Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8, 148-178
2025
Publisher: Learning Gate
DOI: 10.55214/2576-8484.v9i8.9249

Machine learning models for rainfall prediction over arid climatic regions of South Africa

Jeremiah Ayodele Ogunniyi[1], Mohamed A. M. Abd Elbasit[2], Ibidun Christiana Obagbuwa[1*]
[1]Department of Computer Science and Information Technology, Faculty of Natural and Applied Sciences, Sol Plaatje University, South Africa; ibidun.obagbuwa@spu.ac.za (I.C.O.).
[2]Department of Physical and Earth Sciences, Faculty of Natural and Applied Sciences, Sol Plaatje University, South Africa.

Abstract: This study focused on predicting rainfall in four arid climatic zones in South Africa using four machine learning models. Prior to this, numerical models were used for weather forecasts in South Africa. Therefore, this study explores the use of machine learning models for rainfall prediction. The models used include linear regression, random forest, support vector machine, and ridge and lasso regression. The arid climatic zones were divided into four regions using the Koppen-Geiger climate classification system, with three locations selected for each zone. Atmospheric datasets from the South African Weather Service (1991–2023) and NASA (1983–2023) were utilized. These datasets were trained from 1983 to 2014 and tested from 2015 to 2023 using the four models. The monthly rainfall predictions obtained after training and testing were compared with actual data to validate the models' accuracy. Evaluation metrics such as mean absolute error, mean square error, root mean square error, correlation coefficient, and coefficient of determination were used to assess each model's performance. Support vector machine and random forest were the most accurate models across nearly all climatic zones. Linear regression, as well as ridge and lasso regression, also performed well in various regions. The study further indicated that atmospheric parameters such as dew point, cloud cover, and water vapor are essential at similar time lags for improved predictive performance.

Keywords: *Arid climate, Climate data, Climatic regions, Machine learning, Predictive modeling, Rainfall prediction, South Africa, Time series analysis, Weather forecasting.*

1. 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 made using hydrological numerical models [1]. These models are based on physical laws such as momentum, energy, and conservation of mass. Using numerical models, the important physical processes 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 complex models depend highly on correct information, equations, and many supercomputers, making the process challenging. With the increase in the daily data, numerical models will only get more complex in accurately predicting weather and climate parameters while also increasing the probability of errors [2]. Post-processing is done to eliminate these model errors [3]. However, since most quantities to be measured have limited time scales, it reduces the effectiveness of eliminating errors [4]. This leads to the need for machine learning (ML) models and algorithms for weather prediction.

The advent of the 21st century brought about an 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 forecasts have been limited

History: Received: 26 May 2025; Revised: 8 July 2025; Accepted: 11 July 2025; Published: 4 August 2025
* Correspondence: ibidun.obagbuwa@spu.ac.za

in their 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 include big data, machine learning, and artificial intelligence [5]. 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 to build models that can be tested and evaluated. If the result of the model is satisfactory, such models can be used for any 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 [6]. These applications have improved transport systems, healthcare, security, and defense networks 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 levels of uncertainty [5]. In recent times, researchers have suggested the use of machine learning algorithms for weather prediction. Bochenek and Ustrnul [6] 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. [7] also suggested the use of machine learning algorithms for weather prediction due to unsatisfactory performance while using numerical weather prediction. Hewage, et al. [8] 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-driven', they performed better than the numerical models.

In South Africa, numerical weather models are still being used for prediction [9, 10]. South African Weather Service uses the unified model of the United Kingdom Met Office, which is also used in other southern African countries [11]. The forecasts are made using the South African Weather Service Cray XC30, 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, the forecast can be made up to 10 days [11]. This shows the limitation of the numerical method for rainfall and weather prediction, as it cannot predict beyond ten days at most. With the success of machine learning models in rainfall prediction worldwide, this work seeks to explore different models for a long-term (1-year) forecast.

Despite South Africa being bounded by the Atlantic and Indian Oceans, it is susceptible to drought [12]. Extreme droughts that last for years are mostly caused by El Nino Southern Oscillation (ENSO), a quasi-periodic invasion of war sea surface waters into the Pacific Ocean. South Africa experienced drought in 2015 and 2016 due to an intense El Nino event [13]. This resulted in reduced agricultural productivity, leading to importing grains instead of exports, water shortages, and a significant negative economic impact. This drought was the worst in 23 years, from 1992 to 1995. Drought leads to reduced 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, affecting livelihood [14]. 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 affected about 3.2 million people. In South Africa, the damage was about US $250 million, with 2.7 million people affected, resulting in an 8.4% reduction in agricultural production and a 15% reduction in livestock. Accurate rainfall prediction might have reduced the advance effect caused by the drought as farmers and the government would have been better prepared instead of spending almost a billion rand on relief and support for farmers. However,

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

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.

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 systems, 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 challenging to maximise water usage. Therefore, this work is expected to 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.

2. Methodology

The locations chosen for this study were selected based on two criteria. They were chosen using the different climatic zones under the arid climate classification. Cities in different regions of South Africa (four cities in Free State Province, four cities in Northern Cape Province, and one city each in Western Cape Province, Limpopo Province, Eastern Cape Province, and Northwest Province) were also chosen. The purpose of these criteria is to ensure that there is no model bias regarding climate or region. Therefore, some cities were 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 if some regions are more accessible to predict. If the models perform well across all climatic regions, the models can be generalized.

40-year continuous datasets obtained from The National Aeronautics and Space Administration (NASA) website from 1983 were used for this study, obtainable in the Giovanni interactive visualization page. This was combined with South African Weather Service (SAWS) datasets. 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 corresponding to a major city will be picked in each climatic zones for analysis.

Once the data sets were obtained, pre-processing was done to identify the missing values, eliminate them, and duplicate them. Outliers were then identified and removed from the datasets. The daily datasets were then converted to monthly averages. The models were trained using the abovementioned four models on 80% of the datasets. The Mean Square Error (MSE), the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE), the 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.

2.1. Study Location

The study location is based on the arid climatic zone of the Koppen-Gieger climate classification of South Africa. The arid climate classification was divided into the cold and semi-arid steppe, cold arid desert, hot and semi-arid steppe, and hot arid desert. For each division, three locations are randomly selected for rainfall prediction. This climate classification is mostly found in the Free State and Northern Cape Provinces of South Africa. Table 1 shows the subdivisions and locations selected for the study and their average annual temperature and rainfall.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

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 (°C)				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

3 Results and Discussion

3.1. Correlation Between Atmospheric Variables

From the heatmap in Figure 1, dew point, cloud cover, and water vapour are essential for rainfall prediction in Bloemfontein as their correlation with rainfall is 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, corresponding to 0.68, except for temperature, with a correlation coefficient of 0.56. Relative humidity and wind speed had the lowest coefficients of 0.11 and -0.13 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 correlated at 0.29, while rains and windspeed correlated at -0.026. This result reveals that relative humidity and wind speed are not essential for rainfall prediction for cold and semi-arid steppes. Atmospheric parameters such as dewpoint, water vapour, cloud cover, and temperature should be used for accurate predictions.

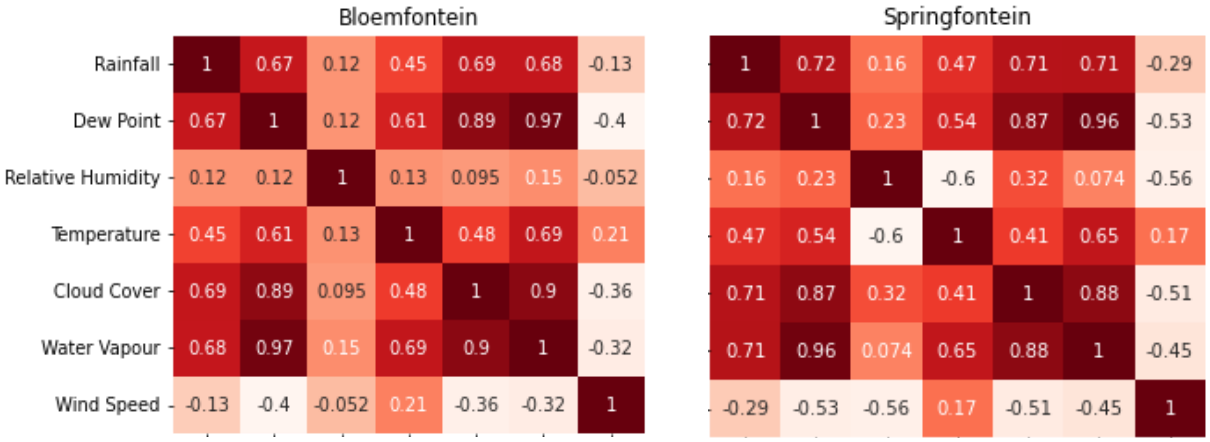
The heatmap in Figure 1 also showed that no atmospheric parameter correlated with rainfall in Alexander Bay. This may be expected as this city has little or no rainfall. The atmospheric variables used for this study had correlation coefficients of 0.084, 0.018, -0.078, 0.20, 0.11, and -0.096 for dewpoint, 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 correlated 0.91, 0.92, and 0.76 with relative humidity, water vapour, and temperature. For Beaufort West, only dew point and water vapour correlated 0.50 with rainfall. The 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, and 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 for rainfall prediction are dewpoint, water vapour, and cloud cover in the cold arid desert.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

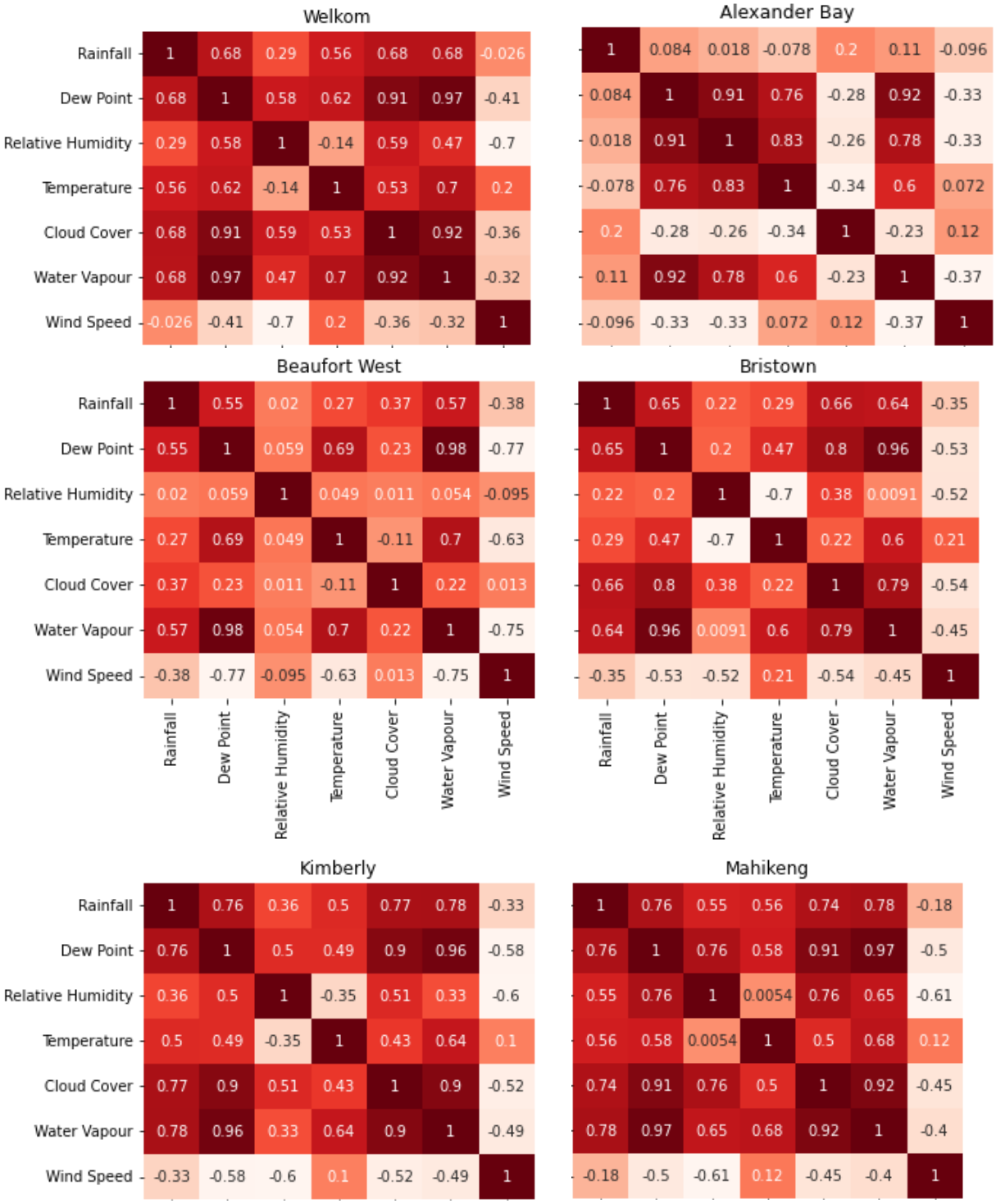
The heatmap for the correlation between rainfall and other atmospheric variables under the hot and semi-arid steppe is also shown in Figure 1. The result reveals that all atmospheric variables in Mahikeng had a correlation coefficient with rainfall exceeding 0.50 except wind speed, which correlated with -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.

In contrast, temperature had a coefficient of 0.50. Their correlation coefficients with rainfall were 0.36 and -0.33 for relative humidity and wind speed, respectively. Port Elizabeth showed a similar pattern to what was obtained in Alexander Bay, with all atmospheric variables having a low correlation with rainfall. The result indicates 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, and 0.054 coefficients, respectively. Also, as in Alexander Bay, relative humidity, temperature, and water vapour had a high correlation with dewpoint corresponding to 0.91, 0.78, and 0.98, respectively.

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 reaching 0.64, followed by cloud cover and dew point corresponding to 0.62 and 0.60, respectively. Other atmospheric variables also had a 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 a correlation coefficient greater than 0.50 with rains, the correlation coefficient is higher than 0.60 in Lauville only for cloud cover. This is similar to Upington, as only cloud cover correlated with rainfall with a 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 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.



Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249



Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

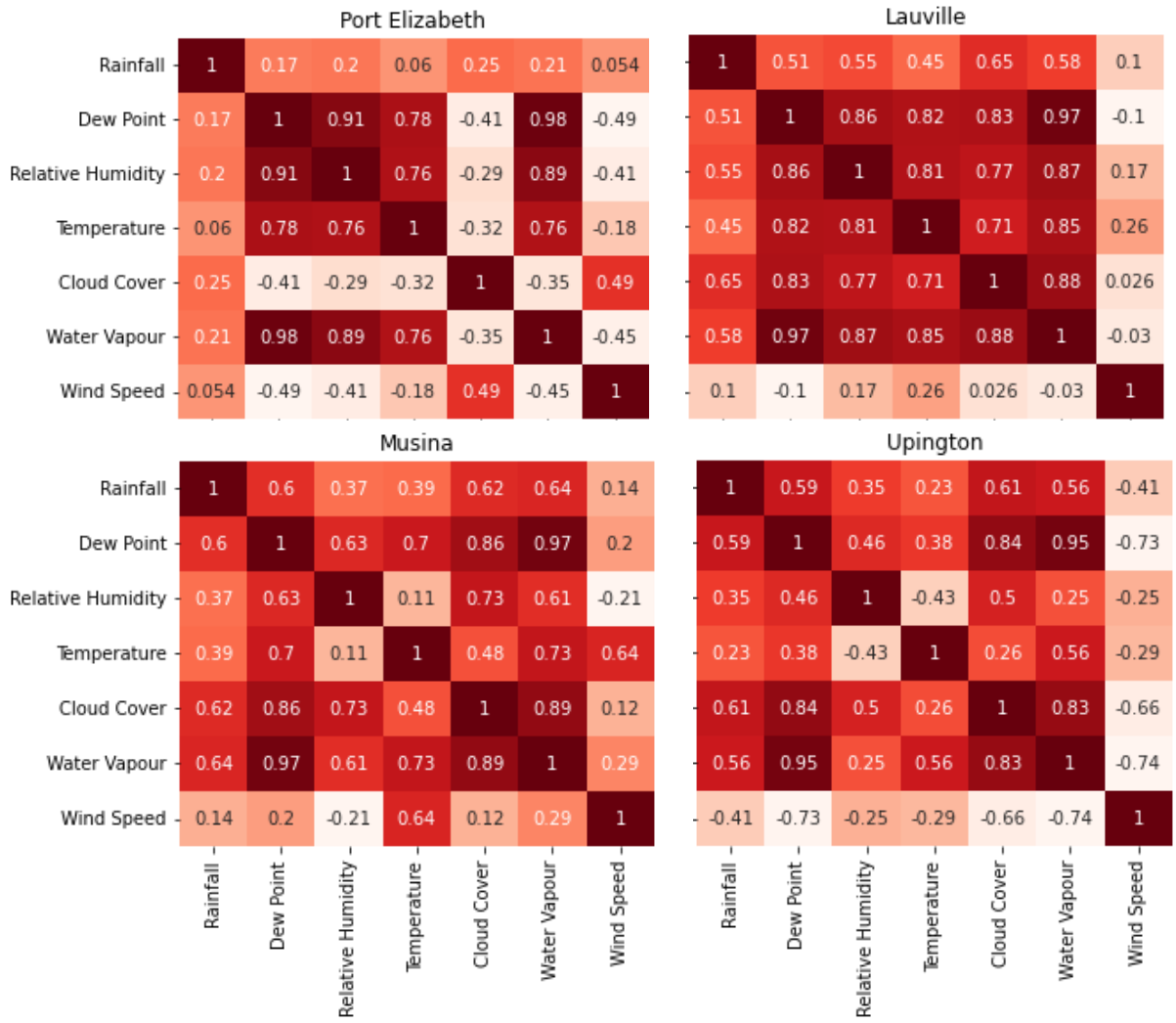


Figure 1.
Correlation coefficient of atmospheric variables used for the study.

3.2. Rainfall Prediction over Different Climatic Zones

3.2.1. Cold and Semi-Arid Steppe

Table 2 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. [15] evaluated an inverse weighting method (IDW) for patching daily and 10-day rainfall for six weather stations in Free State South Africa using 58-year datasets. They showed that their IDW

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

method was highly effective for daily and 10-day predictions. Bloemfontein and Welkom obtained an R2 value of 0.90 and 0.78 and a MAE of 3.17 and 5.60 for both locations, respectively.

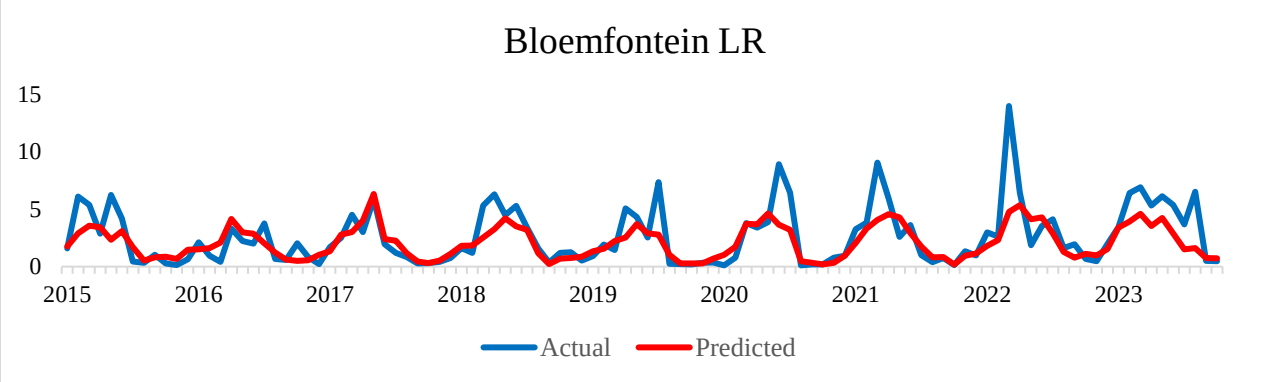
Figure 2 shows temporal variation in rainfall for the three locations under the cold and semi-arid steppe for Linear Regression, Random Forest, Support Vector Machine, and the Ridge and Lasso. The results showed that for all locations and algorithms, the model accurately predicted the seasonal variability 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 2015 and 2019 but underestimated it, especially in 2021 and 2022. Linear regression, random forest, and the support vector machine modelled the spike in 2017 rainfall estimates but overestimated the increase from 2019. 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 a year, an average annual temperature of 22°C, and an average humidity of 48%. Unfortunately, none of the models could accurately predict the unexpected spike in the yearly rainfall. This increase in 2021 was also recorded in Bloemfontein by all models. Bochenek and Ustrnul [6] reported that rainfall in Bloemfontein in 2021 increased by about 150mm compared to the decadal average. Although a similar increase was experienced in Welkom, all the models could predict the spike. However, none of the models could accurately estimate the rainfall received.

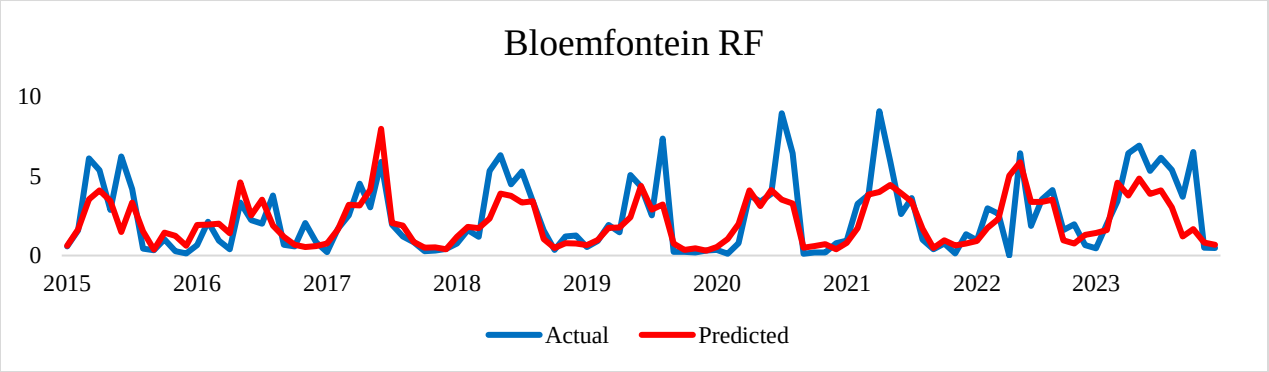
This pattern of high rainfall in 2021/2022 and underestimation of the model was expected in almost all locations as South Africa received above-normal annual rainfall. This was attributed to the El Nino-Southern Oscillation (ENSO) being in a La Nina phase. Sivakumar and Fazel-Rastgar [16] revealed the presence of an active frontal system with continuous rainfall in 2022. They also attributed the increase to the injection of high humidity from extended warmer isotherms.

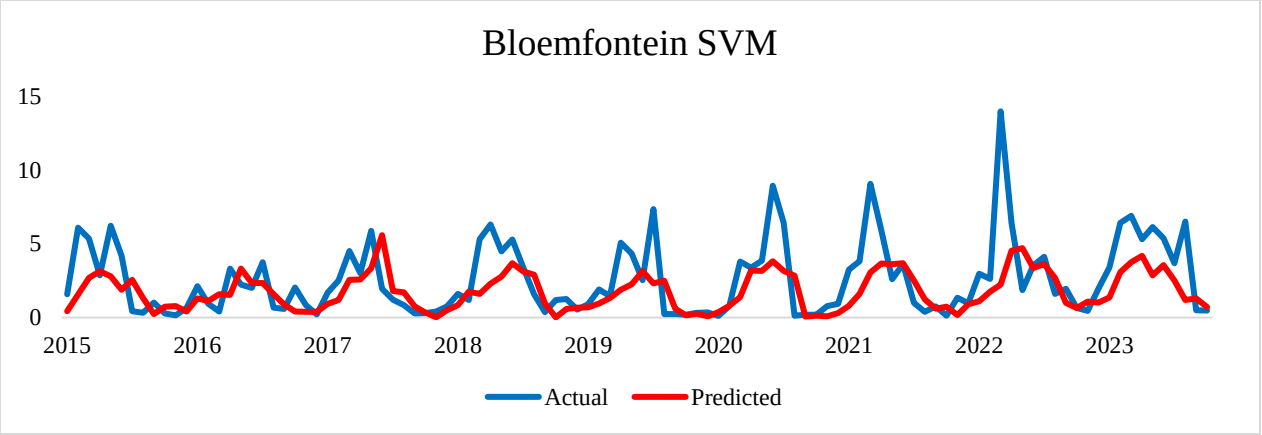
Table 2.
Table showing model evaluation metrics for cold and semi-arid steppe climate classification (Bloemfontein, Springfontein, and Welkom).

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

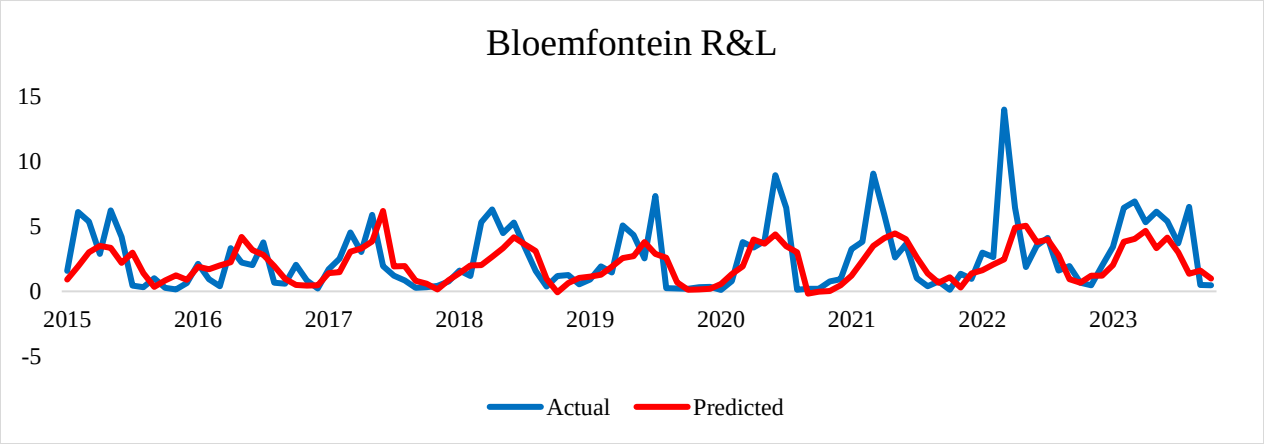
Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

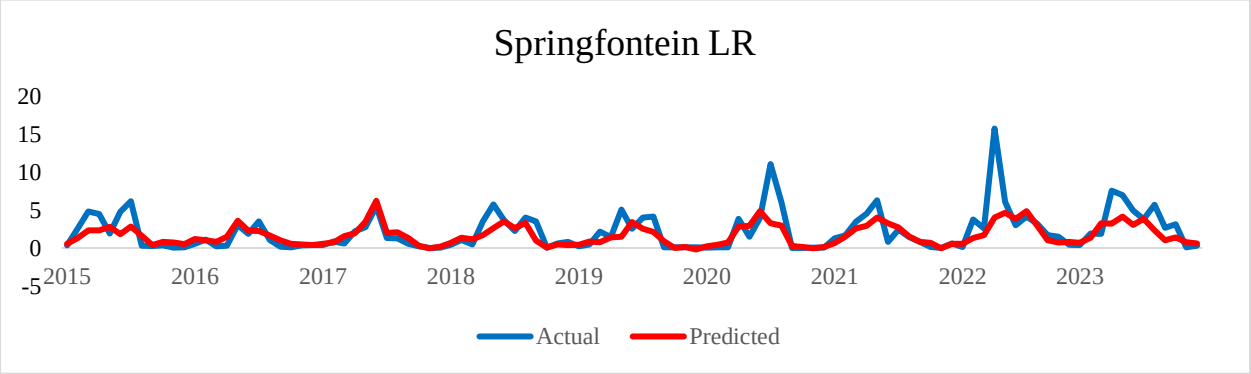


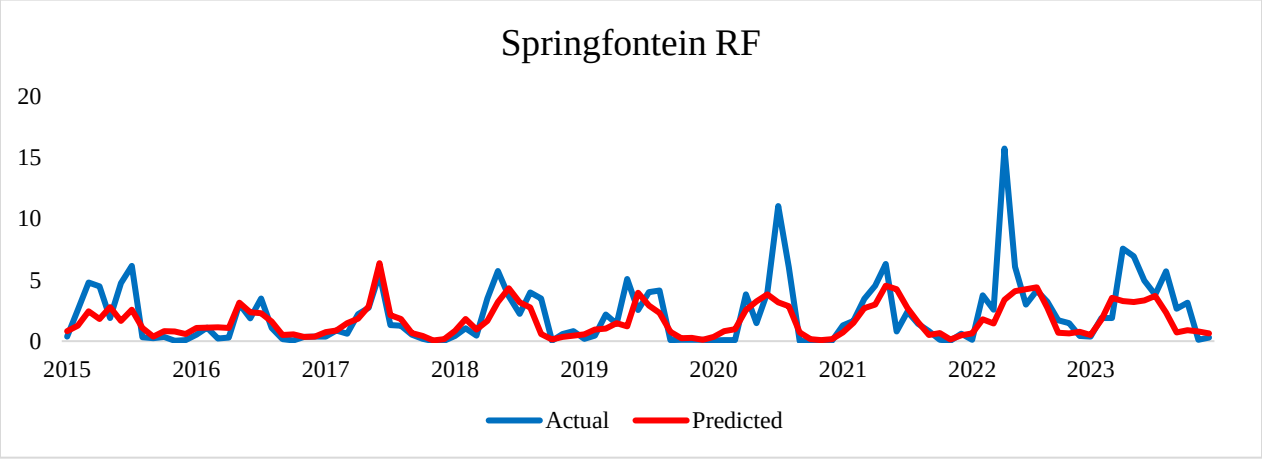




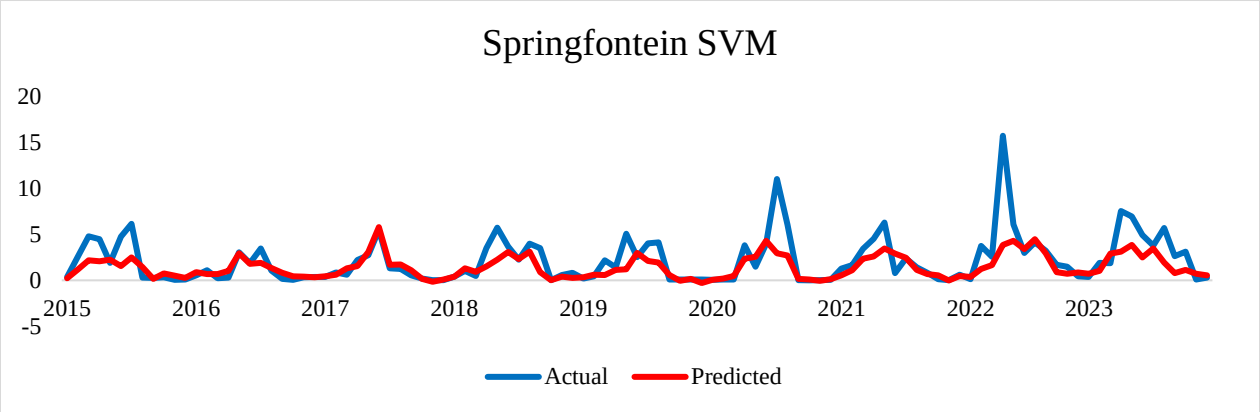
Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

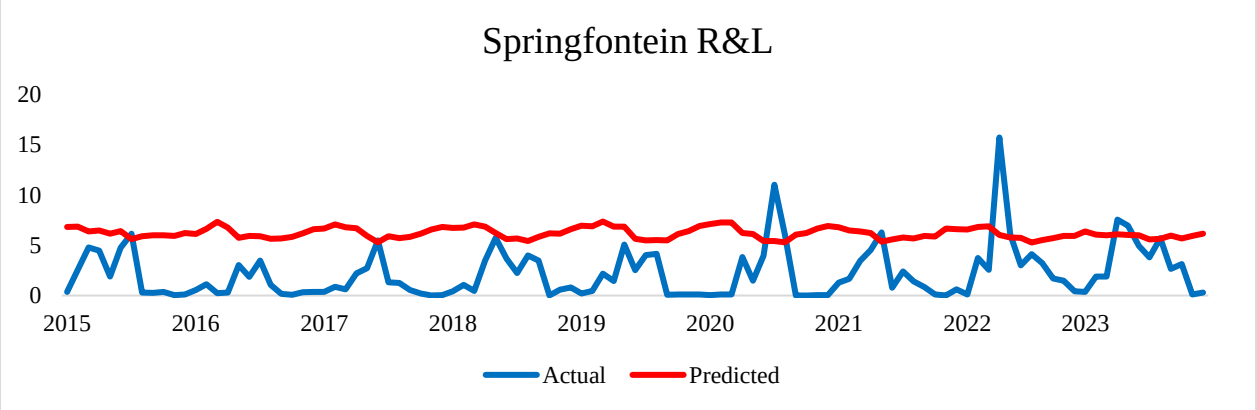




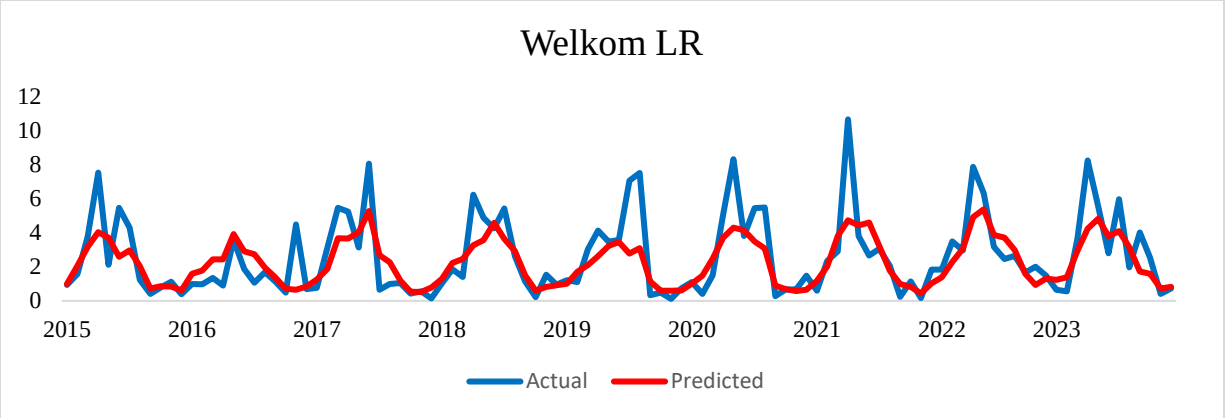


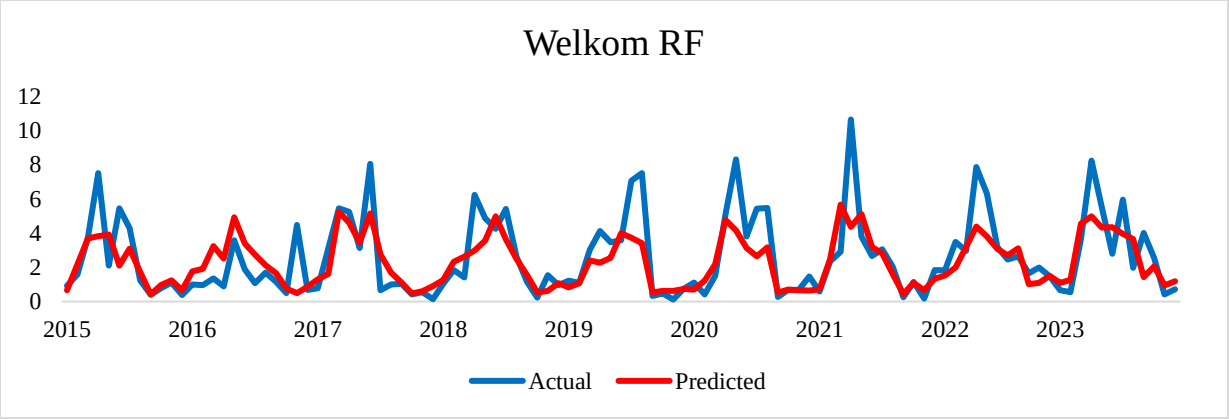
Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

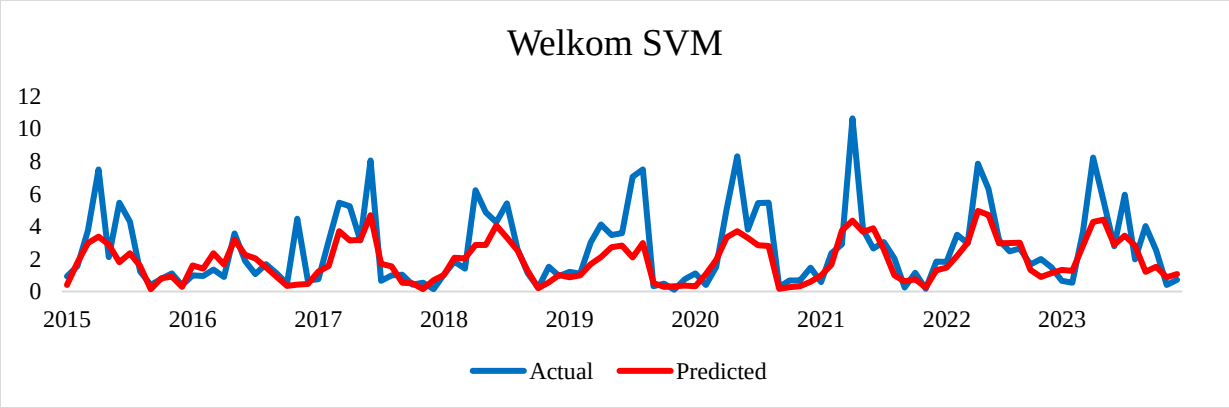




Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249







Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

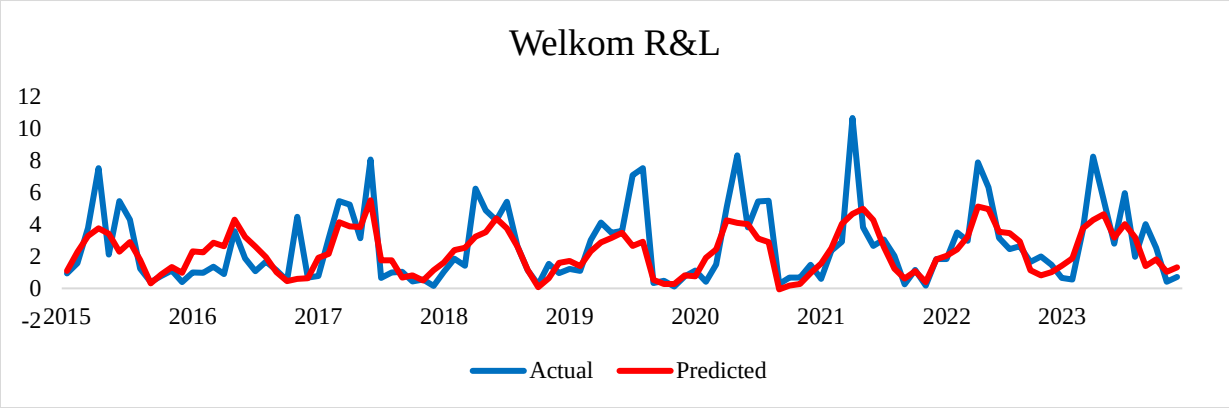


Figure 2.
Shows the predicted and actual rainfall for models used and locations under the cold and semi-arid steppe climate classification.

3.2.2. Cold Arid Desert

For the cold, arid desert, the support vector machine had the highest correlation coefficient of 0.72 in Bristown, 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 3. Similarly, the support vector machine model had a higher correlation coefficient than other models for Beaufort West, which corresponded 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 will be further examined. Perhaps it being officially the driest town in South Africa may contribute partly to this [17]. The models showed a pattern of poor prediction in regions with low rainfall [18]. 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 the 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 a 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 [19]. Notable parts of South Africa have been declared disaster-drought areas [20]. The resultant effect is the increased mortality rate in livestock due to unavailability of water, loss of dairy and livestock production, and disruption in the animal reproduction cycle. It also increases unemployment as industries relying on agricultural produce, such as fertilizer manufacturers, are lost, and food prices increase due to damage to crop quality and reduced food production. The impact of drought is not limited to its economic impacts. Still, it also has environmental impacts as it leads to the degradation of animal habitats, inferior crops, decreased water quality and insufficient drinking water, soil erosion, and fire outbreaks [19].

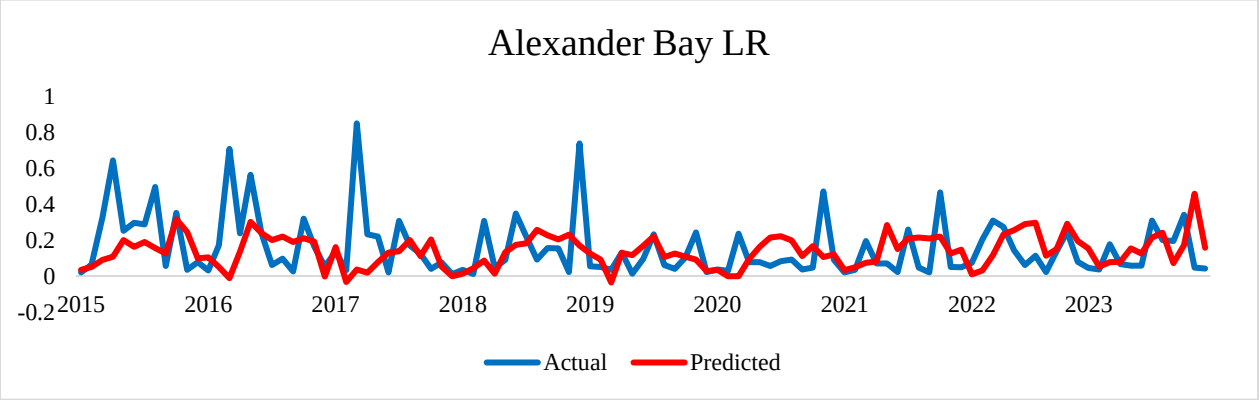
Figure 3 shows the monthly variation in rainfall for Alexander Bay, Beaufort West, and Bristown using different models. As expected, Alexander Bay received the least rainfall, with a monthly average below 1mm. The models could accurately predict the seasonal variation in rainfall for all locations and all models. Random forest better estimated the rainfall received for both Beaufort West and Bristown. While evaluating these regions, Linear regression's estimate for rainfall in Bristown was better. Descamps, et al. [21] revealed the challenge in reliable long-term rainfall prediction due to the area's susceptibility to droughts and floods. They stated the need to use ridge and lasso regression models as they outperformed dynamical climate prediction models. However, they cautioned on independently

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

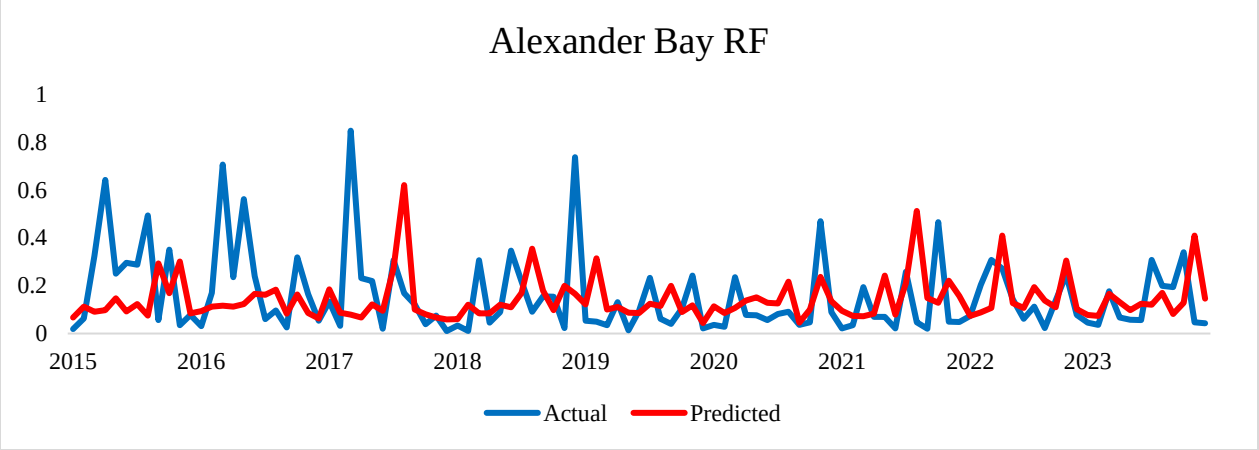
selecting the predictors for both train and test data as they may result in biased results that do not correctly reflect the accuracy of the models.

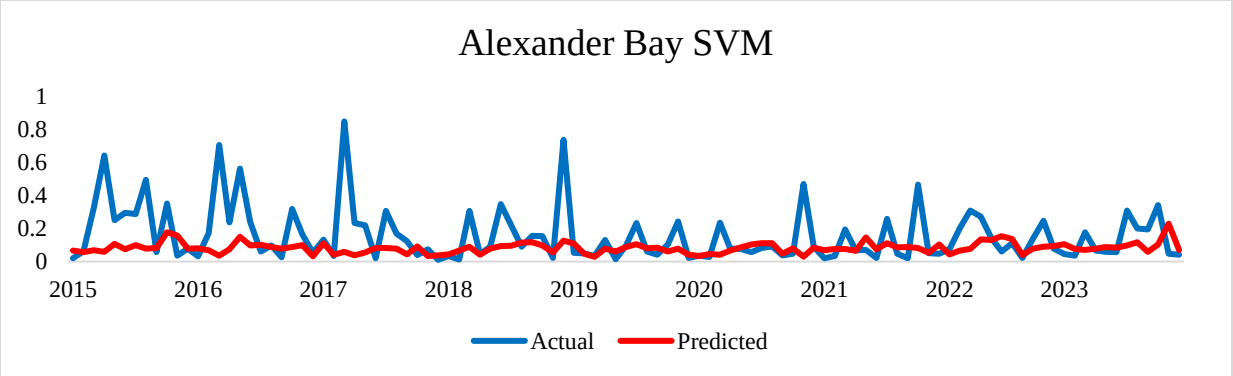
Table 3.
Table showing model evaluation metrics for cold, arid desert climate classification (Alexander Bay, Beaufort West, and Bristown).

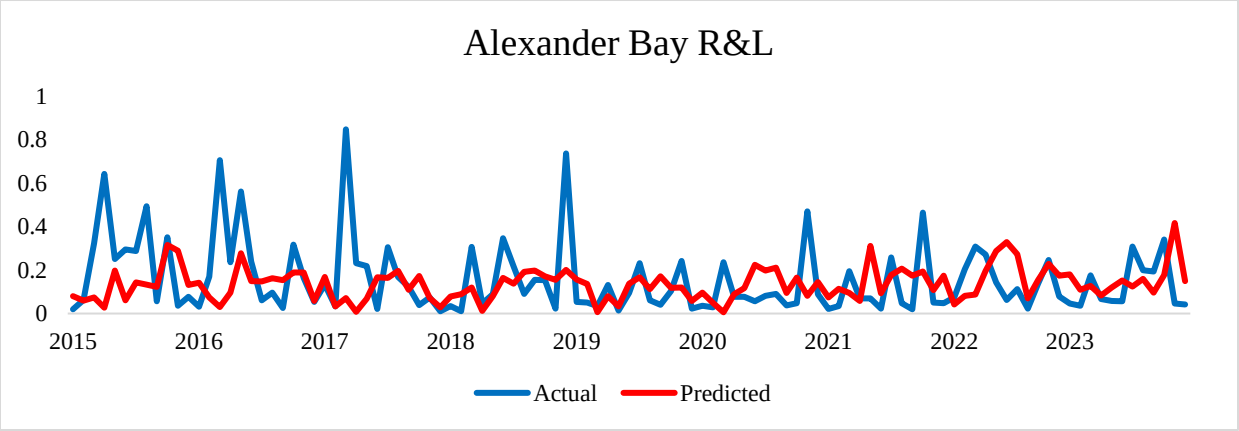
Linear Regression					
Locations	**MAE**	**MSE**	**RMSE**	**r**	**R-Square**
Alexander Bay	1.50	3.75	1.94	0.14	0.12
Beaufort West	0.69	0.78	0.88	0.60	0.14
Bristown	0.77	0.94	0.97	0.66	0.65
Random Forest					
Alexander Bay	0.13	0.04	0.19	0.01	-0.31
Beautfort West	0.62	0.97	0.99	0.55	0.30
Bristown	0.67	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



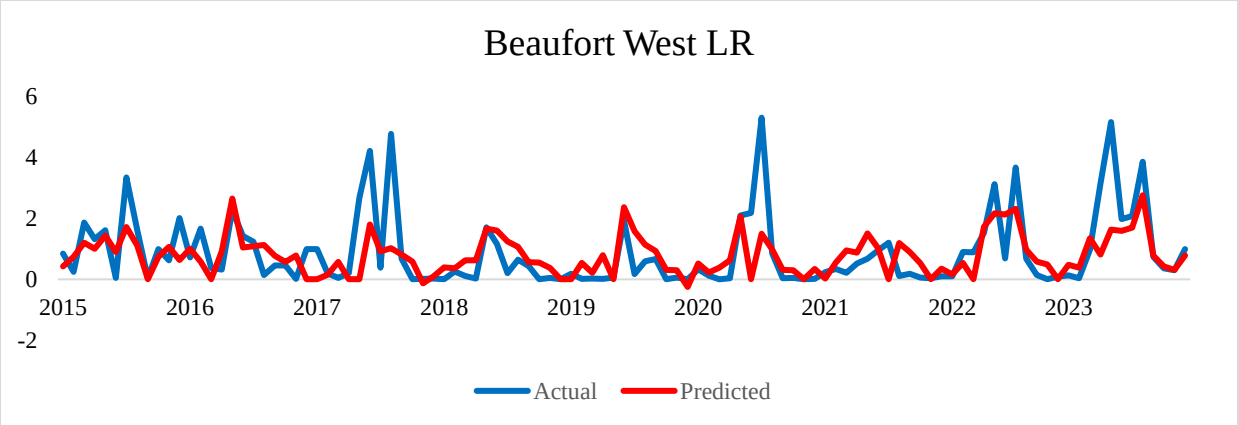
Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

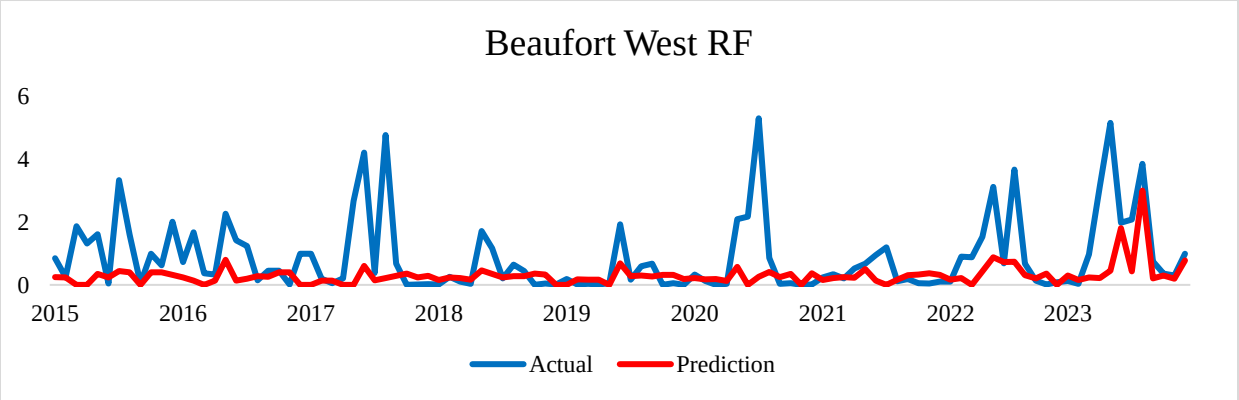


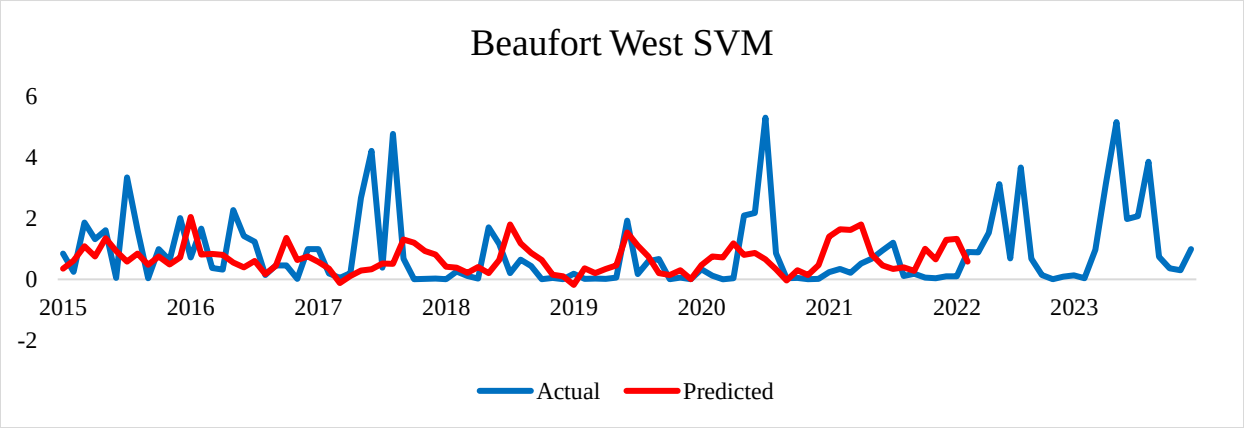




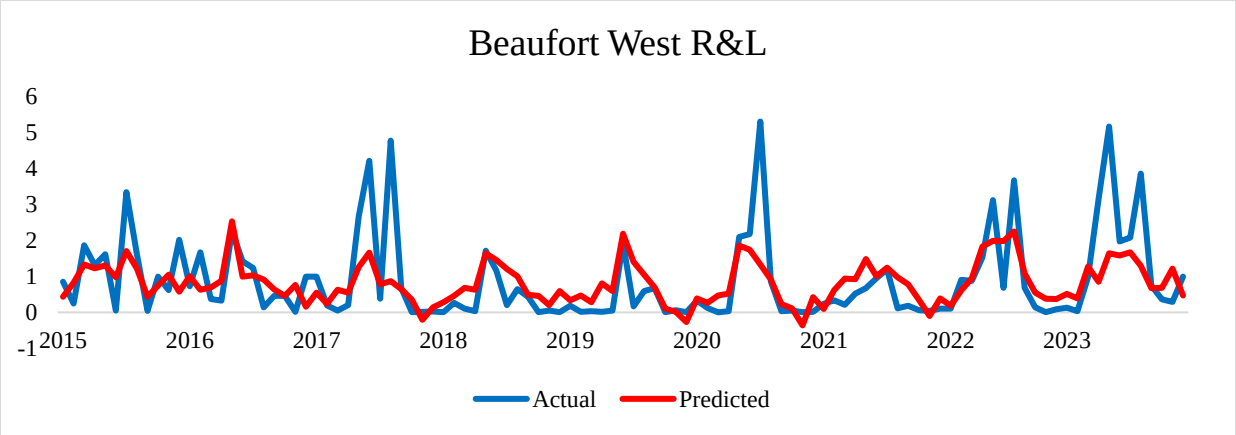
Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

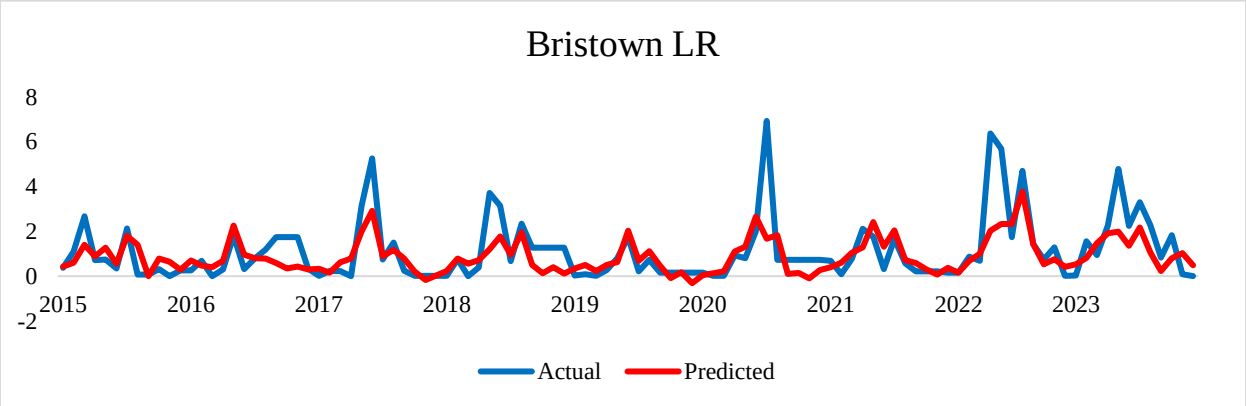


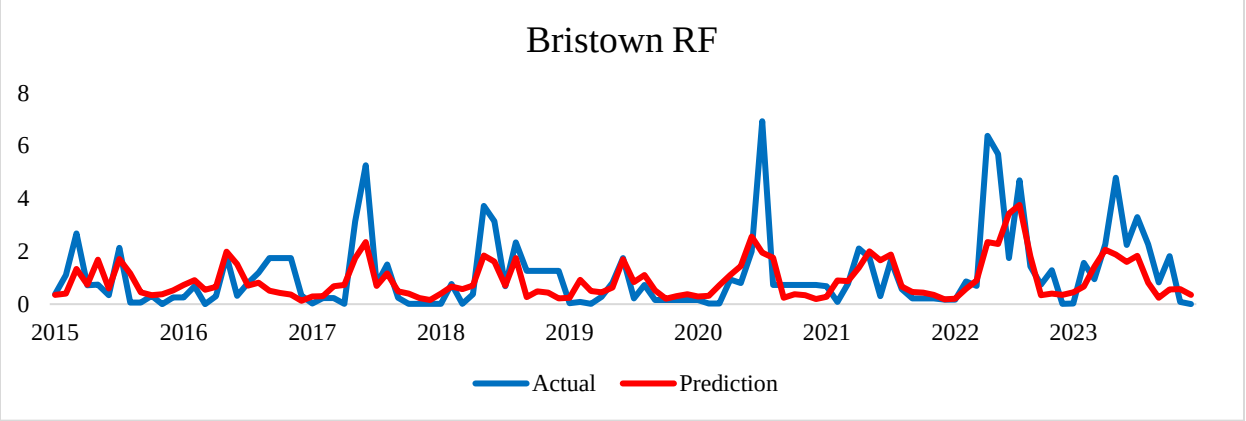




Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249







Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

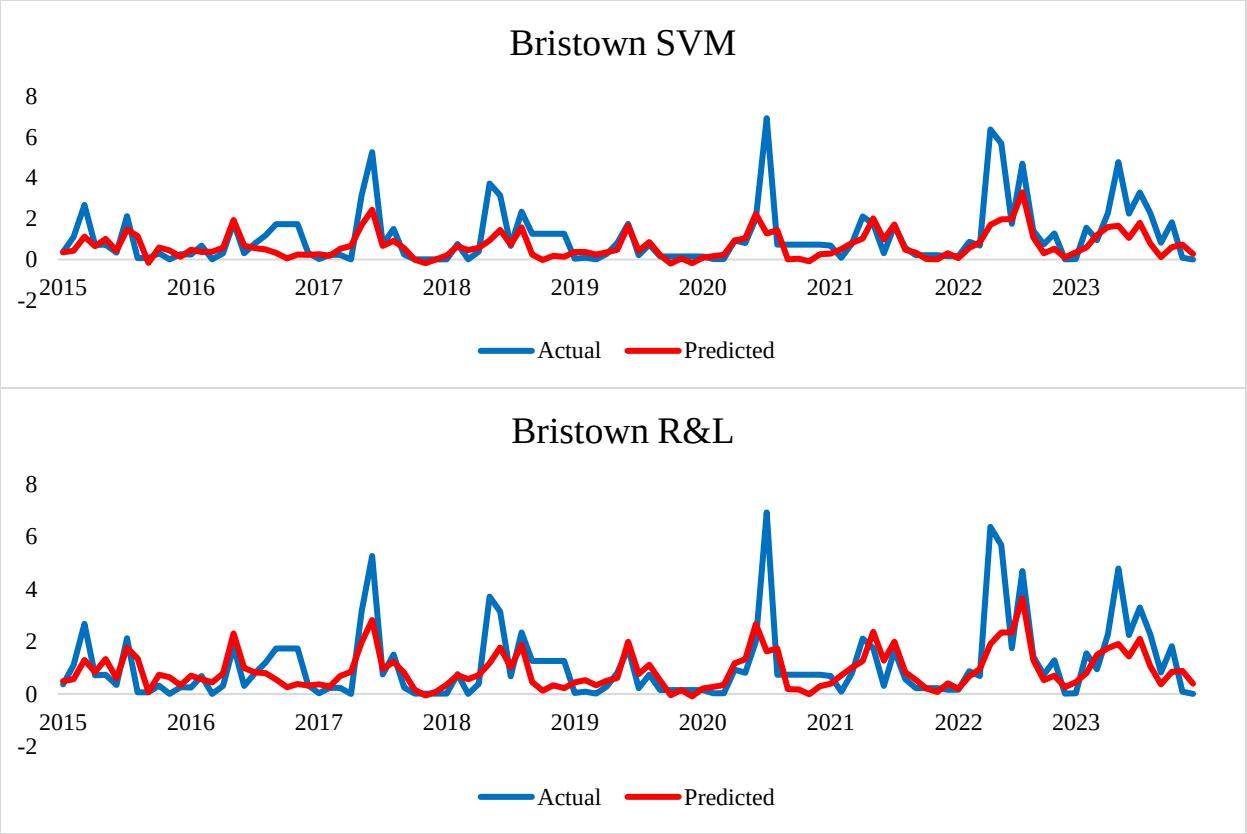


Figure 3.
Figures showing the predicted and actual rainfall for models used and locations under the cold arid desert climate classification.

3.2.3. Hot and Semi-Arid Steppe

Table 4 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 is in the Northern Cape Province, Mahikeng is in the North-West Province, and Port Elizabeth is 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 5. 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 suit machine learning applications in atmospheric sciences. Tladi, et al. [22] conducted 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 a coefficient of determination of about 0.80, mean square values ranging from 0.03 to 0.30, and an 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, which only focuses on groundwater levels. However, the correlation coefficients in Port Elizabeth were 0.13, 0.29, 0.43, and 0.30 for linear regression, random forest, support vector machine, and ridge and lasso, respectively. The coefficient of determination for these models was also significantly low. Yakubu, et al. [23] predicted

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

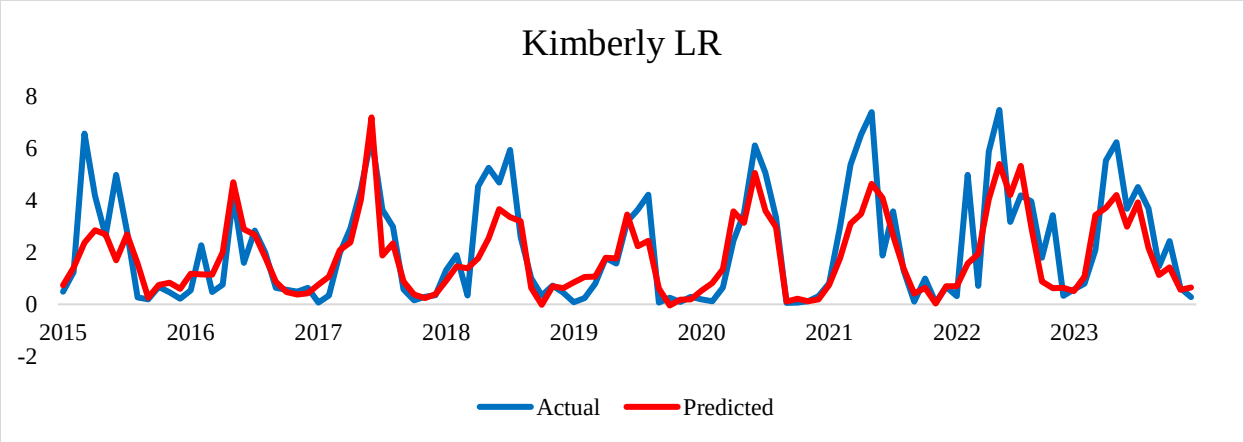
precipitation in Port Elizabeth using two machine learning models – multiple linear regression and multilayered perceptron. Their model used five cloud properties: cloud optical thickness, cloud effective radius, cloud top temperature, cloud top pressure, and liquid water path. Their result showed a 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.

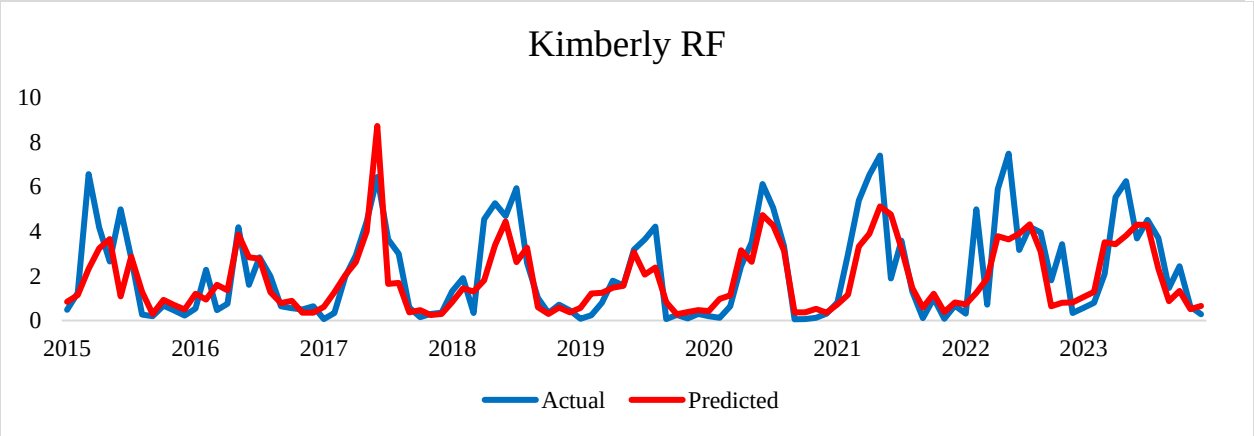
Figure 4 shows the temporal variation in rainfall for various locations using different models. The result also revealed a consistent pattern in modeling all models' interannual variation in rainfall. For Kimberly, random forest, ridge, and lasso almost accurately estimated the amount of rainfall received. However, there was an overestimation of rainfall by random forest in 2015; perhaps the model still understands the datasets. Support vector machine had a better predictive performance in 2023 for Kimberly. A 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. However, it underestimated rainfall in 2019 and 2023.

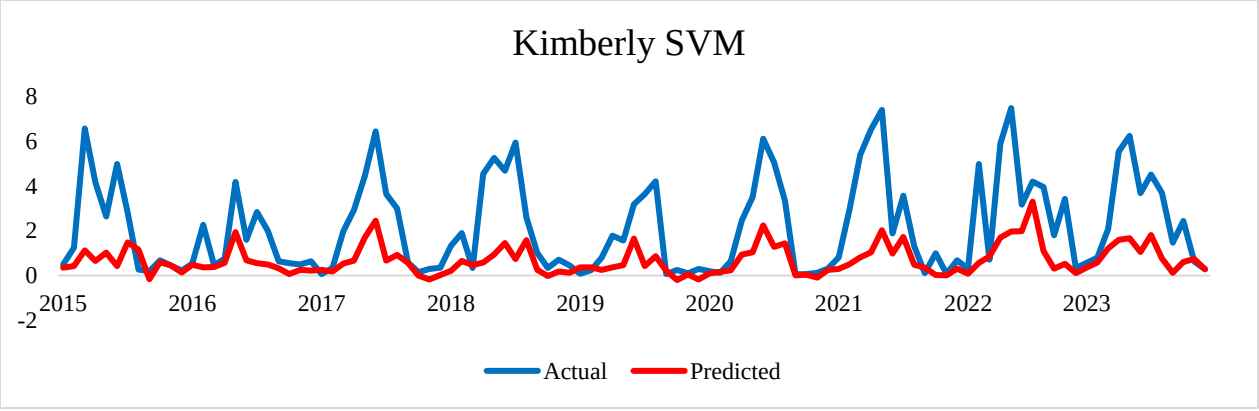
Table 4.
Shows model evaluation metrics for hot and semi-arid steppe climate classification.

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

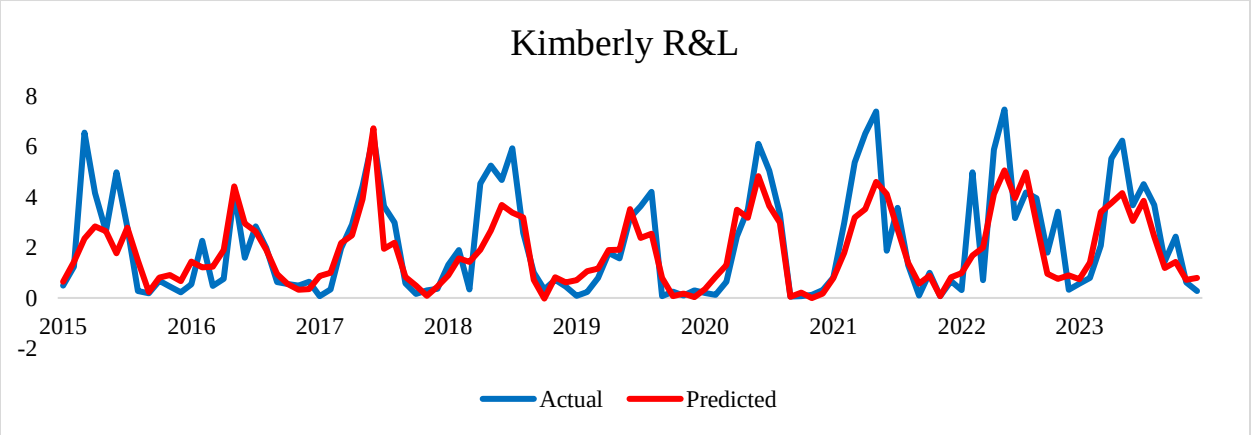
Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

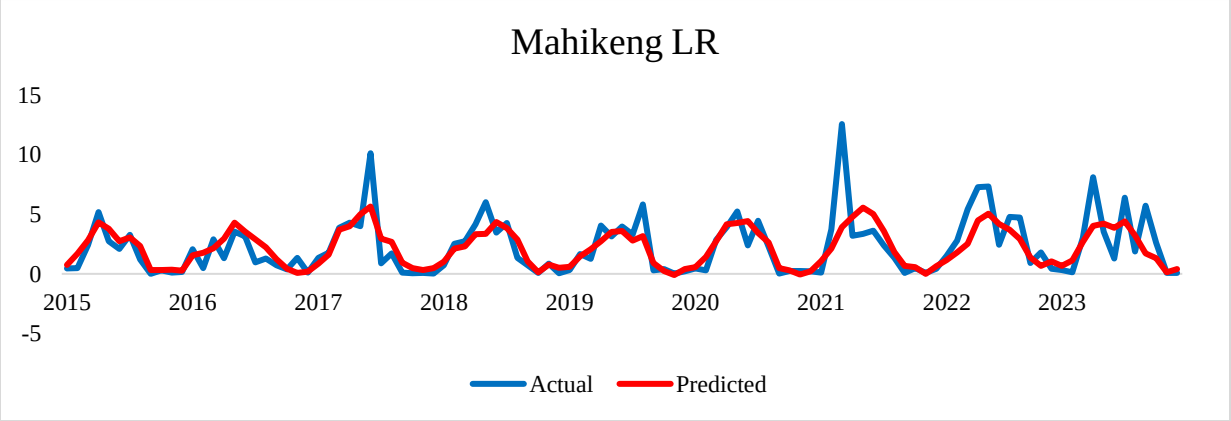


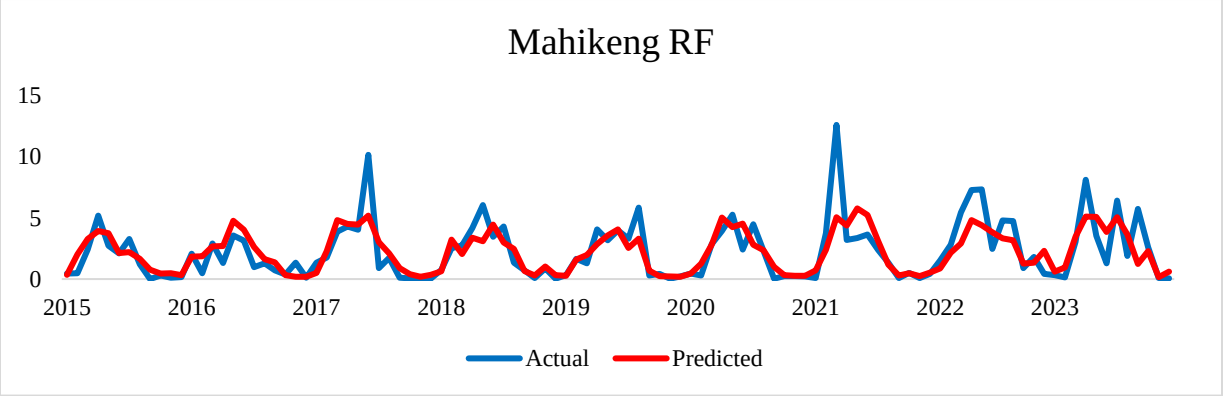




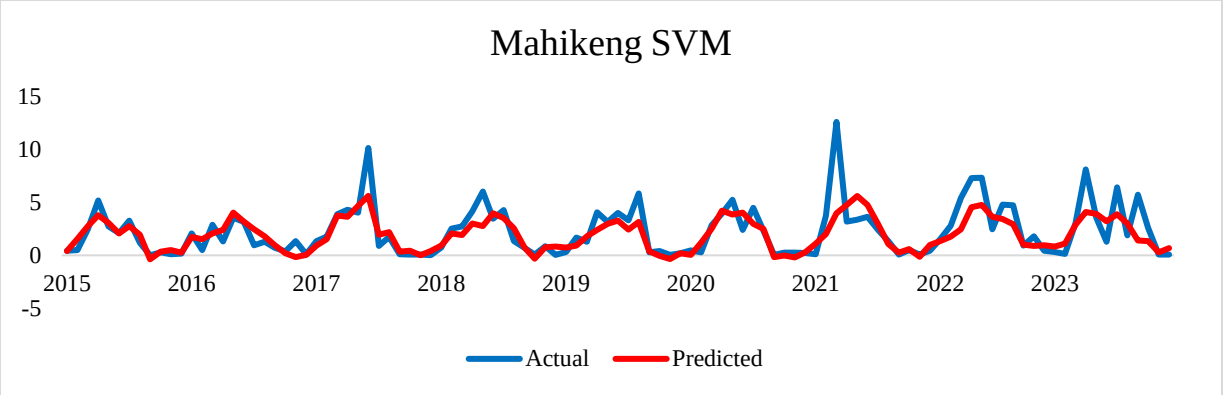
Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

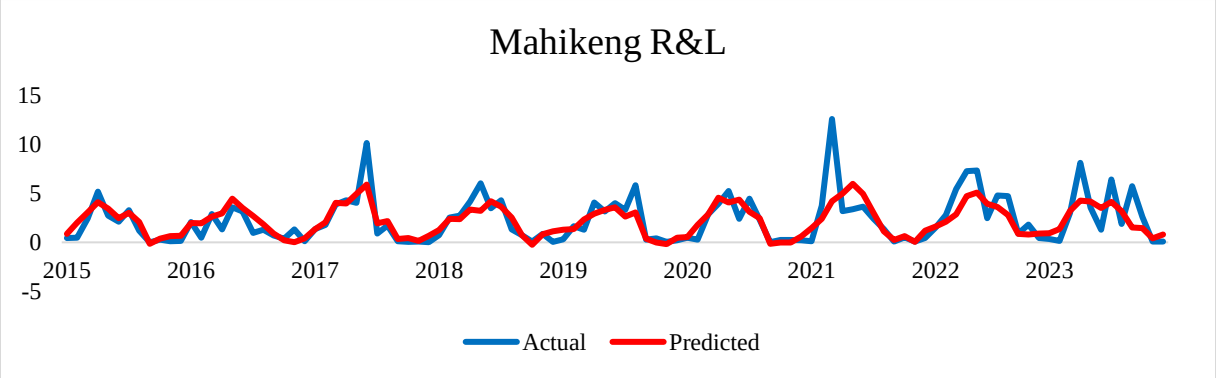


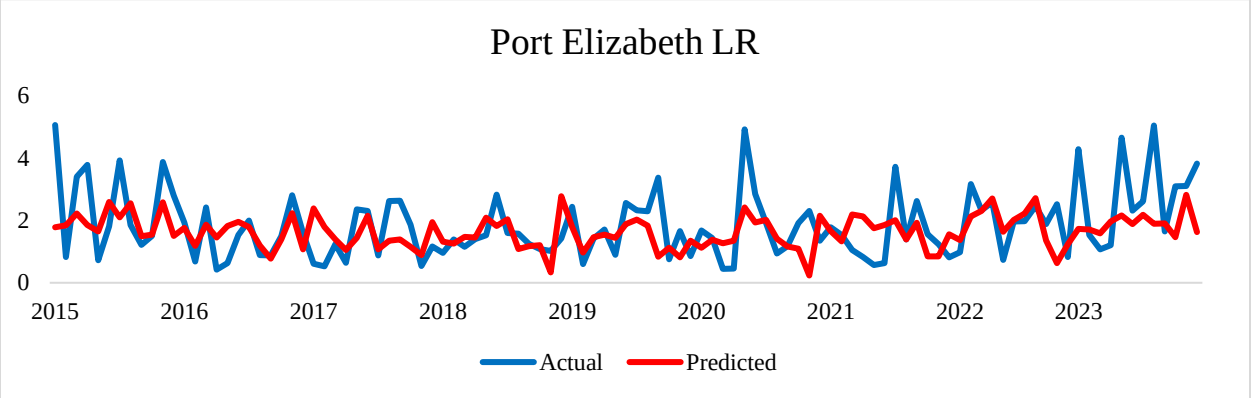




Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249







Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

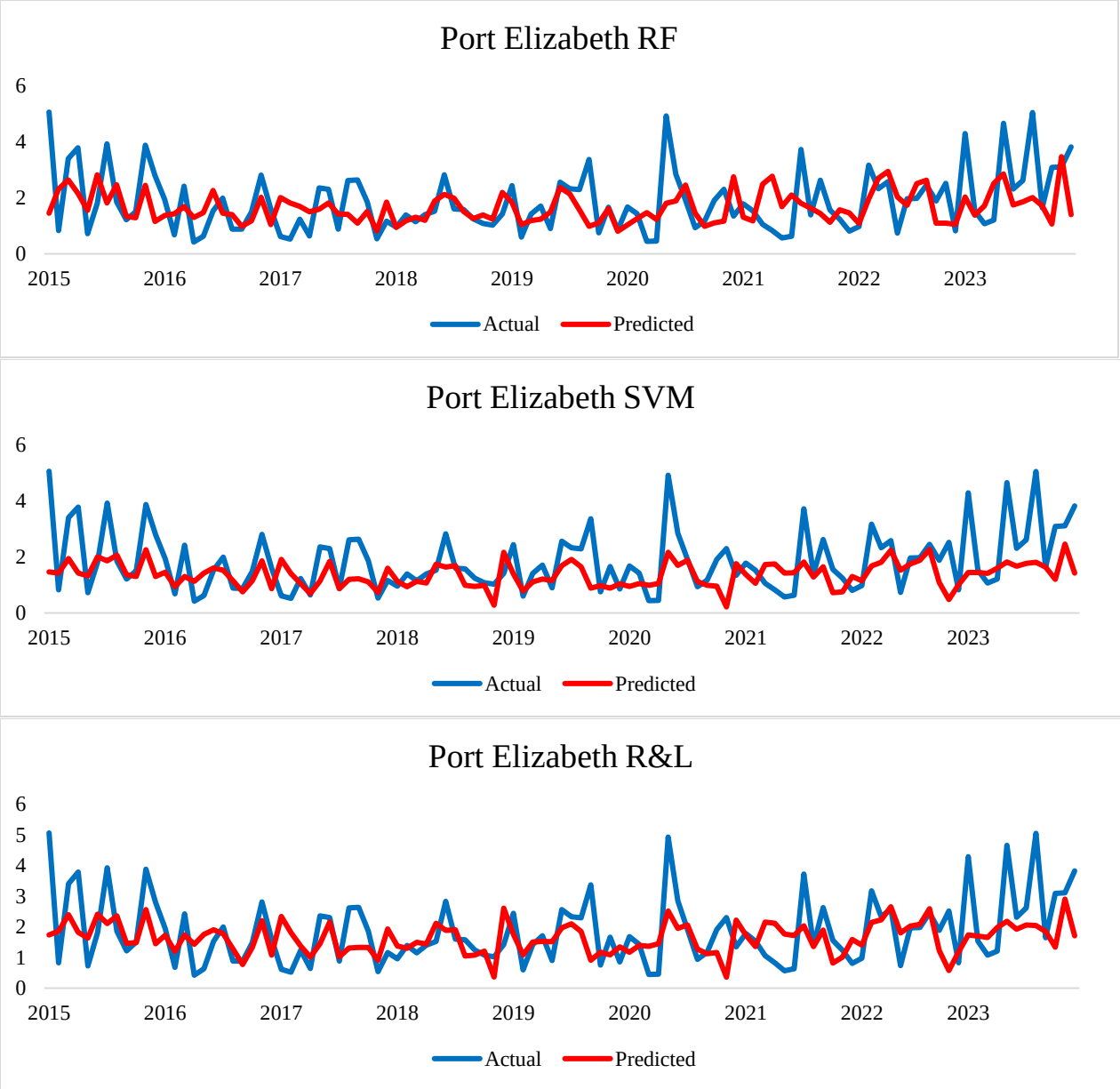


Figure 4.
Shows the predicted and actual rainfall for models used and locations under the hot and semi-arid steppe climate classification.

3.2.4. 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 parts of South Africa. Table 5 shows the evaluation metrics for the models used for the study for the three locations. In Lauville, the 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,

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

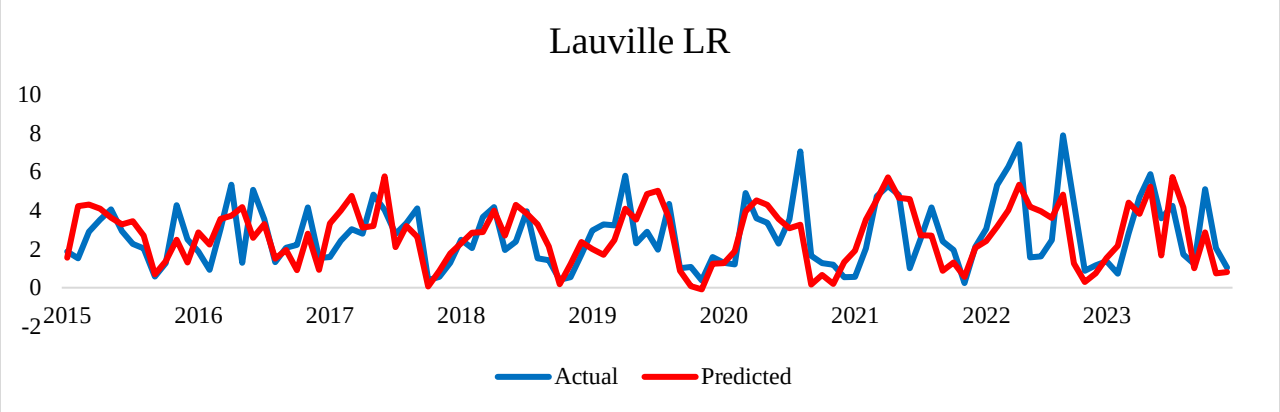
ranging from 0.68 for ridge and lasso to 0.79 for support vector machine. Values for linear regression and random forest are 0.69 and 0.74, respectively. Linear regression had the highest coefficient of determination of 0.75 in Musina, while that of random forest, support vector machine, ridge, and lasso are 0.52, 0.59, and 0.39, respectively. This result presents hope for farmers in Musina as many complain of decreased agricultural production within the last five years due to insufficient rainfall and water scarcity [24]. 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 a need to study Upington more closely due to its changing environmental and atmospheric conditions at a rate different from global and continental estimates [25]. This region's significant decrease in rainfall poses a major challenge for livestock farming [26]. The mean absolute error, mean square error, and root mean square errors 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.

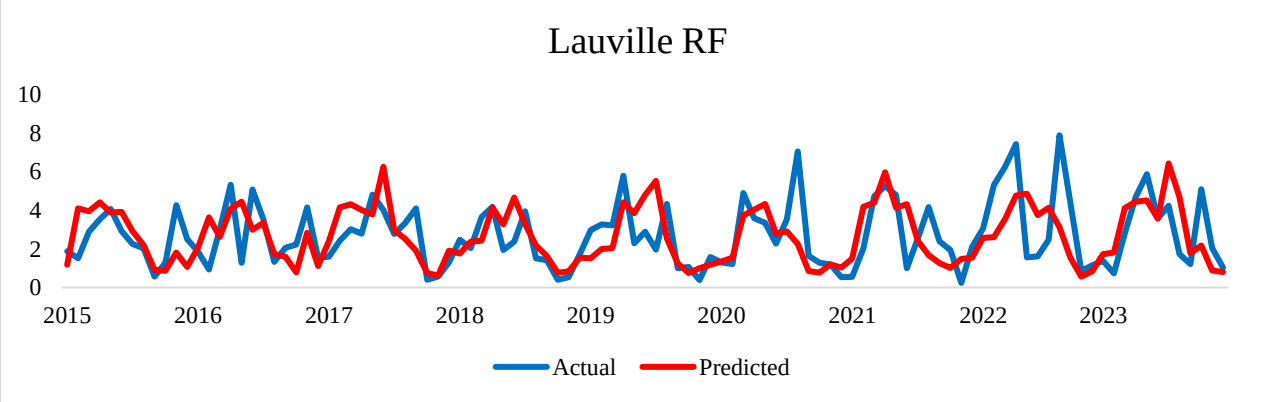
Figure 5 shows the actual and predicted rainfall graph using different models for the three locations under the hot arid climate classification. The result showed that random forest performed best in Upington as it predicted the increased rainfall in 2022 and its seasonal variability. It, however, performed poorly in rainfall estimation between 2015 and 2018, after which the performance was good. Other models have similar patterns in Upington. Musina, Ridge, Lasso, and random forest also performed best compared to linear regression as they accurately predicted the seasonal variation with underestimation of rainfall in 2018, 2021, and 2022. All models performed well in Lauville.

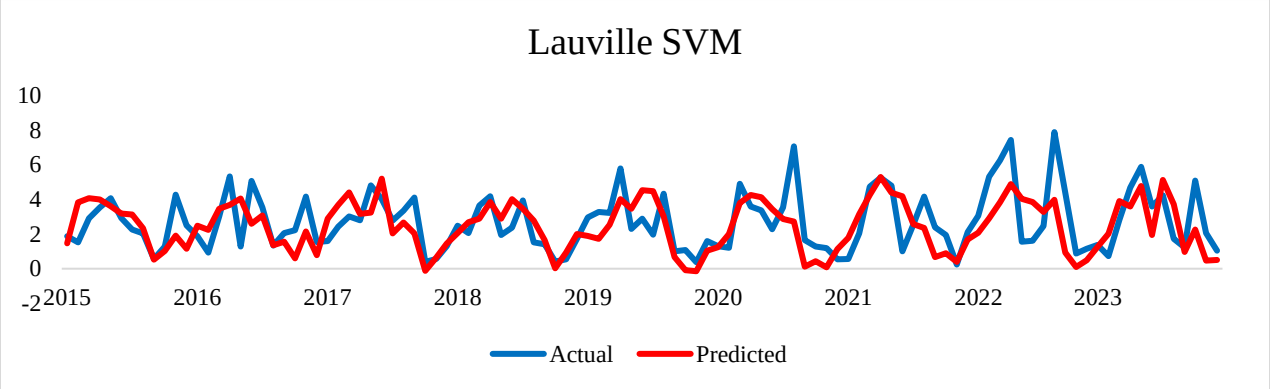
Table 5.
Shows model evaluation metrics for hot, arid desert climate classification.

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

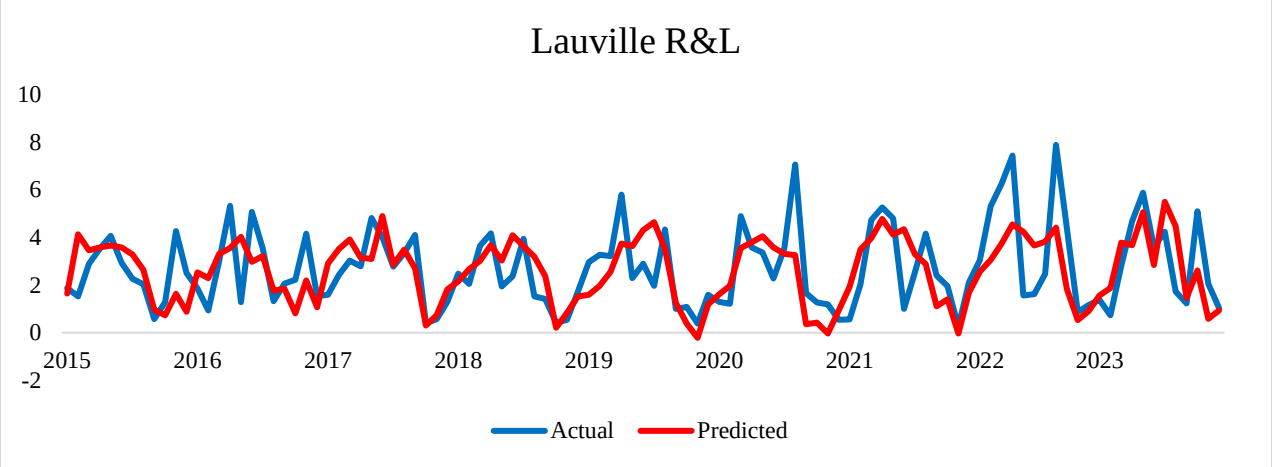
Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

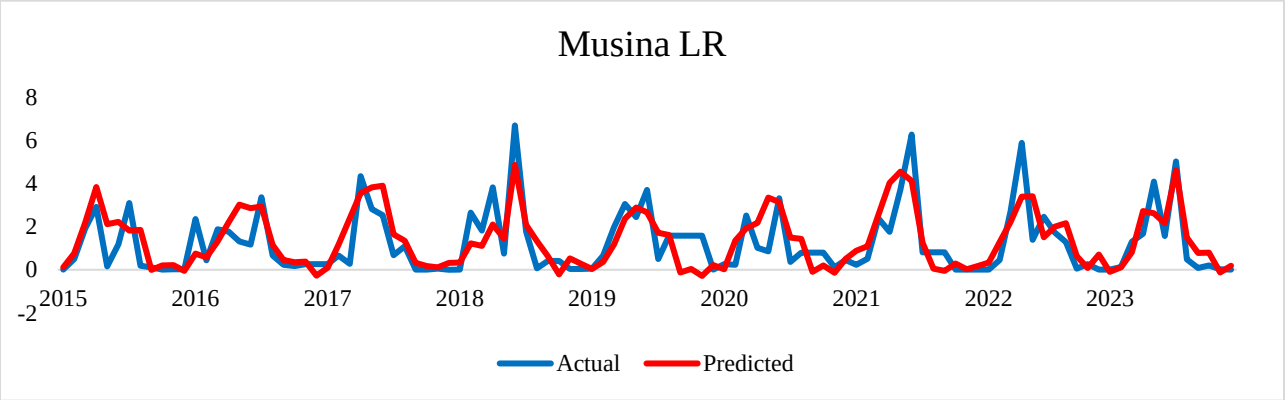


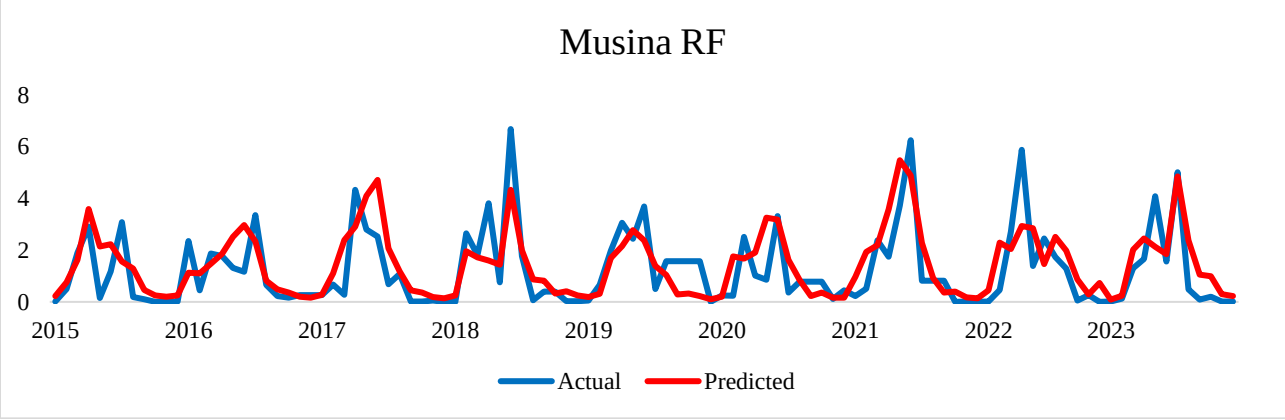




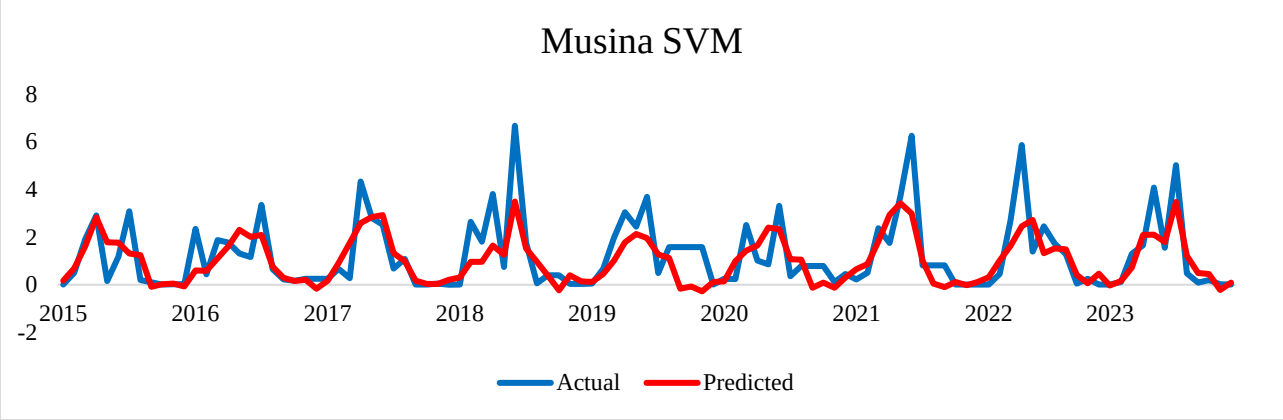
Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

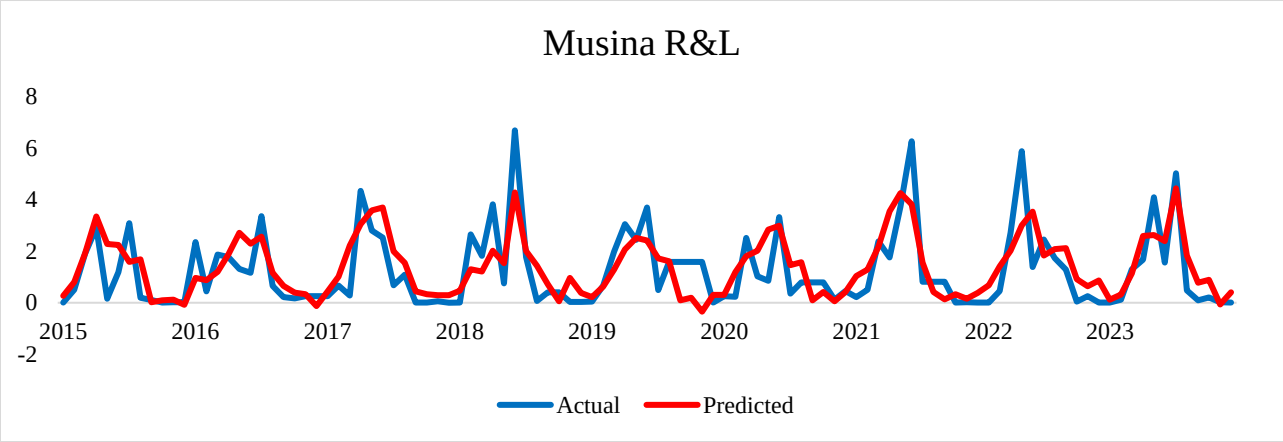


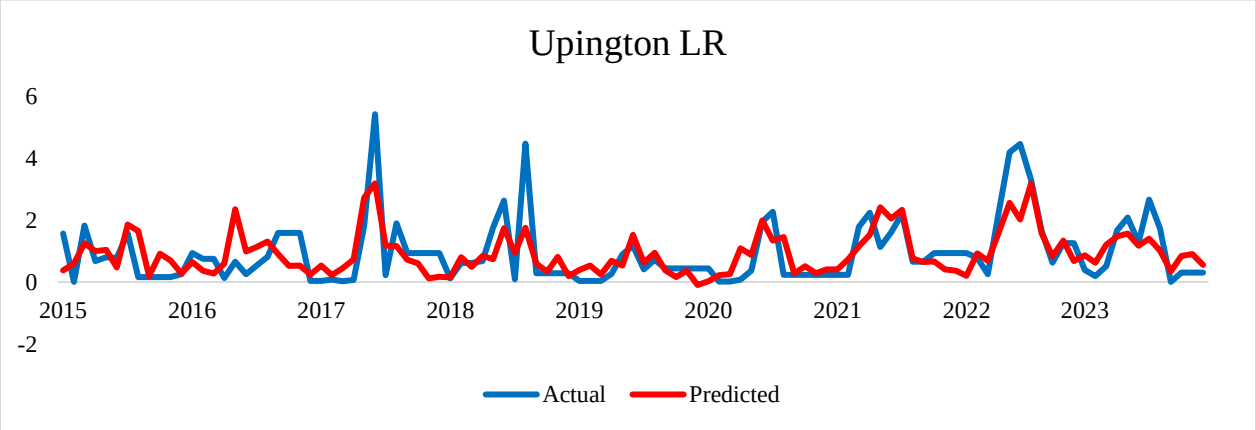




Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249







Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

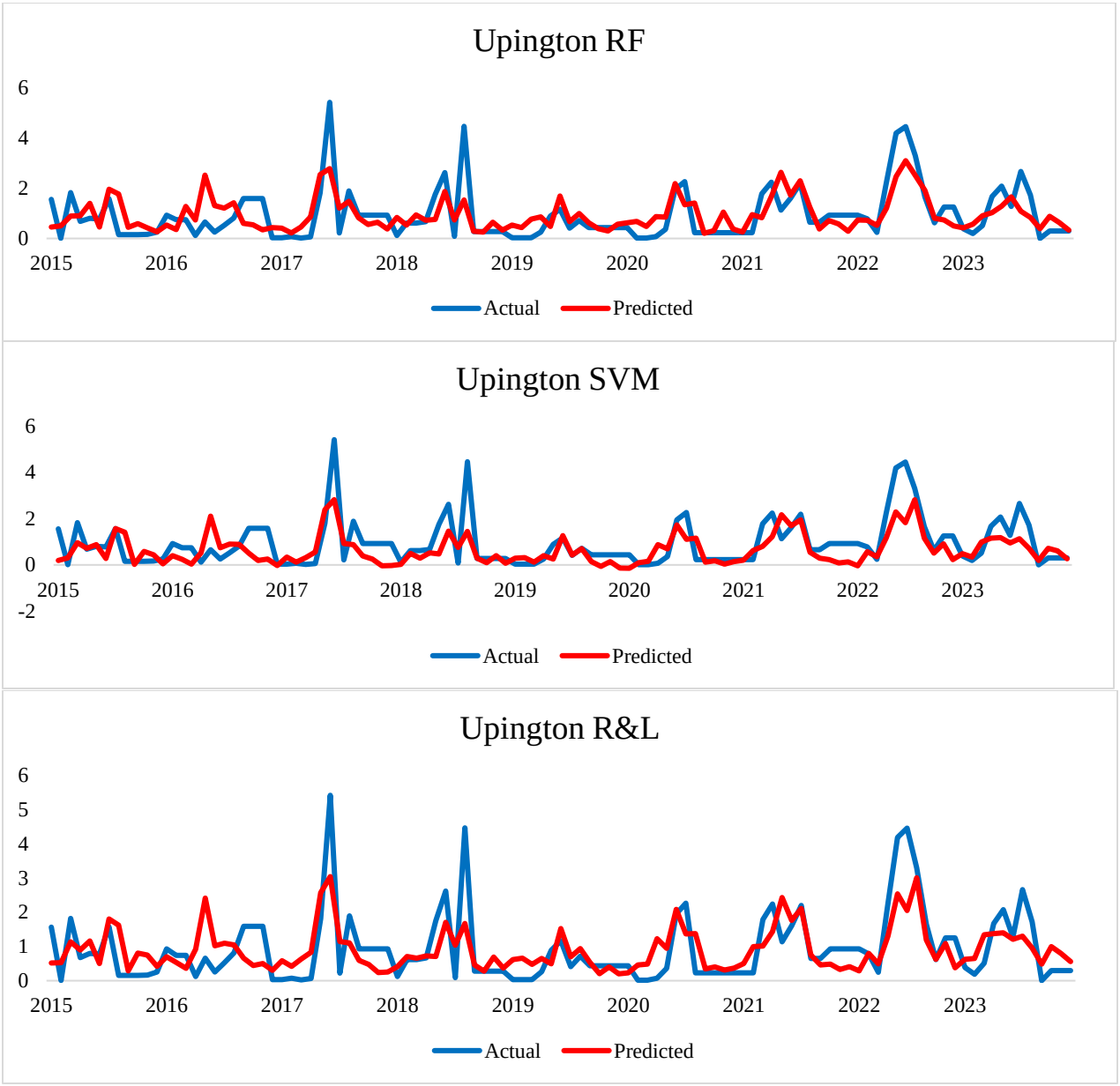


Figure 5.
Shows the predicted and actual rainfall for models used and locations under the hot, arid desert climate classification.

4. 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, the arid region of South Africa was divided into four climate classifications. Three locations were selected for each classification to test four machine learning models. This aims to

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

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. However, much work has not been done in southern and South Africa on predicting 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 accurately predict the seasonality and estimate the rainfall received in the cold and semi-arid steppe climate classification. Random forest, ridge, and lasso also performed well in Springfontein and Welkom. The 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. However, it is suggested that 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 the Northwest Province, and Port Elizabeth in the Eastern Cape Province. In this climatic zone, all models performed poorly in predicting Port Elizabeth's rainfall while excellently forecasting rainfall in Kimberly and Mahikeng. This study suggests that any models can be used for future work in hot and semi-arid climates. 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 in 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.

In addition, the results showed that cloud cover, dew point and water vapour had high correlation with rainfall for rainfall prediction. 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 strongly impacted rainfall prediction in a few locations.

Funding:
This work is funded by the DSCI-NICIS National e-science Postgraduate Teaching and Training Platform (NEPTTP)

Transparency:
The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.

Acknowledgments:
The authors gladly recognize the infrastructure support offered by Sol Plaatje University for this study.

References

[1] A. Ghazikhani, I. Babaeian, M. Gheibi, M. Hajiaghaei-Keshteli, and A. M. Fathollahi-Fard, "A smart post-processing system for forecasting the climate precipitation based on machine learning computations," *Sustainability*, vol. 14, no. 11, p. 6624, 2022. https://doi.org/10.3390/su14116624

[2] C.-H. Jeong and M. Y. Yi, "Correcting rainfall forecasts of a numerical weather prediction model using generative adversarial networks," *Journal of Supercomputing*, vol. 79, no. 2, pp. 1289-1317, 2023.

[3] D. Lucatero, H. Madsen, J. C. Refsgaard, J. Kidmose, and K. H. Jensen, "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*, vol. 22, no. 7, pp. 3601-3617, 2018. https://doi.org/10.5194/hess-22-3601-2018

[4] D. Sexton *et al.*, "Finding plausible and diverse variants of a climate model. Part 1: Establishing the relationship between errors at weather and climate time scales," *Climate Dynamics*, vol. 53, no. 1, pp. 989-1022, 2019.

[5] C. Huntingford, E. S. Jeffers, M. B. Bonsall, H. M. Christensen, T. Lees, and H. Yang, "Machine learning and artificial intelligence to aid climate change research and preparedness," *Environmental Research Letters*, vol. 14, no. 12, p. 124007, 2019. https://doi.org/10.1088/1748-9326/ab4e55

[6] B. Bochenek and Z. Ustrnul, "Machine learning in weather prediction and climate analyses—applications and perspectives," *Atmosphere*, vol. 13, no. 2, p. 180, 2022. https://doi.org/10.3390/atmos13020180

[7] B. Wang *et al.*, "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)*, 2019.

[8] P. Hewage, M. Trovati, E. Pereira, and A. Behera, "Deep learning-based effective fine-grained weather forecasting model," *Pattern Analysis and Applications*, vol. 24, no. 1, pp. 343-366, 2021.

[9] S. Landman, F. A. Engelbrecht, and C. J. Engelbrecht, "A short-range weather prediction system for South Africa based on a multi-model approach," *Water SA*, vol. 38, no. 5, pp. 765-774, 2012.

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

[11] M.-J. M. Bopape *et al.*, "Evaluating South African weather service information on Idai tropical cyclone and KwaZulu-Natal flood events," *South African Journal of Science*, vol. 117, no. 3-4, 2021.

[12] F. Muyambo, A. J. Jordaan, and Y. T. Bahta, "Assessing social vulnerability to drought in South Africa: Policy implication for drought risk reduction," *Jàmbá: Journal of Disaster Risk Studies*, vol. 9, no. 1, pp. 1-7, 2017. https://doi.org/10.4102/jamba.v9i1.326

[13] M.-A. Baudoin, C. Vogel, K. Nortje, and M. Naik, "Living with drought in South Africa: Lessons learnt from the recent El Niño drought period," *International Journal of Disaster Risk Reduction*, vol. 23, pp. 128-137, 2017.

[14] F. Mare, Y. T. Bahta, and W. Van Niekerk, "The impact of drought on commercial livestock farmers in South Africa," *Development in Practice*, vol. 28, no. 7, pp. 884-898, 2018.

[15] M. E.-A. R. C. Moeletsi, Z. P.-A. R. C. Shabalala, G.-A. R. C. De Nysschen, M. E. Moeletsi, and S. Walker, "Evaluation of an inverse distance weighting method for patching daily and dekadal rainfall over the Free State Province, South Africa," *Water SA*, vol. 42, no. 3, pp. 466-474, 2016. https://doi.org/10.4314/wsa.v42i3.12

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

[17] Government Communication and Information System, *South African yearbook 2011/2012*. Pretoria, South Africa: Government of the Republic of South Africa, 2012.

[18] I. Sungkawa and A. Rahayu, "Extreme rainfall prediction using Bayesian quantile regression in statistical downscaling modeling," *Procedia Computer Science*, vol. 157, pp. 406–413, 2019.

[19] R. H. Parker, "Strategies in the Beaufort West region to mitigate the negative financial impacts of a drought," Doctoral Dissertation, Stellenbosch: Stellenbosch University, 2020.

[20] National Drought Task Team, *South African drought report*. Pretoria, South Africa: Department of Agriculture, Forestry and Fisheries, 2015.

[21] L. Descamps, D. Le Roy, and A.-L. Deman, "Microfluidic-based technologies for CTC isolation: a review of 10 years of intense efforts towards liquid biopsy," *International Journal of Molecular Sciences*, vol. 23, no. 4, p. 1981, 2022.

[22] T. M. Tladi, J. M. Ndambuki, T. O. Olwal, and S. S. Rwanga, "Groundwater level trend analysis and prediction in the upper crocodile sub-basin, South Africa," *Water*, vol. 15, no. 17, p. 3025, 2023.

[23] A. T. Yakubu, A. Abayomi, and N. Chetty, "Machine learning-based precipitation prediction using cloud properties," presented at the International Conference on Hybrid Intelligent Systems (pp. 243-252). Cham: Springer International Publishing, 2021.

[24] Z. J. Mokgwathi, "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, 2018.

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249

[25] S. Strydom, M. Savage, and A. Clulow, "Long-term trends and variability in the dryland microclimate of the Northern Cape Province, South Africa," *Theoretical and Applied Climatology*, vol. 137, no. 1, pp. 963-975, 2019.
[26] S. J. Roffe, J. M. Fitchett, and C. J. Curtis, "Investigating changes in rainfall seasonality across South Africa: 1987–2016," *International Journal of Climatology*, vol. 41, pp. E2031-E2050, 2021. https://doi.org/10.1002/joc.6856

Edelweiss Applied Science and Technology
ISSN: 2576-8484
Vol. 9, No. 8: 148-178, 2025
DOI: 10.55214/2576-8484.v9i8.9249