to the potential impact of climate change on the agricultural sector, which can threaten food security [76–79], especially since farmers in the country are estimated to be 70% of the population [74]. Example images of the five climatic parameters used at a specific arbitrary timestamp are provided in Figure 2. The figure shows normalized values of five climatic parameters, namely rainfall, evaporation, humidity, temperature, and wind.

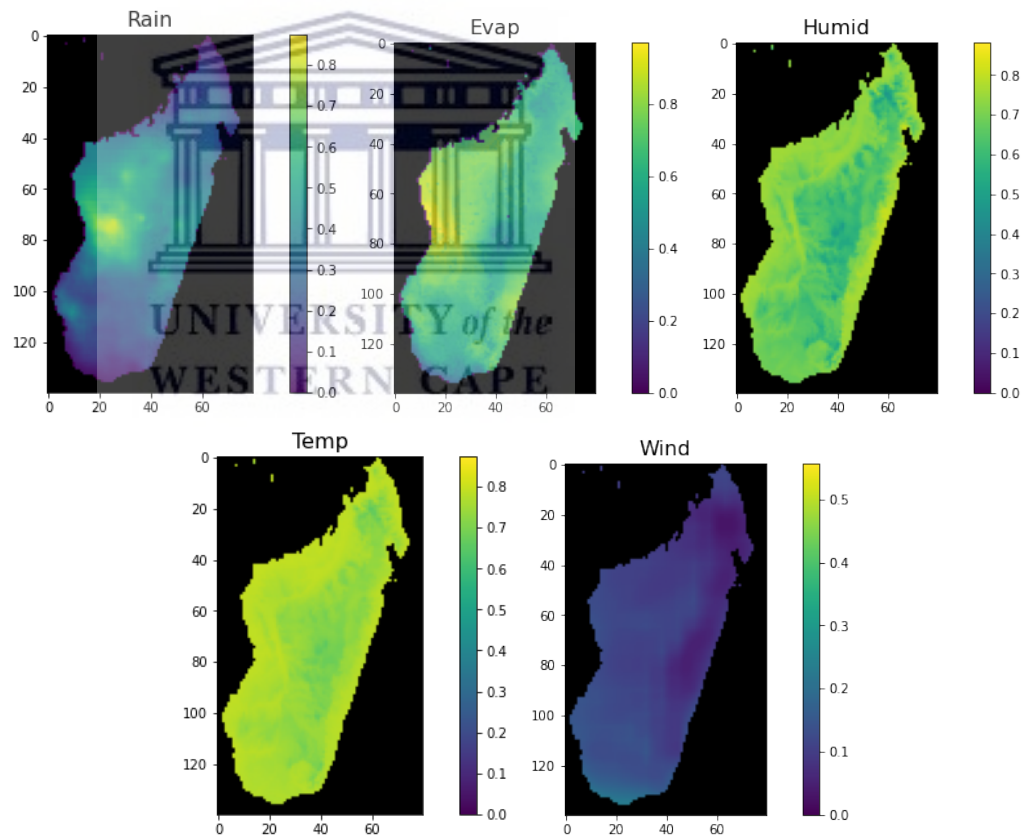


Figure 2. Images showing normalized values of five climatic parameters of Madagascar used in this study (left to right): rainfall (Rain), evaporation (Evap), humidity (Humid), temperature (Temp), and wind.

4. Methodology

Figure 3 shows a flowchart of the system created and used to make predictions in this research. The end goal of the system is to predict monthly rainfall, evaporation, humidity, temperature, and wind speed images on a pixel level, using a sequence of previous images as an input. The rest of this section describes the progression through the flowchart in the figure in detail: first we discuss the pre-processing of the images and the preparation of the data set; then we describe feature selection; and finally, we

indicate the tools used. The code together with the results are available on GitHub at <https://github.com/EslamHussein55/Climatic-parameters> (accessed on 16 April 2021).

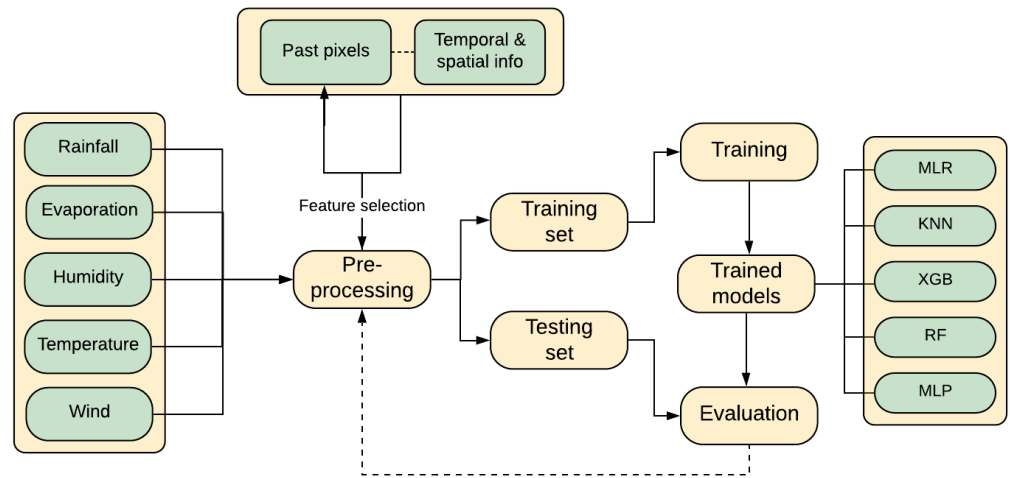


Figure 3. Flowchart showing the implementation process.

4.1. Image Pre-Processing

All images in the parameter datasets were cropped to a rectangle of size 140×80 that includes the Madagascar land area. We transformed the image pixels to greyscale (0–255) and re-sized the images to 70×40 to further reduce their complexity. In view of the fact that extreme values are a common occurrence in geophysical parameter data, pixel values were regularized by replacing them with their square roots, following the example of [57,80,81]. Since our study is concerned with relative performance of different algorithms rather than absolute performance, for simplicity we did not remove over-ocean pixels, which are constant in all images and hence perfectly predicted.

We mentioned previously that some authors recommend transforming time series data to remove seasonality, while many authors do not follow this recommendation. To evaluate the effectiveness of transforming time series, we created two input data sets (denoted as ‘raw’ and ‘de-seasonalized’) for each of the five parameters. The raw dataset contains the original data, while the de-seasonalized data is transformed by subtracting same-month averages. Care was taken to compute monthly averages based only on the training data to prevent data leakage. For illustrative purposes, Figure 4 shows example raw and de-seasonalized images for rainfall.

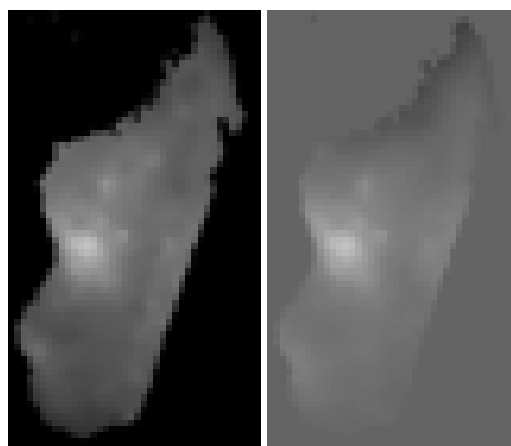


Figure 4. Pre-processed rainfall images (compare first image in Figure 2): raw image (left) and de-seasonalized image (right).

4.2. Data Preparation

We prepared the data in a sliding window fashion, similar to the following studies [35,82–85]. Figure 5 shows how pixels at a given location in a 12 month window are used to predict the corresponding pixel at the same location in the 13th month. In the figure, the symbols $\{f[0], \dots, f[11]\}$ refer, respectively, to the frames $\{12, \dots, 1\}$ months prior to the predicted frame, respectively.

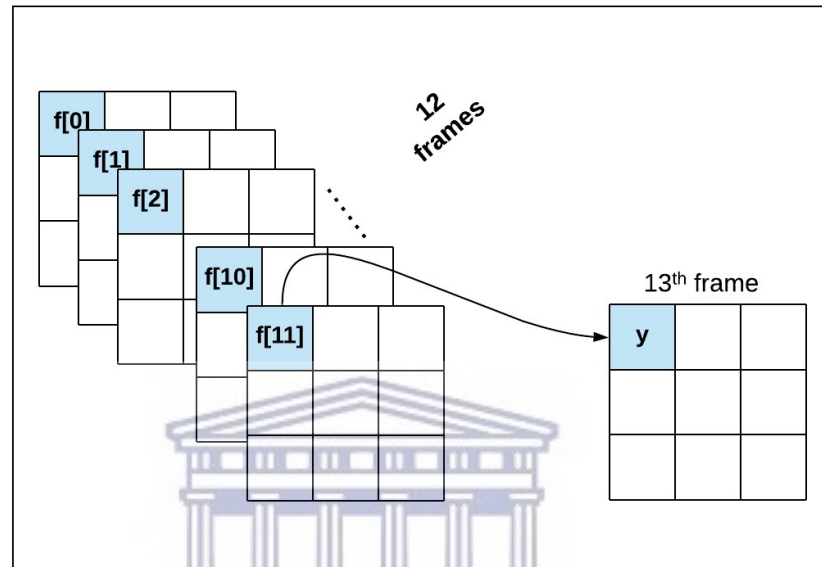


Figure 5. Overview of dataset preparation: Notation for same-pixel features used in image prediction.

As shown in Figure 5, the datasets were used to produce sequences consisting of 12 consecutive months. All datasets were divided into training and testing sets, where the training set was made up of sequences occurring earlier in the dataset, and the testing set made up of sequences that followed those in the training set. This technique of maintaining chronological order when dividing the datasets into training and testing sets helps avoid the problem of information leakage into the trained model from the future [66]. Applying the sliding window generated 216 sequences with the first 156 used for training, and the rest as testing. Although the number of images appears relatively small, the training task is nonetheless computationally expensive since the training process utilizes $156 \times 70 \times 40$ input vectors. This explains why previous similar studies also use relatively few images; for example [86] trains on only 47 images.

4.3. Feature Selection

Feature selection is critical to increasing training efficiency and model accuracy. Based on the reviewed literature for monthly prediction, we tested a variety of feature sets to understand the system mechanism. We also added in features systematically and assessed whether or not added features gave clearly better performances to ensure model parsimony and avoid overfitting [87]. The feature sets are described in the following subsections.

We first created a list of 12 candidate features for image prediction consisting of pixel values at the same location for the 12 prior months. To select a variety of these features, we prepared the data using the sliding window algorithm, where each 12-month window was used to predict the 13th month. Based on previous literature [13,17–20,22–28,44–49,58,59,61,62,67–69,88], we included the following feature sets:

- $f[0]$: same-pixel values from frames 12 months previous;
- $f[11]$: same-pixel values from the previous month;
- $f[0, 11]$: same-pixel values from 12 months previous and the previous month;

- **f[0, 1, 2, 11]**: same-pixel values from {12,11,10} months previous and the previous month;
- **f[0, 1, 10, 11]**: same-pixel values from {12,11} months previous and the previous two months.

Given the geographical variation and seasonal nature of the dataset used, the following spatio-temporal features are also used in this study:

- The (i, j) coordinates of the pixel of interest;
- Monthly time stamp $t \in \{0, \dots, 11\}$ where $\{0 = \text{January}, \dots, 11 = \text{December}\}$.

The five past-pixel feature sets and the two spatio-temporal features were combined to form the following feature set variants:

- Past-pixel features only (five variants, as listed above);
- (i, j) feature set only;
- (i, j, t) feature set only;
- Past-pixel features (five variants) plus (i, j, t) .

These 12 feature set variants were applied to both the raw and de-seasonalized training data.

4.4. Tools and Evaluation Methods

4.4.1. Machine Learning Algorithms

A total of five ML techniques are used for image prediction: (a) multivariate linear regression (MLR); (b) k -nearest neighbor (KNN); (c) random forest (RF); (d) extreme gradient boosting (XGB); (e) multilayer perceptron (MLP). Since the training set consisted of less than 200 sequences, we did not use deep learning, which typically requires much larger training sets [89–92]. For all ML tools except for MLR, parameters were optimized via grid search with three fold validation, using the time series cross-validator implemented in scikit-learn [93]. The purpose of cross-validation is to avoid overfitting by making sure that the model is not overly dependent on the particular training data used to construct the model. Additionally, for MLP, a regularization parameter was used as an additional measure to counteract overfitting. Grid search optimizations to optimize ML parameters were performed separately for each feature set applied to each climatic parameter used on the raw data and separately again on the de-seasonalized data. All optimized parameters for all ML tools can be found in the GitHub link provided above.

Altogether, a total of (5 climate parameters \times 2 data variants (raw/de-seasonalized) \times 12 feature set variants \times 5 ML tools) = 600 optimization experiments were performed.

4.4.2. Performance Metrics

One commonly used measure of the accuracy of a predictor's error is the mean absolute error (MAE). The MAE is calculated as the mean of the absolute values of prediction errors for all predicted pixel values:

$$\text{MAE} = \frac{1}{M} \sum_{m=1}^M |y_m^{\text{obs}} - y_m^{\text{pre}}| \quad (1)$$

where M is the number of observations, and y_m^{obs} and y_m^{pre} refer to the observed and predicted value of the m th output, respectively.

There is a long-running debate over whether or not MAE is superior to root mean squared error (RMSE) in geophysical studies [94–97]. It is generally acknowledged that MAE is more robust, since it puts less weight on outliers. In view of the number of comparisons made in the current research, we settled on MAE as our principal measure of forecasting error, rather than reporting both MAE and RMSE.

In order to obtain error bars for differences between estimated MAE values for different ML estimates, we used the jackknife variance estimator [98]. The jackknife was implemented by obtaining $M - 1$ different MAE values by omitting successively the first,

second, third, ... image in the testing set. It is important to note that entire images were omitted and not single pixels, because pixel errors in the images are highly correlated: a jackknife estimator based on omitting single pixels will greatly underestimate the variance. Since we are interested in relative performance of the ML method compared to a selected baseline, we applied jackknife to the difference between the MAEs for the ML estimate and the baseline. This is another critically important point, because the variance for the MAE for individual ML methods is much larger than the variance of difference between ML and baseline MAEs because the MAEs for ML and baseline are highly correlated. A pseudocode for the procedure is given in Algorithm 1.

Algorithm 1 Computation of MAE for the difference between baseline and ML algorithms.

```

diff_tot = totalMAE(ML_estimate) – totalMAE(baseline_estimate)
var_est = 0
for m in range(M) do
  omit image m from list of M images
  diff = MAE(ML_estimate) – MAE(baseline_estimate) ▷ for the reduced list of images
  var_est = var_est + (diff – diff_tot)2
end for
var_est = (M – 1)/M × var_est
std_est = sqrt(var_est)

```

4.4.3. Baselines and Statistical Estimators

For this study, we employed four different untrained predictors as baselines: (1) previous month (denoted ‘base-11’); (2) same month previous year (denoted ‘base-0’); and (3) average of all training set images for the same month (referred to as ‘seasonal baseline’ or ‘base-Se’); (4) the squared mean square root for training set images of the same month, rounded down to the nearest integer (denoted as ‘base-Se(sqrt)’). When evaluating the effectiveness of different ML algorithms in parameter prediction, we compared these baselines against the trained ML models.

The first three baselines have precedents in the literature. The authors in [66] suggested the use of base-1) as the simplest baseline. Base-0 is suggested by the seasonality of the data. As for base-Se, the authors in [45] implemented the use of the monthly averages as a baseline.

The final baseline is justified by an inferred statistical model of the image pixel distributions, which is motivated as follows. It is clear that the distribution of seasonal climatic parameters for any pixel (i, j, n) must depend on the location (i, j) and the time stamp $t = \text{mod}(n, 12)$. It is also clear that neighboring pixels at the same month index n are correlated. Allowing for these correlations, we posit the simplest possible statistical model for the pixel distributions: namely, that all pixels at month n are statistically independent of all pixels at month n' as long as $n' \neq n$; and further, that the probability distribution for the pixel value (i, j, n) depends only on the values of (i, j, t) .

Given this assumed model for the pixel distributions, we may design an estimator for future pixel values as follows. It is a well-known result in theoretical statistics that the true median of the distribution of a random variable minimizes the expected MAE of a random sample. For a nearly symmetric distribution, the median is approximately equal to the mean. To reduce the influence of high outliers and make the distribution more nearly symmetric, we first take the square root of the data before taking the mean: the result will approximate the median of the square-rooted data, which is equal to the square root of the median of the original data. Consequently, the median may be estimated as the square of the mean of the square-rooted data, which is rounded down to reduce the bias produced by high outliers.

5. Results

In this section, we first present performance results for the different predictors, including baselines and ML methods with and without deseasonalization. Then we give error bars on the relative performance of predictors compared to the base-Se(sqrt) baseline prediction. Finally, we describe the effect of data shuffling on predictor accuracy estimates.

In the following discussion, the data is presented graphically for brevity. Data in tabular format is available at <https://github.com/EslamHussein55/Climatic-parameters> (accessed on 16 April 2021).

5.1. Performance Comparisons for Different Baselines, Feature Sets, and Preprocessing Methods

Figure 6 gives residual plots and R^2 values for the three baselines base-11, base-0, and base-Se(sqrt) for the five climatic parameters (base-Se is not shown, but strongly resembles base-Se(sqrt)). As seen in the figure, Base-Se(sqrt) gives the most accurate estimations across all parameters (predictions lie closer to the 45° line), as well as giving larger R^2 values. Indeed, the R^2 performance for base-Se(sqrt) is almost perfect, with all values over 0.96.

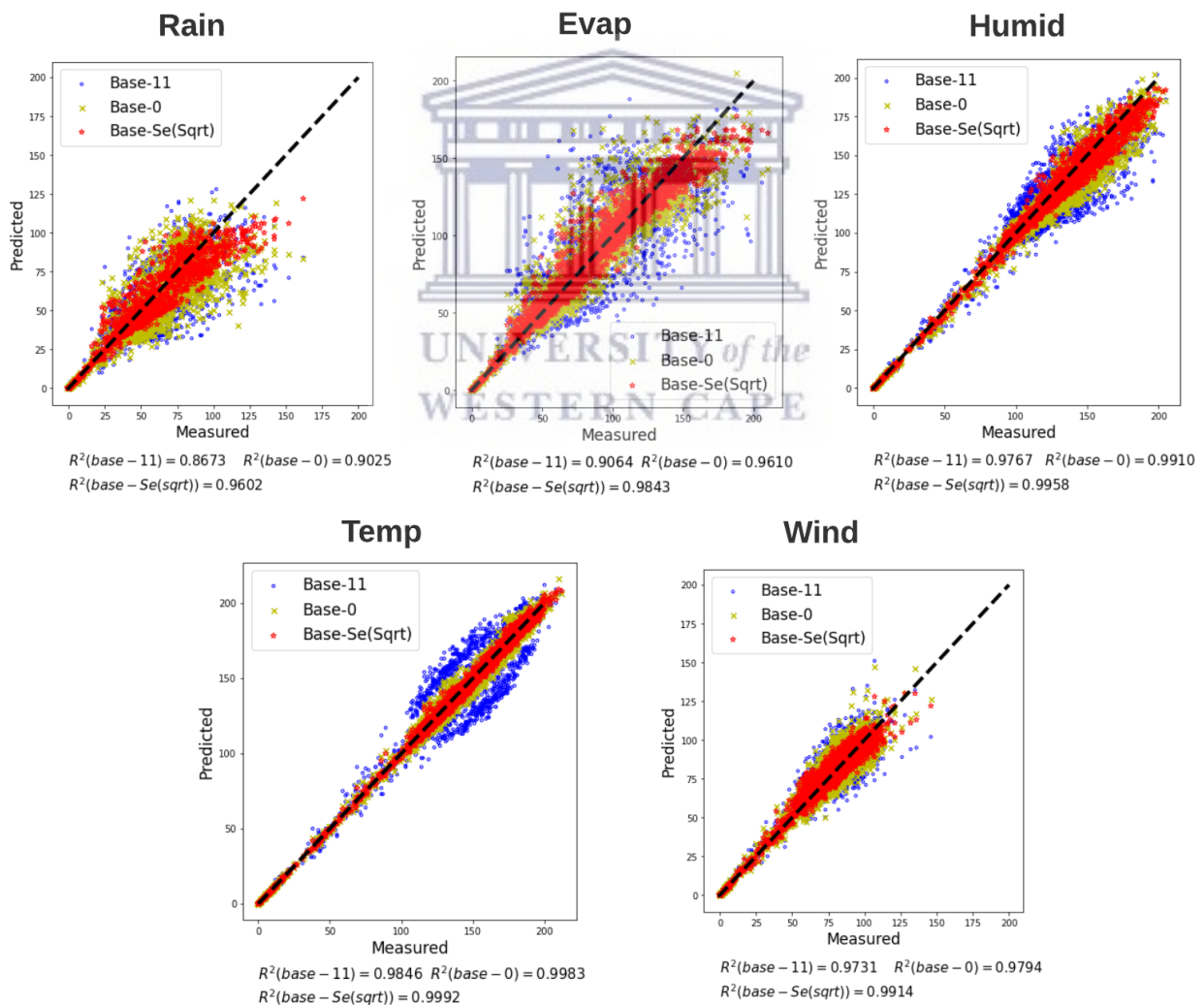


Figure 6. Residual plots and R^2 values for three proposed baselines on five different parameters. The scatter plots show 5000 randomly-selected point for each baseline, for each parameter.

Figures 7–11 summarize the MAE results for models trained using different feature sets for each of the climatic parameters. The corresponding RMSE values were also generated, but since they closely resemble the MAE results, they are omitted here. Each figure contains

two line graphs for raw and de-seasonalized data sets, respectively. For the raw data, the [i,j] feature set performed very badly, so we omitted these results from the figures to avoid stretching the vertical scale. In addition, the base-11 baseline was above the vertical scale for all parameters except rainfall, and is not shown.

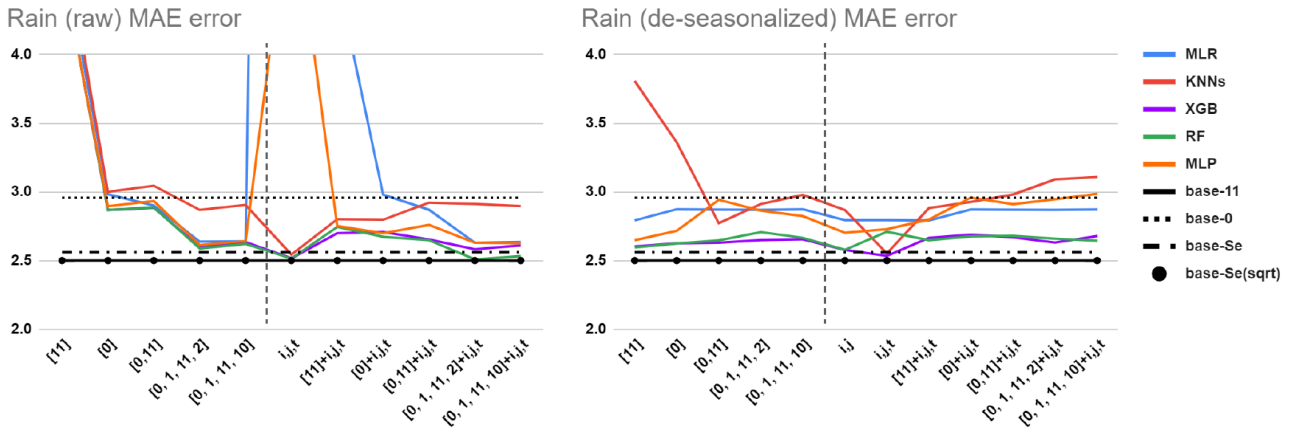


Figure 7. MAE for rainfall predictions with different feature sets, for raw and de-seasonalized data sets.

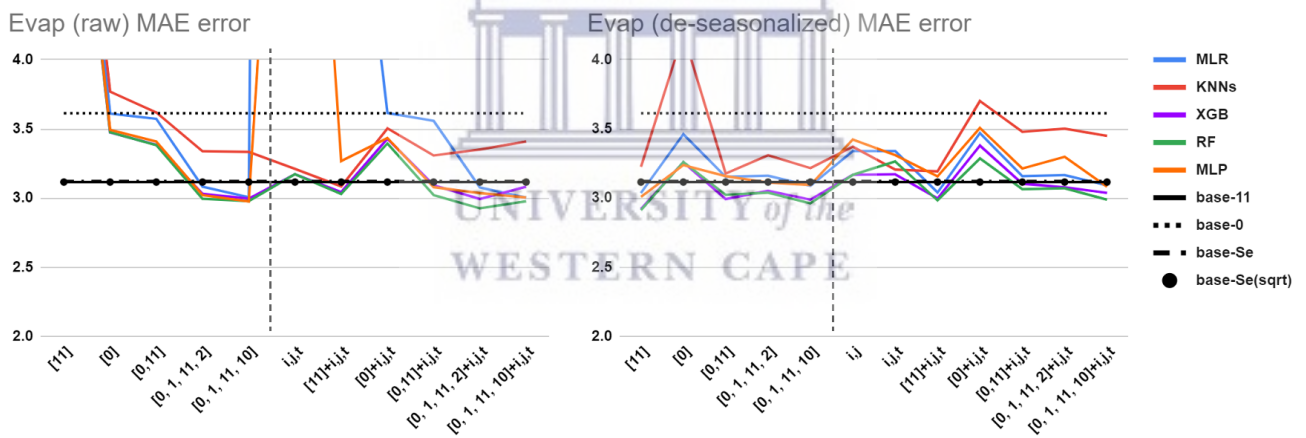


Figure 8. MAE for evaporation predictions with different feature sets, for raw and de-seasonalized data sets.

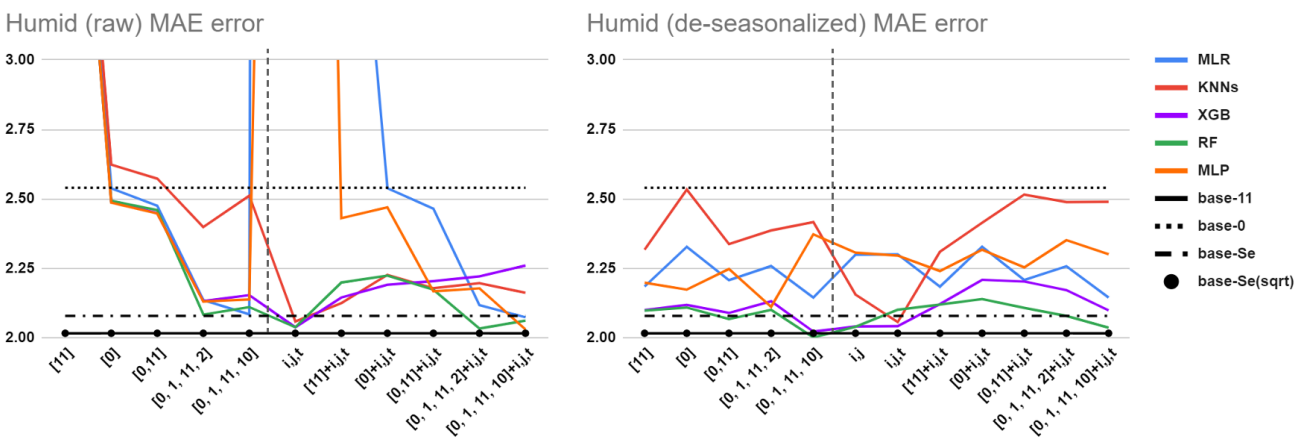


Figure 9. MAE for humidity predictions with different feature sets, for raw and de-seasonalized data sets.

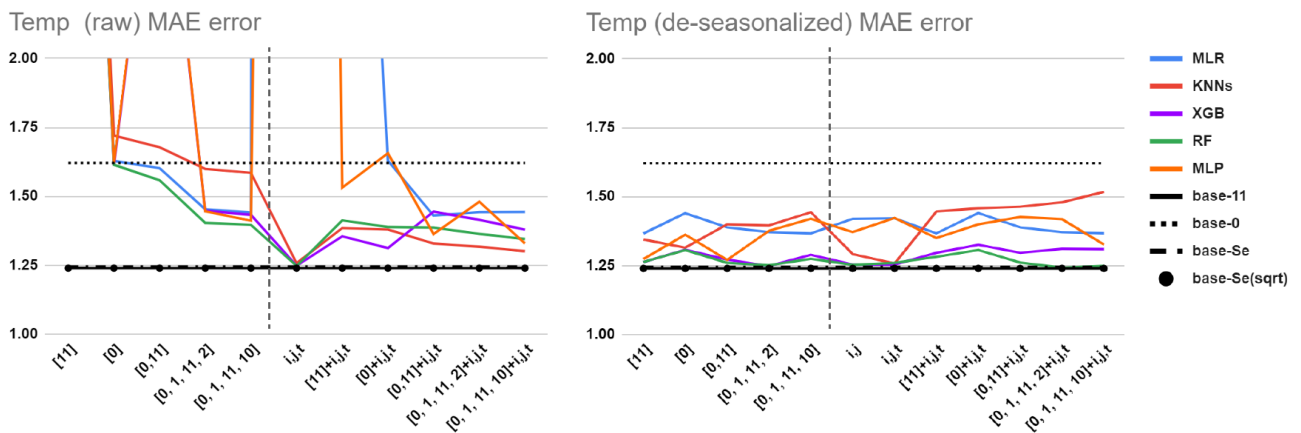


Figure 10. MAE for temperature predictions with different feature sets, for raw and de-seasonalized data sets.

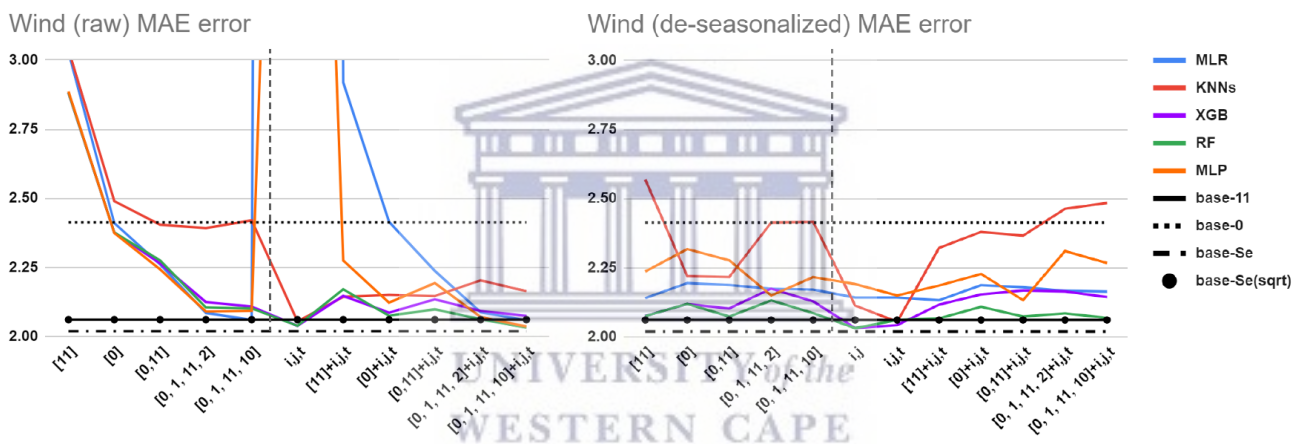


Figure 11. MAE for wind predictions with different feature sets, for raw and de-seasonalized data sets.

Of the four baselines, base-Se(sqrt) is always the best, followed by base-Se, base-0, and base-11, in that order. In fact, Base-Se(sqrt) is also better than all ML tools for all parameters and feature sets, except evaporation for a few feature sets.

Next, comparison between raw-based and de-seasonalized-based predictions shows that de-seasonalizing tends to stabilize the performance, so that it is less dependent on the feature set used. If the feature set contains [0], then de-seasonalizing makes little difference. De-seasonalizing does not always improve the feature sets' performances, as will be discussed in more detail below.

A comparison of feature sets shows that the feature sets [i,j,t],[11,i,j,t], and [0,11,i,j,t] are consistently the best performers, both for raw-based and de-seasonalized-based predictions. In our detailed performance analysis below, we focus on these three feature sets.

It is significant that the above observations apply consistently to all five parameters, which suggests that the same observations can generalize to other climatic parameters.

5.2. Detailed Comparison of ML Tools and Feature Sets

Figures 12 and 13 show the percentage error reductions for different ML algorithms for the 5 climatic parameters, using raw and de-seasonalized data, respectively. Only the three best feature sets are represented, namely i,j,t, [11]+i,j,t, and [0,11]+i,j,t. In the figures, the 100% level corresponds to the Base-Se(sqrt) MAE error: so, for example, the MLR value of 120% for rain (raw) with feature set [0,11] + i,j,t indicates that the MAE error for MLR is 1.2 times the corresponding Base-Se(sqrt) error. Error bars in the figures correspond to

± 2 standard deviations, and were computed using the jackknife procedure described in Section 4.4.2, using the different ML methods and the Base-Se(sqrt) baseline.

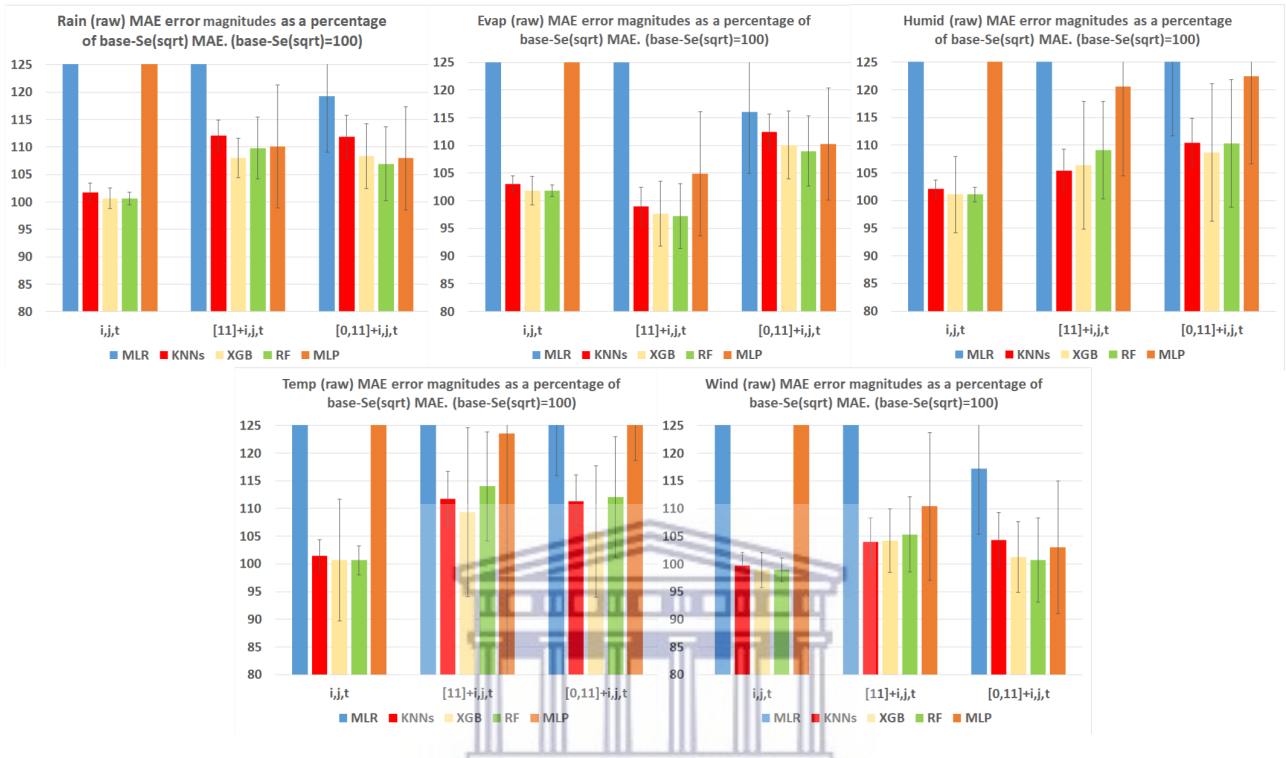


Figure 12. MAE of all trained models with features $[i,j,t, i,j,t+[11], i,j,t+[0,11]]$, compared to base-Se(sqrt) on the raw climate datasets. On the vertical scales, 100 corresponds to the MAE for the base-Se(sqrt) estimator.

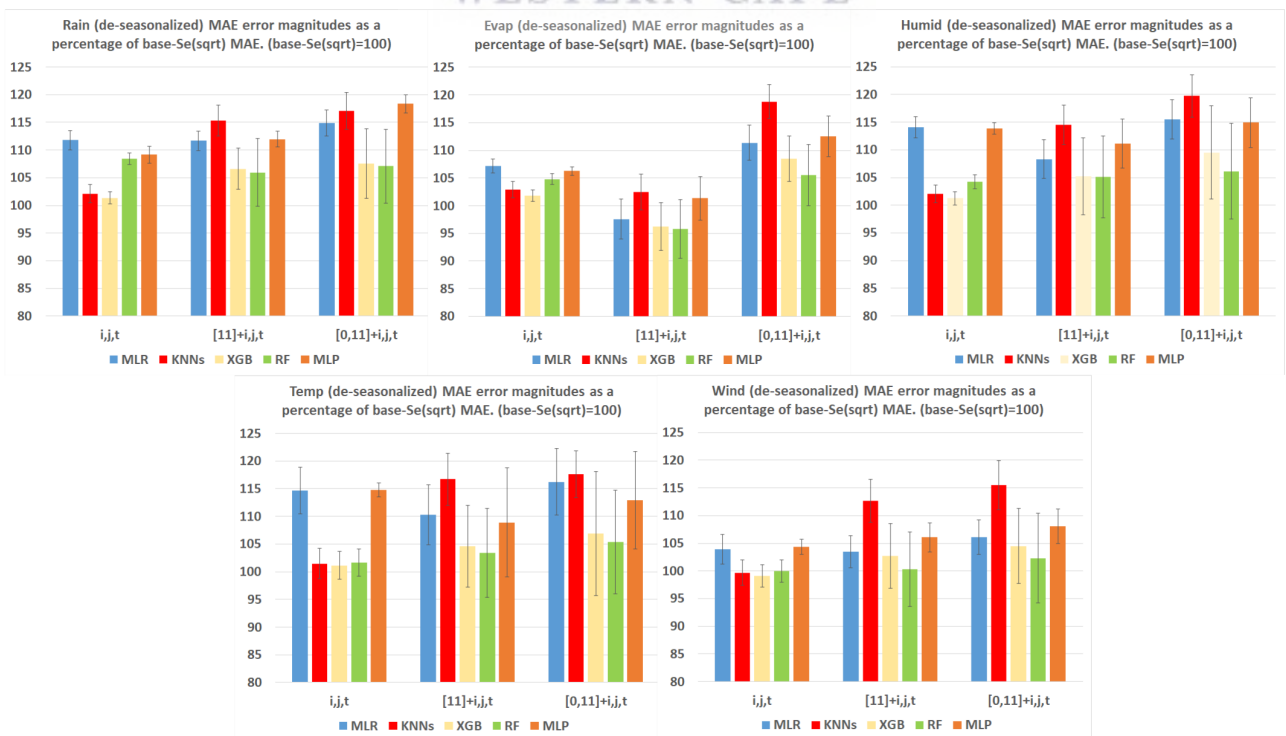


Figure 13. MAE of all trained models with features $[i,j,t, i,j,t+[11], i,j,t+[0,11]]$, compared to base-Se(sqrt) on the de-seasonalized climate datasets. On the vertical scales, 100 corresponds to the MAE for the base-Se(sqrt) estimator.

For raw-based predictions, Figure 12 shows that the KNN, XGB, and RF algorithms typically attain between 100 and 110% of base-Se(sqrt) across all parameters, while MLR and MLP exceed 125% in several cases. For evaporation with the [11]+i,j,t feature set and for wind with the i,j,t feature set, the KNN XGB and RF algorithms are slightly better than base-Se(sqrt), but the error bars show that this relative improvement is not statistically significant.

For deseasonalized-based predictions, the accuracy of XGB and RF is nearly the same as for raw-based, but KNN performance is degraded by up to 10%. The errors for MLR and MLP are reduced to below 115%, but still tend to be 5–10% higher than errors for XGB and RF.

For all parameters except evaporation, the ML methods of KNN, XGB, and RF applied to the feature set i,j,t give the best performance on both raw and de-seasonalized data. This implies that (surprisingly) including lag-based features actually worsens prediction accuracy for these parameters. It is also surprising that the most and least sophisticated methods (MLR and MLP) have similar (and sub-optimal) performance in most cases.

5.3. Data Shuffling

In Section 1, we mentioned that several references shuffle the image sequences. In order to gauge the effects of this shuffling, we used RF with features set [11]+i,j,t to predict all climatic parameters with both shuffled and unshuffled data. For both shuffled and unshuffled data, 156 of the 216 total 12-month sequences were used for training and the rest for testing. The unshuffled data used the first 156 sequences for training and rest for testing, as described in Section 4.2, while the shuffled data took 156 sequences randomly from the entire dataset, thus producing overlap between training and testing sequences. Results showed that MAE obtained from shuffled data was 2–10% lower than from unshuffled data, due to data leakage.

6. Discussion

The results demonstrate that when doing seasonal parameter prediction on monthly time scales, it is important to use a well-motivated simple baseline, e.g., a statistical estimator computed from the source data. This finding is consistent with the points made in [9]. Baselines that depend on lags do not perform as well. Furthermore, a simple same-month average baseline which does not take into account the statistical properties of MAE cannot match the performance of baseline that is designed to estimate median values, which in theory will minimize MAE. For the seasonal parameters we tested, a carefully designed statistical estimator outperforms even highly sophisticated ML models. This finding raises concerns about positive results reported in previous papers that fail to supply statistical baselines.

The results also show that care must be taken in selecting seasonal features as inputs. In the literature, same month previous year (corresponding to our feature [0]) is commonly used [13,17,19,20,24,25,27,28,44,49,51,58,60,62,67]. However, we found that using [0] scarcely outperforms base-0, and is much worse than base-Se(sqrt). Indeed, we found that time stamp t (where t runs from 0 to 11) gave much better results, although it is rarely used in the literature. In addition, using both features typically gave worse performance than using t only.

Aside from using seasonal features, another way to account for seasonality is to de-seasonalize the input data by subtracting monthly averages. The results show that de-seasonalization tends to reduce model complexity: for example, when data is de-seasonalized, then feature [0] becomes unnecessary. However, whether or not de-seasonalization lowers the error depends which algorithm and which features set is used. For example, the best-performing feature-algorithm combinations in our study used i,j,t with RF, or XGB, and for these combinations de-seasonalization of inputs made no difference. We conclude that appropriate feature and algorithm selection has more of an effect on performance than de-seasonalization.

A study similar to ours may be found in [45], in which base-Se is used to standardize the performance of different ML models in predicting 1–6 months ahead rainfall using past rainfall, temperature, and climate index. Compared to base-Se, the following ML algorithms had worse performance: MLR, RF, support vector machine (SVM), artificial neural network (ANN), long short term memory neural networks (LSTM), and convolutional LSTM (ConvLSTM). It follows that including additional climatic parameters as features and doing joint prediction may yield no benefits. Only when the authors used wavelets during pre-processing did their accuracy improve. Even with wavelets, the basic MLR model gave results that nearly matched a sophisticated LSTM model (no error bars for the difference are given, so it is impossible to tell whether there is a significant difference).

For the climatic parameters that were examined in this paper, using previous month (denoted as feature [11]) was not effective, and could even degrade predictive performance when added. However, this conclusion is not applicable to other parameters such as groundwater [7,57], which involves conditions that last over multiple months. The slight improvements seen when adding [11] to evaporation may be due to this effect.

Unlike most prior research in this area, we established the significance of differences in predictive performance between ML methods using error bars that were calculated using statistically rigorous jackknife estimates. The error bars for differences between MAE values for different estimation methods were much smaller than error bars on the MAE values themselves (such as those calculated in [45]). The jackknife methods employed are quite general, and can be used for other ML applications.

Finally, we established that images used for training and testing must be strictly separated and timed. Shuffling of image sequences, which has been employed in some prior research, leads to data leakage, which produces artificial reductions of prediction errors.

7. Conclusions

In this paper, we studied the application of machine learning to the prediction of seasonal climatic parameters on a monthly basis. Our conclusions may be briefly summarized as follows. First, a well-thought out baseline based on a simple statistical estimator will often outperform all ML models. Hence, studies of ML prediction algorithms that do not provide a baseline comparison are not sufficiently demonstrating the effectiveness of the algorithms. Second, the use of time stamp (i.e., month index) as a feature can replace de-seasonalization, and often yields better results than lags (i.e., previous month, or same month previous year). Third, we have demonstrated that jackknife estimation can be used to calculate error bars on algorithms' relative performance, which until now have not been generally reported in the literature. Fourth, we have shown that the practice of data shuffling produces error estimates that are artificially lowered. The methods we have used are quite general, and can be readily applied to other situations. The fact that our results are consistent over five widely different climatic parameters suggests that similar results may be expected for other climatic parameters measured on other regions. This conclusion is reinforced by the fact that similar results have been observed in another study of rainfall conducted in China [45].

In the current research, we have considered only single parameter prediction, using local spatio-temporal based features. For future work, we may apply similar methods to predictions based on other features. Reference [45] for instance shows that using wavelets can lead to better predictions—the question remains whether ML applied to these features can bring significant improvements, or whether simple statistics are sufficient.

Another possibility for future research is the application of deep learning. However, since most monthly datasets available are not large, deep learning may be of limited applicability for monthly prediction. Furthermore, the authors of [45] found that deep learning did not significantly improve on multi-linear regression for monthly rainfall prediction. Nonetheless, since the field of deep learning is developing rapidly, future techniques may produce algorithms that perform well even on datasets of limited size.

Author Contributions: Conceptualization, M.G., A.B. and E.A.H.; methodology, M.G. and E.A.H.; software, E.A.H.; validation, M.G. and C.T.; formal analysis, C.T. and E.A.H.; computing resources, M.G., M.V., and A.B.; writing—original draft preparation, E.A.H.; writing—review and editing, M.G., M.V., C.T. and A.B.; visualization, M.G., C.T. and E.A.H.; supervision, M.G., C.T. and A.B.; funding acquisition, M.G. and M.V. All authors have read and agreed to the published version of the manuscript.

Funding: E.A.H. acknowledges financial support from the South African National Research Foundation (NRF CSUR Grant Number 121291 for the HIPPO project) and from the Telkom-Openserve-Aria Technologies Center of Excellence at the Department of Computer Science of the University of the Western Cape.

Data Availability Statement: Data available in a publicly accessible repository The data presented in this study are openly available in the GES DISC repository at doi:10.5067/5NHC22T9375G.

Acknowledgments: This work made use of the Meerkat Cluster (<http://docs.meerkat.uwc.ac.za> (accessed on 16 April 2021)) provided by the University of the Western Cape’s eResearch Office (<https://eresearch.uwc.ac.za> (accessed on 16 April 2021)).

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

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine learning
ANNs	Artificial neural networks
Base-0, . . . ,base-11	Baseline estimators based on previous lags (see Section 4.4.3)
Base-Se	Seasonal baseline computed from same-month averages (see Section 4.4.3)
Base-Se(sqrt)	Seasonal baseline computed from regularized same-month averages (see Section 4.4.3)
CNNs	Convolution neural networks
LSTMs	Long short term memory
ConvLSTMs	Convolutions layers with Long short term memory
MLP	Multilayer perceptron
RF	Random forest
SVMs	Support vector machines
XGB	Extreme gradient boosting
MLR	Multi linear regression
KNN	K-nearest neighbour
RMSE	Root mean square error
MAE	Mean absolute error
Rain	Rainfall
Temp	Temperature
Evap	Evaporation
Humid	Humidity
FEWS NET	Famine Early Warning Systems Network
FLDAS	FEWS NET Land Data Assimilation System

References

1. Miller, H.J.; Goodchild, M.F. Data-driven geography. *GeoJournal* **2015**, *80*, 449–461. [[CrossRef](#)]
2. Hey, T.; Tansley, S.; Tolle, K.; *The Fourth Paradigm: Data-Intensive Scientific Discovery*; Microsoft research Redmond: Redmond, WA, USA, 2009; Volume 1.
3. Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C.; Hung Byers, A. *Big Data: The Next Frontier for Innovation, Competition, and Productivity*; McKinsey Global Institute: New York, NY, USA, 2011.
4. Kitchin, R. Big Data, new epistemologies and paradigm shifts. *Big Data Soc.* **2014**, *1*, 2053951714528481. [[CrossRef](#)]
5. Ardabili, S.; Mosavi, A.; Dehghani, M.; Várkonyi-Kóczy, A.R. Deep learning and machine learning in hydrological processes climate change and earth systems a systematic review. In *International Conference on Global Research and Education*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 52–62.
6. Monteleoni, C.; Schmidt, G.A.; McQuade, S. Climate informatics: Accelerating discovering in climate science with machine learning. *Comput. Sci. Eng.* **2013**, *15*, 32–40. [[CrossRef](#)]

7. Buontempo, C.; Hewitt, C.D.; Doblas-Reyes, F.J.; Dessai, S. Climate service development, delivery and use in Europe at monthly to inter-annual timescales. *Clim. Risk Manag.* **2014**, *6*, 1–5. [[CrossRef](#)]
8. Steinert, M.; Leifer, L. Scrutinizing Gartner's hype cycle approach. In Proceedings of the Picmet 2010 Technology Management for Global Economic Growth, Phuket, Thailand, 18–22 July 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 1–13.
9. Dacrema, M.F.; Cremonesi, P.; Jannach, D. Are we really making much progress? A worrying analysis of recent neural recommendation approaches. In Proceedings of the 13th ACM Conference on Recommender Systems, Copenhagen, Denmark, 16–20 September 2019; pp. 101–109.
10. Lin, J. The neural hype and comparisons against weak baselines. In *ACM SIGIR Forum*; ACM: New York, NY, USA, 2019; Volume 52, pp. 40–51.
11. Hussein, E.A.; Ghaziasgar, M.; Thron, C. Regional Rainfall Prediction Using Support Vector Machine Classification of Large-Scale Precipitation Maps. *arXiv* **2020**, arXiv:2007.15404.
12. Ludewig, M.; Jannach, D. Evaluation of session-based recommendation algorithms. *User Model. User-Adapt. Interact.* **2018**, *28*, 331–390. [[CrossRef](#)]
13. Cristian, M. Average monthly rainfall forecast in Romania by using K-nearest neighbors regression. *Analele Univ. Constantin Brâncuși Din Târgu Jiu Ser. Econ.* **2018**, *1*, 5–12.
14. Karimi, H.A. *Big Data: Techniques and Technologies in Geoinformatics*; CRC Press: Boca Raton, FL, USA, 2014.
15. Armstrong, T.G.; Moffat, A.; Webber, W.; Zobel, J. Improvements that do not add up: Ad-hoc retrieval results since 1998. In Proceedings of the 18th ACM conference on Information and knowledge management, Hong Kong, China, 2–6 November 2009; pp. 601–610.
16. Du, Y.; Berndtsson, R.; An, D.; Zhang, L.; Yuan, F.; Uvo, C.B.; Hao, Z. Multi-Space Seasonal Precipitation Prediction Model Applied to the Source Region of the Yangtze River, China. *Water* **2019**, *11*, 2440. [[CrossRef](#)]
17. Lakshmaiah, K.; Krishna, S.M.; Reddy, B.E. Application of referential ensemble learning techniques to predict the density of rainfall. In Proceedings of the 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), Mysuru, India, 9–10 December 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 233–237.
18. Lee, J.; Kim, C.G.; Lee, J.E.; Kim, N.W.; Kim, H. Application of artificial neural networks to rainfall forecasting in the Geum River basin, Korea. *Water* **2018**, *10*, 1448. [[CrossRef](#)]
19. Beheshti, Z.; Firouzi, M.; Shamsuddin, S.M.; Zibarzani, M.; Yusop, Z. A new rainfall forecasting model using the CAPSO algorithm and an artificial neural network. *Neural Comput. Appl.* **2016**, *27*, 2551–2565. [[CrossRef](#)]
20. Duong, T.A.; Bui, M.D.; Rutschmann, P. A comparative study of three different models to predict monthly rainfall in Ca Mau, Vietnam. In *Wasserbau-Symposium Graz 2018. Wasserwirtschaft–Innovation aus Tradition. Tagungsband. Beiträge Zum 19; Gemeinschafts-Symposium der Wasserbau-Institute TU München, TU Graz und ETH Zürich: Graz, Austria, 2018; p. Paper-G5.*
21. Gao, L.; Wei, F.; Yan, Z.; Ma, J.; Xia, J. A Study of Objective Prediction for Summer Precipitation Patterns Over Eastern China Based on a Multinomial Logistic Regression Model. *Atmosphere* **2019**, *10*, 213. [[CrossRef](#)]
22. Mishra, N.; Kushwaha, A. Rainfall Prediction using Gaussian Process Regression Classifier. *Int. J. Adv. Res. Comput. Eng. Technol. (IJARCT)* **2019**, *8*.
23. Aguasca-Colomo, R.; Castellanos-Nieves, D.; Méndez, M. Comparative analysis of rainfall prediction models using machine learning in islands with complex orography: Tenerife Island. *Appl. Sci.* **2019**, *9*, 4931. [[CrossRef](#)]
24. Sulaiman, J.; Wahab, S.H. Heavy rainfall forecasting model using artificial neural network for flood prone area. In *IT Convergence and Security 2017*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 68–76.
25. Chhetri, M.; Kumar, S.; Pratim Roy, P.; Kim, B.G. Deep BLSTM-GRU Model for Monthly Rainfall Prediction: A Case Study of Simtokha, Bhutan. *Remote Sens.* **2020**, *12*, 3174. [[CrossRef](#)]
26. Bojang, P.O.; Yang, T.C.; Pham, Q.B.; Yu, P.S. Linking Singular Spectrum Analysis and Machine Learning for Monthly Rainfall Forecasting. *Appl. Sci.* **2020**, *10*, 3224. [[CrossRef](#)]
27. Canchala, T.; Alfonso-Morales, W.; Carvajal-Escobar, Y.; Cerón, W.L.; Caicedo-Bravo, E. Monthly Rainfall Anomalies Forecasting for Southwestern Colombia Using Artificial Neural Networks Approaches. *Water* **2020**, *12*, 2628. [[CrossRef](#)]
28. Mehr, A.D.; Nourani, V.; Khosrowshahi, V.K.; Ghorbani, M.A. A hybrid support vector regression–firefly model for monthly rainfall forecasting. *Int. J. Environ. Sci. Technol.* **2019**, *16*, 335–346. [[CrossRef](#)]
29. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.; Wong, W.; Woo, W.C. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. *arXiv* **2015**, arXiv:1506.04214.
30. Manandhar, S.; Dev, S.; Lee, Y.H.; Meng, Y.S.; Winkler, S. A data-driven approach for accurate rainfall prediction. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 9323–9331. [[CrossRef](#)]
31. Jing, J.; Li, Q.; Peng, X. MLC-LSTM: Exploiting the Spatiotemporal Correlation between Multi-Level Weather Radar Echoes for Echo Sequence Extrapolation. *Sensors* **2019**, *19*, 3988. [[CrossRef](#)]
32. Sato, R.; Kashima, H.; Yamamoto, T. Short-term precipitation prediction with skip-connected prednet. In *International Conference on Artificial Neural Networks*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 373–382.
33. Ayzel, G.; Heistermann, M.; Sorokin, A.; Nikitin, O.; Lukyanova, O. All convolutional neural networks for radar-based precipitation nowcasting. *Procedia Comput. Sci.* **2019**, *150*, 186–192. [[CrossRef](#)]

34. Singh, S.; Sarkar, S.; Mitra, P. A deep learning based approach with adversarial regularization for Doppler weather radar ECHO prediction. In Proceedings of the 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Fort Worth, TX, USA, 23–28 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 5205–5208.
35. Chen, L.; Cao, Y.; Ma, L.; Zhang, J. A Deep Learning Based Methodology for Precipitation Nowcasting with Radar. *Earth Space Sci.* **2020**, *7*, e2019EA000812. [[CrossRef](#)]
36. Tran, Q.K.; Song, S.K. Computer vision in precipitation nowcasting: Applying image quality assessment metrics for training deep neural networks. *Atmosphere* **2019**, *10*, 244. [[CrossRef](#)]
37. Shi, E.; Li, Q.; Gu, D.; Zhao, Z. Convolutional Neural Networks Applied on Weather Radar Echo Extrapolation. *DEStech Trans. Comput. Sci. Eng.* **2017**. [[CrossRef](#)]
38. Castro, R.; Souto, Y.M.; Ogasawara, E.; Porto, F.; Bezerra, E. STConvS2S: Spatiotemporal Convolutional Sequence to Sequence Network for weather forecasting. *Neurocomputing* **2020**, *426*, 285–298. [[CrossRef](#)]
39. Wang, Y.; Long, M.; Wang, J.; Gao, Z.; Philip, S.Y. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. *Adv. Neural Inf. Process. Syst.* **2017**, 879–888.
40. Tran, Q.K.; Song, S.K. Multi-Channel Weather Radar Echo Extrapolation with Convolutional Recurrent Neural Networks. *Remote Sens.* **2019**, *11*, 2303. [[CrossRef](#)]
41. Zhang, P.; Jia, Y.; Gao, J.; Song, W.; Leung, H.K. Short-term rainfall forecasting using multi-layer perceptron. *IEEE Trans. Big Data* **2018**, *6*, 93–106. [[CrossRef](#)]
42. Oswal, N. Predicting rainfall using machine learning techniques. *arXiv* **2019**, arXiv:1910.13827.
43. Balamurugan, M.; Manojkumar, R. Study of short term rain forecasting using machine learning based approach. *Wirel. Netw.* **2019**, 1–6. [[CrossRef](#)]
44. Nourani, V.; Uzelaltinbulat, S.; Sadikoglu, F.; Behfar, N. Artificial intelligence based ensemble modeling for multi-station prediction of precipitation. *Atmosphere* **2019**, *10*, 80. [[CrossRef](#)]
45. Xu, L.; Chen, N.; Zhang, X.; Chen, Z. A data-driven multi-model ensemble for deterministic and probabilistic precipitation forecasting at seasonal scale. *Clim. Dyn.* **2020**, *54*, 1–20. [[CrossRef](#)]
46. Mehdizadeh, S.; Behmanesh, J.; Khalili, K. New approaches for estimation of monthly rainfall based on GEP-ARCH and ANN-ARCH hybrid models. *Water Resour. Manag.* **2018**, *32*, 527–545. [[CrossRef](#)]
47. Shenify, M.; Danesh, A.S.; Gocić, M.; Taher, R.S.; Wahab, A.W.A.; Gani, A.; Shamsirband, S.; Petković, D. Precipitation estimation using support vector machine with discrete wavelet transform. *Water Resour. Manag.* **2016**, *30*, 641–652. [[CrossRef](#)]
48. Banadkooki, F.B.; Ehteram, M.; Ahmed, A.N.; Fai, C.M.; Afan, H.A.; Ridwam, W.M.; Sefelnasr, A.; El-Shafie, A. Precipitation forecasting using multilayer neural network and support vector machine optimization based on flow regime algorithm taking into account uncertainties of soft computing models. *Sustainability* **2019**, *11*, 6681. [[CrossRef](#)]
49. Haidar, A.; Verma, B. Monthly rainfall forecasting using one-dimensional deep convolutional neural network. *IEEE Access* **2018**, *6*, 69053–69063. [[CrossRef](#)]
50. Zhan, C.; Wu, F.; Wu, Z.; Chi, K.T. Daily Rainfall Data Construction and Application to Weather Prediction. In Proceedings of the 2019 IEEE International Symposium on Circuits and Systems (ISCAS), Hokkaido, Japan, 26–29 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
51. Weesakul, U.; Kaewprapha, P.; Boonyuen, K.; Mark, O. Deep learning neural network: A machine learning approach for monthly rainfall forecast, case study in eastern region of Thailand. *Eng. Appl. Sci. Res.* **2018**, *45*, 203–211.
52. Chattopadhyay, A.; Hassanzadeh, P.; Pasha, S. Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data. *Sci. Rep.* **2020**, *10*, 1–13. [[CrossRef](#)]
53. Patel, M.; Patel, A.; Ghosh, D. Precipitation nowcasting: Leveraging bidirectional lstm and 1d cnn. *arXiv* **2018**, arXiv:1810.10485.
54. Zhuang, W.; Ding, W. Long-lead prediction of extreme precipitation cluster via a spatiotemporal convolutional neural network. In Proceedings of the 6th International Workshop on Climate Informatics: CI, Boulder, CO, USA, 22–23 September 2016.
55. Boonyuen, K.; Kaewprapha, P.; Srivihok, P. Daily rainfall forecast model from satellite image using Convolution neural network. In Proceedings of the 2018 IEEE International Conference on Information Technology, Bhubaneswar, India, 19–21 December 2018; pp. 1–7.
56. Boonyuen, K.; Kaewprapha, P.; Weesakul, U.; Srivihok, P. Convolutional Neural Network Inception-v3: A Machine Learning Approach for Leveling Short-Range Rainfall Forecast Model from Satellite Image. In *International Conference on Swarm Intelligence*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 105–115.
57. Hussein, E.A.; Thron, C.; Ghaziasgar, M.; Bagula, A.; Vaccari, M. Groundwater Prediction Using Machine-Learning Tools. *Algorithms* **2020**, *13*, 300. [[CrossRef](#)]
58. Aswin, S.; Geetha, P.; Vinayakumar, R. Deep learning models for the prediction of rainfall. In Proceedings of the 2018 International Conference on Communication and Signal Processing (ICCSP), Tamilnadu, India, 3–5 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 0657–0661.
59. Amiri, M.A.; Amerian, Y.; Mesgari, M.S. Spatial and temporal monthly precipitation forecasting using wavelet transform and neural networks, Qara-Qum catchment, Iran. *Arab. J. Geosci.* **2016**, *9*, 421. [[CrossRef](#)]
60. Abbot, J.; Marohasy, J. Forecasting Monthly Rainfall in the Western Australian Wheat-Belt up to 18-Months in Advance Using Artificial Neural Networks. In *Australasian Joint Conference on Artificial Intelligence*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 71–87.

61. Damavandi, H.G.; Shah, R. A Learning Framework for An Accurate Prediction of Rainfall Rates. *arXiv* **2019**, arXiv:1901.05885.
62. Abbot, J.; Marohasy, J. Application of artificial neural networks to forecasting monthly rainfall one year in advance for locations within the Murray Darling basin, Australia. *Int. J. Sustain. Dev. Plan.* **2017**, *12*, 1282–1298. [[CrossRef](#)]
63. Mohamadi, S.; Ehteram, M.; El-Shafie, A. Accuracy enhancement for monthly evaporation predicting model utilizing evolutionary machine learning methods. *Int. J. Environ. Sci. Technol.* **2020**, *17*, 3373–3396. [[CrossRef](#)]
64. Delleur, J.W.; Kavvas, M.L. Stochastic models for monthly rainfall forecasting and synthetic generation. *J. Appl. Meteorol.* **1978**, *17*, 1528–1536. [[CrossRef](#)]
65. Barnett, A.G.; Baker, P.; Dobson, A. Analysing seasonal data. *R J.* **2012**, *4*, 5–10. [[CrossRef](#)]
66. Nielsen, A. *Practical Time Series Analysis: Prediction with Statistics and Machine Learning*; O'Reilly: Newton, MA, USA, 2020.
67. Kumar, D.; Singh, A.; Samui, P.; Jha, R.K. Forecasting monthly precipitation using sequential modelling. *Hydrol. Sci. J.* **2019**, *64*, 690–700. [[CrossRef](#)]
68. Ramsundram, N.; Sathya, S.; Karthikeyan, S. Comparison of decision tree based rainfall prediction model with data driven model considering climatic variables. *Irrig. Drain. Syst. Eng.* **2016**. [[CrossRef](#)]
69. Sardeshpande, K.D.; Thool, V.R. Rainfall Prediction: A Comparative Study of Neural Network Architectures. In *Emerging Technologies in Data Mining and Information Security*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 19–28.
70. Shi, X.; Gao, Z.; Lausen, L.; Wang, H.; Yeung, D.Y.; Wong, W.k.; Woo, W.c. Deep learning for precipitation nowcasting: A benchmark and a new model. *Adv. Neural Inf. Process. Syst.* **2017**, 5617–5627.
71. McNally, A. FLDAS Noah Land Surface Model L4 Global Monthly 0.1 × 0.1 degree (MERRA-2 and CHIRPS). *Atmos. Compos. Water Energy Cycles Clim. Var.* **2018**. [[CrossRef](#)]
72. Loeser, C.; Rui, H.; Teng, W.L.; Ostrenga, D.M.; Wei, J.C.; McNally, A.L.; Jacob, J.P.; Meyer, D.J. Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (LDAS) and Other Assimilated Hydrological Data at NASA GES DISC. In Proceedings of the 100th American Meteorological Society Annual Meeting, St. Boston, MA, USA, 12–16 January 2020.
73. Nematchoua, M.K. A study on outdoor environment and climate change effects in Madagascar. *J. Build. Sustain.* **2017**, *1*, 12.
74. Tadross, M.; Randriamarolaza, L.; Rabefitia, Z.; Zheng, K. *Climate Change in Madagascar; Recent Past and Future*; World Bank: Washington, DC, USA, 2008; Volume 18.
75. Szabó, A.; Raveloson, A.; Székely, B. Landscape evolution and climate in Madagascar: Lavakization in the light of archive precipitation data. *Cuad. Investig. GeogrÁfica/Geogr. Res. Lett.* **2015**, *41*, 181–204. [[CrossRef](#)]
76. Harvey, C.A.; Rakotobe, Z.L.; Rao, N.S.; Dave, R.; Razafimahatratra, H.; Rabarijohn, R.H.; Rajaofara, H.; MacKinnon, J.L. Extreme vulnerability of smallholder farmers to agricultural risks and climate change in Madagascar. *Philos. Trans. R. Soc. B Biol. Sci.* **2014**, *369*, 20130089. [[CrossRef](#)]
77. Ingram, J.C.; Dawson, T.P. Climate change impacts and vegetation response on the island of Madagascar. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2005**, *363*, 55–59. [[CrossRef](#)]
78. Sanchez-Pi, N.; Marti, L.; Abreu, A.; Bernard, O.; de Vargas, C.; Eveillard, D.; Maass, A.; Marquet, P.A.; Sainte-Marie, J.; Salomon, J.; et al. Artificial Intelligence, Machine Learning and Modeling for Understanding the Oceans and Climate Change. In *NeurIPS 2020 Workshop-Tackling Climate Change with Machine Learning*; 2020. Available online: <https://hal.archives-ouvertes.fr/hal-03138712> (accessed on 19 April 2021) .
79. Stein, A.L. Artificial Intelligence and Climate Change. *Yale J. Reg.* **2020**, *37*, 890.
80. Abudu, S.; Cui, C.; King, J.P.; Moreno, J.; Bawazir, A.S. Modeling of daily pan evaporation using partial least squares regression. *Sci. China Technol. Sci.* **2011**, *54*, 163–174. [[CrossRef](#)]
81. Pinheiro, A.; Vidakovic, B. Estimating the square root of a density via compactly supported wavelets. *Comput. Stat. Data Anal.* **1997**, *25*, 399–415. [[CrossRef](#)]
82. Qiu, M.; Zhao, P.; Zhang, K.; Huang, J.; Shi, X.; Wang, X.; Chu, W. A short-term rainfall prediction model using multi-task convolutional neural networks. In Proceedings of the 2017 IEEE International Conference on Data Mining (ICDM), New Orleans, LA, USA, 18–21 November 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 395–404.
83. Cramer, S.; Kampouridis, M.; Freitas, A.A.; Alexandridis, A.K. An extensive evaluation of seven machine learning methods for rainfall prediction in weather derivatives. *Expert Syst. Appl.* **2017**, *85*, 169–181. [[CrossRef](#)]
84. Cao, Y.; Li, Q.; Shan, H.; Huang, Z.; Chen, L.; Ma, L.; Zhang, J. Precipitation Nowcasting with Star-Bridge Networks. *arXiv* **2019**, arXiv:1907.08069.
85. Klein, B.; Wolf, L.; Afek, Y. A dynamic convolutional layer for short range weather prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4840–4848.
86. Mukhopadhyay, A.; Shukla, B.P.; Mukherjee, D.; Chanda, B. A novel neural network based meteorological image prediction from a given sequence of images. In Proceedings of the 2011 Second International Conference on Emerging Applications of Information Technology, Kolkata, India, 19–20 February 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 202–205.
87. Vandekerckhove, J.; Matzke, D.; Wagenmakers, E.J. Model comparison and the principle of parsimony. In *Oxford Handbook of Computational and Mathematical Psychology*; Oxford University Press: Oxford, UK, 2015; pp. 300–319.
88. Dash, Y.; Mishra, S.K.; Panigrahi, B.K. Rainfall prediction for the Kerala state of India using artificial intelligence approaches. *Comput. Electr. Eng.* **2018**, *70*, 66–73. [[CrossRef](#)]
89. Purushotham, S.; Meng, C.; Che, Z.; Liu, Y. Benchmarking deep learning models on large healthcare datasets. *J. Biomed. Inform.* **2018**, *83*, 112–134. [[CrossRef](#)] [[PubMed](#)]

90. Leming, M.; Górriz, J.M.; Suckling, J. Ensemble deep learning on large, mixed-site fMRI datasets in autism and other tasks. *arXiv* **2020**, arXiv:2002.07874.
91. Chen, X.W.; Lin, X. Big data deep learning: Challenges and perspectives. *IEEE Access* **2014**, *2*, 514–525. [[CrossRef](#)]
92. Zhang, Q.; Yang, L.T.; Chen, Z.; Li, P. A survey on deep learning for big data. *Inf. Fusion* **2018**, *42*, 146–157. [[CrossRef](#)]
93. Buitinck, L.; Louppe, G.; Blondel, M.; Pedregosa, F.; Mueller, A.; Grisel, O.; Niculae, V.; Prettenhofer, P.; Gramfort, A.; Grobler, J.; et al. API Design for machine learning software: Experiences from the scikit-learn project. *arXiv* **2013**, arXiv:1309.0238.
94. Stigler, S.M. Studies in the History of Probability and Statistics. XXXII: Laplace, Fisher, and the discovery of the concept of sufficiency. *Biometrika* **1973**, *60*, 439–445. [[CrossRef](#)]
95. Willmott, C.J.; Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.* **2005**, *30*, 79–82. [[CrossRef](#)]
96. 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](#)]
97. Brassington, G. Mean absolute error and root mean square error: Which is the better metric for assessing model performance? *Egu Gen. Assem. Conf. Abstr.* **2017**, *19*, 3574.
98. Efron, B.; Stein, C. The jackknife estimate of variance. *Ann. Stat.* **1981**, *9*, 586–596. [[CrossRef](#)]



Chapter 6

Conclusion

This study aimed to investigate the question “How effective is ML at predicting geophysical parameter data, given the spatio-temporal aspect of this kind of data?”. The background to this question was an abundance of research studies in the literature in the same context that make use of very complex ML methods without proper or any baselining. This results in reports of high levels of prediction success, with very few exceptions.

The aim of this study was to objectively investigate the task of geophysical parameter forecasting using ML tools to provide insight into the debate that is ongoing [10].

Accordingly, the research investigated the use of six ML tools towards the prediction of several geo-physical parameters based on image data. Specifically, the predictive performance of these algorithms applied to six climatic parameters using various combinations of features was quantified. Furthermore, the predictive success of the resulting trained ML models in each case was objectively compared to that of basic statistical estimators computed directly from the training data.

The main research question was broken down into four research sub-questions.

Research sub-question 1 was “What are the evaluation and pre-processing pitfalls when it comes to spatio-temporal parameter forecasting?”. The key findings obtained towards this question were as follows:

- Many studies lack a baseline predictor to which to compare the ML methods used. This makes it difficult to objectively quantify the performance of the ML methods in real terms.

- Many studies do not provide error bars for prediction errors, so that the significance of differences between prediction methods cannot be determined.
- Some references utilize practices that permit data leakage, leading to overestimates of predictive accuracy.

Accordingly, research sub-question 1 can be answered as follows: The evaluation and pre-processing pitfalls relating to spatio-temporal parameter forecasting include data leakage, lack of baselining and a lack of error bars when reporting errors.

Research sub-question 2 was phrased as “What features/parameters affect the prediction success of the geophysical parameter prediction models?”. Key findings made towards this sub-question include the following:

- A variety of features can be used including global features derived from GMMs, and local features including previous lags, pixel location, timestamp and others. A complete and detailed description of these features were provided in Chapters 3–5.
- Global features constructed from GMMs fitted to the pixel intensity distribution brought further improvements when applied to groundwater. Although GMMs gave some improvement in groundwater prediction, GMMs are not suitable for highly seasonal parameters since GMMs captures the slow changes that occurs on a month-to-month basis. This conclusion was supported by preliminary investigations in which GMMs were used for rainfall and temperature prediction. Based on these preliminary results, GMMs were not investigated further for highly seasonal parameters.
- More complicated feature sets do not necessarily contribute towards better prediction success, and in most cases contribute towards a decline in prediction success.
- For all parameters used in this study except for groundwater, making use of features based on previous lags was not effective and degraded the predictive performance when utilized. This is due to the highly seasonal nature of these parameters i.e. rain, temperature, evaporation, humidity and wind. This finding is consistent with [45], in which the authors showed that using previous lags was not effective for rainfall prediction. Groundwater, on the other hand, changes at a much slower pace than other parameters, and groundwater conditions are generally observed to last for multiple months, making it easier to predict.

Based on the above findings, research sub-question 2 can be answered as follows: while a range of features, both local and global, can be used towards prediction: a) increasing

the feature complexity does not provide better performance and can even contribute towards reducing it; b) local features based on previous lags may only be effective for parameters that are not highly seasonal like groundwater, and are certainly not effective for highly seasonal parameters like rainfall, temperature etc.; and c) global features provided by means of GMMs are promising and could help improve on prediction performance, although this was only investigated for groundwater.

Research sub-question 3 was phrased as “How do various ML techniques compare in terms of geophysical parameter forecasting success?”. Key findings made towards this sub-question include the following:

- On groundwater data, SVMs were shown to outperform all other tools such as XGB, RF, MLR and MLP when using local features. However, XGB provided better prediction success when coupled with global features provided by GMMs.
- On highly seasonal parameters i.e. rainfall, temperature, evaporation, wind and humidity, the techniques compared i.e. MLR, MLP, kNNs, RF, and XGB, showed very similar performance, with a small advantage to RF.
- It was also shown that data leakage, which is regularly to be found in the majority of related studies in this field, can lead to over estimation in model performance, and steps should therefore always be taken to prevent data leakage.

Based on the above findings, research sub-question 3 can be answered as follows: In general, most ML techniques provide comparable prediction performance for most parameters, and it is in fact the choice of features used that has the most prominent effect on prediction performance. While this is the general trend across all parameters (highly-seasonal or not), for groundwater which is not a highly-seasonal parameter, some techniques may outperform others.

Finally, research sub-question 4 was phrased as “How do the various ML techniques compare to appropriate simple baselines?”. Towards answering this question, and unlike most prior research in this area, the significance of differences in predictive performance between the ML methods and the baselines used were clearly established in this research using error bars that were calculated using statistically rigorous jackknife estimates. The error bars for differences between MAE values for different estimation methods were much smaller than error bars on the MAE values themselves (such as those calculated in [45]). The jackknife methods employed in this research are quite general, and can be used for other ML applications. The jackknife estimates helped to objectively quantify performance of the ML methods in real-terms.

Key findings made towards this sub-question include the following:

- For groundwater, it was shown that ML can out-perform a simple baseline given an appropriate feature set.
- For all five other parameters considered, which are highly-seasonal parameters that are commonly used, it was shown definitively that ML never significantly outperforms the statistical baseline. In fact, it was shown that ML methods actually perform worse than the baseline for most feature sets. This finding is consistent with [10] and [45]. This demonstrates the crucial importance of using a simple baseline in such studies, which helps to objectively quantify the success of current prevailing algorithms in climate research and other fields so as to generate an appropriate evaluation of the performance which is neither a hype nor depreciation.

Based on the above, research sub-question 4 can be answered as follows: Simple baselines outperform trained ML methods on all highly-seasonal geophysical parameters, while the baseline did not outperform the ML methods for groundwater, which is not highly seasonal. This points to difficulty of predicting highly-seasonal parameters which are fast-changing. This also demonstrates the importance of using well thought-out baselines to measure the true performance of the ML methods.

It should be noted that the methods used were shown to be readily and generally applicable, and can therefore be readily applied to other parameters and regions. The results were shown to be consistent over five widely different parameters i.e. rain, temperature, evaporation, humidity and wind, over a variety of different regions, and this suggests that similar results may be expected for other climatic parameters measured in other regions. This conclusion is reinforced by the fact that similar results have been observed in another study of rainfall conducted in China [45].

Based on the answers to the research sub-questions above, an answer to the main research question of this thesis can be formulated, which was phrased as “How effective is ML at predicting geophysical parameter data, given the spatio-temporal aspect of this kind of data?”. In response to the main research question and in conclusion to this thesis, it is stated that: “The ML tools considered in this study have limited effectiveness at predicting geophysical parameter data; they may be more effective at predicting slower-changing parameters such as groundwater, but not effective when predicting highly-seasonal parameters such as rainfall, temperature, humidity, evaporation and wind. The features used to this end are very important and can help attain an effective prediction model, where the use of more complex feature sets should not be primarily preferred to simpler ones, and global features using GMMs should be considered as they are

promising. Finally, it is important to avoid various pitfalls such as data leakage, lack of error bars, and the lack of an appropriate statistical baseline in order to obtain accurate indications of the prediction performance of the trained models.”

6.1 Future Work

The following provides several directions for future work.

- Investigating the use of deep learning on limited data, even though the authors in [45] found that deep learning did not significantly improve over MLR for monthly rainfall prediction. The field is continuously evolving, and future techniques may produce algorithms that perform well even on datasets of limited size.
- Conduct joint parameter prediction, and try other features based on wavelets. Reference [45] for instance shows that using wavelets can lead to better predictions. The question remains whether ML applied to these features can bring significant improvements, or whether simple statistics are sufficient.

6.2 Concluding Comments

Ultimately, this research demonstrates the importance of using well-grounded statistical methods and techniques when producing and analyzing the results of ML predictive models.

Bibliography

- [1] M. A. Amiri, Y. Amerian, and M. S. Mesgari, "Spatial and temporal monthly precipitation forecasting using wavelet transform and neural networks, qara-qum catchment, iran," *Arabian Journal of Geosciences*, vol. 9, no. 5, p. 421, 2016.
- [2] S. Ardabili, A. Mosavi, M. Dehghani, and A. R. Várkonyi-Kóczy, "Deep learning and machine learning in hydrological processes climate change and earth systems a systematic review," in *International Conference on Global Research and Education*. Springer, 2019, pp. 52–62.
- [3] G. Ayzel, M. Heistermann, A. Sorokin, O. Nikitin, and O. Lukyanova, "All convolutional neural networks for radar-based precipitation nowcasting," *Procedia Computer Science*, vol. 150, pp. 186–192, 2019.
- [4] Z. Beheshti, M. Firouzi, S. M. Shamsuddin, M. Zibarzani, and Z. Yusop, "A new rainfall forecasting model using the capso algorithm and an artificial neural network," *Neural Computing and Applications*, vol. 27, no. 8, pp. 2551–2565, 2016.
- [5] K. Boonyuen, P. Kaewprapha, and P. Srivihok, "Daily rainfall forecast model from satellite image using convolution neural network," in *2018 IEEE International Conference on Information Technology*, 2018, pp. 1–7.
- [6] K. Boonyuen, P. Kaewprapha, U. Weesakul, and P. Srivihok, "Convolutional neural network inception-v3: A machine learning approach for leveling short-range rainfall forecast model from satellite image," in *International Conference on Swarm Intelligence*. Springer, 2019, pp. 105–115.
- [7] L. Chen, Y. Cao, L. Ma, and J. Zhang, "A deep learning based methodology for precipitation nowcasting with radar," *Earth and Space Science*, p. e2019EA000812, 2020.
- [8] M. Chhetri, S. Kumar, P. Pratim Roy, and B.-G. Kim, "Deep blstm-gru model for monthly rainfall prediction: A case study of simtokha, bhutan," *Remote Sensing*, vol. 12, no. 19, p. 3174, 2020.

- [9] M. Cristian *et al.*, “Average monthly rainfall forecast in romania by using k-nearest neighbors regression,” *Analele Universităţii Constantin Brâncuşi din Târgu Jiu: Seria Economie*, vol. 1, no. 4, pp. 5–12, 2018.
- [10] M. F. Dacrema, P. Cremonesi, and D. Jannach, “Are we really making much progress? a worrying analysis of recent neural recommendation approaches,” in *Proceedings of the 13th ACM Conference on Recommender Systems*, 2019, pp. 101–109.
- [11] Y. Dash, S. K. Mishra, and B. K. Panigrahi, “Rainfall prediction for the kerala state of india using artificial intelligence approaches,” *Computers & Electrical Engineering*, vol. 70, pp. 66–73, 2018.
- [12] J. Diez-Sierra and M. del Jesus, “Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods,” *Journal of Hydrology*, p. 124789, 2020.
- [13] J. Du, Y. Liu, Y. Yu, and W. Yan, “A prediction of precipitation data based on support vector machine and particle swarm optimization (psosvm) algorithms,” *Algorithms*, vol. 10, no. 2, p. 57, 2017.
- [14] Y. Du, R. Berndtsson, D. An, L. Zhang, F. Yuan, C. B. Uvo, and Z. Hao, “Multi-space seasonal precipitation prediction model applied to the source region of the yangtze river, china,” *Water*, vol. 11, no. 12, p. 2440, 2019.
- [15] Felix Landerer, “Jpl tellus grace level-3 monthly land water-equivalent-thickness surface mass anomaly release 6.0 version 03 in netcdf/ascii/geotiff formats. ver. rl06 v03. po.daac, ca, usa.” 2020, dataset accessed [2019-08/15] - <https://doi.org/10.5067/TELND-3AJ63>.
- [16] A. Haidar and B. Verma, “Monthly rainfall forecasting using one-dimensional deep convolutional neural network,” *IEEE Access*, vol. 6, pp. 69 053–69 063, 2018.
- [17] M. Huang, R. Lin, S. Huang, and T. Xing, “A novel approach for precipitation forecast via improved k-nearest neighbor algorithm,” *Advanced Engineering Informatics*, vol. 33, pp. 89–95, 2017.
- [18] E. A. Hussein, M. Ghaziasgar, and C. Thron, “Regional rainfall prediction using support vector machine classification of large-scale precipitation maps,” in *IEEE 23rd International Conference on Information Fusion (FUSION)*. IEEE, 2020, pp. 1–8.
- [19] E. A. Hussein, M. Ghaziasgar, C. Thron, M. Vaccari, and A. Bagula, “Basic statistical estimation outperforms machine learning in monthly prediction of

- seasonal climatic parameters,” *Atmosphere*, vol. 12, no. 5, 2021. [Online]. Available: <https://www.mdpi.com/2073-4433/12/5/539>
- [20] E. A. Hussein, M. Ghaziasgar, C. Thron, M. Vaccari, and Y. Jafta, *Rainfall Prediction Using Machine Learning Models: Literature Survey*. Springer.
- [21] E. A. Hussein, C. Thron, M. Ghaziasgar, A. Bagula, and M. Vaccari, “Groundwater prediction using machine-learning tools,” *Algorithms*, vol. 13, no. 11, p. 300, 2020.
- [22] M. Jacobson, R. J. Charlson, H. Rodhe, and G. H. Orians, *Earth System Science: from biogeochemical cycles to global changes*. Academic Press, 2000.
- [23] J. Jing, Q. Li, and X. Peng, “Mlc-1stm: Exploiting the spatiotemporal correlation between multi-level weather radar echoes for echo sequence extrapolation,” *Sensors*, vol. 19, no. 18, p. 3988, 2019.
- [24] H. A. Karimi, *Big Data: techniques and technologies in geoinformatics*. Crc Press, 2014.
- [25] D. Kumar, A. Singh, P. Samui, and R. K. Jha, “Forecasting monthly precipitation using sequential modelling,” *Hydrological sciences journal*, vol. 64, no. 6, pp. 690–700, 2019.
- [26] K. Lakshmaiah, S. M. Krishna, and B. E. Reddy, “Application of referential ensemble learning techniques to predict the density of rainfall,” in *2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT)*. IEEE, 2016, pp. 233–237.
- [27] J. Lin, “The neural hype and comparisons against weak baselines,” in *ACM SIGIR Forum*, vol. 52, no. 2. ACM New York, NY, USA, 2019, pp. 40–51.
- [28] M. Mallika and M. Nirmala, “Chennai annual rainfall prediction using k-nearest neighbour technique,” *International Journal of Pure and Applied Mathematics*, vol. 109, no. 8, pp. 115–120, 2016.
- [29] H. J. Miller and M. F. Goodchild, “Data-driven geography,” *GeoJournal*, vol. 80, no. 4, pp. 449–461, 2015.
- [30] V. Nourani, S. Uzelaltinbulat, F. Sadikoglu, and N. Behfar, “Artificial intelligence based ensemble modeling for multi-station prediction of precipitation,” *Atmosphere*, vol. 10, no. 2, p. 80, 2019.
- [31] B. Pan, K. Hsu, A. AghaKouchak, and S. Sorooshian, “Improving precipitation estimation using convolutional neural network,” *Water Resources Research*, vol. 55, no. 3, pp. 2301–2321, 2019.

- [32] M. Patel, A. Patel, D. Ghosh *et al.*, “Precipitation nowcasting: Leveraging bidirectional lstm and 1d cnn,” *arXiv preprint arXiv:1810.10485*, 2018.
- [33] Y. Peng, H. Zhao, H. Zhang, W. Li, X. Qin, J. Liao, Z. Liu, and J. Li, “An extreme learning machine and gene expression programming-based hybrid model for daily precipitation prediction,” *International Journal of Computational Intelligence Systems*, vol. 12, no. 2, pp. 1512–1525, 2019.
- [34] Q. B. Pham, S. I. Abba, A. G. Usman, N. T. T. Linh, V. Gupta, A. Malik, R. Costache, N. D. Vo, and D. Q. Tri, “Potential of hybrid data-intelligence algorithms for multi-station modelling of rainfall,” *Water Resources Management*, vol. 33, no. 15, pp. 5067–5087, 2019.
- [35] N. Ramsundram, S. Sathya, and S. Karthikeyan, “Comparison of decision tree based rainfall prediction model with data driven model considering climatic variables,” *Irrigation and Drainage Systems Engineering*, 2016.
- [36] R. Sato, H. Kashima, and T. Yamamoto, “Short-term precipitation prediction with skip-connected prednet,” in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 373–382.
- [37] E. Shi, Q. Li, D. Gu, and Z. Zhao, “Convolutional neural networks applied on weather radar echo extrapolation,” *DEStech Transactions on Computer Science and Engineering*, no. csae, 2017.
- [38] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. chun Woo, “Convolutional lstm network: A machine learning approach for precipitation nowcasting,” *ArXiv*, vol. abs/1506.04214, 2015.
- [39] X. Shi, Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, “Deep learning for precipitation nowcasting: A benchmark and a new model,” in *Advances in neural information processing systems*, 2017, pp. 5617–5627.
- [40] S. Singh, S. Sarkar, and P. Mitra, “Leveraging convolutions in recurrent neural networks for doppler weather radar echo prediction,” in *International Symposium on Neural Networks*. Springer, 2017, pp. 310–317.
- [41] J. Sulaiman and S. H. Wahab, “Heavy rainfall forecasting model using artificial neural network for flood prone area,” in *IT Convergence and Security 2017*. Springer, 2018, pp. 68–76.
- [42] Q.-K. Tran and S.-k. Song, “Multi-channel weather radar echo extrapolation with convolutional recurrent neural networks,” *Remote Sensing*, vol. 11, no. 19, p. 2303, 2019.

- [43] C. Valencia-Payan and J. C. Corrales, “A rainfall prediction tool for sustainable agriculture using random forest,” in *Mexican International Conference on Artificial Intelligence*. Springer, 2018, pp. 315–326.
- [44] Y. Wang, M. Long, J. Wang, Z. Gao, and S. Y. Philip, “Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms,” in *Advances in Neural Information Processing Systems*, 2017, pp. 879–888.
- [45] L. Xu, N. Chen, X. Zhang, and Z. Chen, “A data-driven multi-model ensemble for deterministic and probabilistic precipitation forecasting at seasonal scale,” *Climate Dynamics*, pp. 1–20, 2020.
- [46] P.-S. Yu, T.-C. Yang, S.-Y. Chen, C.-M. Kuo, and H.-W. Tseng, “Comparison of random forests and support vector machine for real-time radar-derived rainfall forecasting,” *Journal of Hydrology*, vol. 552, pp. 92–104, 2017.
- [47] P. Zhang, Y. Jia, J. Gao, W. Song, and H. K. Leung, “Short-term rainfall forecasting using multi-layer perceptron,” *IEEE Transactions on Big Data*, 2018.
- [48] W. Zhuang and W. Ding, “Long-lead prediction of extreme precipitation cluster via a spatiotemporal convolutional neural network,” in *Proceedings of the 6th International Workshop on Climate Informatics: CI*, 2016.