Application of Diverse Machine Learning Models for Rainfall Prediction over various Climatic Regions of South Africa

Master's thesis submitted to the

Faculty of Natural and Applied Sciences,

Department of Computer Sciences, and Information Technology,

Sol Plaatje University, Kimberly

In partial fulfilment of the requirements for the degree of

MASTER OF SCIENCE IN E-SCIENCE

By

Jeremiah Ayodele Ogunniyi

Dr. Ibidun Obagbuwa	Main Supervisor
Prof. Mohamed Ahmed	Co-supervisor

December 2023

Declaration

I, Jeremiah Ayodele Ogunniyi, declare that this research work is my own, unaided work. It is being submitted for the degree of Master of e-Science at Sol Plaatje University, Kimberly. It has not been submitted for any degree or examination in any other university.

Abstract

This study focused on the prediction of rainfall in different climatic zones in South Africa using four machine learning models. The models used in this study are the linear regression, random forest, support vector machine, and the ridge and lasso regression. South Africa was divided into nine using the Koppen-Geiger climate classification system and three cities were selected for each climatic zone. Atmospheric datasets from the South African Weather Service from 1991 to 2023 and the National Aeronautics and Space Agency from 1983 to 2023 were used for this study. These datasets were trained and tested using the four models. The monthly rainfall predictions obtained after training and testing are then compared with the actual datasets to validate the accuracy of the models. Evaluation metrics such as mean average error, mean square error, root mean square error, correlation coefficient, and coefficient of determination were used to access the accuracy of each model. The best model for almost all climatic zones were the support vector machine and by random forest. Linear regression and ridge and lasso regression also performed well in various regions. It was however difficult to accurately predict rainfall under the warm and dry summer. This was attributed to the unpredictable atmospheric variability in this region. Also, in regions where there is little rainfall, the models performed worse compared to climatic zones with rainfall above 5mm. This study also showed that for better predictive performance, atmospheric parameters such as dew point, cloud cover, and water vapour are the most essential. Using the random forest model, monthly rainfall for 2024 was predicted and compared with 2022 and 2023 rainfall.

Dedication

This research work is dedicated to the Almighty God, the Author and Finisher of my faith, my Strong Tower and Potentate.

Acknowledgements

My sincere gratitude goes to the Almighty for His help, strength, grace, and mercies that are new every morning. I would like to appreciate my supervisors also, Dr Ibidun Obagbuwa for her patience and understanding. Her guidance, input, contributions to this research work are not unnoticed. I would also like to thank my co-supervisor, Prof Ahmed for his technical input which improved the quality of this work. Dr Collin Chibaya for ensuring that I got funding for the project and Dr Olusanya Micheal for assisting in securing admission for the program.

I would love to thank the Department of Science and Technology for providing me with funding for my studies, the South African Weather Service and the National Aeronautics and Space Agency for providing datasets for this study. Dr Peter and Dr Mrs Seun Olukanmi for housing me during the period of my course work and assisting me when I had challenges with some modules. Mr Peter Popoola of his tremendous assistance during my coursework, helping me learn the ropes in Python, Mr Victor Uzor, Dr Luke Joel for their assistance.

My wife who stood my me through it all, encouraging and praying for me, I appreciate and love you dearly. My beloved elder brother, Dr Daniel, and big sister Dunsin, my loving dad and mum for always supporting me, thanks so much, your baby made it. My in-laws for the unconditional love, encouragement, and support, thanks so much. Finally, to the body of Christ, Deeper Christian Life Ministry in South Africa, especially in Westville, thank you very much.

Table of Contents

List of Figures

List of Tables

List of Equations

Appendix

List of Output

Jeremiah Ayodele Ogunniyi, Ibidun Christiana Obagbuwa, and Mohamed Ahmed, Machine Learning Models for Rainfall Prediction over Arid Climatic Regions of South Africa submitted to International Journal of Cognitive Computing in Engineering. Manuscript Number: IJCCE-D-24-00030.

Chapter 1

1.0 Introduction

The impact of weather and climate on daily activities can not be over emphasized. Different human activities and programs such as environmental, economic, social, and political often rely on accurate weather predictions. These predictions are done using hydrological numerical models (Ghazikhani, 2022). These models are based on physical laws such as momentum, energy, conservation of mass. Using numerical models, the important physical processes taking place at all levels (atmosphere, ground-level, and soil) are described with their impact on variables such as wind, water vapour, temperature, precipitation, clouds, and pressure. However, these models are quite complex and depend highly on correct information, equations and many super computers which makes the process challenging. With the increase in the amount of data on daily basis, numerical models will only get more complex in predicting accurately weather and climate parameters, while also increasing the probability of errors (Jeong & Yi, 2023). To eliminate these model errors, post-processing is done (Lucatero et al., 2018). However, since most quantities to be measured have limited time scales, it reduces the effectiveness of the elimination of errors (Sexton et al., 2019). This leads to the need of machine learning (ML) models and algorithms for weather prediction.

The advent of the 21st century brought about increase in big data, supercomputers with high computational power. This led to the era of machine learning and artificial intelligence with one of its applications being weather forecasting. Machine learning models in weather forecast in time past have been limited in its application due to computational architecture and power constraints. However, in recent times, these constraints have been overcome with the use of the graphic processing unit (which is faster) and increased computer memory which makes calculations efficient. These new computational methods can be described as big data, machine learning or artificial intelligence. (Huntingford et al., 2019).

Many researchers have written about the importance and significance of machine learning algorithms for various applications. When labelled datasets are available, they can be used as training datasets which can be used to build models that can be tested and evaluated. If the result of the model is satisfactory, such models can be used for any type of classification and regression. These models are called supervised learning. Under supervised learnings, we have models such as Support Vector Machines (SVM), Random Forest (RF), Logistic Regression, K-Nearest Neighbor (KNN), Neural Networks (NN), XGBoost (XGB), Linear Regression Models (LRM), Generalized Linear Models (GLM) among others. There is another group of machine learning algorithms that do not need the datasets to be labelled for prediction. This is known as unsupervised learning. Examples include Principal Component Analysis (PCA), K-means, Hierarchical Clustering (Bochenek & Ustrnul, 2022). These applications have brought improvement to transport systems, healthcare, security, and defence networks, and in every area of life. The availability of these datasets and supercomputers with high computational power and speed have made prediction accuracy better and faster with reduced level of uncertainty (Huntingford et al., 2019).

In recent times, researchers have suggested the use of machine learning algorithms for weather prediction. Bochenek and Ustrnul (2022) reviewed about 500 publications from 2018 to determine the future of weather and climate prediction and concluded that machine learning models are the future of weather forecasting. Wang et al (2019) also suggested the use of machine learning algorithms for weather prediction due to unsatisfactory performance while using the numerical weather prediction. Hewage et al (2021) pointed to the high computational power and complex mathematical equations to solve as reasons to change from numerical weather prediction to machine learning models. Their results showed that though machine learning models were ‘lightweight data-drive’, they performed better than the numerical models.

In south Africa, numerical weather model is still being used for prediction (Landman et al., 2012; Sumbiri & Afullo, 2021). South African Weather Service uses the unified model of the United Kingdom Met Office which is also used among the southern African countries (Bopape et al., 2021). The forecasts are made using the South African Weather Service Cray XC30 which is a high performing computing system. Using a grid spacing of 4.4km, the united model can forecast up to 3 days. If the grid spacing is expanded to 16km, forecast can be made up to 10 days (Bopape et al., 2021). This shows the limitation of numerical method for rainfall and weather prediction as it cannot predict beyond 10 days at most. With the success of machine learning models in rainfall prediction across the world, this work seeks to explore different models for a long-term (1-year) forecast.

Despite South Africa being bounded by the Atlantic and Indian Ocean, South Africa is susceptible to drought (Muyambo et al., 2017). Extreme droughts which last for years are mostly caused by El Nino Southern Oscillation (ENSO) which is a quasi-periodic invasion of war sea surface waters into Pacific Ocean. South Africa experienced drought in 2015 and 2016 which was attributed to a strong El Nino event (Baudoin et al., 2017). This resulted in reduction in agricultural productivity which led to importing of grains instead of exports, water shortages and significant negative impact on the economy. This drought was termed the worst in 23 years after the drought of 1992 to 1995. Drought leads to reduction in crop yields and animal productivity, and it is expected that the frequency, intensity, and duration of droughts will increase due to climate change and anthropogenic activities thereby affecting livelihood (Mare et al., 2018). In 2015, the economic damage attributed to drought in Africa was estimated to be about US $2.4 billion and US$ 354 million in the Southern African region affecting about 3.2 million people. In South Africa, the damage was about US $250 million with 2.7 million people affected, resulting in 8.4% reduction in agricultural production and 15% reduction in livestock. Accurate prediction of rainfall might have reduced the advance effect caused by the drought as farmers and government would be better prepared instead of the government spending almost a billion rand on relief and support for farmers. However, due to global warming and climate change, it is becoming increasingly difficult to accurately predict rainfall as it depends on several physical factors of the hydrological cycle.

1.1 Problem Statement

Most research carried out on weather forecasting especially rainfall prediction has been carried out outside Africa. Most weather stations in Africa still rely on numerical and statistical methods for their forecast (Meque et al., 2021; Milton et al., 2017). These methods are time consuming and very expensive (Chen et al., 2022). Recent research in Africa has shown the potential for machine learning applications. Bamisile et al (2020) used machine and deep learning models for solar radiation comparison. These models were applied to four northern states (Borno, Kano, Yobe, Zamfara) in Nigeria using an hourly time step 12 years datasets. Their result revealed a coefficient of determination value of 0.89 when using the support vector regression and 0.95 while using the recurrent neural network.

Bouras et al (2021) used four ML algorithms (MLR, SVM, RF, and XGB) to forecast crop yield (cereal). Their results showed the effectiveness of applying machine learning models for cereal yield prediction. Their models showed that they can accurately predict crop yield with an R2 of 0.88. Cedric et al (2022) used different ML models were also used to predict crop yield production in six West African countries. Their result revealed coefficients of determination of 95.3%, 93.15%, and 89.78% for decision tree, K-Nearest Neighbor, and logistic regression respectively. Machine learning algorithms have also been used in Sub-Saharan Africa to predict malaria occurrence using climate variability (Nkiruka, 2021). They stated that the result of their research will aid decision making as well as adequate preparation in future outbreaks.

Researchers have shown the possibility of using machine learning models for weather prediction across the world. They have also shown that it is faster, more accurate, does not need high computational power, and most importantly, it is the future of weather prediction. This work therefore seeks to explore the use of machine learning models for rainfall prediction in South Africa since it still makes use of numerical weather prediction.

1.2 Research Questions

This research work aims to answer the following questions:

i. Which machine learning algorithm can best predict daily rainfall in different South African climatic zones using historical datasets?

ii. Using various evaluation metrics, which model is best suitable for rainfall in each climatic zone?

iii. Based on the best models using evaluation metrics, can rainfall pattern for the different climatic zones be predicted?

iv. Using the best model, can 2024 rainfall be predicted?

1.3 Research Objectives

The objectives of this research are to

i. compare how different machine learning models predict daily rainfall over various South African climatic zones.

ii. use different evaluation metrics to determine the best model for rainfall prediction for the different climatic zones.

iii. compare how the different models predict rainfall for different climatic zones in South Africa.

iv. use the best model to predict 2024 rainfall.

1.4 Study Locations

The study location is based on the Koppen-Gieger climate classification of South Africa. South Africa is broadly classified into three: Arid, subtropical wet (fully humid and dry winter) and subtropical dry (hot summer). This study also considered the sub-group climate classification such as the cold and semi-arid steppe, cold arid desert, hoot and semi-arid steppe, and the hot arid desert under the arid classification. The subtropical highland with dry winter and the humid subtropical with dry winter are considered under the subtropical wet. For subtropical dry, warm and dry summer, temperate oceanic without dry season, humid subtropical without dry season are the sub-divisions. For each sub-division, three locations are randomly selected for rainfall prediction. Table 1 shows the subdivisions, locations selected for the study as well as their average annual temperature and rainfall. Figure 1 shows the Koppen-Geiger climate classification for South Africa as well as the study locations.

Table 1: Table showing the Koppen-Geiger climate classification of South Africa the sub-divisions, study locations, annual average temperature, and annual average rainfall.

Climate Classification	Sub-division	Location	Average Temperature (ºC)				Annual Rainfall (mm)
			Summer (January)		Winter (July)		
			Max	Min	Max	Min	
Arid	Cold and semi-arid steppe	Bloemfontein	29	15	15	-2	469
		Springfontein	29	16	14	3	287
		Welkom	32	17	20	2	577

	Cold arid desert	Alexander Bay	25	16	23	9	50
		Beautfort West	33	16	19	6	392
		Bristown	31	20	15	5	168
	Hot and semi-arid steppe	Kimberly	33	18	19	3	350
		Mahikeng	31	17	22	4	571
		Port Elizabeth	25	18	20	9	563
	Hot arid desert	Lauville	21	13	16	10	301
		Musina	34	21	25	7	372
		Upington	36	20	21	4	219
Subtropical wet	Humid subtropical with dry winter	Dundee	26	16	19	5	684
		Louis Trichardt	30	12	22	3	540
		Nelspruit	29	19	23	6	934
	Subtropical highland with dry winter	Harrismith	26	14	19	2	973
		Johannesburg	26	15	17	4	543
		Newcastle	30	17	22	4	895
Subtropical dry	Humid subtropical without dry season	Durban	28	21	23	11	828
		Port Edward	27	21	22	13	1044
		Richards Bay	29	21	23	12	1228
	Temperate oceanic without dry season	East London	26	18	19	10	732
		George	25	15	15	7	657
		Mthatha	27	16	21	4	1044
	Warm and dry summer	Bredasdorp	25	18	18	11	463
		Cape Town	26	16	16	7	475
		Clanvilliam	32	17	18	6	420

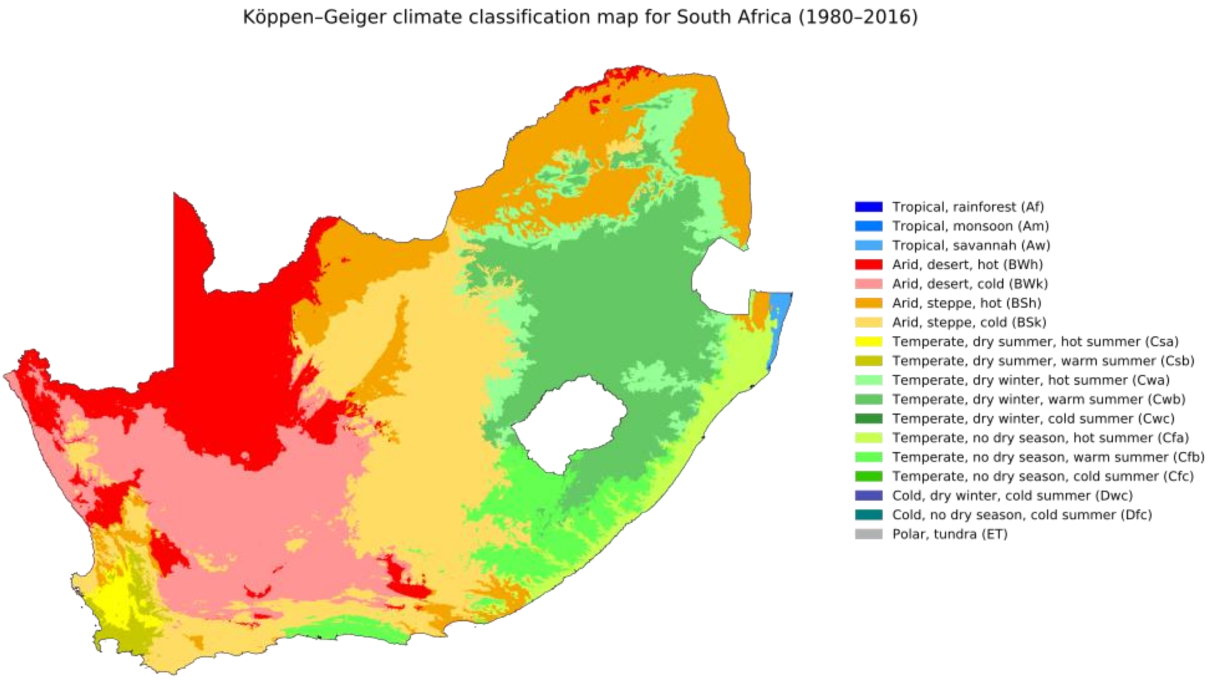

Köppen–Geiger climate classification map for South Africa (1980–2016)
Tropical, rainforest (Af)
Tropical, monsoon (Am)
Tropical, savannah (Aw)
Arid, desert, hot (BWh)
Arid, desert, cold (BWk)
Arid, steppe, hot (BSh)
Arid, steppe, cold (BSk)
Temperate, dry summer, hot summer (Csa)
Temperate, dry summer, warm summer (Csb)
Temperate, dry winter, hot summer (Cwa)
Temperate, dry winter, warm summer (Cwb)
Temperate, dry winter, cold summer (Cwc)
Temperate, no dry season, hot summer (Cfa)
Temperate, no dry season, warm summer (Cfb)
Temperate, no dry season, cold summer (Cfc)
Cold, dry winter, cold summer (Dwc)
Cold, no dry season, cold summer (Dfc)
Polar, tundra (ET)

Figure 1: Koppen-Geiger South Africa weather classification showing the various climates.

1.5 Significance of the Study

The results of accurate rainfall prediction can be applied to various aspects and sectors such as water management planning, mitigating natural disaster and early warning system, water allocation for agricultural purposes, water storage in dams for hydroelectricity, monitoring of droughts, inflow, and outflow of water in dams and reservoirs among others. With adequate information on the amount of water in a region, it will be difficult to maximise water usage.

It is therefore expected that this work will reveal the potential of machine learning algorithms for weather prediction in South Africa. It will also show which model is best suited for the various provinces. This work is expected to build a foundation for exploring machine learning models for weather prediction in Southern Africa.

In addition, since 60% of sub-Saharan Africa is susceptible to drought, and 2015 drought in South Africa cost farmers about R250 million in losses, this work will help farmers in identification of sowing dates and government agencies to prepare for drought if predicted.

1.6 Research justification

The use of machine learning models in rainfall prediction have increased across the world due to its advantages. The atmospheric variables affecting rainfall are complex and nonlinear. Rainfall prediction requires the use of variables such as dew point, relative humidity, windspeed, temperature, atmospheric pressure, water vapour among others. With an ever-growing database, machine learning models can easily reveal patterns and relationships between the variables using historical datasets. These models can handle the complexities in the datasets better than the numerical method of rainfall prediction. In addition, multiple models can be combined to improve the accuracy of the forecast. With evolving weather conditions due to climate change, machine learning models can be designed to continuously learn and adapt to this change. Since the main aim is to improve forecasting accuracy, the use of machine learning models will enhance both precision and reliability of prediction which has resultant effects on the economy, agriculture, disaster, and water resource management. However, in South Africa, no work has been done using machine learning models to predict rainfall. Predictions have been done using numerical methods. This work therefore aims to explore the possibility of using machine learning models to predict rainfall since it has been tested in other continents and found accurate.

1.7 Scope of the study

Despite the advantages of machine learning models in rainfall prediction, the models depended on the availability and reliability of the datasets used. In some climatic zones, it was difficult to get datasets for all the atmospheric variables needed for rainfall prediction. Also, three cities were selected in each climatic zones to examine the best model for rainfall prediction. These cities only represent a fraction of the whole climatic zone and can only give indications on the best models. Therefore, though this work suggests models to be used in different climatic zones, some models may perform better in cities that were not examined.

This work was carried out using four machine learning models. Although these models were selected based on their accuracy from research conducted in other countries and continents, there is the possibility that other models may perform better than the four models used.

1.8 Contribution to Knowledge

Machine learning models for rainfall prediction has been applied in different parts of the world and has shown to be more reliable and accurate compared to the traditional numerical method of weather forecasting. A significant contribution of this work to knowledge is the prediction of 2024 rainfall across 27 locations in South Africa using machine learning models. This research also assessed the performance of 4 machine learning models over different climatic zones in South Africa. This research forms a basis upon which other researcher can be carried out. It shows the potential to enhance prediction accuracy that will benefit different areas of the economy if properly applied. This work also gives a background to local forecast in 27 cities in South Africa under the nine climatic zones, three cities per climatic zones. This therefore gives an idea into what is happening in these zones. This work also helps researchers identify the key components for rainfall prediction in the different climatic zones which will help them refine selection of input features to improve the performance of the models.

1.9 Conclusion of the chapter

This chapter presents a background to the study, the importance of accuracy in rainfall prediction and its implications for the economy, agriculture, and infrastructure. It shows the ways in which rainfall is being predicted in South Africa, its limitations, and the need to explore the use of machine learning models. The objectives of this study and questions this research work aims to carry out are stated as well as the justification, limitation, and contribution to knowledge of this work.

In the next chapter, literatures are reviewed on the state-of-the-art machine learning models for rainfall prediction on a global and local stage. The evaluation metrics for these models and the gap in research will be discussed.

Chapter 2

2.0 Literature Review

2.1 State-of-the-art of Machine Learning Models for Rainfall Prediction over various Climatic Regions globally

Due to the computational cost and time-consuming nature of numerical weather models, researchers have explored alternative ways of forecasting weather on a large scale using various machine learning algorithms and have found them accurate. Ridwan et al (2020) used four machine learning methods for rainfall forecasting in Terengganu Malaysia. They used datasets from 10 stations around Terengganu for the study using autocorrelation function and projected error based on historical rainfall data. Their result revealed that Boosted decision tree regression had the best coefficient of determination compared to decision tree forest, neural network regression, and Bayesian linear regression. He et al (2021) examined the use of machine learning algorithms for sub-seasonal climate forecasting. They focused on predicting temperature and rainfall on two-week to two-month time scale. They used two non-deep learning (DL) models (AutoKNN, MultiLLR) and five machine learning models (Multitask Lasso, Gradient boosted trees (XGBoost), State of the art baselines, Encoder (LSTM)-Decoder (FNN) and CNN-LSTM models) for their research. Their results showed that machine learning models captured the predictability on sub-seasonal time scales with the ability to outperform the baselines set while with the best designed models for deep learning, machine learning model results were better.

Due to various catastrophic events in South America attributed to with weather and climate, Anochi et al (2021) deployed the use of ML algorithms for weather forecasting modelling in South America. They observed that using numerical methods for precipitation prediction could not accurately show precipitation patterns due to the absence of datasets specific for the regions. Their work showed the possibility of using machine learning algorithms for accurate precipitation prediction. They trained their models using 36-year datasets from 1980 to predict 2018 and 2019 rainfall. However, they observed large errors in summer months during the rainy season. This was attributed to local processes in the region that the algorithms could not learn as well as high energy during that period. They then showed that for all seasons except spring, training the networks with Tensor flow will make it perform better compared to those trained with neural networks.

Diez-Sierra and del Jesus (2020) investigated how 8 statistical and machine learning models performed while predicting daily precipitation in a semi-arid region. They used a 36-year rainfall data divided into training and test datasets using the 80-20 principle. The methods the used include the LRM, GLM, LR, RF, K-NN, SVM, K-means, and NN. They stated two advantages of machine learning methods over others to be that they do not need a known priori, neither do they need to make assumption on the distribution errors. Their result showed that the hyperparameter chosen affects the machine learning model. They further stated that if wrong parameters are used, it will affect the predictive capability and overfitting of the training models. They suggested that the hyperparameters should take higher values to avoid overfitting. The summary of their results showed that NN performed better than other models while predicting the intensity of rainfall. This was closely followed by SVM, K-NN and RF while the worst of their models was WT.

As a result of the successes of ML approaches in weather forecast in South America, Monego et al (2022) also applied these models to precipitation prediction using Gradient-Boosting (GB). They used the extreme gradient boosting (XGB) and TensorFlow (TF) models to train datasets from January 1980 to February 2020 using 75-25 principle. They considered meteorological variables such as air temperature at the surface and at 850 hPa, surface pressure, specific humidity at 850 hPa, zonal wind component at 500 hPa and at 850 hPa, meridional wind component at 850 hPa, as well as rainfall. Their result showed excellent performance with regards to pattern recognition for precipitation. They also showed that XGB performed better than TensorFlow deep neural network for all seasons except autumn while TensorFlow performed better than XGB only in autumn.

Baran et al (2020) compared machine learning models with parametric classification techniques using datasets from 2002 to 2014. Their result showed that when average rainfall is used as additional covariate, multilayer perceptron performed best. They also used these models to predict cloud clover. They stated systematic errors in calibration when several probabilistic classification methods are used. For their work, POLR had the best performance for two days, however, for long-term forecast, MLP performed best. They stated that the inclusion of atmospheric parameters such as pressure, humidity, temperature can improve the predictive performance of any model.

Bamisile et al (2020) compared the results of global and diffuse solar radiation using machine and deep learning models. For deep learning models, they used the artificial neural network (ANN), convolutional neural network (CNN), and the recurrent neural network (RNN) while for machine learning algorithms, they selected the SVM, polynomial regression (PR), and the RF. These models were applied to four northern states (Borno, Kano, Yobe, Zamfara) in Nigeria using an hourly time step 12 years datasets. They stated that the time spent in training the machine learning models reduced compared to the time spent in the training of deep learning models, however, from their results, the deep learning models performed better.

Appiah-Badu et al (2021) predicted precipitation using five ML models (Decision Tree, K-Nearest Neighbour, Multilayer Perceptron, Extreme Gradient Boosting, and Random Forest) in Ghana using 41-year climatic datasets from 2018. They divided the datasets into training and test sets using three different ratios (70:30, 80:20; 90:10) to assess the performance of the models. They analysed four climatic zones in Ghana: Coastal zone, Forest zone, Transitional zone, and Savanah zone. Their result showed that over the coastal zone of Ghana, MLP performed best using the 90:10 principle while XGB performed best using 80:20. For the three different ratios in the coastal zone of Ghana, KNN performed worst. Similarly, for the forest zone, at 90:10, MLP performed better than other models with KNN performing worst. Using 70:30, DT, RF, and XGB all had better precision, recall and f1 score. Over the transitional zones, all three ratios of RF and XGB performed best in the metrics used. The result also showed that MLP performed best using the 90:10 ratio, perhaps the best ratio to be used while considering multilayer perceptron. Finally, over the Savannah zone, RF and XGB also showed the best performance in 90:10, 80:20, and 70:30 both with rain and without rain. The reason for KNN performing worst in all zones was not investigated in the study. However, their study showed that machine learning models are good for rainfall prediction especially RF, XGB, and MLP.

In various parts of the world, machine learning algorithms have been applied for rainfall prediction. In India, Jose et al (2022) used various machine learning models to predict daily rainfall. Their result showed that long short-term memory performed best with a coefficient of determination of 0.9. In Australia, noted for the highest extreme temperature in the world, Polishchuk et al (2021) used random forest model to predict rainfall to know when there will be wildfire. Their results revealed an accuracy of 85.9% in their training model and 84.7% accuracy in the prediction of rainfall. Similarly, Raval et al (2021) used different machine learning models to predict rainfall using 10 years datasets. Their result revealed that logistic regression model had the highest classification with and f1 and precision of 86.9% and 97.1% respectively. Other researchers such as He et al (2022), Islam et al (2023), Sachindra et al (2018) among others have all used machine learning models to predict rainfall in Australia with accuracy above 80%. In South America, Ferreira and Reboita (2022) showed that the application of machine learning models in rainfall prediction has led to 75% error reduction in rainfall estimates. Their result also indicated the ability of their models to reproduce both the spatial and temporal variation for dry and wet seasons across south America. Despite high topographic gradients and unstable climatic conditions in south America, researchers have shown the possibility of deploring machine learning models for rainfall prediction with high degree of accuracy (Anochi et al., 2021; Gómez et al., 2023).

Based on the successful application of machine learning algorithms in different parts of the world, this research work seeks to explore its application for daily rainfall prediction in South Africa. The results from the literatures reviewed have shown the accuracy of machine learning models as well as their better performance compared with numerical weather forecasts. They have also indicated the various machine learning models for various climates and the possibility for improvements in their forecast. It is however important to note that for different regions, some machine learning models will be more appropriate. With this research work predicting rainfall in different climates across South Africa, it will give better indication on which machine learning model is appropriate for a climatic condition.

Table 1 shows different machine learning models that have been applied to rainfall prediction across the world and their correlation coefficient to prediction. Mekanik et al (2013) used both linear regression models and artificial neural networks for long-term rainfall forecasting in Australia. Their result showed poor generalization and forecasting ability in the east compared to other areas, with a correlation coefficient 0f 0.06 to 0.69 obtained. Also, Australia, a southern midlatitude country like South Africa Hossain et al (2020), used linear and nonlinear models for long-term seasonal rainfall forecasting. They used multiple linear regression for linear forecasting, while the artificial neural network was used for nonlinear regression. They applied the models to three stations and evaluated them using the typical statistical parameters. They obtained a correlation coefficient ranging from 0.35 to 0.83 for the linear regression models, while the artificial neural network performed much better with a correlation coefficient ranging from 0.76 to 0.90. Their mean square and absolute percentage error values also showed that nonlinear models performed better than linear models for their study. However, neither model was able to estimate the extreme spring rainfall accurately.

Similarly, Peter and Precious (2018) used multiple linear regression and artificial neural networks for seasonal rainfall prediction in Nigeria, west Africa using data from 1986 to 2017. They obtained a correlation coefficient of 0.66 for the linear regression model and 0.93 for the artificial neural network, showing that it performed better for predicting seasonal rainfall. In India, Swain et al (2017) also used a linear regression model for precipitation forecasting and obtained a correlation coefficient of 0.96 and a coefficient of determination of 0.97.

Support vector machine and multilayer perceptron were used in Zhang et al (2020) to predict annual and non-monsoon rainfall in India using relative humidity and annual rainfall data from 1991 to 2015. They obtained a correlation coefficient ranging from 0.59 in April to 0.80 in October, giving a yearly average of 0.71 for the support vector machine model. They suggested the combination of both models for better prediction accuracy. Pham et al (2020) developed advanced artificial intelligence models for daily rainfall prediction in Vietnam. They used relative humidity, wind speed, temperature, and solar radiation as the atmospheric parameters and rainfall as the output. Their result revealed that SVM best-predicted rainfall with a correlation coefficient of 0.85. This is like the value we obtained using the support vector machine. Other evaluation metrics, such as the mean absolute error, skill score, probability of detection, false alarm ratio, and critical success index, proved that the SVM performed better than other models. In Taiwan Yen et al (2019) obtained a correlation coefficient of 0.49 using the SVM, while Kisi and Cimen (2012) obtained a correlation coefficient 0.78 in Turkey.

The complete empirical ensemble mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecast in Pakistan was investigated in Ali et al (2020). They obtained a correlation coefficient of 0.74-0.99. They proposed a hybrid of random forest and kernel ridge regression for higher accuracy. In Iran, Lotfirad et al (2022) also used random forests to monitor and predict drought in various climates. They obtained a correlation coefficient ranging from 0.81 to 0.95 depending on the climatic region. Both values are also within the values obtained in this study when a random forest was used to predict rainfall. Zhang et al (2021) obtained a correlation coefficient of 0.71 in China, while Orellana-Alvear et al (2019) got a value of 0.83 in Ecuador while using the random forest to predict rainfall, similar to that obtained in this work.

Sajan and Kumar (2021) examined the forecasting and analysis of train delays and impact of weather using different machine learning models with the available historical data in India. Their result showed that lasso had a correlation coefficient of 0.82 while that of the support vector machine was 0.79. Still in India, Tiwari and Singh studied rainfall in Indian states and their predictive analysis using machine learning models. Their objective was to improve harvest of crops which highly depended on the pattern of rainfall the country receives especially the monsoon. Their result showed application of lasso for this purpose as they obtained a correlation coefficient of 0.65 when they compared their result with historical datasets from 1901 to 2017. Sungkawa and Rahuyu (2019) predicted extreme rainfall using the Bayesian Quantile regression in statistical downscaling modelling in Indonesia. This was the first study in Indonesia using the adaptive lasso in their modelling. For moderate rainfall, they obtained a correlation coefficient of 0.75 which increased to 0.90 during high extreme rainfall. This points to the possibility of the model performing better in regions with high amount of rainfall compared to regions that experience intermittent drought.

Table 2: Table showing the correlation coefficients from selected previous studies for linear regression, random forest, support vector machine, and the ridge and lasso regression models.

	Models (Correlation Coefficient)			
	LR	**RF**	**SVM**	**Ridge & Lasso**
This Study	0.85	0.812	0.849	-
Mekanik et al (2013)	0.06-0.69	-	-	-
Hossain et al (2020)	0.35-0.83	-	-	-
Peter and Precious (2018)	0.66	-	-	-
Swain et al (2017)	0.96	-	-	-
Zhang et al. (2020)	-	-	0.71	-
Pham et al. (2020)	-	-	0.85	-
Yen et al (2019)	-	-	0.49	-
Kisi and Cimen (2012)	-	-	0.78	-
Ali et al (2020)	-	0.74-0.99	-	-
Lotfirad et al (2022)	-	0.81-0.95	-	-
Zhang et al. (2021)	-	0.71	-	-
Orellana-Alvear et al (2019)	-	0.69	-	-
Tiwari and Singh (2020)	-	-	-	0.65
Sungkawa (2019)	-	-	-	0.90
Zaikarina et al (2016)	-	-	-	0.75-0.90
Sajan and Kumar (2021)	-	-	0.79	0.82
He et al (2019)	-	-	-	0.66

2.2 Research gap

From the literatures reviewed, it is evident that most research carried out on rainfall prediction using machine learning models have been done in either Asia, North America, South America, and Europe. Researched carried out in Africa using machine learning models have been to evaluate wheat yield, global irradiance with no identifiable literature exploring the potential use of machine learning models for rainfall prediction. This research work therefore seeks to explore the potential use of machine learning models for rainfall prediction beginning in South Africa by examining it under the different climatic zones.

2.3 Machine Learning Models for Rainfall Prediction over various Climatic Regions of South Africa

This study explored the use of four machine learning models over the different climatic zones of South Africa.

2.3.1 Linear Regression

Regression analysis is a statistical tool for estimating the value of a dependent variable from an independent variable (Su et al., 2012). When two variables have a linear relationship, linear regression is used. Multiple linear regression is used when two or more independent variables have a linear relationship, while polynomial regression is used when the variables have a polynomial relationship (Maulud & Abdulazeez, 2020). Linear regression is one of the most common statistical techniques used in modelling for observational studies (Kumari & Yadav, 2018). It provides satisfactory approximation in modelling for small sample-size datasets. Maulud and Abdulazeez (2020) presented the theoretical background for linear regression. The first is a regression commonly used for forecasting, predicting, and determining the relationship between variables. Next is the regression model, where the independent variables predict the dependent variable, which can be a simple, multivariate polynomial regression. The least-square error in linear regression ensures that the predicted values approach the minimum of all possible regression coefficients. Kumari and Yadav (2018) gave five reasons for using linear regression. Linear regression is used for descriptive studies as they help to analyse the strength of relationships between outcomes and predictors. They are also used to adjust the effects of covariates and to estimate the important factors affecting the dependent variables. They added that it helps analyse the extent of prediction and quantify new predictions.

Linear Regression Algorithm

Linear regression provides a linear relationship between an independent variable (x-axis) and dependent variable (y-axis) to predict future events. It is used to show relationships between continuous variables. It therefore shows how the dependent variable changes with respect to the independent variable. A simple linear regression is shown in figure (2). Mathematically, it is represented as

$$y = a_0 + a_1 x + \varepsilon$$

Equation 1: Mathematical representation of linear regression

Where y is the dependent variable or the target variable,

x is the independent variable or the predictor variable,

a_0 is the line intercept,

a_1 is the linear regression coefficient, and

ε is the random error.

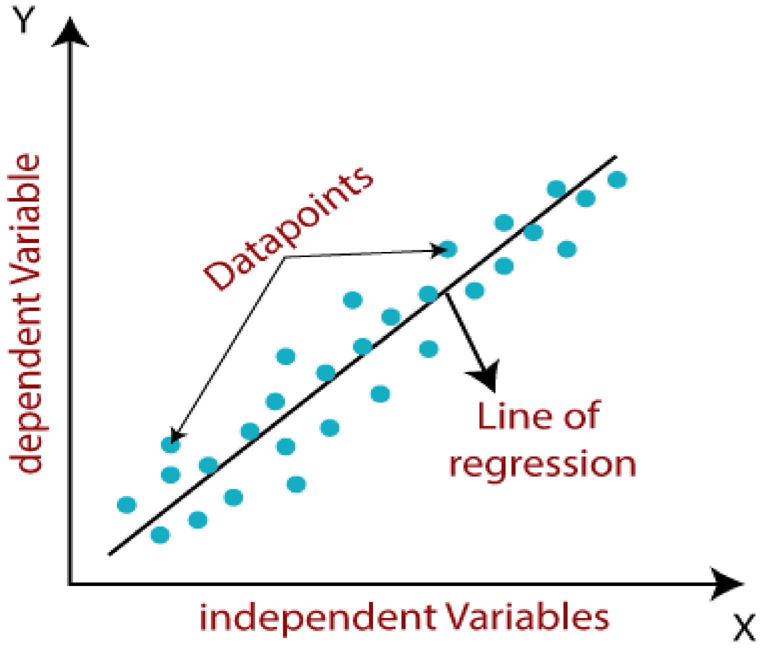


Figure 2. 1: Simple linear regression.

2.3.2 Random Forest

Random Forest was proposed by Leo Breiman in 2001 and is mainly used for classification and regression problems (Biau & Scornet, 2016). It is mostly used when there are more variables compared to observations. An ensemble method uses several decision trees and averages their aggregate predictions. It can rank variables based on their relevance and ability to discriminate target classes (Belgiu & Dragut, 2016). In Breiman's approach, each decision tree is formed by randomly selecting variables at each node. These variables are then split, after which the best split is calculated based on the features in the training set. This decision tree is then developed by maximizing the size without pruning, known as the Classification and Regression Trees (CART) methodology (Biau, 2012). Details on the theoretical framework for random forest can be found in (Breiman, 2001). Despite the growing interest in random forest and its accuracy, the mathematical forces behind it are not well known, nor are its statistical properties (Biau, 2012). Recent works have still not been able to explain the behaviour of random forests, only that they are accurate (Biau & Scornet, 2016).

Random Forest Algorithm

Random forest is a widely used machine learning model that combines the result of multiple decision trees to obtain a single result. Although decision trees are prone to bias and overfitting, however, random forest can predict more accurately since it uses multiple decision trees. Random forest algorithm can handle datasets containing both continuous variables (regression) and categorical variables (classification). Random forest is an ensemble learning technique since it makes use of multiple models. ensemble can either be bagging or boosting, however, random forest makes use of the bagging method. Steps involved in bagging:

Subset selection: A random sample is chosen from the datasets.

Bootstrap sampling: From the random samples, each model called bootstrap sample is created.

Independent model training: Each model is then trained independently on its corresponding bootstrap sample.

Majority voting: The results of all models are combined to determine the final output through majority voting. Through majority voting, the most predicted result is then selected.

Summarily, in the random forest algorithm, random samples and random features are selected from the datasets and individual decision trees are created from each sample. Each decision tree then generates an output which is aggregated, and the most predicted result is selected.

For example, figure (3) shows how random forest works. There is a fruit basket from which samples are taken to construct individual decision trees. Each decision generates an output of apple, apple, and banana. Since apple is the most common decision, the final output is apple.

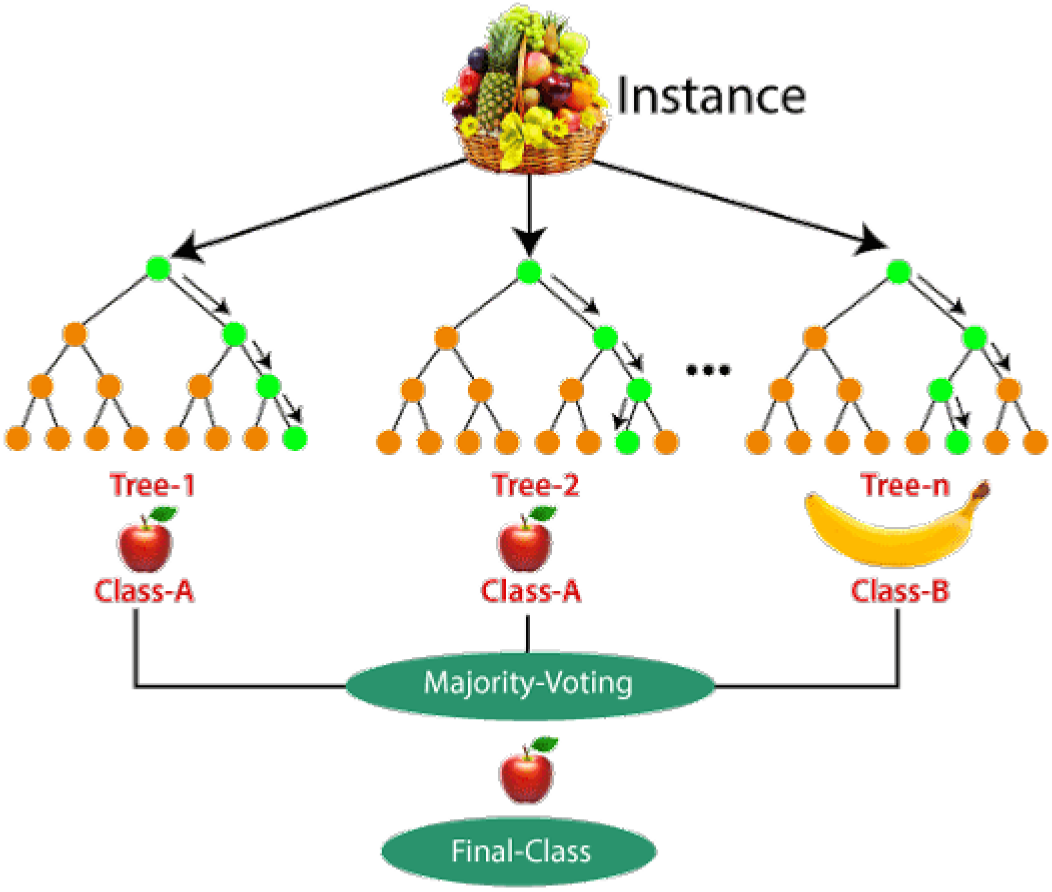


Figure 2. 2: Random Forest model used for fruit classification.

2.3.3 Support Vector Machine

The support vector machine was developed by Vapnik and coworkers (Mammone et al., 2009). It is an algorithm that assigns labels to objects through learning by example with the primary goal of prediction (Noble, 2006). The support vector machine is used for regressions with three outcomes: continuous, multinomial, and binary. It can generalize well even when trained with limited samples and is most suitable in remote sensing where limited reference data is provided (Mountrakis et al., 2011). Support vector machine does not make any assumption on the datasets as it is a supervised non-parametric statistical model. The model learns through a process known as structural risk minimization, which minimizes classification error. However, studies have shown a common limitation in the support vector machine: the selection of the kernel functions. A small kernel value can lead to overfitting, while high values can lead to over smoothing (Boswell, 2002). This problem is not limited to support vector machine but to all kernel models. Also, when the datasets are noisy due to distortions in the atmosphere and instruments, training the model becomes difficult, and the performance of the support vector machine classifier reduces (Mountrakis et al., 2011). Its application varies from face detection, remote sensing, handwriting recognition, pattern detection, text categorization, protein prediction, and gene expression analysis (Mammone et al., 2009). The theoretical breakdown of support vector machine can be found in (Boswell, 2002; Moguerza & Munoz, 2006).

For this study, four different kernels (linear, radial, polynomial, and sigmoid) were tested experimentally on the data. The optimal values for the model parameters, cost, gamma, and epsilon were determined experimentally by assessing the effect on the forecasting performance (in terms of mean absolute error, mean square error and root mean square error) of the different possible combinations of the values for the three parameters from a predefined set of values. For this study, the radial kernel was selected as it outperformed other kernels. The optimal values for the parameters were also determined experimentally in the model.

Support Vector Machine Algorithm

Support vector machine algorithm is also used for both regression and classification problems. It has the advantage of higher speed and better performance with limited datasets. The support vector regression approximates the relationship between the input variable and continuous variable while minimizing prediction error. It maps input variables into a high-dimensional feature as well as determines the hyperplane which maximizes the distance between the hyperplane and the nearest datapoint. The mapping of the input variable to high-dimensional feature is done by the kernel function. The process: once the libraries are imported and read, feature scaling is done to normalize the data. The model is then fitted to the datasets and a new result is predicted.

2.3.4 Ridge and Lasso

The ridge and lasso models are regularization techniques for preventing overfitting. This is achieved by adding a penalty to the loss function. This penalty is added to the square of coefficient in ridge regression and to the absolute value of the coefficient in lasso regression (Yang & Wen, 2018). The theory of ridge regression was first introduced by Hoerl and Kennard in 1970 as a possible solution to the error in least square estimators in linear regression when independent variables are highly correlated known as collinearity. The negative impact of collinearity is well documented (McDonald, 2009; Yang & Wen, 2018). Various approaches were developed to reduce the negative impact, but they mostly centred on variable elimination where one or more of the independent variables are removed to improve the performance of the model. However, with ridge and lasso, this problem is solved without removing any independent variable with the introduction of the penalty function (McDonald, 2009). The ridge and lasso is an improvement on the ridge regression. Despite improvement in prediction while using ridge regression, it does not perform covariate selection which makes models difficult to interpret. However, with the ridge and lasso regression, models become easier to interpret. Lasso stands for Least Absolute Shrinkage and Selection Operator (Tibshirani, 1996).

Ridge and Lasso Algorithm

Ridge and Lasso are commonly used for large datasets that have the tendency to overfit or cause computational challenges. This is done using the penalty function called regularization. The magnitude of the coefficient features is penalized and the error between actual and predicted values are minimized. This is achieved by adding the sum of squared coefficients.

Objective = RSS + α *(sum of squared coefficients)

Equation 2: Formula for ridge and lasso model.

Where α is the regularization parameter and RSS is the Residual Sum of Squares.

If α = 0, the objective is the same as simple linear regression and we obtain the same coefficient.

If α = ∞, the coefficients will be zero since anything less than zero will make the objective infinity

If 0 < α < ∞, the magnitude of α will determine the weight given to the objective and the coefficient will be between 0 and that in a simple linear regression.

2.4 Evaluation metrics for Machine Learning Models for Rainfall Prediction

The necessity to assess the accuracy and reliability of machine learning models for rainfall prediction gives rise to the evaluation metrics. These metrics help to identify the most suitable model for specific tasks. The most common models for rainfall prediction are the root mean square error, mean absolute error, mean absolute percentage error, coefficient of determination, and the normalized root mean square error. While the root mean square error and the mean absolute error are mostly used for continuous variables, mean absolute percentage error and coefficient of determination are used to measure the accuracy of the model's predictions.

2.4.1 Root Mean Square Error

The Root Mean Square Error is one of the two major performance indicators for a regression model along with the Mean Absolute Error. It measures the average of the squared errors between the values predicted by a model and its actual values. This gives an estimation of how well the model can predict the target value. The lower the values, the better the predictive performance of the model.

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - x_j)^2}{N}}$$

Equation 3: Formula for root mean square error.

Where N is the number data points, x_i is the actual observation time series, x_j is the estimated time series and i is the variable.

2.4.2 Mean Absolute Error

It is the second major performance indicator for a regression model. It calculates the average of the absolute errors between the predicted values by the model and the actual values. Also, the lower the values, the better the predictive performance of the model.

$$\text{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Equation 4: Formula for mean absolute error.

where y_i is the prediction, x_i is the true value and n is the total number of data points.

2.4.3 Mean Absolute Percentage Error

This is also known as the mean absolute percentage deviation. This measures the percentage of the absolute errors relative to the actual values. It is given by the formula:

$$\text{M} = \frac{1}{N} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{A_t} \right|$$

Equation 5: Formula for mean absolute percentage error.

Where M is the mean absolute percentage error, n is the number of times the summation iteration occurs, A_t is the actual value and F_t is the forecast value.

2.4.4 Coefficient of Determination

This is commonly referred to as R-squared which is a proportion of the variance in the dependent variable that is predictable from the independent variable. The value ranges from 0 to 1, measuring how well the model predicts the outcome with one indicating a perfect fit. It is represented with this formula:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Equation 6: Formula for Coefficient of Determination

Where R^2 is the coefficient of determination, RSS is the sum of square residuals and TSS is the total sum of squares.

2.4.5 Normalized Root Mean Square Error

This is a fraction of the overall range resolved by the model. It is the root mean square error that is normalized by the range of observed values. This provides a measure of the error relative to the range of the data. It is represented by:

$$\text{NRMSD} = \frac{RMSD}{y}$$

Equation 7: Formular for normalized root mean square error.

Where y is the range (maximum value minus the minimum value)

2.5 Conclusion

Researchers have used various machine learning models for rainfall prediction across the world. However, in the selection of the models used by these researchers, nothing indicates the criteria for selection. Most researchers either compared different models to see which one is best applicable to their region or select a model that has been proven to be accurate for their use. Since there has been no criteria for model selection, this research work picks four common model that have been applied in different countries and have shown high degree of accuracy. The four models picked have been discussed with results obtained by researchers that made use of the models. Other models that were not used in this work were also discussed in this chapter.

Chapter 3

3.1 Methodology

3.1.1 Research Design

The locations chosen for this study will be selected based on two criteria. They will first be chosen using the different climatic zones. Cities in different regions in South Africa will also be chosen. The purpose of these criteria is to ensure that there is no model bias with regards to climate or region. Therefore, some cities will be selected with the same climatic conditions but in different regions. This will also assist in determining whether some models perform better under certain climatic conditions or whether some regions are easier to predict. If the models perform well across all climatic regions, it will indicate that the models can generalize. Once the climatic zones and cities have been identified, the data collection process will begin.

3.1.2 Data collection

40-year continuous datasets obtained from The National Aeronautics and Space Administration (NASA) website from 1983 was used for this study obtainable in the giovanni interactive visualization page. This was combined with datasets from the South African Weather Service (SAWS). Seven weather parameters daily measurements (dew point, temperature, rainfall, wind speed, relative humidity, water vapour, and cloud cover) were retrieved for each climatic zone and used for this study. Using the Koppen-Geiger climate classification, a point which corresponds to a major city will be picked in each climatic zones for analysis. Details on each dataset from NASA as well as timeseries examples can be found here:

https://giovanni.gsfc.nasa.gov/giovanni/#service=TmAvMp&starttime=&endtime=

3.1.3 Data analysis

Atmospheric parameter datasets: Datasets from 1983 to 2023 were retrieved from the NASA website to predict month rainfall. This was combined with datasets from the SAWS. Climatological parameters such as rainfall, temperature, dew point, relative humidity, water vapour, cloud cover and wind speed were retrieved over the different climatic zones corresponding to different cities in South Africa. The flowchart of the data analysis process is shown in figure 2 below.

3.1.4 Pre-processing

Once the data sets are obtained, pre-processing was done to identify the missing values and eliminate them as well as duplicate values. Outliers were then identified and removed from the datasets. Both datasets were combined to have a long-term historical data as data from SAWS was only made available from 1991.

3.1.5 Correlation analysis

A heatmap was done to determine the correlation between rainfall and the atmospheric variables used for this study.

3.1.6 Data splitting

The datasets were divided into training and test sets using the ratio 80:20.

3.1.7 Trained models

The models were trained using the four different models mentioned above on 80% of the datasets.

3.1.8 Test set

The remaining 20% will be tested to see how good each model performs and then evaluated using various metrics.

3.1.9 Evaluation models

The Mean Square Error (MSE), the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), correlation coefficient and the coefficient of determination for each model in all selected locations will also be determined to assess the predictive accuracy of the models. All these will be done using Python 3.9.

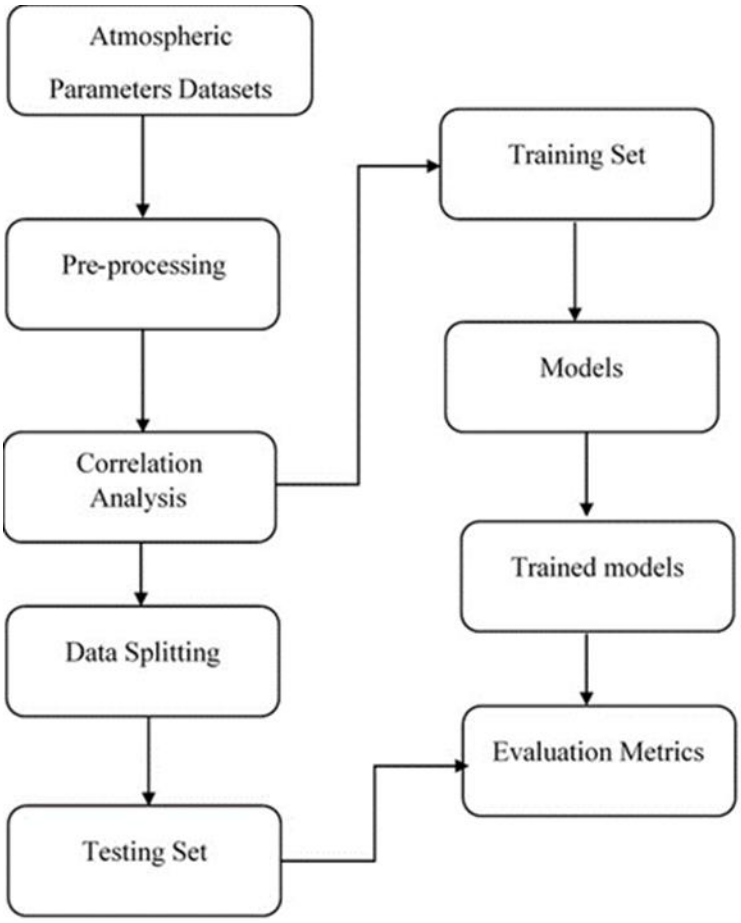


Figure 3. 1: Schematic view of the data analysis process

Chapter 4

4.0 Results and Discussion on Arid Climate Classification

4.1 Arid Climate Classification

4.1.1 Cold and Semi-Arid Steppe

Table 4.1 shows the evaluation metrics for rainfall prediction for three locations (Bloemfontein, Springfontein, and Welkom) under the cold and semi-arid steppe climate classification. The result reveals that all four machine learning models can be employed in the region. The support vector machine had the highest correlation coefficient for all three locations corresponding to 0.79, 0.74, and 0.80 for Bloemfontein, Springfontein, and Welkom respectively. This is closely followed by Linear regression and then Ridge and Lasso. However, for the coefficient of determination (R^2), the best model was the linear regression with values of 0.80, 0.76, and 0.82 for Bloemfontein, Springfontein, and Welkom respectively, followed by the Ridge and Lasso. For both Random Forest and Support Vector Machine, the values were below 0.50 except for Random Forest in Welkom. Previously, Moeletsi et al (2016) evaluated an inverse weighting method (IDW) for patching daily and 10-days rainfall for six weather stations in Free State South Africa using 58-year datasets. They showed that their IDW method was highly effective for daily and 10-day predictions. For Bloemfontein and Welkom, they obtained a R^2 value of 0.90 and 0.78 and a MAE of 3.17 and 5.60 for both locations respectively.

Figure 4.1 shows that temporal variation in rainfall for the three locations under the cold and semi-arid steppe for Linear Regression, Random Forest, Support Vector Machine as well as the Ridge and Lasso. The results showed that for all locations and algorithm, the model accurately predicted the seasonal variability of rainfall, though they all underestimated the amount of rainfall received. For Springfontein, the model correctly predicted the seasonal variation with

90% accuracy for the first five years between 2016 and 2020 but underestimated especially in 2021 and 2022. Linear regression, random forest and the support vector machine all modelled the spike in 2017 rainfall estimates but overestimated the increase while from 2020. They all underestimated the amount of rainfall received in Springfontein. The underestimation in 2021 and 2022 may be due to the region receiving more rainfall than usual. In a semi-arid climate, it experiences over 230 dry days in a year, an average annual temperature of 22^{o}C, and an average humidity of 48%. Unfortunately, none of the models could accurately predict the unexpected spike in the annual rainfall. This increase in 2021 was also recorded in Bloemfontein by all models. Weather Sa (2022) reported that rainfall in Bloemfontein in 2021 increased by about 150mm compared to the decadal average. Although similar increase was experienced in Welkom, all the models were able to predict the spike. However, none of the models could still accurately estimate the amount of rainfall received.

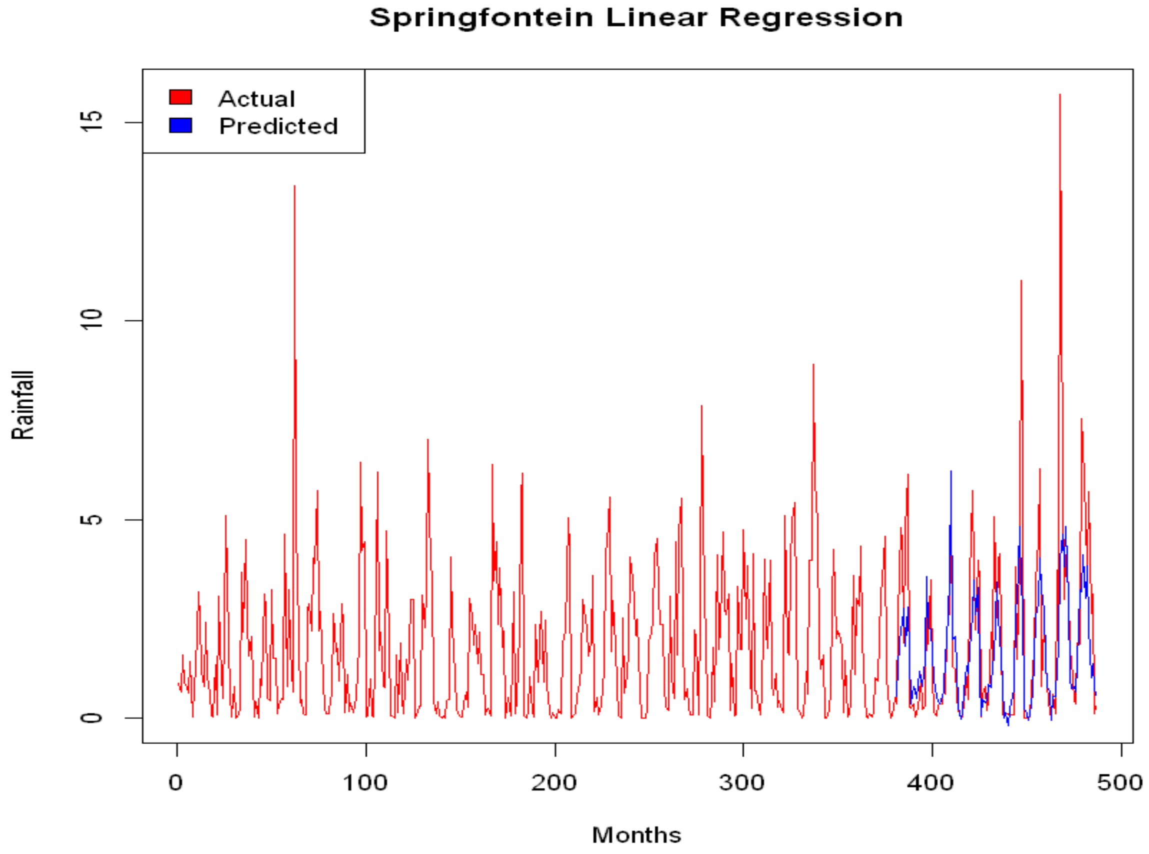
This pattern of high rainfall in 2021/2022 and underestimation of the model was expected in almost all locations as South Africa received above normal annual rainfall. This was attributed to the El Nino-Southern Oscillation (ENSO) being in a La Nina phase. Sivakumar and Fazel-Rastgar (2023) revealed the presence of active frontal system with continuous rainfall in 2022. They also attributed the increase to the injection of high humidity from extended warmer isotherms.

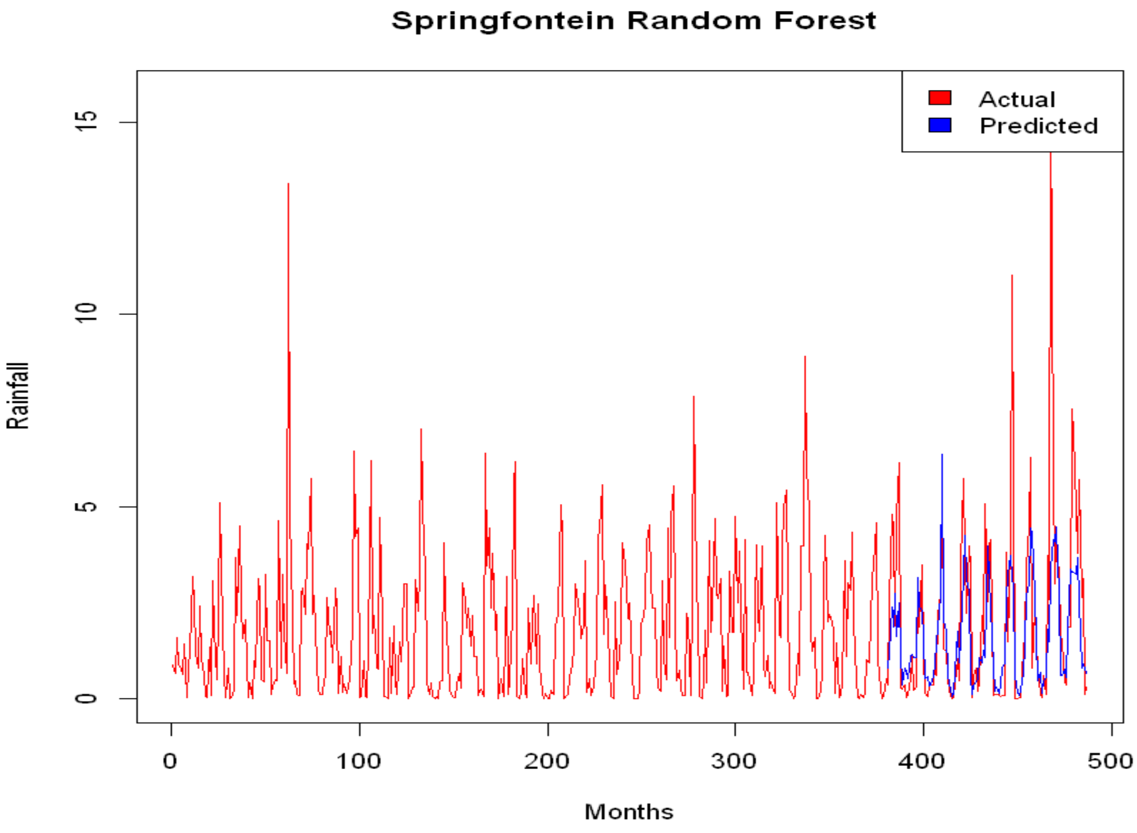
From the heatmap in appendix A, dew point, cloud cover, and water vapour are essential for rainfall prediction in Bloemfontein as their correlation with rainfall are 0.67,0.69, and 0.68 respectively. These same parameters correlated with rainfall with coefficients of 0.72,0.71, 0.71 for dew point, cloud cover and water vapour respectively for Springfontein. In Welkom, the parameters that correlated best with rainfall were dew point, cloud cover, water vapour and temperature with coefficients all corresponding to 0.68 except for temperature which has a correlation coefficient of 0.56. Relative humidity and wind speed had the lowest coefficients of 0.11 and -0.13 respectively in Bloemfontein. These two variables also had the lowest in Springfontein corresponding to 0.16 and -0.29 for relative humidity and wind speed respectively. In Welkom, relative humidity and rainfall had a correlation of 0.29 while rainfall and windspeed had a correlation of -0.026. This result reveals that relative humidity and wind speed are not very essential for rainfall prediction for cold and semi-arid steppe. Atmospheric parameters such as dewpoint, water vapour, cloud cover, and temperature should be used for accurate predictions.

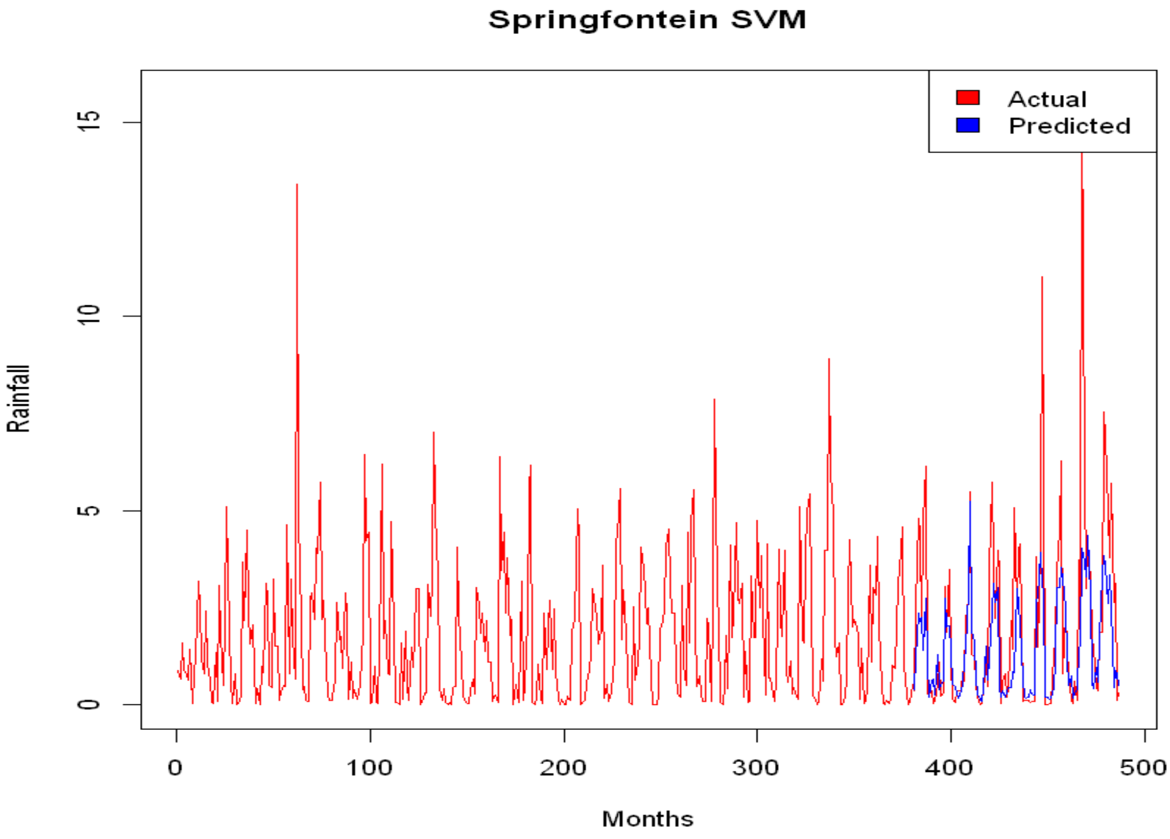
Table 4. 1: Table showing models evaluation metrics for cold and semi-arid steppe climate classification (Bloemfontein, Springfontein, and Welkom)

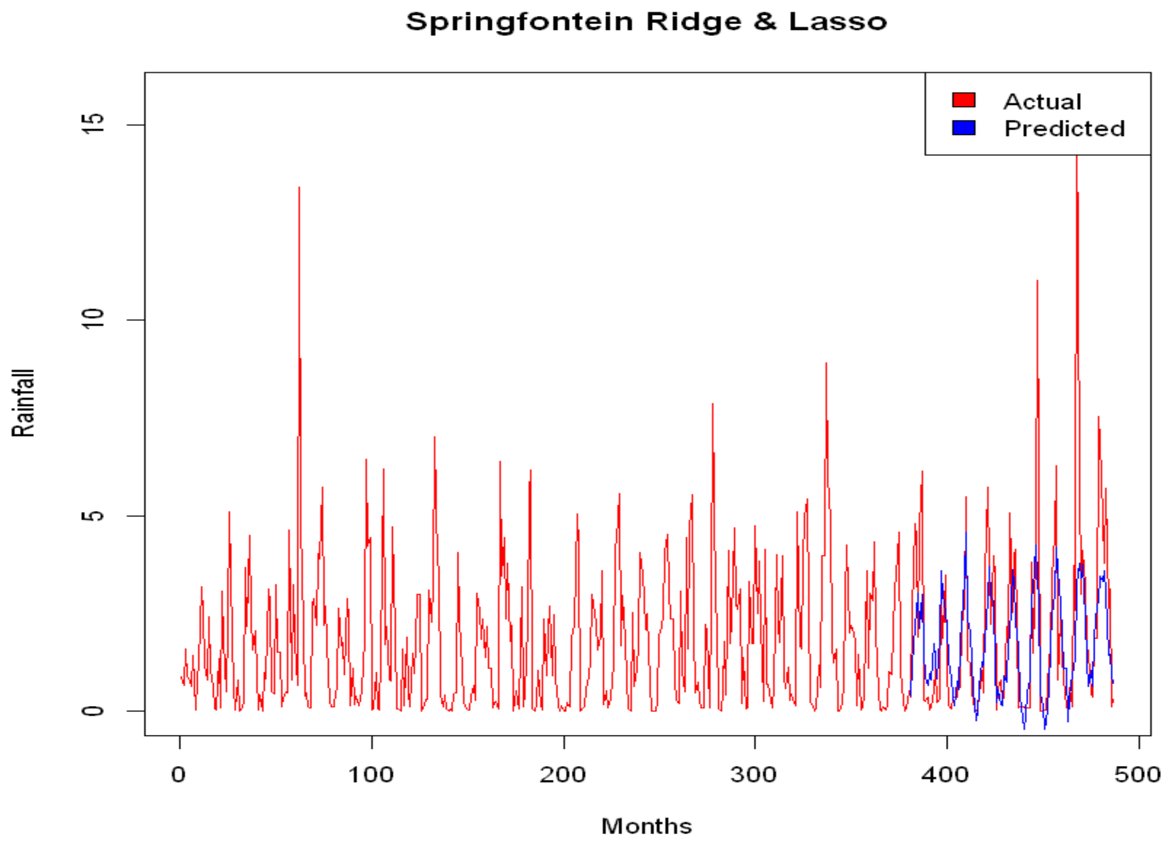
Linear Regression					
Locations	**MAE**	**MSE**	**RMSE**	**r**	**R-Square**
Bloemfontein	0.95	1.40	1.18	0.76	0.80
Springfontein	0.80	0.98	0.99	0.72	0.76
Welkom	1.01	1.70	1.30	0.76	0.82
Random Forest					
Bloemfontein	1.07	3.13	1.77	0.76	0.49
Springfontein	1.11	3.82	1.96	0.67	0.39
Welkom	1.06	2.53	1.59	0.74	0.51
Support Vector Machine					
Bloemfontein	1.10	3.48	1.87	0.79	0.44
Springfontein	1.07	3.69	1.92	0.74	0.41

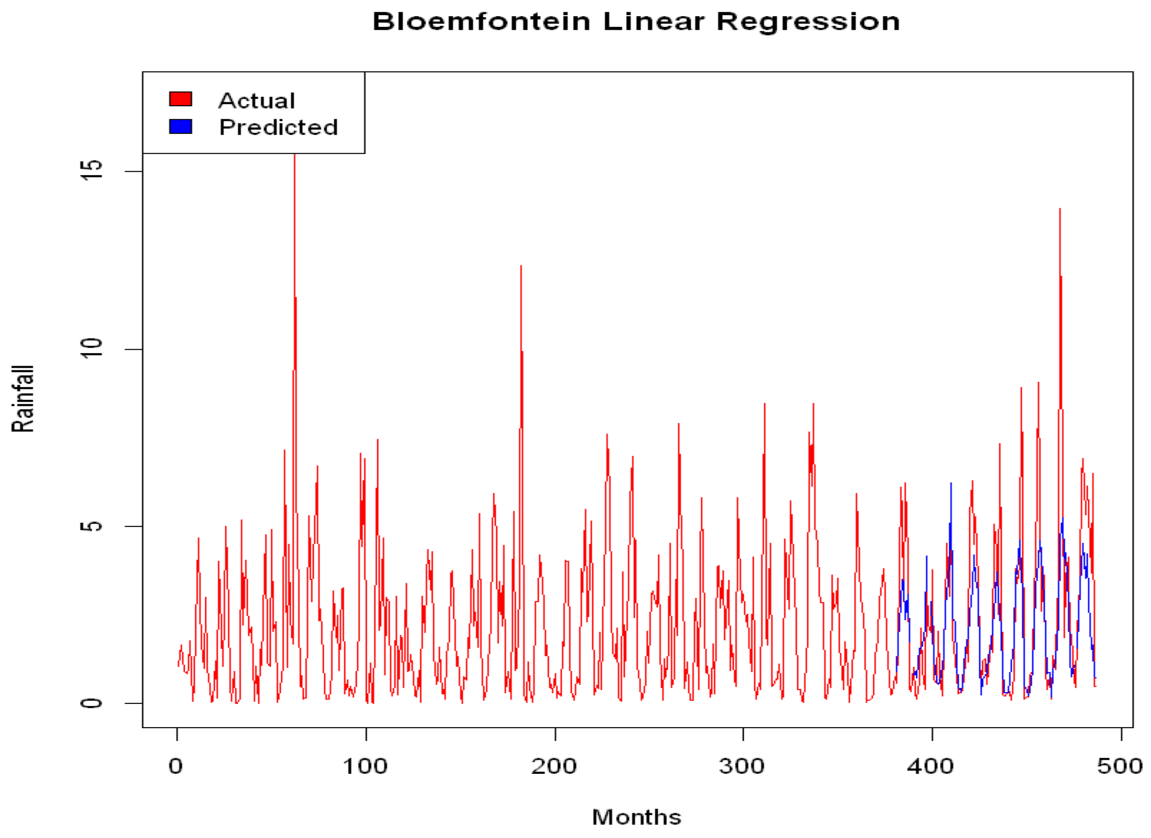
Welkom	1.06	2.69	1.64	0.80	0.48
Ridge and Lasso					
Bloemfontein	1.24	3.63	1.90	0.75	0.69
Springfontein	1.23	4.09	2.02	0.69	0.69
Welkom	1.12	2.64	1.62	0.79	0.67

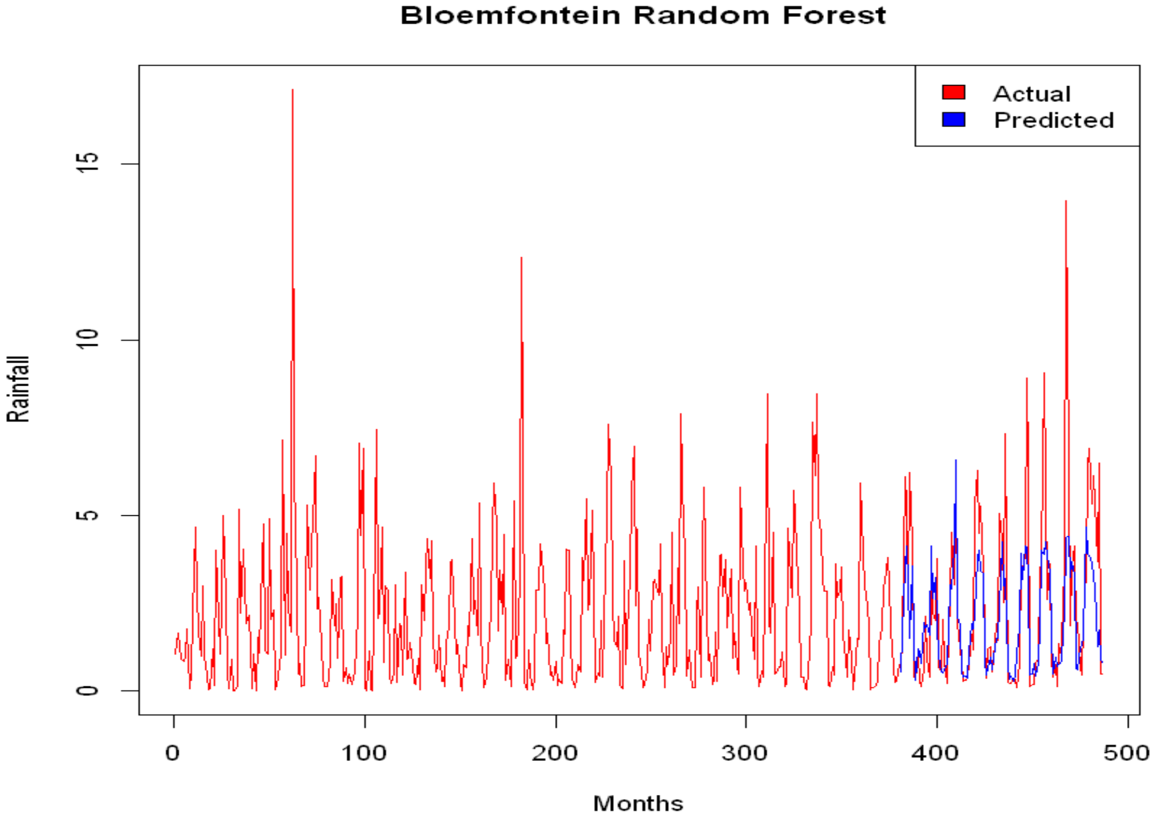


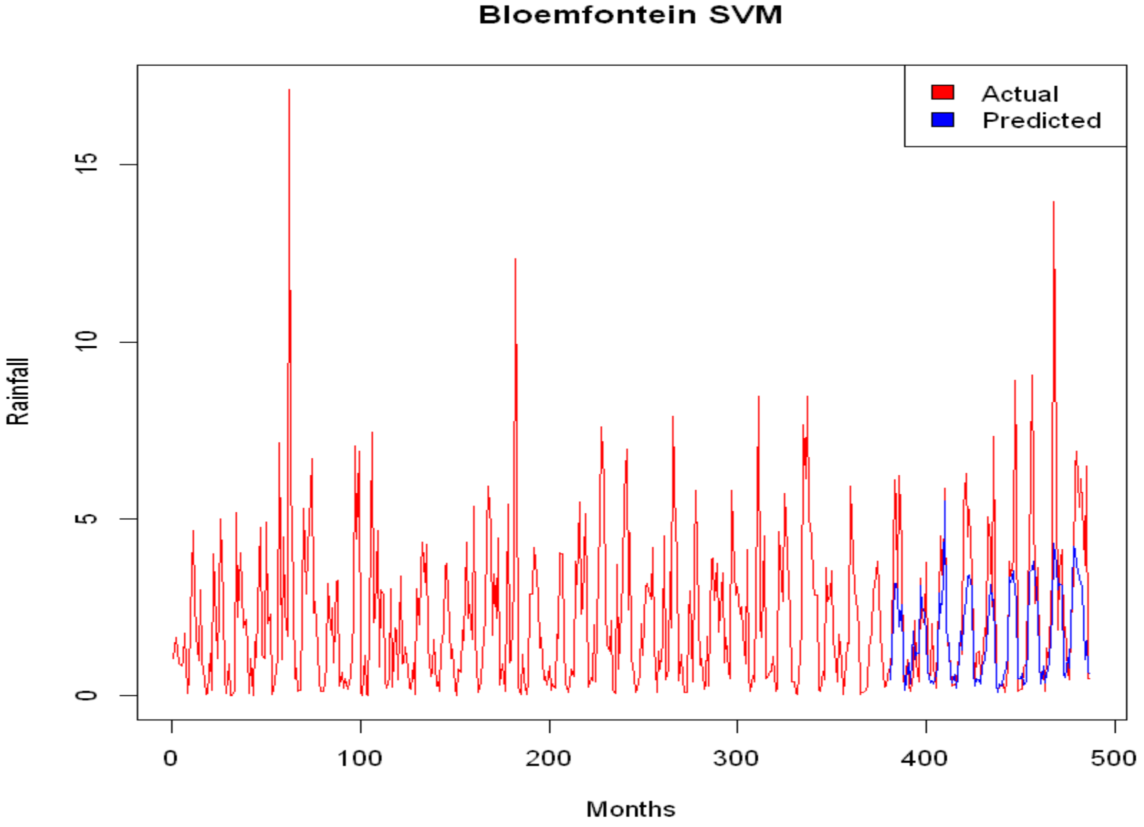

Springfontein Random Forest
🟥 Actual
🟦 Predicted
Rainfall
15
10
5
0
0
100
200
300
400
500
Months

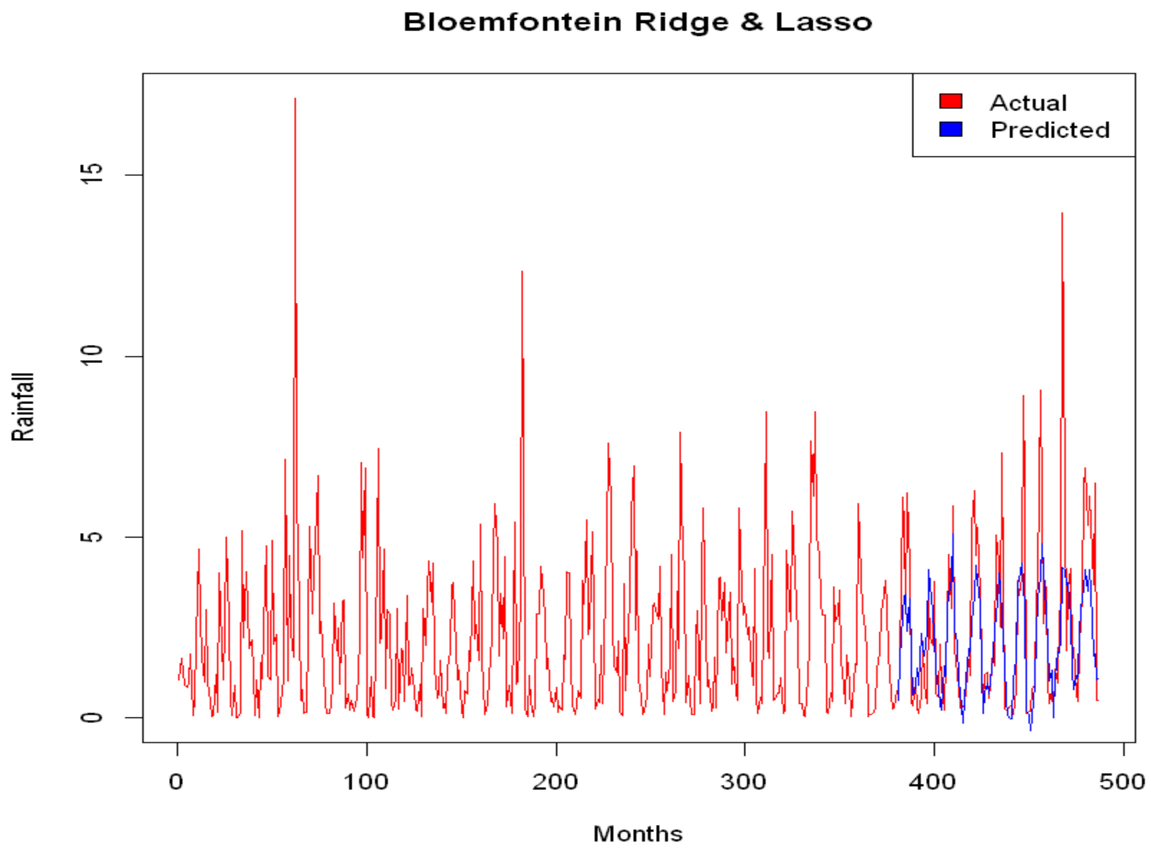

Springfontein SVM
🟥 Actual
🟦 Predicted
Rainfall
15
10
5
0
0
100
200
300
400
500
Months

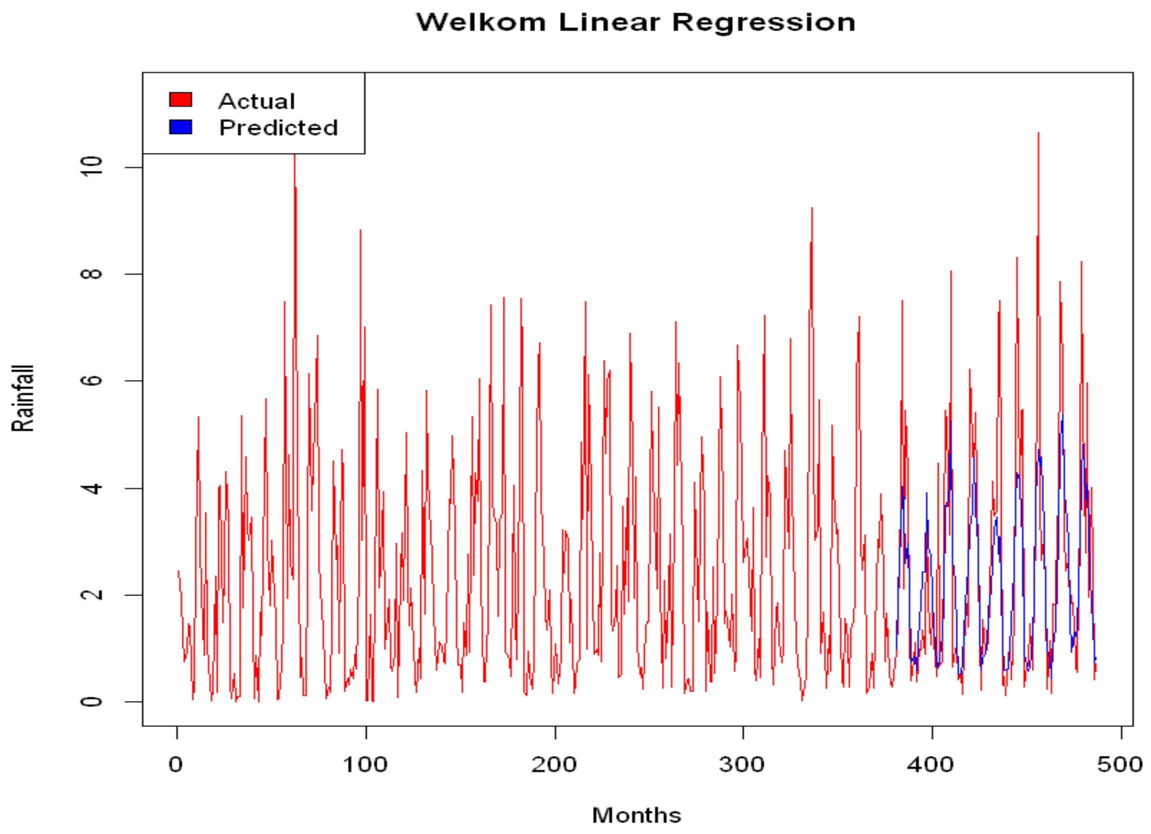

Springfontein Ridge & Lasso
🟥 Actual
🟦 Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months

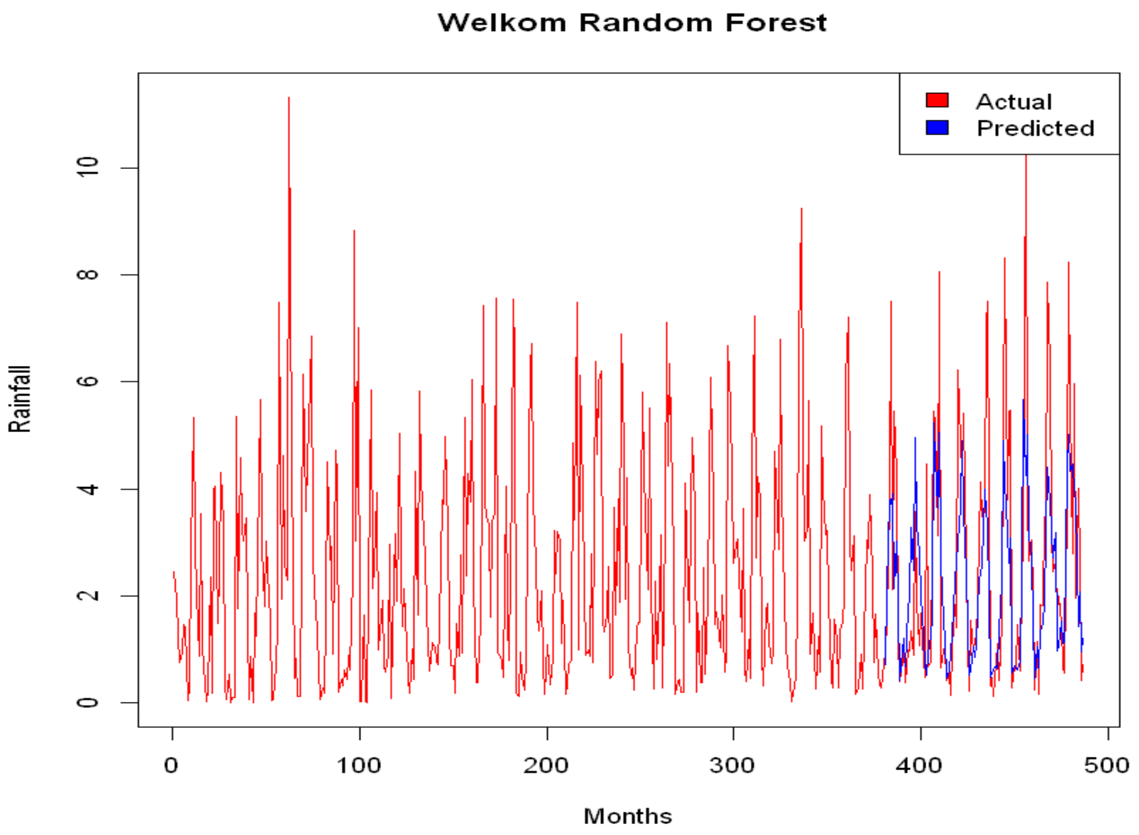

Bloemfontein Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months

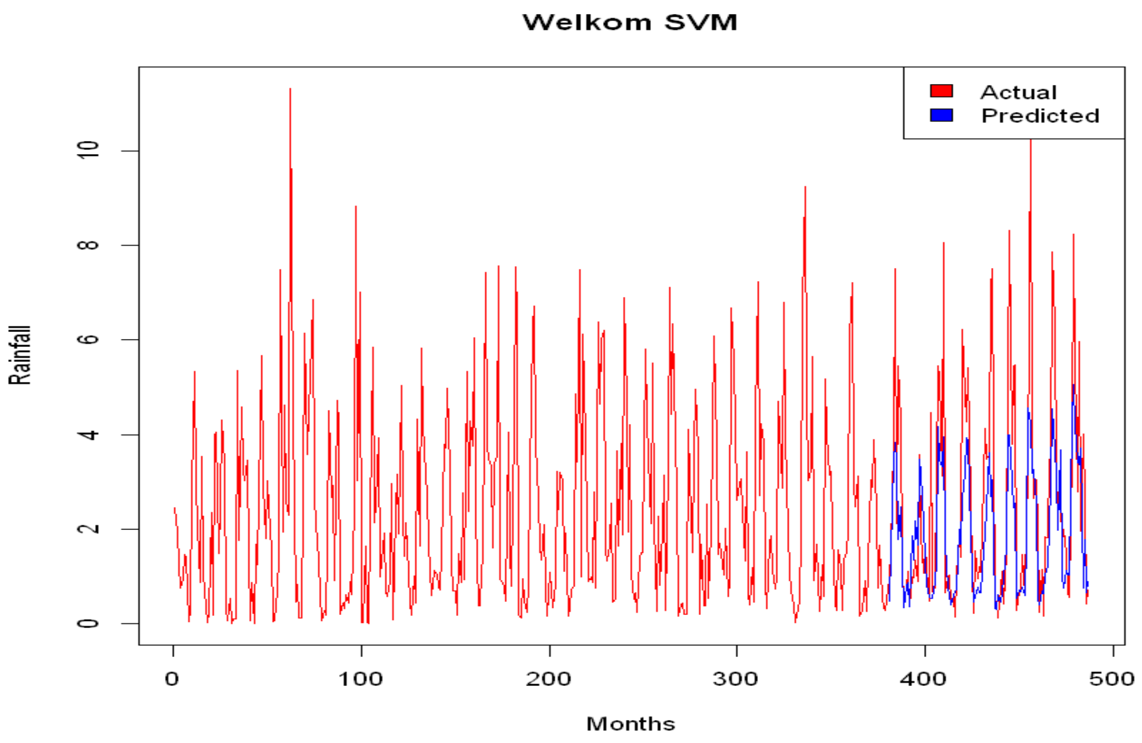

Bloemfontein Random Forest
Actual
Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months


Bloemfontein SVM
Actual
Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months


Bloemfontein Ridge & Lasso
Actual
Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months


Welkom Linear Regression
Actual
Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months


Welkom Random Forest
Actual
Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months


Welkom SVM
Actual
Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months

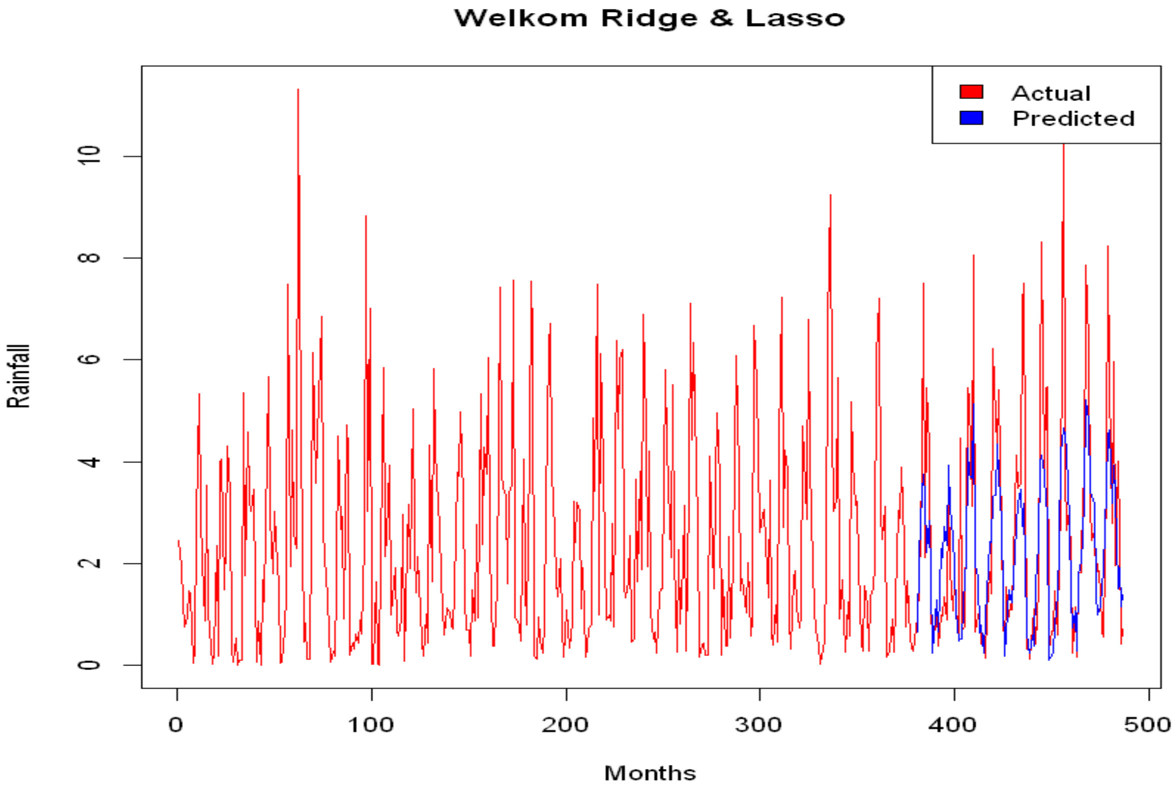


Figure 4. 1: Figures showing the predicted and actual rainfall for models used and locations under the cold and semi-arid steppe climate classification.

4.1.2 Cold Arid Desert

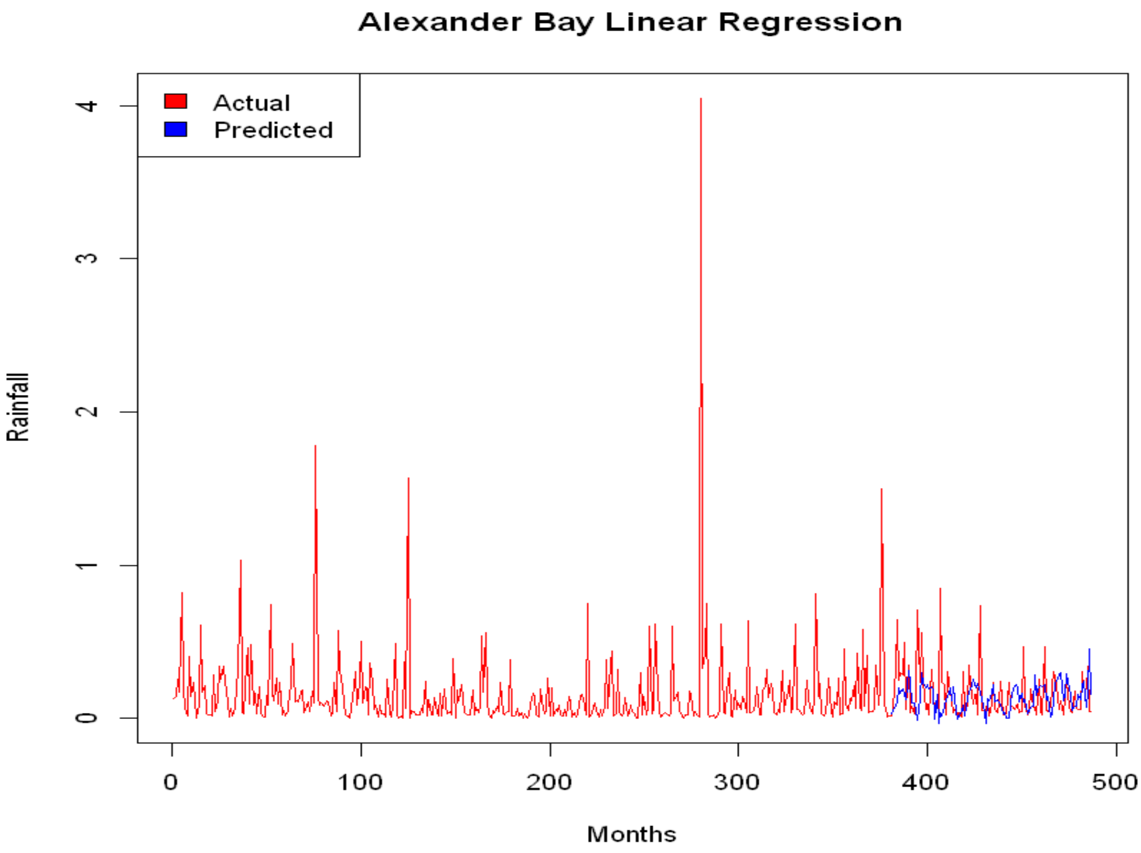
For the cold arid desert, the support vector machine had the highest correlation coefficient of 0.72 in Bristown closely followed by random forest model with a coefficient of 0.71 while those of linear regression, ridge and lasso were 0.66 and 0.65 respectively as shown in table 4.2 Similarly, the support vector machine model had higher correlation coefficient compared to other models for Beaufort West corresponding to 0.65 closely followed by ridge and lasso and the linear regression corresponding to 0.64 and 0.60 respectively. However, for Alexander Bay, none of the models had a correlation coefficient higher than 0.15 nor a coefficient of determination higher than 0.14. The reason for this is yet to be determined and would be further examined. Perhaps, it being officially the driest town in South Africa may contribute partly to this (South African yearbook, 2012). The models showed pattern of poor prediction in regions with low rainfall (Sungkawa & Rahuyu, 2019). Alexander Bay receives an annual rainfall of less than 51mm with the cold Benguela current influencing its climate. The mean absolute error, mean square error, and the root mean square error for random forest, support vector machine, ridge and lasso were all similar for all locations, however, the values were higher for linear regression in all locations. For ridge and lasso, Beaufort west had a coefficient of determination of 0.86 and correlation coefficient of 0.64. This is great as trend prediction and analysis using ridge and lasso can help farmers mitigate the negative impact of droughts on their sheep farming and aid proper planning (Parker, 2020). Notable parts of South Africa have been declared as disaster drought area (National Drought Task Team, 2015). The resultant effect of this is the increased mortality rate in livestock due to unavailability of water, loss of diary and livestock production, as well as disruption in the reproduction cycle in animals. It also increases unemployment as industries relying on agricultural produce such as fertilizer manufacturers are lost, food prices increase due to damage to crop quality and reduced food production. The impact of drought is not limited to its economic impacts, but it also has environmental impacts as it leads to degradation of animal habitats, inferior crops, decreased water quality and insufficient drinking water, soil erosion and fire outbreak (Parker, 2020).

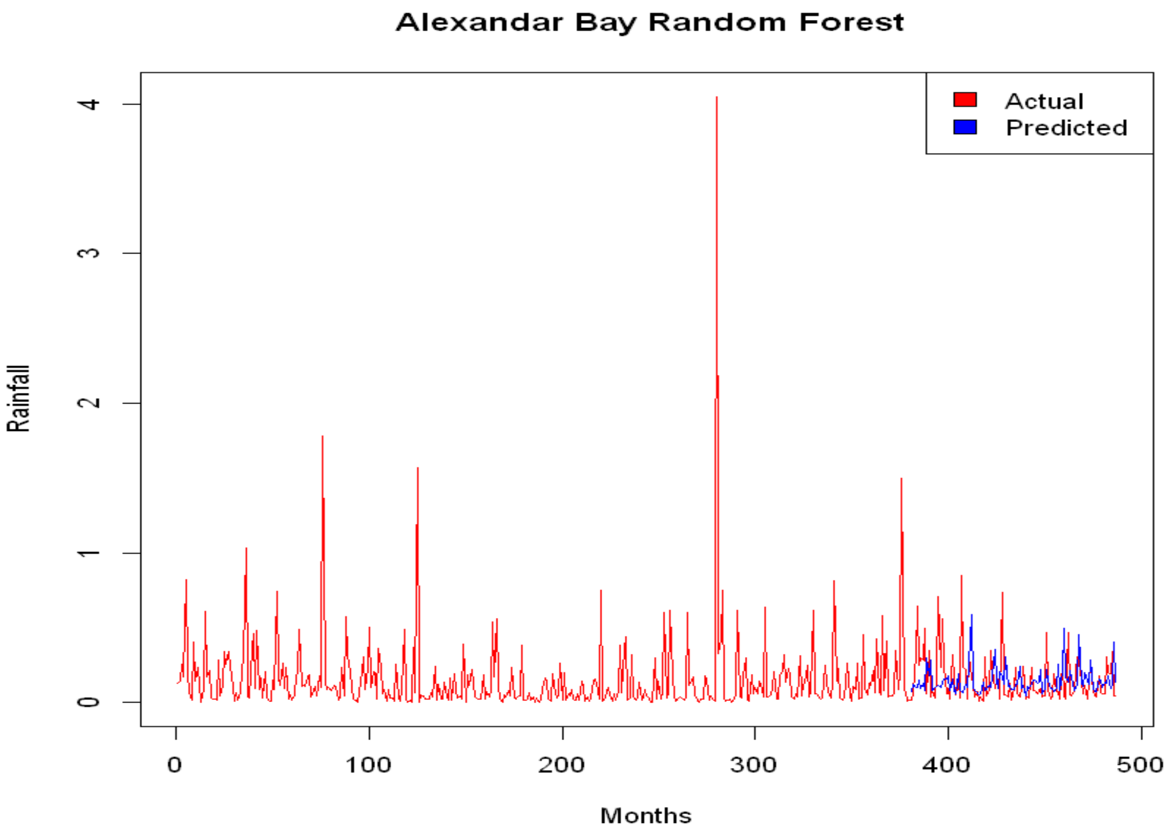
Figure 4.2 shows the monthly variation in rainfall for Alexander Bay, Beaufort West, and Bristown using different models. As expected, Alexander Bay received least amount of rainfall with a monthly average below 1mm. The models were able to accurately predict the seasonal variation in rainfall for all locations and all models. Random forest performed better in estimating the amount of rainfall received for both Beaufort West and Bristown. While evaluating these regions, Linear regression's estimate for rainfall in Bristown was better. Deman et al (2022) revealed the challenge in reliable long-term rainfall prediction due to the area susceptibility to droughts and floods. They stated the need to use ridge and lasso regression models as they outperformed dynamical climate prediction models. They however cautioned on selecting independently the predictors for both train and test data as they may result in biased results which do not correctly reflect the accuracy of the models.

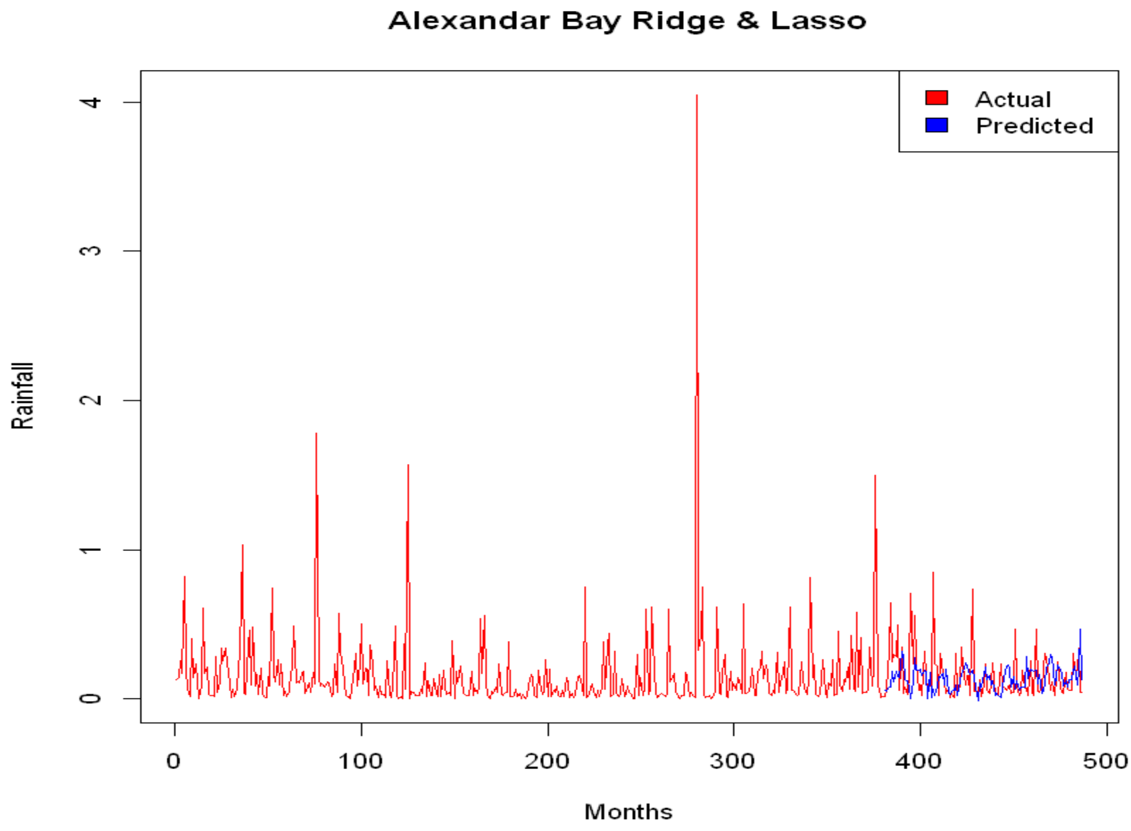
The heatmap in the appendix A showed that no atmospheric parameter correlated with rainfall in Alexander Bay. This may be expected as there is little or no rainfall in this city. The atmospheric variables used for this study had correlation coefficients of 0.084, 0.018, -0.078, 0.20, 0.11, and -0.096 for dewpoint, relative humidity, temperature, cloud cover, water vapour, and wind speed respectively. Therefore, to make any prediction, historical rainfall datasets would be the best datasets to be used. However, other atmospheric variables had high correlation between them, especially with dewpoint. Dewpoint had a correlation of 0.91, 0.92, and 0.76 with relative humidity, water vapour, and temperature. For Beaufort West, only dew point and water vapour had a correlation above 0.50 with rainfall. Dew point corresponded to 0.55 while water vapour corresponded to 0.57. For other atmospheric variables, their correlation with rainfall corresponds to 0.045, 0.27, 0.37, and -0.38 for relative humidity, temperature, cloud cover, and wind speed. Similar to locations under the cold and semi-arid steppe climate, rainfall in Bristown had the best correlation coefficient corresponding to 0.65, 0.66, 0.64 for dew point, cloud cover, and water vapour respectively. Its correlation with relative humidity, temperature, and wind speed are 0.22, 0.29, and -0.35. This shows that the most important variables to consider for rainfall prediction are dewpoint, water vapour, and cloud cover in the cold arid desert.

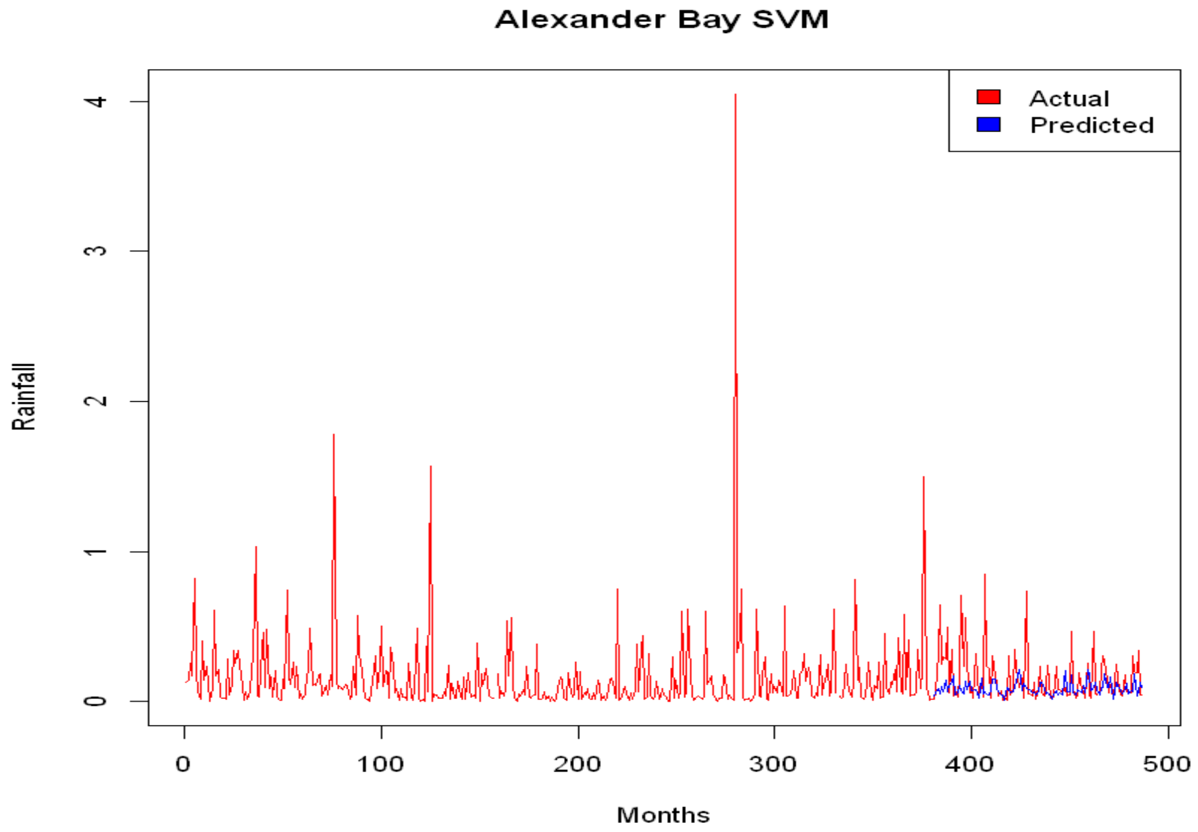
Table 4. 2: Table showing models evaluation metrics for cold arid desert climate classification (Alexander Bay, Beaufort West, and Bristown)

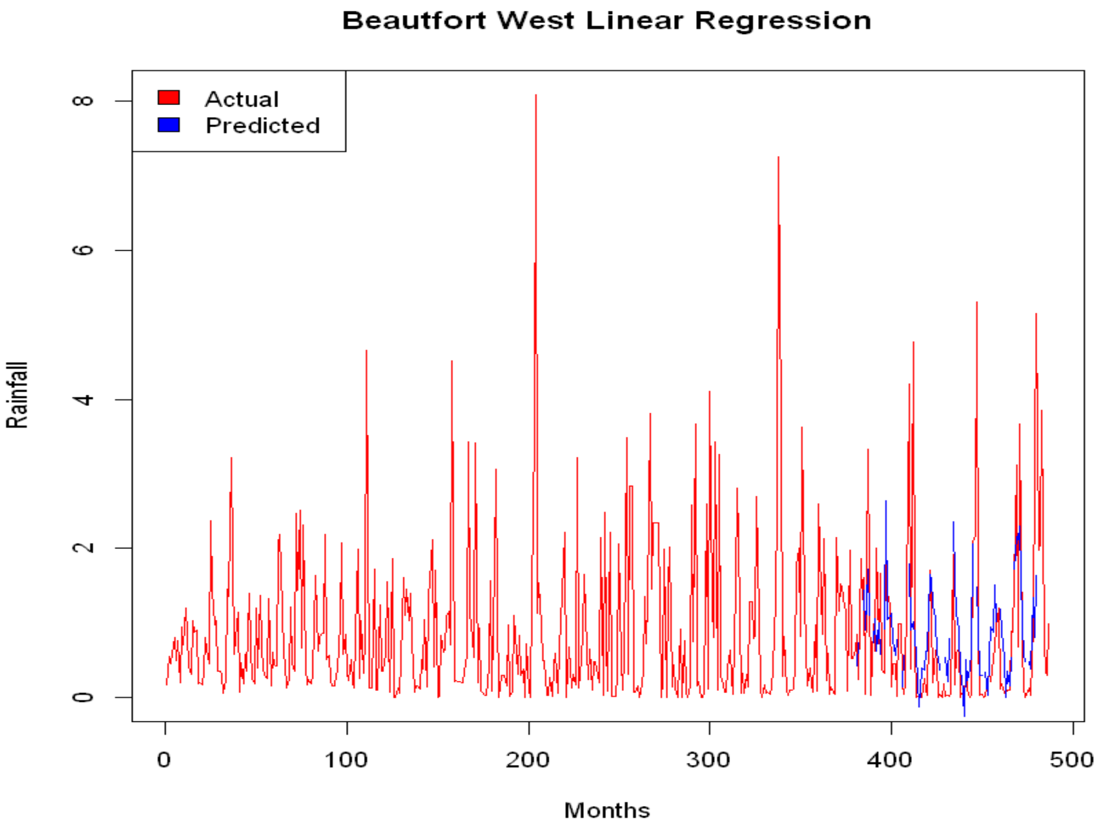
Linear Regression					
Locations	**MAE**	**MSE**	**RMSE**	**r**	**R-Square**
Alexander Bay	1.50	3.75	1.94	0.14	0.12
Beaufort West	0.69	0.78	0.88	0.60	0.14
Bristown	0.77	0.94	0.97	0.66	0.65
Random Forest					
Alexander Bay	0.13	0.04	0.19	0.01	-0.31
Beautfort West	0.62	0.97	0.99	0.55	0.30
Bristown	0.67	1.08	1.04	0.71	0.44
Support Vector Machine					
Alexander Bay	0.11	0.03	0.18	0.15	-0.22
Beaufort West	0.55	0.94	0.97	0.65	0.32
Bristown	0.65	1.23	1.11	0.72	0.36
Ridge and Lasso					
Alexander Bay	0.12	0.03	0.18	0.06	0.14
Beaufort West	0.62	0.92	0.96	0.64	0.86
Bristown	0.73	1.30	1.14	0.65	0.74

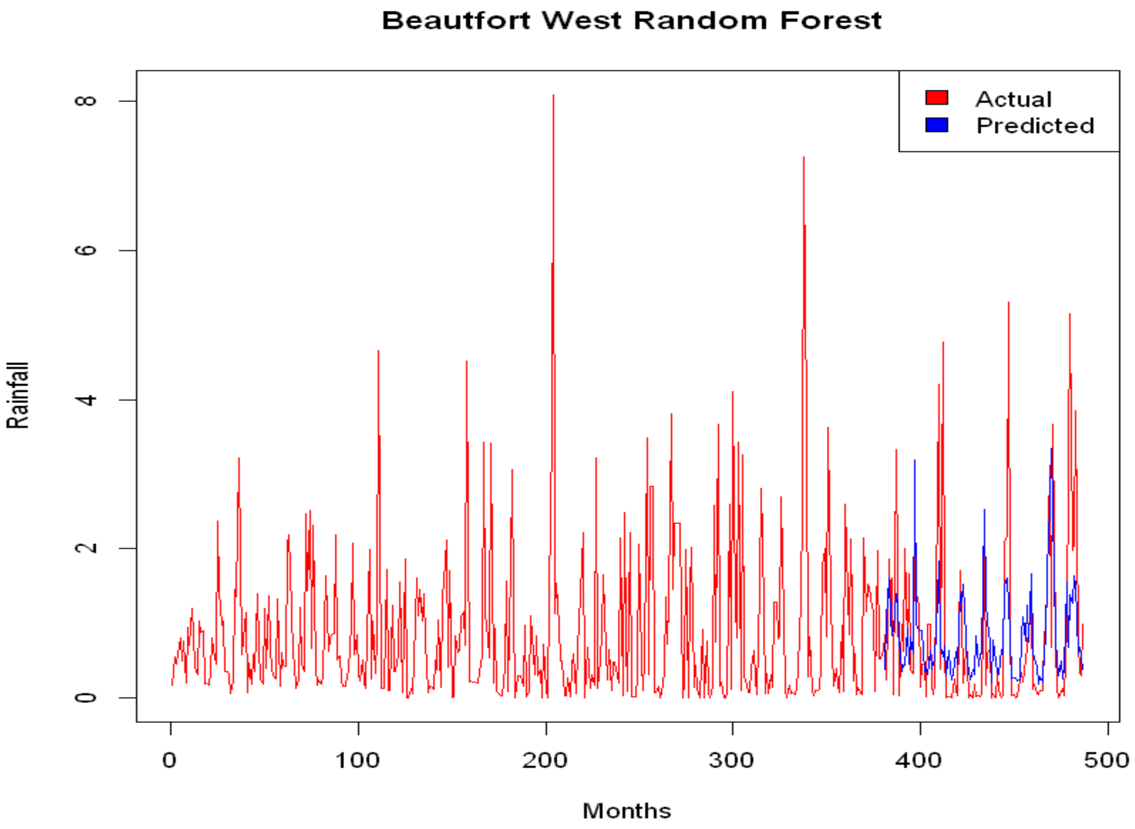

Alexander Bay Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
1
2
3
4
0
100
200
300
400
500
Months

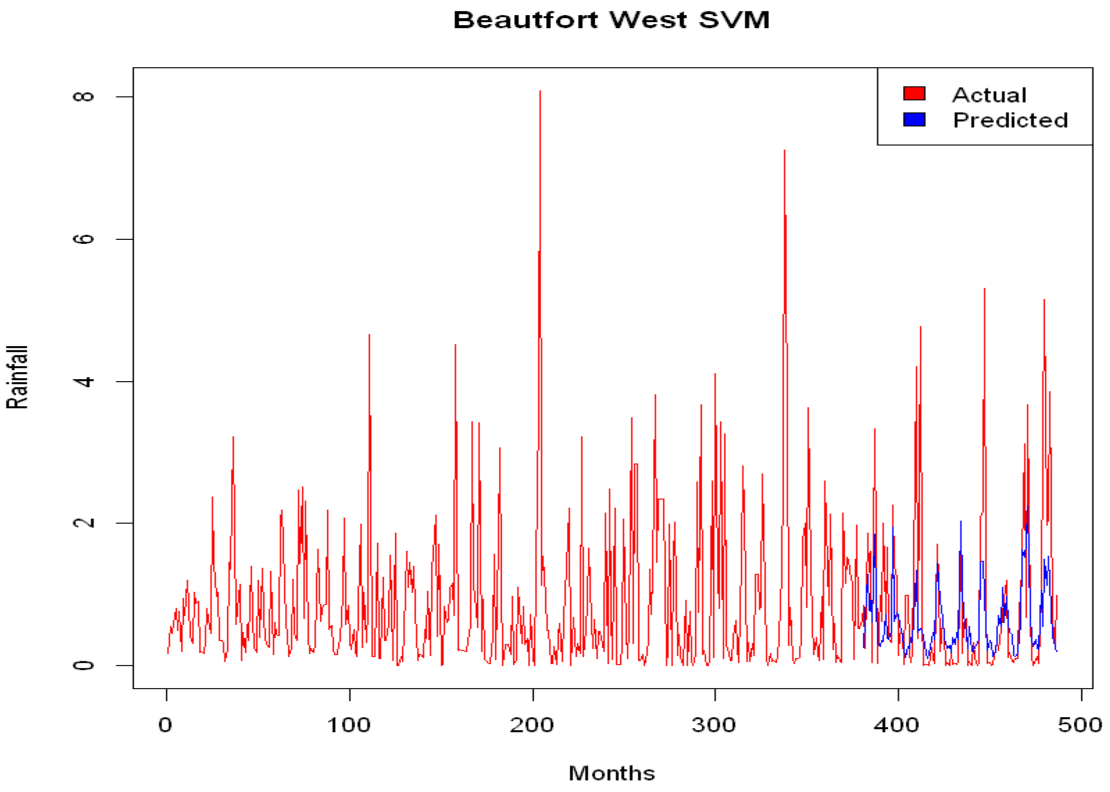

Alexandar Bay Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
1
2
3
4
0
100
200
300
400
500
Months

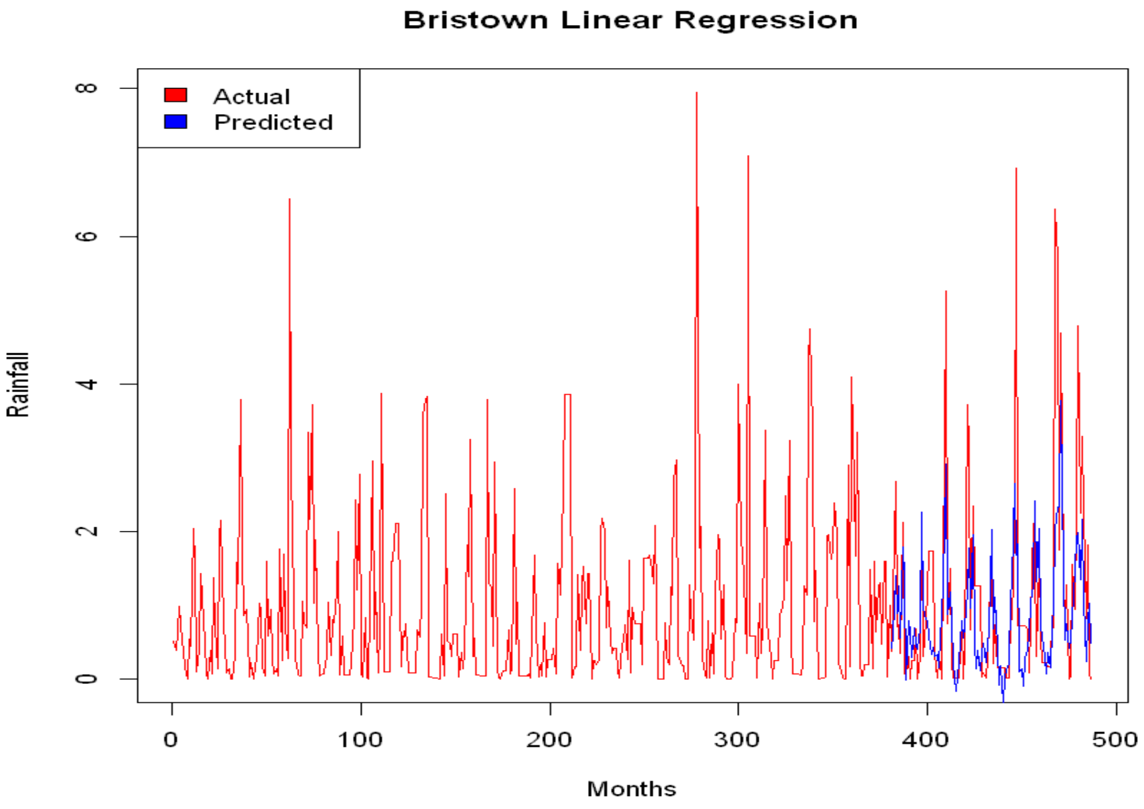

Alexandar Bay Ridge & Lasso
Actual
Predicted
Rainfall
0
1
2
3
4
0
100
200
300
400
500
Months

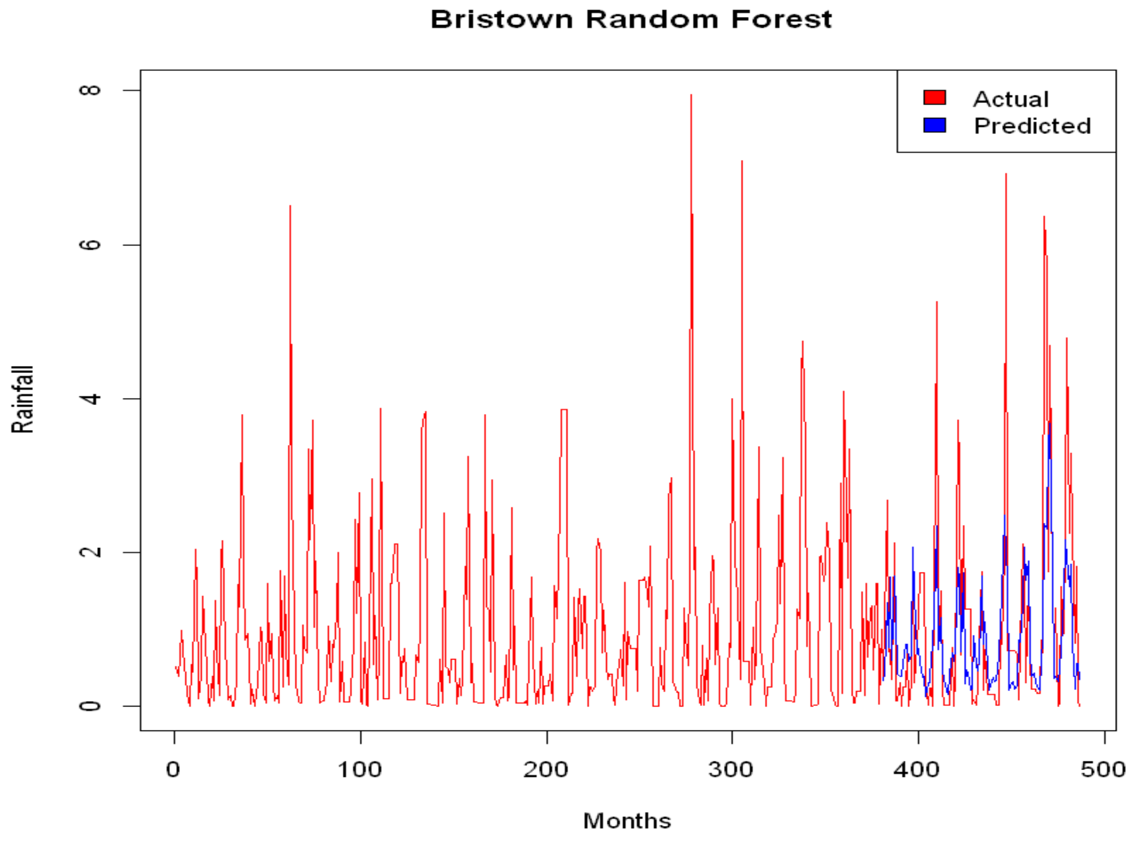

Alexander Bay SVM
Actual
Predicted
Rainfall
0
1
2
3
4
0
100
200
300
400
500
Months

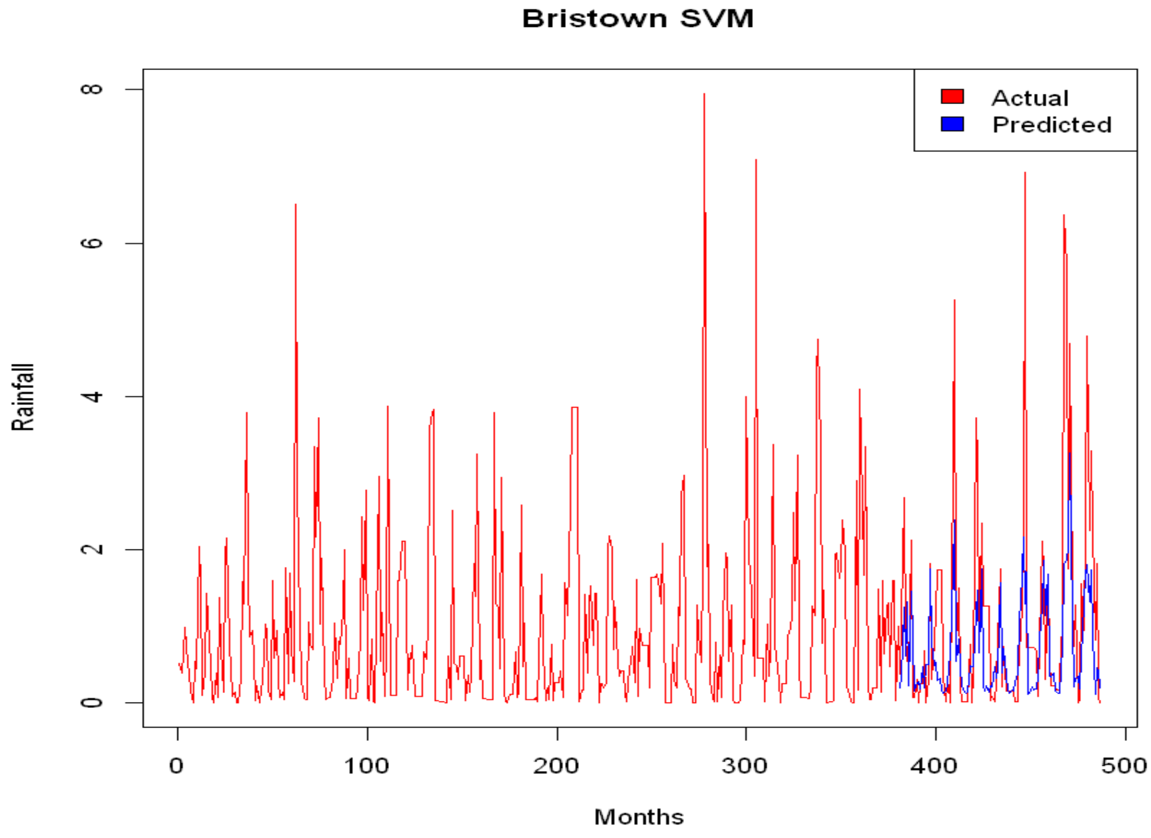

Beautfort West Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
0
100
200
300
400
500
Months


Beautfort West Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
0
100
200
300
400
500
Months


Beautfort West SVM
Actual
Predicted
Rainfall
0
2
4
6
8
0
100
200
300
400
500
Months


Bristown Linear Regression
Actual
Predicted
Rainfall
0
2
4
6
8
0
100
200
300
400
500
Months


Bristown Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
0
100
200
300
400
500
Months


Bristown SVM
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
0
100
200
300
400
500
Months

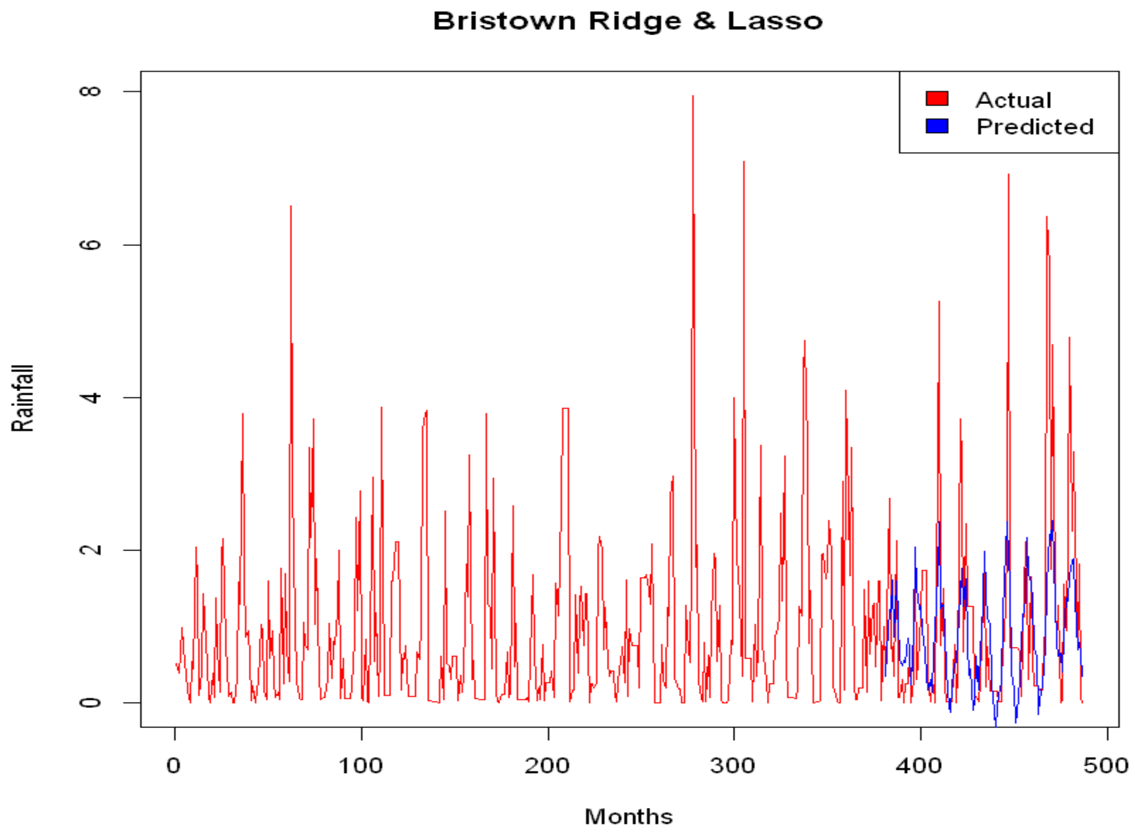


Figure 4. 2: Figures showing the predicted and actual rainfall for models used and locations under the cold arid desert climate classification.

4.1.3 Hot and Semi-Arid Steppe

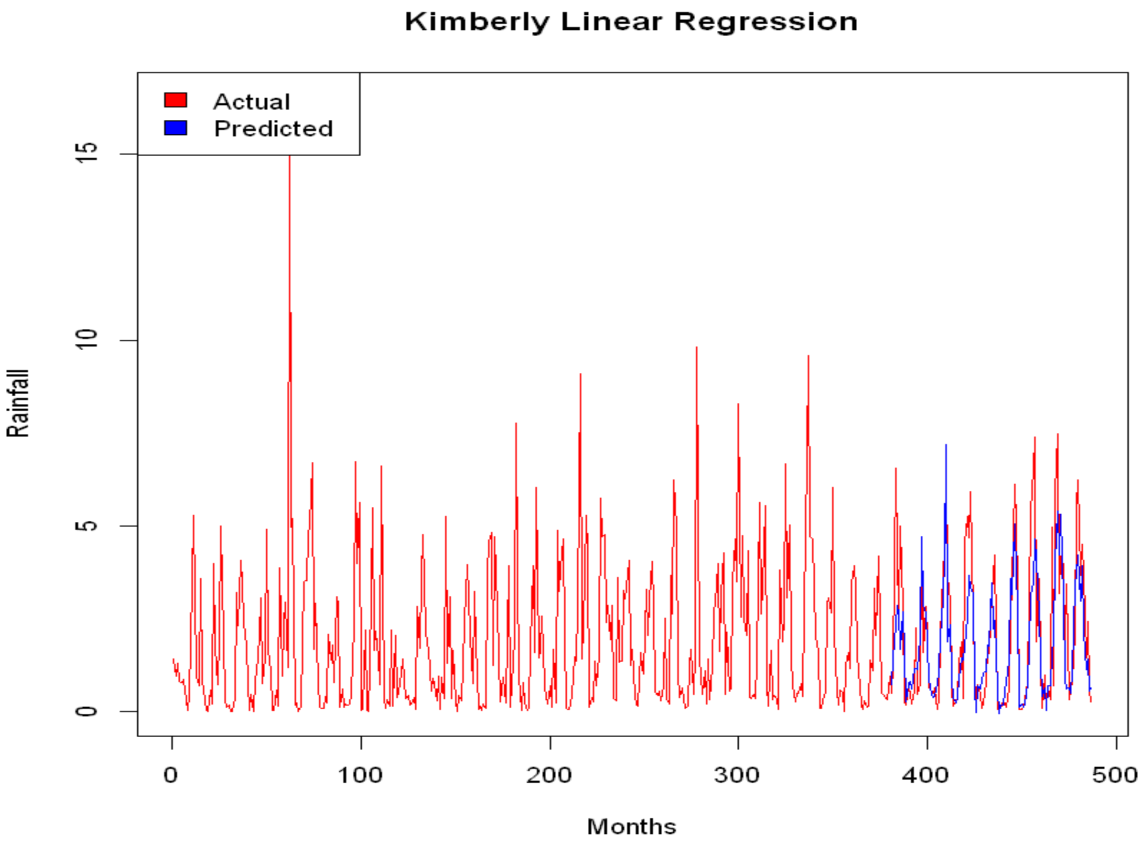
Table 4.3 presents the evaluation metrics for the four models used in this study for Kimberly, Mahikeng, and Port Elizabeth, under the hot and semi-arid steppe climate classification. These cities are situated in three different provinces. Kimberly in Northern Cape Province, Mahikeng in North-West Province, and Port Elizabeth in the Eastern Cape Province of South Africa. The result revealed that linear regression was best for rainfall prediction in Kimberly with a correlation coefficient of 0.85 and coefficient of determination of 0.82 as seen in table 4.3. These values were equally high for other models. For random forest, support vector machine, ridge and lasso, the correlation coefficients are 0.79, 0.84, and 0.82 respectively while the coefficients of determination for these models are 0.59,0.61, and 0.68. Also, high values were obtained for all models in Mahikeng. This shows that these two locations are suitable for machine learning applications in atmospheric sciences. Tladi et al (2023) carried out a correlation analysis using gradient boosting regression on the upper crocodile sub-basin which cuts across Mahikeng. Their results also showed that the atmospheric parameters in Mahikeng are suitable for rainfall prediction. They obtained coefficient of determination of about 0.80, mean square values ranging from 0.03 to 0.30 and mean average error ranging from 0.12 to 0.50. These values were lower than what was obtained in this study. This may be attributed to their study only focusing on groundwater levels. However, in Port Elizabeth, the correlation coefficients were 0.13, 0.29, 0.43, 0.30 for linear regression, random forest, support vector machine, and ridge and lasso respectively. The coefficient of determination for these models were also significantly low. Yakubu et al (2021) predicted precipitation in Port Elizabeth using two machine learning models – multiple linear regression and multilayered perceptron. They made use of five cloud properties, cloud optical thickness, cloud effective radius, cloud top temperature, cloud top pressure, and liquid water path for their model. Their result showed a correlation coefficient above 0.70. This may be responsible for the low values obtained in Port Elizabeth as only one cloud property was used in this study. In subsequent studies, other cloud properties can be used as parameters for rainfall prediction to determine if they will perform better than the selected atmospheric parameters.

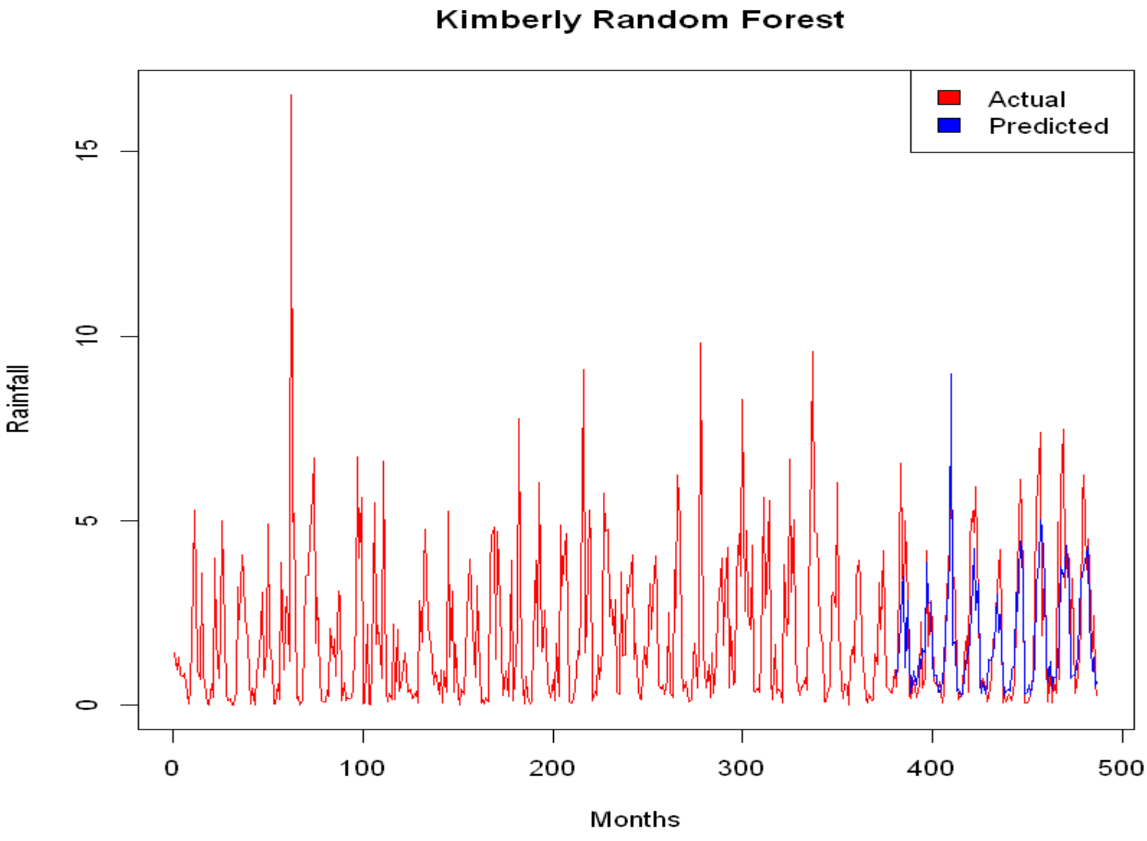
Figure 4.3 shows the temporal variation in rainfall for various locations using different models. The result also revealed consistent pattern in modelling the interannual variation in rainfall for all models. For Kimberly, random forest as well as ridge and lasso almost accurately estimated the amount of rainfall received. However, there was an overestimation of rainfall by random forest in 2016, perhaps the model was still understanding the datasets. Support vector machine had a better predictive performance in 2023 for Kimberly. Similar pattern is also observed in Mahikeng. However, there was an overestimation of rainfall in 2021 which none of the models could accurately predict. All the models estimated the seasonal variability well. Despite low evaluation metrics for Port Elizabeth, the models also performed well in predicting the seasonality of rainfall for the right years of prediction. It however underestimated rainfall in 2019 and 2023.

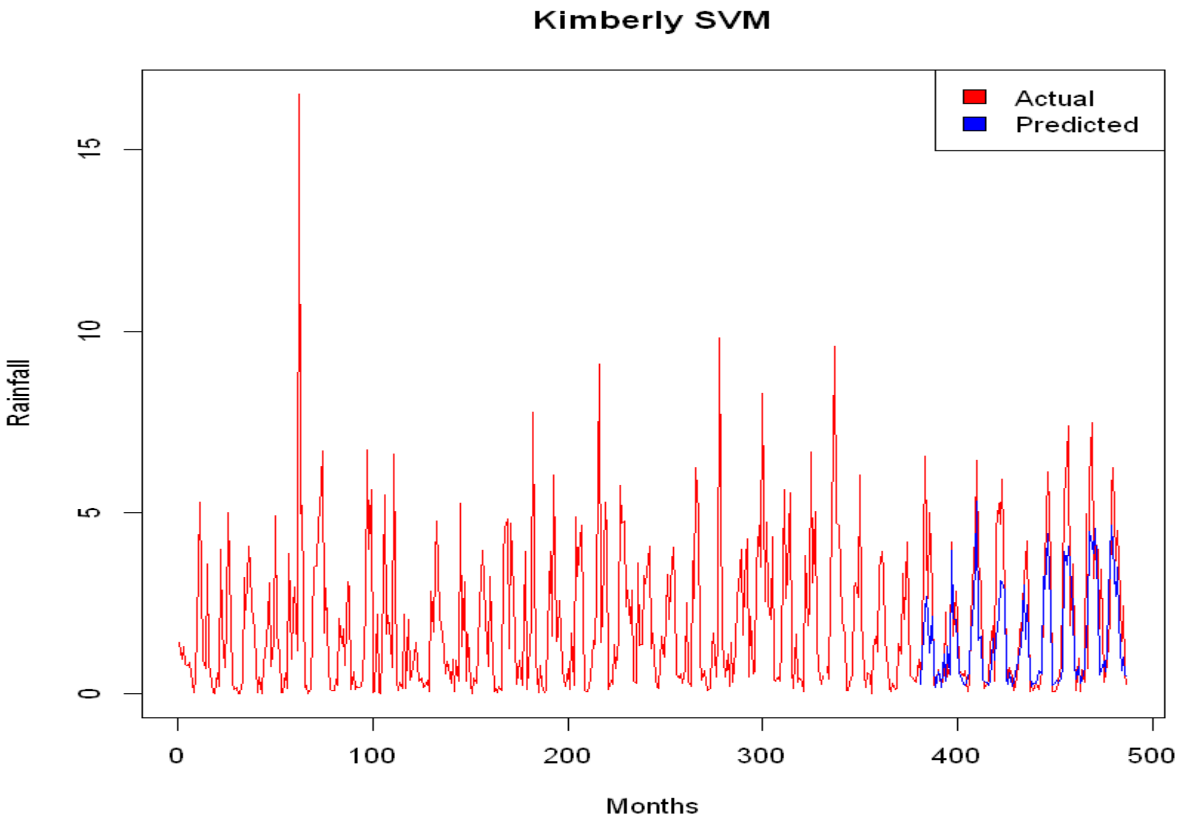
The heatmap for the correlation between rainfall and other atmospheric variables under the hot and semi-arid steppe is shown in appendix A. The result reveals that all atmospheric variables in Mahikeng had a correlation coefficient with rainfall exceeding 0.50 except wind speed which had a correlation of -0.18. Dewpoint, relative humidity, temperature, cloud cover, and water vapour correlation coefficients correspond to 0.76, 0.55, 0.56, 0.74, and 0.78 respectively. As with most cities under the arid climate classification, dewpoint, water vapour and cloud clover had the highest correlation with rainfall. In Kimberly, water vapour had the highest correlation with rainfall with a coefficient of 0.78 closely followed by cloud cover and dew point corresponding to 0.77 and 0.76 respectively, while temperature had a coefficient of 0.50. For relative humidity and wind speed, their correlation coefficients with rainfall were 0.36, and -0.33 respectively. Port Elizabeth showed similar pattern to what was obtained in Alexander Bay with all atmospheric variables having low correlation with rainfall. The result shows that dewpoint, relative humidity, temperature, cloud cover, water vapour, and wind speed correlated with rainfall with 0.17, 0.20, 0.06, 0.25, 0.21, 0.054 coefficients respectively. Also, just as it happened in Alexander Bay, relative humidity, temperature, and water vapour had high correlation with dewpoint corresponding to 0.91, 0.78, and 0.98 respectively.

Table 4. 3: Table showing models evaluation metrics for hot and semi-arid steppe climate classification.

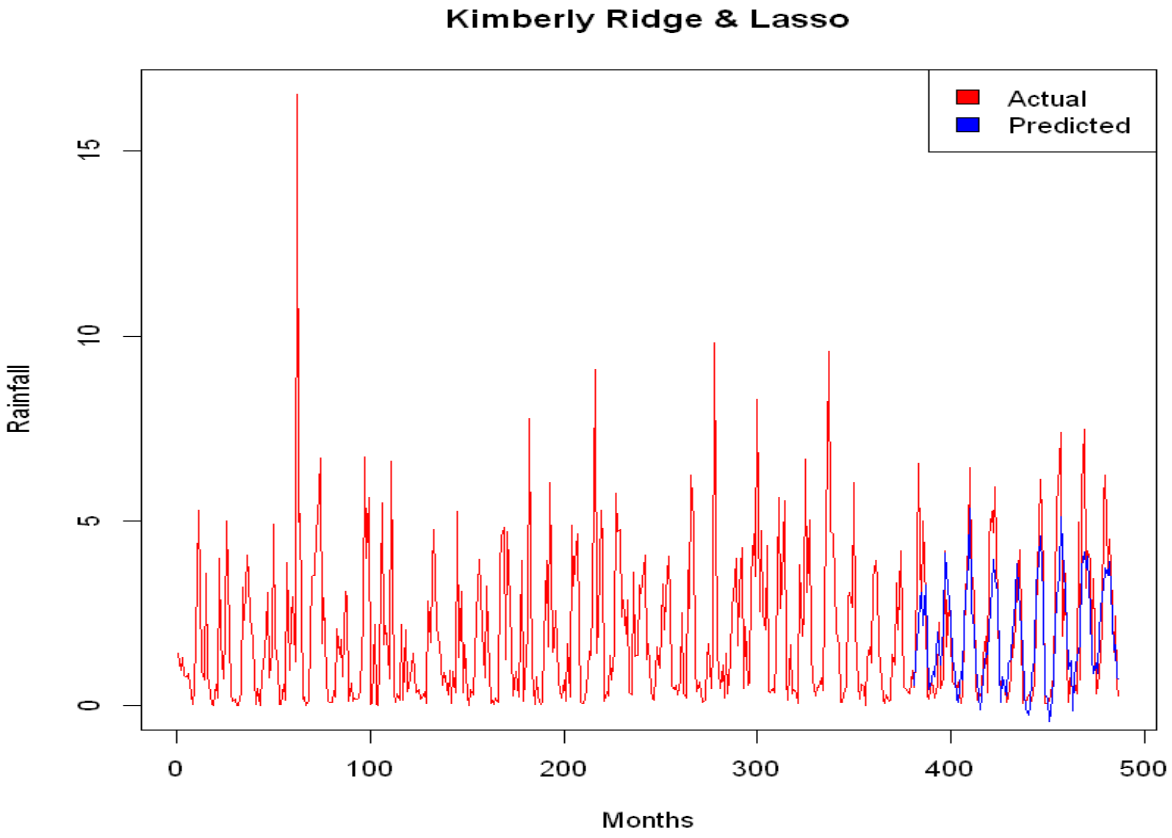
Linear Regression					
Locations	**MAE**	**MSE**	**RMSE**	**r**	**R-Square**
Kimberly	0.93	1.35	1.16	0.85	0.82
Mahikeng	1.09	1.85	1.36	0.75	0.84
Port Elizabeth	0.76	0.94	0.97	0.13	0.28
Random Forest					
Kimberly	0.89	1.75	1.32	0.79	0.59
Mahikeng	0.94	2.14	1.46	0.78	0.60
Port Elizabeth	0.82	1.17	1.08	0.29	0.01
Support Vector Machine					
Kimberly	0.86	1.65	1.29	0.84	0.61
Mahikeng	0.91	2.22	1.49	0.78	0.59
Port Elizabeth	0.78	1.19	1.09	0.43	-0.01
Ridge and Lasso					
Kimberly	0.97	1.83	1.35	0.82	0.68
Mahikeng	1.06	2.55	1.60	0.75	0.67
Port Elizabeth	0.78	1.10	1.05	0.30	0.47


Kimberly Linear Regression
Actual
Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months

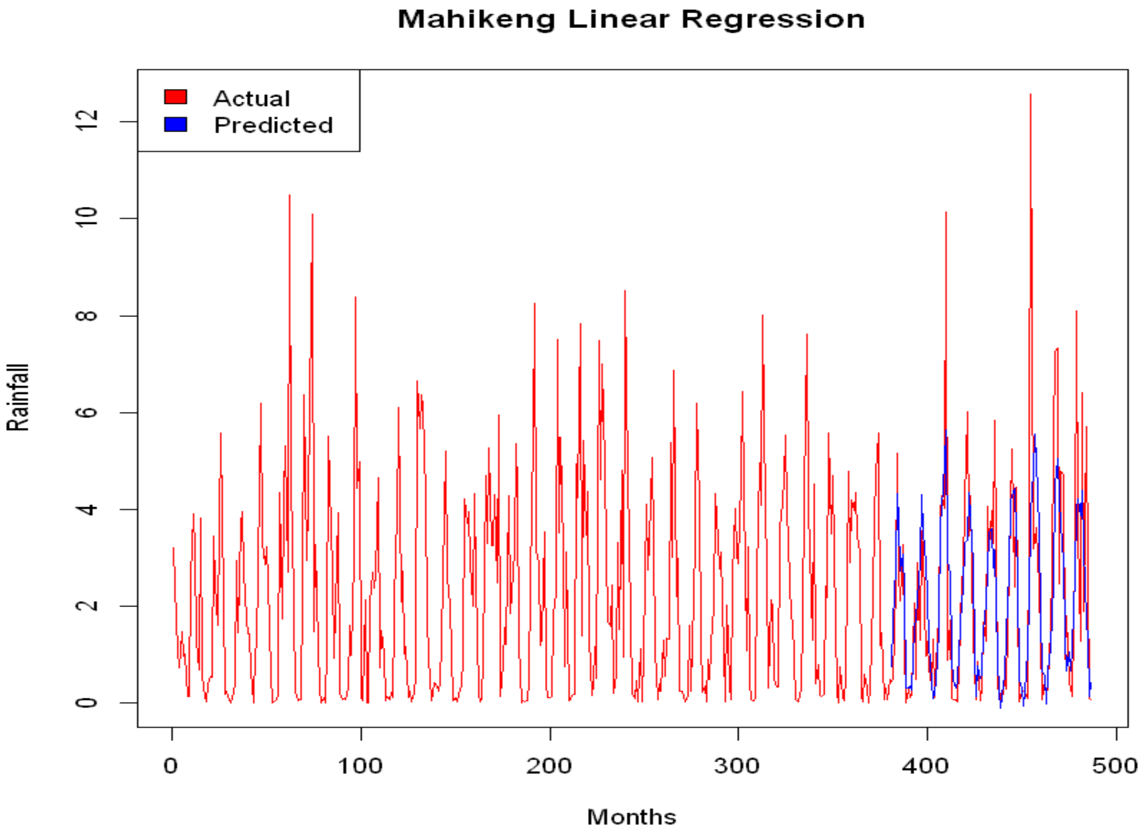

Kimberly Random Forest
Actual
Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months

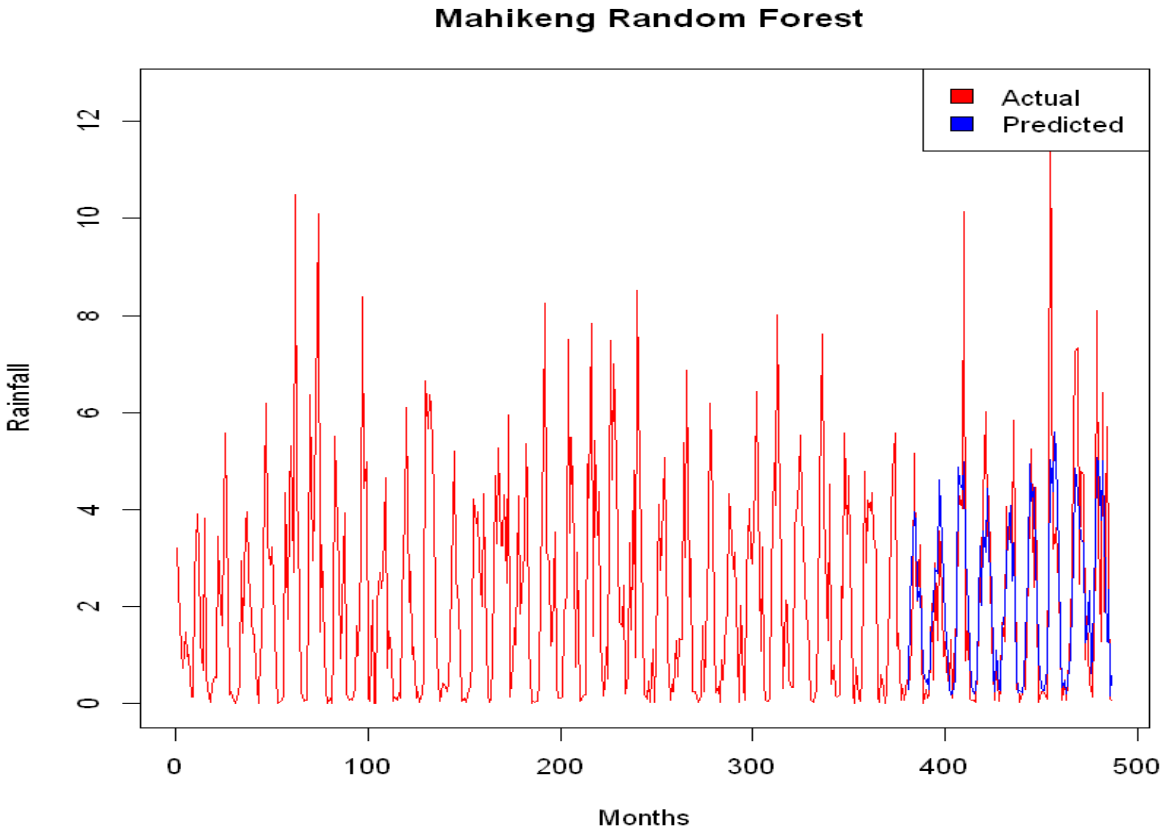


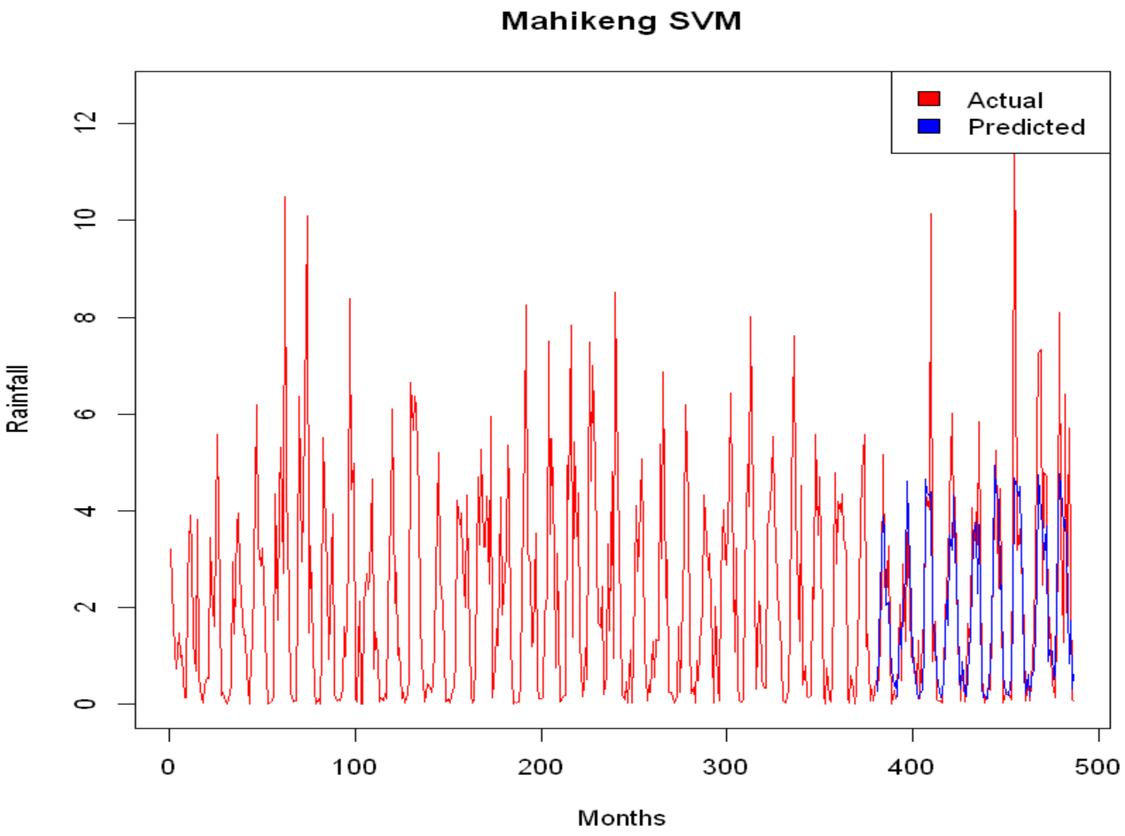
Kimberly SVM

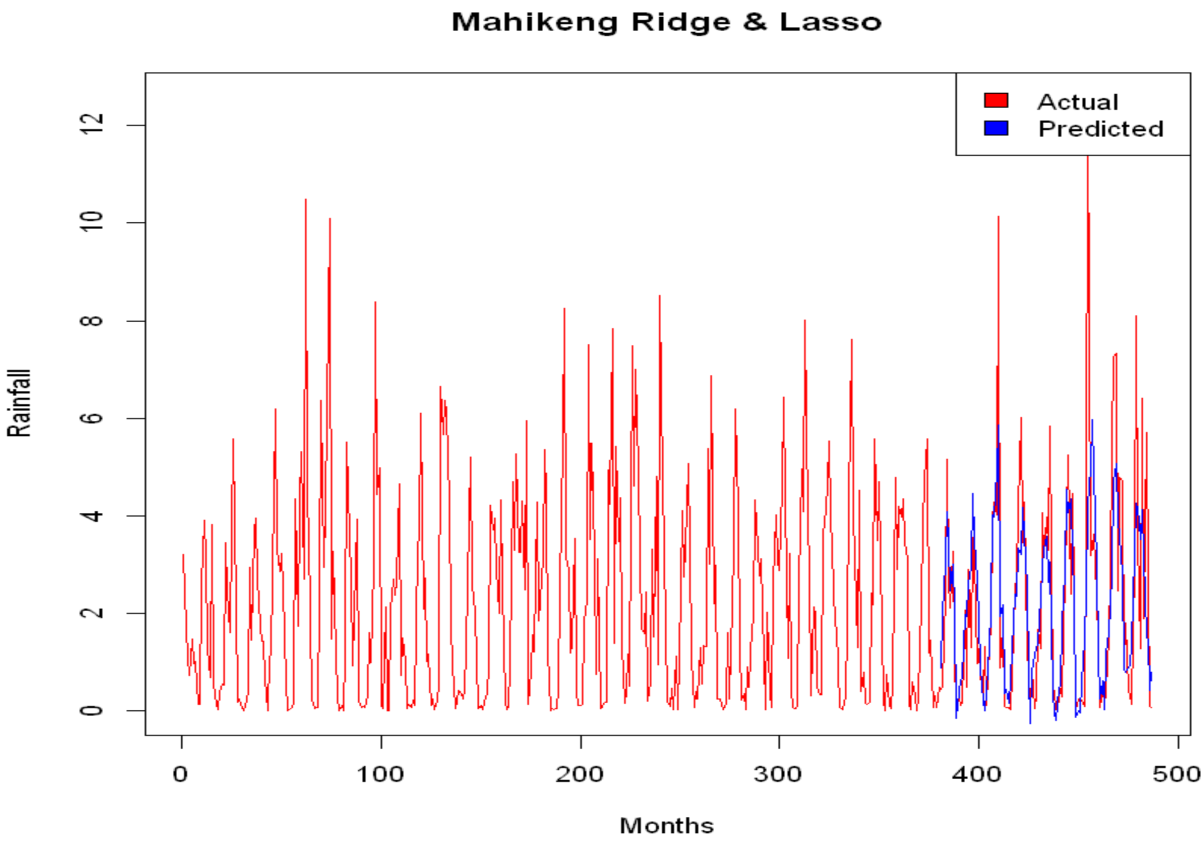


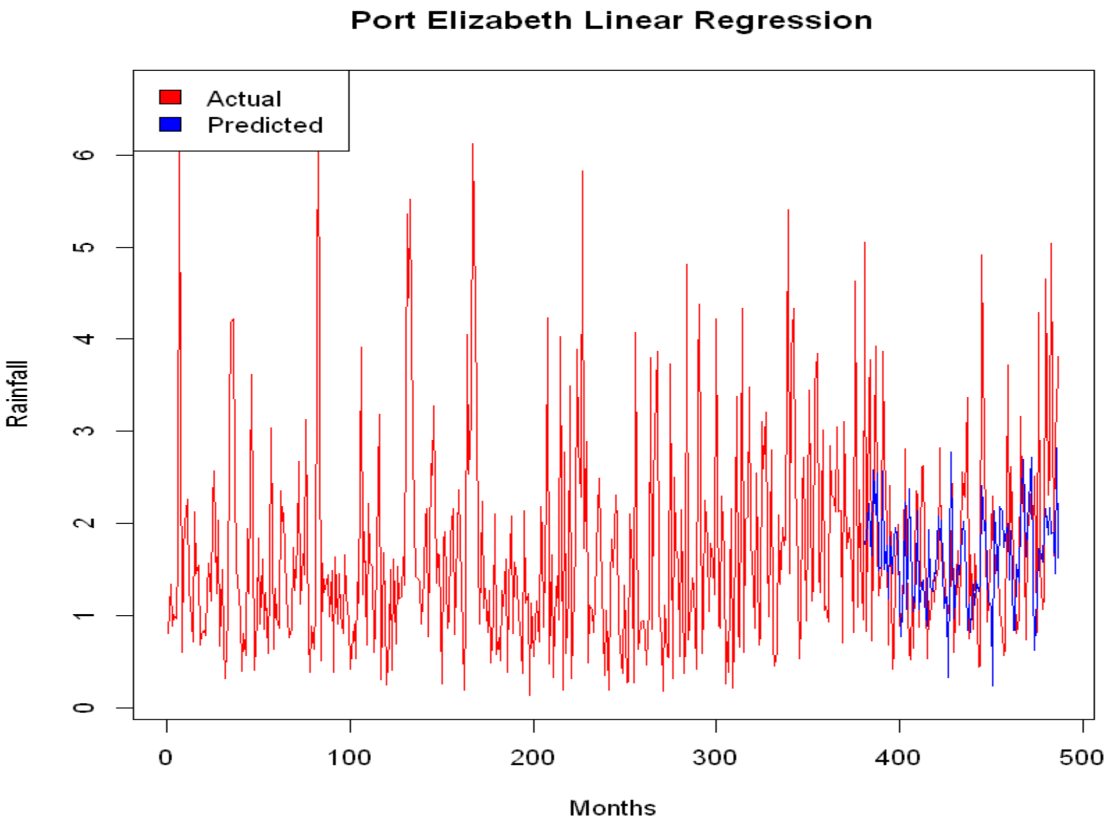
Kimberly Ridge & Lasso

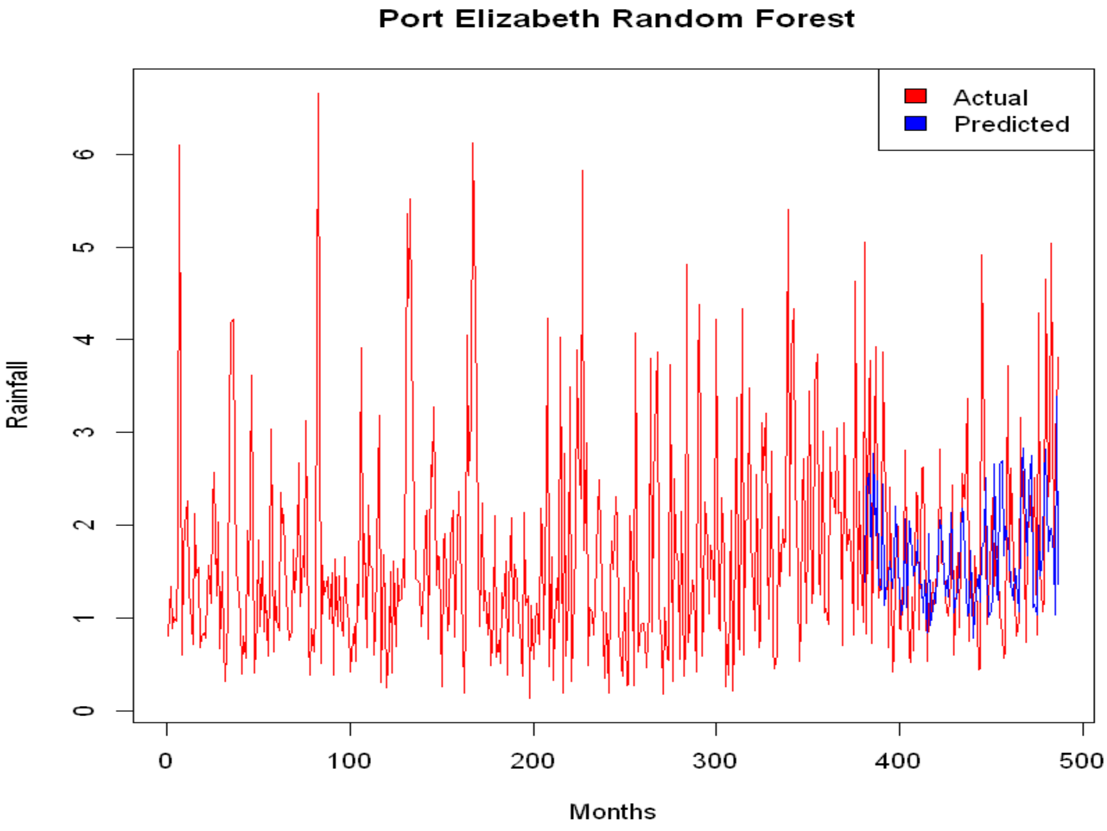

Mahikeng Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
0
100
200
300
400
500
Months

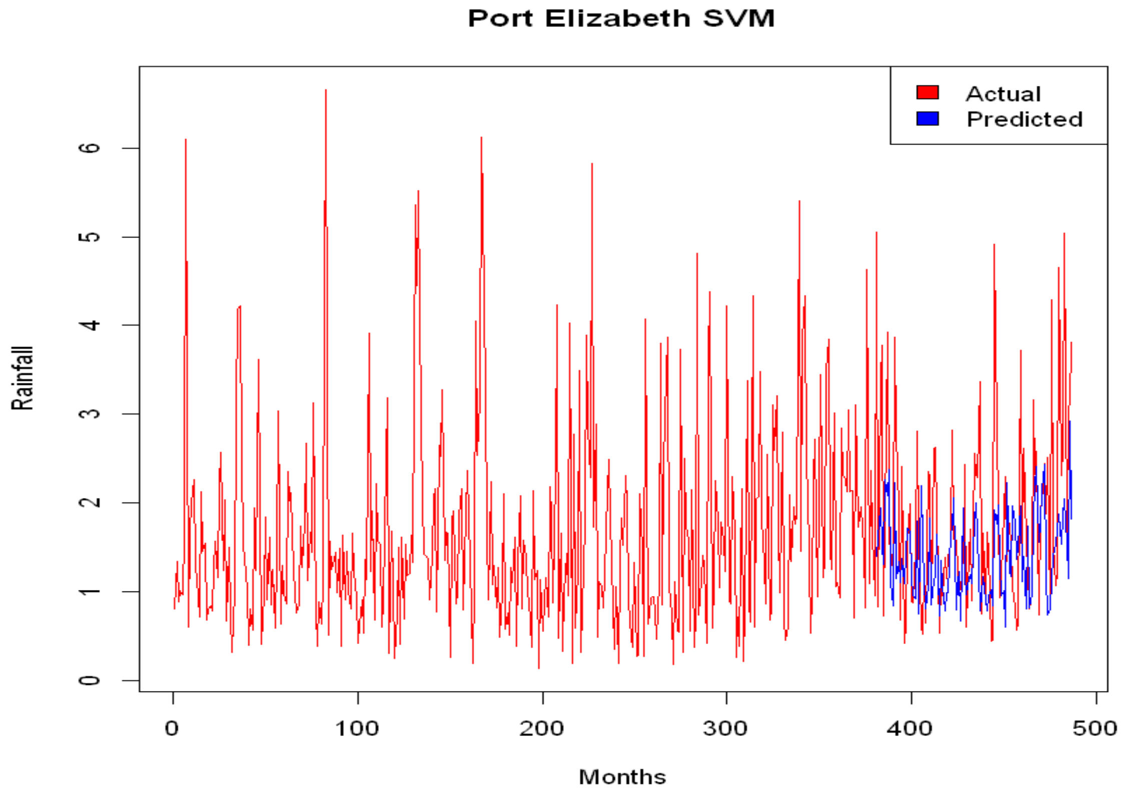

Mahikeng Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
0
100
200
300
400
500
Months


Mahikeng SVM
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
0
100
200
300
400
500
Months


Mahikeng Ridge & Lasso
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
0
100
200
300
400
500
Months


Port Elizabeth Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
1
2
3
4
5
6
0
100
200
300
400
500
Months


Port Elizabeth Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
1
2
3
4
5
6
0
100
200
300
400
500
Months


Port Elizabeth SVM
🟥 Actual
🟦 Predicted
Rainfall
0
1
2
3
4
5
6
0
100
200
300
400
500
Months

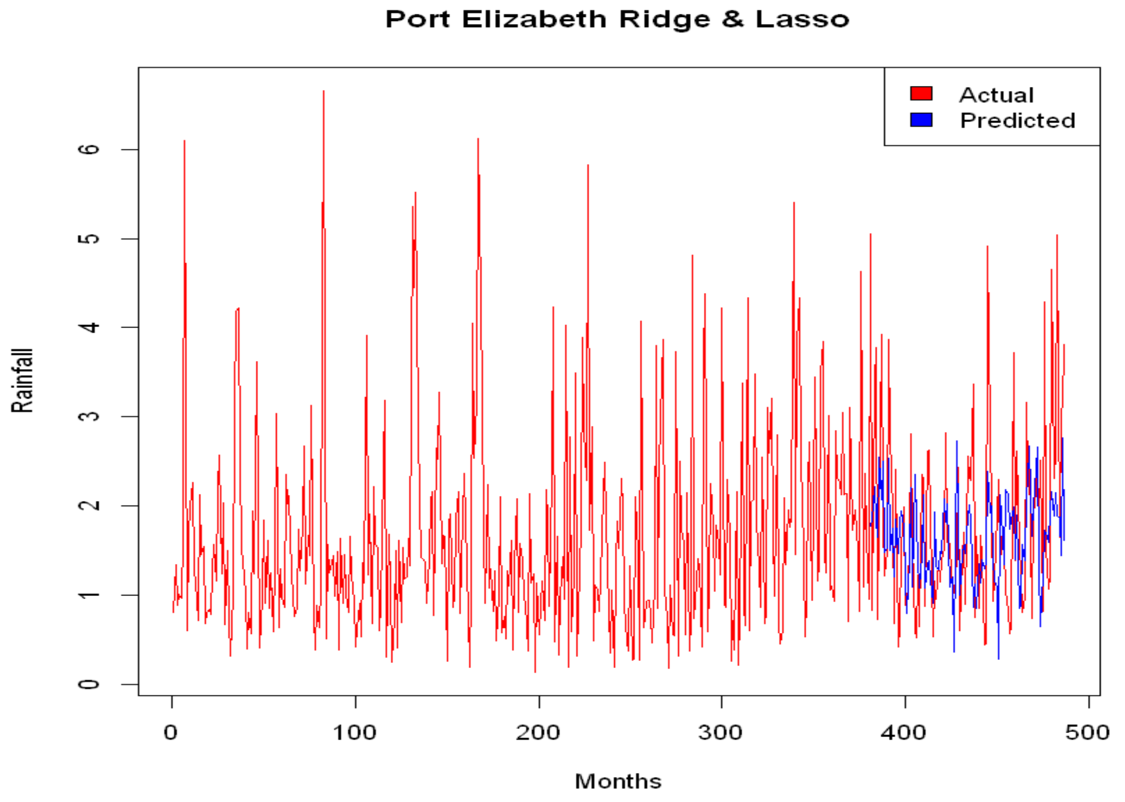

Port Elizabeth Ridge & Lasso
🟥 Actual
🟦 Predicted
Rainfall
0
1
2
3
4
5
6
0
100
200
300
400
500
Months

Figure 4. 3: Figures showing the predicted and actual rainfall for models used and locations under the hot and semi-arid steppe climate classification.

4.1.4 Hot Arid Desert

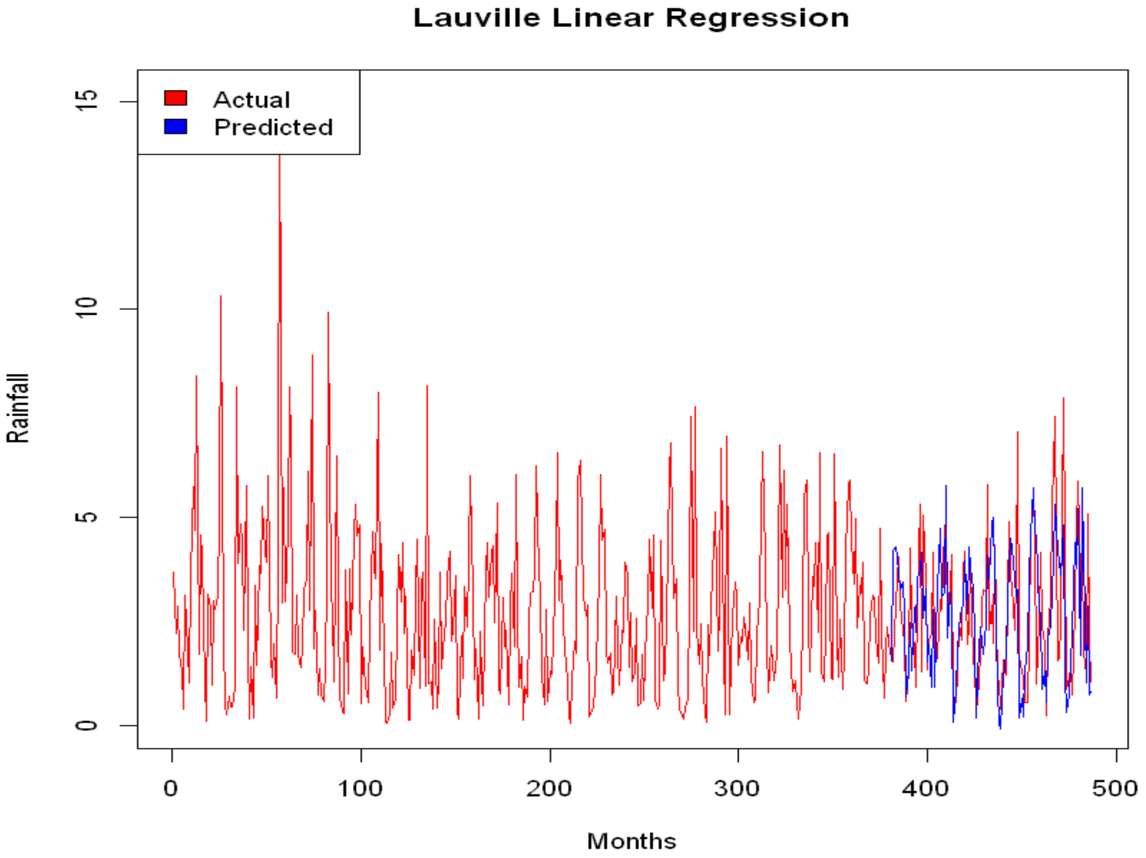
This is the last zone under the arid climate classification. Lauville in Mpumalanga Province, Musina in Limpopo Province, and Upington in Northern Cape Province were selected for the hot arid desert classification. These regions are in the northern and northwestern part of South Africa. Table 4.4 shows the evaluation metrics for the models used for the study for the three locations. In Lauville, support vector machine had the highest correlation coefficient of 0.60 while linear regression had the highest coefficient of determination of 0.76. Musina had a relatively high correlation coefficient for all models ranging from 0.68 for ridge and lasso to 0.79 for support vector machine. Values for linear regression and random forest are 0.69 and 0.74 respectively. Linear regression had the highest coefficient of determination of 0.75 in Musina while that of random forest, support vector machine, ridge and lasso are 0.52, 0.59, 0.39 respectively. This result presents hope for farmers in Musina as many of them complain of decreased agricultural produce within the last five years due to insufficient amount of rainfall and water scarcity (Mokgwathi, 2018). With accurate prediction, farmers can be better prepared for drought. In Upington, the correlation coefficients are 0.64, 0.66, 0.69, 0.63 and coefficient of determination of 0.71, 0.43, 0.40, and 0.5 for linear regression, random forest, support vector machine, and ridge and lasso respectively. There is the need to study Upington more closely due to its changing environmental and atmospheric condition at a rate different from global and continental estimates (Strydom et al., 2019). The significant decrease in the amount of rainfall in this region poses a major challenge for livestock farming (Roffe et al., 2021). The mean absolute error, mean square error and the root mean square for all models in Lauville and Upington were in a similar range. The values were also similar for all models in Musina except for linear regression where the value is about double.

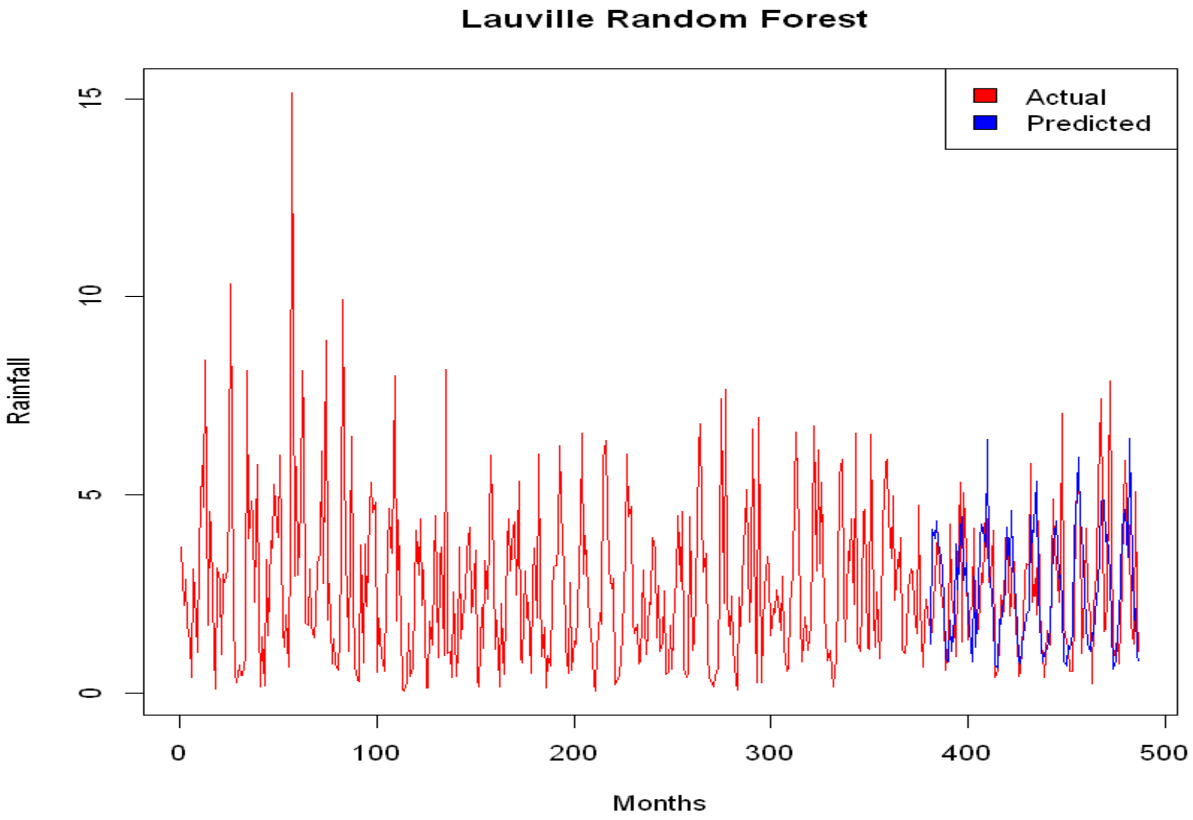
Figure 4.4 shows the graph of actual and predicted rainfall using different models for the three locations under the hot arid climate classification. The result showed that random forest performed best in Upington as it was able to predict the increased rainfall in 2022 as well as its seasonal variability. It however performed poorly in rainfall estimation between 2015 and 2018 after which the performance was good. Other models have similar pattern in Upington. For Musina, Ridge and Lasso as well as random forest also performed best compared to linear regression as they also accurately predicted the seasonal variation with underestimation of rainfall in 2018, 2021, and 2022. All models performed well in Lauville.

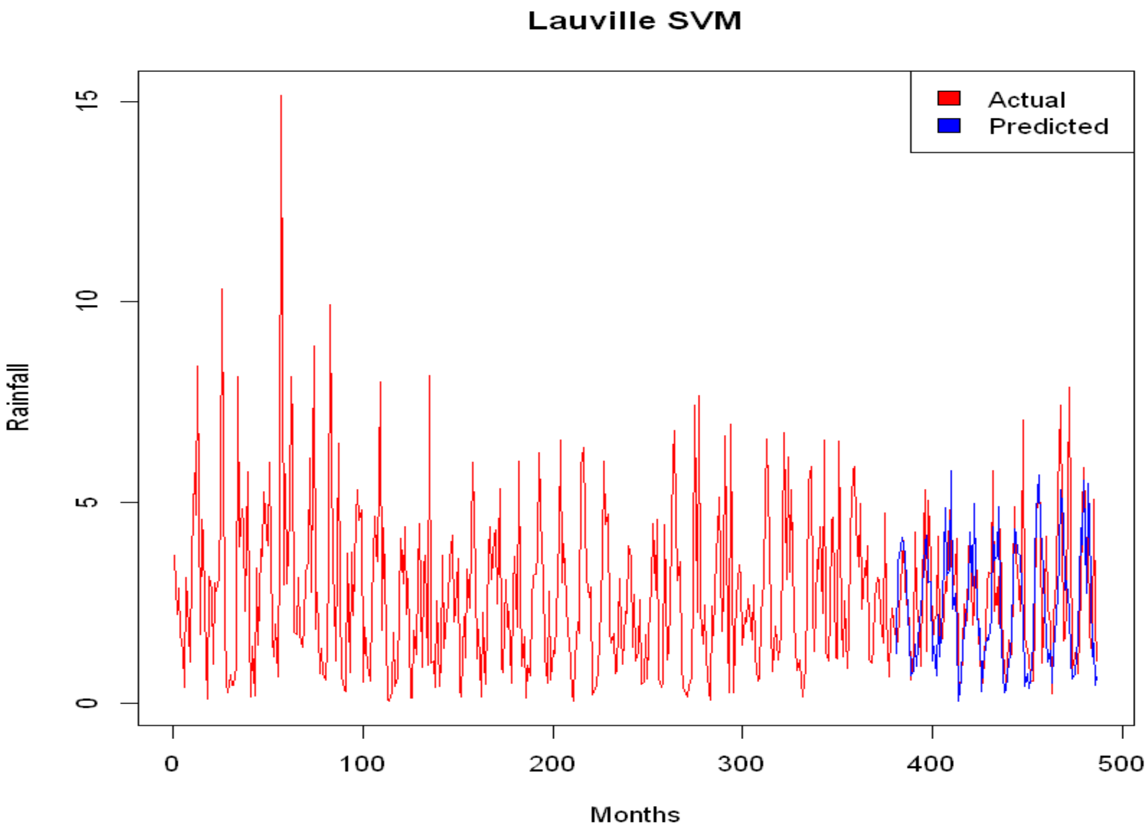
The heatmap of the correlation of atmospheric parameters with rainfall showed that cloud cover had the best coefficient in Lauville and Upington corresponding to 0.65 and 0.61 respectively. In Musina, the best correlation with rainfall was with water vapour corresponding to 0.64 closely followed with cloud cover and dew point corresponding to 0.62 and 0.60 respectively. Other atmospheric variables also had good correlation with rainfall in Lauville. Dewpoint, relative humidity, temperature, and water vapour all correlated with rainfall with coefficients corresponding to 0.51, 0.55, 0.45, and 0.58 respectively. Although most atmospheric variables had correlation coefficient greater than 0.50 with rainfall, only for cloud cover is the correlation coefficient higher than 0.60 in Lauville. This is similar to Upington as only cloud cover correlated with rainfall with coefficient greater than 0.60. Dewpoint, relative humidity, temperature, water vapour, and wind speed had correlation coefficients of 0.59, 0.35, 0.23, 0.56, and -0.41 respectively with rainfall at Upington. In Musina, only relative humidity, temperature, and wind speed had correlated with rainfall with coefficients below 0.60. The values correspond to 0.37, 0.39, and 0.14 for relative humidity, temperature, and wind speed respectively.

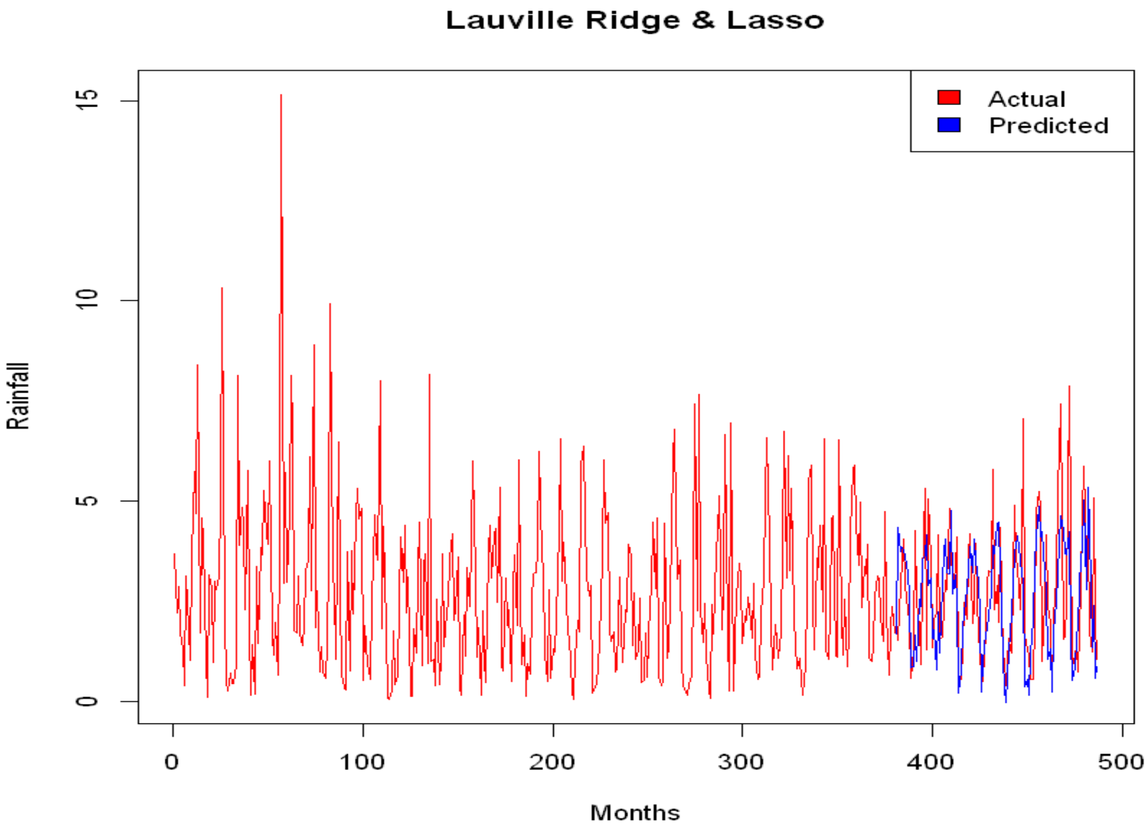
Table 4. 4: Table showing models evaluation metrics for hot arid desert climate classification.

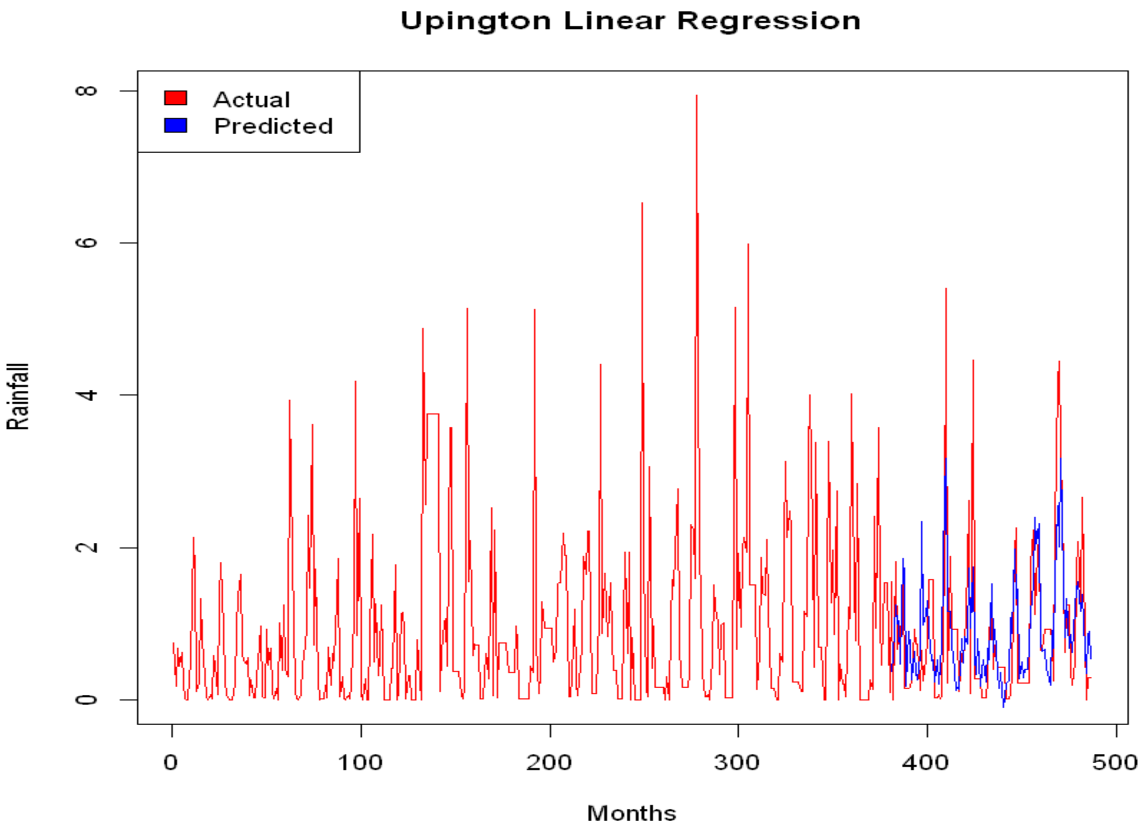
Linear Regression					
Lauville	1.07	2.04	1.43	0.53	0.76
Musina	1.44	3.04	1.74	0.69	0.75
Upington	0.93	1.36	1.17	0.64	0.71
Random Forest					
Lauville	1.17	2.41	1.55	0.51	0.12
Musina	0.76	1.01	1.01	0.74	0.52
Upington	0.57	0.59	0.77	0.66	0.43
Support Vector Machine					
Lauville	1.13	2.02	1.42	0.60	0.26
Musina	0.66	0.86	0.93	0.79	0.59
Upington	0.56	0.63	0.79	0.69	0.40
Ridge and Lasso					
Lauville	1.09	1.94	1.39	0.57	0.51
Musina	0.80	1.09	1.04	0.68	0.39
Upington	0.58	0.63	0.79	0.63	0.51

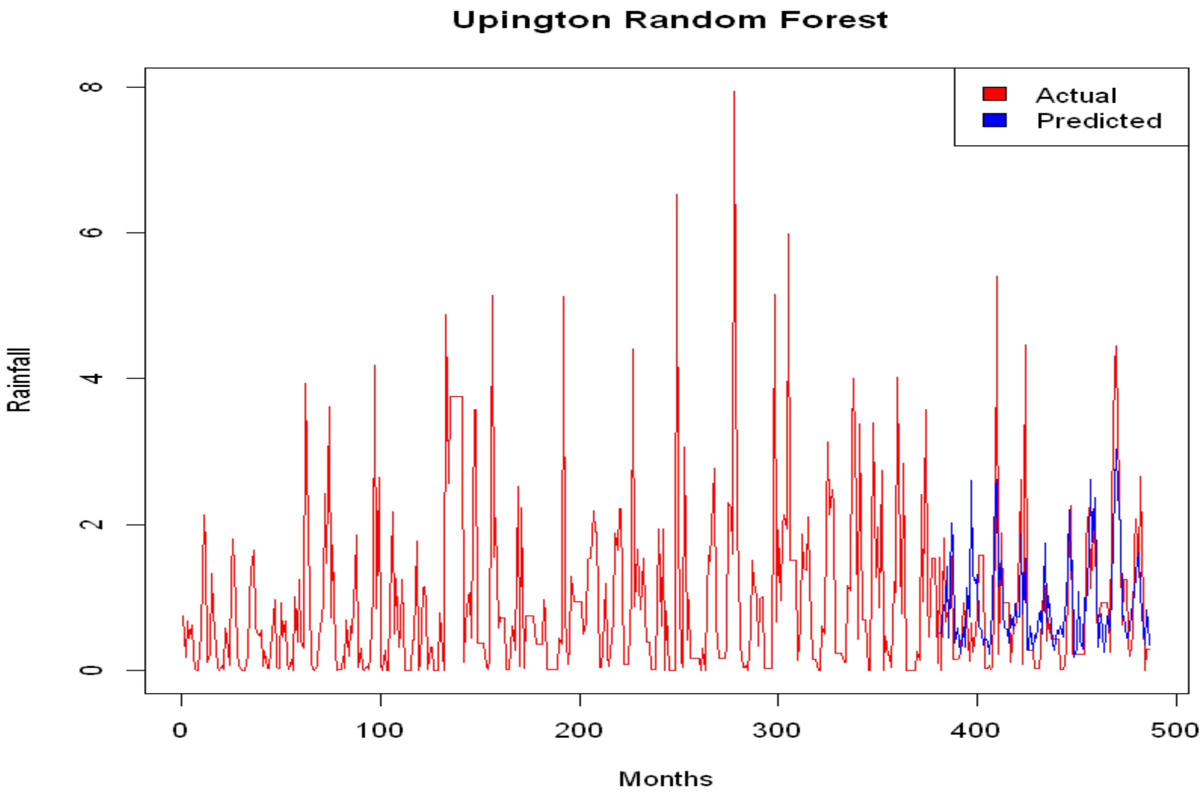

Lauville Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months

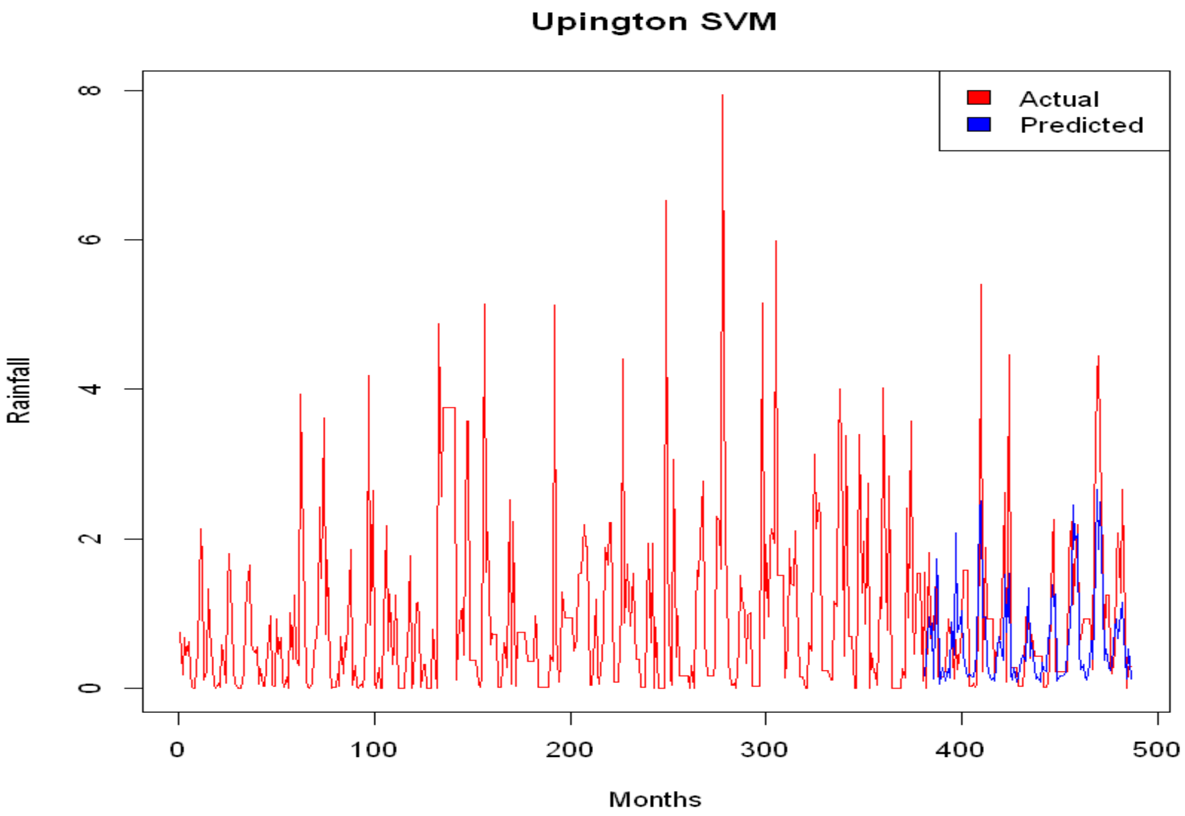

Lauville Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months

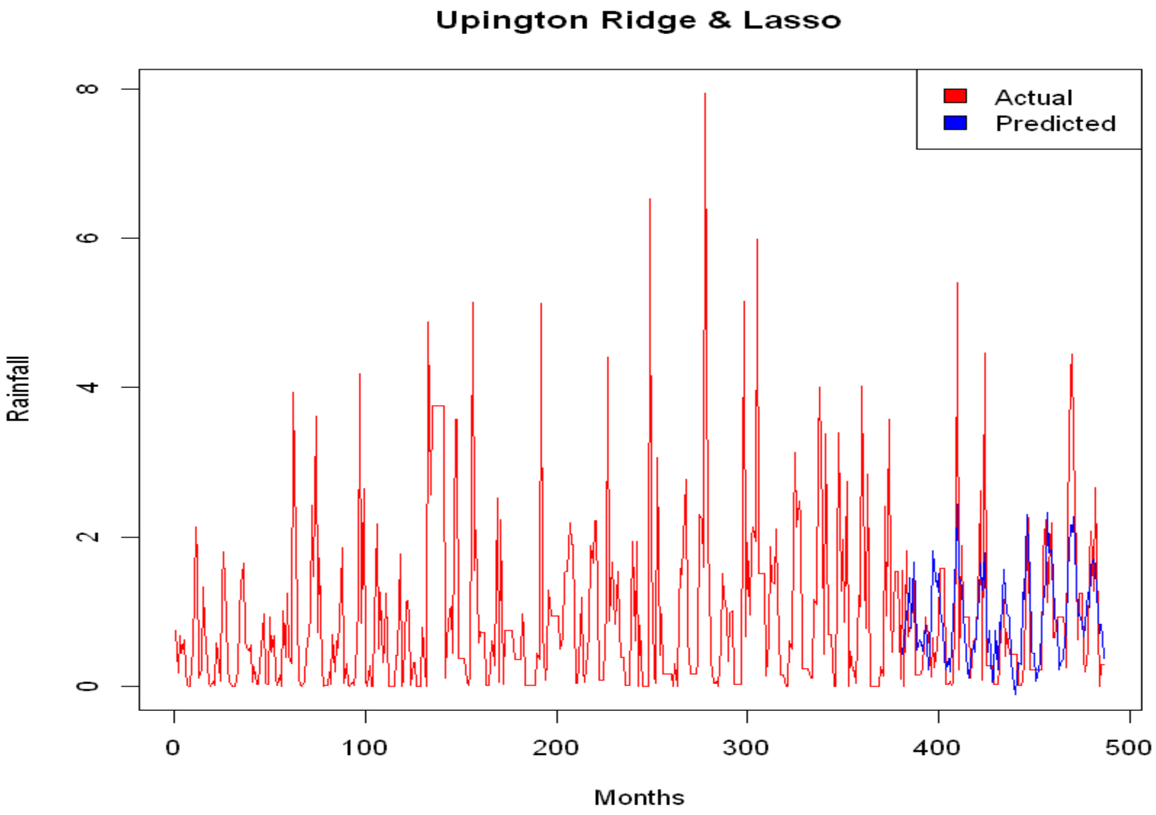

Lauville SVM
Actual
Predicted
Rainfall
15
10
5
0
0
100
200
300
400
500
Months

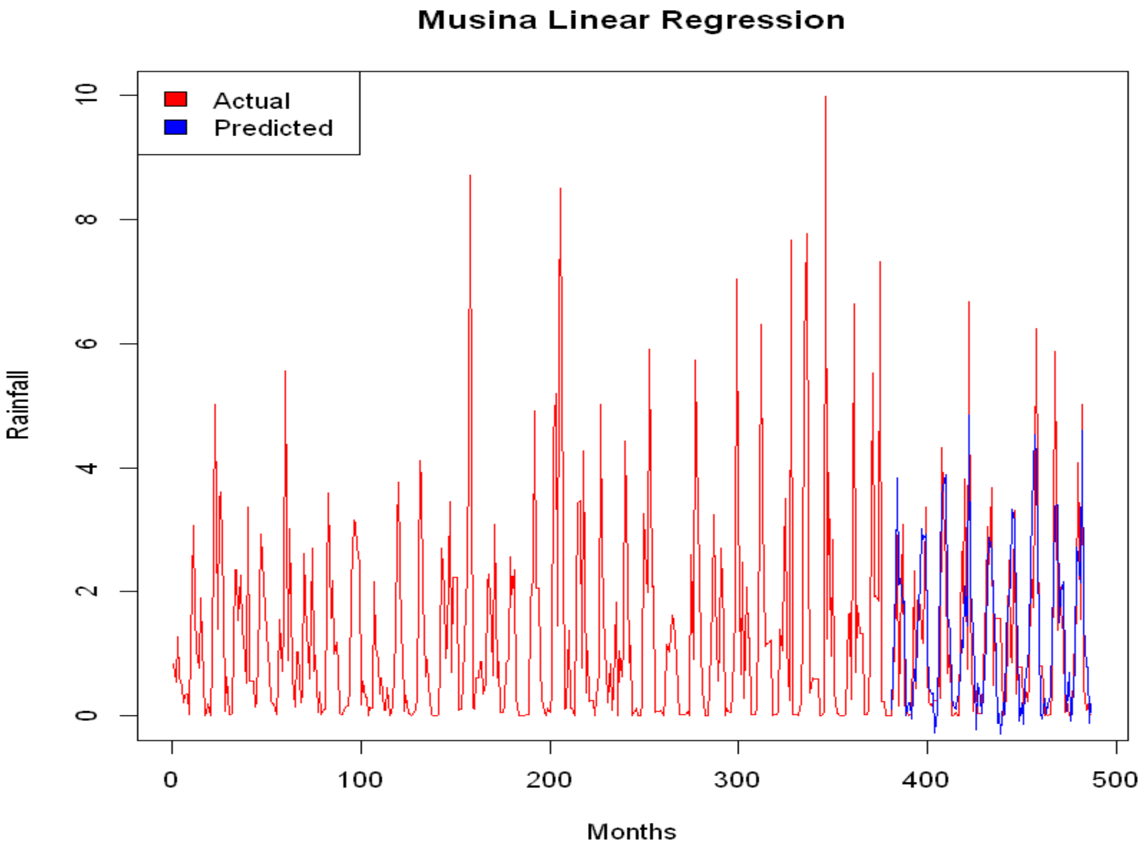

Lauville Ridge & Lasso
Actual
Predicted
Rainfall
15
10
5
0
0
100
200
300
400
500
Months

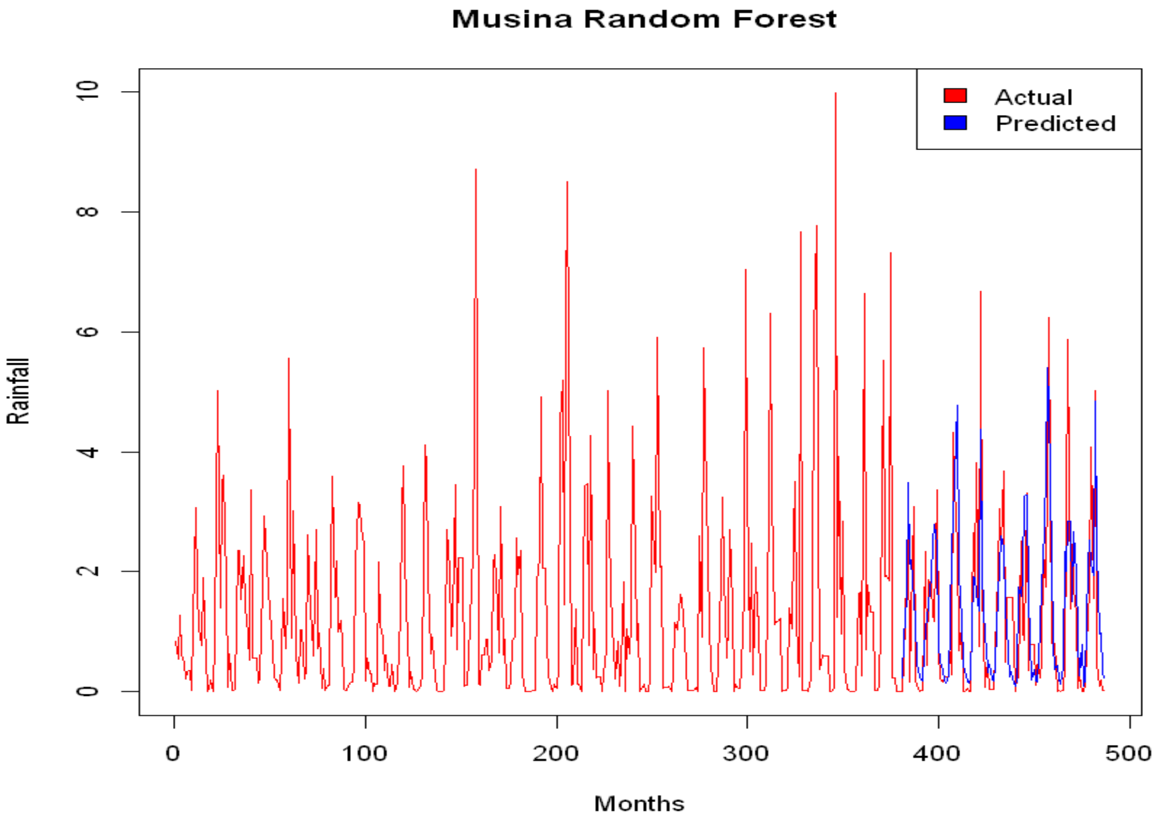

Upington Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
0
100
200
300
400
500
Months


Upington Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
0
100
200
300
400
500
Months


Upington SVM
Rainfall
Months
0
2
4
6
8
0
100
200
300
400
500
🟥 Actual
🟦 Predicted


Upington Ridge & Lasso
Rainfall
Months
0
2
4
6
8
0
100
200
300
400
500
🟥 Actual
🟦 Predicted


Musina Linear Regression
Actual
Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months


Musina Random Forest
Actual
Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months

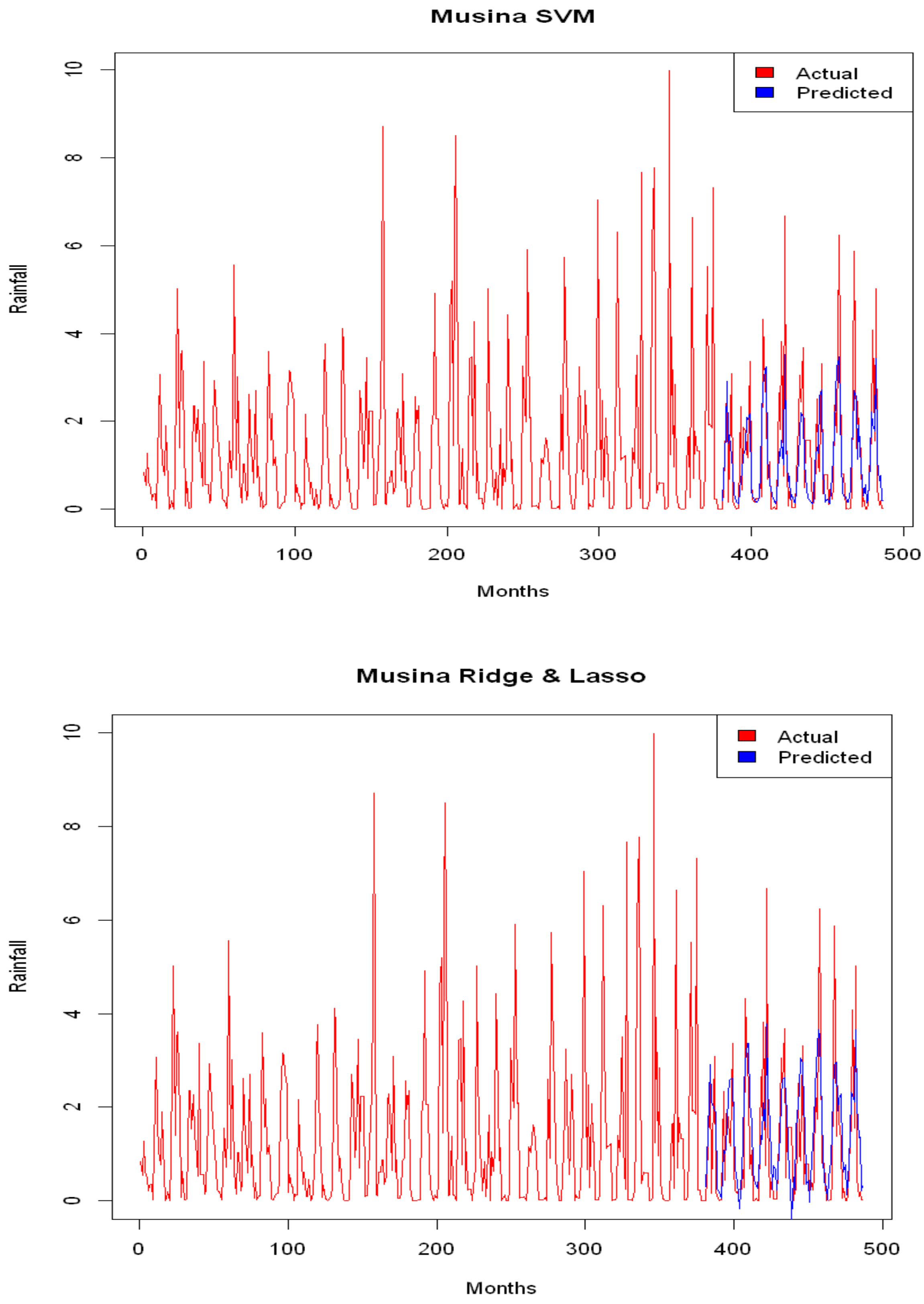


Figure 4. 4: Figure 9: Figures showing the predicted and actual rainfall for models used and locations under the hot arid desert climate classification.

Chapter 5: Results and Discussion on Subtropical Wet Weather Classification

5.0 Subtropical Wet Weather Classification

5.1 Humid Subtropical with Dry Winter

This section presents the rainfall prediction models for the subtropical wet climate classification. This climate classification is divided into two, the humid subtropical climate with dry winter and the subtropical highland climate with dry winter. For the humid subtropical climate with dry winter, Dundee in KwaZulu Natal Province, Louis Trichardt in Limpopo Province, and Nelspruit in Mpumalanga Province were selected. Table 5.1 reveals the evaluation metrics for the four models used in this study – Linear regression, random forest, support vector machine, and the ridge and lasso models. The table indicates that random forest had the highest correlation coefficient with a value of 0.77 in Dundee closely followed by the support vector machine model with a value of 0.76. For linear regression and ridge and lasso models in Dundee, the correlation coefficients were 0.70 and 0.73 respectively. For the coefficient of determination in Dundee, the values were 0.66, 0.65, 0.54, 0.53 for ridge and lasso, support vector machine, linear regression, and random forest respectively. Charpentier et al (2023) assessed the rainfall pattern in KwaZulu Natal province using datasets from 1970 to 2017. They linked the years of extreme dryness with cyclonic events induced by the El Nino and reported Dundee as one of the most affected areas during time of extreme drought. With agriculture being the major occupation in this region, it is necessary to correctly predict the amount of rainfall expected and to prepare farmers for drought seasons.

In Nelspruit, the best model with respect to the correlation coefficient is the support vector machine with a value of 0.76 followed by random forest corresponding to 0.72. The values of these coefficients for ridge and lasso as well as linear regression are 0.71 and 0.69 respectively. However, linear regression had the highest coefficient of determination of 0.84 followed by ridge and lasso with a value of 0.61. Comparing the locations selected for the humid subtropical climate with dry winter, Louis Trichardt had the highest correlation coefficient corresponding to 0.81, 0.76, 0.76, 0.74 for support vector machine, random forest, ridge and lasso, and linear regression while their coefficient of determination corresponds to 0.65, 0.52, 0.59, and 0.78 respectively. This result performed better than what was obtained by van Tol et al (2020) when they estimated the hydrological response in Stevenson Hamilton research supersite of the Kruger National Park, a park which spans across Mpumalanga and Limpopo provinces. They used parametric data model using measured properties of soil, soil matric potentials, and evapotranspiration data. Their model had a correlation coefficient ranging from 0.58 to 0.69.

In Nelspruit, the values of the mean absolute error, mean square error and root mean square error for linear regression, random forest, as well as the ridge and lasso models were similar while the values for the support vector machine were lower. Louis Trichardt also had similar values for both ridge and lasso and random forest models. While the values of these metrics in the support vector machine model were lower, they were much higher for linear regression. However, in Dundee, support vector machine and random forest had similar values in these metrics while values for these metrics in linear regression as well as ridge and lasso were not far apart.

Figure 5.1 shows the actual and predicted rainfall for these locations using different models. The datasets available in Dundee were from 1983 to 2017 compared to other datasets available to 2023. The datasets were later updated to 2023. However, using the datasets to 2017, the prediction was from 2012 to 2017 and all models accurately modelled the seasonal variations as well as rainfall estimates for all years except 2016. The same can be said for both Louis Trichardt and Nelspruit as all models predicted the seasonal variability and underestimation of rainfall in 2022. Reasons for underestimation in 2022 had been discussed in earlier chapter. Linear regression performed best in Nelspruit followed by random forest, then ridge and lasso.

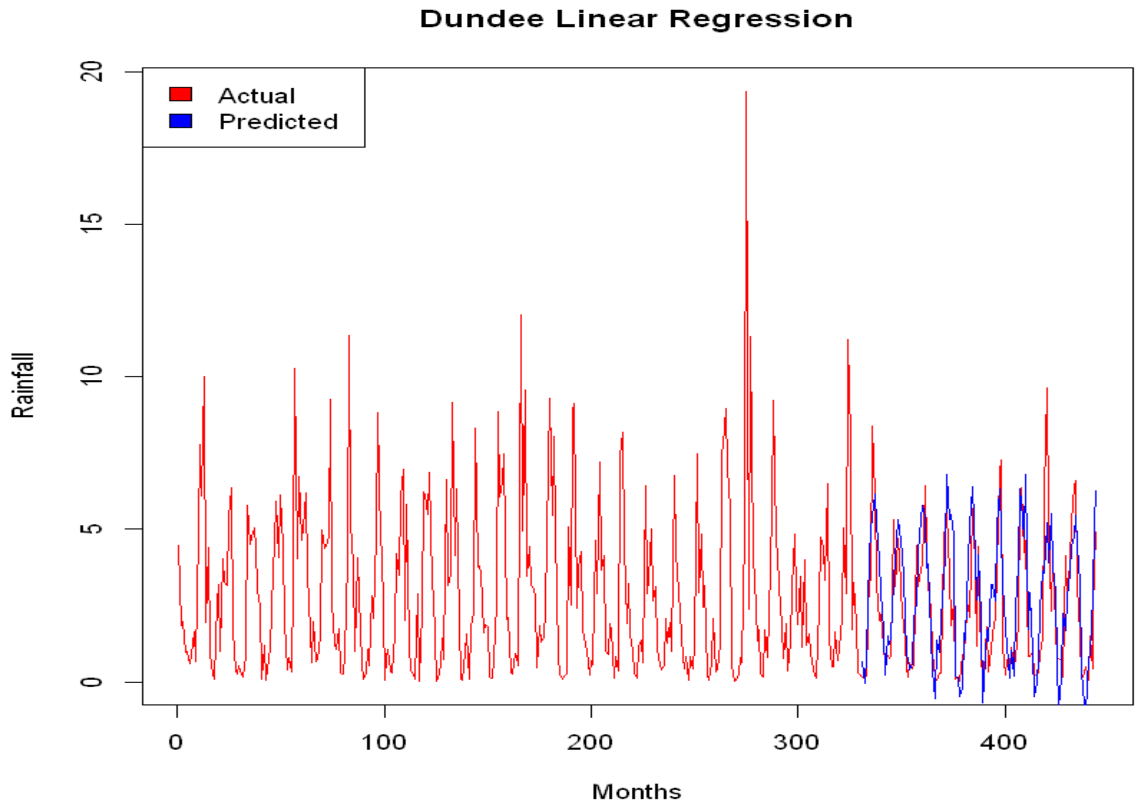
While in Louis Trichardt, ridge and lasso performed best in predicting the amount of rainfall received followed by the linear regression model.

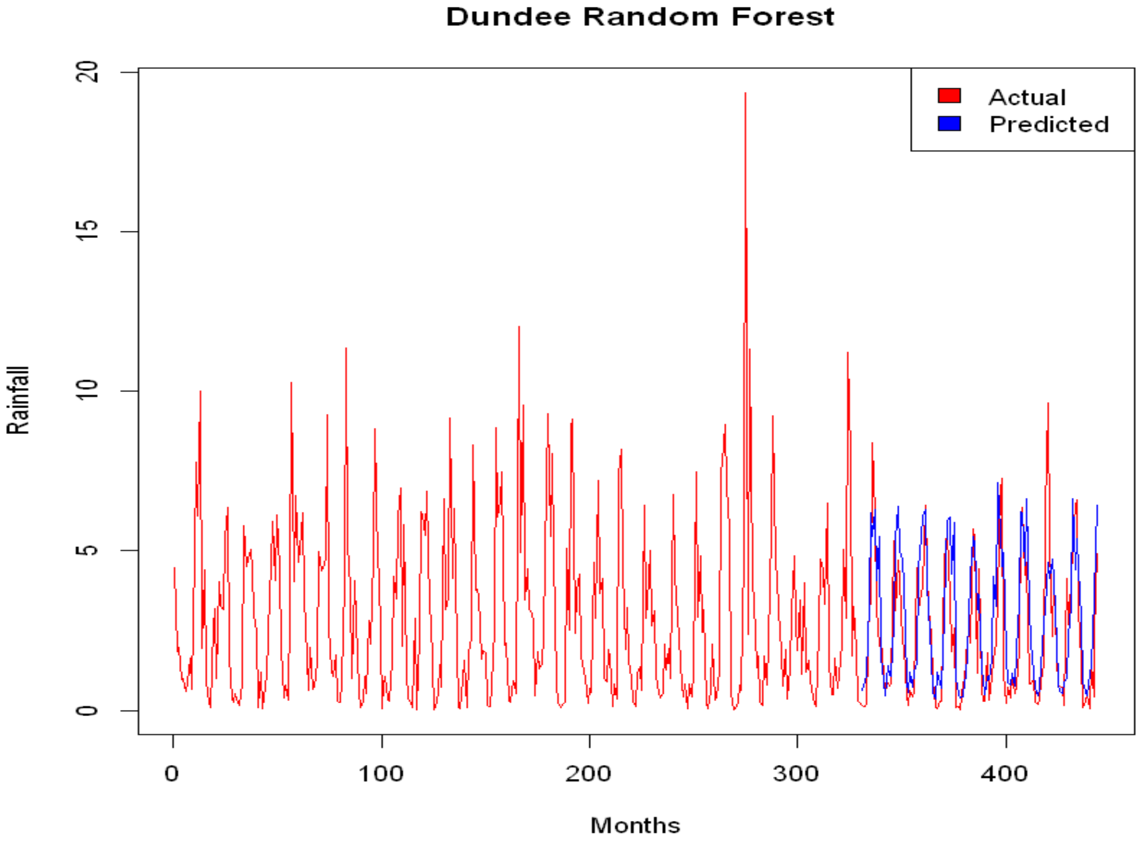
The heatmap as presented in the appendix showed that rainfall and water vapour both at Dundee and Nelspruit had the best correlation coefficient with a value of 0.71 in both cities. Water vapour also had the best correlation with rainfall in Louis Trichardt with a value of 0.66. Dew point, temperature, and cloud cover had a correlation coefficient of 0.66, 0.60, and 0.69 respectively with rainfall in Nelspruit while in Dundee, their correlation with rainfall corresponds to 0.59, 0.66, -0.49 respectively. Relative humidity had a correlation coefficient of 0.66 with rainfall in Dundee. The result of the correlation between cloud cover and rainfall in Dundee is quite strange as cloud cover and rainfall have shown to have one of the highest degrees of correlation in other cities. This reason for this is yet to be determined. In Louis Trichardt, apart from water vapour, only dew point and cloud cover had a correlation coefficient greater than 0.50 with rainfall corresponding to 0.62 and 0.63 respectively. Relative humidity, temperature, and wind speed all had a coefficient corresponding to 0.44, 0.45, and 0.11 with rainfall respectively.

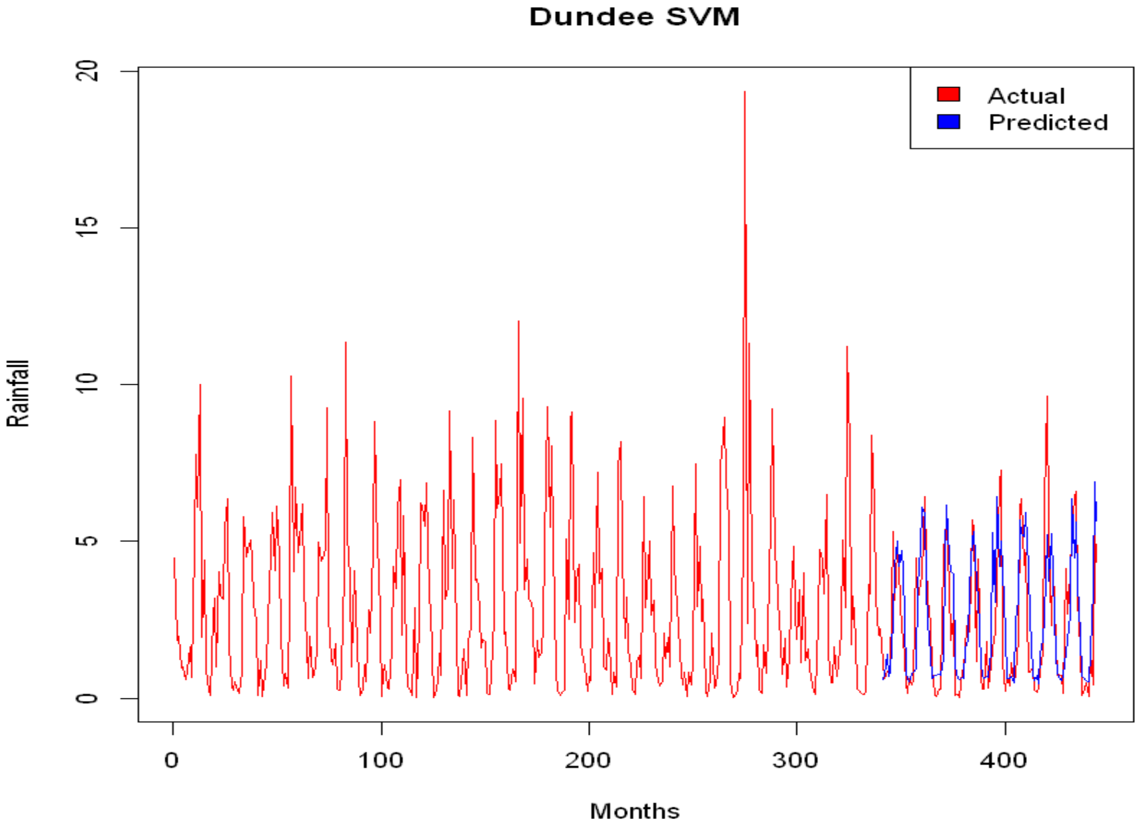
Table 5. 1: Table showing models evaluation metrics for humid subtropical with dry winter climate classification.

Linear Regression					
Dundee	1.59	3.72	1.93	0.70	0.54
Louis Trichardt	1.32	2.67	1.63	0.74	0.78
Nelspruit	1.13	2.21	1.49	0.69	0.84
Random Forest					

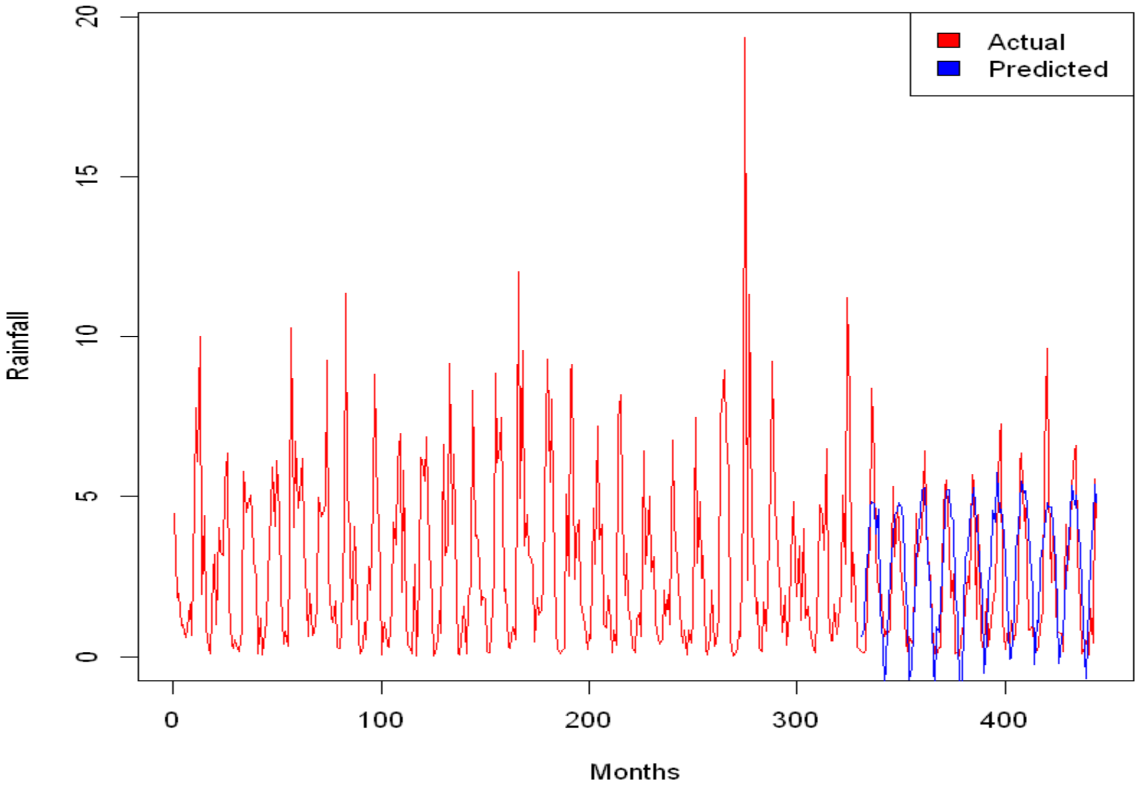
Dundee	1.11	2.30	1.52	0.77	0.53
Louis Trichardt	0.72	1.04	1.02	0.76	0.52
Nelspruit	1.10	2.35	1.53	0.72	0.49
Support Vector Machine					
Dundee	1.04	2.13	1.46	0.76	0.65
Louis Trichardt	0.62	0.77	0.88	0.81	0.65
Nelspruit	0.93	1.95	1.40	0.76	0.58
Ridge and Lasso					
Dundee	1.40	2.85	1.69	0.73	0.66
Louis Trichardt	0.79	1.06	1.03	0.76	0.59
Nelspruit	1.13	2.41	1.55	0.71	0.61



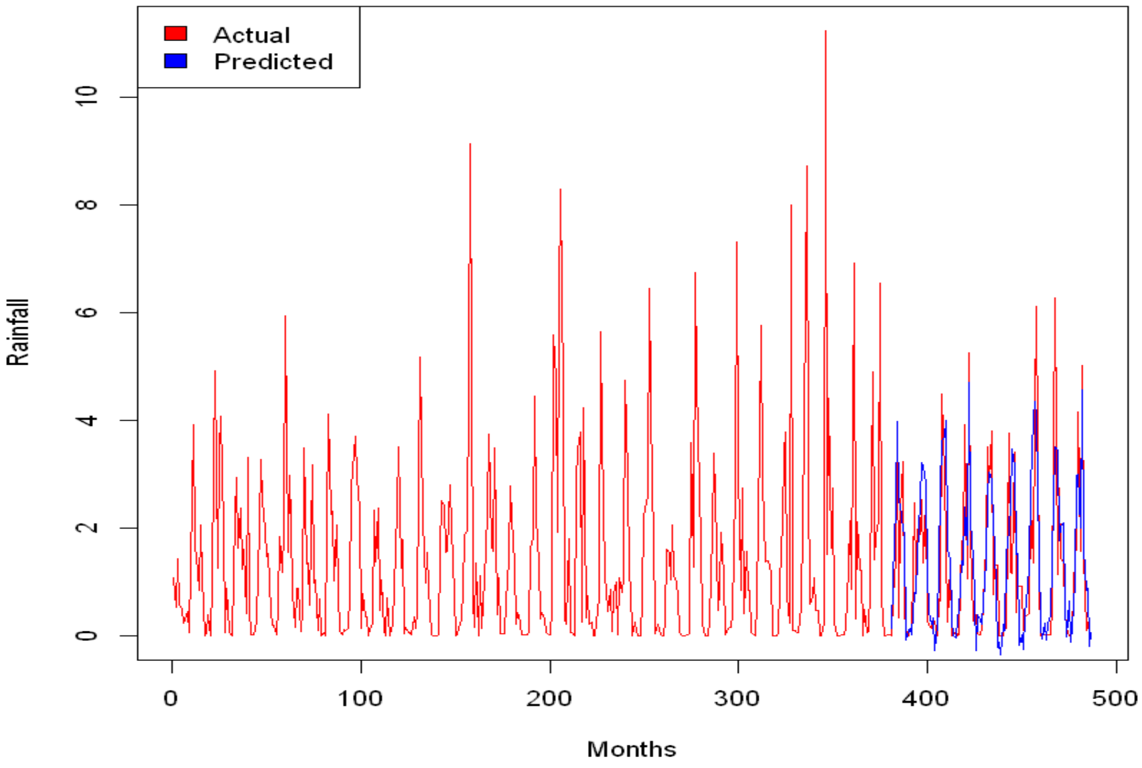

Dundee Random Forest
Actual
Predicted
Rainfall
0
5
10
15
20
0
100
200
300
400
Months

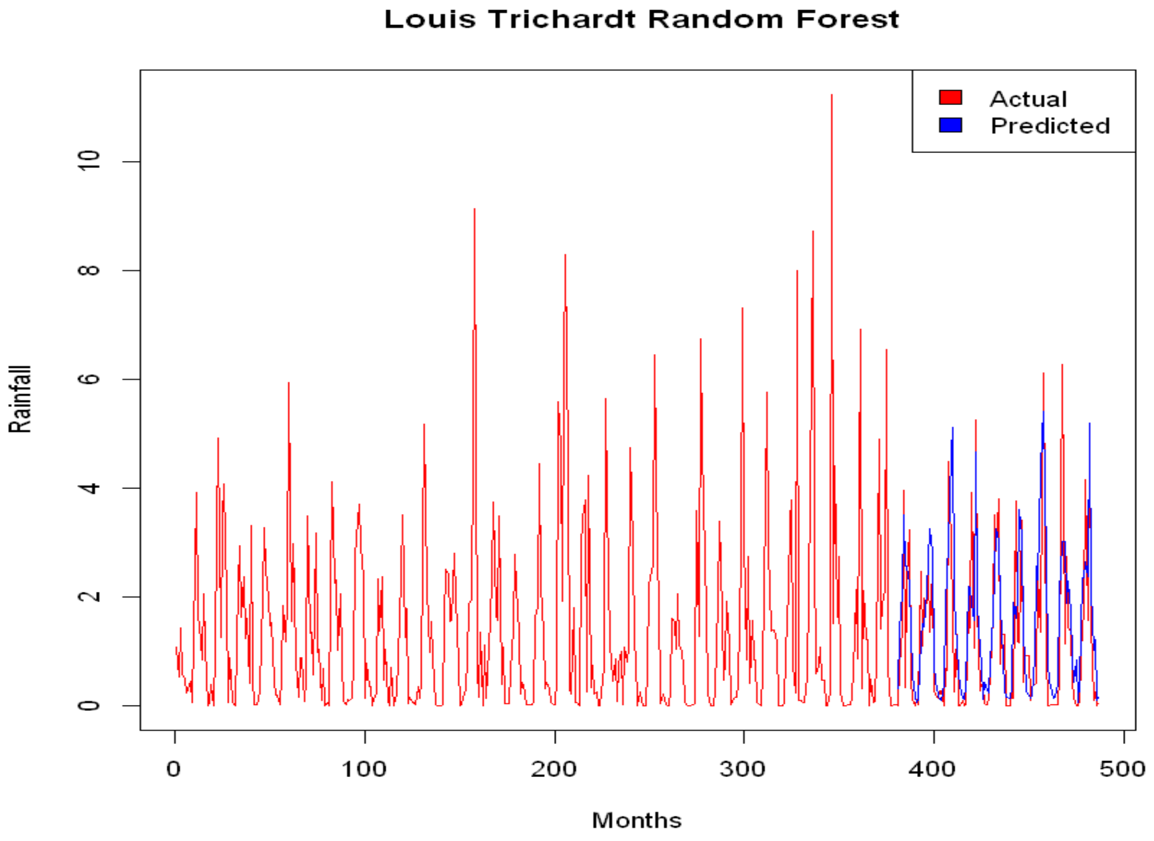

Dundee SVM
Actual
Predicted
Rainfall
0
5
10
15
20
0
100
200
300
400
Months

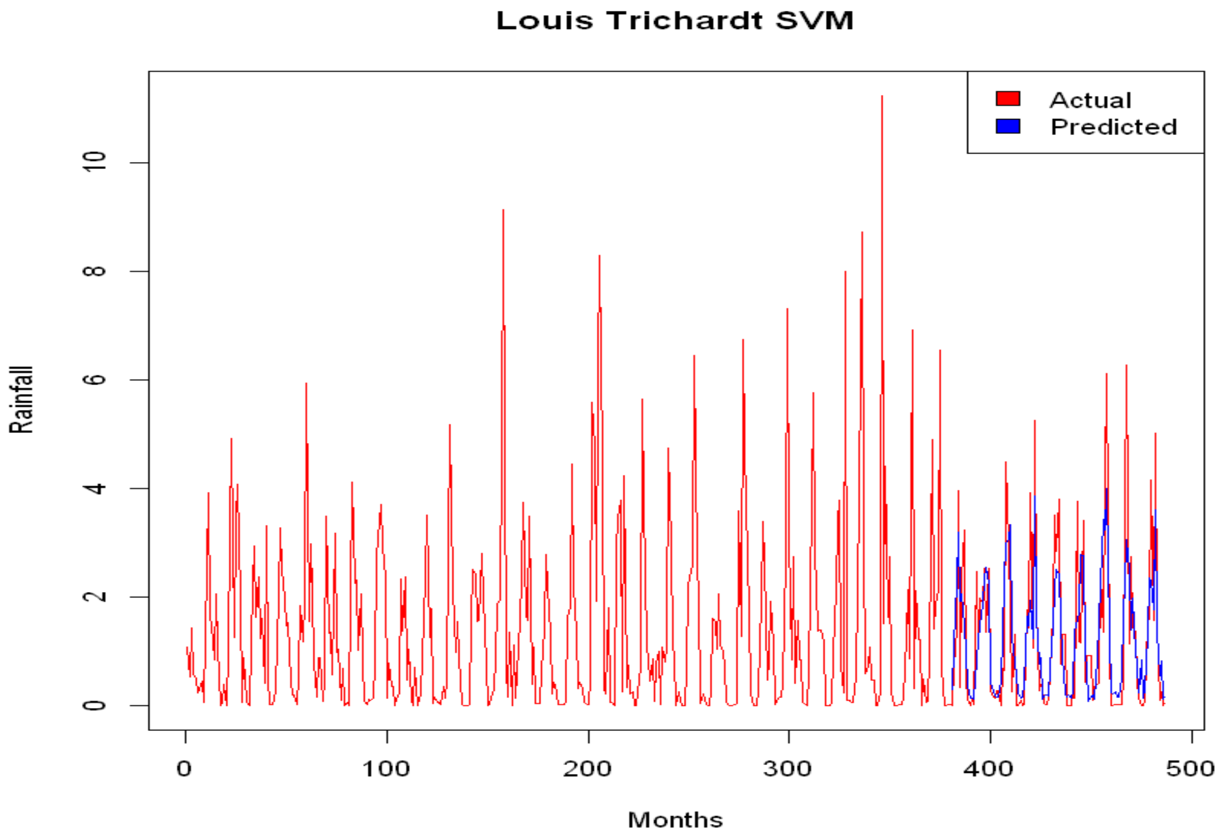
Dundee Ridge & Lasso

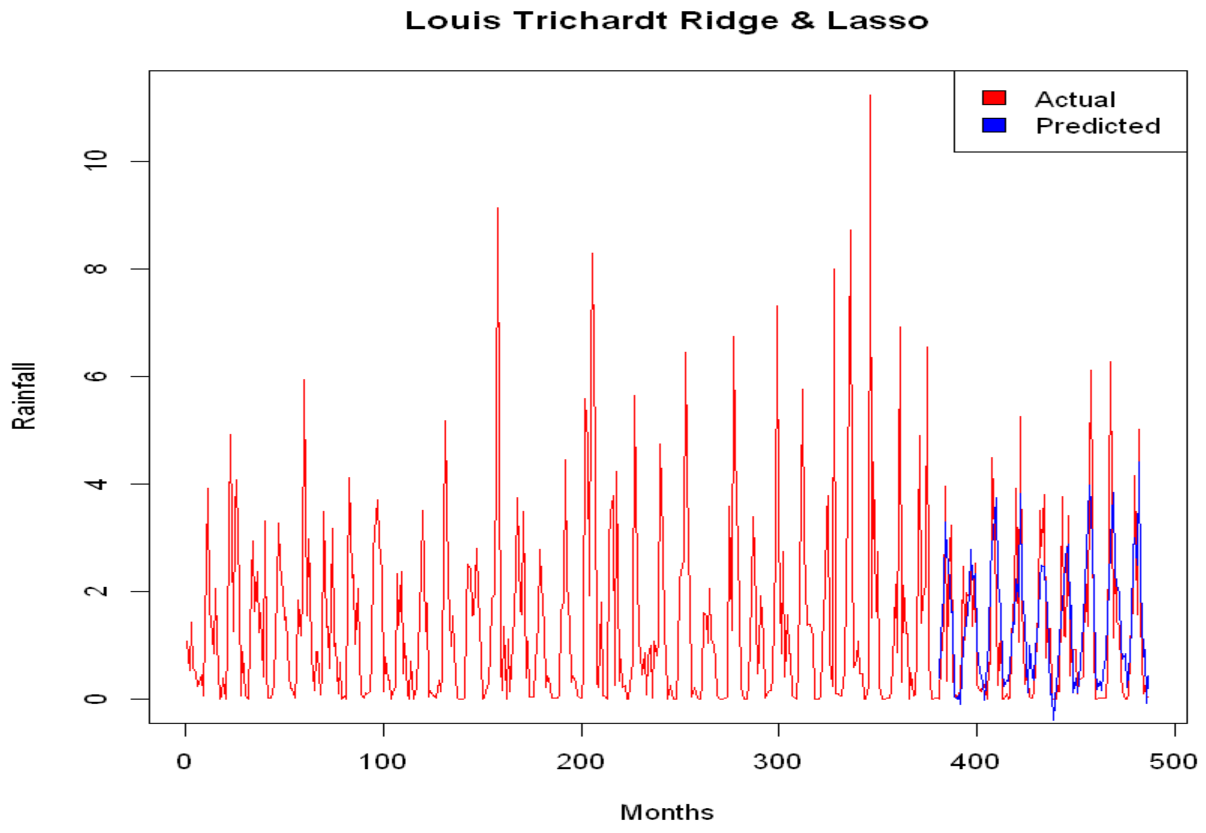


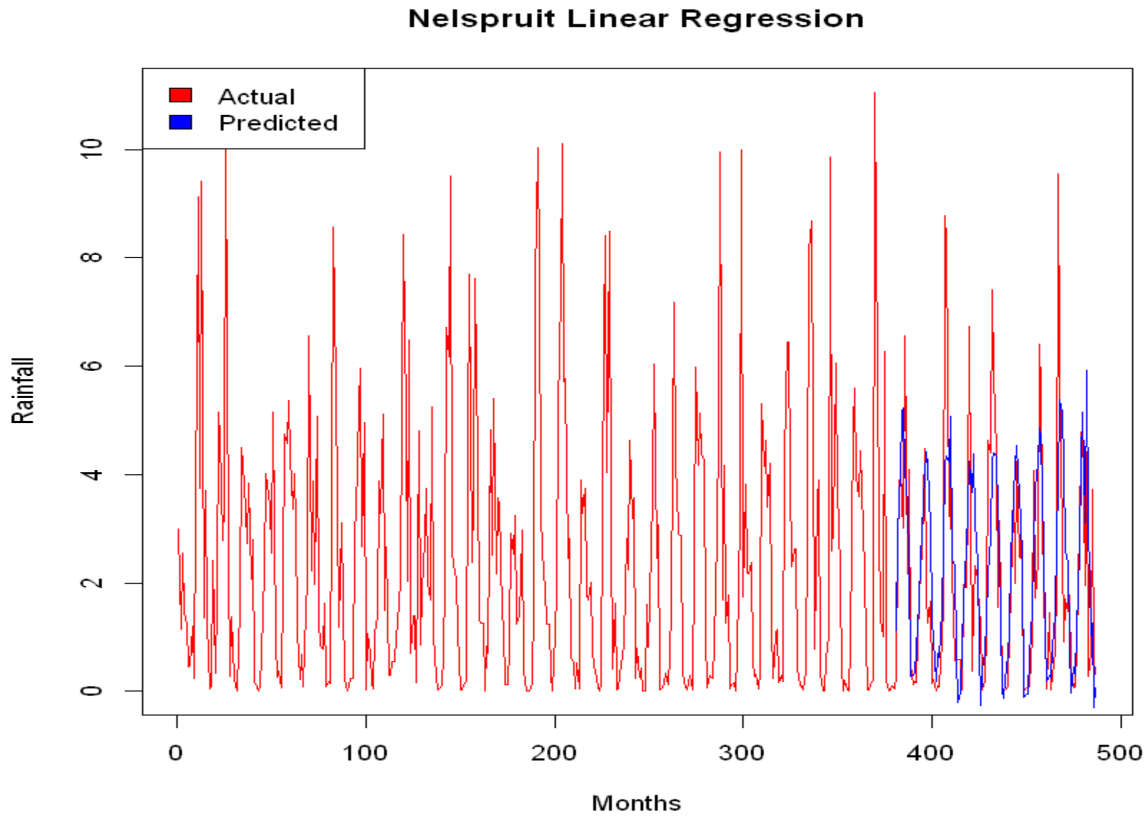
Louis Trichardt Linear Regression

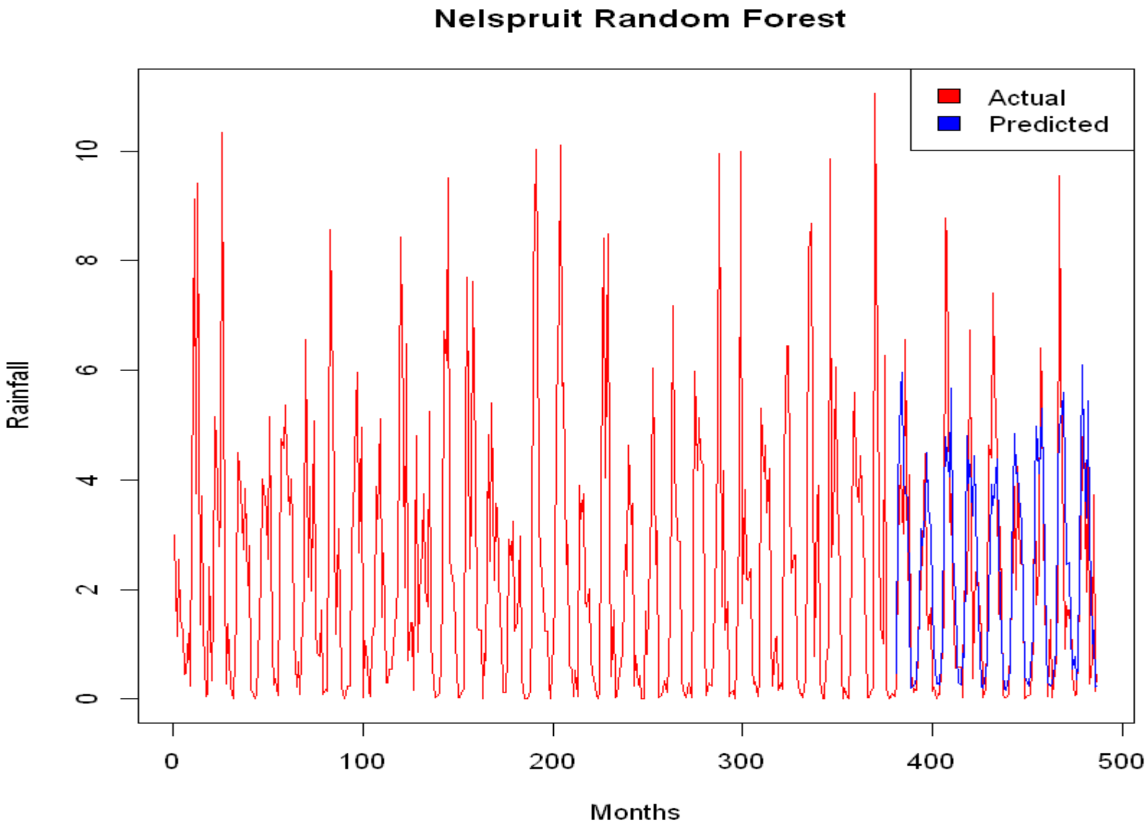


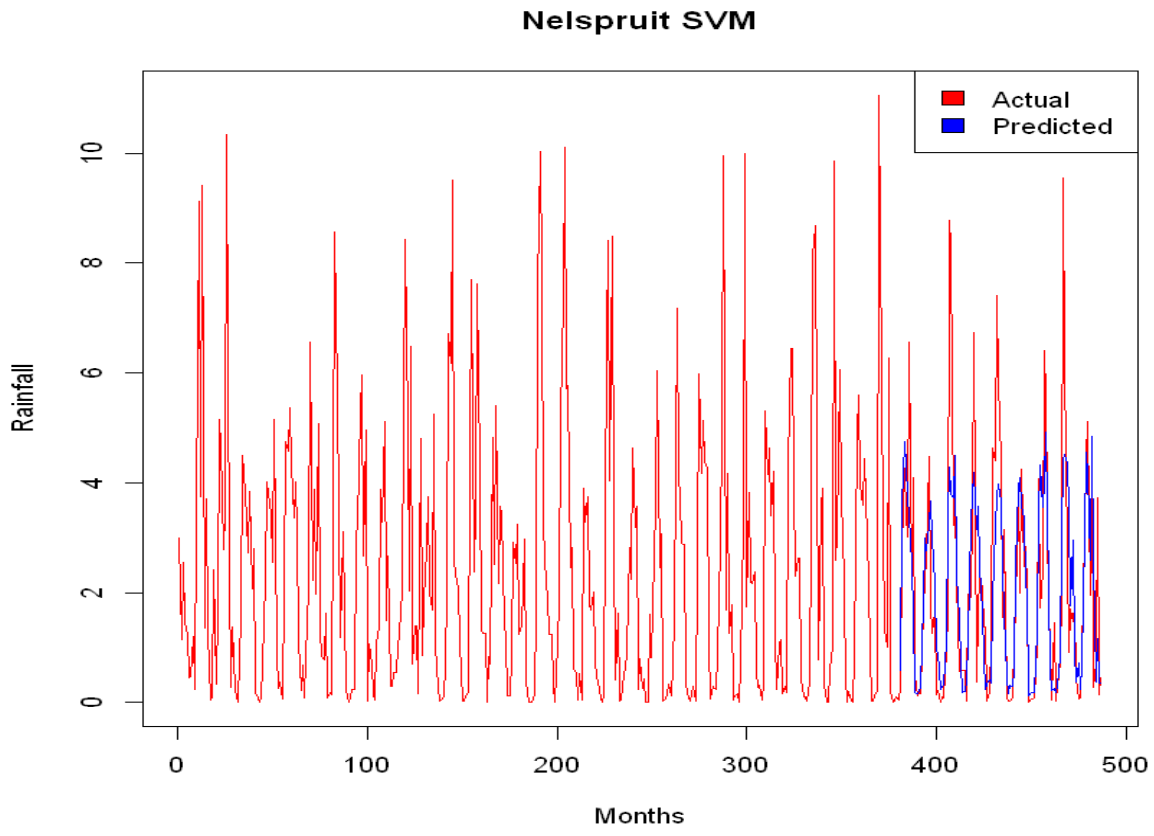

Louis Trichardt Random Forest
Actual
Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months


Louis Trichardt SVM
Actual
Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months


Louis Trichardt Ridge & Lasso
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months


Nelspruit Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months


Nelspruit Random Forest
Actual
Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months


Nelspruit SVM
Actual
Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months

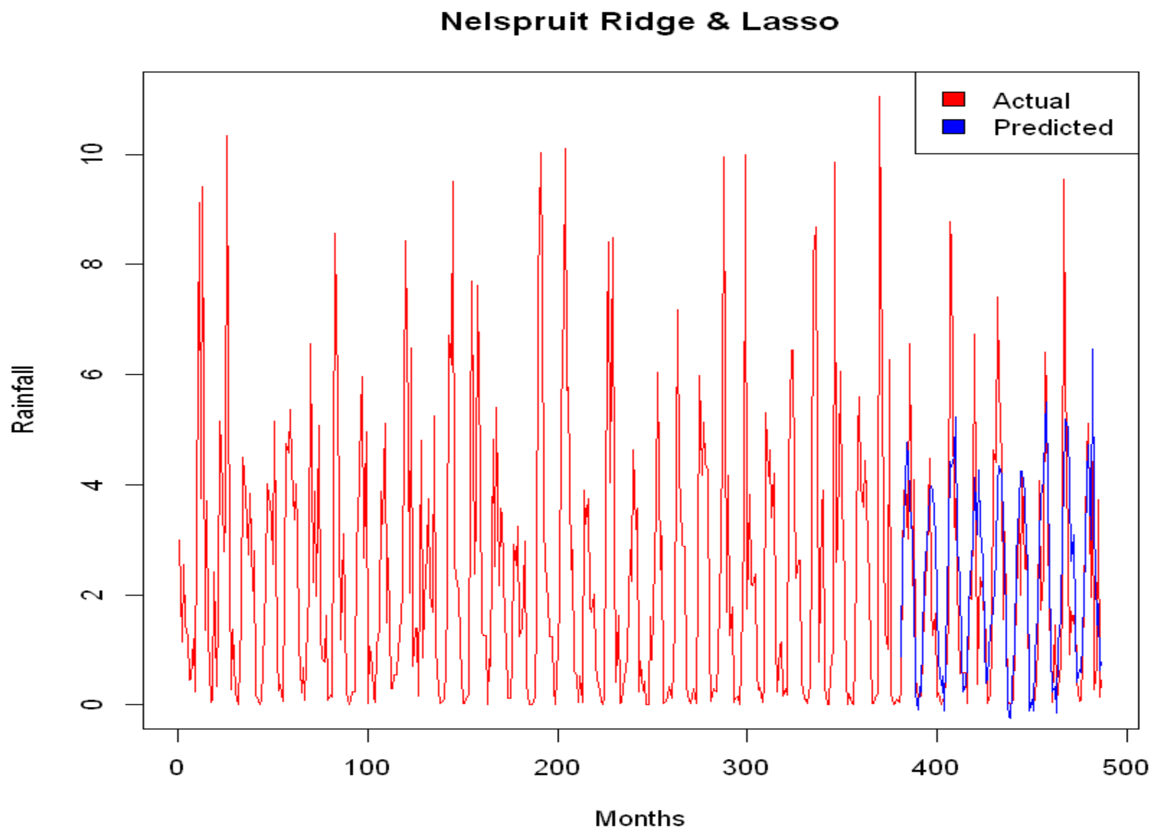

Nelspruit Ridge & Lasso
Actual
Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months

Figure 5. 1: Figures showing the predicted and actual rainfall for models used and locations under the humid subtropical with dry winter climate classification.

5.2 Subtropical Highland with Dry Winter

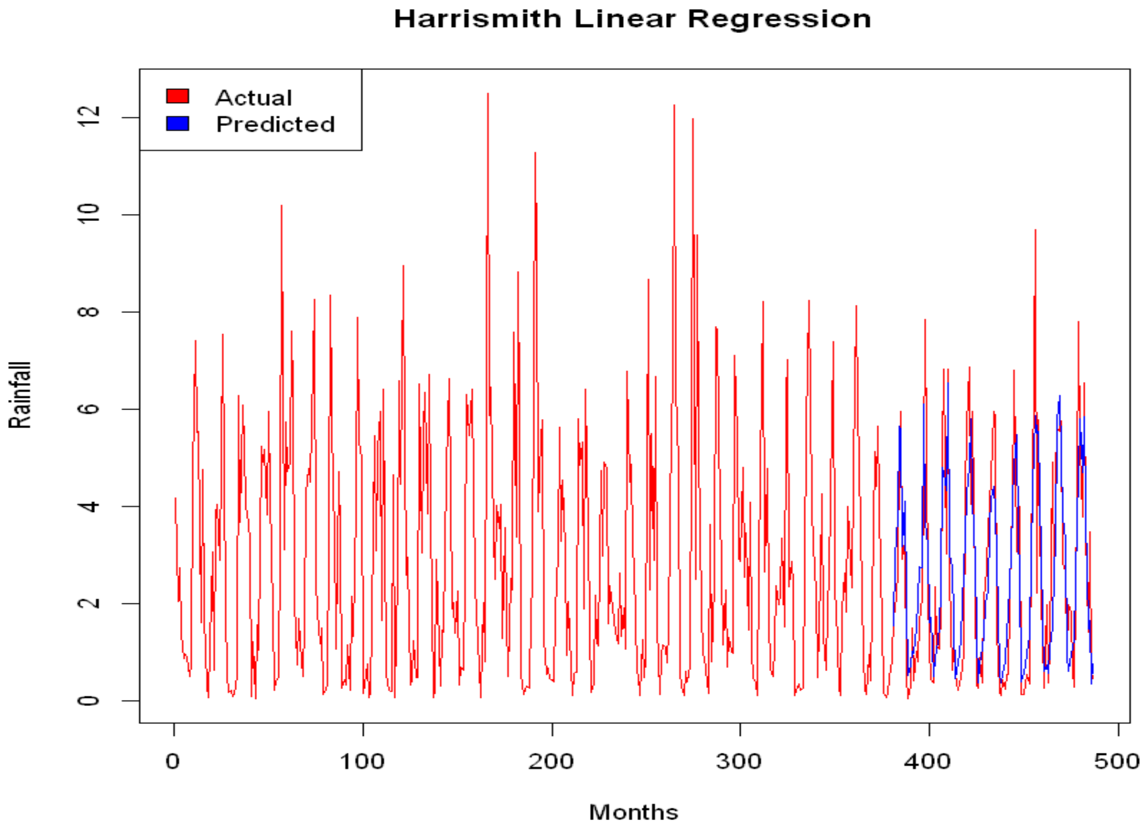
The subtropical highland climate with dry winter, Harrismith in the Free State Province, Johannesburg in the Gauteng Province, and Newcastle in KwaZulu Natal Province were selected. This climatic condition is mostly found around the centre of South Africa. Table 5.2 shows the evaluation metrics for the locations using the four models employed in this study. Apart from Johannesburg, the other two locations had high correlation coefficient and coefficient of determination. For all four models, the correlation coefficient in Harrismith was not less than 0.80. It corresponds to 0.84, 0.82, 0.82, and 0.82 for support vector machine, linear regression, random forest, and ridge and lasso. All models apart from linear regression had a coefficient of determination of 0.68 in Harrismith while that of linear regression was 0.89. Random forest and the support vector machine were very similar for all evaluation metrics for Newcastle. Ndlovu et al (2021) performed an assessment on the impact of climate variability change in six cities of KwaZulu Natal province including Newcastle due to the risk rainfall and air temperature variability pose to environmental change. Their result showed that Newcastle experience distinct changes in both inter and intra-seasonal rainfall fluctuations. Their result also revealed a decreasing trend in the number of rainy days from 1986 to 2016 indicating the necessity for proper planning. They obtained a coefficient of determination of 0.32 which is lower than what was obtained in this study. In Johannesburg, the support vector machine as well as ridge and lasso had close values for the evaluation metrics, the same with linear regression and random forest also in Johannesburg. Gidey and Mhangara (2023) used random forest to analyse the impact of land use change on surface water resources in Johannesburg and obtained a correlation coefficient of 0.60. However, Obiora et al (2020) suggested the use of support vector machine for solar irradiance prediction in Johannesburg. It is suggested that for better predictive performance, other atmospheric variables and machine learning models be explored in Johannesburg.

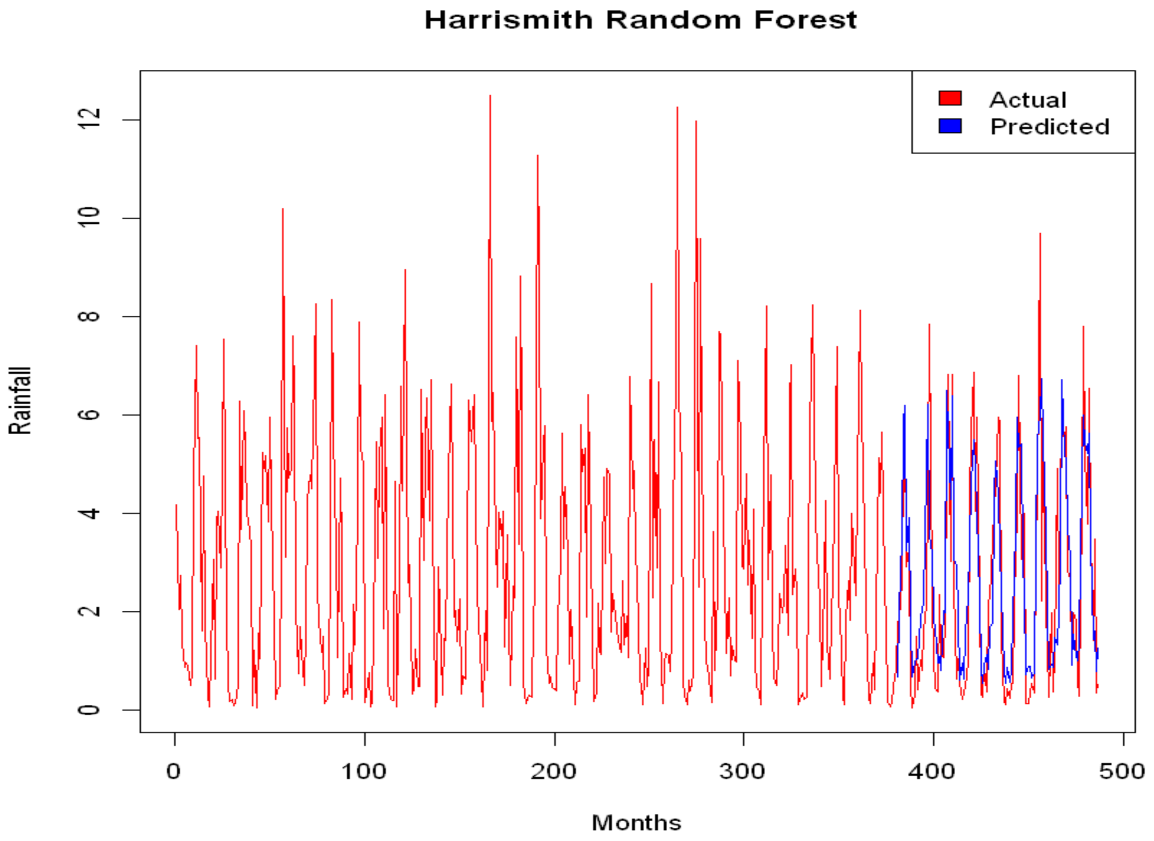
Figure 5.2 shows the temporal variation in rainfall with predicted values. The result indicated that all model correctly models the trend in rainfall variation from one season to another. High rainfall in 2021 were not accurately modelled by all models in Harrismith with random forest being the closest in estimating the amount of rainfall received. In Newcastle, the models also could not correctly estimate the amount of rainfall received in 2022 in Newcastle, however, they performed better in other years. This same trend can also be seen in Johannesburg for all models.

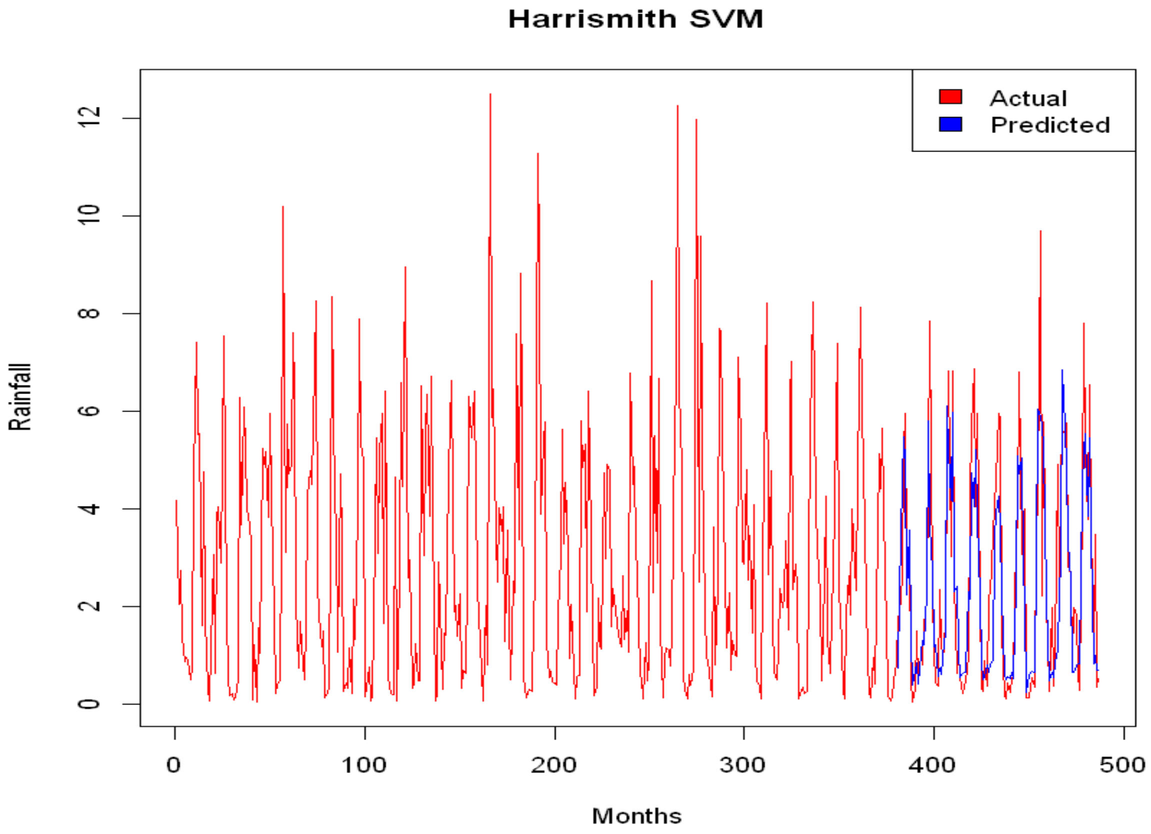
The heatmap showing the correlation between rainfall and other atmospheric parameters in Harrismith showed high correlation coefficients of 0.78, 0.72, 0.61, and 0.50 with cloud cover, water vapour, dew point, temperature, and relative humidity respectively, while that of rainfall and wind speed is -0.42. There was also a negative correlation coefficient of -0.31 and -0.019 between rainfall and wind speed in Newcastle and Johannesburg respectively. Cloud cover had the best correlation coefficient of 0.69 with rainfall in Newcastle, closely followed by 0.67, 0.64, 0.62 for water vapour dew point, and temperature while that of relative humidity corresponds to 0.42. In Johannesburg, only cloud cover had a correlation coefficient greater than 0.50 corresponding to 0.51. Dewpoint and water vapour both had a correlation coefficient of 0.50 with rainfall. The correlation between other atmospheric variables and rainfall was below 0.50 with the coefficients of relative humidity, temperature, and wind speed corresponding to 0.27, 0.40, and -0.019 respectively.

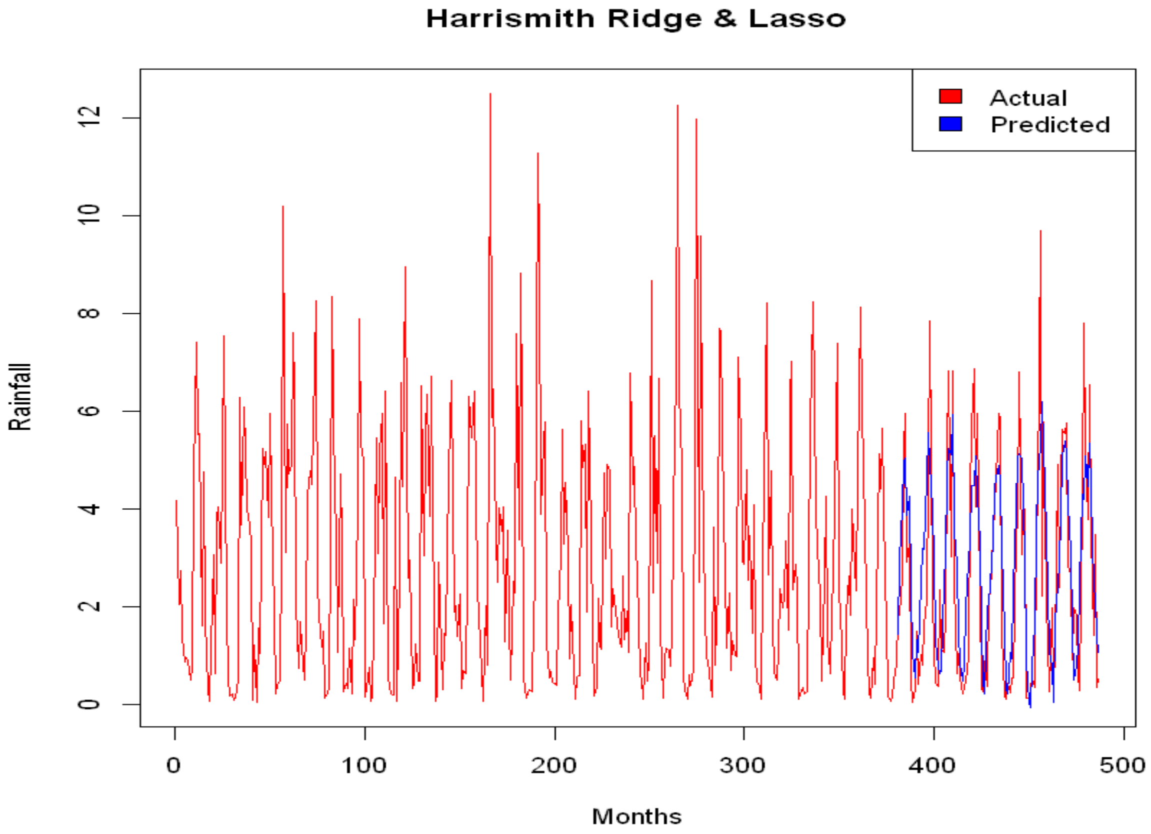
Table 5. 2: Table showing models evaluation metrics for subtropical highland with dry winter climate classification.

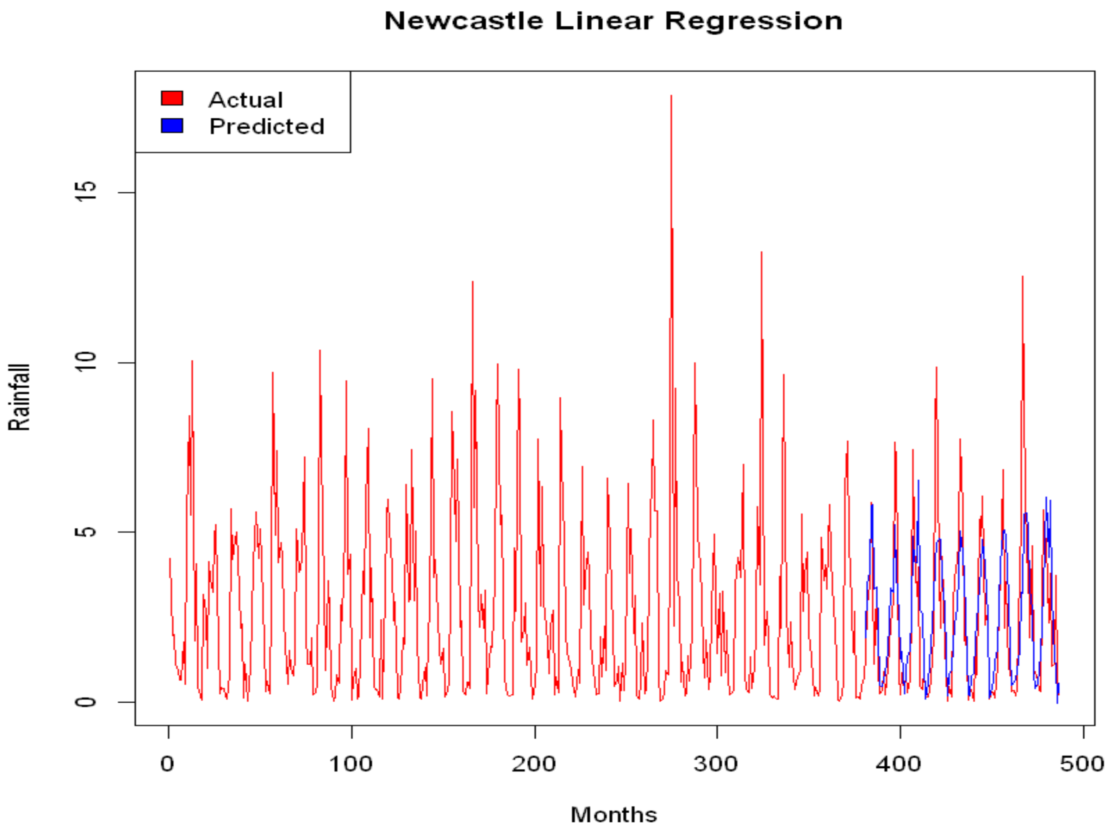
Linear Regression					
Harrismith	0.80	1.25	1.12	0.82	0.89
Johannesburg	1.17	2.26	1.50	0.50	0.49
Newcastle	0.97	1.77	1.34	0.69	0.86
Random Forest					
Harrismith	0.96	1.63	1.28	0.82	0.68
Johannesburg	1.04	1.94	1.39	0.42	0.08
Newcastle	1.08	2.53	1.59	0.77	0.59
Support Vector Machine					
Harrismith	0.86	1.59	1.26	0.84	0.68
Johannesburg	0.95	1.66	1.29	0.54	0.21
Newcastle	1.00	2.54	1.60	0.78	0.59
Ridge and Lasso					
Harrismith	1.13	2.10	1.45	0.80	0.68
Johannesburg	0.99	1.60	1.27	0.50	0.34
Newcastle	1.35	3.43	1.85	0.70	0.60

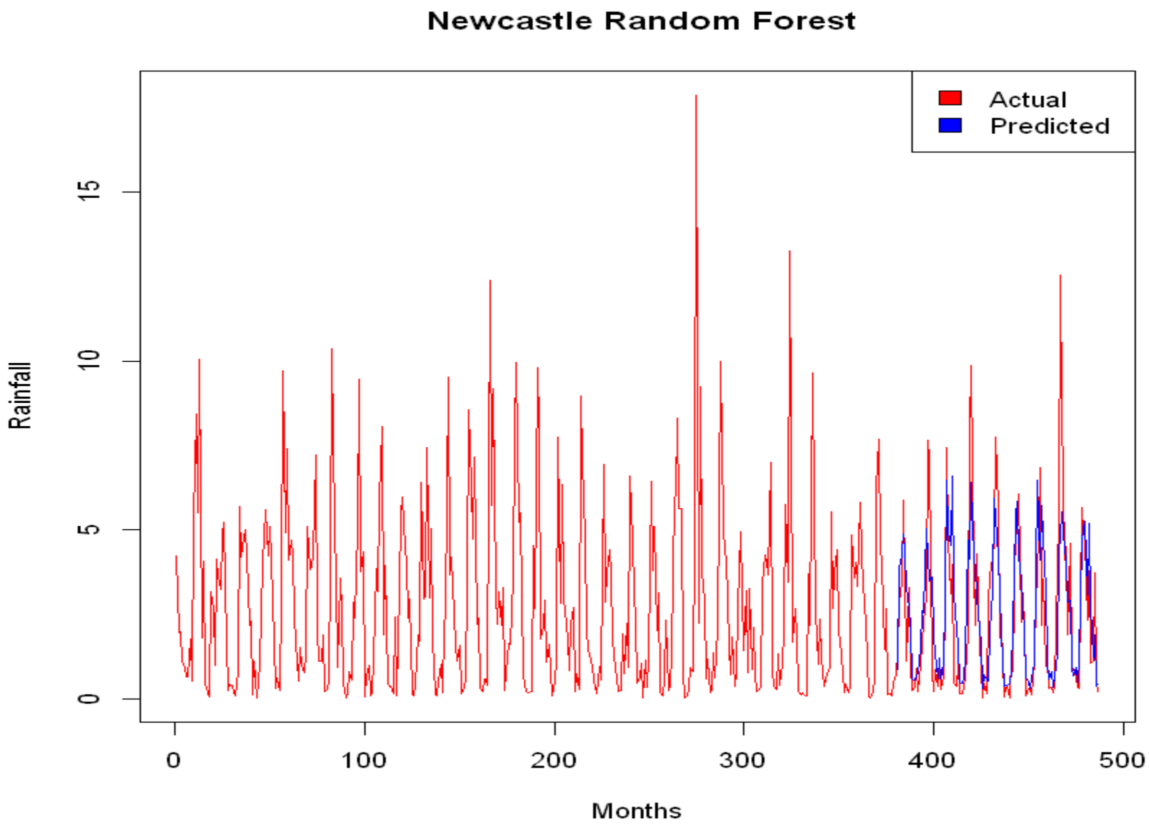

Harrismith Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
0
100
200
300
400
500
Months

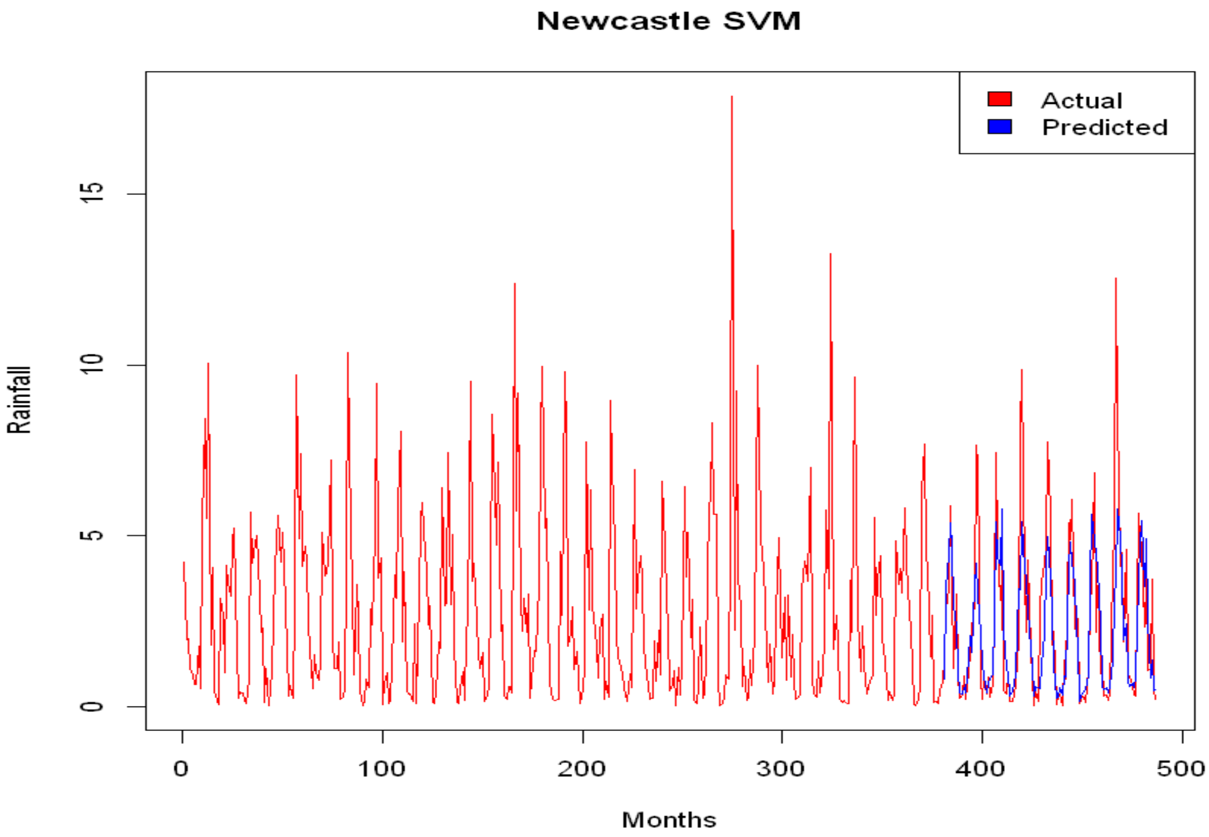

Harrismith Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
0
100
200
300
400
500
Months

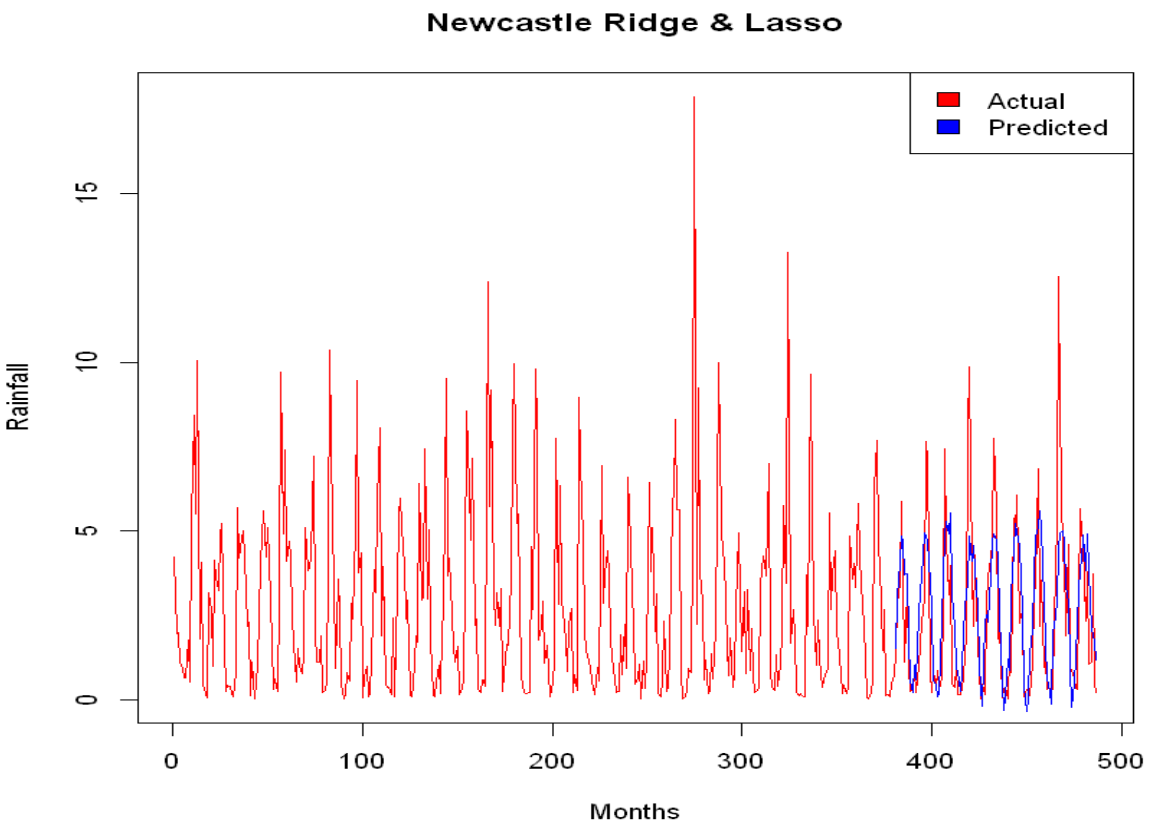

Harrismith SVM
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
0
100
200
300
400
500
Months

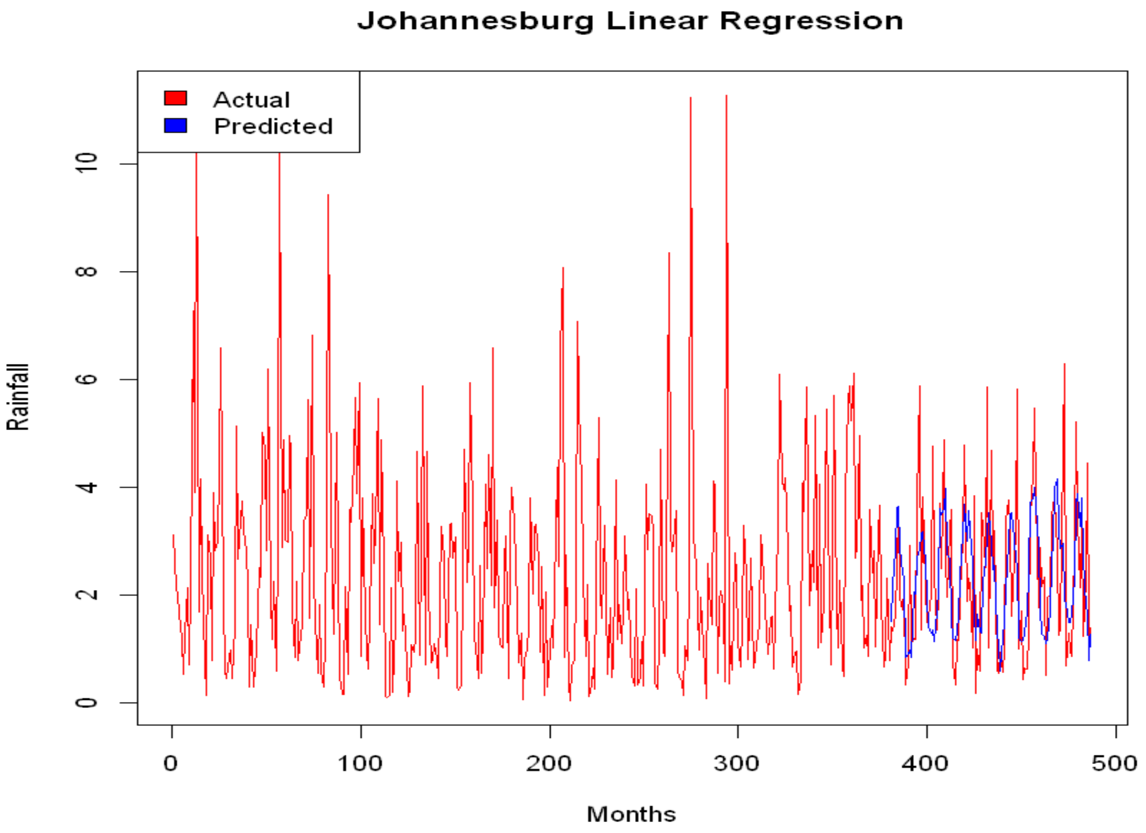

Harrismith Ridge & Lasso
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
0
100
200
300
400
500
Months

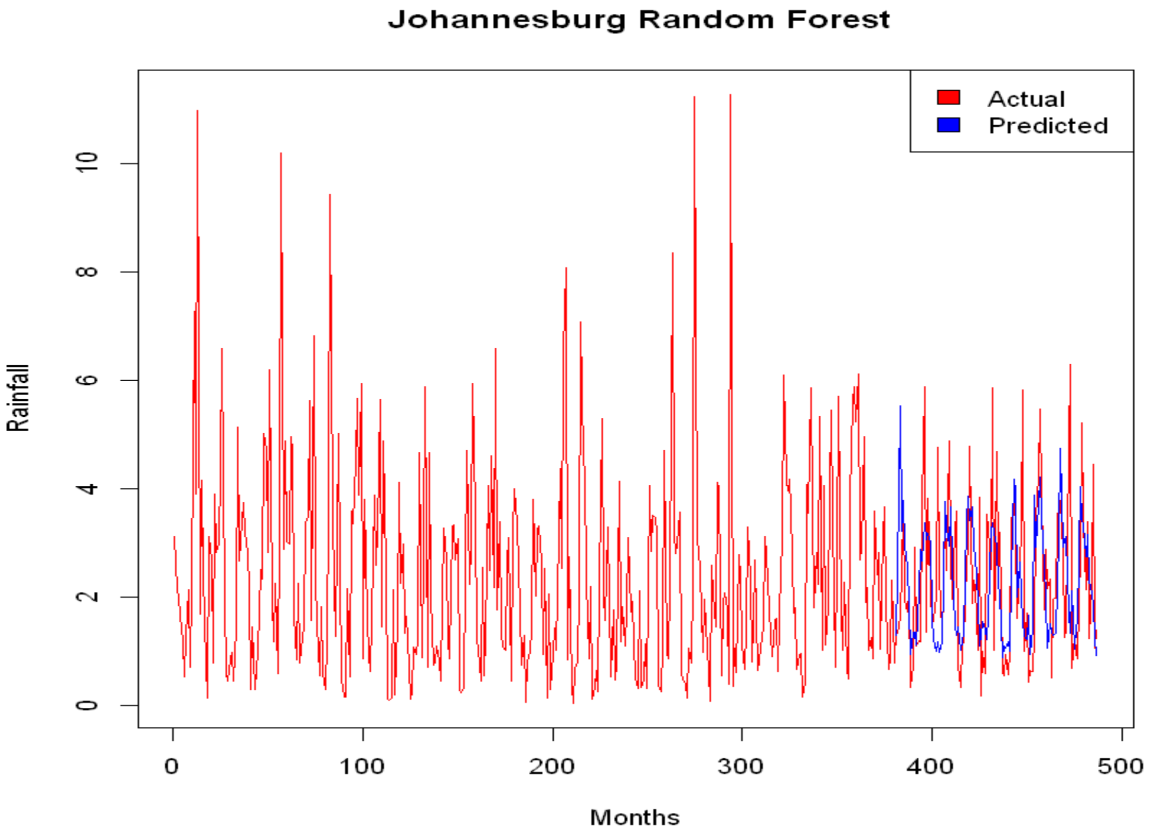

Newcastle Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months


Newcastle Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months


Newcastle SVM
Actual
Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months


Newcastle Ridge & Lasso
Actual
Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months


Johannesburg Linear Regression
Actual
Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months


Johannesburg Random Forest
Actual
Predicted
Rainfall
0
2
4
6
8
10
0
100
200
300
400
500
Months

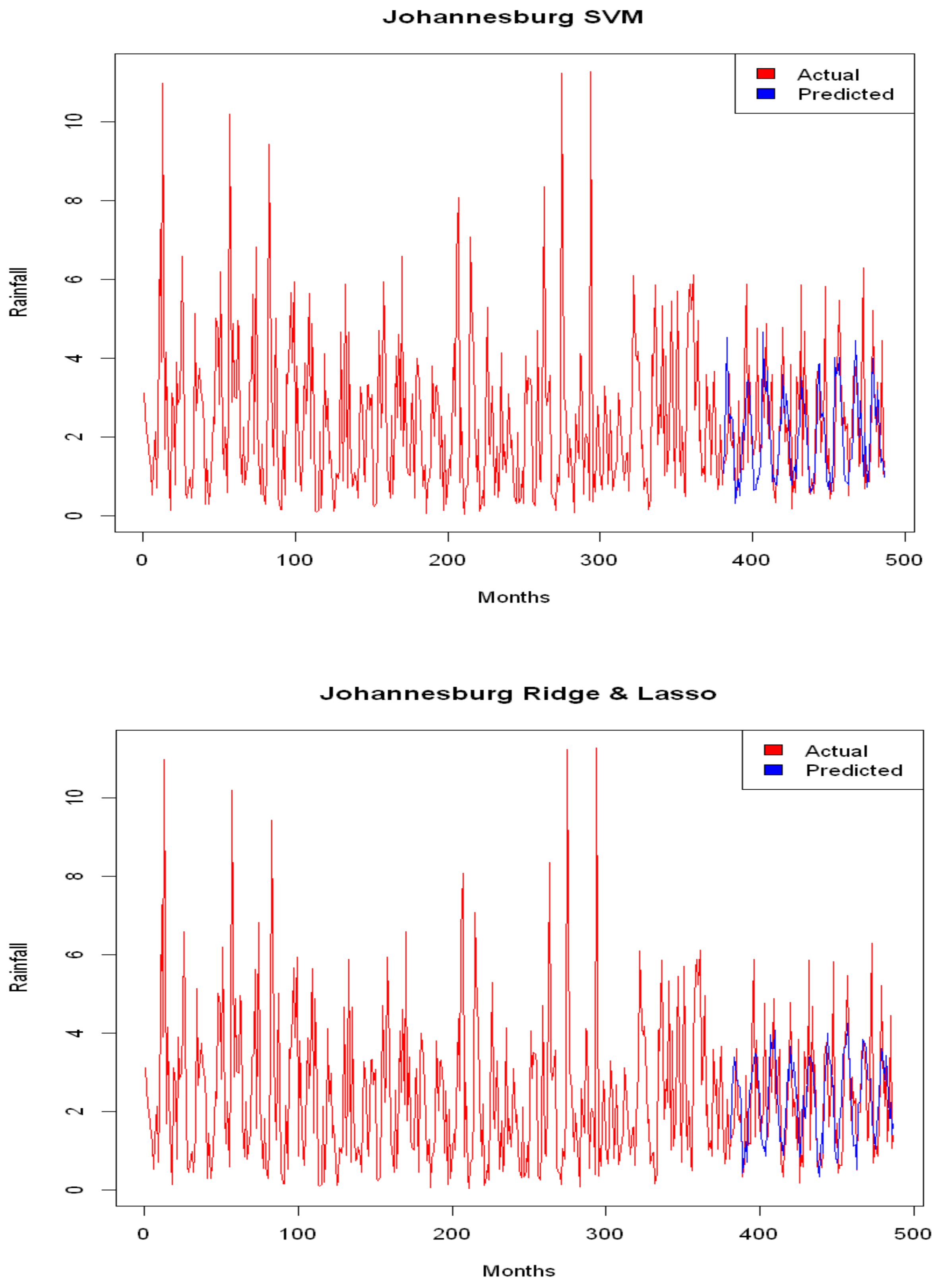


Figure 5. 2: Figures showing the predicted and actual rainfall for models used and locations under the subtropical highland with dry winter climate classification.

Chapter 6: Results and Discussion on Subtropical Dry Climate Classification

6.0 Subtropical Dry

6.1 Humid Subtropical without Dry Season

This chapter presents the models' performance for rainfall prediction for the subtropical dry climate. This climate condition is sub-divided into three: the humid subtropical without dry season, temperate oceanic without dry season, and the warm and dry summer. For the humid subtropical without dry season, three locations were selected: Durban, Port Edwards, and Richards Bay. Table 6.1 presents the evaluation metrics for different machine learning models. For all models, the correlation coefficient in Port Edward was higher than other locations. The coefficients 0.69, 0.66, 0.64, 0.63 corresponded to support vector machine, random forest, ridge and lasso, and linear regression. However, the coefficient of determination was below 0.50 for all models in Port Edwards apart from linear regression which had a value of 0.79. Ridge and Lasso had the highest coefficient of determination in Durban corresponding to 0.89 and correlation coefficient of 0.60. High coefficient of determination value of 0.76 was obtained with linear regression in Durban and a correlation coefficient of 0.53. using the backpropagation neural network, Ahuna et al (2019) predicted rain attenuation in Durban and obtained a correlation coefficient of 0.83 for their model. This model performed better than the models used for this research. For their work, they made use of only rainfall data ranging from drizzle to super storms. On the east coast of Durban and Port Elizabeth, Ingreso (2022) developed an open loop nonlinear autoregressive network with exogenous inputs to predict sea level using five atmospheric inputs. She observed that their model performed better when all inputs were considered than when anyone was left out of the model. She obtained a correlation coefficient of 0.85 for her prediction.

In Richards Bay, all model performed badly with respect to the correlation coefficient as they were all below 0.50 for all models. Similarly, the coefficient of determination value was also below 0.40 for all models except linear regression whose value corresponds to 0.60. This result shows that either the support vector machine or linear regression can be used to predict rainfall in Port Edwards while the ridge and lasso model is advisable to be used for prediction in Durban.

Figure 6.1 shows the model performance in rainfall prediction for the selected locations. All models predicted the seasonality of rainfall accurately as well as estimated values. None of the models could accurately predict the flood that took place in Durban in 2022. The extreme rainfall experienced in Durban in 2022 has been well documented. Mashao et al (2023) reported that this was related with the mid-tropospheric cut-off low pressure system which affected the entire east coast of South Africa. The result also shows that underestimation of rainfall was more for the models in Port Edwards, perhaps due to the low values of the coefficient of determination in this region. The same can be observed in Richards Bay with low estimation of amount of rainfall received but accurate prediction of its seasonality. The result also showed that although the ridge and lasso model performed best in Durban, the support vector machine model was close, and it performed better in Port Edwards and Richards Bay.

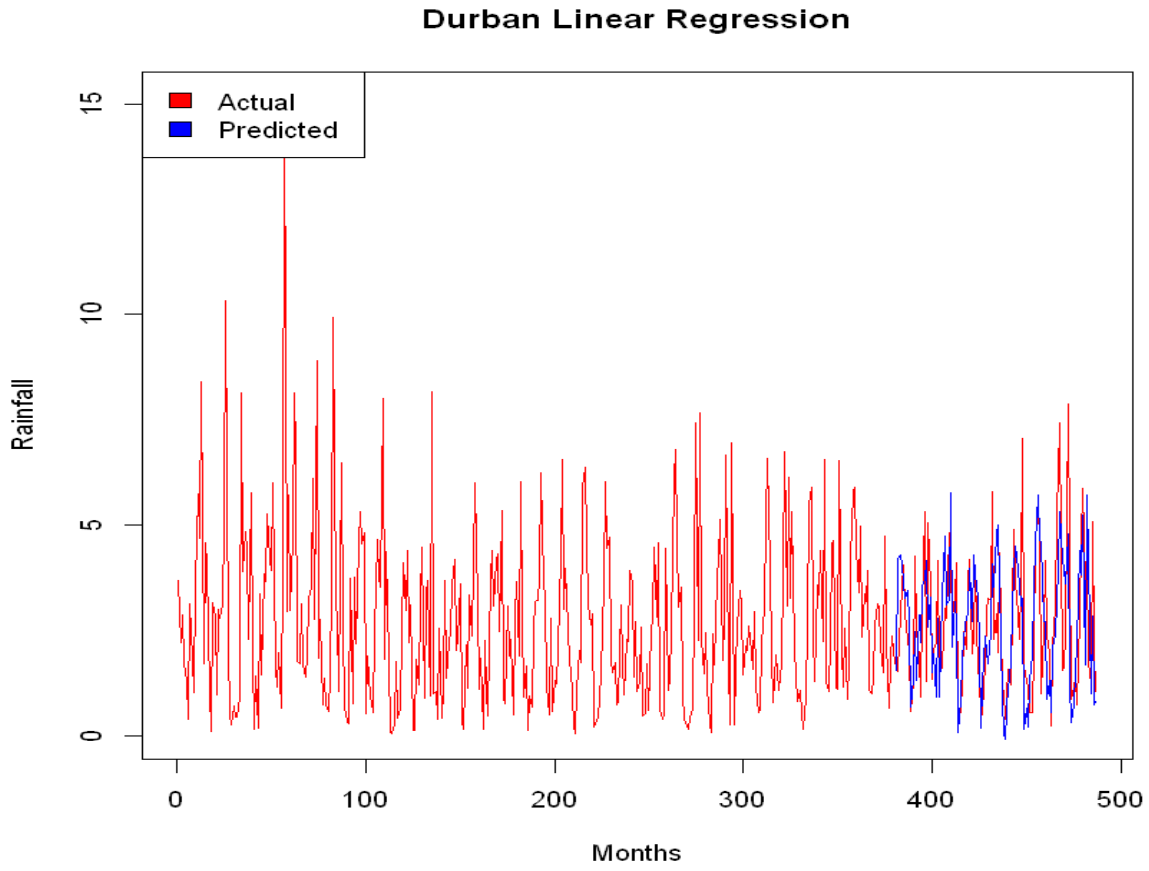
The heatmap on the correlation between rainfall and other atmospheric parameters for humid subtropical without dry season is seen in Appendix A. The result shows that in Durban, only cloud cover had a correlation above 0.60 corresponding to 0.65. However, other atmospheric variables such as dewpoint, relative humidity, and water vapour also had high correlation corresponding to 0.51, 0.55, and 0.58 respectively. Temperature and wind speed had correlation coefficient less than 0.50 with rainfall corresponding to 0.45 and 0.10 respectively in Durban. In Port Edwards, the highest correlation between rainfall and other atmospheric variables is with cloud cover corresponding to 0.69 followed by water vapour corresponding to 0.64. Dew point and relative humidity also had high correlation corresponding to 058 and 0.59 respectively while temperature and wind speed had correlations of 0.49 and -0.051 with rainfall.

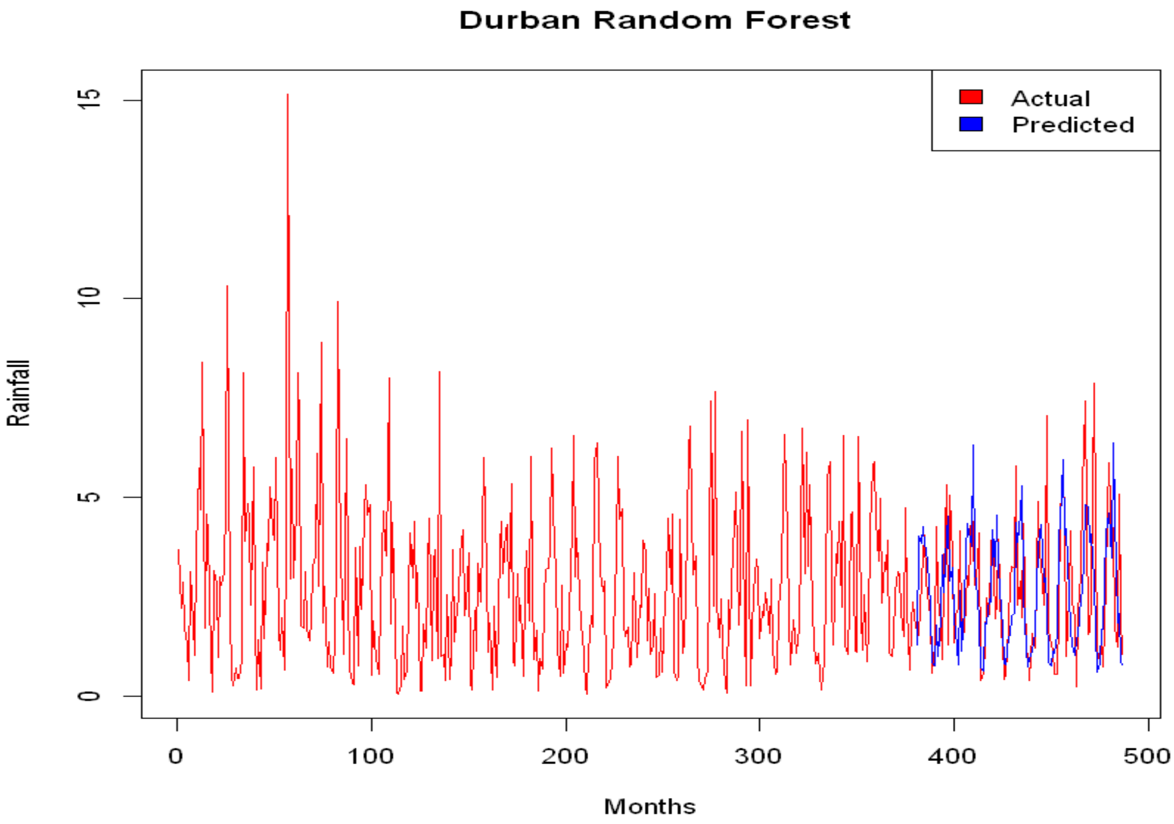
However, in Richards Bay, all atmospheric variables showed negative correlation with rainfall. The values of this correlation are -0.36, -0.25, -0.48, -0.30, -0.40, and -0.23 for dewpoint, relative humidity, temperature, cloud cover, water vapour, and wind speed respectively. The reason for this result is yet to be determined. As in Alexander Bay and Port Elizabeth, high correlation can be seen among other atmospheric variables especially with dew point. Relative humidity, temperature, cloud cover, and water vapour all correlated with dewpoint with coefficients corresponding to 0.79, 0.74, 0.59, and 0.86 respectively.

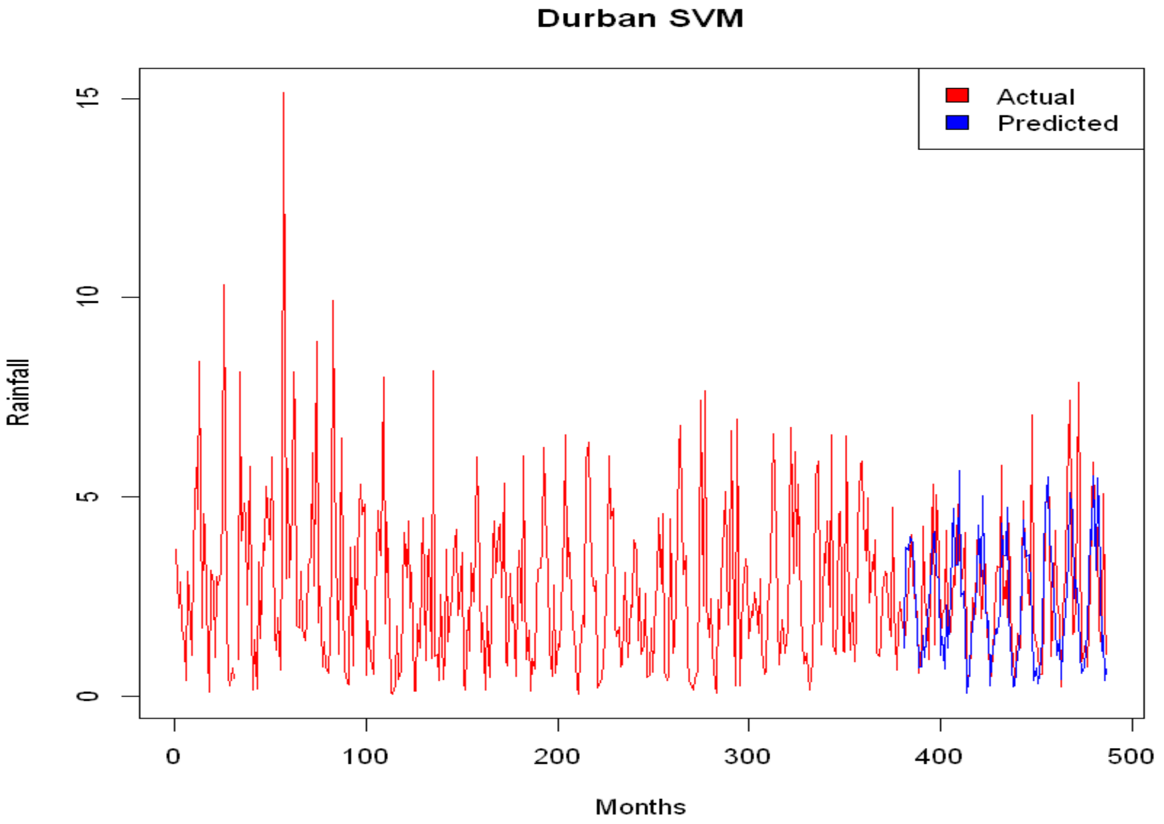
Table 6. 1: Table showing models evaluation metrics for humid subtropical without dry season climate classification.

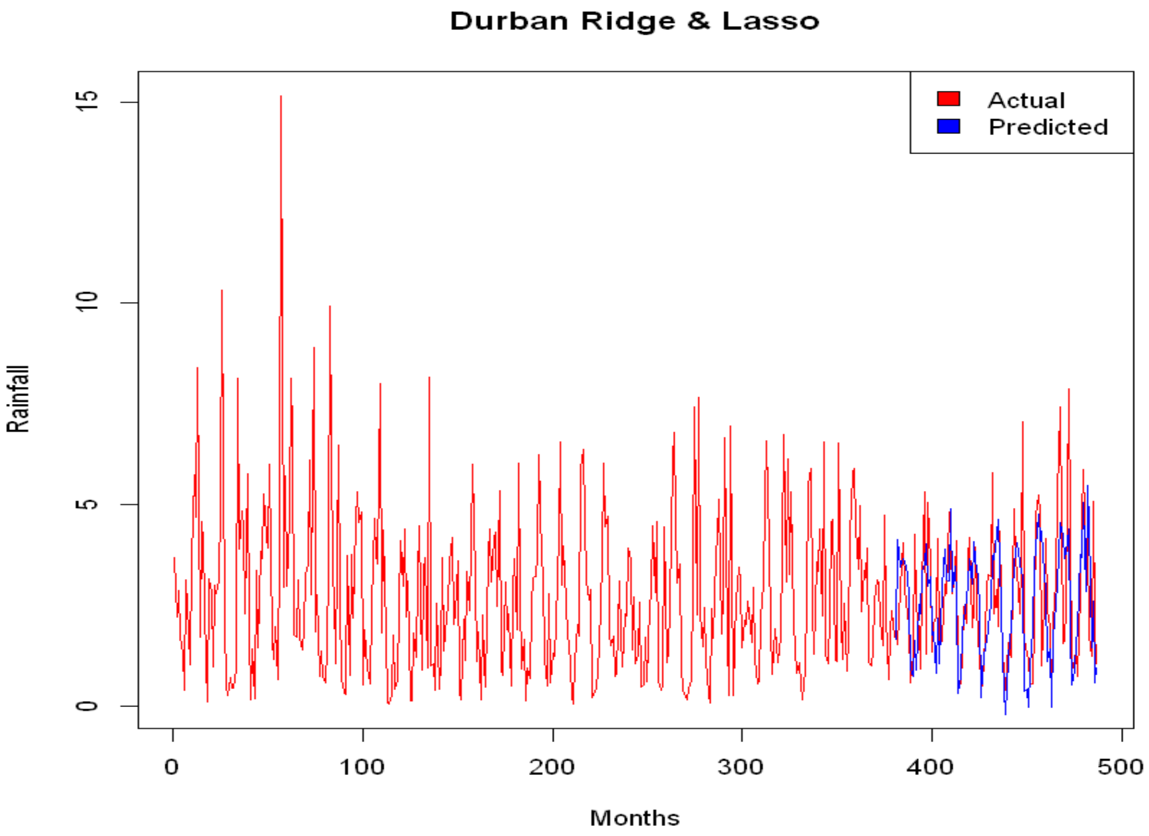
Linear Regression					
Durban	1.07	2.02	1.42	0.53	0.76
Port Edward	0.98	1.63	1.28	0.63	0.79
Richards Bay	2.46	10.61	3.26	0.37	0.60
Random Forest					
Durban	1.17	2.40	1.55	0.51	0.13
Port Edward	1.07	1.96	1.40	0.66	0.42
Richards Bay	0.60	0.67	0.82	0.41	0.13
Support Vector Machine					

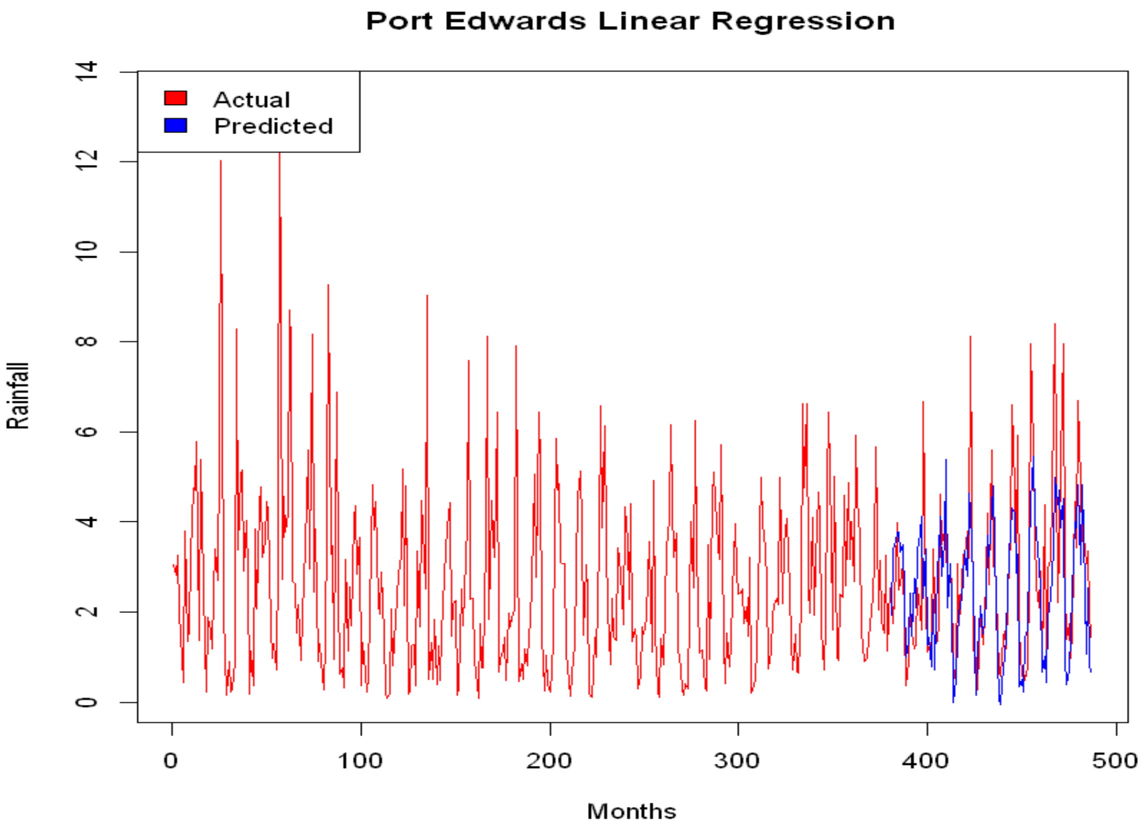
Durban	1.11	2.03	1.42	0.58	0.26
Port Edward	1.03	2.02	1.42	0.69	0.39
Richards Bay	0.56	0.66	0.81	0.44	0.15
Ridge and Lasso					
Durban	1.08	1.92	1.38	0.60	0.89
Port Edward	1.03	2.02	1.42	0.64	0.37
Richards Bay	0.61	0.67	0.82	0.37	0.35

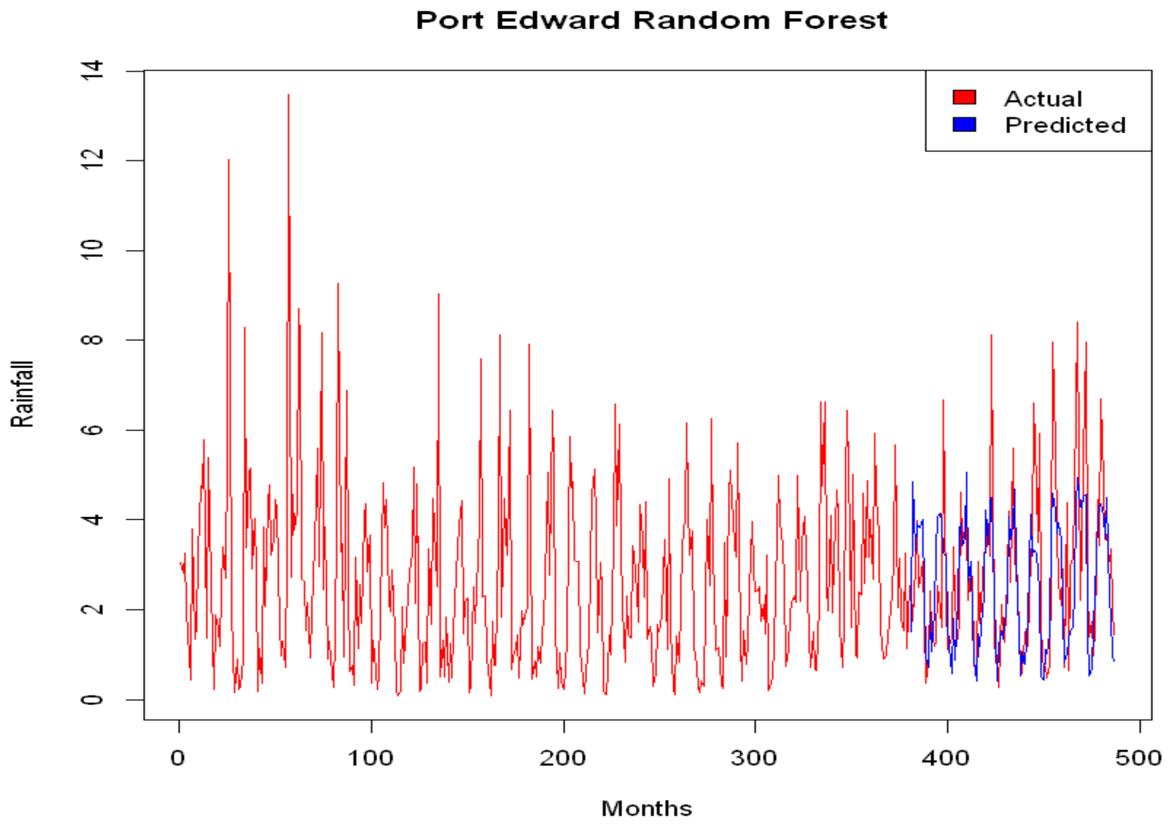


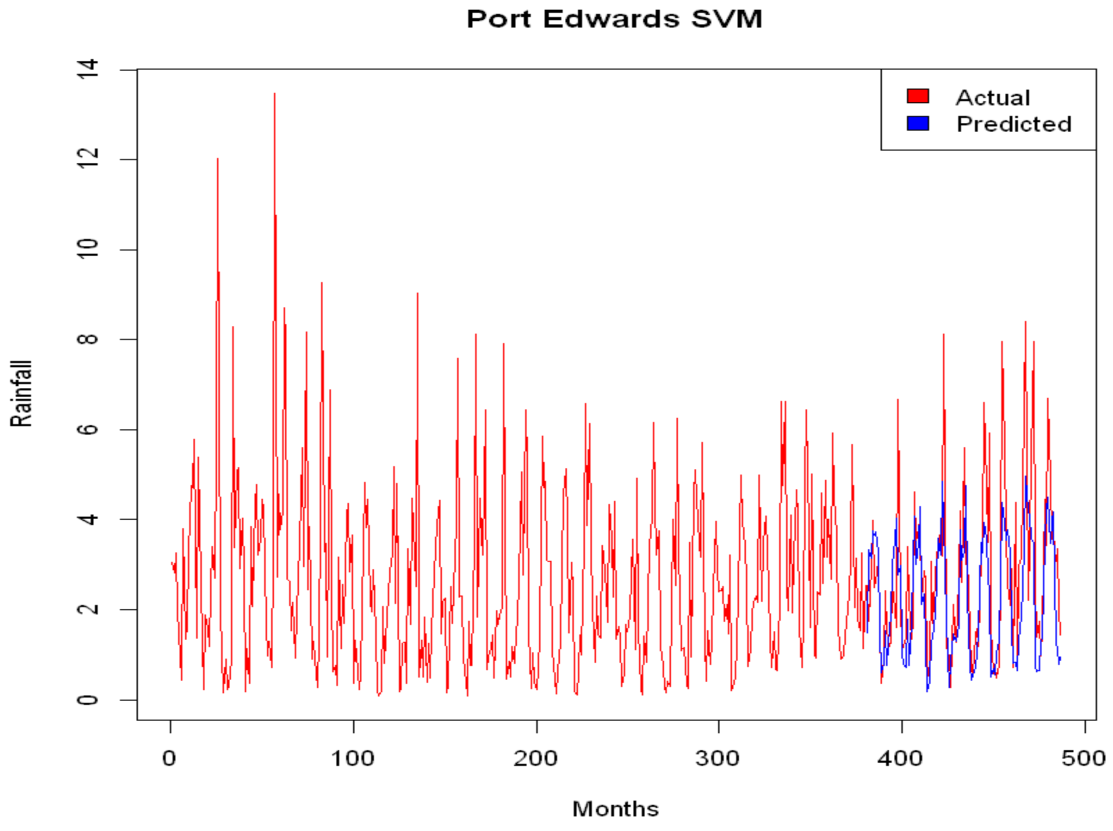

Durban Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months

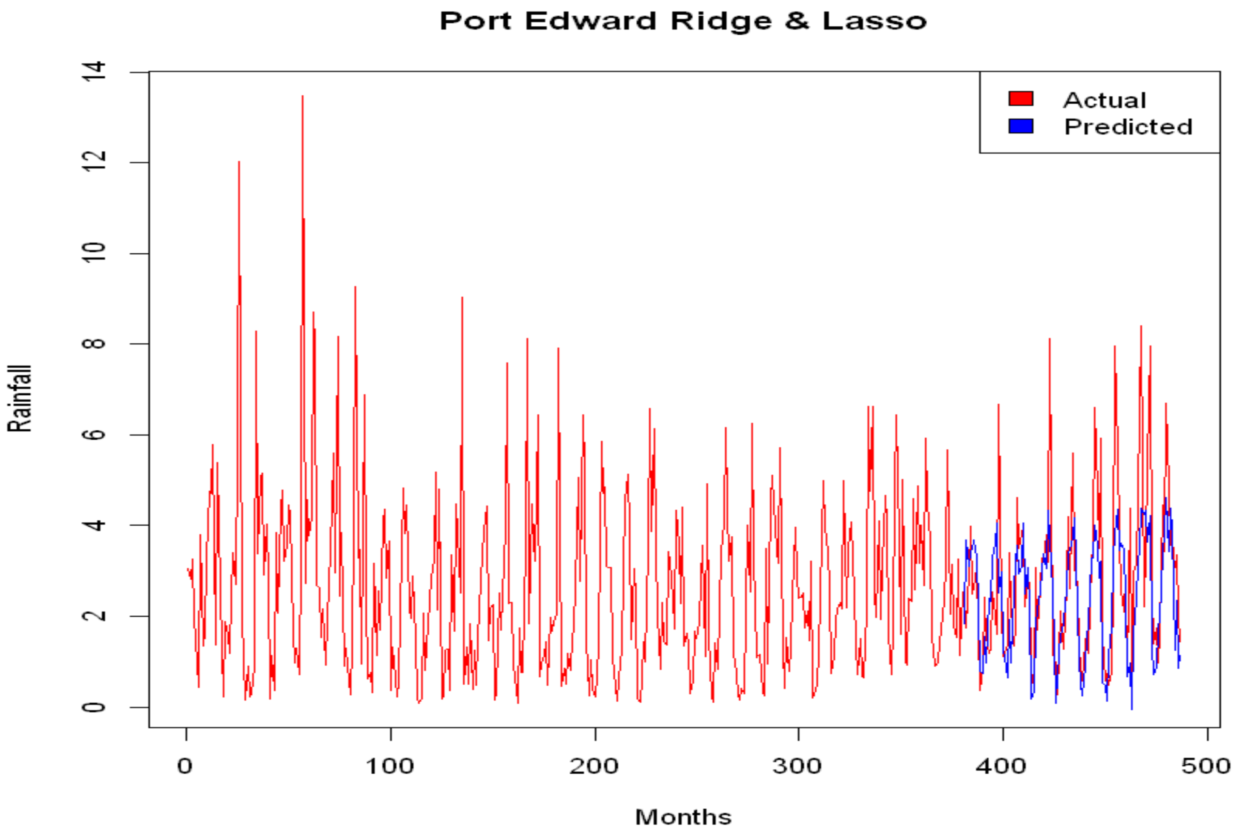

Durban SVM
🟥 Actual
🟦 Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months

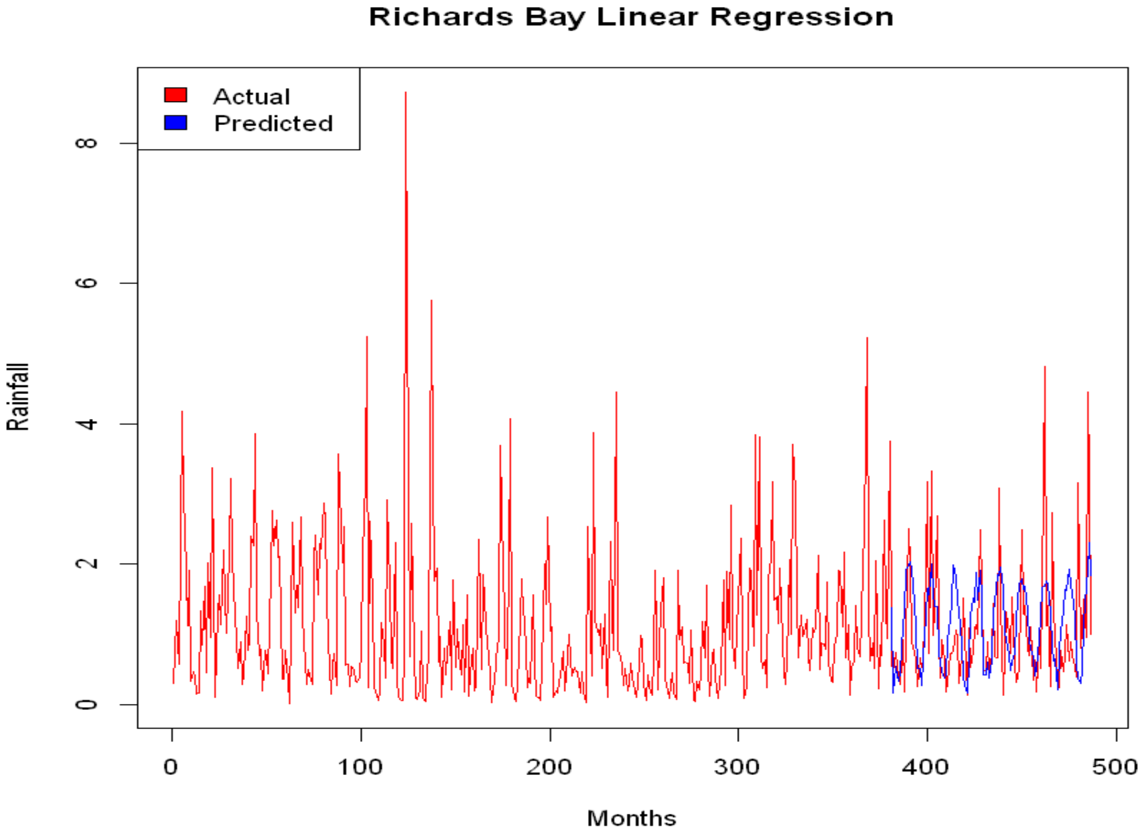

Durban Ridge & Lasso
🟥 Actual
🟦 Predicted
Rainfall
0
5
10
15
0
100
200
300
400
500
Months

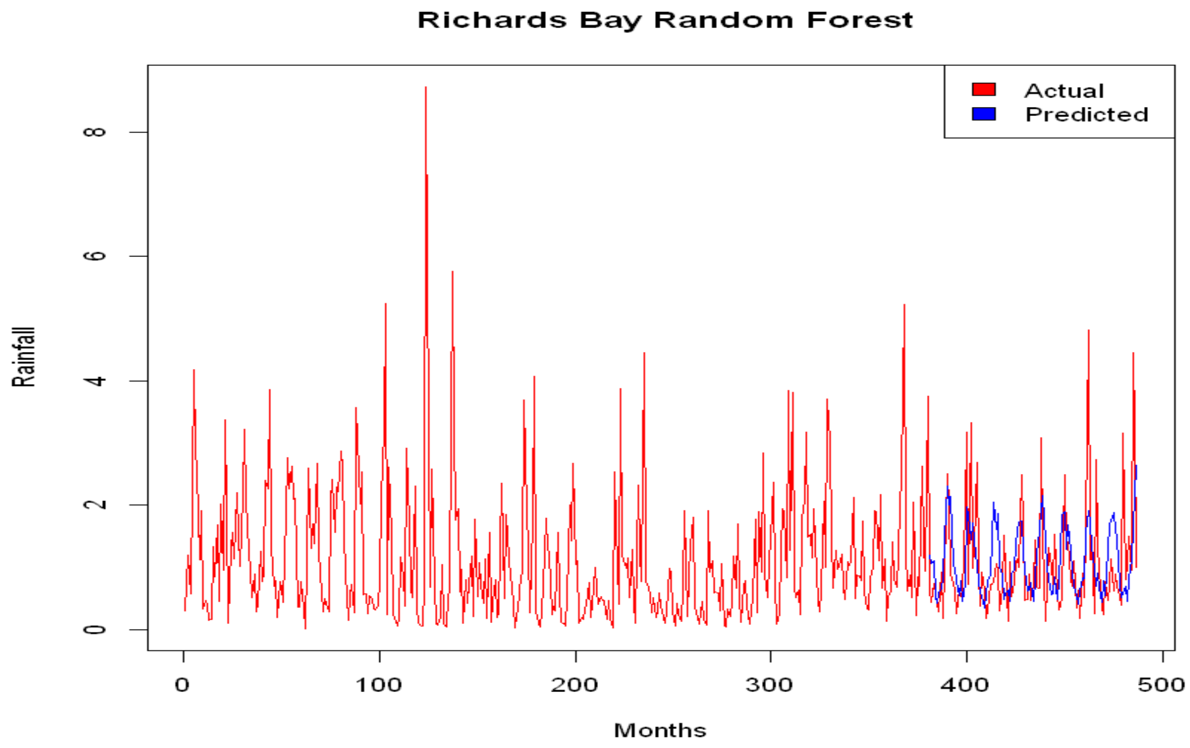

Port Edwards Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
14
0
100
200
300
400
500
Months


Port Edward Random Forest
Rainfall
0
2
4
6
8
10
12
14
Months
0
100
200
300
400
500
🟥 Actual
🟦 Predicted


Port Edwards SVM
Rainfall
0
2
4
6
8
10
12
14
Months
0
100
200
300
400
500
🟥 Actual
🟦 Predicted


Port Edward Ridge & Lasso
Rainfall
0
2
4
6
8
10
12
14
Months
0
100
200
300
400
500
🟥 Actual
🟦 Predicted


Richards Bay Linear Regression
Rainfall
0
2
4
6
8
Months
0
100
200
300
400
500
🟥 Actual
🟦 Predicted



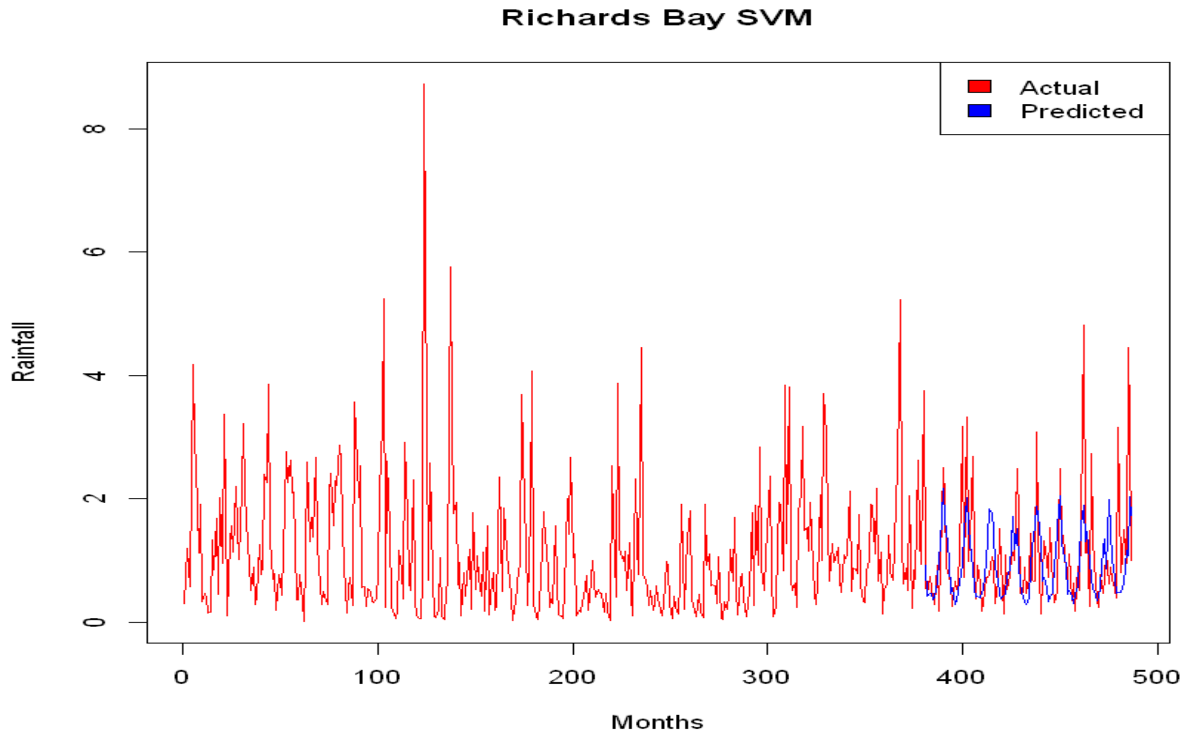


Figure 6. 1: Figures showing the predicted and actual rainfall for models used and locations under the humid subtropical without dry season climate classification.

6.2 Temperate Oceanic without Dry Season

The temperate oceanic climate classification without dry season is mostly found in the Eastern Cape Province of South Africa. This is located in the southern part of the country. The locations, East London, George, and Mthatha were selected for this section. While East London and Mthatha are situated in the Eastern Cape Province, George is in the Western Cape Province. Table 6.2 presents the evaluation metrics for the four models used in this study. The result shows low correlation coefficient and coefficient of determination for all models in George with both metrics having a value below 0.40 for all models. The mean average error, mean square error and root mean square error for all models in George were within the same range for all models. The models performed best in Mthatha with a coefficient of determination and correlation of 0.82 and 0.72 with linear regression, 0.51 and 0.74 respectively for random forest. For support vector machine, the values correspond to 0.46 and 0.73 while ridge and lasso had values corresponding to 0.45 and 0.72 for the R-squared and correlation coefficient. For East London, all values of the coefficient of determination were less than 0.45 while the correlation coefficient ranged from 0.53 using linear regression to 0.62 with support vector machines. Ridge and lasso and random forest had values corresponding to 0.58 and 0.59 respectively.

Figure 6.2 shows the performance of the models in predicting and estimating the amount of rainfall received for the three East London, George, and Mthatha. The models performed well with predicting and estimating amount of rainfall received in East London with underestimation in 2023 and 2023. The underestimation was more pronounced with random forest in Mthatha while ridge and lasso as well as the support vector machine performed better in rainfall estimation. This study in Mthatha is particularly important as it has always been affected by drought. Nkamisa et al (2022) analysed the trends of recurrences of drought in Mthatha and other locations in Eastern Cape and reported its high intensity. Mahlalela et al

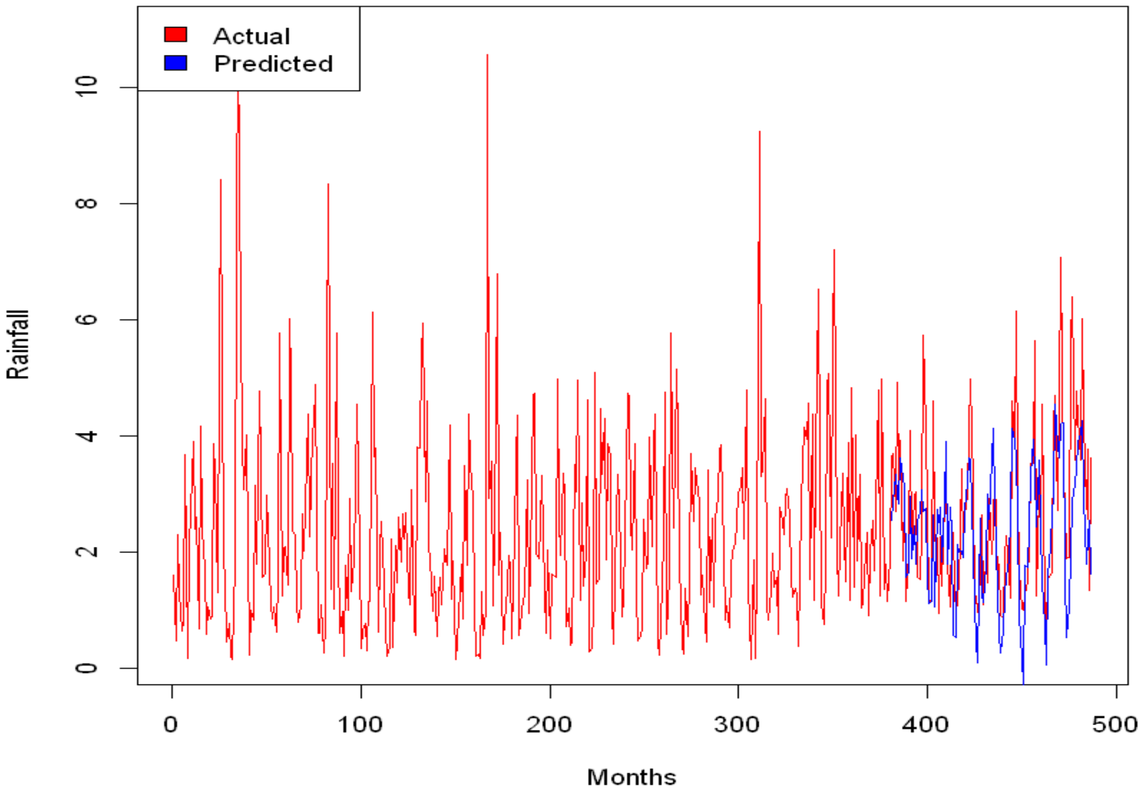
(2020) reported decreasing trend in the amount of rainfall received as well as in the number of rainfall days. In George, the models seem to accurately predict both seasonality and estimation of rainfall received for all years except 2023 where there was an unusual increase in the amount of rainfall received. Values of rainfall below 3mm reveals the perplexing situation in George, a town notable for severe drought. According to Lottering (2015), the drought in George is primarily due to lack of rainfall for a prolonged period. Botai et al (2017) reported that due George's drought, water reservoirs are below 30% capacity resulting in socio-economic impacts.

The heatmap of rainfall correlation with other atmospheric variables under the temperate oceanic without dry season is seen in Appendix A for East London, Mthatha, and George. In East London, only temperature and wind speed had correlation coefficients lower than 0.50 corresponding to 0.36 and -0.20 respectively. Dewpoint, relative humidity, cloud cover, and water vapour had coefficients corresponding to 0.50, 0.51, 0.56, and 0.54 respectively in East London. These values are just about averaged. Compared with dewpoint in East London, its correlation with relative humidity, temperature, cloud cover and water vapour correspond to 0.89, 0.79, 0.57, and 0.98 respectively. Better results are seen in Mthatha compared to East London with regards to correlation between atmospheric variables and rainfall. Most atmospheric variable had correlation coefficients exceeding 0.50 except for wind speed which corresponded to -0.38. For other atmospheric variables, their correlation with rainfall corresponds to 0.678, 0.63, 0.57, 0.73, and 0.70 for dewpoint, relative humidity, temperature, cloud cover, and water vapour. In George, similar pattern observed in Alexander Bay, Port Elizabeth, and Richards Bay is seen as all atmospheric variables had low correlation with rainfall. The coefficients of their correlation with rainfall are 0.10, 0.13, -0.018, 0.25, 0.14, 0.063 for dewpoint, relative humidity, temperature, cloud cover, water vapour, and wind speed respectively. Relative humidity, temperature, and water vapour all had high correlation with dewpoint corresponding to 0.89, 0.79, and 0.98 respectively.

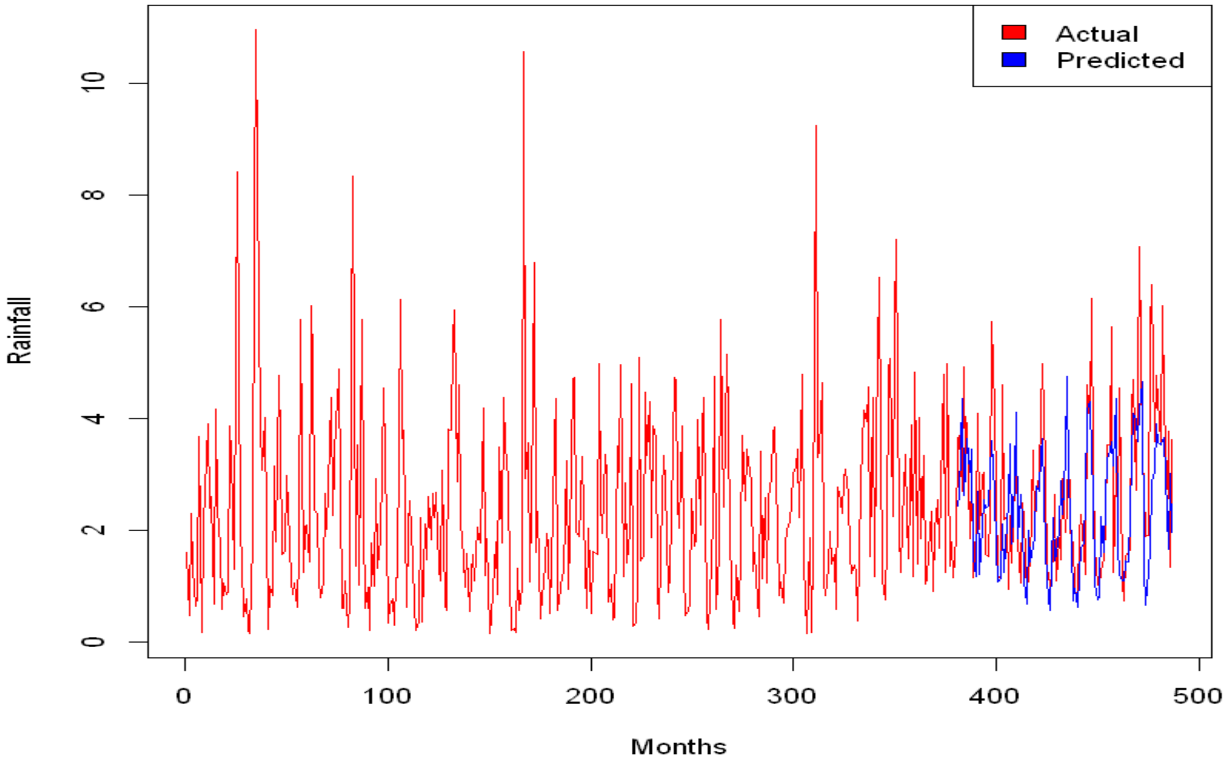
Table 6. 2: Table showing models evaluation metrics for temperate oceanic without dry season climate classification.

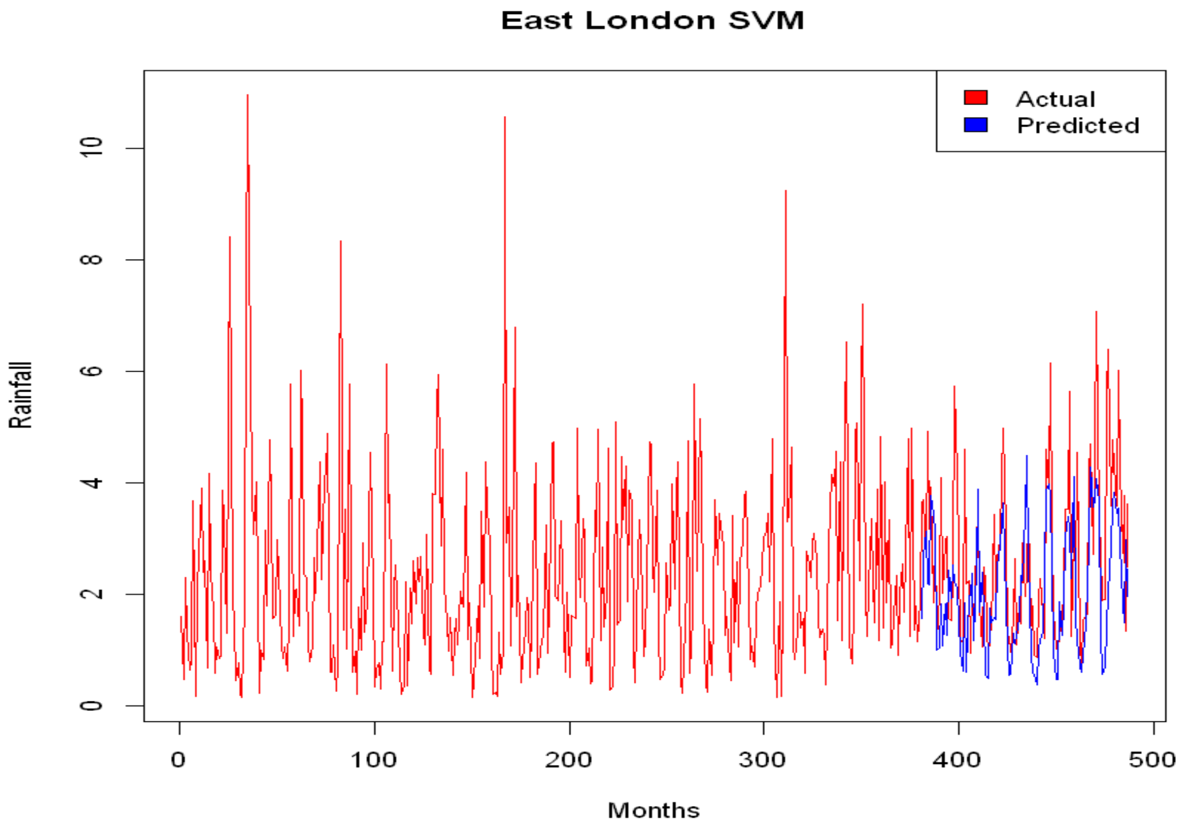
Linear Regression					
East London	0.89	1.26	1.12	0.53	0.43
George	0.61	0.58	0.76	0.03	0.39
Mthatha	0.90	1.25	1.12	0.72	0.82
Random Forest					
East London	0.92	1.52	1.23	0.59	0.27
George	0.67	0.90	0.95	0.30	0.07
Mthatha	1.11	2.44	1.56	0.74	0.51
Support Vector Machine					
East London	0.98	1.70	1.30	0.62	0.19
George	0.60	0.89	0.95	0.38	0.08
Mthatha	1.14	2.73	1.65	0.73	0.46
Ridge and Lasso					
East London	0.94	1.55	1.25	0.58	0.40
George	0.63	0.85	0.92	0.35	0.29
Mthatha	1.14	2.69	1.64	0.72	0.45

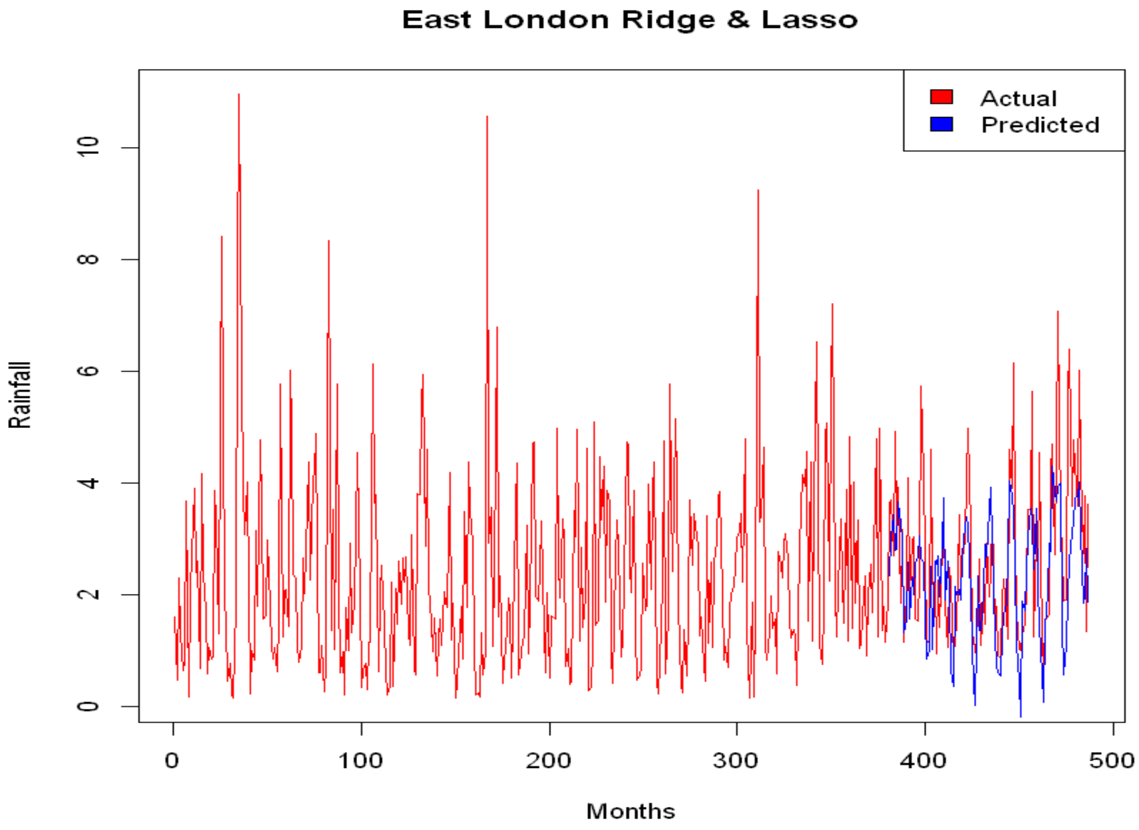
East London Linear Regression

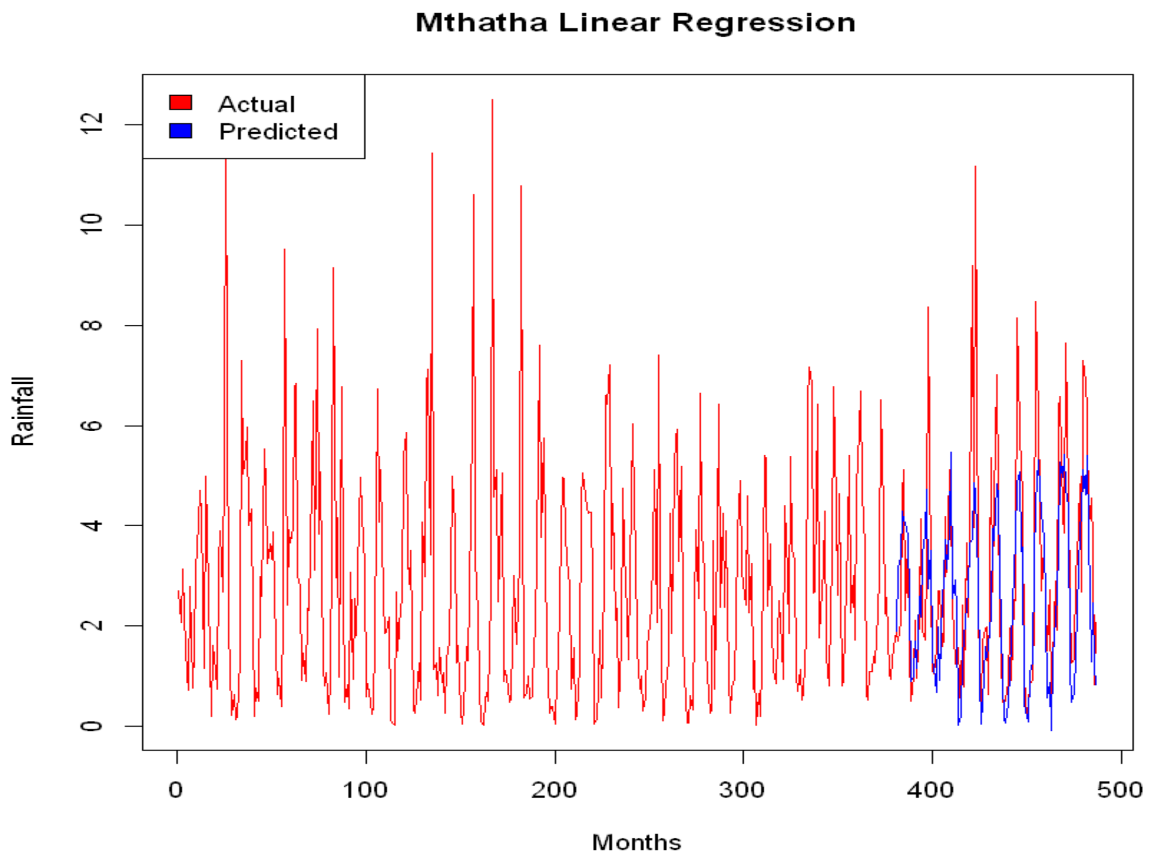


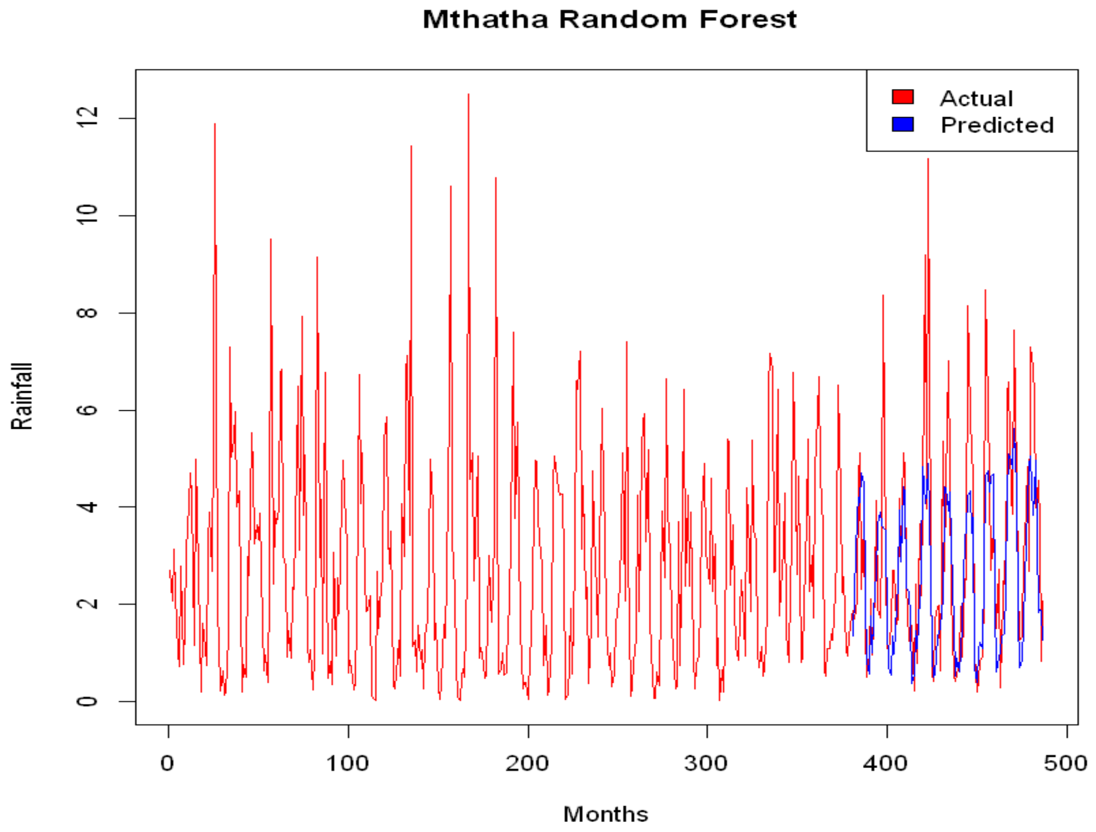
East London Random Forest

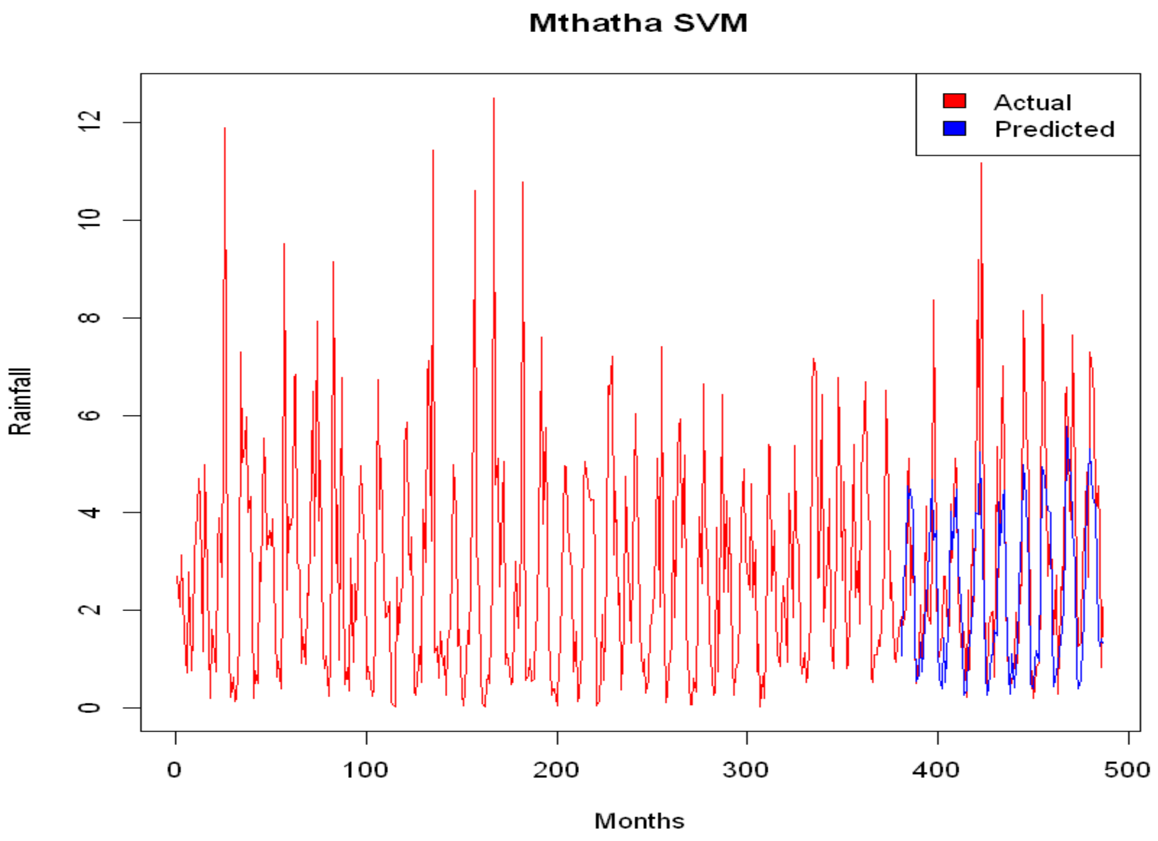


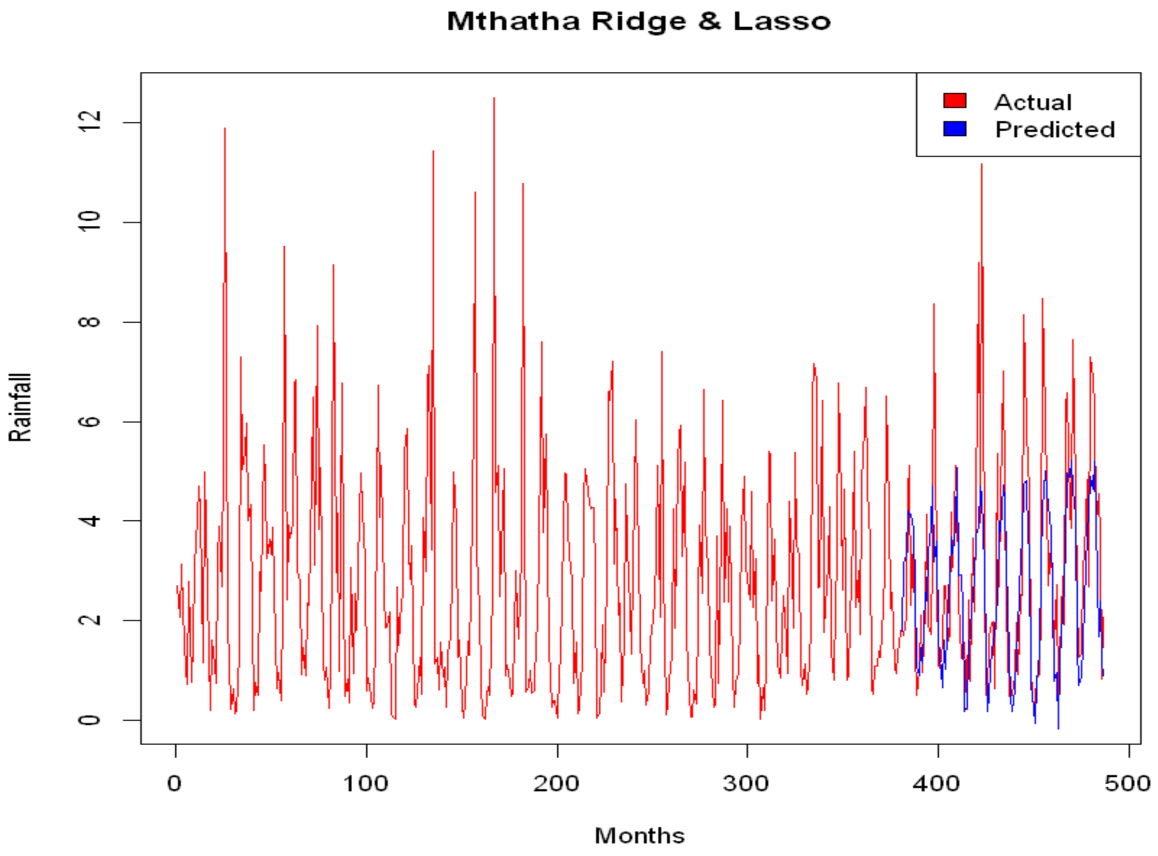

East London SVM
Rainfall
0
2
4
6
8
10
Months
0
100
200
300
400
500
🟥 Actual
🟦 Predicted

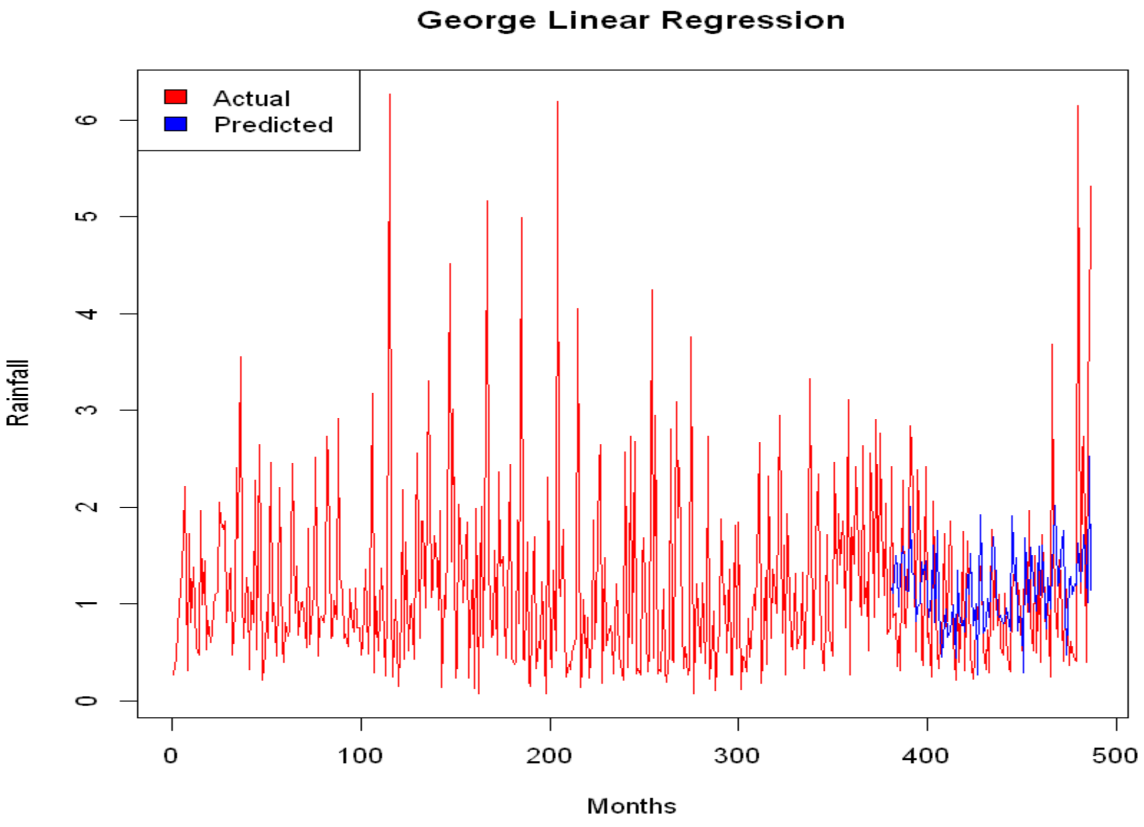

East London Ridge & Lasso
Rainfall
0
2
4
6
8
10
Months
0
100
200
300
400
500
🟥 Actual
🟦 Predicted

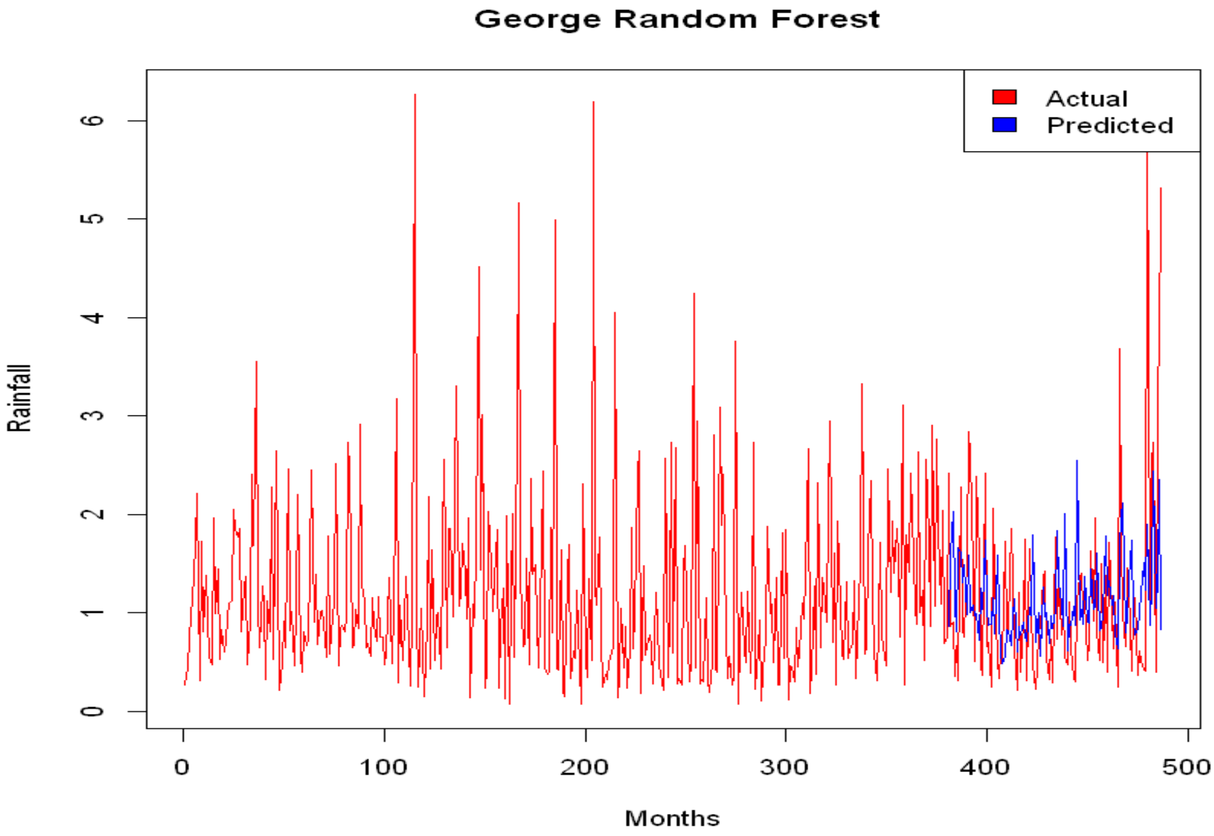

Mthatha Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
0
100
200
300
400
500
Months

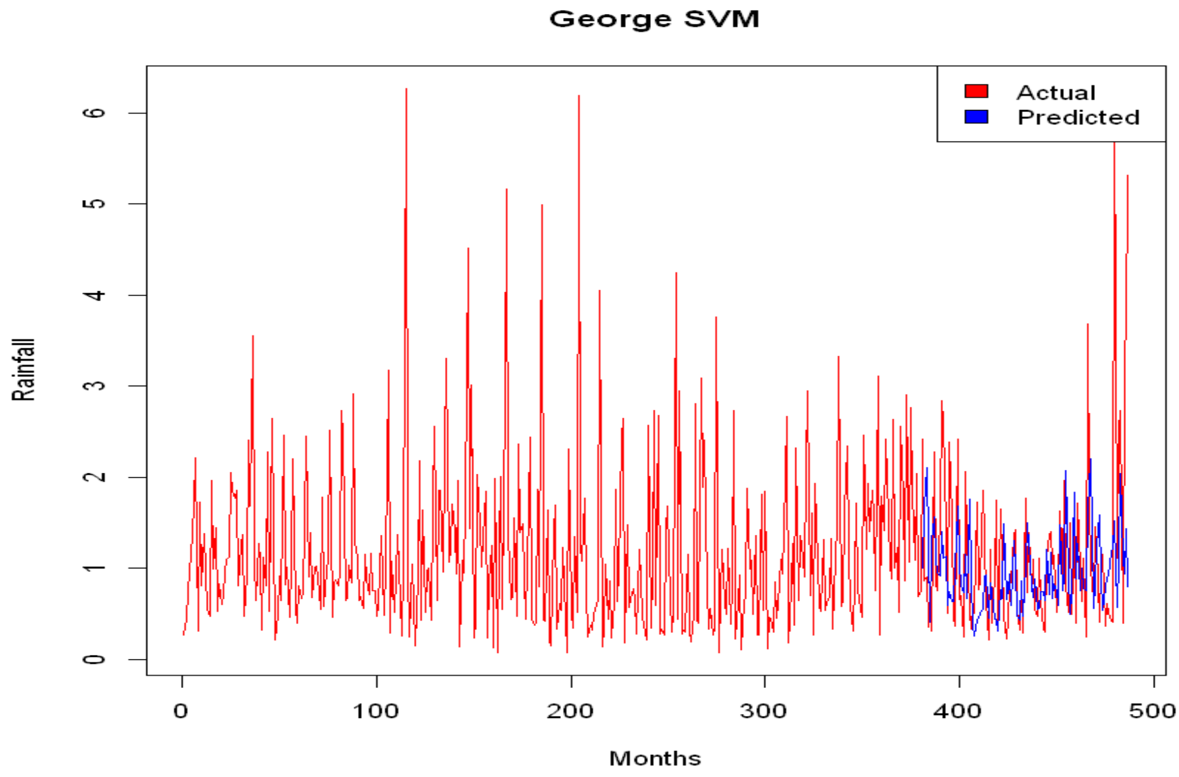

Mthatha Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
0
100
200
300
400
500
Months


Mthatha SVM
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
Months
0
100
200
300
400
500


Mthatha Ridge & Lasso
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
10
12
Months
0
100
200
300
400
500


George Linear Regression
Actual
Predicted
Rainfall
0
1
2
3
4
5
6
0
100
200
300
400
500
Months


George Random Forest
Actual
Predicted
Rainfall
0
1
2
3
4
5
6
0
100
200
300
400
500
Months



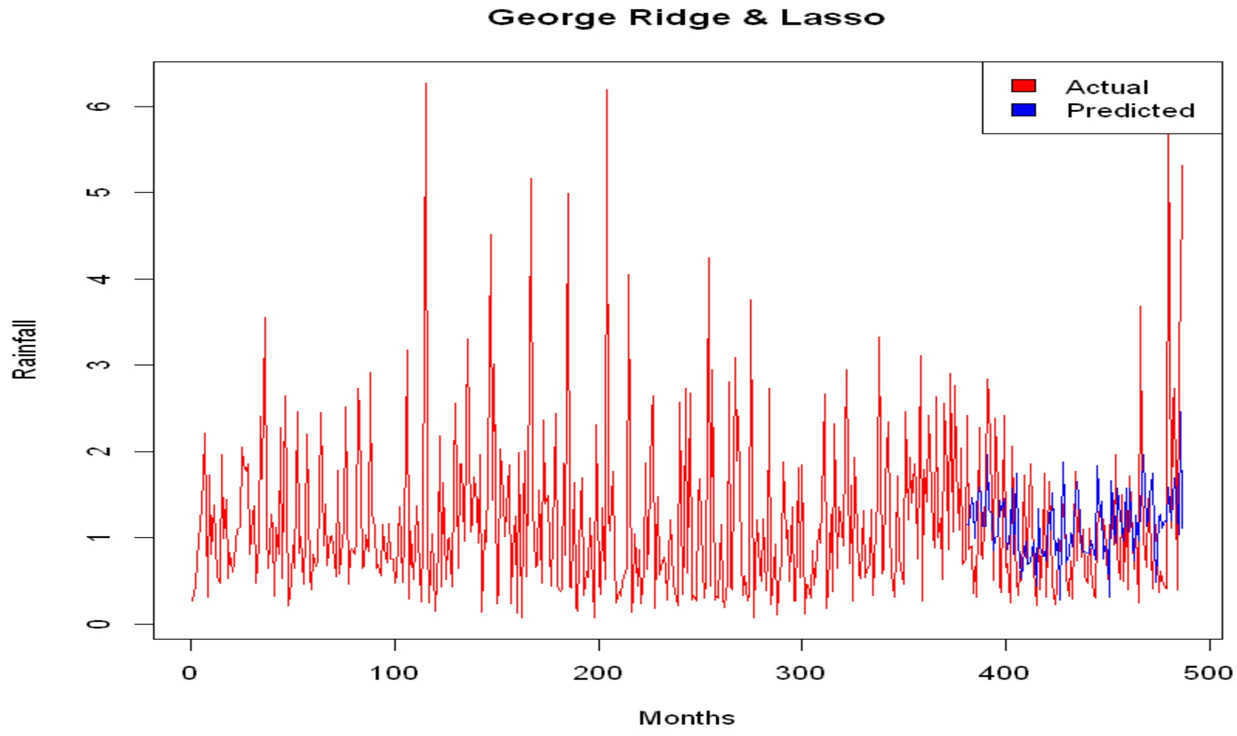


Figure 6. 2: Figures showing the predicted and actual rainfall for models used and locations under the temperate oceanic without dry season climate classification.

6.3 Warm and Dry Summer

The warm and dry summer climate classification is mainly found in the Western Cape Province of South Africa. Three locations in this province were selected: Bredasdorp, Cape Town, and Clanvilliam. Table 6.3 shows the performance of models using the metrics stated above. Random forest, support vector machine, and ridge and lasso had metric values within the same range while that of linear regression were completely different. The correlation coefficient in linear regression for the three locations were below 0.30, however, the coefficient of determination in Bredasdorp was surprisingly high with a value of 0.94. For the remaining models, no other location had a R square value higher than 0.40. The models generally performed poorly in these locations. The highest correlation coefficient recorded was 0.58 while using the support vector machine in Clanvilliam, while 0.54 and 0.53 were recorded for ridge and lasso and random forest. For Cape Town, correlation coefficients of 0.54, 0.53, 0.49 were recorded while using the support vector machine, random forest, and ridge and lasso models. Cash et al (2023) studied the predictable and unpredictable components of Cape Town winter rainfall using datasets from observational and seasonal forecast for their north American multi-model ensemble model. Their result showed that rainfall in Cape Town is dominated by unpredictable atmospheric variability which results in failure to accurately simulate.

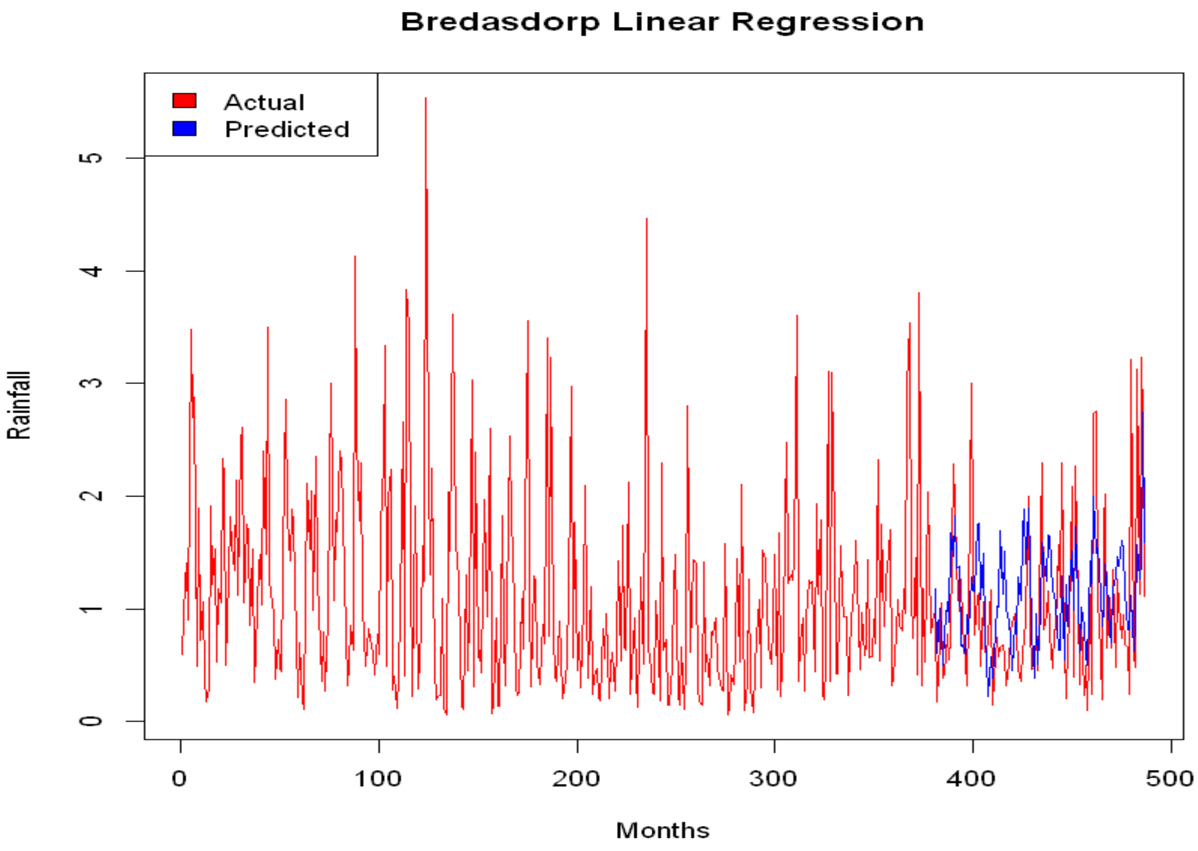
However, the models could predict the seasonality of rainfall in these locations as seen in figure 6.3, the models also estimated the amount of rainfall received. Perhaps this was made easier as all locations received very little amount of rainfall. Jury (2020) reported that the amount of rainfall Cape Town receives has reduced over the years resulting more dry months. With this correct estimation and seasonality prediction, these models can be used in predicting future estimate of rainfall expected despite the low metrics discussed above. As with other locations, the models could not accurately predict the increase in the amount of rainfall received in 2022.

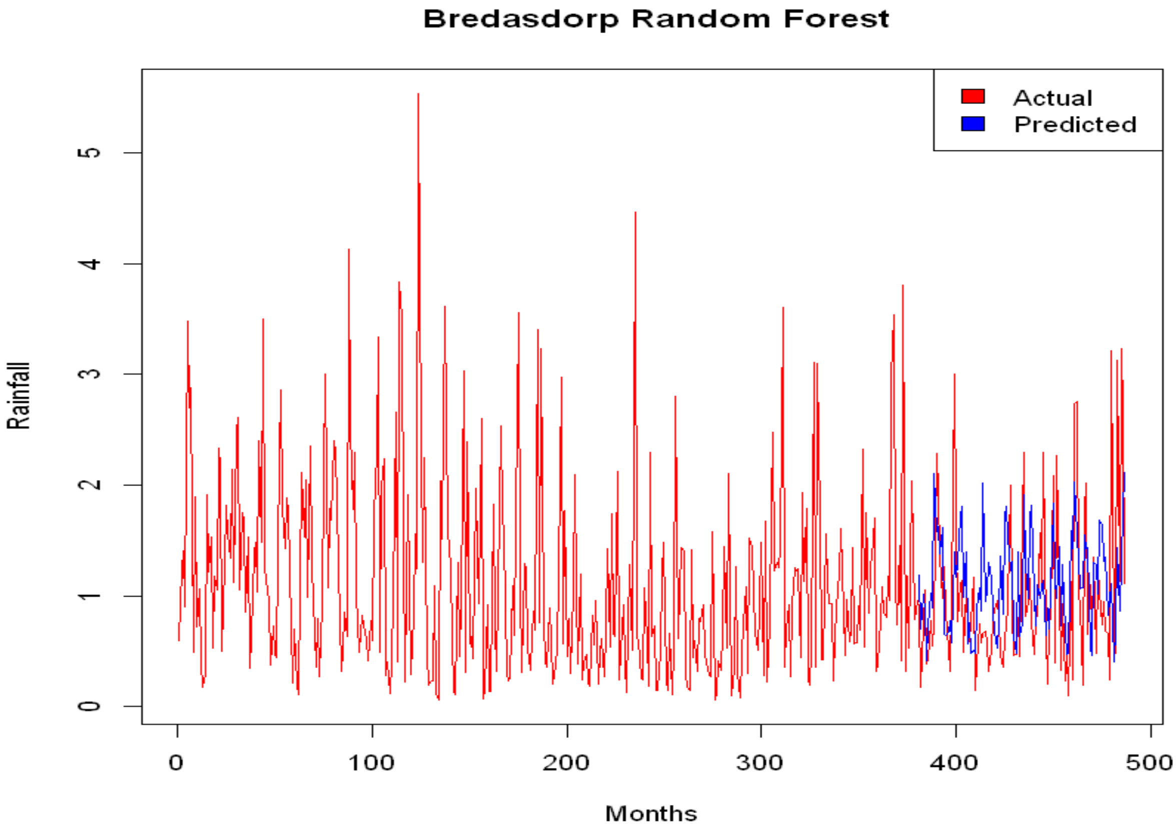
Since the evaluation metrics were low, it is not surprising that that the correlation between rainfall and other parameters were low. Contrary to most results observed earlier, wind speed had the highest correlation with rainfall in Cape Town with a value of 0.54. Other parameters such as temperature, cloud clover, relative humidity, dew point, and water vapour had a negative correlation with rainfall corresponding to -0.48, -0.45, -0.40, -0.36, -0.2 respectively. Jury (2020) reported an increase in easterly winds in Cape Town. This may be responsible for the high correlation between wind speed and rainfall in Cape Town compared to other locations. Similarly, a negative correlation was recorded between rainfall and dew point, relative humidity, temperature, and water vapour in Bredasdorp. Only cloud cover and wind speed correlated positively with rainfall with low values of 0.40 and 0.16 respectively. Though the atmospheric parameters in Clanvilliam had positive correlation with rainfall except wind speed, the coefficients were quite low with water vapour having the highest value of 0.31. Therefore, Alexander Bay, Port Elizabeth, Richards Bay and Geroge all show similar pattern in their correlation with rainfall with the cities under the warm and dry summer weather classification. The reason for this was not explored in this study.

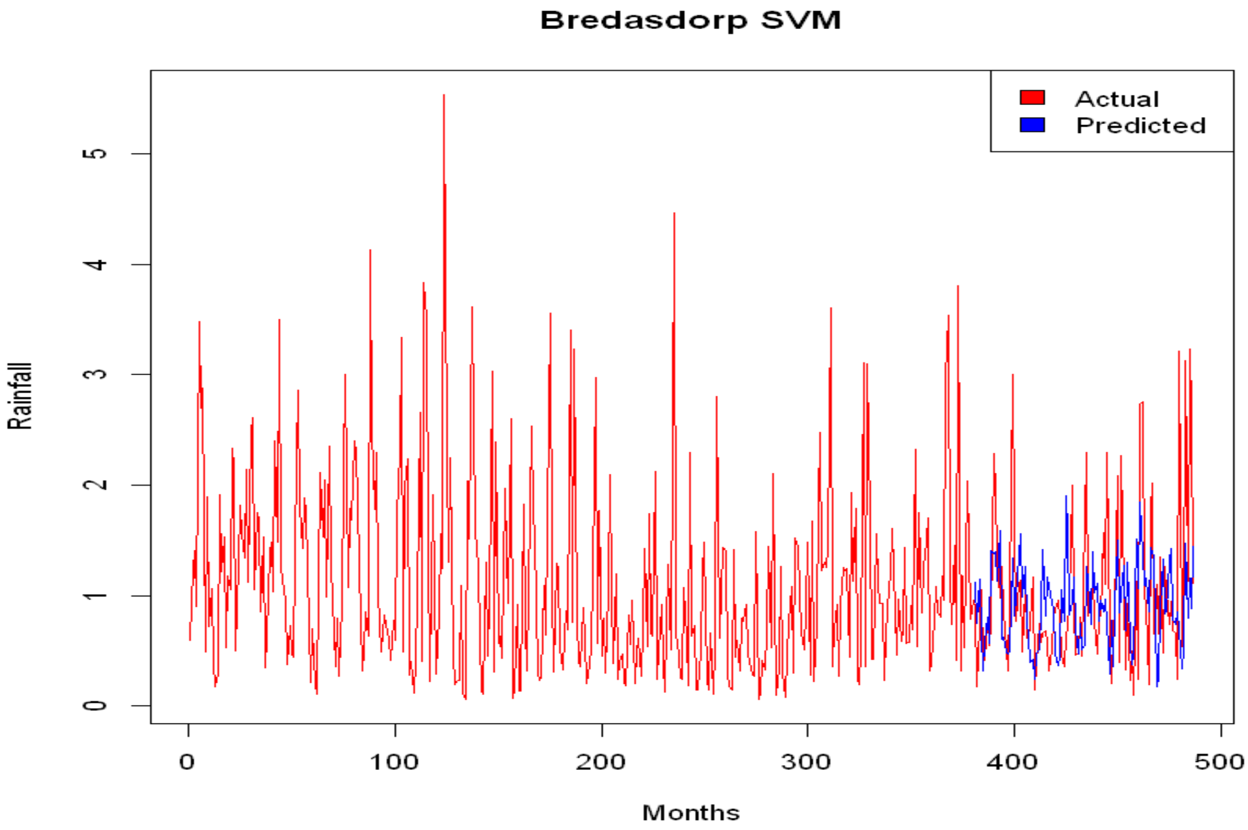
Table 6. 3: Table showing models evaluation metrics for warm and dry summer climate classification.

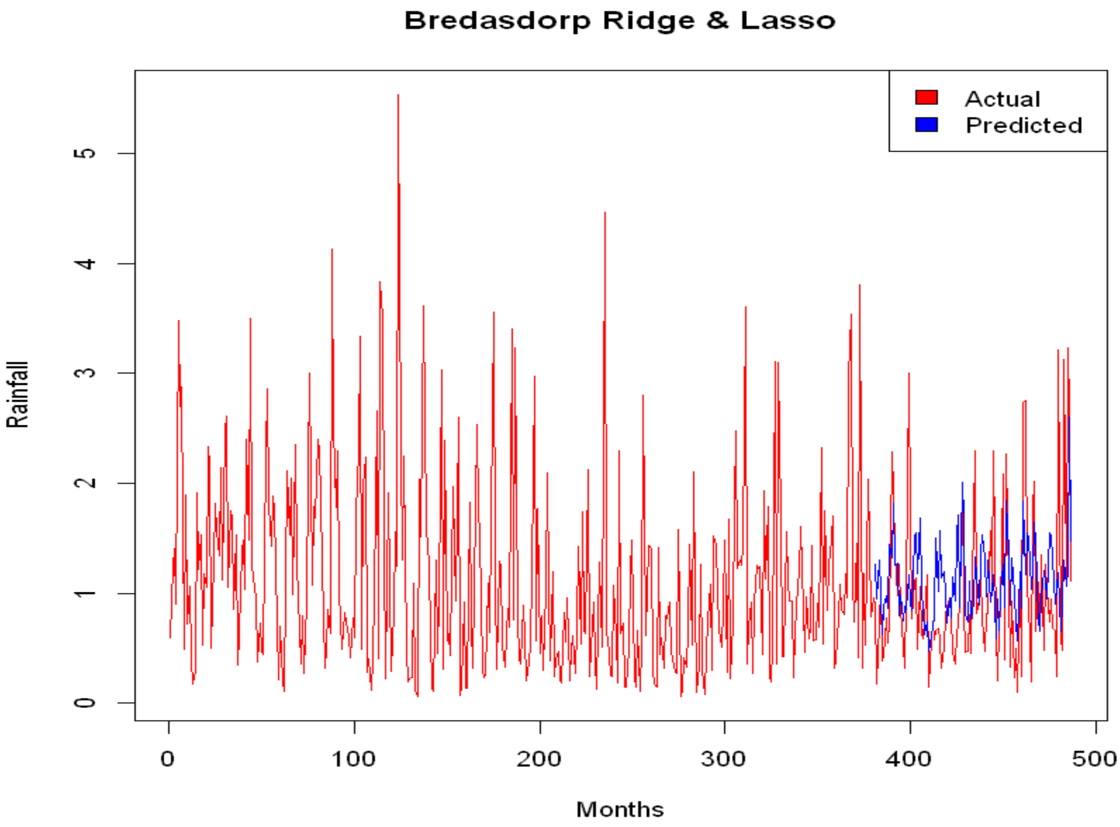
Linear Regression					
Bredasdorp	0.64	0.66	0.81	0.01	0.94
Cape Town	0.77	0.91	0.96	0.21	0.58
Clanvilliam	1.05	1.70	1.31	0.29	0.25
Random Forest					
Bredasdorp	0.53	0.45	0.67	0.41	0.11

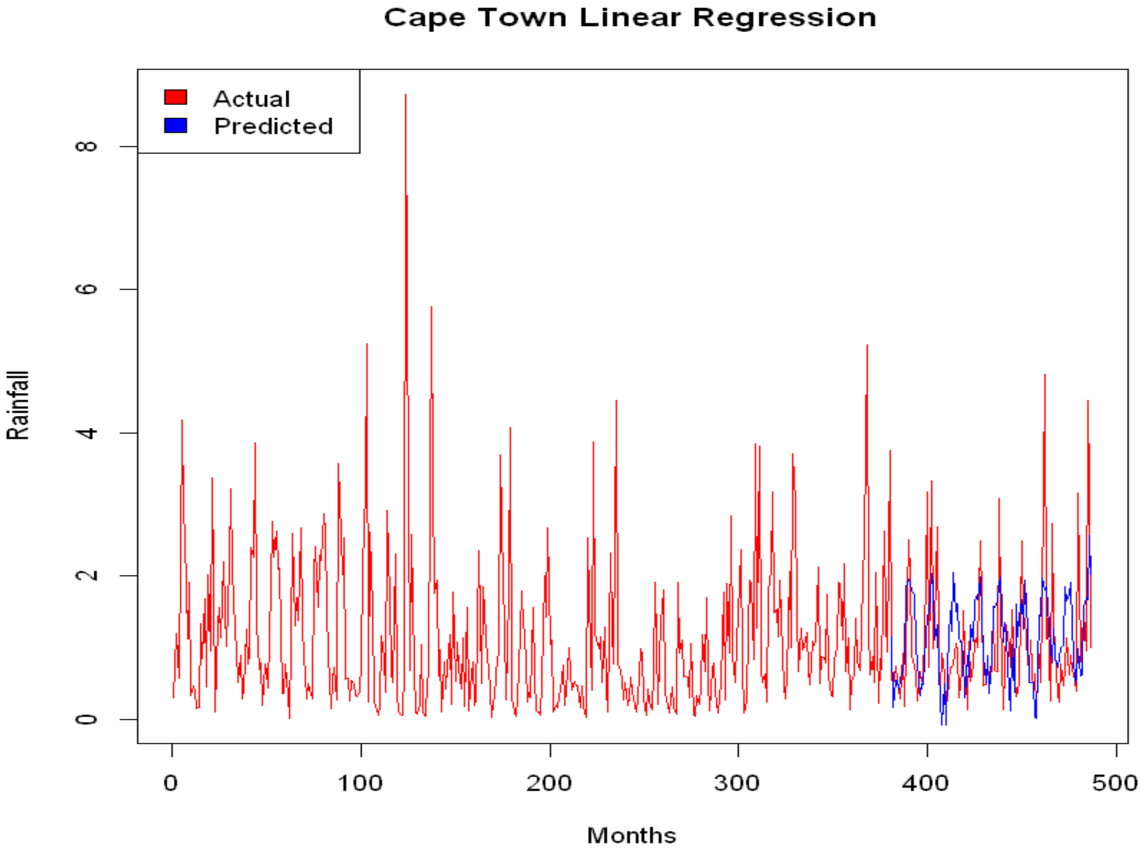
Cape Town	0.55	0.58	0.76	0.53	0.25
Clanvilliam	0.29	0.17	0.41	0.53	0.27
Support Vector Machine					
Bredasdorp	0.45	0.41	0.64	0.47	0.20
Cape Town	0.49	0.57	0.75	0.54	0.27
Clanvilliam	0.24	0.16	0.41	0.58	0.28
Ridge and Lasso					
Bredasdorp	0.51	0.43	0.66	0.36	0.31
Cape Town	0.53	0.58	0.76	0.49	0.40
Clanvilliam	0.28	0.17	0.42	0.54	0.32

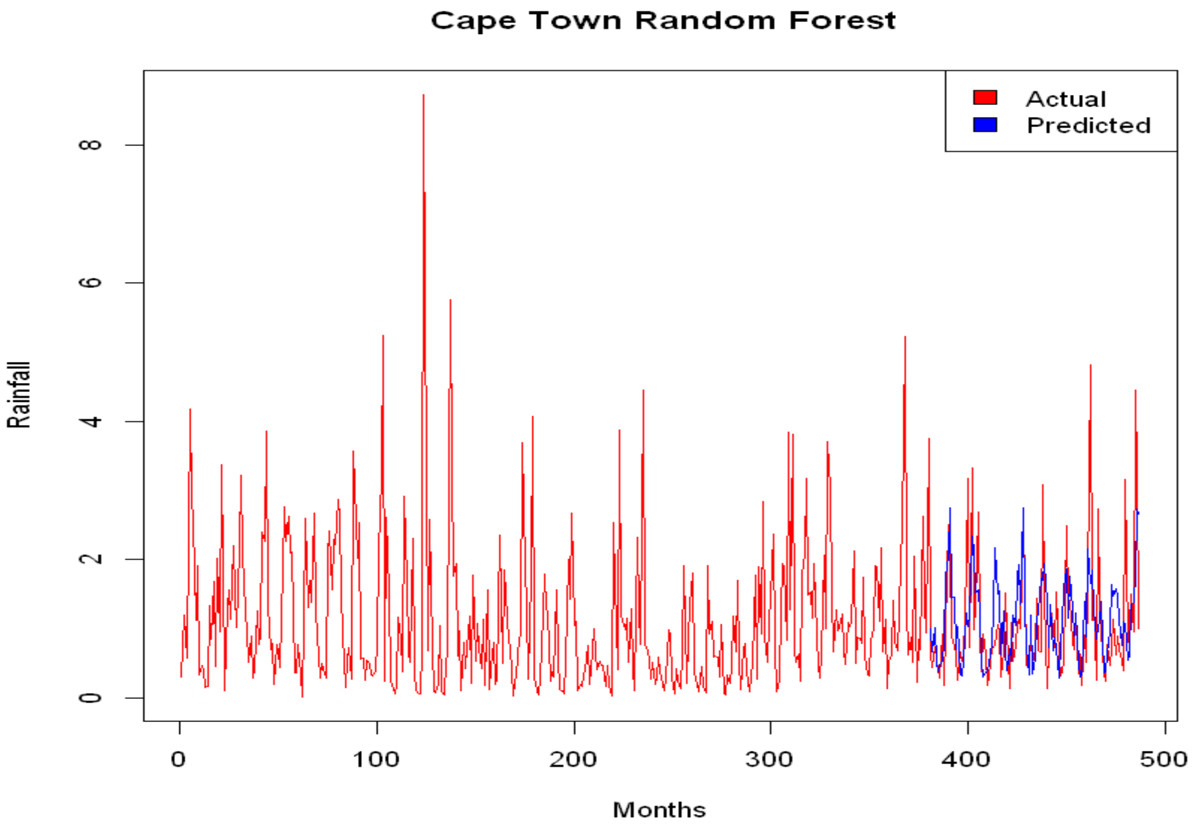


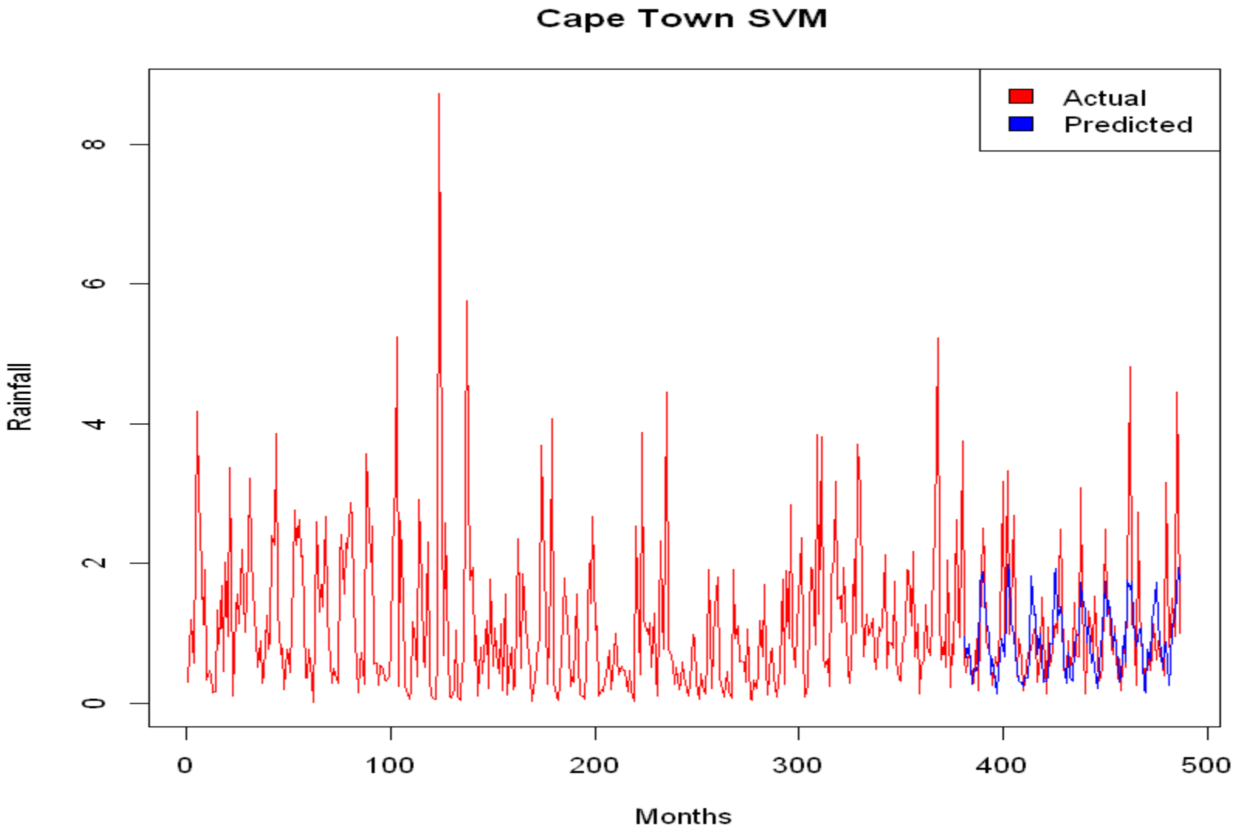

Bredasdorp Random Forest
Actual
Predicted
Rainfall
0
1
2
3
4
5
0
100
200
300
400
500
Months

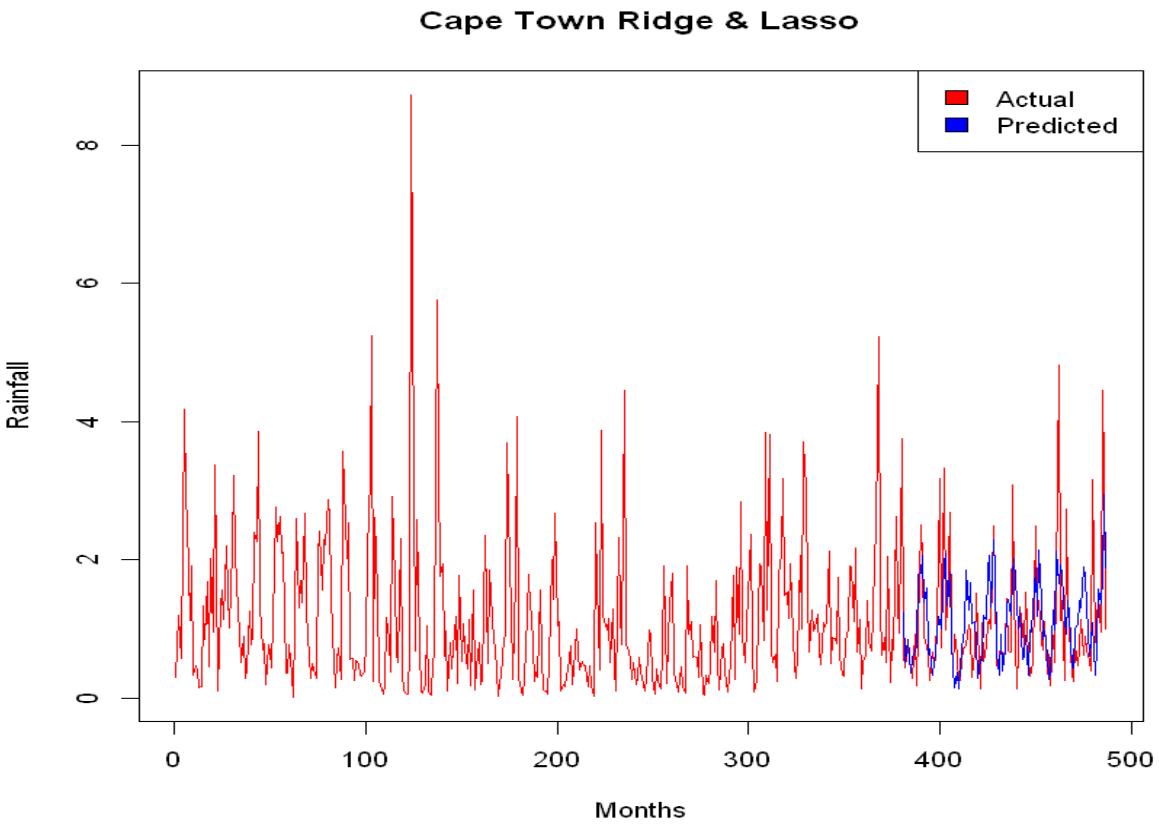

Bredasdorp SVM
Actual
Predicted
Rainfall
0
1
2
3
4
5
0
100
200
300
400
500
Months

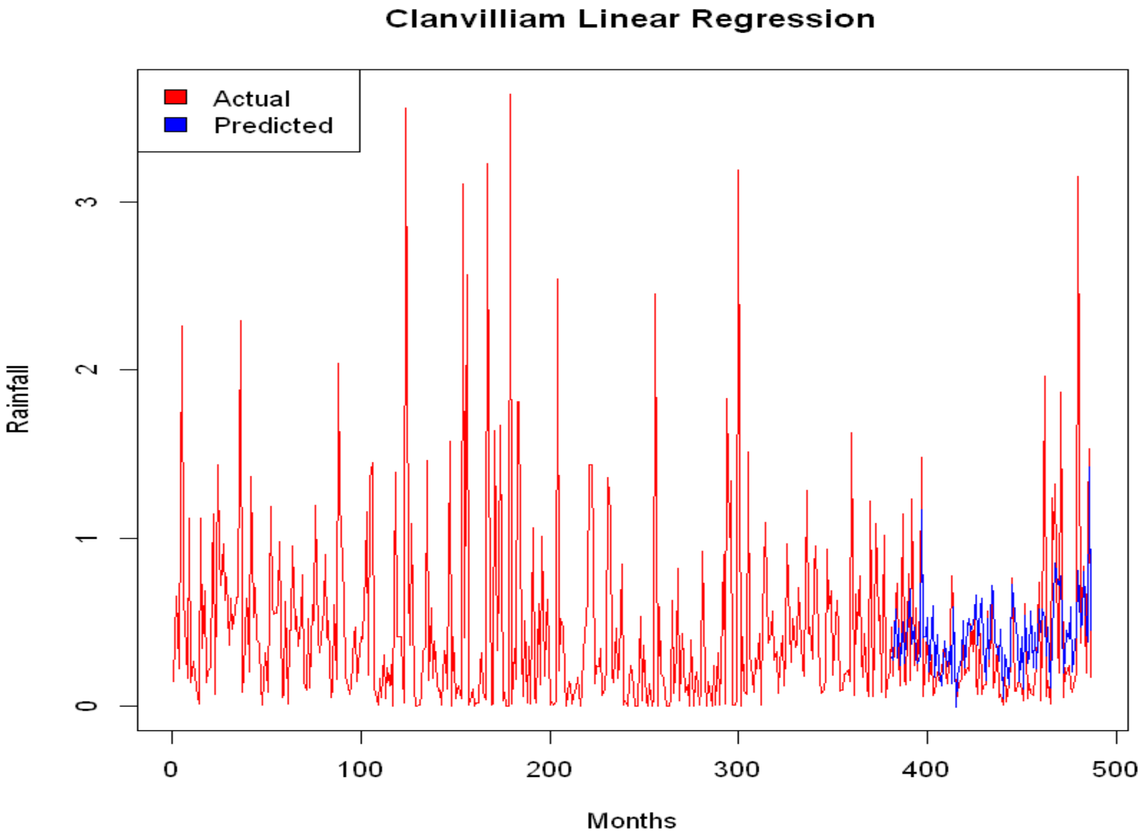

Bredasdorp Ridge & Lasso
🟥 Actual
🟦 Predicted
Rainfall
0
1
2
3
4
5
0
100
200
300
400
500
Months

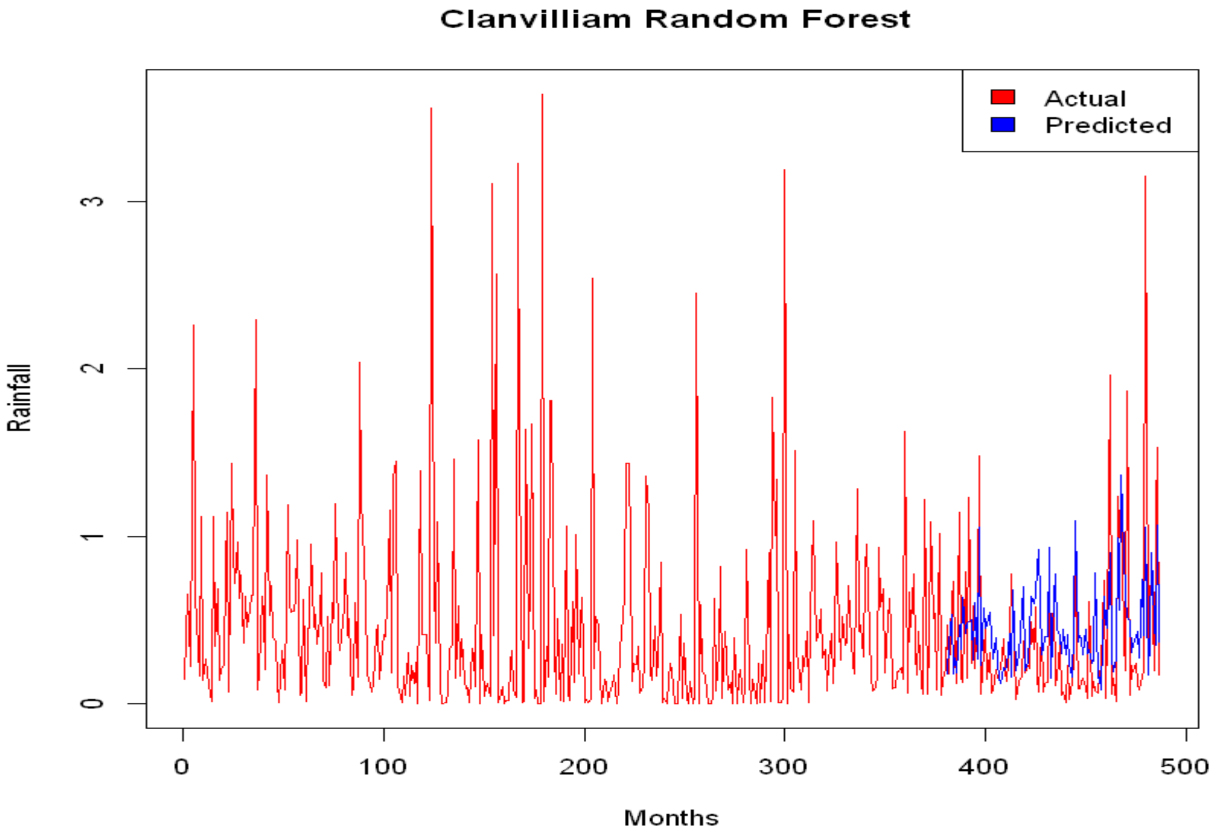

Cape Town Linear Regression
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
0
100
200
300
400
500
Months

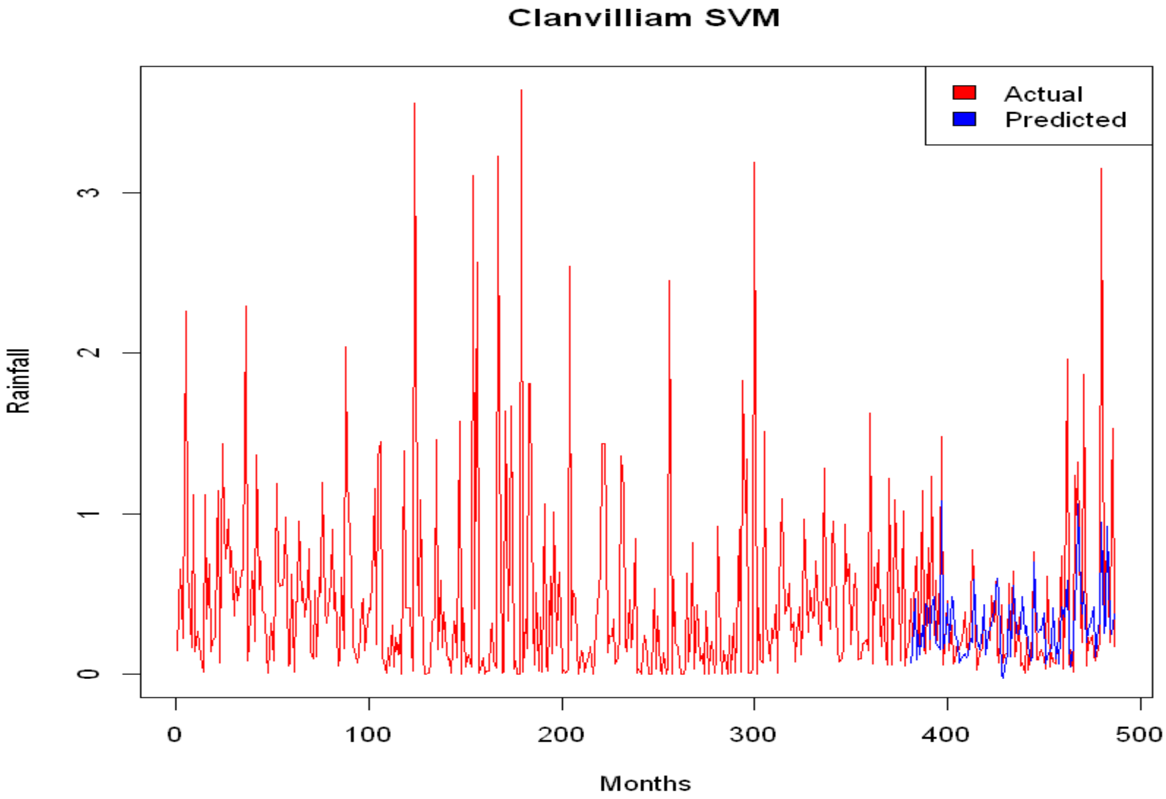

Cape Town Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
0
100
200
300
400
500
Months


Cape Town SVM
🟥 Actual
🟦 Predicted
Rainfall
0
2
4
6
8
0
100
200
300
400
500
Months


Cape Town Ridge & Lasso
Actual
Predicted
Rainfall
0
2
4
6
8
0
100
200
300
400
500
Months


Clanvilliam Linear Regression
Actual
Predicted
Rainfall
0
1
2
3
0
100
200
300
400
500
Months


Clanvilliam Random Forest
🟥 Actual
🟦 Predicted
Rainfall
0
1
2
3
0
100
200
300
400
500
Months


Clanvilliam SVM
🟥 Actual
🟦 Predicted
Rainfall
0
1
2
3
0
100
200
300
400
500
Months

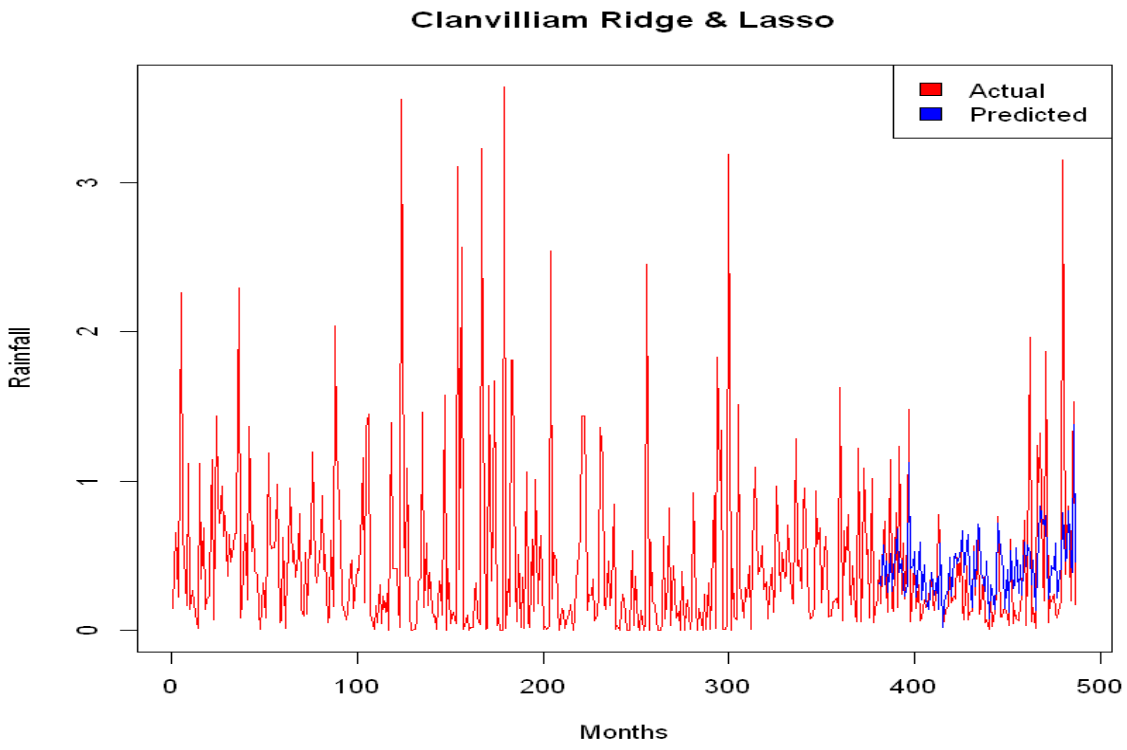


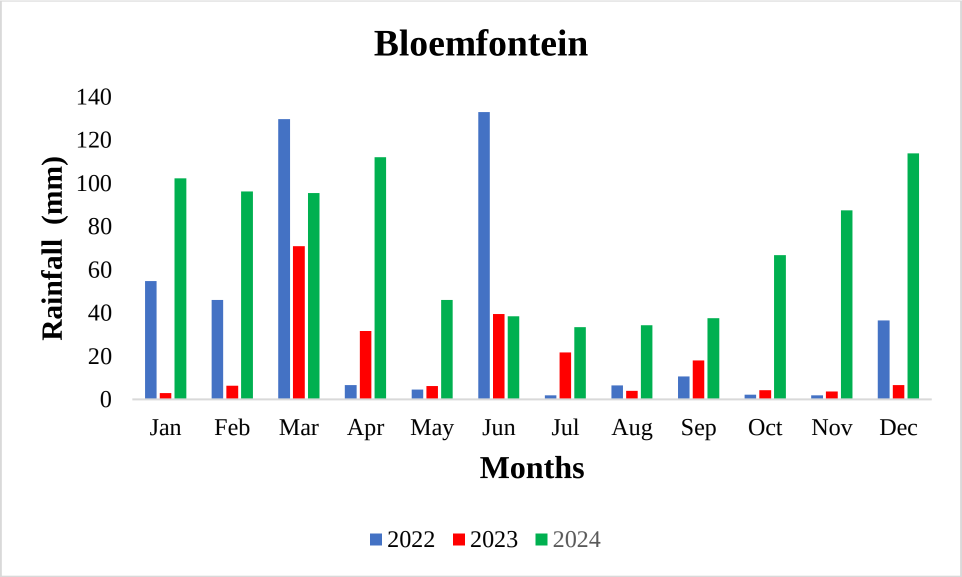
Figure 6. 3: Figures showing the predicted and actual rainfall for models used and locations under the warm and dry summer climate classification.

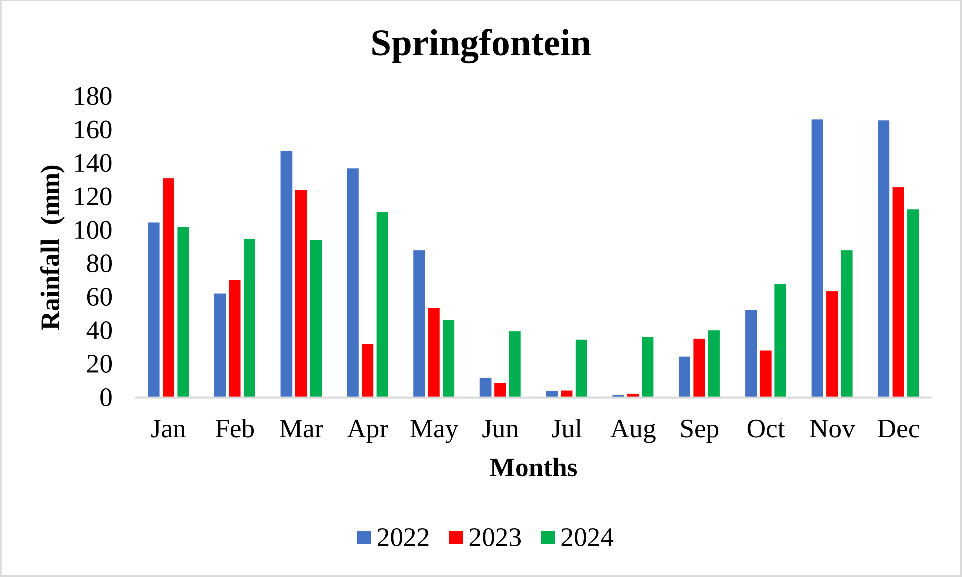
Chapter 7: Rainfall Prediction

7.1 Arid Climate Classification Rainfall Prediction

7.1.1 Cold and Semi-Arid Steppe Rainfall Prediction

This chapter presents monthly rainfall prediction for different climatic zones in South Africa for 2024 using random forest model. The prediction was compared to monthly rainfall of 2022 and 2023 as shown in figure 7.1. Under the cold and semi-arid steppe, in Bloemfontein, the figure shows that more rainfall of over 100mm will be received in the first four months of 2024. Similar amount of rainfall received in January, March, May, and October is expected in Springfontein in 2024. Compared to the previous years, more rainfall is expected in February and winter months. While there is an increase in the predicted amount of rainfall in November that in 2023, the amount of rainfall in December will be lower than what was experienced in 2023 and about half of 2022 rainfall. In Springfontein, it is predicted that there will be more rainfall in 2024 than 2023 throughout the years, and except for March and June, it is also expected that more rainfall will be received than 2024. This is good news for farmers as they can plan appropriately as rainfall above 80mm is expected in the last three months of 2024 while during winter, rainfall of about 40mm is expected monthly. In Welkom, the figure shows similar pattern of rainfall to 2023 in 2024 with 2024 expected to receive more rainfall. The amount of rainfall expected from February to April is also similar to the amount of rainfall Welkom received in 2022. However, more rainfall is expected in the autumn and early spring months (June to September) compared to previous years. Although the rainfall expected in October will be more than that of 2023, the amount of rainfall expected in November and December are similar to that of 2023 and much lower than rainfall in 2022 during summer. For the three cities under this climatic zone, more rainfall is expected during autumn in 2024 compared to other years. While Springfontein predicted that there would be more rainfall during spring and summer of 2024, similar amounts of rainfall are expected in Bloemfontein and Welkom compared to 2023.





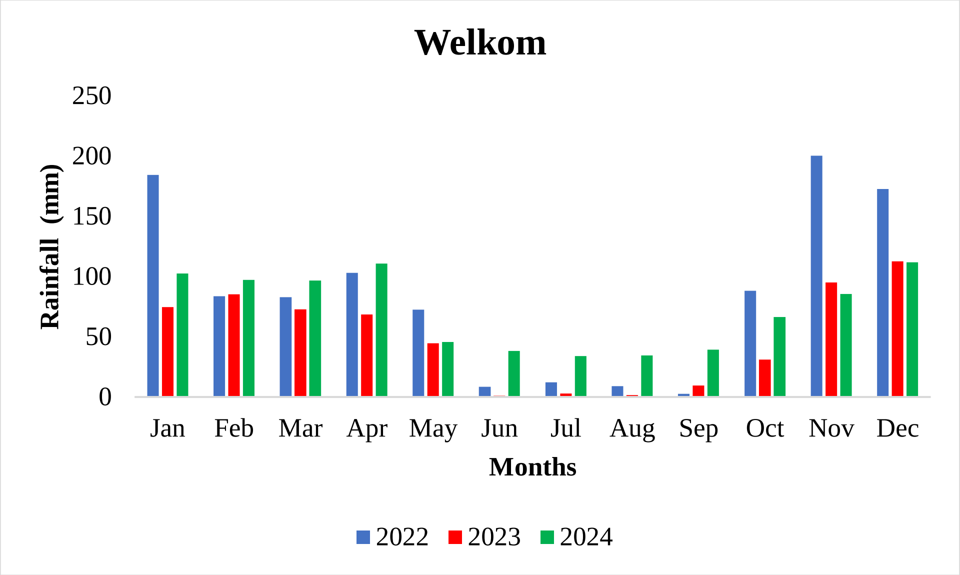
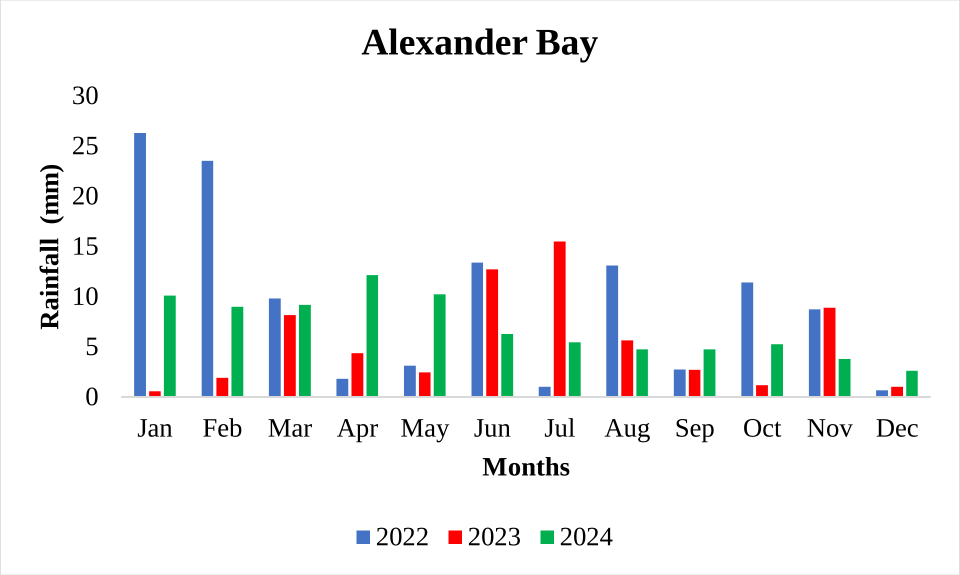
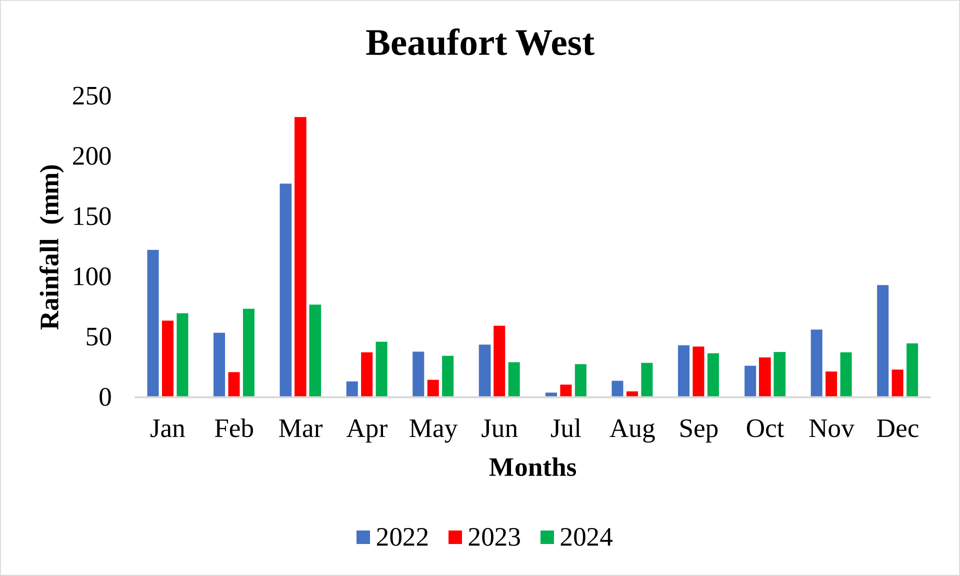


Figure 7. 1: Rainfall prediction for 2024 compared with 2023 and 2023 rainfall under the cold and semi-arid steppe (Bloemfontein, Springfontein, and Welkom).

7.1.2 Cold Arid Desert Rainfall Prediction

Figure 7.2 shows rainfall prediction for Alexander Bay, Beaufort West, and Bristown under the cold arid desert. The amount of rainfall expected in Alexander Bay is more in 2024 compared to 2023 for the first two months as there was little or no rainfall those months in 2023, however, compared to 2022, rainfall expected in January and February is less than half of what was received in 2022. However, in March, similar amount of rainfall is expected as that received in 2022 and 2023. In April and May, more rainfall is expected in 2024, more than double what was received in previous years. While rainfall in Alexander Bay was about 2mm and 4mm for April and May of 2022, in 2023, it was about 4mm and 3mm for April and May respectively. It is expected that rainfall for those months in 2024 will be about 13mm and 11mm indicating that more rainfall in expected in the winter of 2024 compared to other years. However, this amount of rainfall is still significantly small as it may all be showers. In Autumn, 2023 received more rainfall compared to what is expected in 2024 autumn with the amount of rainfall expected reducing by about half compared to 2023. There is a slight increase in the amount of rainfall expected in September, October, and December compared to 2023, though this increase is negligible. 2024 is therefore expected to be another dry year in Alexander Bay. In Beaufort West, 2024 is expected to be another dry. Only in the first three months is rainfall above 50mm expected, but all under 80mm. The amount of rainfall received in 2022 January reduced by almost half compared to 2023 January. This amount was maintained in 2023 however, it is predicted that February 2024 rainfall would be more than that of 2022 and 2023. The amount of rainfall expected in February is similar to that expected in March in Beaufort West. Compared to other years, this is less than half of what was received in March. Although, there was significant decrease from March to April by over 150mm in 2022 and 2023, the decrease in 2024 was about 30mm. The amount of rainfall expected in late winter, early autumn, and spring, is quite similar to that of 2024. More rainfall is expected is the last three months of the year compared to 2023. In Bristown, similar patten of rainfall is seen across the months with higher rainfall in the early parts of the year, reduces during winter and autumn, then increases late spring to early summer. It is predicted that over 100mm monthly rainfall will be received in 2024 from January to April, an increase from 2023 with no month receiving up to 75mm except in December. Compared with 2022 with high rainfall received, only in January, March May, September, and December did 2022 received more rainfall than what is predicted for 2024. The total amount of rainfall expected in 2024 is similar to the amount of rainfall received in 2022 which is almost twice the amount of rainfall received in 2023. That means, it is expected that 2024 rainfall in Bristown doubles the amount of rainfall received in 2023. Also, in 2024, more rainfall is expected during later winter and early autumn compared to other years. January, April, and December are expected to receive the hight amount of rainfall in 2024 in Bristown.


Alexander Bay
30
25
20
15
10
5
0
Rainfall (mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2022
2023
2024


Beaufort West
250
200
150
100
50
0
Rainfall (mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2022
2023
2024

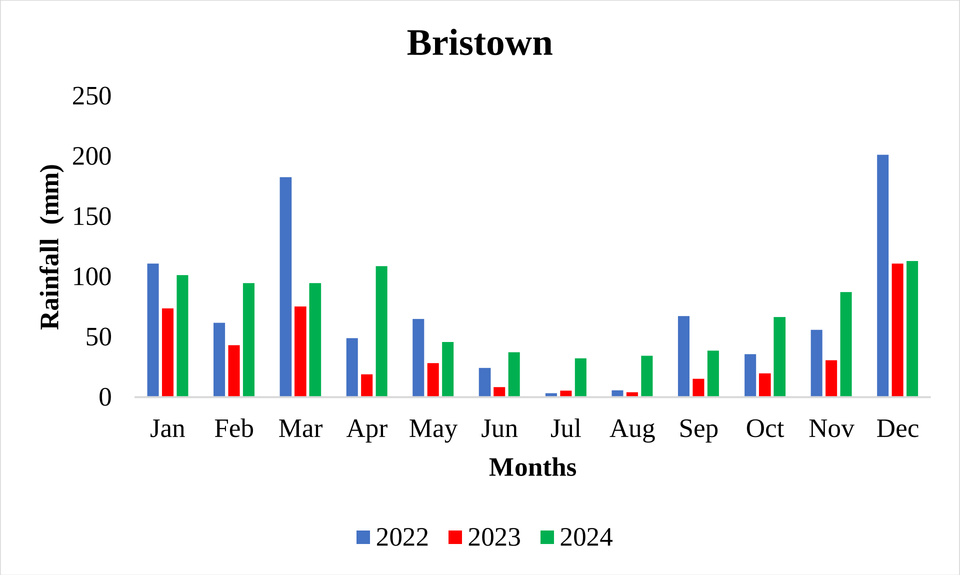
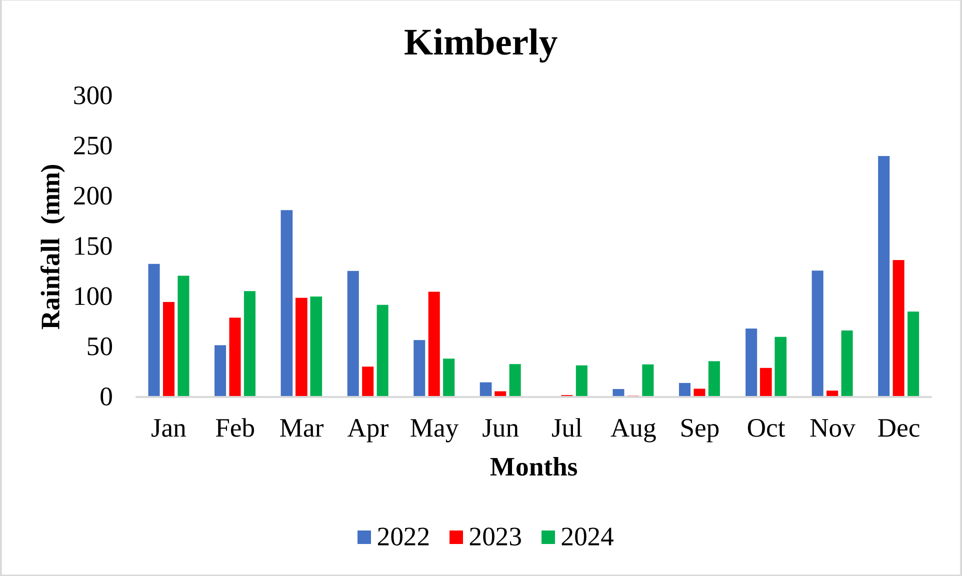

Bristown
250
200
150
100
50
0
Rainfall (mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2022
2023
2024

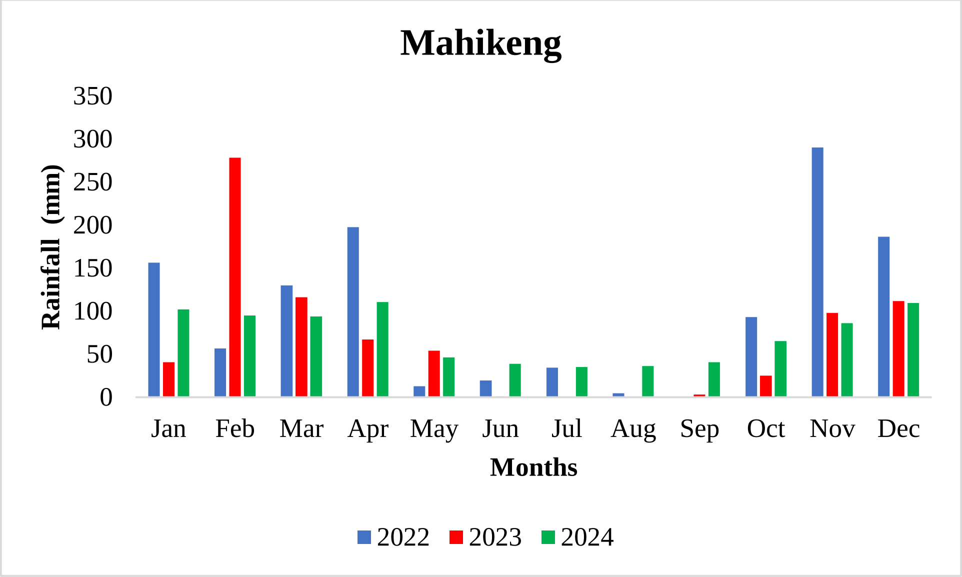
Figure 7. 2: Rainfall prediction for 2024 compared with 2023 and 2023 rainfall under the cold arid desert (Alexander Bay, Beaufort West, and Bristown).

7.1.3 Hot and Semi-Arid Steppe Rainfall Prediction

Figure 7.3 shows the rainfall prediction for 2024 compared to 2022 and 2023 rainfall for Kimberly, Mahikeng, and Port Elizabeth under the hot and semi-arid steppe climate classification. In Kimberly, the figure shows that it is expected to receive about 20mm more rainfall more in January than what was received in 2023 January and 10mm less what was received 2022 January. However, more rainfall is expected in February. There was an increase in rainfall by about 25mm in February rainfall in 2022 and 2023, this estimated increase is predicted for February 2023. In March, similar amount of rainfall compared to 2023 is expected in Kimberly in 2024 with little reduction in April. However, estimated rainfall for April 2024 will be about three times what was experienced in April 2023 and about 30mm lower than 2022 April. Compared to May 2022 and May 2023, lower rainfall is expected in Kimberly in 2024. From autumn (June) to early spring (September), more rainfall is expected in 2024 although, monthly estimates are below 35mm during this period. It is predicted that there will be gradual increase in the amount of rainfall expected till the end of the year. The predicted values are higher than the amount of rainfall Kimberly received in 2023 except from December but lower than 2022 rainfall. It is also predicted that there will be an annual increase of about 200mm in 2024 rainfall compared to 2023, however, this is still about 250mm lower than 2022 rainfall. In 2022, Mahikeng received rainfall of over 150mm in January, April, November, and December while in 2023, in February, the amount of rainfall Mahikeng received is estimated to be over 270mm. However, in 2024, only in January, April, and December are the predicted monthly estimate above 100mm. There is an increase by about 60mm in January rainfall between 2023 and predicted 2024 values but about 200mm difference in February. The amount of rainfall predicted for March though lower in 2024 is quite similar to that of 2023. In April, Mahikeng received about 200mm rainfall in 2022, about 70mm in 2023 and predicted estimate of over 110mm in 2024. Similar values in May are seen in both 2023 and 2024 estimates. Similar to Kimberly, there is expected to be more rainfall in autumn and early spring compared to other moths and then a gradual increase in 2024 rainfall till the end of the year. It is also predicted that the amount of rainfall in November and December will be similar to 2023 rainfall in those months. Predicted annual rainfall in 2024 is within the same range as that of 2023 but much lower than the amount of rainfall received in 2022 in Mahikeng.

In Port Elizabeth, more rainfall is predicted for 2024 for middle to late spring and summer months compared to 2023. Among all locations, this is the first city that has predicted more rainfall in 2024 than what was received both in 2022 and 2023. In January 2022, Port Elizabeth received about 20mm rainfall which increased to 70mm in January 2023, January 2024 estimates are predicted to be over 100mm. Similarly for February, it is expected that there will be more rainfall compared to 2022 and 2023. However, lesser rainfall is predicted for March compared with the previous two years. The amount of rainfall expected in April is more than twice what was experienced in 2022 and 2023. From May to September, except for August, it is predicted that lesser rainfall will be experienced compared to 2023. For October and November, rainfall estimates for 2024 will be at least thrice what was received in 2022 and 2023 while that of December will be similar to the amount of rainfall received December 2022.


Kimberly
300
250
200
150
100
50
0
Rainfall (mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2022
2023
2024


Mahikeng
350
300
250
200
150
100
50
0
Rainfall (mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2022
2023
2024

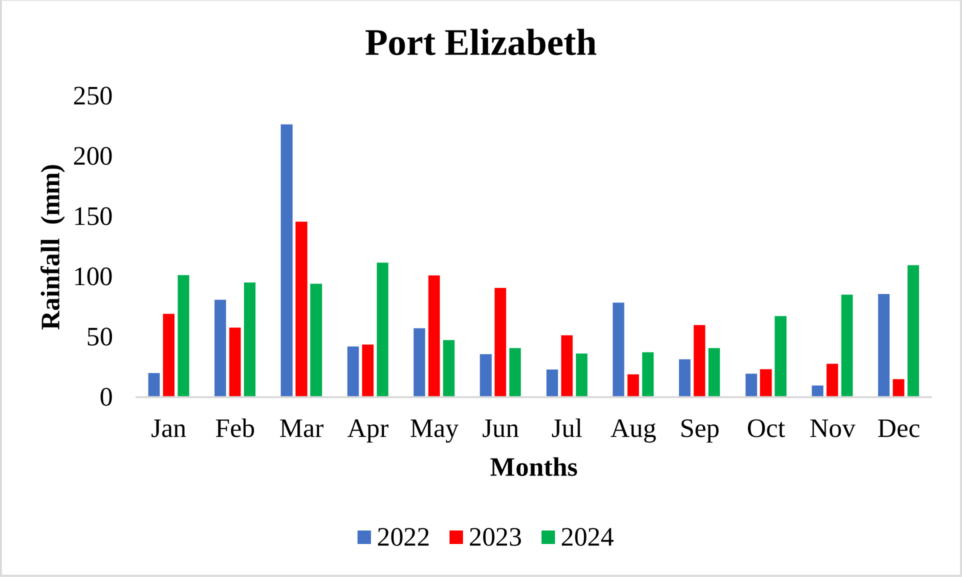
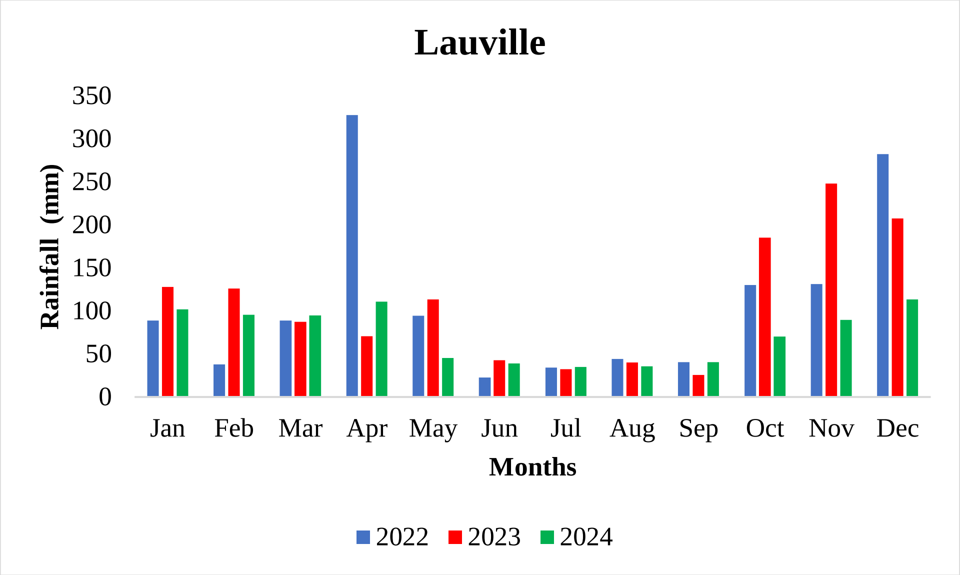

Port Elizabeth
250
200
150
100
50
0
Rainfall (mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2022
2023
2024

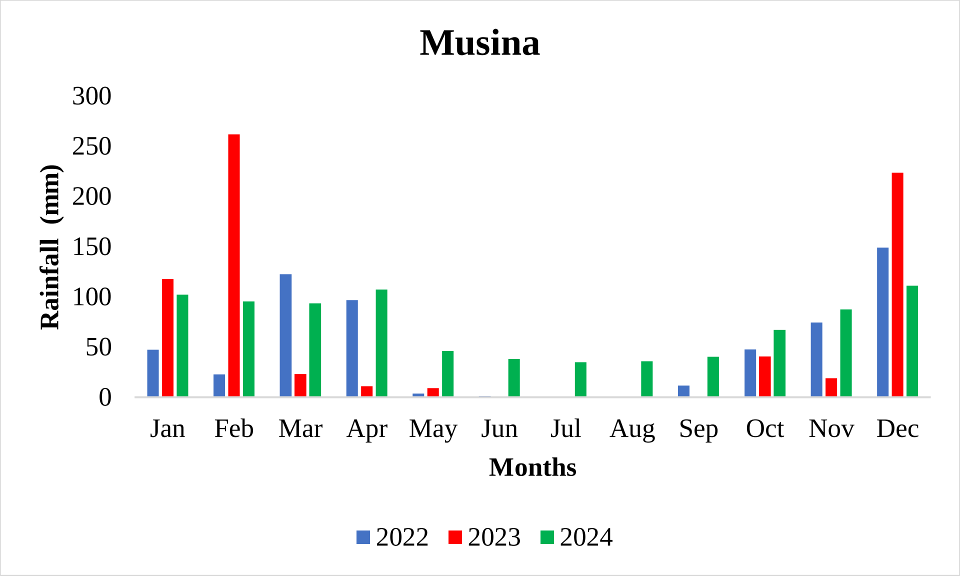
Figure 7. 3: Rainfall prediction for 2024 compared with 2023 and 2023 rainfall under the hot and semi-arid desert (Kimberly, Mahikeng, and Port Elizabeth).

7.1.4 Hot arid desert Rainfall Prediction

In Lauville, there is a predicted decrease of about 450mm in the amount of rainfall expected in 2024 compared to 2022 and 2023 rainfall. Therefore, with monthly rainfall from January to April 2024 around 100mm, it is suggested that proper planning be made especially for those into farming to mitigate the shortage of water. Although similar amount of rainfall is observed from the graph as shown in figure 7.4 for 2023 and 2024, the most significant change can be seen in the last three months of the year where expected rainfall in 2024 is less than half the amount received in 2023.

Contrary to Lauville, the amount of rainfall estimated for 2024 in Musina is more than the amount received in 2023 while the predicted rainfall for 2024 in Upington is more than twice the amount received in 2023. In 2023, Musina received significant amount of rainfall in January, February, and December with February and December rainfall over 200mm. While Musina have only few months of rainfall above 100mm as predicted, the amount of rainfall expected is well spread compared to other months. While there was almost drought between May and September of 2022 and 2023, monthly rainfall of about 40mm is expected during this period in 2024 which will gradually increase to over 110mm by the end of the year. In Upington as shown in figure 7.4, only in January, February, March, and December did it receive rainfall above 40mm in 2023. 2023 rainfall in other months were insignificant. Though there was more rainfall in 2022 compared to 2023, only in March does the 2022 rainfall exceed the predicted rainfall for 2024. Similar amount of rainfall in January, February, May, June, and December 2022 is predicted in 2024.


Lauville
350
300
250
200
150
100
50
0
Rainfall (mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2022
2023
2024


Musina
300
250
200
150
100
50
0
Rainfall (mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2022
2023
2024

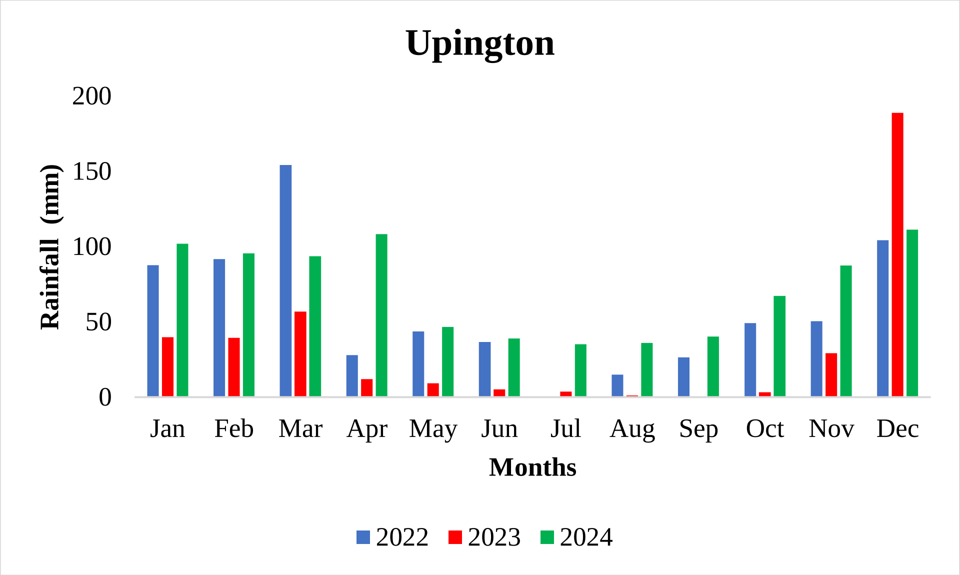

Upington
200
150
100
50
0
Rainfall (mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2022
2023
2024

Figure 7. 4: Rainfall prediction for 2024 compared with 2023 and 2023 rainfall under the hot arid desert (Lauville, Musina, and Upington).

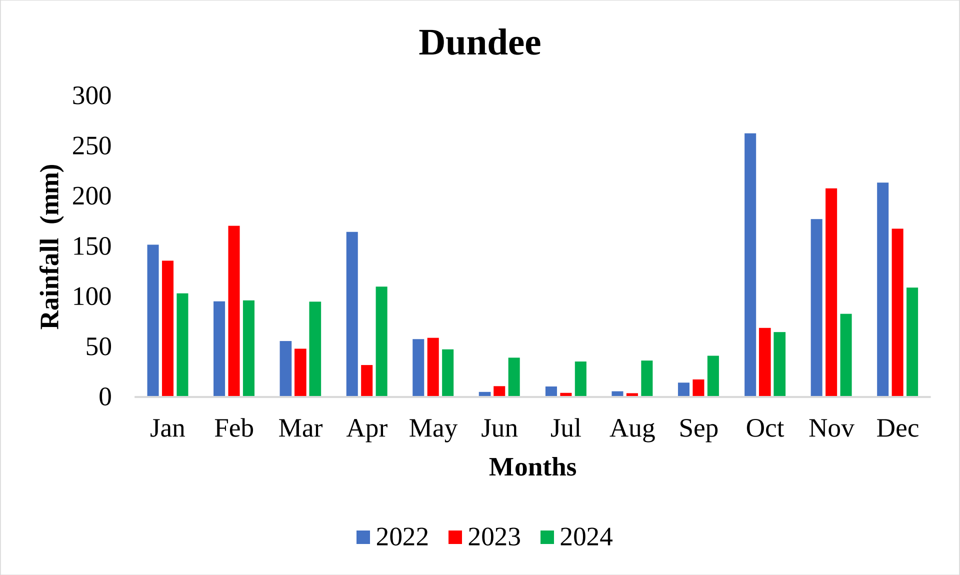
7.2 Subtropical Wet Rainfall Prediction

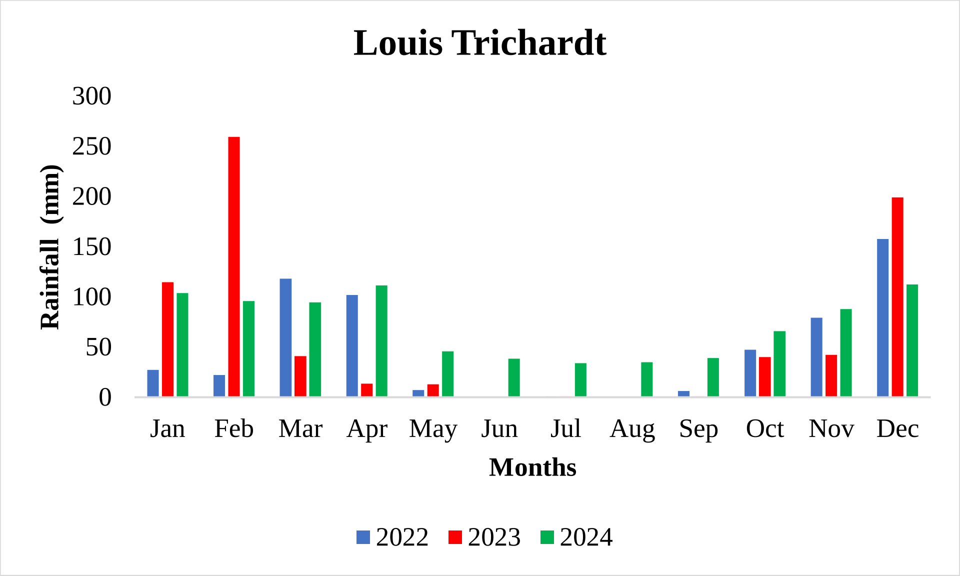
7.2.1 Humid Subtropical with Dry Winter Rainfall Prediction

Figure 7.5 shows the predicted rainfall for 2024 compared with rainfall received in 2022 and 2023 in Dundee, Louis Trichardt, and Nelspruit. In Dundee, there is an observed decrease in rainfall received in January 2022 and January 2023 by about 20mm, and further decrease by 20mm in the estimated values of January 2024. However, in February, the amount of rainfall predicted for 2024 is similar to rainfall received in 2022 and about 80mm lesser than the February rainfall of 2023. It is predicted that there will be more rainfall in March 2024 compared to previous years. It is also predicted that the rainfall expected in April will be thrice the amount of rainfall received in 2023 April but less than that of 2022. However, in May, the amount of rainfall received in the past two years and predicted rainfall for 2024 are similar. Between June and September, it is predicted that there will be more rainfall that the same period combined in 2022 and 2023. From mid-spring to December, it is predicted that there will be lesser rainfall in 2024 compared to previous years. The amount of rain expected in November is less than half what was received in the past two years and the amount expected in December will be about 60mm lower than what was received in 2023 and 100mm lower than December 2022 rainfall. Summarily, it is predicted that there would be lower rainfall in 2024 compared to 2022 and 2023 despite having more rainfall in the autumn of 2024.

In Louis Trichardt, it is predicted that the annual rainfall of 2024 would be more than those of 2022 and 2023. In 2023, significant months of rainfall were January, February, and December all receiving rainfall above 100mm while March, October, and November received about 40mm rainfall. In 2022, significant amount of rainfall was received in March, April, November, and December while in 2023, it is predicted that the first four months and the last three months will experience significant amount of rainfall. It is also predicted that there will not be any dry month in 2024 compared to other years. predicted rainfall for winter months would be about 40mm. However, despite 2024 having more rainfall than the previous two years, the predicted monthly rainfall would still be lower than the months of heavy rainfall in 2022 and 2023. For instance, February 2023, about 250mm rainfall was received while the predicted rainfall for 2024 would be about 100mm, although the amount of rainfall received January 2023 and predicted January 2024 are similar. The amount of rainfall expected In March, April, October, and November are also similar to the amount received in 2022 during these months.

The amount of rainfall expected in 2024 is similar to the annual rainfall of 2022 in Nelspruit and lower than 2023 rainfall. In January 2024, it is predicted that there will be more rain compared to January of 2022 and 2023 by about 30mm. 2023 had months of heavy rain like February, November, and December where rainfall exceeded 200mm and even 250mm in November. Also, in 2022, rainfall in November and December were above 150mm, however, in 2024, months with heavy rainfall would receive only about 100mm rainfall. These months are January, April, and November, although significant amount of rainfall is predicted for February, March, and November. It is also predicted that there will be more rainfall for 2024 winter and early spring compared to 2022 and 2023.


Dundee
300
250
200
150
100
50
0
Rainfall (mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2022
2023
2024


Louis Trichardt
300
250
200
150
100
50
0
Rainfall (mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2022
2023
2024

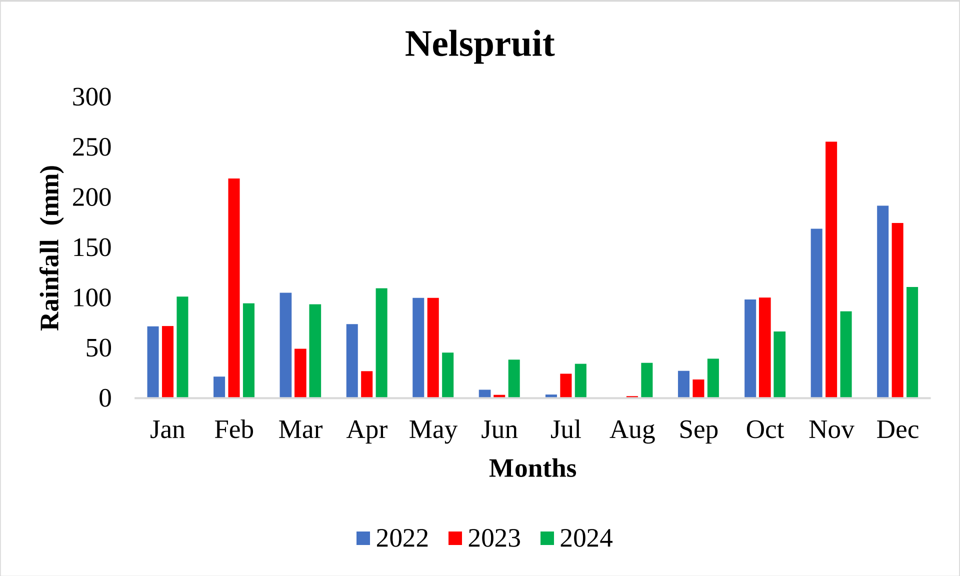

Nelspruit
300
250
200
150
100
50
0
Rainfall (mm)
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Months
2022
2023
2024

Figure 7. 5: Rainfall prediction for 2024 compared with 2023 and 2023 rainfall under the humid subtropical with dry winter (Dundee, Louis Trichardt, and Nelspruit).

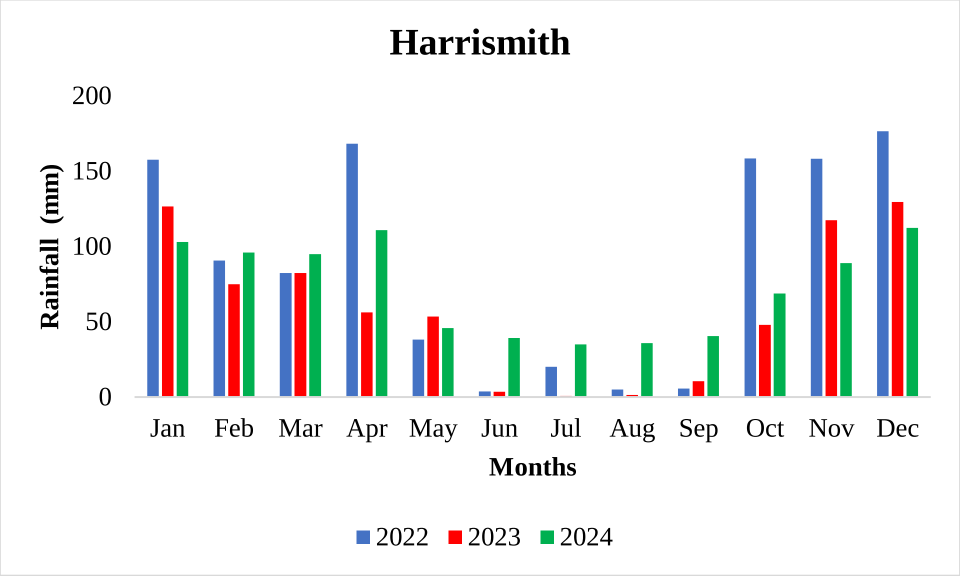
7.2.2 Subtropical Highland with Dry Winter Rainfall Prediction

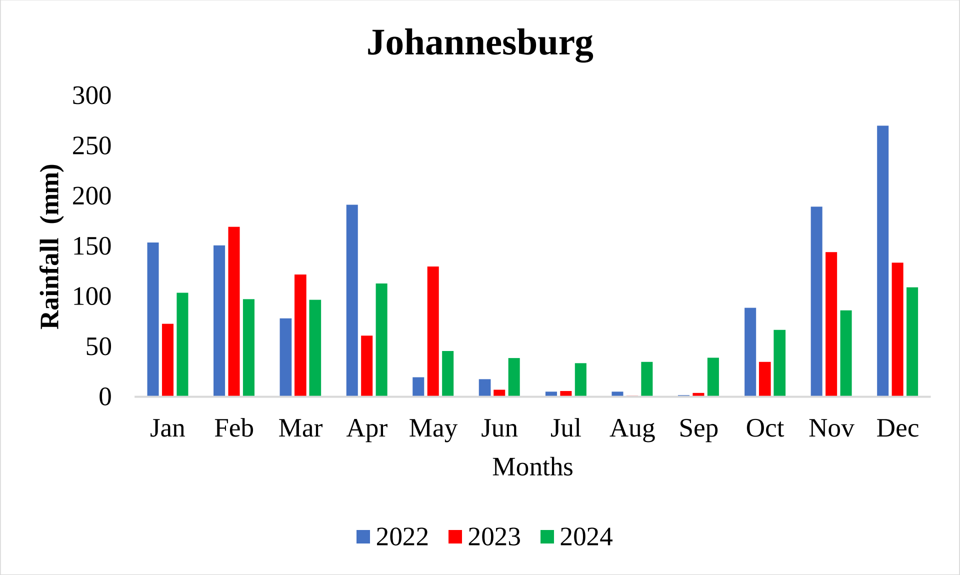
Figure 7.6 shows the predicted rainfall for 2024 for the subtropical highland with dry winter and compared with 2022 and 2023 results for Harrismith, Johannesburg, and Newcastle. The result shows that there was more rainfall in 2022 than in 2023 and what is predicted for 2024, although 2024 estimates are higher than 2023 for Harrismith. In 2022, there were five months with monthly rainfall above 150mm (January, April, October, November, and December) with February and March also experiencing high rainfall. Compared to 2023, rainfall above 100mm were only received in January, November, and December. While only January, April, and December of 2024 are predicted to have rainfall above 100mm, other months are predicted to equally have high amount of rain. Between January 2022 and January 2023, there is a decrease of about 30mm in rain received, a further decrease of about 20mm is predicted for January 2024. The predicted amount of rain for February and March are within 20mm of what was received both in 2022 and 2023 with 2024 predicted to have more rain. It is also predicted that there will be more rain in April of 2024 by about 50mm to what was received April 2023, but lower than the amount of rainfall received in 2022 in Harrismith. In May, a decrease of about 10mm is predicted while compared to 2023 rainfall. However, more rainfall is predicted for 2024 from June to September. Though this climatic zone is dry winter, it is predicted that there would be more rain during winter than previous years. It is also predicted that there would be a gradual increase in 2024 rainfall from September till the end of the year, however, it would still be lower than 2022 and 2023 rainfall.

In the commercial city of Johannesburg, the total amount of rainfall predicted for 2024 would be a little lower than the annual rainfall of 2023 and much lower than the annual rainfall of 2022. Similar to Harrismith, Johannesburg had five months of rain above 150mm in 2022

(January, February, April, November, and December) while there were four months of rain above 100mm in 2023 (February, May, November, and December). This reduced to three months in the predicted rainfall for 2024 (January, April, and December). This reveals that more rain should be expected in the early and late months of the year with varying intensity. Compared to 2024, in the first five months of the year, only in January and April will the predicted rainfall be more than that of 2023 while compared with 2022, 2024 predicted rainfall will only be higher in March and May. A monthly average of 100mm rainfall is predicted for the first four months of 2024. It is also predicted that there will be more rain in winter and early spring compared to other years. the predicted amount of rainfall for the last two months of 2024 is much lower than rainfall of 2023 and 2022 during this period.

Newcastle also had four months (April, October, November, and December) of rainfall above 150mm with rainfall in two of those months exceeding 200mm (October and December) in 2022. In 2023 rainfall over 150mm were recorded only in February and November with December equally having high amount of rainfall. In 2024, none of the months is predicted to have rainfall of over 150mm with only three months (January, April, and December) having rainfall above 100mm. Predicted rain for January 2024 is similar to what was received both in 2022 and 2023 while that of February is similar to 2022 February rainfall and much lower than 2023 February rainfall. While for March, predicted rainfall is about twice what was received in the previous two years. Similarly for April, predicted rainfall is more than twice what was received in 2023 though lower than 2022 April rainfall while the predicted rainfall for May is within the same range as what was recorded in 2022 and 2023. As with other cities under the dry winter climatic zone, more rainfall is expected in 2024 compared to other years. Rainfall predicted for October is similar to 2023 October rainfall but much lower than that of 2023. However, for the last two months of the year, lesser rainfall is predicted compared to these months in 2022 and 2023 with November receiving about half the amount of rainfall recorded in 2022 and 2023.





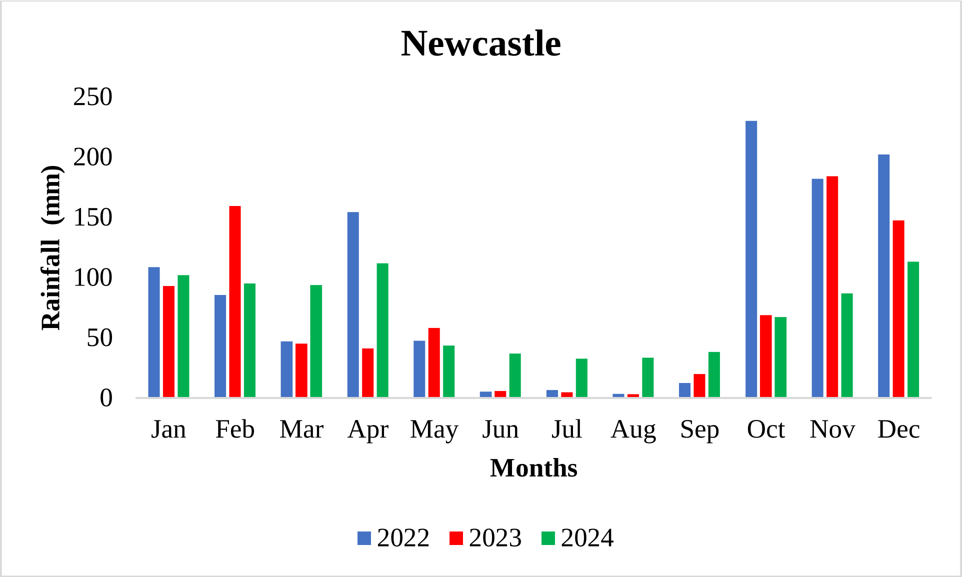


Figure 7. 6: Rainfall prediction for 2024 compared with 2023 and 2023 rainfall under the subtropical highland with dry winter (Harrismith, Johannesburg, and Newcastle).

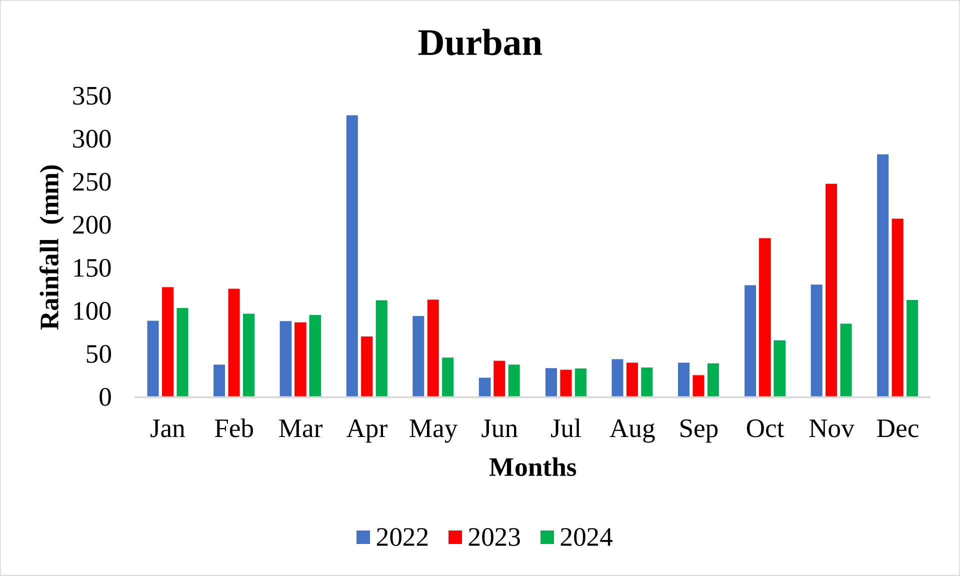
7.3 Subtropical dry Rainfall Prediction

7.3.1 Humid subtropical without dry season Rainfall Prediction

In figure 7.7, 2024 rainfall prediction with 2022 and 2023 rainfall is shown for Durban, Port Edward, and Richards Bay under the humid subtropical without dry season. These locations were the primary motivation for this study due to the flooding that occurred April 2022. Parts of KwaZulu natal province including Durban and Richards Bay and Port Edward in Eastern Cape received rainfall above expected average which led to the loss of lives, destruction of properties and infrastructures. As seen in the three figures in figure 7.7, all through the year, there is no dry season as rain is expected to fall with varying degrees of intensity. In Durban, predicted rainfall in January and February is about 20mm lower than 2023 rainfall but higher than 2022 rainfall in those months. However, in March, both the predicted rainfall and previously recorded rainfalls of 2022 and 2023 fall within the same range. As mentioned earlier in the study, April 2022 received an unusual amount of rainfall. Durban received almost 350mm of rain with most of them happening 11th and 12th April. This is responsible for the high rainfall profile in April as seen in the graph. December 2022 also received high rainfall, although it was more spread across the days compared to what happened in April. In 2023 April, rainfall received in Durban drastically reduced to about 70mm, it is predicted that this will increase in 2024 to over 110mm but will experience a decrease to about 50mm in May. The predicted rainfall for winter and early spring compared with those of 2022 and 2023 are within the same range. Therefore, the same amount of rainfall experienced and recorded in Durban in 2022 and 2023 winter and early spring should be expected in 2024. However, from October to December 2024, predicted rainfall is less than half what was recorded in 2023. This means that lesser rainfall should be expected towards the end of the year.

In Port Edward, 2024 rainfall predicted is within the same range as rainfall recorded in 2022 January and February. This was lower than 2023 January rainfall and more than 2023 February rainfall. In March, the predicted rainfall is about 40mm lower than what was recorded in 2022 and 2023. As with other cities under this climatic condition, 2022 April rainfall was about 300mm compared to 60mm recorded in 2023. It is predicted that there would be an increase by about 50mm in April rainfall from 2023. May predicted rainfall is about 20mm lower than rainfall recorded in 2022 and 2023. Similar to Durban, predicted 2024 rainfall and recorded rainfall for 2022 and 2023 during winter months are similar. Although rain would still be expected to fall, the intensity would be low. There was still high rainfall recorded in the spring of 2022 and December which is also predicted for 2024, but at lower amount. The amount of rain expected in September and October is less than a third of what was recorded during these months in 2022 and less than half of what was recorded in 2023 November is predicted for 2024 April. However, similar amount of rainfall should be expected in December of 2024 with December of 2023.

In Richards Bay, much rainfall was received in January, April, October, November, and December of 2022 with four months (April, October, November, and December) receiving over 200mm of rain. In 2023, there were five months (February, May, October, November, and December) that received rainfall exceeding 150mm. However, in 2024, it is predicted that only in January, April, and December will monthly rainfall exceed 100mm. This is reflected in the amount of annual rainfall expected which is over 300mm lower than the annual rainfall of 2023 and over 500mm lower than the average rainfall of 2022. Despite 2024 being predicted to have lower amounts of rainfall compared with the previous years, significant amount of rain is expected in the first four months of 2024 with March rainfall exceeding those of March 2022 and March 2023. It is also predicted that there will be more rain in April compared to 2023 April. The figure also shows that there would be rain during winter months which will gradually increase till the end of the year. Rainfall for October, November, and December is predicted to be at least 80mm lower than that of 2023 and 130mm lower than that of 2022.



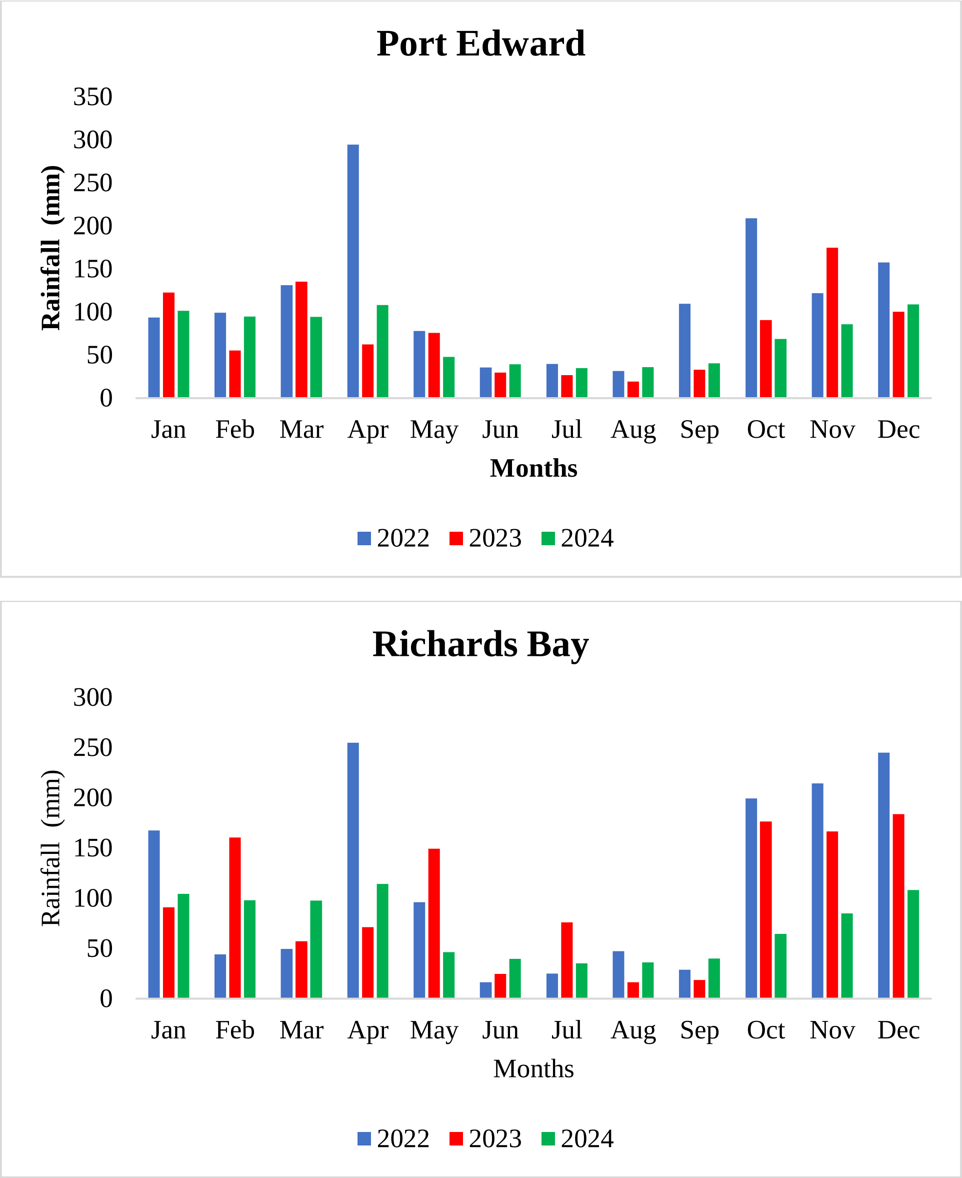


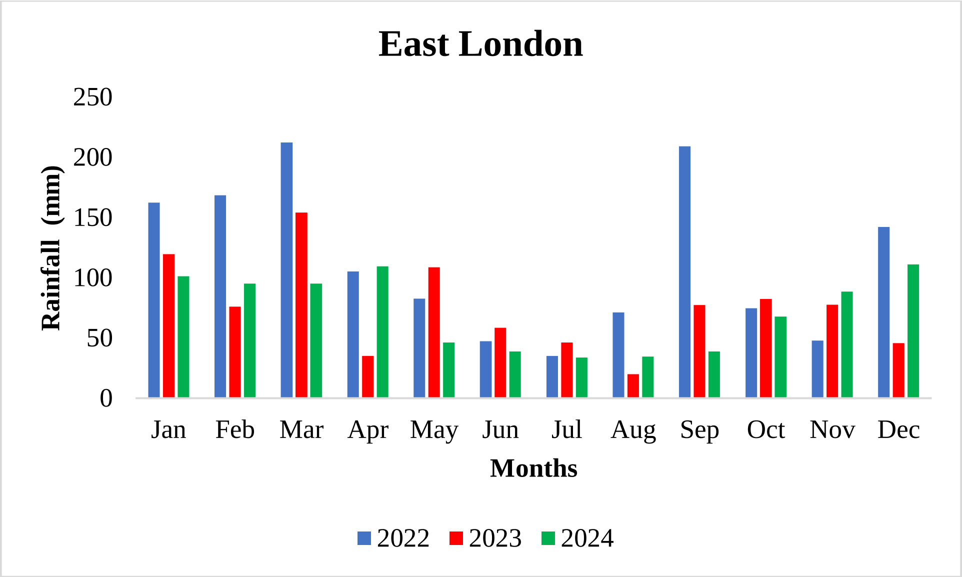
Figure 7. 7: Rainfall prediction for 2024 compared with 2023 and 2023 rainfall under the humid subtropical without dry season (Durban, Port Edward, and Richards Bay).

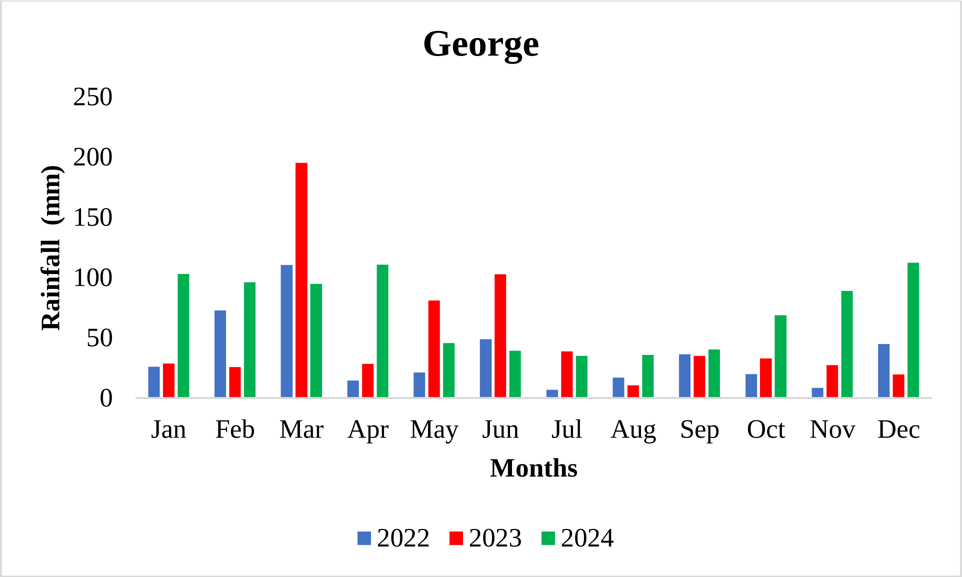
7.3.2 Temperate oceanic without dry season rainfall prediction

Rainfall prediction for the temperate oceanic without dry season (East London, George, and Mthatha) for 2024 compared with rainfall values for 2022 and 2023 is shown in figure 7.8. The first figure in figure 7.8 shows that while a steady increase in the amount of rainfall received in 2022 for the first three months is observed, in 2024, there is a slight decrease of less than 5mm between January and March. In 2023, there was a decrease by about 40mm between January and February and an increase of about 80mm between February and March. In April, the predicted values are within the same range with 2022 rainfall and thrice 2023 rainfall. Though there is no dry season in this climatic region, the amount of rainfall predicted and recorded for previous years reduced during winter months compared to other months. It is also seen in the graph that there would be lesser amount of rainfall between May and July in 2024 compared to 2023 as well as in September and October. The result also reveals that the annual average rainfall for 2024 will be lower than that of 2023 and much lower than 2022 annual rainfall.

However, in Geroge, there is an increase of about 200mm between 2022 annual rainfall and 2023 annual rainfall. Further 200mm increase rainfall is predicted between 2023 and 2024. This is an exciting news for a region that has been known to receive little amount of rainfall over the years. Predicted January rainfall is about four times what was recorded in 2022 and 2023. Also, predicted rainfall for February is about four times that of 2023 and about 20mm more than 2022 February rainfall in George. However, in March, there would lower rainfall compared to other years and about 100mm lower than 2023 March rainfall. This would be compensated in April as it is predicted that 2024 April rainfall would be about 90mm more than what was recorded in 2023 and 100mm more than 2022 April rainfall. Though for this climatic zone, there is no dry season, the result shows that in five months (April, July, August, October, and November) of 2022, rainfall was below 20mm with July receiving about 5mm the entire month. August 2023 also received rainfall below 10mm. it is expected that this value will rise to about 40mm in 2024. Finally, the results in George indicate that there would be more rainfall between October and December than what was recorded in 2022 and 2023 by a wide margin.

While in George, there was an increase of about 200mm from 2022 to 2023 and from 2023 to 2024, in Mthatha, the reverse is the case. There was an annual decrease of about 250mm from 2022 rainfall to 2023 rainfall. It is predicted that there would be a further decrease by about 300mm between 2023 rainfall and 2024 rainfall. As with this climatic zone, there was no dry season as rainfall was experienced and predicted during the winter season. It is predicted that January 2024 rainfall would be about half of what was recorded in 2023 as well as in March and May. February rainfall would be about 10mm more than 2023 February rainfall but 10mm lower than 2022 February rainfall. It is also predicted that there would be more rainfall in April 2024 by about 20mm compared to 2023 April. Between June and August, rainfall prediction for 2024 and recorded values for 2022 and 2023 are within the same range. Compared to 2023, lower rainfall is predicted in September and November while in October and December, the amount of rainfall expected would be similar to that of 2023.





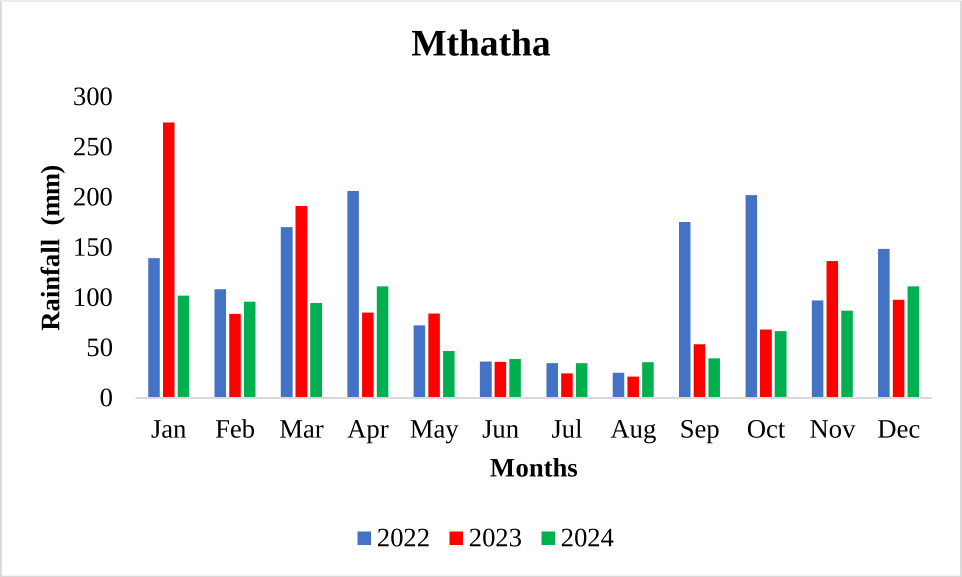


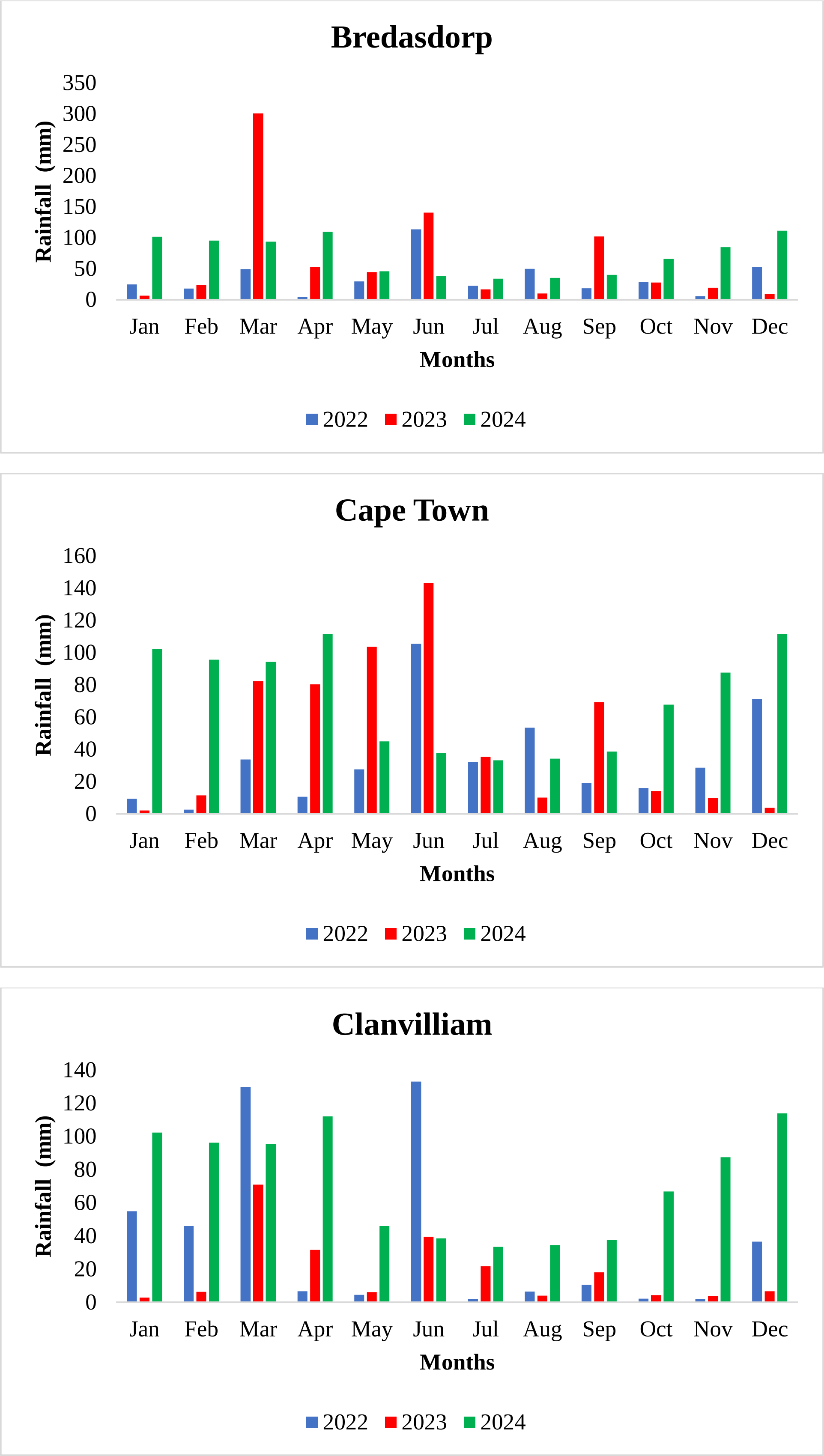
Figure 7. 8: Rainfall prediction for 2024 compared with 2023 and 2023 rainfall under the temperate oceanic without dry season (East London, George, and Mthatha).

7.3.3 Warm and dry summer

For the warm and dry summer, rainfall prediction for 2024 compared with recorded values of 2022 and 2023 is presented in figure 7.9. In Bredasdorp it is shown that there was little rainfall in summer months of December and January of 2023 while low rainfall was recorded in April and November of 2022. In the predicted values of rainfall, it is estimated that there would be much more rainfall from October to December than the previous years. Similarly, more rainfall is predicted in 2024 for January, February, and April than the previous years. it is predicted that the rain in March would be considerably lower than that of 2023 by almost 200mm while predicted rainfall for June 2024 would be about 100mm lower than the amount of rain received in 2023. Similar amount of rainfall to that of 2023 May and July is predicted for 2024.

In Cape Town, dry summer was only experienced in 2023 among the three years with 2022 December receiving considerable amount of rainfall. Most rainfall recorded in Cape Town in 2023 was during autumn, early winter, and early spring. In January 2023, less than 5mm rainfall was received compared to a predicted value of about 100mm in 2024 January. Similarly for February, the predicted value is about five times what it was in 2023 February. In March, from year to year, there is a steady increase in the amount of rainfall received from 2022 to 2024 although 2022 rainfall in March was much lower than in 2023 and 2024. This pattern is also seen in April with 2023 and 2024 values closer than what was obtained in 2024. While there was an increase in rain from April to June for both 2023 and 2024, the predicted values in 2024 is expected to reduce during this period. Predicted values in July for 2024 and recorded rainfall values for 2022 and 2023 are quite similar. The observed increase in 2022 rainfall from October to December can also be seen in the predicted values for 2024 while in 2023, there was a decrease in these months.

In Clanvilliam, significant amount of rainfall was recorded in 2022 in January, February, March, June, and December. Despite it being dry summer, in 2023, it was mostly a dry year with monthly rainfall below 20mm recorded in nine months (January, February, May, July, August, September, October, November, and December). In 2022, seven months recorded rainfall below 20mm (April, May, July, August, September, October, and November). In the predicted rainfall estimates of 2024, only four months (May, June, July, and August) are expected to have rainfall values below 40mm. While in 2023, only April had monthly rainfall of over 40mm recorded, in 2022, there was rainfall of about 130mm recorded in March and June. The predicted values of 2024 shows that rainfall of over 90mm is expected in five months (January, February, March, April, and December).



Bredasdorp

Cape Town

Clanvilliam

Figure 7. 9:Rainfall prediction for 2024 compared with 2023 and 2023 rainfall under the warm and dry summer (Bredasdorp, Cape Town, and Clanvilliam).

Chapter 8

8.0 Conclusion, Recommendations and Future Direction

8.1 Conclusion

This research work forms a foundation for future weather prediction in South Africa using machine learning models as it investigates the performance of four machine learning models over different climatic zones in South Africa. The machine learning models used for this study are linear regression, random forest, support vector machines, and ridge and lasso regression. Based on the Koppen-Geiger climate classification system, South Africa was divided into three groups and further subdivided. The broad three climatic zones are the arid, subtropical wet, and subtropical dry climate classification. The arid climate classification was further divided into four, the cold and semi-arid steppe, cold arid desert, hot and semi-arid steppe, and the hot arid desert. For each category, three locations were selected to test four machine learning models. The purpose of this is to determine if the same model can be used for rainfall prediction within the same climatic zone. These models were selected as they have shown to be accurate in rainfall prediction when tested in other regions of the world. However, much work had not been done in southern Africa and South Africa on prediction of atmospheric variables using machine learning methods.

The cold and semi-arid steppe climate is mainly found in the Free State Province, therefore, all three locations, Bloemfontein, Springfontein and Welkom are in the Free State Province. The result showed that either linear regression or support vector machines can be used to accurately predict the seasonality and estimate the amount of rainfall received in the cold and semi-arid steppe climate classification. Random forest and ridge and lasso also performed well in Springfontein and Welkom. Cold arid desert is majorly found in the western part of South Africa. locations selected are Alexander Bay and Bristown in the Northern Cape Province, and

Beaufort West in the Western Cape Province. All models performed poorly in Alexander Bay, in the prediction of rainfall. The reason for the poor performance could not be established in this study, however, all other models performed well in Beaufort West and Bristown. It is however suggested that either the support vector machine or ridge and lasso be used in Beaufort West. Hot and semi-arid climate classification was found in three provinces, Kimberly in the Northern Cape Province, Mahikeng in North-west Province, and Port Elizabeth in the Eastern Cape Province. In this climatic zone, all models performed poorly in predicting Port Elizabeth's rainfall while they all performed excellently in forecasting rainfall in Kimberly and Mahikeng. This study suggests that any of the model can be used for future work in the hot and semi-arid climate. Towards the northern part of South Africa is the hot arid desert. Locations selected for this study are Lauville in Mpumalanga Province, Musina in Limpopo province and Upington in Northern Cape province. The performance of the models in the hot arid desert was not as good compared to other regions in the arid classification except in Musina where the models performed well especially the support vector machine and random forest. Although other models still effectively predicted rainfall, the support vector machine is suggested to be used in other locations.

The subtropical wet climate classification was identified with dry winter. This was sub-divided into two, the humid subtropical highland with dry winter and the humid highland with dry winter. Dundee in KwaZulu Natal Province, Louis Trichardt in Limpopo Province, and Nelspruit in Mpumalanga Province were selected for the humid subtropical while Harrismith in the Free State Province, Johannesburg in Gauteng Province, and Newcastle in KwaZulu Natal Province were selected for the subtropical highlands. The regions of the subtropical highland are around the central part of South Africa. The support vector machine performed best in Louis Trichardt and Nelspruit while random forest performed best in Dundee. All four models can be used for rainfall prediction as they all had high predictive performance in the three locations under the humid subtropical highland. The performance of all models in Johannesburg was not too accurate. In regions where the model performance is not good enough, it is suggested that other atmospheric variables be considered such as different cloud properties. The performance of all models to predict rainfall were high in both Harrismith and Newcastle, therefore, any of the models can be used for future work.

For the third climate classification, Durban and Richards Bay from KwaZulu Natal Province, Port Edward from the Eastern Cape Province were selected for the humid subtropical without dry season while East London and Mthatha in Eastern Cape Province and George in Western Cape Province were selected for the temperate oceanic without dry season. The temperate oceanic without dry season is mostly found in the southern part of South Africa. The final classification under the subtropical dry is the warm and dry summer and locations selected are Bredasdorp, Cape Town, and Clanvilliam all in the Western Cape Province. The performance of the models in the humid subtropical was not too good except in Port Edwards, although results in Durban using ridge and lasso was better compared to other models. Model performance in Richards Bay was not good enough as with George in the temperate oceanic. Random forest performed best in Mthatha closely followed by the support vector machine while linear regression and ridge and lasso also performed well in Mthatha. In East London, support vector machine performed best closely followed by random forest then, ridge and lasso regression. The performance of the model in the warm and dry summer was also poor. This may be attributed to the amount of zonal wind and unpredictable atmospheric variability in this region. Also, the models' performance was not too great in regions where there is little rainfall or pronounced drought seasons.

In addition, the results showed that for rainfall prediction, cloud cover, dew point and water vapour had high correlation with rainfall. It is suggested that for future modelling and predictions, these atmospheric variables should be included to increase the performance of the models. Wind speed, temperature, and relative humidity only had strong impact on rainfall prediction is few locations.

Finally, using the random forest model, 2024 monthly rainfall was predicted for all 27 locations and compared with values obtained from 2022 and 2023. In some locations, it is expected that there would be more rainfall in 2024 compared to previous years while in some locations, lesser amount of rainfall was predicted. Where there would be more rainfall, it is good news for farmers, while in regions where lower rainfall were predicted, farmers need to look for alternative ways to combat the foreseeable challenges with low rainfall.

8.2 Recommendations

This work recommends the use of machine learning models for future studies in South Africa for all climatic zones.

This work also recommends that, in areas susceptible to drought and low annual rainfall, other atmospheric variables, especially cloud properties be explored for rainfall prediction and estimation.

For accurate prediction, dew point, cloud cover, and water vapour are important atmospheric variables that must be considered which will improve the accuracy of the models. However, in Cape Town, wind speed should be included in the atmospheric variables.

Government and organizations should look more into the results of this study to discover areas that were predicted to have both more and less rainfall and appropriate actions should be taken to avoid its adverse effect on the people.

8.3 Future Direction

Although predictions have been made for twenty-seven cities in South Africa, the prediction was done on monthly basis. This work can be extended by predicting daily rainfall as this will help guide more in decision making.

The reason for the low correlation between rainfall and other atmospheric variables in Alexander Bay, Port Elizabeth, Richards Bay, George, and the three locations under the warm and dry summer, Bredasdorp, Cape Town and Clanvilliam needs to be explored.

While machine learning models have proved to be useful, accurate and reliable in rainfall predictions, the potential of deep learning models such as long-short term memory has also been demonstrated as well as ensemble methods. In the future, deep learning models as well as a combination of different machine learning models would be used to daily rainfall prediction and compare their performance and potential advantages over traditional machine learning models.

Also, since Africa is lagging in the use of machine learning models for rainfall prediction, this work will be expanded to Southern African countries and Africa at large to help mitigate the impact of drought and enhance water resource management.

8.4 Limitations of the Study

This study was carried out in 27 cities located in different climatic zones using four machine learning models. Since only four models were examined, there is the possibility that better models may be adaptable to South Africa. Also, despite the 27 cities in the nine climatic zones in South Africa selected, many other cities were not included in the study.

References

Ali, M., Prasad, R., Xiang, Y. and Yaseen, Z.M., 2020. Complete ensemble empirical mode decomposition hybridized with random forest and kernel ridge regression model for monthly rainfall forecasts. Journal of Hydrology, 584, p.124647.

Ahuna, M.N., Afullo, T.J. and Alonge, A.A., 2019. Rain attenuation prediction using artificial neural network for dynamic rain fade mitigation. SAIEE Africa Research Journal, 110(1), pp.11-18.

Anochi, J.A., de Almeida, V.A. and de Campos Velho, H.F., 2021. Machine learning for climate precipitation prediction modeling over South America. Remote Sensing, 13(13), p.2468.

Appiah-Badu, N.K.A., Missah, Y.M., Amekudzi, L.K., Ussiph, N., Frimpong, T. and Ahene, E., 2021. Rainfall prediction using machine learning algorithms for the various ecological zones of Ghana. IEEE Access, 10, pp.5069-5082.

Bamisile, O., Oluwasanmi, A., Ejiyi, C., Yimen, N., Obiora, S. and Huang, Q., 2021. Comparison of machine learning and deep learning algorithms for hourly global/diffuse solar radiation predictions. International Journal of Energy Research.

Baran, Á., Lerch, S., El Ayari, M. and Baran, S., 2021. Machine learning for total cloud cover prediction. Neural Computing and Applications, 33(7), pp.2605-2620.

Baudoin, M.A., Vogel, C., Nortje, K. and Naik, M., 2017. Living with drought in South Africa: lessons learnt from the recent El Niño drought period. International journal of disaster risk reduction, 23, pp.128-137.

Belgiu, M. and Drăguţ, L., 2016. Random forest in remote sensing: A review of applications and future directions. ISPRS journal of photogrammetry and remote sensing, 114, pp.24-31.

Biau, G., 2012. Analysis of a random forests model. The Journal of Machine Learning Research, 13(1), pp.1063-1095.

Biau, G. and Scornet, E., 2016. A random forest guided tour. Test, 25, pp.197-227.

Bochenek, B. and Ustrnul, Z., 2022. Machine learning in weather prediction and climate analyses—applications and perspectives. Atmosphere, 13(2), p.180.

Bopape, M.J.M., Sebego, E., Ndarana, T., Maseko, B., Netshilema, M., Gijben, M., Landman, S., Phaduli, E., Rambuwani, G., Van Hemert, L. and Mkhwanazi, M., 2021. Evaluating South African weather service information on idai tropical cyclone and KwaZulu-natal flood events. South African Journal of Science, 117(3-4), pp.1-13.

Boswell, D., 2002. Introduction to support vector machines. Department of Computer Science and Engineering University of California San Diego, 11.

Botai, C.M., Botai, J.O., De Wit, J.P., Ncongwane, K.P. and Adeola, A.M., 2017. Drought characteristics over the western cape province, South Africa. Water, 9(11), p.876.

Bouras, E.H., Jarlan, L., Er-Raki, S., Balaghi, R., Amazirh, A., Richard, B. and Khabba, S., 2021. Cereal yield forecasting with satellite drought-based indices, weather data and regional climate indices using machine learning in Morocco. Remote Sensing, 13(16), p.3101.

Breiman, L., 2001. Random forests. Machine learning, 45, pp.5-32.

Cash, B.A., Burls, N.J. and Howar, L.V., 2023. Predictable and Unpredictable Components of Cape Town Winter Rainfall. Journal of Climate, pp.1-26.

Cedric, L.S., Adoni, W.Y.H., Aworka, R., Zoueu, J.T., Mutombo, F.K., Krichen, M. and Kimpolo, C.L.M., 2022. Crops yield prediction based on machine learning models: case of west african countries. Smart Agricultural Technology, p.100049.

Charpentier, G.V.M., Lafont, U. and Teixeira De Freitas, S., 2023. A settings Order Article Reprints Open AccessArticle Assessing the Long-Term Performance of Adhesive Joints in Space Structures during Interplanetary Exploration. Materials, 16(14).

Chen, C.Y., Yeh, N.C., Chuang, Y.C. and Lin, C.Y., 2022. Development of a Low-Cost Portable Cluster for Numerical Weather Prediction. Electronics, 11(17), p.2769.

Diez-Sierra, J. and del Jesus, M., 2020. Long-term rainfall prediction using atmospheric synoptic patterns in semi-arid climates with statistical and machine learning methods. Journal of Hydrology, 586, p.124789.

Ferreira, G.W. and Reboita, M.S., 2022. A new look into the South America precipitation regimes: Observation and Forecast. Atmosphere, 13(6), p.873.

Ghazikhani, A., Babaeian, I., Gheibi, M., Hajiaghaei-Keshteli, M. and Fathollahi-Fard, A.M., 2022. A smart post-processing system for forecasting the climate precipitation based on machine learning computations. Sustainability, 14(11), p.6624.

Gidey, E. and Mhangara, P., 2023. An Application of Machine-Learning Model for Analyzing the Impact of Land-Use Change on Surface Water Resources in Gauteng Province, South Africa. Remote Sensing, 15(16), p.4092.

Gómez, D., Aristizábal, E., García, E.F., Marín, D., Valencia, S. and Vásquez, M., 2023. Landslides forecasting using satellite rainfall estimations and machine learning in the Colombian Andean region. Journal of South American Earth Sciences, 125, p.104293.

He, S., Li, X., DelSole, T., Ravikumar, P. and Banerjee, A., 2021, May. Sub-seasonal climate forecasting via machine learning: Challenges, analysis, and advances. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 1, pp. 169-177).

He, R.R., Chen, Y., Huang, Q. and Kang, Y., 2019. LASSO as a tool for downscaling summer rainfall over the Yangtze River valley. Hydrological Sciences Journal, 64(1), pp.92-104.

He, R., Zhang, L. and Chew, A.W.Z., 2022. Modelling and predicting rainfall time series using seasonal-trend decomposition and machine learning. Knowledge-Based Systems, 251, p.109125.

Hewage, P., Trovati, M., Pereira, E. and Behera, A., 2021. Deep learning-based effective fine-grained weather forecasting model. Pattern Analysis and Applications, 24(1), pp.343-366.

Hossain, I., Rasel, H.M., Imteaz, M.A. and Mekanik, F., 2020. Long-term seasonal rainfall forecasting using linear and nonlinear modelling approaches: a case study for Western Australia. Meteorology and Atmospheric Physics, 132, pp.131-141.

Huntingford, C., Jeffers, E.S., Bonsall, M.B., Christensen, H.M., Lees, T. and Yang, H., 2019. Machine learning and artificial intelligence to aid climate change research and preparedness. Environmental Research Letters, 14(12), p.124007.

Ingreso, M.K., 2022. Short-term sea level forecasting using machine learning techniques: A case study for South Africa (Master's thesis, Faculty of Science).

Islam, M.S., Hossain, A., Khatun, A. and Kor, A.L., 2023, February. Evaluation of the Performance of Machine Learning and Deep Learning Techniques for Predicting Rainfall: An Illustrative Case Study from Australia. In 2023 International Conference on Electrical, Computer and Communication Engineering (ECCE) (pp. 1-5). IEEE.

Jeong, C.H. and Yi, M.Y., 2023. Correcting rainfall forecasts of a numerical weather prediction model using generative adversarial networks. The Journal of Supercomputing, 79(2), pp.1289-1317.

Jose, D.M., Vincent, A.M. and Dwarakish, G.S., 2022. Improving multiple model ensemble predictions of daily precipitation and temperature through machine learning techniques. Scientific Reports, 12(1), pp.1-25.

Jury, M.R., 2020. Climate trends in the Cape Town area, South Africa. Water SA, 46(3), pp.438-447.

Kisi, O. and Cimen, M., 2012. Precipitation forecasting by using wavelet-support vector machine conjunction model. Engineering Applications of Artificial Intelligence, 25(4), pp.783-792.

Kumari, K. and Yadav, S., 2018. Linear regression analysis study. Journal of the practice of Cardiovascular Sciences, 4(1), p.33.

Landman, S., Engelbrecht, F.A. and Engelbrecht, C.J., 2012. A short-range weather prediction system for South Africa based on a multi-model approach. Water SA, 38(5), pp.765-774.

Lotfirad, M., Esmaeili-Gisavandani, H. and Adib, A., 2022. Drought monitoring and prediction using SPI, SPEI, and random forest model in various climates of Iran. Journal of Water and Climate Change, 13(2), pp.383-406.

Lottering, N., Du Plessis, D. and Donaldson, R., 2015. Coping with drought: the experience of water sensitive urban design (WSUD) in the George Municipality. Water Sa, 41(1), pp.1-8.

Lucatero, D., Madsen, H., Refsgaard, J.C., Kidmose, J. and Jensen, K.H., 2018. Seasonal streamflow forecasts in the Ahlergaarde catchment, Denmark: the effect of preprocessing and post-processing on skill and statistical consistency. Hydrology and Earth System Sciences, 22(7), pp.3601-3617.

Mahlalela, P.T., Blamey, R.C., Hart, N.C.G. and Reason, C.J.C., 2020. Drought in the Eastern Cape region of South Africa and trends in rainfall characteristics. Climate Dynamics, 55, pp.2743-2759.

Mammone, A., Turchi, M. and Cristianini, N., 2009. Support vector machines. Wiley Interdisciplinary Reviews: Computational Statistics, 1(3), pp.283-289.

Mare, F., Bahta, Y.T. and Van Niekerk, W., 2018. The impact of drought on commercial livestock farmers in South Africa. Development in Practice, 28(7), pp.884-898.

Mashao, F.M., Mothapo, M.C., Munyai, R.B., Letsoalo, J.M., Mbokodo, I.L., Muofhe, T.P., Matsane, W. and Chikoore, H., 2023. Extreme rainfall and flood risk prediction over the East Coast of South Africa. Water, 15(1), p.50.

Maulud, D. and Abdulazeez, A.M., 2020. A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4), pp.140-147.

McDonald, G.C., 2009. Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), pp.93-100.

Mekanik, F., Imteaz, M.A., Gato-Trinidad, S. and Elmahdi, A., 2013. Multiple regression and Artificial Neural Network for long-term rainfall forecasting using large scale climate modes. Journal of Hydrology, 503, pp.11-21.

Meque, A., Gamedze, S., Moitlhobogi, T., Booneeady, P., Samuel, S. and Mpalang, L., 2021. Numerical weather prediction and climate modelling: Challenges and opportunities for improving climate services delivery in Southern Africa. Climate Services, 23, p.100243.

Milton, S., Diongue-Niang, A., Lamptey, B., Bain, C., Birch, C. and Bougeault, P., 2017. Numerical weather prediction over Africa. Meteorology of tropical West Africa: The forecasters' handbook, pp.380-422.

Moeletsi, M.E.A.R.C., Shabalala, Z.P.A.R.C., De Nysschen, G.A.R.C., Moeletsi, M.E. and Walker, S., 2016. Evaluation of an inverse distance weighting method for patching daily and dekadal rainfall over the Free State Province, South Africa. Water SA, 42(3), pp.466-474.

Moguerza, J.M. and Muñoz, A., 2006. Support vector machines with applications.

Mokgwathi, Z.J., 2018. Water Scarcity, Food Production & Dietary Choices of Rural Populations in Limpopo Province: A Study of Musina Local Municipality (Doctoral dissertation, University of the Witwatersrand, Faculty of Science).

Mountrakis, G., Im, J. and Ogole, C., 2011. Support vector machines in remote sensing: A review. ISPRS journal of photogrammetry and remote sensing, 66(3), pp.247-259.

Muyambo, F., Jordaan, A.J. and Bahta, Y.T., 2017. Assessing social vulnerability to drought in South Africa: Policy implication for drought risk reduction. Jàmbá: Journal of Disaster Risk Studies, 9(1), pp.1-7.

Nkamisa, M., Ndhleve, S., Nakin, M.D., Mngeni, A. and Kabiti, H.M., 2022. Analysis of trends, recurrences, severity, and frequency of droughts using standardised precipitation index: Case of OR Tambo District Municipality, Eastern Cape, South Africa. Jàmbá-Journal of Disaster Risk Studies, 14(1), p.1147.

Nkiruka, O., Prasad, R. and Clement, O., 2021. Prediction of malaria incidence using climate variability and machine learning. Informatics in Medicine Unlocked, 22, p.100508.

Noble, W.S., 2006. What is a support vector machine? Nature biotechnology, 24(12), pp.1565-1567.

Obiora, C.N., Ali, A. and Hassan, A.N., 2020, October. Predicting hourly solar irradiance using machine learning methods. In 2020 11th International Renewable Energy Congress (IREC) (pp. 1-6). IEEE.

Parker, R.H., 2020. Strategies in the Beaufort West region to mitigate the negative financial impacts of a drought (Doctoral dissertation, Stellenbosch: Stellenbosch University).

Peter, E.E. and Precious, E.E., 2018. Skill comparison of multiple-linear regression model and artificial neural network model in seasonal rainfall prediction-Northeast Nigeria. Asian Research Journal of Mathematics, 10(8), pp.1-10.

Pham, B.T., Le, L.M., Le, T.T., Bui, K.T.T., Le, V.M., Ly, H.B. and Prakash, I., 2020. Development of advanced artificial intelligence models for daily rainfall prediction. Atmospheric Research, 237, p.104845.

Polishchuk, B., Berko, A., Chyrun, L., Bublyk, M. and Schuchmann, V., 2021, September. The rain prediction in Australia based Big Data analysis and machine learning technology. In 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT) (Vol. 1, pp. 97-100). IEEE.

Raval, M., Sivashanmugam, P., Pham, V., Gohel, H., Kaushik, A. and Wan, Y., 2021. Automated predictive analytics tool for rainfall forecasting. Scientific Reports, 11(1), p.17704.

Ridwan, W.M., Sapitang, M., Aziz, A., Kushiar, K.F., Ahmed, A.N. and El-Shafie, A., 2021. Rainfall forecasting model using machine learning methods: Case study Terengganu, Malaysia. Ain Shams Engineering Journal, 12(2), pp.1651-1663.

Roffe, S.J., Fitchett, J.M. and Curtis, C.J., 2021. Investigating changes in rainfall seasonality across South Africa: 1987–2016. International Journal of Climatology, 41, pp.E2031-E2050.

Sachindra, D.A., Ahmed, K., Rashid, M.M., Shahid, S. and Perera, B.J.C., 2018. Statistical downscaling of precipitation using machine learning techniques. Atmospheric research, 212, pp.240-258.

Sajan, G.V. and Kumar, P., 2021, July. Forecasting and analysis of train delays and impact of weather data using machine learning. In 2021 12th International conference on computing communication and networking technologies (ICCCNT) (pp. 1-8). IEEE.

Sexton, D.M.H., Karmalkar, A.V., Murphy, J.M., Williams, K.D., Boutle, I.A., Morcrette, C.J., Stirling, A.J. and Vosper, S.B., 2019. Finding plausible and diverse variants of a climate model. Part 1: establishing the relationship between errors at weather and climate time scales. Climate Dynamics, 53, pp.989-1022.

Sivakumar, V. and Fazel-Rastgar, F., 2023, February. Heavy rainfall resulting from extreme weather disturbances in eastern coastal parts of South Africa: 11 April 2022. In Proceedings of the Earth and Environmental Sciences International Webinar Conference (pp. 161-186). Cham: Springer International Publishing.

Strydom, S., Savage, M.J. and Clulow, A.D., 2019. Long-term trends and variability in the dryland microclimate of the Northern Cape Province, South Africa. Theoretical and Applied Climatology, 137, pp.963-975.

Su, X., Yan, X. and Tsai, C.L., 2012. Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3), pp.275-294.

Sumbiri, D. and Afullo, T.J., 2021. An overview of rainfall fading prediction models for satellite links in Southern Africa. Progress In Electromagnetics Research B, 90, pp.187-205.

Sungkawa, I. and Rahayu, A., 2019. Extreme rainfall prediction using bayesian quantile regression in statistical downscaling modeling. Procedia Computer Science, 157, pp.406-413.

Swain, S., Patel, P. and Nandi, S., 2017, April. A multiple linear regression model for precipitation forecasting over Cuttack district, Odisha, India. In 2017 2nd international conference for convergence in technology (I2CT) (pp. 355-357). IEEE.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology, 58(1), pp.267-288.

Tiwari, N. and Singh, A., 2020, July. A novel study of rainfall in the Indian states and predictive analysis using machine learning algorithms. In 2020 International Conference on Computational Performance Evaluation (ComPE) (pp. 199-204). IEEE.

Tladi, T.M., Ndambuki, J.M., Olwal, T.O. and Rwanga, S.S., 2023. Groundwater Level Trend Analysis and Prediction in the Upper Crocodile (West) Basin, South Africa.

Van Tol, J., Julich, S., Bouwer, D. and Riddell, E.S., 2020. Hydrological response in a savanna hillslope under different rainfall regimes. Koedoe: African Protected Area Conservation and Science, 62(2), pp.1-10.

Wang, B., Lu, J., Yan, Z., Luo, H., Li, T., Zheng, Y. and Zhang, G., 2019, July. Deep uncertainty quantification: A machine learning approach for weather forecasting. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 2087-2095).

Yakubu, A.T., Abayomi, A. and Chetty, N., 2021, December. Machine Learning-Based Precipitation Prediction Using Cloud Properties. In International Conference on Hybrid Intelligent Systems (pp. 243-252). Cham: Springer International Publishing.

Yang, X. and Wen, W., 2018. Ridge and lasso regression models for cross-version defect prediction. IEEE Transactions on Reliability, 67(3), pp.885-896.

Yen, M.H., Liu, D.W., Hsin, Y.C., Lin, C.E. and Chen, C.C., 2019. Application of the deep learning for the prediction of rainfall in Southern Taiwan. Scientific reports, 9(1), p.12774.

Zaikarina, H., Djuraidah, A. and Wigena, A.H., 2016. Lasso and ridge quantile regression using cross validation to estimate extreme rainfall. Global Journal of Pure and Applied Mathematics, 12(3), pp.3305-3314.

Zhang, X., Mohanty, S.N., Parida, A.K., Pani, S.K., Dong, B. and Cheng, X., 2020. Annual and non-monsoon rainfall prediction modelling using SVR-MLP: an empirical study from Odisha. IEEE Access, 8, pp.30223-30233.

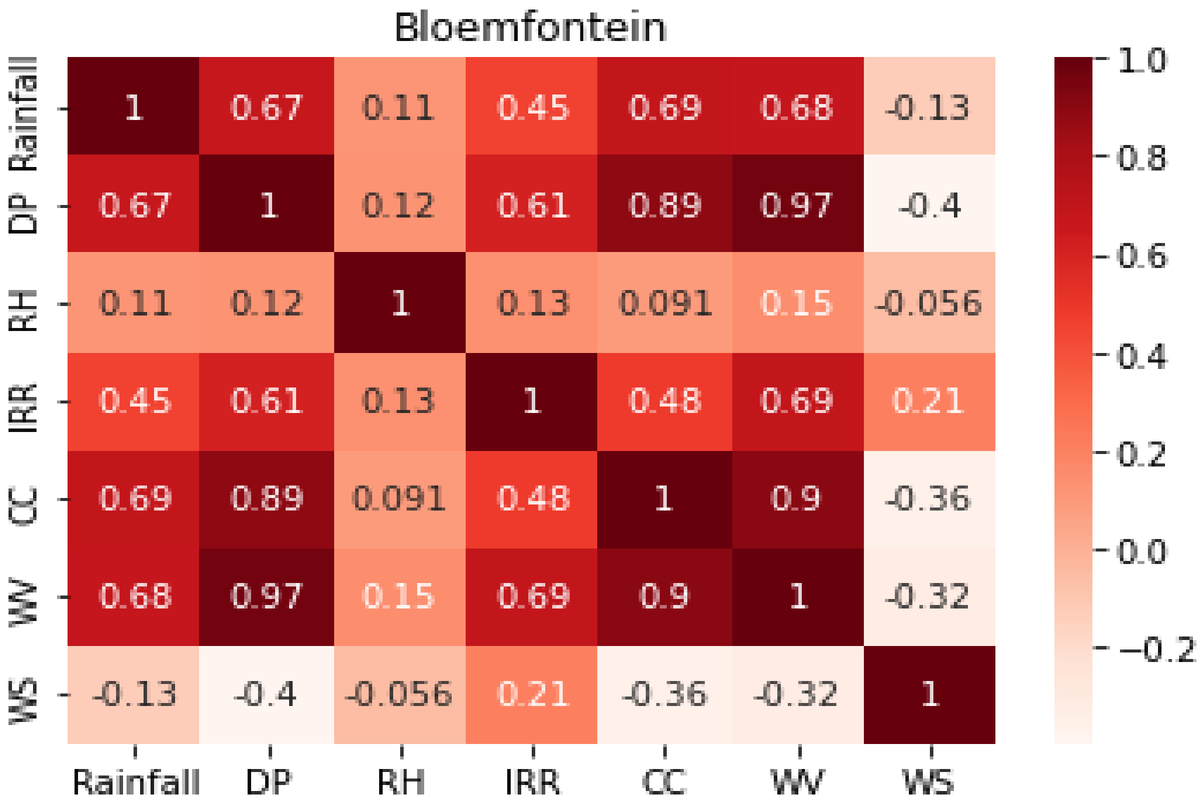
Zhang, Y., Yao, X., Wu, Q., Huang, Y., Zhou, Z., Yang, J. and Liu, X., 2021. Turbidity prediction of lake-type raw water using random forest model based on meteorological data: A case study of Tai lake, China. Journal of Environmental Management, 290, p.112657.

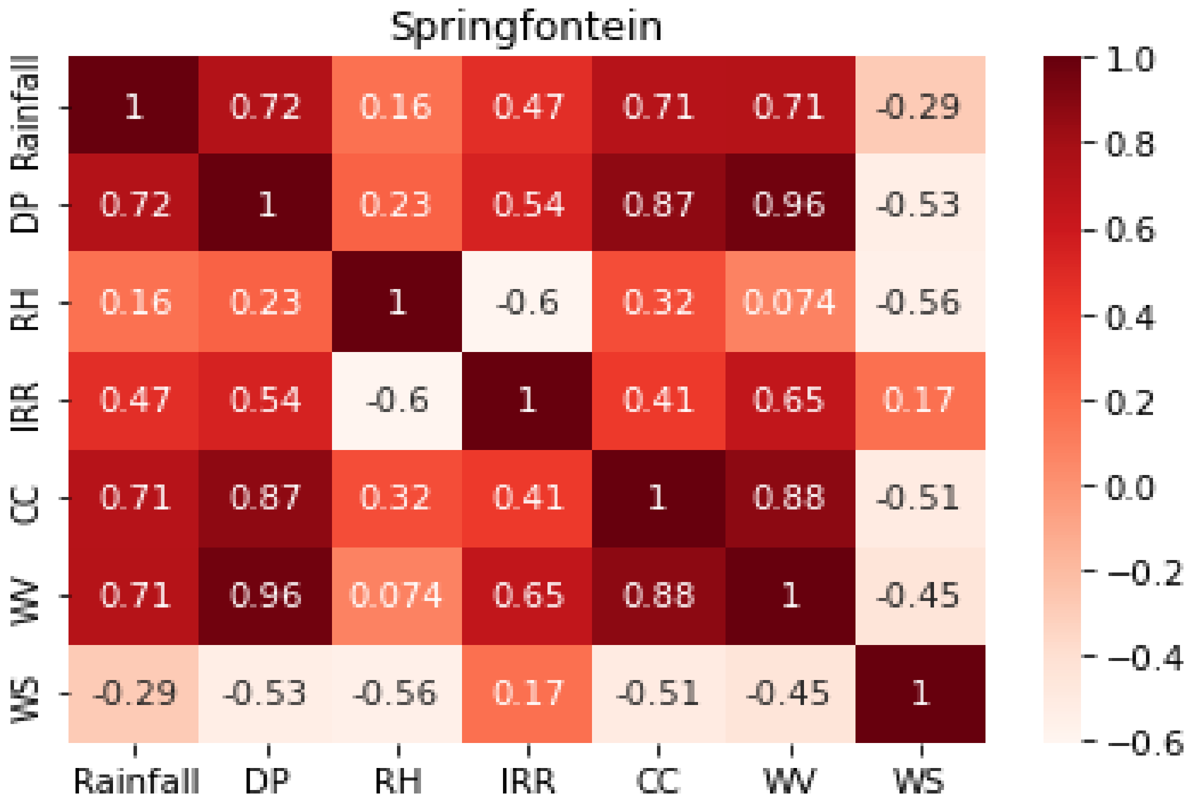
https://www.weathersa.co.za/Documents/Corporate/Annual%20State%20of%20the%20Climate%202021_04042022114230.pdf (last visited 14 December 2023)

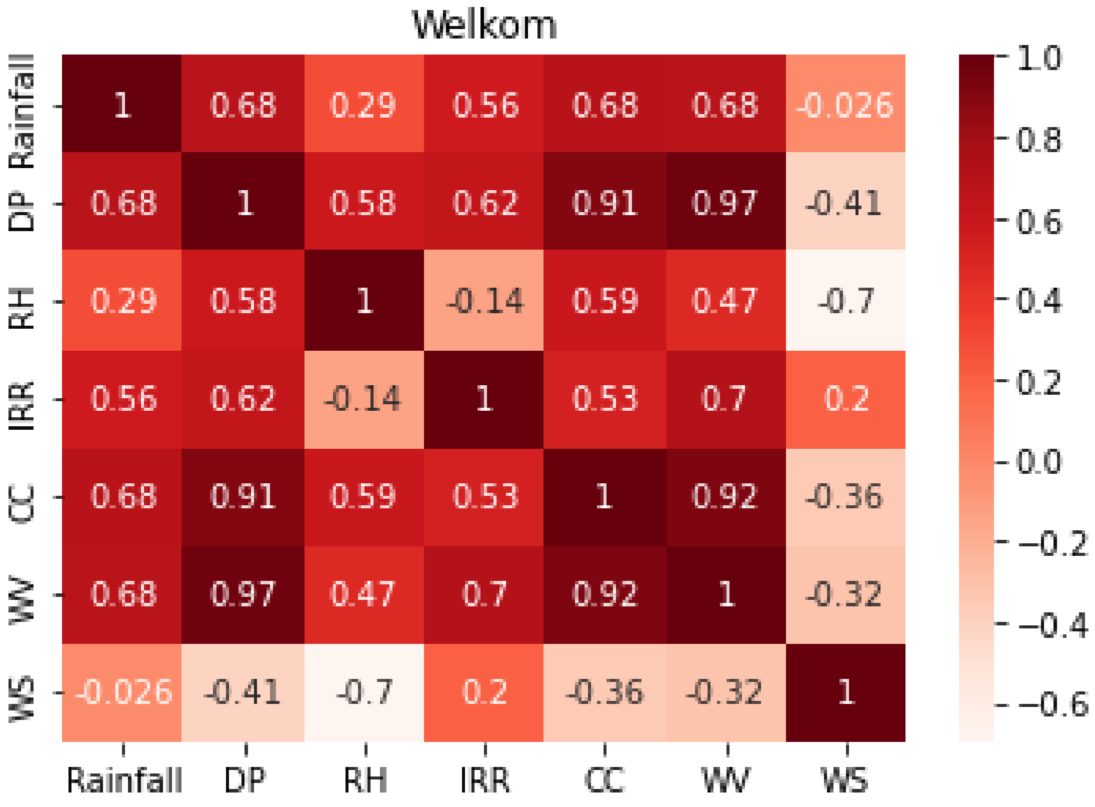
https://www.gcis.gov.za/sites/default/files/docs/resourcecentre/yearbook/2011/06_Land%20and%20its%20people.pdf (last visited 14 December 2023)

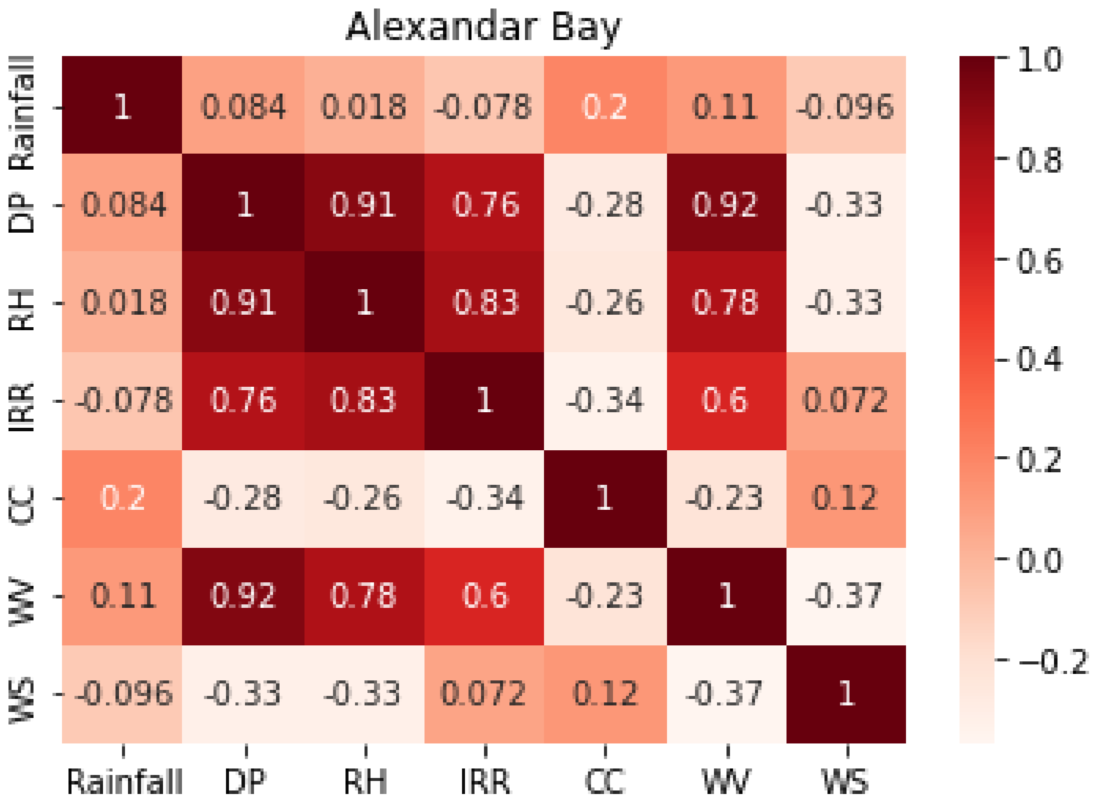
Appendix A

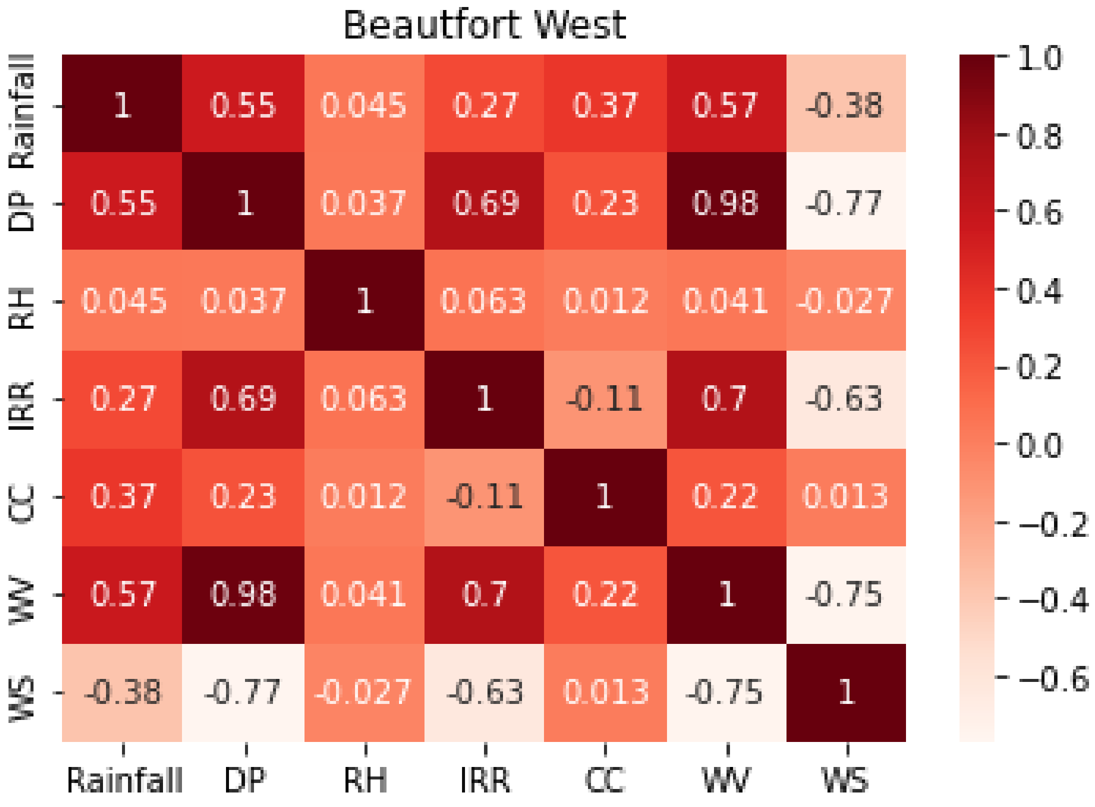
Heat map of the correlation between rainfall and atmospheric variables for various locations

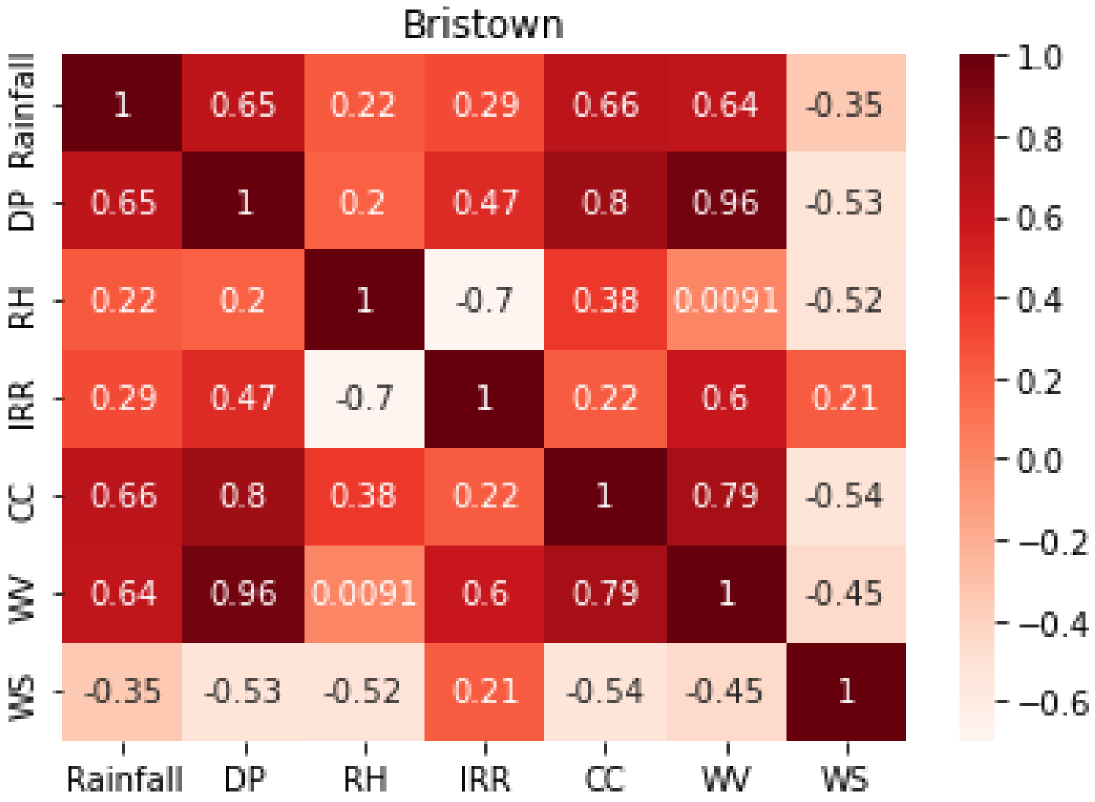


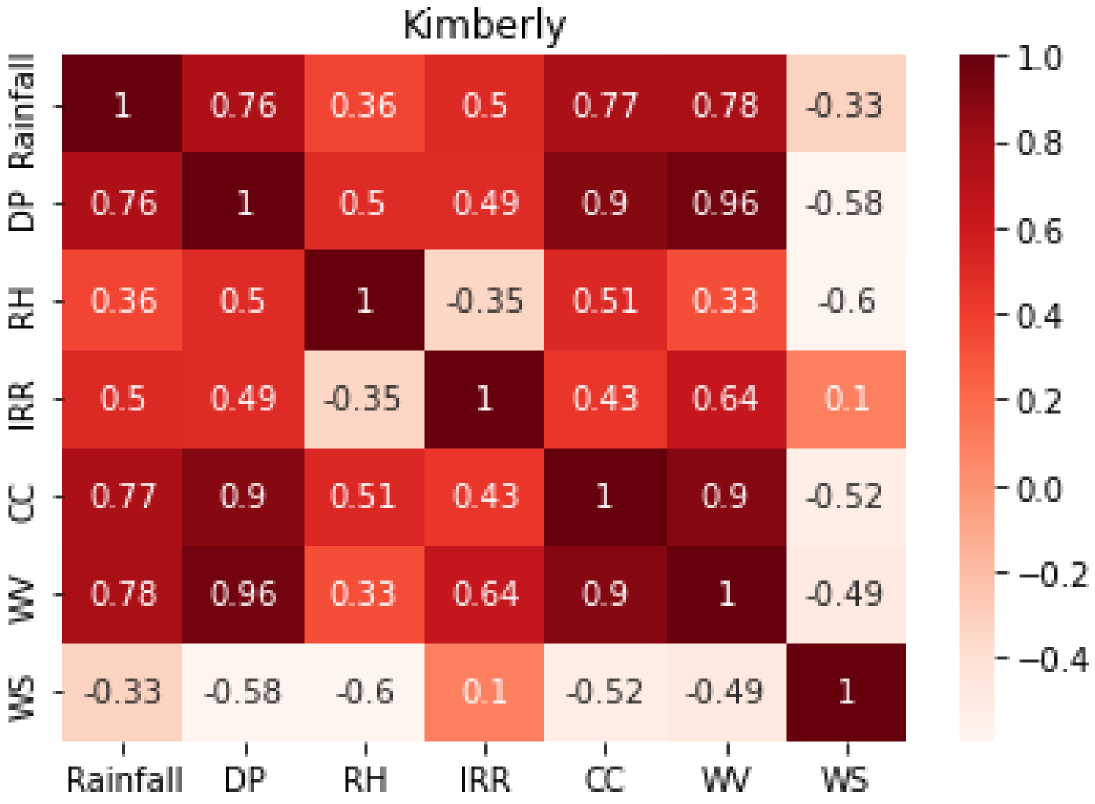

Springfontein
Rainfall
DP
RH
IRR
CC
WV
WS
1 0.72 0.16 0.47 0.71 0.71 -0.29
0.72 1 0.23 0.54 0.87 0.96 -0.53
0.16 0.23 1 -0.6 0.32 0.074 -0.56
0.47 0.54 -0.6 1 0.41 0.65 0.17
0.71 0.87 0.32 0.41 1 0.88 -0.51
0.71 0.96 0.074 0.65 0.88 1 -0.45
-0.29 -0.53 -0.56 0.17 -0.51 -0.45 1
Rainfall
DP
RH
IRR
CC
WV
WS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Welkom
Rainfall
DP
RH
IRR
CC
WV
WS
1 0.68 0.29 0.56 0.68 0.68 -0.026
0.68 1 0.58 0.62 0.91 0.97 -0.41
0.29 0.58 1 -0.14 0.59 0.47 -0.7
0.56 0.62 -0.14 1 0.53 0.7 0.2
0.68 0.91 0.59 0.53 1 0.92 -0.36
0.68 0.97 0.47 0.7 0.92 1 -0.32
-0.026 -0.41 -0.7 0.2 -0.36 -0.32 1
Rainfall
DP
RH
IRR
CC
WV
WS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

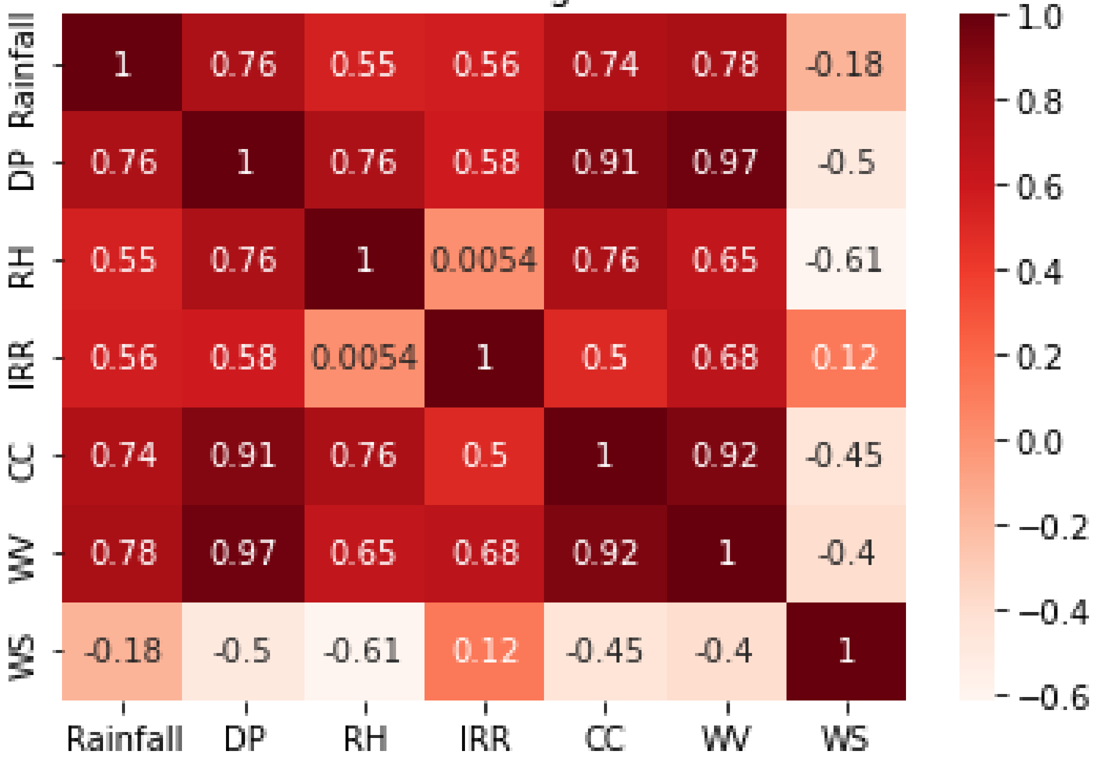

Alexandar Bay
Rainfall 1 0.084 0.018 -0.078 0.2 0.11 -0.096
DP 0.084 1 0.91 0.76 -0.28 0.92 -0.33
RH 0.018 0.91 1 0.83 -0.26 0.78 -0.33
IRR -0.078 0.76 0.83 1 -0.34 0.6 0.072
CC 0.2 -0.28 -0.26 -0.34 1 -0.23 0.12
WV 0.11 0.92 0.78 0.6 -0.23 1 -0.37
WS -0.096 -0.33 -0.33 0.072 0.12 -0.37 1
Rainfall DP RH IRR CC WV WS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Beautfort West
Rainfall 1 0.55 0.045 0.27 0.37 0.57 -0.38
DP 0.55 1 0.037 0.69 0.23 0.98 -0.77
RH 0.045 0.037 1 0.063 0.012 0.041 -0.027
IRR 0.27 0.69 0.063 1 -0.11 0.7 -0.63
CC 0.37 0.23 0.012 -0.11 1 0.22 0.013
WV 0.57 0.98 0.041 0.7 0.22 1 -0.75
WS -0.38 -0.77 -0.027 -0.63 0.013 -0.75 1
Rainfall DP RH IRR CC WV WS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

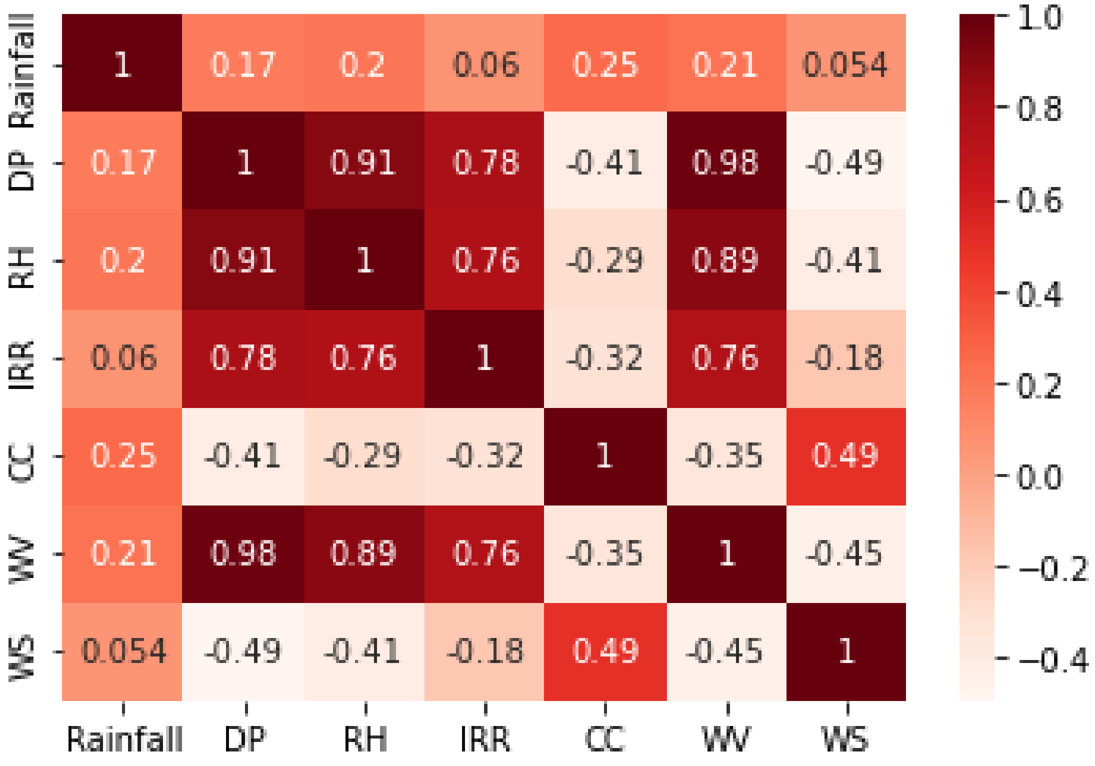

Bristown
Rainfall 1 0.65 0.22 0.29 0.66 0.64 -0.35
DP 0.65 1 0.2 0.47 0.8 0.96 -0.53
RH 0.22 0.2 1 -0.7 0.38 0.0091 -0.52
IRR 0.29 0.47 -0.7 1 0.22 0.6 0.21
CC 0.66 0.8 0.38 0.22 1 0.79 -0.54
WV 0.64 0.96 0.0091 0.6 0.79 1 -0.45
WS -0.35 -0.53 -0.52 0.21 -0.54 -0.45 1
Rainfall
DP
RH
IRR
CC
WV
WS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

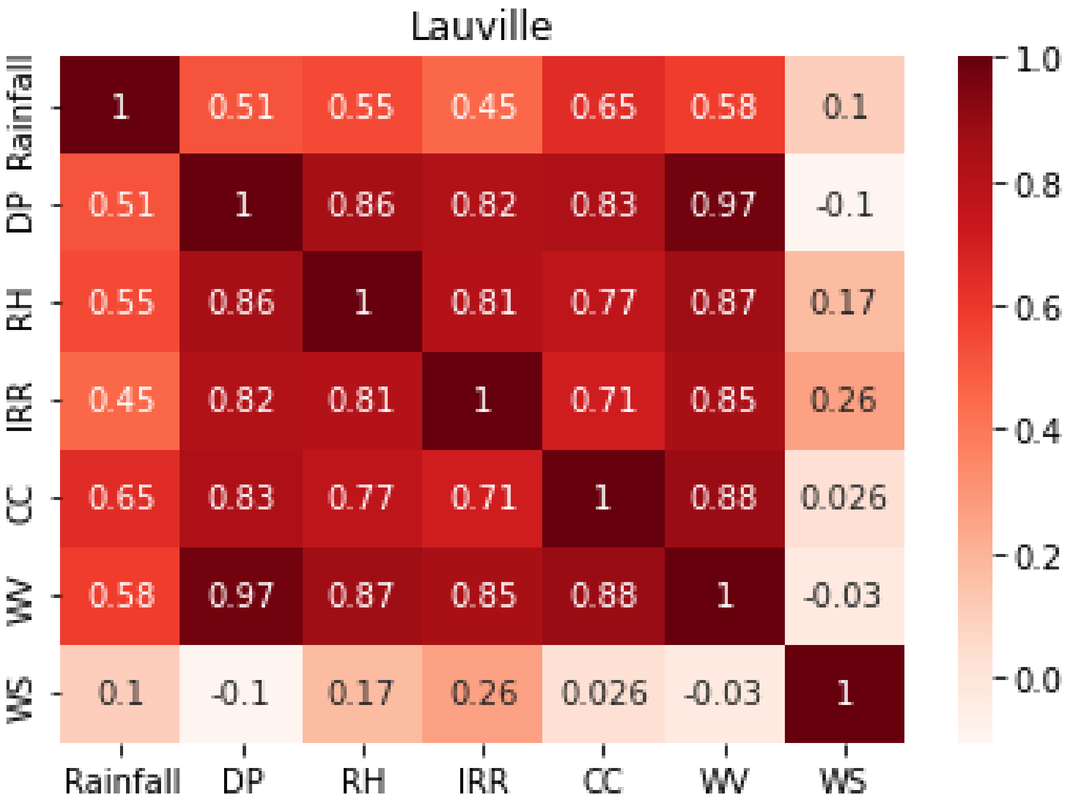

Kimberly
Rainfall 1 0.76 0.36 0.5 0.77 0.78 -0.33
DP 0.76 1 0.5 0.49 0.9 0.96 -0.58
RH 0.36 0.5 1 -0.35 0.51 0.33 -0.6
IRR 0.5 0.49 -0.35 1 0.43 0.64 0.1
CC 0.77 0.9 0.51 0.43 1 0.9 -0.52
WV 0.78 0.96 0.33 0.64 0.9 1 -0.49
WS -0.33 -0.58 -0.6 0.1 -0.52 -0.49 1
Rainfall
DP
RH
IRR
CC
WV
WS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

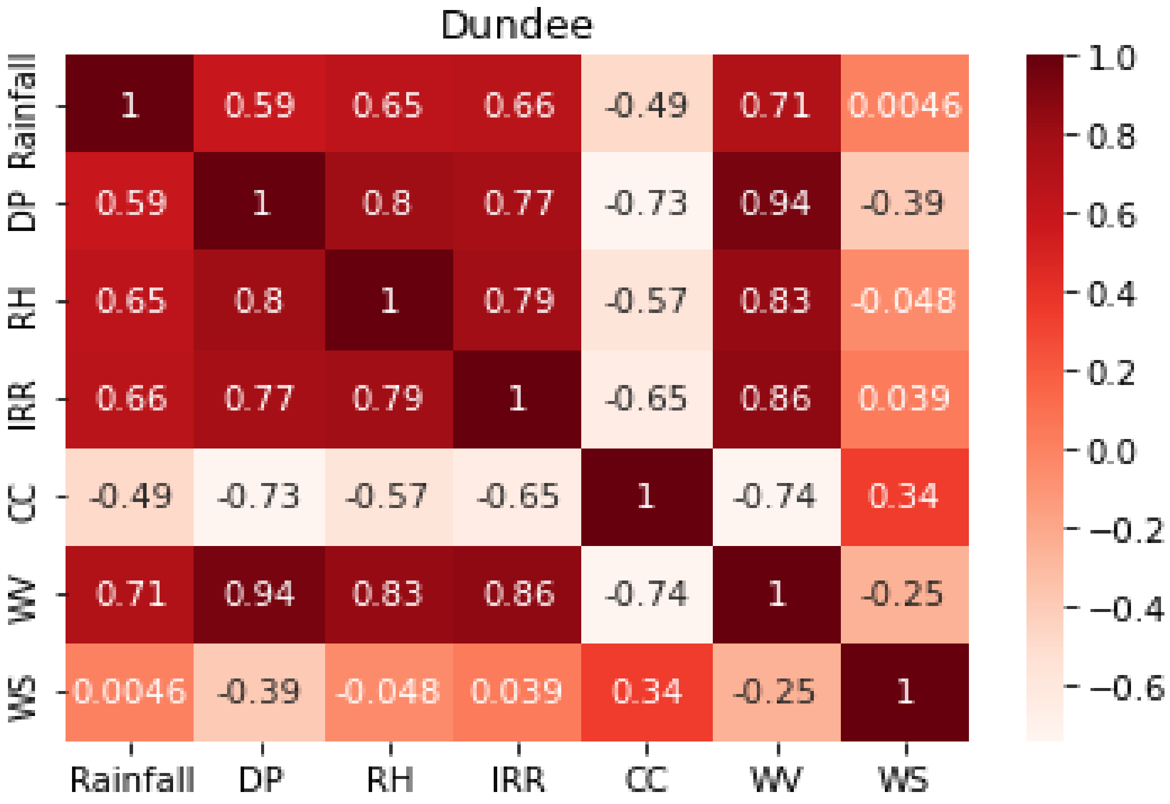
Mahikeng



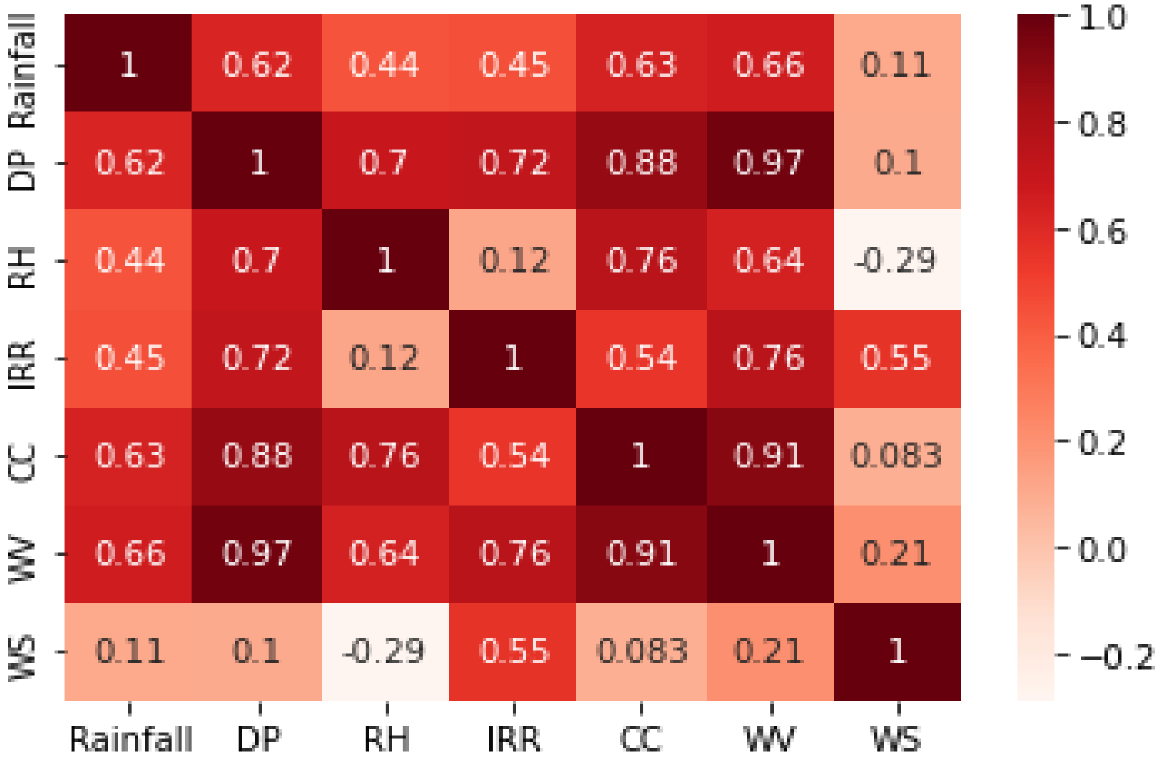
Port Elizabeth




Lauville
Rainfall 1 0.51 0.55 0.45 0.65 0.58 0.1
DP 0.51 1 0.86 0.82 0.83 0.97 -0.1
RH 0.55 0.86 1 0.81 0.77 0.87 0.17
IRR 0.45 0.82 0.81 1 0.71 0.85 0.26
CC 0.65 0.83 0.77 0.71 1 0.88 0.026
WV 0.58 0.97 0.87 0.85 0.88 1 -0.03
WS 0.1 -0.1 0.17 0.26 0.026 -0.03 1
Rainfall DP RH IRR CC WV WS
1.0
0.8
0.6
0.4
0.2
0.0

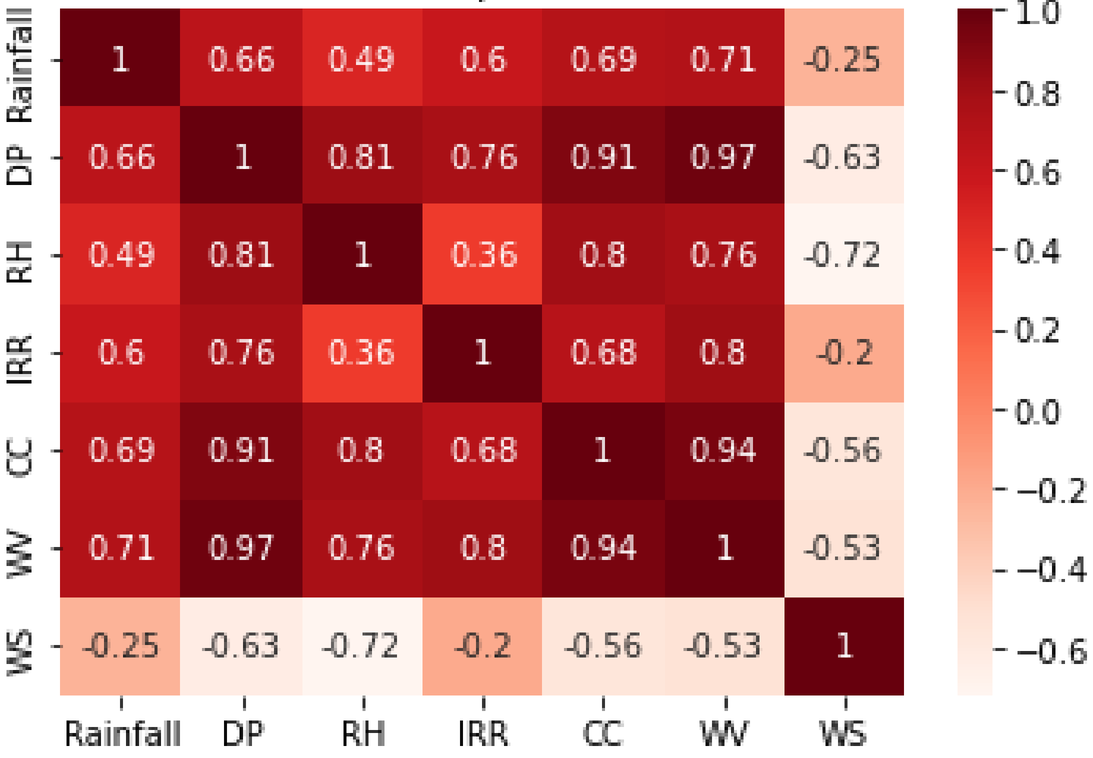

Dundee
Rainfall 1 0.59 0.65 0.66 -0.49 0.71 0.0046
DP 0.59 1 0.8 0.77 -0.73 0.94 -0.39
RH 0.65 0.8 1 0.79 -0.57 0.83 -0.048
IRR 0.66 0.77 0.79 1 -0.65 0.86 0.039
CC -0.49 -0.73 -0.57 -0.65 1 -0.74 0.34
WV 0.71 0.94 0.83 0.86 -0.74 1 -0.25
WS 0.0046 -0.39 -0.048 0.039 0.34 -0.25 1
Rainfall DP RH IRR CC WV WS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

Louis Trichardt

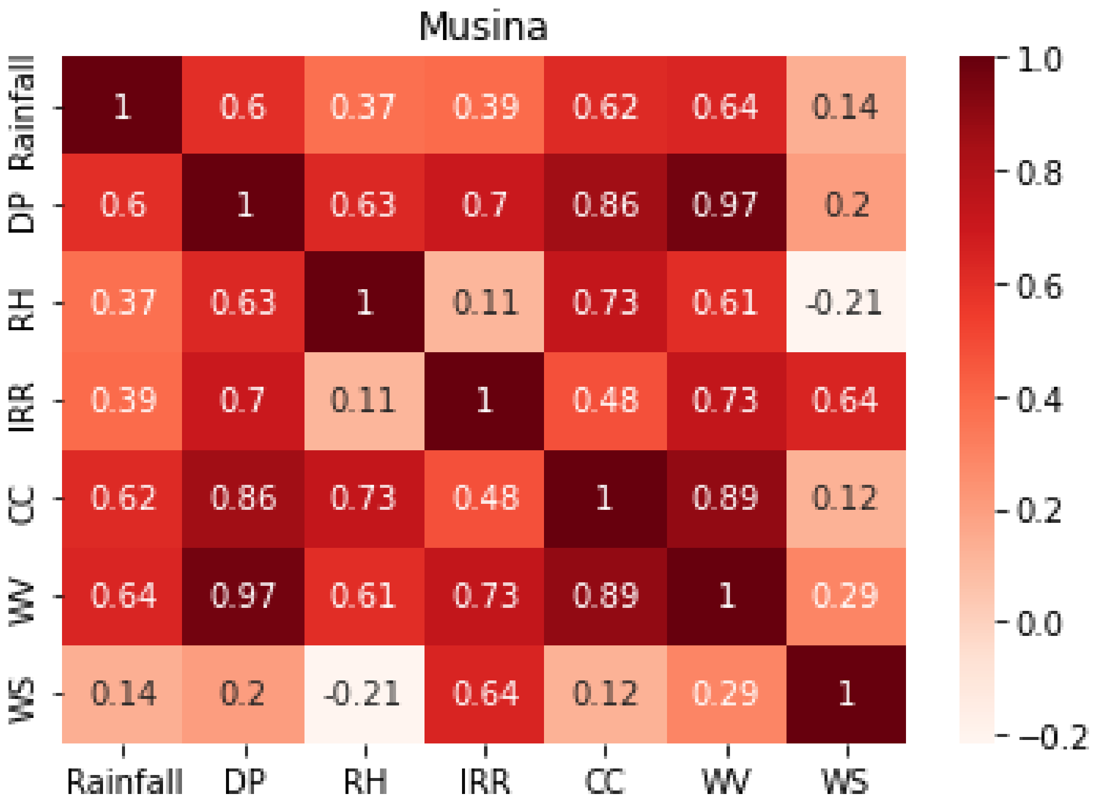


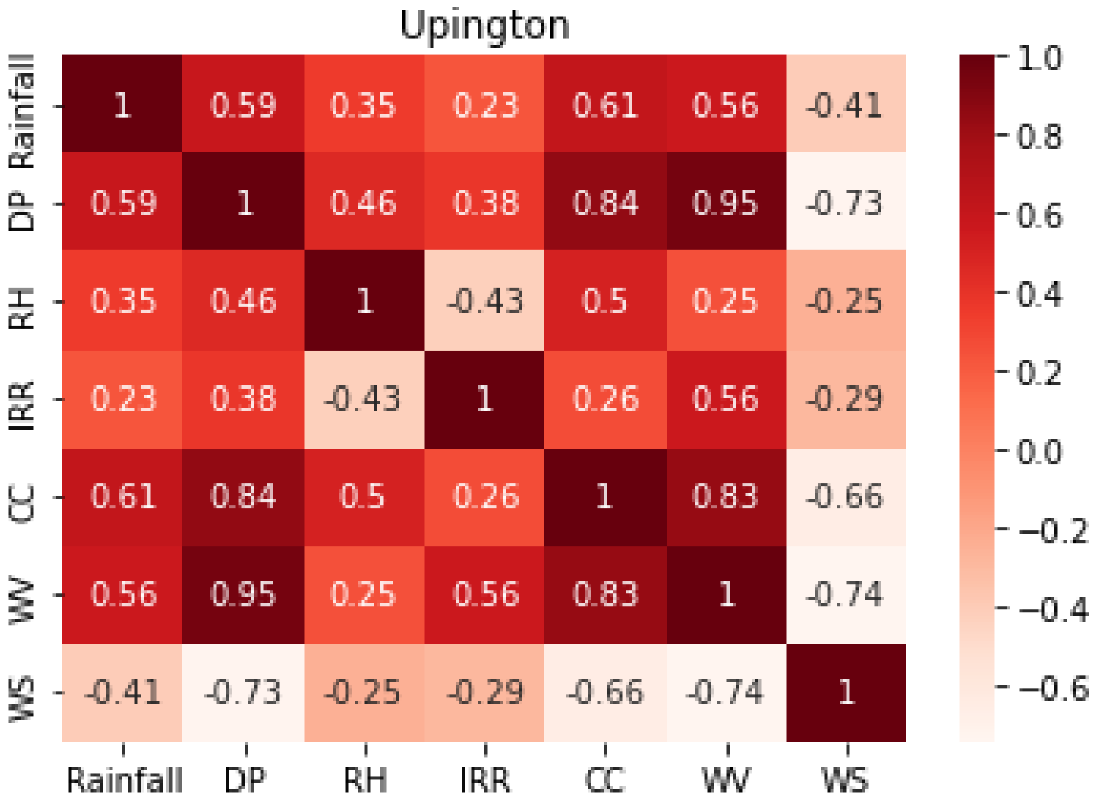
	Rainfall	DP	RH	IRR	CC	WV	WS
Rainfall	1	0.62	0.44	0.45	0.63	0.66	0.11
DP	0.62	1	0.7	0.72	0.88	0.97	0.1
RH	0.44	0.7	1	0.12	0.76	0.64	-0.29
IRR	0.45	0.72	0.12	1	0.54	0.76	0.55
CC	0.63	0.88	0.76	0.54	1	0.91	0.083
WV	0.66	0.97	0.64	0.76	0.91	1	0.21
WS	0.11	0.1	-0.29	0.55	0.083	0.21	1

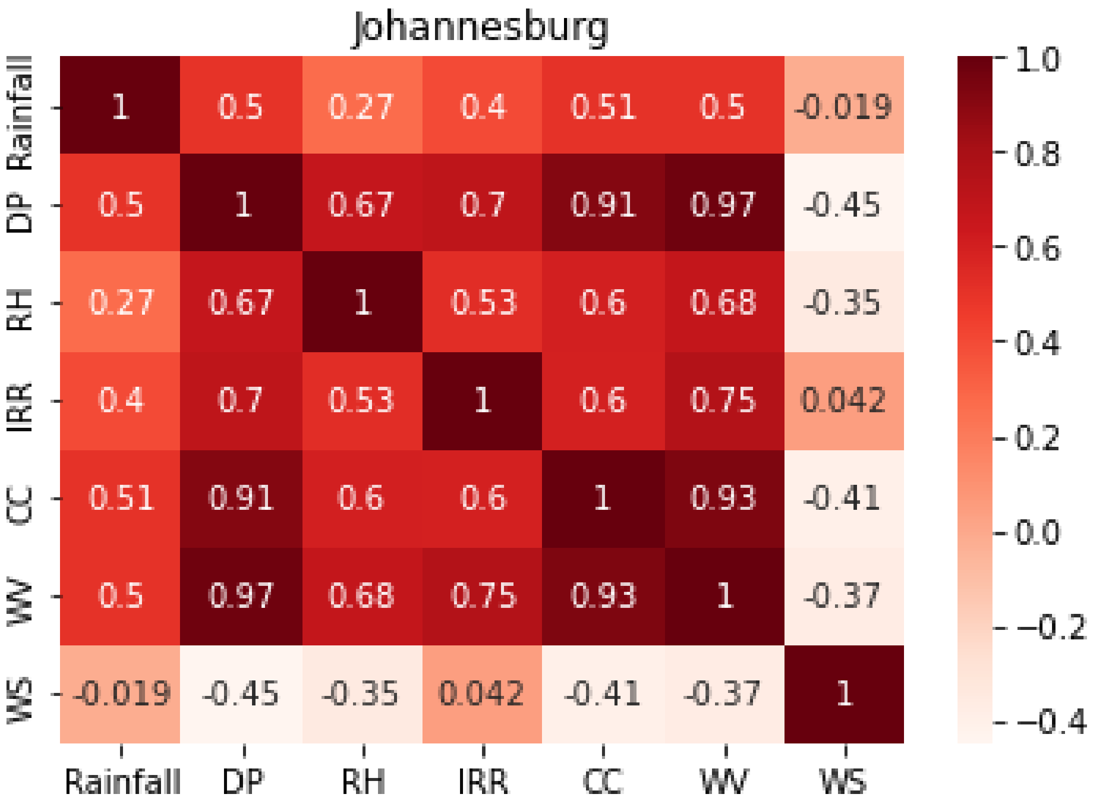
Nelspruit

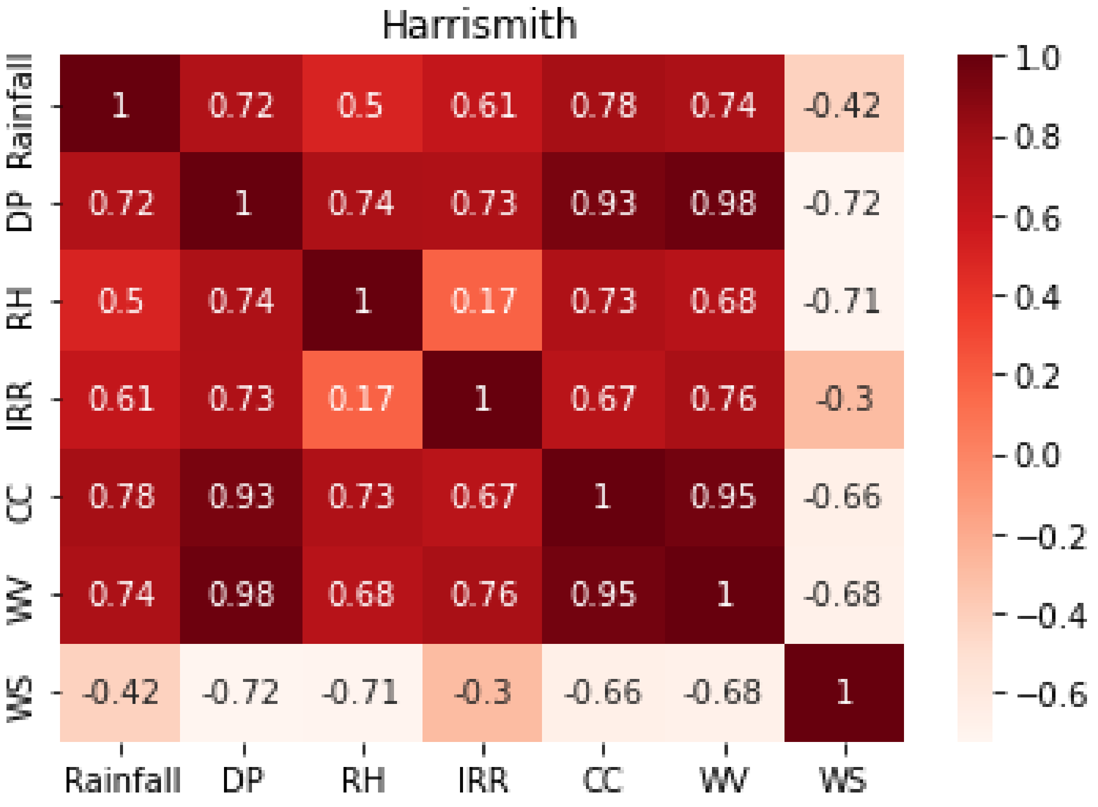


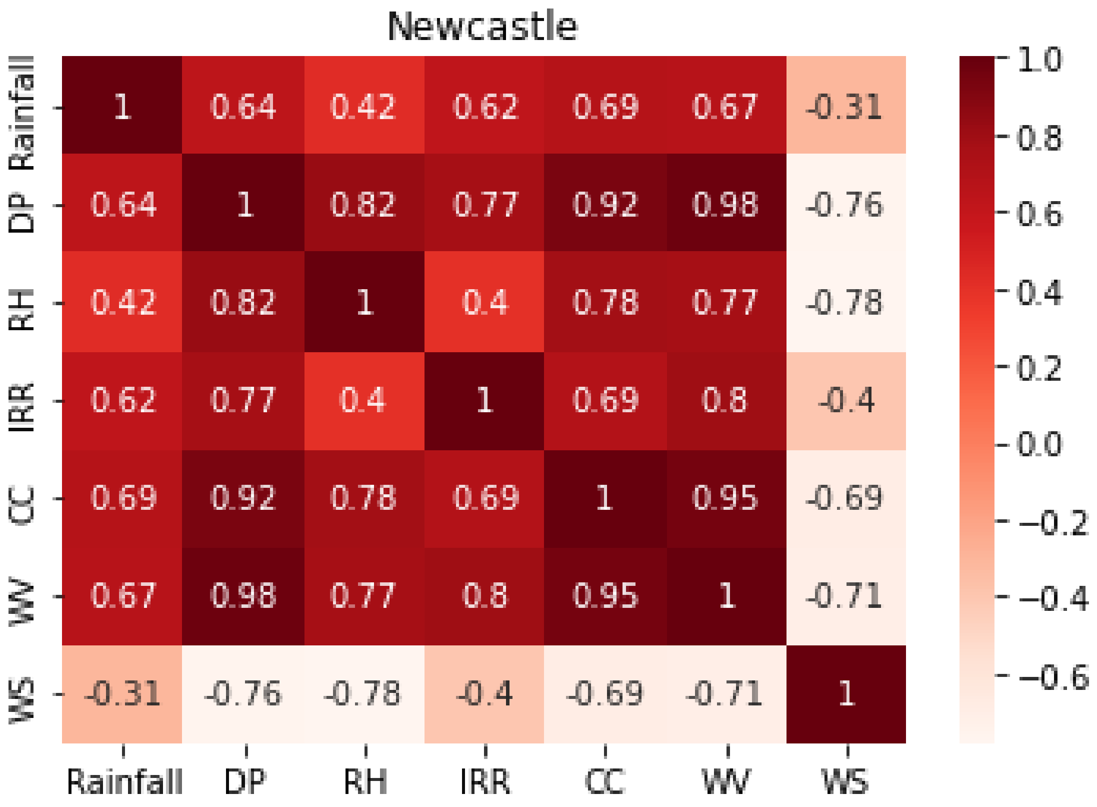
	Rainfall	DP	RH	IRR	CC	WV	WS
Rainfall	1	0.66	0.49	0.6	0.69	0.71	-0.25
DP	0.66	1	0.81	0.76	0.91	0.97	-0.63
RH	0.49	0.81	1	0.36	0.8	0.76	-0.72
IRR	0.6	0.76	0.36	1	0.68	0.8	-0.2
CC	0.69	0.91	0.8	0.68	1	0.94	-0.56
WV	0.71	0.97	0.76	0.8	0.94	1	-0.53
WS	-0.25	-0.63	-0.72	-0.2	-0.56	-0.53	1

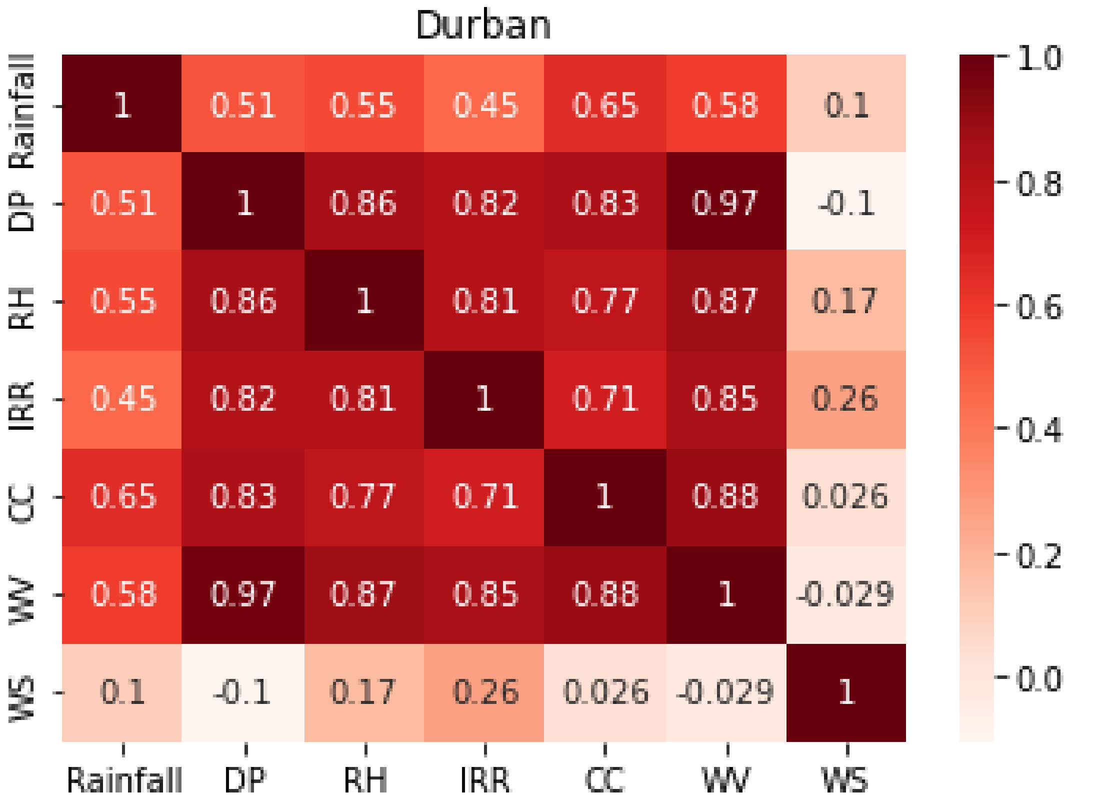

Musina
Rainfall 1 0.6 0.37 0.39 0.62 0.64 0.14
DP 0.6 1 0.63 0.7 0.86 0.97 0.2
RH 0.37 0.63 1 0.11 0.73 0.61 -0.21
IRR 0.39 0.7 0.11 1 0.48 0.73 0.64
CC 0.62 0.86 0.73 0.48 1 0.89 0.12
WV 0.64 0.97 0.61 0.73 0.89 1 0.29
WS 0.14 0.2 -0.21 0.64 0.12 0.29 1
Rainfall DP RH IRR CC WV WS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

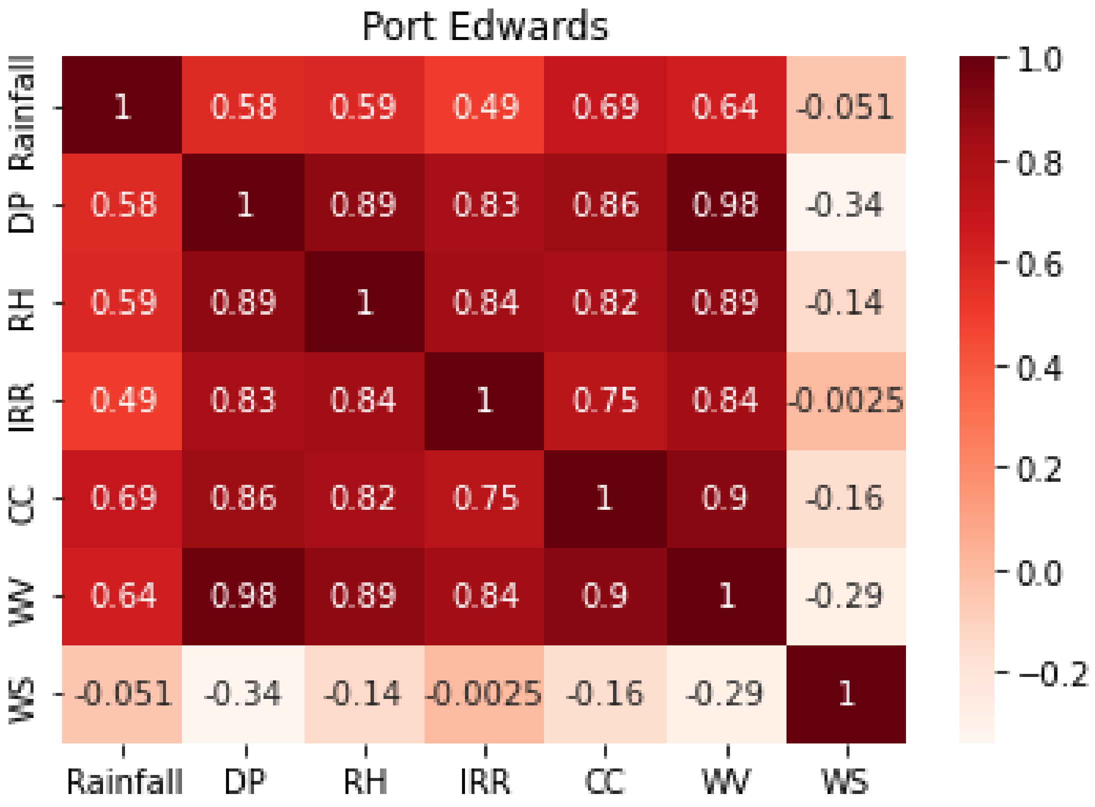

Upington
Rainfall 1 0.59 0.35 0.23 0.61 0.56 -0.41
DP 0.59 1 0.46 0.38 0.84 0.95 -0.73
RH 0.35 0.46 1 -0.43 0.5 0.25 -0.25
IRR 0.23 0.38 -0.43 1 0.26 0.56 -0.29
CC 0.61 0.84 0.5 0.26 1 0.83 -0.66
WV 0.56 0.95 0.25 0.56 0.83 1 -0.74
WS -0.41 -0.73 -0.25 -0.29 -0.66 -0.74 1
Rainfall DP RH IRR CC WV WS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


Johannesburg
Rainfall 1 0.5 0.27 0.4 0.51 0.5 -0.019
DP 0.5 1 0.67 0.7 0.91 0.97 -0.45
RH 0.27 0.67 1 0.53 0.6 0.68 -0.35
IRR 0.4 0.7 0.53 1 0.6 0.75 0.042
CC 0.51 0.91 0.6 0.6 1 0.93 -0.41
WV 0.5 0.97 0.68 0.75 0.93 1 -0.37
WS -0.019 -0.45 -0.35 0.042 -0.41 -0.37 1
Rainfall DP RH IRR CC WV WS
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4

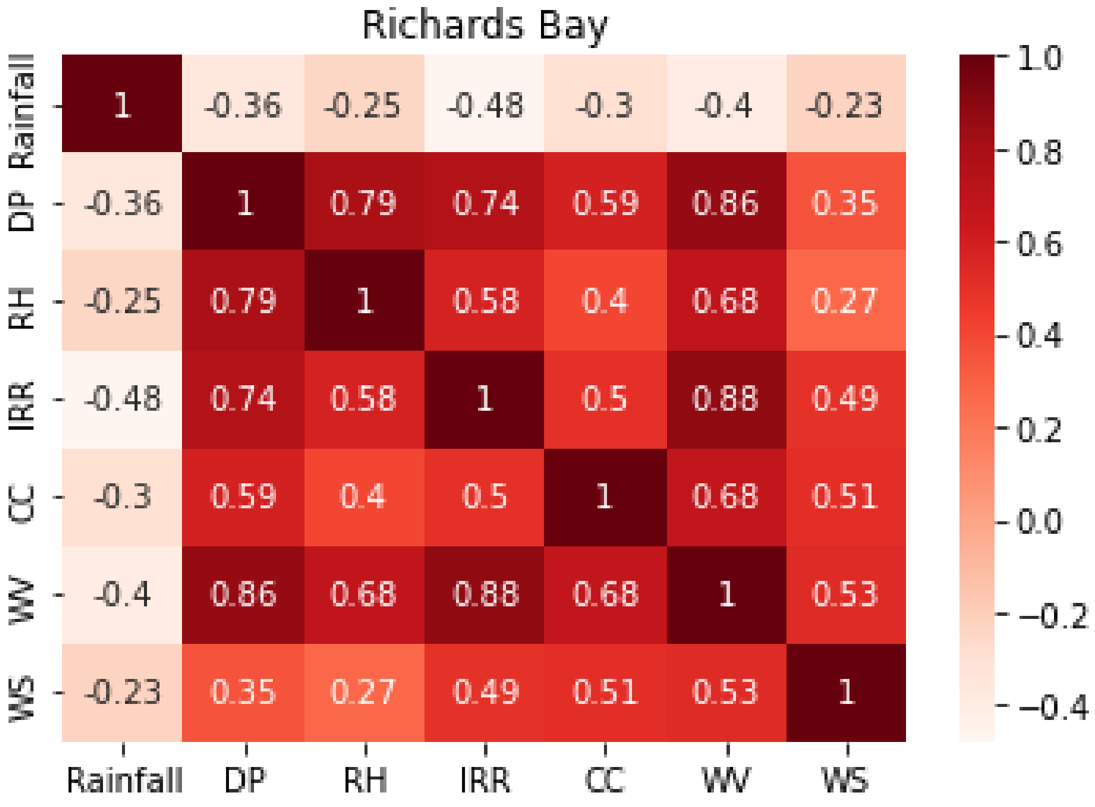

Harrismith
Rainfall 1 0.72 0.5 0.61 0.78 0.74 -0.42
DP 0.72 1 0.74 0.73 0.93 0.98 -0.72
RH 0.5 0.74 1 0.17 0.73 0.68 -0.71
IRR 0.61 0.73 0.17 1 0.67 0.76 -0.3
CC 0.78 0.93 0.73 0.67 1 0.95 -0.66
WV 0.74 0.98 0.68 0.76 0.95 1 -0.68
WS -0.42 -0.72 -0.71 -0.3 -0.66 -0.68 1
Rainfall DP RH IRR CC WV WS
1.0 0.8 0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6


Newcastle
Rainfall
DP
RH
IRR
CC
WV
WS
1 0.64 0.42 0.62 0.69 0.67 -0.31
0.64 1 0.82 0.77 0.92 0.98 -0.76
0.42 0.82 1 0.4 0.78 0.77 -0.78
0.62 0.77 0.4 1 0.69 0.8 -0.4
0.69 0.92 0.78 0.69 1 0.95 -0.69
0.67 0.98 0.77 0.8 0.95 1 -0.71
-0.31 -0.76 -0.78 -0.4 -0.69 -0.71 1
Rainfall
DP
RH
IRR
CC
WV
WS
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6

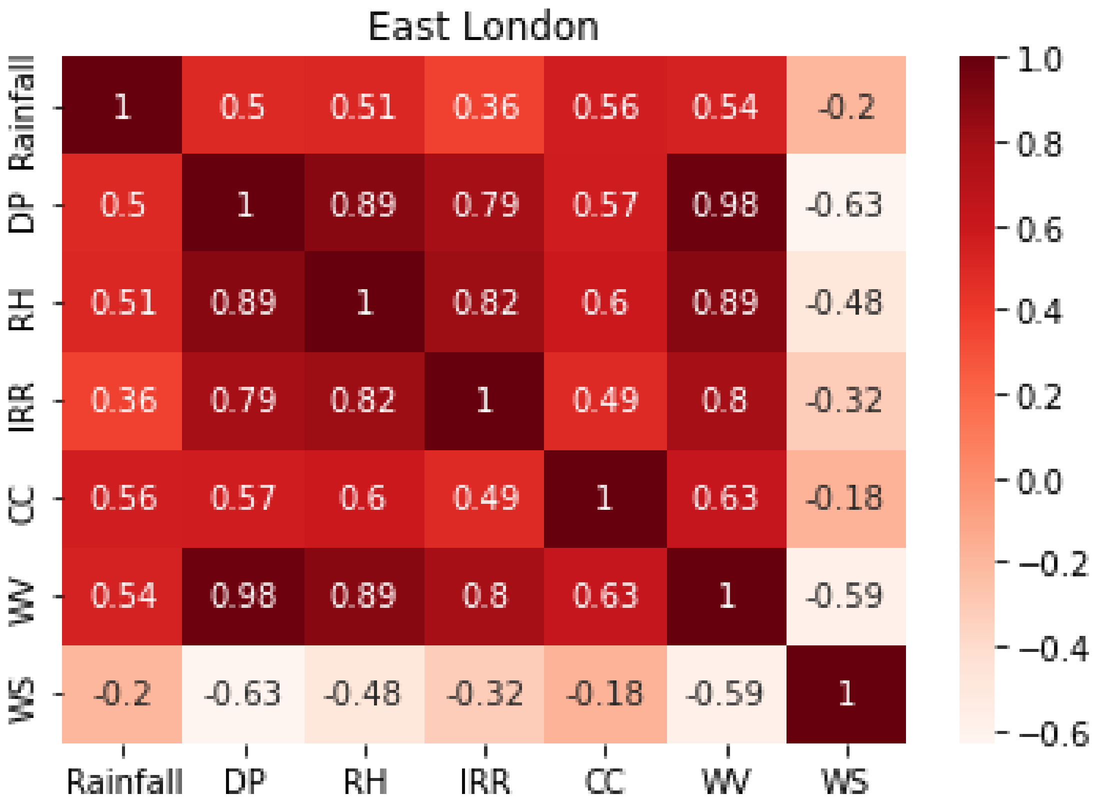

Durban
Rainfall
DP
RH
IRR
CC
WV
WS
1 0.51 0.55 0.45 0.65 0.58 0.1
0.51 1 0.86 0.82 0.83 0.97 -0.1
0.55 0.86 1 0.81 0.77 0.87 0.17
0.45 0.82 0.81 1 0.71 0.85 0.26
0.65 0.83 0.77 0.71 1 0.88 0.026
0.58 0.97 0.87 0.85 0.88 1 -0.029
0.1 -0.1 0.17 0.26 0.026 -0.029 1
Rainfall
DP
RH
IRR
CC
WV
WS
1.0
0.8
0.6
0.4
0.2
0.0



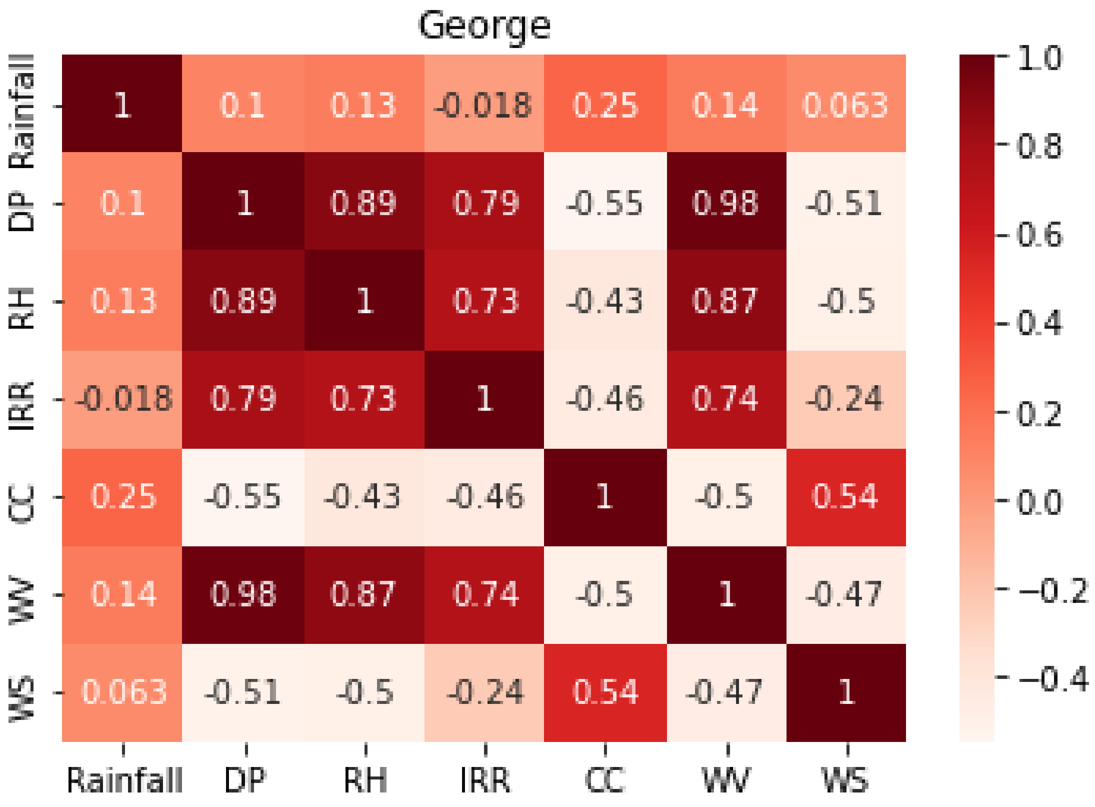
Port Edwards



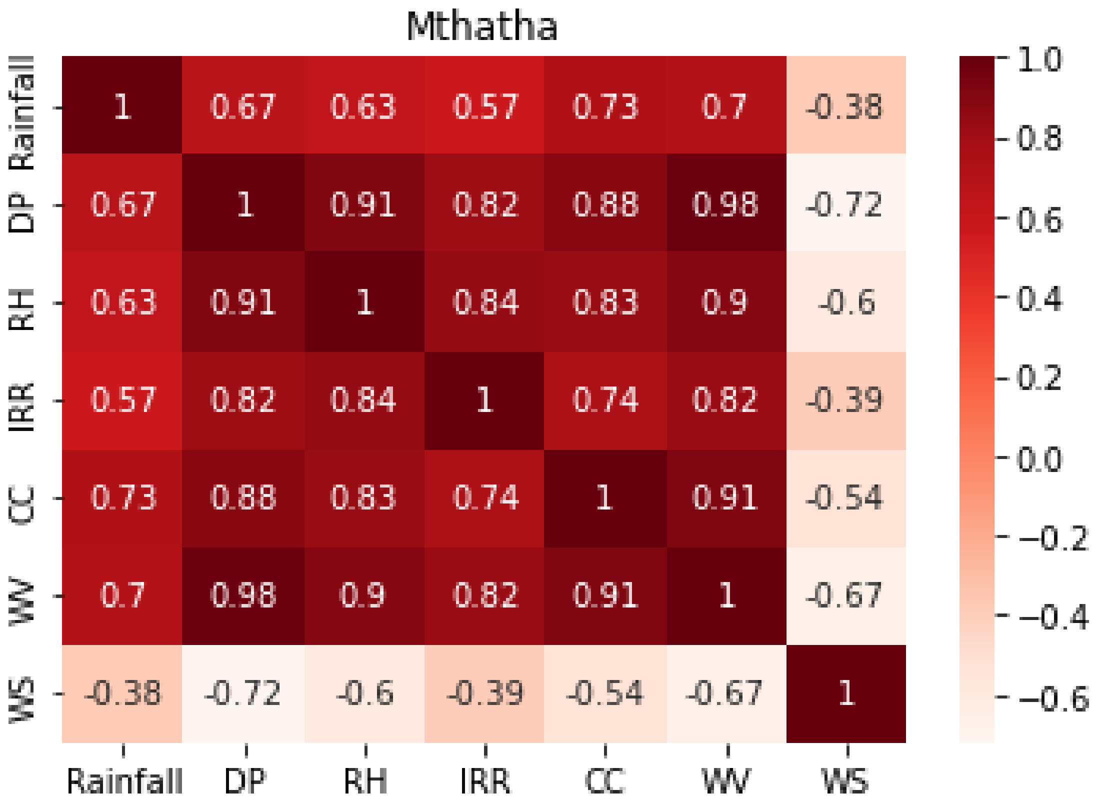
Richards Bay

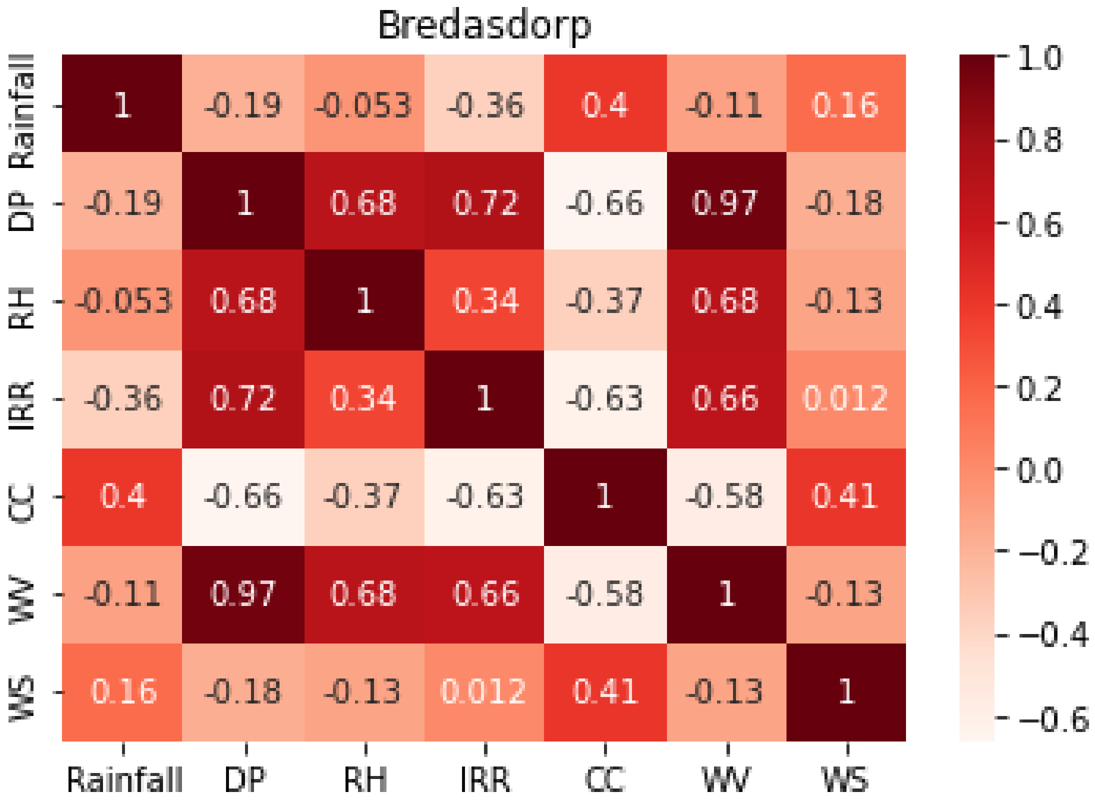


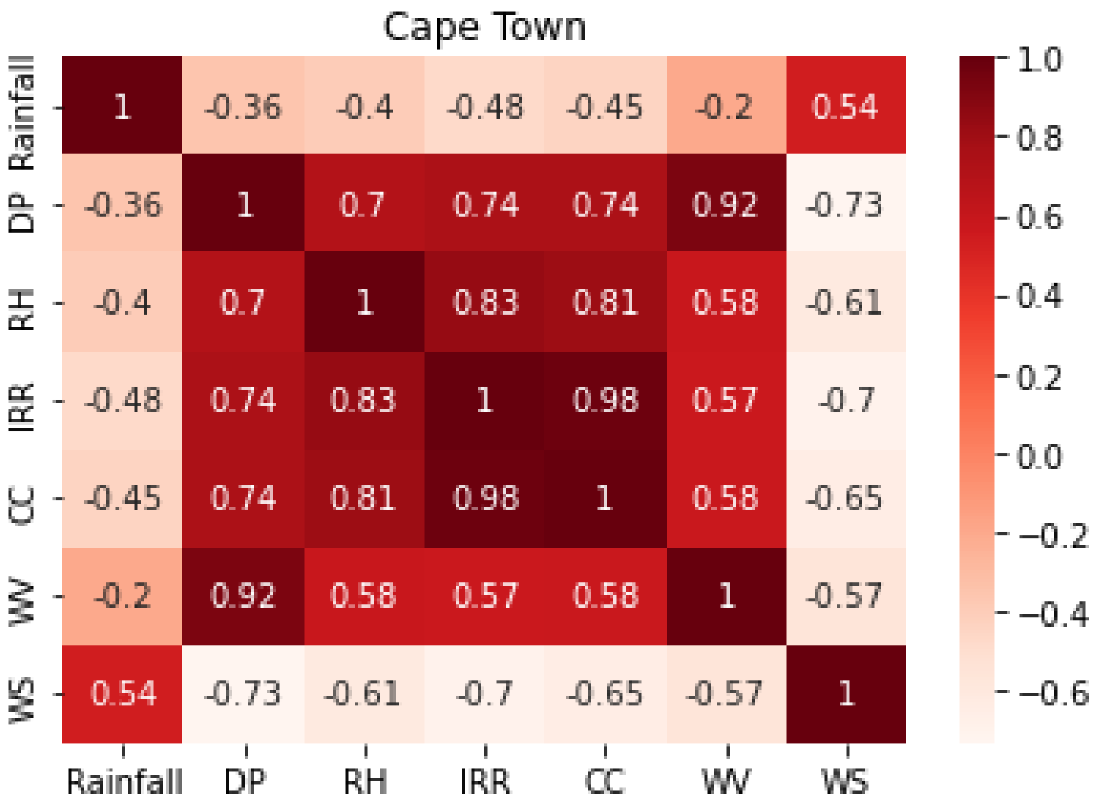
East London



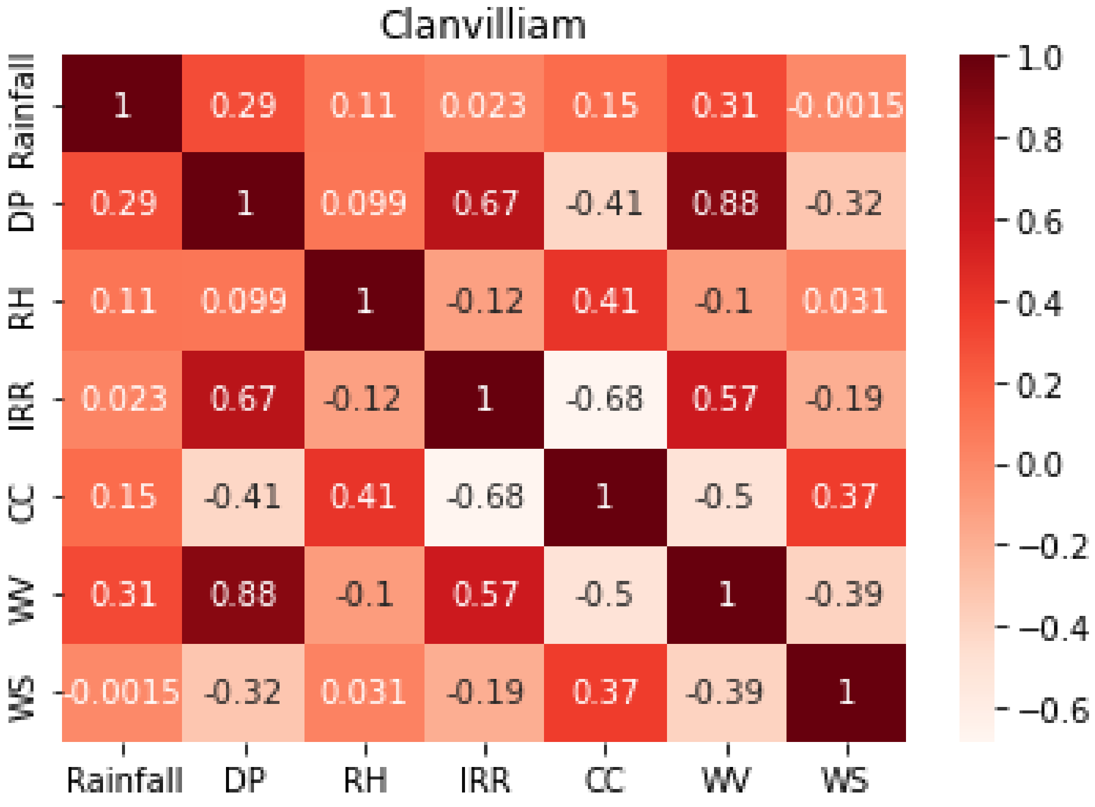
George


Mthatha


Bredasdorp



Cape Town



Clanvilliam