Received: 15 September 2022 | Revised: 22 April 2023 | Accepted: 18 September 2023

DOI: 10.1002/gdj3.228

Geoscience Data Journal | RMetS | WILEY

DATA ARTICLE

Unsupervised segmentation and clustering time series approach to Southern Africa rainfall regime changes

Lovemore Chipindu[1] | **Walter Mupangwa**[2] | **Isaiah Nyagumbo**[1] | **Mainassara Zaman-Allah**[1]

[1]International Maize and Wheat Improvement Center (CIMMYT), Southern Africa Regional Office, Harare, Zimbabwe

[2]Marondera University of Agricultural Sciences and Technology, Faculty of Earth and Environmental Sciences, Marondera, Zimbabwe

Correspondence
Lovemore Chipindu, International Maize and Wheat Improvement Center (CIMMYT), Southern Africa Regional Office, P.O. Box MP163, Mt. Pleasant, Harare, Zimbabwe.
Email: l.chipindu@cgiar.org and lovemore.datascience@gmail.com

Funding information
CGIAR Fund Council

Abstract

Analysis of hydro-climatological time series and spatiotemporal dynamics of meteorological variables has become critical in the context of climate change, especially in Southern African countries where rain-fed agriculture is predominant. In this work, we compared modern unsupervised time series and segmentation approaches and commonly used time series models to analyse rainfall regime changes in the coastal, sub-humid and semi-arid regions of Southern Africa. Rainfall regimes change modelling and prediction inform farming strategies especially when choosing measures for mixed crop–livestock farming systems, as farmers can decide to do rainwater harvesting and moisture conservation or supplementary irrigation if water resources are available. The main goal of this study was to predict/identify rainfall cluster trends over time using regression with hidden logistic process (RHLP) or hidden Markov model regression (HMMR) supplemented by autoregressive integrated moving average (ARIMA) and Facebook Prophet models. Historical time series rainfall data was sourced from meteorological services departments for selected site over an average period of 55 years. Commonly used approaches forecasted an upward rainfall trend in the coastal and sub-humid regions and a declining trend in semi-arid areas with high variability between and within seasons. For all sites, Ljung-Box Test Statistics suggested the existence of autocorrelation in rainfall time series data. Prediction capabilities were investigated using the root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) which indicated not much difference between ARIMA and Facebook Prophet models. RHLP and HMMR offered a unique clustering and segmentation approach examining between and within-season rainfall variability. A maximum of 20 unique rainfall clusters with similar trend characteristics were determined as going beyond this

Dataset The datasets were sourced from Southern Africa countries namely Malawi, Mozambique, South Africa, and Zimbabwe where recorded historical rainfall data was available. In each country different meteorological stations were selected and were then categorized into coastal, sub-humid and semi-arid zones based on agro-ecological regions. The following information was collected for each meteorological service station; country, meteorological service station name, agro-ecological region, year, recorded rainfall, and stations geographical coordinates (Data S1).

brought non-significant difference to regime changes. A clear trend was exhibited from 1980 going backwards as compared to recent years signifying how unpredictable is rainfall in Southern Africa. The unsupervised approaches predicted a clear cluster trend in coastal than in sub-humid and semi-arid and the performance was assessed using Akaike information criteria and log-likelihood which showed improvement in prediction power as the number of segmentation clusters approaches 20.

KEYWORDS

ARIMA, coastal, Facebook Prophet, hidden Markov model regression, regression with hidden logistic process, semi-arid, sub-humid

1 | INTRODUCTION

Rainfall is one of the most limiting factors in rain-fed farming systems of Africa, since it determines availability of soil moisture required for crop productivity (Mkuhlani et al., 2018; Mlenga, 2016; Nyagumbo et al., 2020). High rainfall variability has severe implications for food production and livelihoods (Muktar et al., 2020). The amount and distribution of rainfall determines suitability of crop types and varieties and related agronomic management at different locations (Babaousmail et al., 2021). Low or sub-optimal rainfall cause agricultural drought that retard plant growth and reduce yields (Su et al., 2021), while extremely high rainfall events cause floods that damage crops (Guhathakurta & Rajeevan, 2008). The distribution of the global rainfall is shifting as our climate changes resulting in wet areas becoming wetter, dry areas becoming drier, storms becoming more intense and all this leading to more unpredictable weather around the world (Precipitation Measurement Missions, 2020). Much of the Southern Africa's population is exposed to climate change impacts, including changes in the frequency and severity of droughts, heat waves and cyclones (Shah et al., 2021; Su et al., 2021). Climatic extremes with adverse effects on crops and ecosystems include droughts, flooding, hailstorms, heat waves and frost or their combinations and these have significant effect on structure, functions, land use patterns and livelihoods in agroecosystems (Adhikari et al., 2015). Accurate prediction of rainfall pattern changes is particularly important in regions such as Southern Africa where people are highly exposed and sensitive to environmental changes (Twenefour et al., 2018). In fact, understanding the spatial and temporal patterns of climate change and variability is a key step towards designing and targeting appropriate adaptation strategies.

Rainfall forecasting has commonly been done using traditional methods that employ statistical techniques such as ARIMA models, Fuzzy Time Series (FTS), Prophet and Theil's Regression. Among those methods, ARIMA and Prophet were the most used because of less forecasting and prediction errors based on root mean square error (RMSE), Mean absolute error (MAE),and mean absolute percentage error (MAPE). More recently, emphasis was put on the modelling of time series as a tool to improve the management and forecasting of the earth's meteorological and hydrological resources, including rainfall patterns (Asfaw et al., 2018). Time series represents a dynamic measure of a physical process over a given period and may be discrete or continuous (Singh, 2018). It is becoming more difficult to predict rainfall changes using traditional time series models, hence the need for modern approaches. While all the numerous advanced tools and techniques are employed for data analysis such as machine learning (ML), Internet of Things (IoT) and others, one of the techniques frequently preferred for analysing such data is statistical time series (Babaousmail et al., 2021). Time series algorithms are used extensively for analysing and forecasting time-based data such as rainfall (Wimhurst & Greene, 2021). However, given the complexity of other factors apart from time, ML has emerged as a powerful method for understanding hidden complexities in time series data and generating good forecasts (Zhang et al., 2021). Furthermore, examining the spatiotemporal dynamics of meteorological variables in the context of changing climate, particularly in Southern African countries where rain-fed agriculture is predominant, is vital to assess climate-induced changes and suggest feasible adaptation strategies (Asfaw et al., 2018). The commonly used time series models include the ARIMA and the Prophet which was developed by Facebook to manage extreme events and other external circumstances such as effects of holiday (Akdag & Bozma, 2021; Khayyat et al., 2021).

In this study, two approaches are suggested to supplement the traditional models in analysing complex rainfall trends that are currently being witnessed in Southern Africa. Regression with hidden logistic

process (RHLP) and hidden Markov model regression (HMMR) are powerful tools in modelling time series and are classified as unsupervised ML approaches as they can analyse and cluster unlabelled time series data (Akdag & Bozma, 2021). Hidden Markov model (HMM) offers a natural tool for dealing with one of the fundamental problems in stochastic modelling (Chamroukhi et al., 2011). The source of strength of the HMM seems to be due to its ability to acknowledge the relationships between changing regimes on a short-term basis, one could adequately model the observed data by a homogeneous process (Huang et al., 2018). The second source of strength is their exceptional ability to incorporate structural feature of the phenomena under study into a structural feature of the model. The topology of the HMM (the number of states, the transition matrix structure and observed sequence distributions) is designed to incorporate as many features of the observed process as the underlying science can justify (Chamroukhi et al., 2009). The RHLP is a new approach for signal parameterization in the context of the rainfall regime changes (Chamroukhi et al., 2009). This approach is based on a regression model incorporating a discrete hidden logistic process or abrupt switching between polynomial regressive components overtime (Fridman & Angeles, 2010).

In 2008, the WCRP Working Group on Coupled Modelling (WGCM), endorsed the CMIP5 (CMIP Phase 5) protocol (Di Luca et al., 2020), which defined a set of 35 model experiments designed to be useful in: (a) assessing the mechanisms responsible for model differences in poorly understood feedbacks associated with carbon cycle and with clouds, (b) examining climate predictability and exploring the ability of the models to predict climate on decadal time scale and (c) determining why similarly forced models produce a range of responses. However, several studies revealed that CMIP models systematically exaggerate the magnitude of daily rainfall anomalies, as they are designed to suit decadal hindcasts and predictions simulations, long-term simulations, and atmosphere-only (prescribed SST) simulations for computationally demanding models (Di Luca et al., 2020; Lei et al., 2023). The application of unsupervised ML time series approaches supplement CMIP models, as they do not require labelled dataset. As a replacement for that, the models are in such a way that they themselves recognize the hidden patterns and insights from the given time series data (Atiqul Haq et al., 2021). Unlike, other commonly used time series models, the unsupervised ML approach clusters the time series according to the trend similarities which makes it easy to identify extreme events which largely contribute to biased predictions or simulations and moreover, they are not guided by several assumptions.

This study was designed to explore the use of modern unsupervised learning approaches in analysing and predicting rainfall trends in coastal, sub-humid and semi-arid regions of Southern Africa. Rainfall regimes change modelling and prediction help in decision making especially when choosing adaptation measures for mixed crop–livestock farming systems of southern Africa, as well as sub–Saharan Africa. The study focussed on (a) assessing RHLP and HMMR models prediction accuracy of rainfall regime/cluster changes in Southern Africa over a given period was assessed (b) evaluating the effectiveness of the proposed models in clustering historical frequency rainfall regime changes as compared to specific predictions offered by ARIMA and Prophet models and (c) testing the prediction and clustering power of each of the proposed approach in understanding the fluctuating rainfall regimes in Southern Africa. The specific objectives of this study were (a) to define/identify a shift in rainfall trends overtime using the modern recommended approaches in comparison with the traditional model and (b) to recommend the best models that can handle the complexity of rainfall regime changes. The accuracy level of the ML models used in predicting rainfall based on historical data has been one of the most critical concerns in hydrological studies. An accurate ML forecasting model can give early alerts of severe weather to help prevent natural disasters and destruction. Hence, there is need to develop ML algorithms.

2 | MATERIALS AND METHODS

2.1 | Study sites

The analysis mainly focused on four Southern Africa countries namely Malawi, Mozambique, South Africa and Zimbabwe where historical rainfall data were available. In each country, different meteorological stations were selected and were then categorized into coastal, sub-humid and semi-arid zones based on agro-ecological regions. This category guided the analysis of the gathered rainfall data as it was based on these major classifications. Table 1 and Figure 1 below summarize the geographical information of the selected sites and the years considered in the analysis. Historical rainfall data was measured by relevant meteorological services departments in different sites.

2.2 | Statistical analysis

The analysis was conducted in R version 4.1.0 (2021-05-18) using the samurais' package which is a toolbox including many original and flexible user-friendly

TABLE 1 Geographical information of the selected sites and the years considered in the analysis.

Country	Station	Years	Latitude	Longitude	Altitude (m)	Region
Zimbabwe	Harare	1963–2001	−17.72	31.02	1475	Sub-humid
	Marondera	1952–2000	−18.93	31.54	1658	Sub-humid
	Bulawayo	1931–2001	−20.16	28.61	1356	Semi-arid
	Matopos	1940–2015	−20.51	28.44	1347	Semi-arid
	West Nich.	1963–2001	−21.06	29.36	864	Semi-arid
	Beitbridge	1952–2001	−22.21	29.99	462	Semi-arid
Malawi	Chitala	1948–1999	−13.68	34.25	606	Semi-arid
	Chitedze	1981–2013	−13.98	33.64	1100	Sub-humid
	Dedza	1959–1999	−14.32	34.25	1632	Sub-humid
Mozambique	Chimoio	1952–2012	−19.25	33.43	693	Sub-humid
	Pemba	1952–2005	−12.59	40.52	70	Coastal
	Quelimane	1961–2008	−17.86	36.87	5	Coastal
	Xai Xai	1952–1989	−25.09	33.53	2	Coastal
South Africa	Harmony	1905–2000	−23.08	29.85	517	Sub-humid
	Levubu	1966–2004	−23.08	30.28	706	Sub-humid
	Mertz	1905–2000	−26.5	28.36	1521	Sub-humid
	Polokwane	1961–2006	−23.73	29.59	1194	Sub-humid

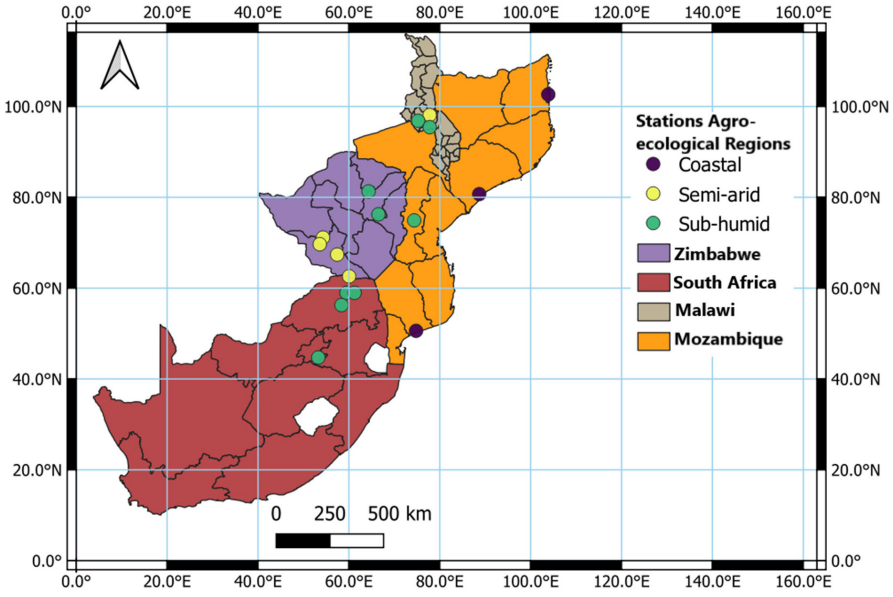

FIGURE 1 Map of part of southern Africa showing agro-ecological regions of the meteorological stations used in the study.

statistical latent variable models and efficient unsupervised algorithms to segment and represent time series data (univariate or multivariate), and more generally, longitudinal data, which include regime changes (Chamroukhi et al., 2013). To generalize the measured rainfall time series data in terms of skewness, kurtosis and existence of extreme events, probability distributions were plotted for each site. The following are major statistical approaches used to answer the raised research questions.

2.2.1 | Autoregressive integrated moving average (ARIMA)

The autoregressive model was used as a traditional time series approach in modelling the rainfall historical patterns in the costal, sub-humid and semi-arid regions of Southern Africa. A prediction period of 50 years was considered to be realistic. The approach was implemented first before the un-supervised machine learning algorithms to get a general understanding to the rainfall

trends. An autoregressive integrated moving average (ARIMA) is a statistical analysis model that uses time series data to either better understand the dataset or to predict future trends (Wang et al., 2014). A statistical model is autoregressive if it predicts future values based on past values (Shao et al., 2021). ARIMA model assume that past values have some residual effect on current or future values, and it is generally given by the following formula.

$$\Delta Y_t = c + \emptyset_1 \Delta Y_{t-1} + \theta_1 \varepsilon_{t-1} + \varepsilon_t$$

Y_t and Y_{t-1} represent the rainfall values in the current period and 1 period ago, respectively. ε_t and ε_{t-1} are the error terms for the same two periods. $\emptyset_1$ and θ_1, express what parts of the value Y_{t-1} and error ε_{t-1} last rainfall period are relevant in estimating the current one.

Ljung-box test statistic

The Ljung-box test statistic was implemented to assess if a group of autocorrelations of the rainfall historical time series are different from zero that is if autocorrelation exists (Kim et al., 2004).

The Ljung-Box test uses the following hypotheses:

- H_0: The residuals are independently distributed (the model does not exhibit lack of fit)
- H_1: The residuals are not independently distributed; they exhibit serial correlation (the model exhibits lack of fit)

As a rule of thumb, we would like to fail to reject the null hypothesis that is the residuals are independently distributed (Burns, 2002). If the *p*-value of the test is greater than 0.05, it means the residuals for rainfall time series model are independent.

2.2.2 | Facebook Prophet model

To further understand the general rainfall trend, the Prophet model developed by Facebook was used to supplement the ARIMA approach. Understanding that some of the extreme rainfall events are because of cyclones, implementing this approach was an appropriate method to handling rare events and give a better prediction. The model is based on a decomposable additive model where nonlinear trends are fit with seasonality, and it also considers the effects of rare events (Hossain et al., 2021). In general, it is given by the formula:

$$y(t) = g(t) + s(t) + h(t) + \epsilon_i$$

where, $y(t)$=future rainfall values, $g(t)$=rainfall trend changes that do not repeat, $s(t)$=repeated seasonal changes, $h(t)$=irregular changes like cyclones and ϵ_i=leftover unique errors that cannot be explained.

2.2.3 | ARIMA and Facebook Prophet performance evaluation

To evaluate the performance of ARIMA and Facebook Prophet models, three major statistics were used namely root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) (Luo et al., 2021). The RMSE is a quadratic scoring rule which measures the average magnitude of the error of a model in predicting quantitative data. The lower the RMSE, the better a given model can fit a dataset. The MAE measures the average magnitude of the errors in a set of forecasts, without considering their direction and it is the average over the verification sample of the absolute values of the differences between forecast and the corresponding observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average. The mean absolute percentage error (MAPE) is a measure of how accurate a forecast system is. It measures this accuracy as a percentage and can be calculated as the average absolute per cent error for each time minus actual values divided by actual values.

2.2.4 | Regression with hidden logistic process (RHLP)

Regression with hidden logistic process is a new approach for signal parametrization, which consists of a specific regression model incorporating a discrete hidden logistic process. The model parameters are estimated by the maximum likelihood method performed by a dedicated expectation maximization (EM) algorithm (Landwehr et al., 2005). The parameters of the hidden logistic process, in the inner loop of the EM algorithm, are estimated using a multi-class Iterative Reweighted Least-Squares (IRLS) algorithm (Mohan & Fazel, 2010).

Global regression model

The random sequence $\boldsymbol{x} = (x_1, \ldots, x_n)$ represents the rainfall changes of n real observed or recorded rainfall observations, where x_i is observed at time t_i (years). The sample was assumed to be generated by the following regression model with a discrete hidden logistic process.

$\boldsymbol{z} = (z_1, \ldots, z_n)$, where $z_i \in \{1, \ldots, K\}$:

$$x_i = \beta_{z_i}^T \boldsymbol{r_i} + \varepsilon_i; \quad i = 1, \ldots, n \tag{1}$$

In this model, β_{z_i} is the $(p+1)$ dimensional coefficients vector of a p degree polynomial.

$r_i = \left(1, t_i, \ldots, (t_i)^p\right)^T$ Is the time-dependent $(p+1)$-dimensional covariate vector associated to the parameter β_{z_i} and theε_i are independent random variables distributed according to a Gaussian distribution with zero mean and variance $\delta^2_{z_i}$ (Chamroukhi et al., 2009).

The hidden logistic process

The probability distribution of the process $\boldsymbol{z} = (z_1, \ldots, z_n)$ that allows the switching from one regression model to another is defined. The proposed hidden logistic process supposes that the variables z_i, given the vector $t = (t_1, \ldots, t_n)$, are generated independently according to the multinomial distribution $M\left(1, \pi_{i1}(\boldsymbol{w}), \ldots, \pi_{iK}(\boldsymbol{w})\right)$, where;

$$\pi_{iK}(\boldsymbol{w}) = p(z_i = k; w) = \frac{e^{(w^T{}_k\, v_i)}}{\sum_{l=1}^{K} e^{(w^T{}_l\, v_i)}} \tag{2}$$

Is the logistic transformation of a linear function of the time-dependent covariate? $v_i = \left(1, t_i, \ldots, (t_i)^q\right)^T$, $w_k = \left(w_{ko}, \ldots, w_{kq}\right)^T$ is the $(q+1)$ dimensional coefficients vector associated to the covariate v_i and $w = (w_1, \ldots, w_k)$. Thus, given the vector $\boldsymbol{t} = (t_1, \ldots, t_n)$, the distribution of z can be written as:

$$p(z; w) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\frac{e^{(w^T{}_k\, v_i)}}{\sum_{l=1}^{K} e^{(w^T{}_l\, v_i)}} \right)^{z_{ik}} \tag{3}$$

where $z_{ik} = 1$ if $z_i = k$, that is when x_i is generated by the k^{th} regression model, and 0 otherwise.

2.2.5 | Hidden Markov model regression (HMMR)

Hidden Markov model regression (HMMR) is an extension of the hidden Markov model (HMM) to regression analysis. We assume that the parameters of the regression model are determined by the outcome of a finite-state Markov chain and that the error terms are conditionally independent normally distributed with mean zero and state-dependent variance (Chamroukhi et al., 2009; Lal & Bhat, 1988). We consider the problem of maximum likelihood estimation of the HMMR parameters and develop analogues for the methods used in HMM's for our regression case. Hidden Markov model (HMM).

The Markov model was defined as $\{Y_t\}$ the observed sequence of an HMM process there exist a Markov chain $\{Q_t\}$ on the state space $S = (S_1, \ldots, S_n)$ and a cumulative distribution function $F_1, \ldots, F_N$ (Gupta, 2011) such that:

$$P(Y_1 \le c_1, \ldots, Y_t \le c_T \setminus Q_t = S_i) = P(Y_1 \le c_1, \ldots, Y_{t-1} \le c_{t-1}, Q_t = S_i)$$

$$F_i(c_t).P(Y_{t+1} \le c_{t+1}, \ldots, Y_T \le c_T \setminus Q_t = S_i).$$

2.2.6 | RHLP and HMMR model evaluation

Two major statistics were used to evaluate the models in determining the rainfall regime changes over time namely, Log-likelihood and Akaike's information criterion (AIC).

Log-likelihood

The likelihood ratio test assesses the goodness of fit of competing statistical models based on the ratio of their likelihoods, specifically one found by maximization over the entire parameter space and another found after imposing some constraint (Shimodaira, 2000). The statistics was used to assess the changes in goodness of fit of the models at different clusters.

Akaike's information criterion

The Akaike information criterion (commonly referred to simply as AIC) is a criterion for selecting among nested statistical models. The AIC was used as an estimated measure of the quality of each of the available time series models as they relate to one another for a certain change in rainfall over time, making it an ideal method for model cluster selection (Akaike, 2011).

2.2.7 | RHLP and HMMR implementation in R

The models computation was implemented in R as follows:

emHMMR / emRHLP $(X, Y, K, p = 3, n\_\text{tries} = 1, \text{max\_iter} = 1500, \text{threshold} = 1e-06, \text{verbose} = \text{FALSE / TRUE})$

where, emHMMR/RHLP implements the EM (Baum-Welch) algorithm to fit a HMMR or RHLP model, X=numeric vector of length m representing the covariates/inputs (years) (x_1, ..., x_m), Y=numeric vector of length m representing the observed response/output (rainfall) (y_1, ..., y_m), K=the number of regimes/segments (HMMR components), p=the order of the polynomial regression, which is optional, $n_$tries=number of runs of the EM algorithm, max_iter=The maximum number of iterations for

the EM algorithm, threshold is a numeric value specifying the threshold for the relative difference of log-likelihood between two steps of the EM as stopping criteria, and verbose is logical value indicating whether values of the log-likelihood should be printed during EM iterations.

3 | RESULTS AND DISCUSSION

3.1 | Measured rainfall distributions

In sub-humid sites, the rainfall distributions are not perfectly symmetrical with the right tail more prolonged than the left and that is skewedness greater than zero or positively skewed as shown in Figure 2a below. The frequency of the distributions mostly lies between 450 and 1000 mm and the mean, median and mode are defined within this range. In Dedza and Chitedze the rainfall distributions are approximately symmetrical (mesokurtic), unlike in Chimoi, Levubu, Harare and Marondera where the rainfall is unevenly distributed and suggest the possibility of extreme rainfall events. In general, the sub-humid sites distributions are leptokurtic indicating a positive excess kurtosis and this suggests that this agro-ecological region is prone to extreme or excess rainfall. Excess rainfall poses challenges in crop production including waterlogging in cropping systems under conventional systems and improved practices such as conservation agriculture (Mkuhlani et al., 2018; Nyagumbo et al., 2020; Wall et al., 2013).

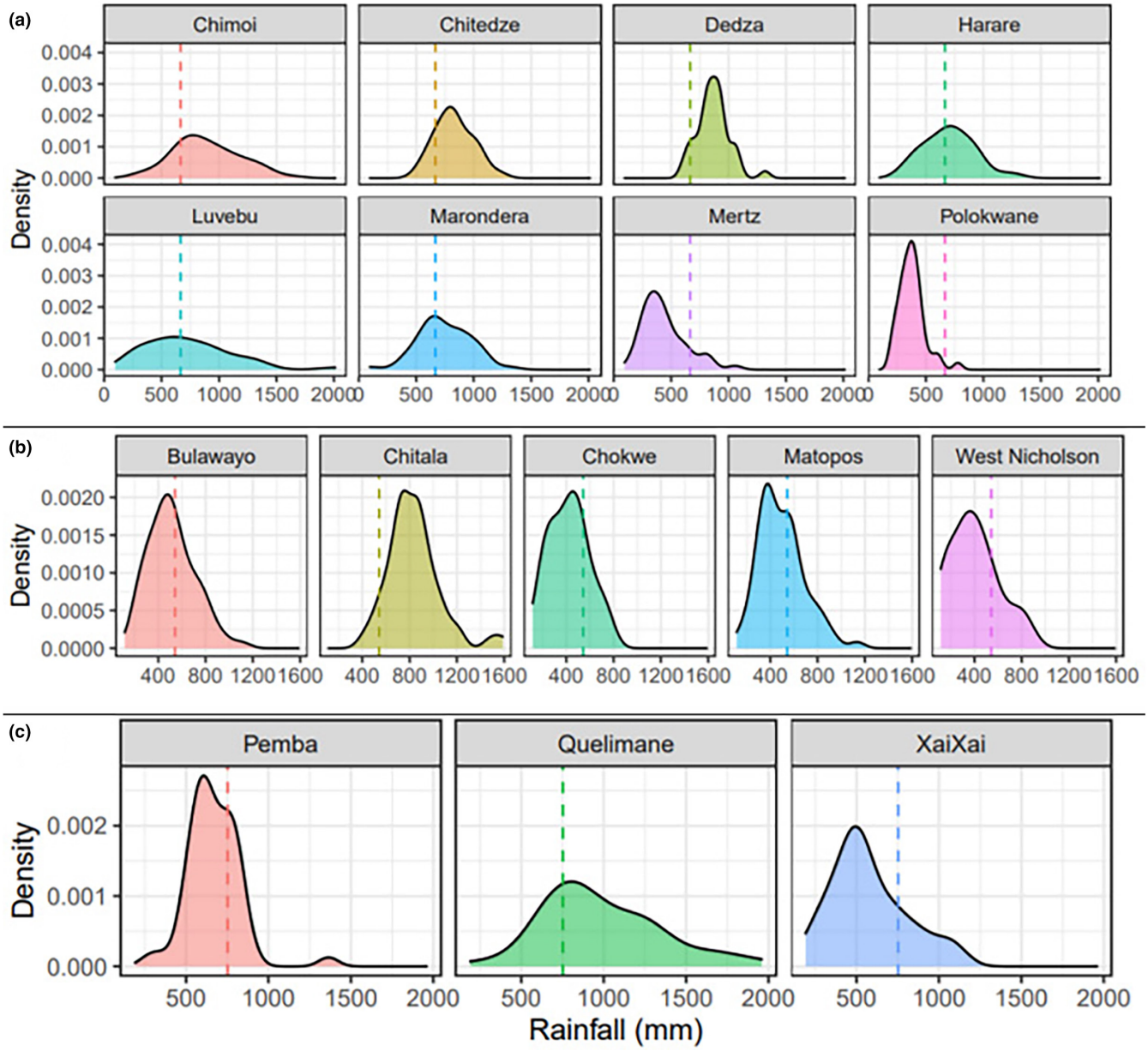

FIGURE 2 Sub-humid (a), semi-arid (b) and coastal (c) recorded rainfall distribution plots (dotted lines indicate the average rainfall).

Furthermore, in semi-arid regions the average rainfall is below 500 mm compared to sub-humid and coastal areas (Figure 2b). The distributions indicate positive skewness with heavy tails on the right which suggest extreme rainfall data points though the peakedness across all the semi-arid site in the study is similar. Extreme rainfall events occur through late season cyclones and events such as the El Nino as observed during the early 2000s (WFP, 2016). Livestock and dry land cereals (sorghum and millet) production are the major agricultural activities viable in semi-arid regions as rainfall is highly variable (Prasad & Staggenborg, 2011). Minimum tillage and mulching are other conservation agriculture techniques which can be practised in these areas as they help to conserve moisture.

In coastal areas, the rainfall distributions are unique and unevenly distributed with heavy tails which suggest several extreme rainfall events over 1500 mm in Pemba and Quelimane (Figure 2c). All the distributions are positively skewed with an average rainfall between 800 and 1000 mm. In general, understanding the rainfall distributions gave an insight about the effectiveness of the suggested time series approaches in modelling the data. Prophet models, HMMR and RHLP offered a special way of handling extreme rainfall events as observed in the distribution plots. Some conservation agriculture systems such as intercropping with different legumes gives better yield when the rainfall is excessive especially in coastal regions (Nyagumbo et al., 2020).

3.2 | RIMA and Facebook Prophet

In the coastal and sub-humid regions of Southern Africa rainfall regime change ranges between 600 mm and 800 mm as suggested by the forecasted values (Figure 3). Rainfall shows large year-to-year variations especially between 1980 and 2000. Although the actual rainfall for both models fluctuated from below 400 mm to above 1000 mm in some cases, the predicted values gave a uniform range estimated between 600 and 800 mm and slightly above after year 2000. The Facebook Prophet model forecasted an upward rainfall trend for the coastal regions from 2000 to 2050 but with a low probability of exceeding 1000 mm. A similar trend is also observed in the sub-humid regions where large variability is within and between seasons as indicated by the actual rainfall trend line while the forecasted trend line exhibited

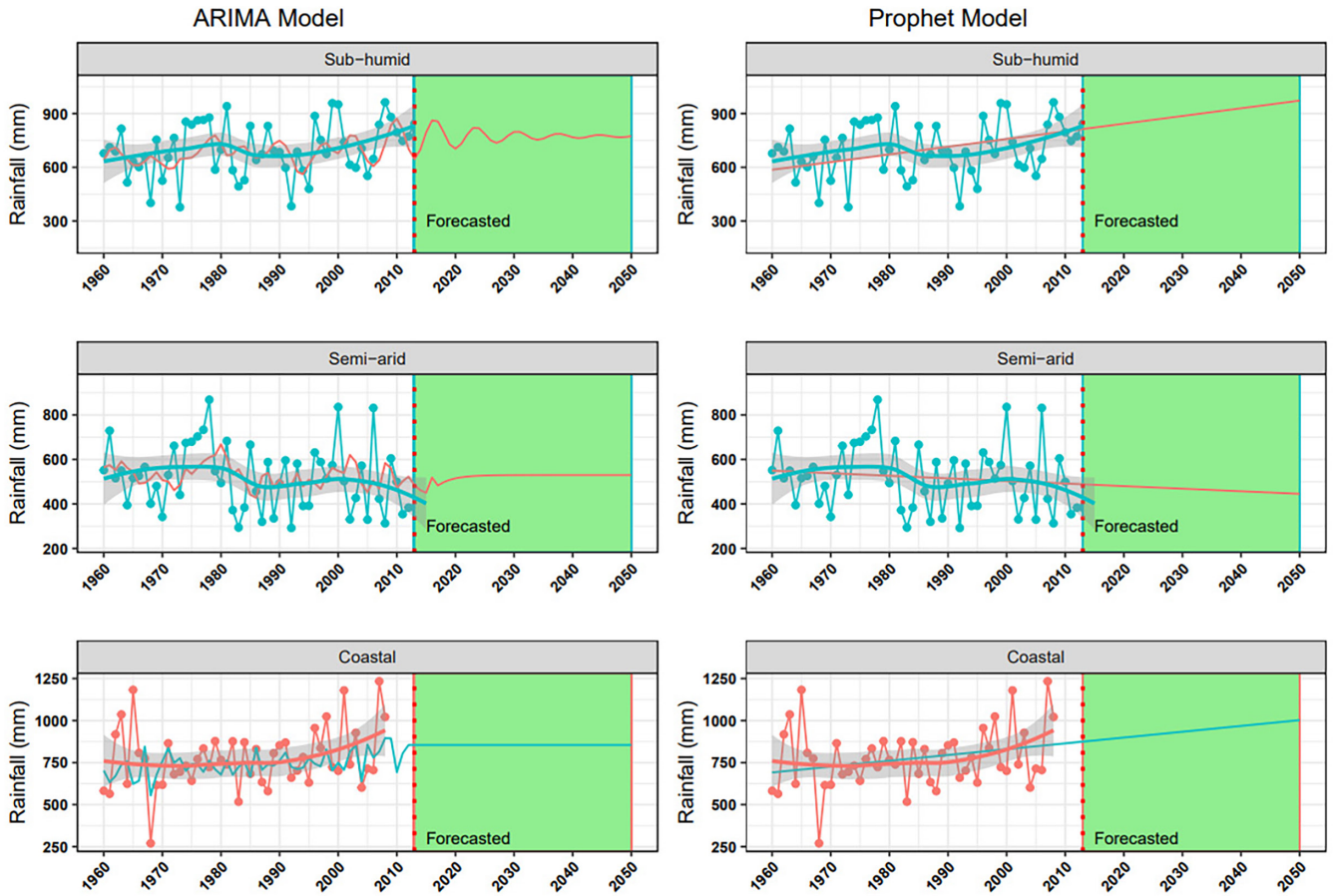

FIGURE 3 ARIMA and Facebook Prophet models actual and predicted rainfall trend based on a 50-year predictions period.

a slight increase in rainfall across years. However, the trend is totally different in the semi-arid regions of southern Africa as both the actual and forecasted rainfall shows a declining trend to less than 400 mm cross years.

The observed trends are supported by previous studies which predicted an increase in rainfall trend in sub-humid and coastal regions but a declining trend in semi-arid regions (Strauch et al., 2018). As rainfall trend is the major climate change indicator in most of the Southern Africa' regions with large seasons and across seasons variability, agriculture activities or practices must be adjusted incorporating adaptation strategies to curb the upwards and declining rainfall change effects on crop and livestock production (Adimassu & Kessler, 2016). Coastal areas are mostly threatened by rising sea levels and more intense cyclones, facing more crop-damaging heat waves, pests and flooding (Kyei-Mensah et al., 2019). Preserving vital ecosystems and species becomes mandatory as the rising seas threaten coastal barrier reefs, which protect communities from storm surges and wetlands, which filter impurities from water. A declining predicted rainfall trend in semi-arid regions suggest that conserving water resources remains important, as water shortages have wide-ranging consequences. For example, as sources of water used for irrigation dry up, the costs of producing food could rise. Lower water levels and higher temperatures in streams and rivers could diminish the capacity of hydropower and cause the collapse of some fisheries. Water prices could rise not only for farmers but also for industry and homeowners, especially in areas where growing populations are already putting stress on water resources, such as in Southern Africa (Malhi et al., 2020). Adoption of conservation agricultural technologies can be one of the important practices to curb the effects of high rainfall variability between and within season (Adimassu & Kessler, 2016). In semi-arid regions conservation agriculture components such mulching can contribute positively towards crop productivity through yield increase. Drought-tolerant hybrid crop varieties that are resilient to climate shocks must be a priority to households or farmers in the agriculture space.

A difference in the performance of the two models was observed in all the study sites across the three regions of Southern Africa (Table 2). As a rule of thumb, the lower the RMSE and MAE the better the model prediction. The performance evaluation results suggested that the ARIMA and Prophet models predict the rainfall changes with a similar accuracy in some cases as most of the computed statistics are in the same range. On average, the Facebook Prophet is better in predicting rainfall trend across agro-ecological regions compared with the ARIMA. This might be attributed to the effectiveness of the model in

TABLE 2 Root mean square error (RMSE), mean absolute error (MAE) and mean absolute percentage error (MAPE) evaluating the prediction performance of both the ARIMA and Facebook Prophet models in different study sites within the semi-arid, sub-humid and coastal regions of Southern Africa.

		ARIMA			Prophet			
Regions	**Site**	**RMSE**	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**No of years**
Semi-Arid	Bulawayo	193.675	152.991	0.364	178.073	144.419	0.343	71
	Chitala	220.954	163.394	0.208	221.302	162.808	0.207	52
	Chokwe	161.901	138.409	0.417	161.734	134.582	0.414	35
	Matopos	187.382	146.560	0.367	190.665	152.613	0.379	76
	W. Nicholson	204.751	166.495	0.616	163.479	135.260	0.482	39
Sub-humid	Chimoio	280.859	228.756	0.311	285.272	230.936	0.312	61
	Chitedze	163.814	131.919	0.168	163.814	131.913	0.168	33
	Dedza	136.100	106.067	0.126	138.017	104.650	0.125	41
	Harare	219.086	173.302	0.295	219.589	174.831	0.298	39
	Luvebu	376.071	297.305	0.564	377.106	297.521	0.564	39
	Marondera	225.447	181.633	0.377	226.398	183.036	0.385	49
	Mertz	190.936	147.188	0.396	190.992	147.744	0.396	96
	Polokwane	108.021	80.246	0.231	109.797	80.018	0.229	46
Coastal	Pemba	158.837	116.054	0.192	144.661	107.370	0.171	54
	Quelimane	349.372	284.225	0.353	352.070	284.380	0.349	48
	Xai Xai	223.714	180.123	0.370	206.622	159.879	0.331	38

handling extreme events as it was specifically developed to manage and take into consideration of the unexpected events (Lounis, 2021). Similarities in forecast accuracy of the two models pave way for advanced approaches such as unsupervised machine learning to investigate the rainfall variability.

3.2.1 | ARIMA model lack of fit test

The results in Figure 4 suggest that the computed Ljung-Box Test Statistic for all sites across agro-ecologies are not significant as the *p*-values are greater than 0.05. This is in line with the ARIMA model assumption that is the residuals must be independently distributed as an indication of lack of serial autocorrelation because rainfall is random. The alternative hypothesis is rejected in favour of the null hypothesis and the conclusion is that the ARIMA does not exhibit lack of fit. The ARIMA model fitness in modelling historical rainfall regime changes was further investigated to this extent as it depends on several assumptions compared to Prophet model. While the two traditional approaches give a general prediction overview of the time series, they lack the clustering and segmentation aspect to understand rainfall regime changes and characteristics over time which is offered by the HMMR and RHLP. In the face of climate changes, prediction within regimes clusters is more accurate than using all historical time series data. The autocorrelation varies between different clusters and collating different cluster characteristics for prediction makes biased hence cluster-based prediction can be realistic. Cluster inference can make climate shock preparedness more accurate and realistic than individual observation prediction (Huang et al., 2018; Table A1).

3.3 | HMMR and RHLP results

The Hidden Markov model Regression provided a natural technique for dealing with one of the fundamental problems of using stochastic modelling, as many naturally generated stochastic processes exhibited temporal heterogeneity that is driven by underlying (but unobservable) change in rainfall pattern. Figure 5 shows the original against predicted time series, prediction and filtering probabilities across the three agro-ecological regions of Southern Africa. The original rainfall patterns are shown in black whereas the predicted are represented by the red line. In the coastal region where the rainfall in some years reached over 1400 mm, the predicted rainfall over time was very close to the actual recorded rainfall which was above 600 mm and between 800 and 1400 mm in most years. Changes in rainfall pattern especially in the coastal region are quite clear but with a more fluctuating trend which makes it difficult to describe with less than 20 clusters. In sub-humid region, the rainfall pattern was clear from 1920 to 1950; however, more changes in pattern transpired between 1960 and 2000 with an unpredictable trend. The maximum rainfall received in the sub-humid areas does not exceed 1200 mm and reaching a minimum of 200 mm in some years. In semi-arid region there is no evidence of clear and predictable rainfall trend from 1940 to 2000 and the received rainfall in most years did not exceed 1000 mm with the minimum going down to 200 mm. Predicted probabilities indicated that in coastal region the chances

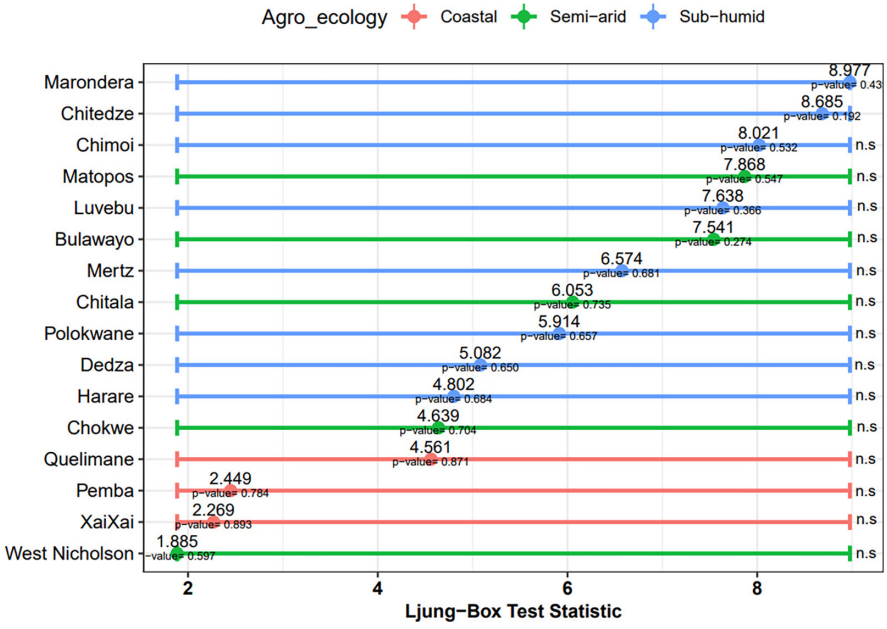

FIGURE 4 Ljung-Box Test Statistic investigating if residuals for Autoregressive Integrated Moving Average (ARIMA) model are independent. (n.s., not significant) (the statistic value is indicated in larger font and the corresponding *p*-values in smaller font).

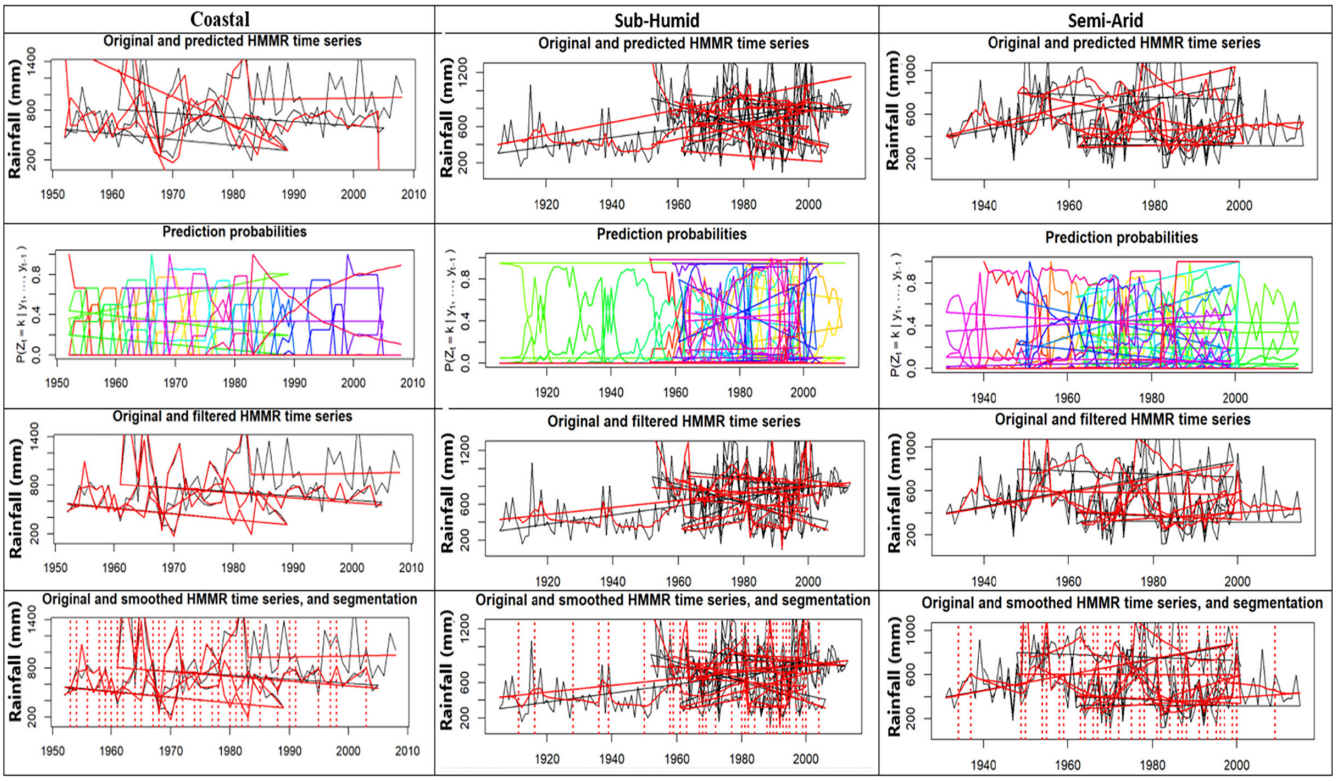

FIGURE 5 Original, predicted time series, prediction probabilities, filtered time series and filtering probabilities of the hidden Markov model regression (HMMR) under coastal, sub-humid and semi-arid regions of Southern Africa.

of receiving the same amount of rainfall are decreasing from 0.8 approaching zero over time across different clusters, while there is evidence of constant rainfall patterns in sub-humid region with an approximately constant probability close to 0.8. However, there is a significant decrease in the chances of receiving predictable rainfall patterns in semi-arid region as the probability decreased from 0.8 to approximately zero. Filtering probabilities smoothen the trend and indicate better predictability under the sub-humid compared to coastal and semi-arid conditions. The approach offered a better way of understanding the rainfall regime changes over time by segmenting the changes in different clusters. This helps to understand variability between and with years compared to the ability of commonly used approaches. As rainfall is highly variable especially in semi-arid regions as suggested by the results, robust approaches such HMMR are needed to predict and understand the rainfall changes (Fridman & Angeles, 2010). A clear understanding of these rainfall trends can help to reduce the effects of excessive precipitation which can degrade water quality, harming human health and ecosystems especially in coastal and sub-humid regions were the probability of receiving excessive rainfall is high (Malhi et al., 2020). Furthermore, in coastal and sub-humid areas storm water runoff is always experienced which often includes pollutants like heavy metals, pesticides, nitrogen and phosphorus, can end up in lakes, streams and bays, damaging aquatic ecosystems and lowering quality for human uses (Arnbjerg-Nielsen et al., 2013).

The regression with hidden logistic process estimated process probabilities across different rainfall regimes as well as the estimated model and segmentation (Figure 6, Table A1). Results of the model indicated the same exhibited trend by the HMMR model with the coastal region having clear pattern compared to sub-humid and semi-arid. Process probabilities suggested a decrease in maintaining the same rainfall pattern per cluster over time, as the probabilities continue to decrease from 0.8 to 0. The highest probabilities were observed in 1950–1970 compared to new and recent years (1990–2000) and this suggested that as the years continue to increase, a decrease in the rainfall predictability will continue to be witnessed in Southern Africa. The yearly rainfall segmentation process continues to be more complicated as years increase, especially in sub-humid regions after year 2000 more fluctuations were predicted because of changes in rainfall patterns. The rainfall pattern direction in semi-arid region promised to continue being less than 100 mm in most years but with unpredictable pattern within the maximum and minimum range. The results suggested a drastic change in rainfall patterns in Southern Africa making it more complicated for ordinary time series approach to extract the trend and predict the direction. This implies that adoption of new

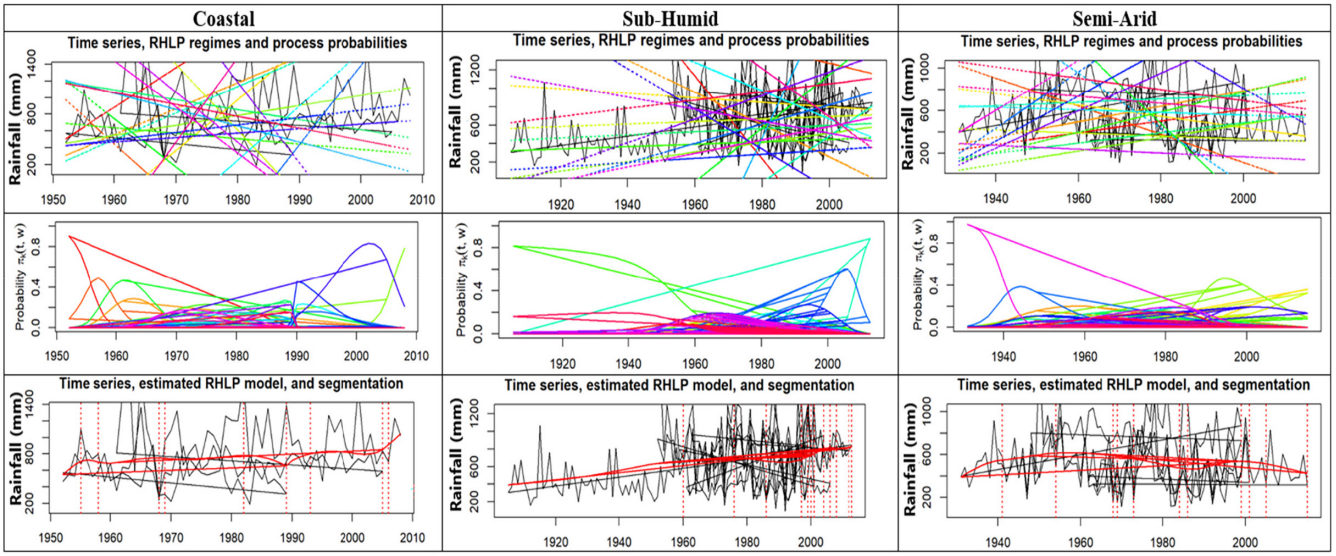

FIGURE 6 Rainfall regimes and process probabilities, estimated model and segmentation of the regression with hidden logistic process (RHLP) in coastal, sub-humid and semi-arid regions of Southern Africa.

agriculture practices is a necessity for the Southern African countries to remain food secure in the midst of the climate change as witnessed by rainfall changes (Adimassu & Kessler, 2016).

3.3.1 | HMMR and RHLP cluster observations and estimated rainfall variability

The higher the standard deviation the more variable or spread in measured rainfall data. In southern Africa, rainfall is highly variable which makes prediction a challenge. The maximum number of estimated cluster observations in sub-humid is 34 which is the highest compared to coastal and semi-arid regions with 8 and 32, respectively, for HMMR model (Figure 7). RHLP model estimated a maximum of 105 cluster observations in sub-humid, 74 in semi-arid and 39 in coastal. Comparing the two models the RHLP model estimated the highest number of cluster observations in all agro-ecologies compared with HMMR. The results also suggest that cluster variability is lower in RHLP model, which is less than 20,000 in all agro-ecologies, while HMMR estimated cluster variability surpasses 50,000 in some the clusters. In general, RHLP proved to be the best model to cluster rainfall variability accordingly and make some inferences in Southern Africa agro-ecologies for better informed climate change decisions. As reported by Ogallo (1979), rainfall trend analysis in Southern Africa showed that most of the annual series indicate some forms of oscillations rather than any particular trend. It is challenging to predict rainfall with a high degree of accuracy on long range or seasonal time scales over many global regions including Southern Africa (Landman & Beraki, 2010; Lyon & Mason, 2009). Southern Africa is subject to high inter-annual rainfall variability and the factors influencing this are generally understood but vary from location to location. They include the influence of oceans, both Indian and Atlantic, atmospheric air circulations, sea surface temperature variations, temperature differences between land and oceans, as well as local topographic features and their (Muktar et al., 2020). Rainfall variability has been linked with various sea surface temperature anomalies (SSTAs) in Southern Africa. On the other hand, increased pressure on natural water systems and artificial water storage systems because of a growing population make Southern Africa vulnerable to potential changes in the hydrological cycle because of global warming, which could lead to extremely negative impacts on societies within agro-ecologies. Studies on long-term changes and variability in rainfall and streamflow are therefore of immense interest in South Africa.

3.3.2 | HMMR and RHLP performance evaluation

The changes in log-likelihood as the number of clusters increase are shown in Figure 8. The statistic was used to evaluate the goodness of fit of the two suggested models HMMR and RHLP per each cluster across the three agro-ecological regions. In the coastal region the log-likelihood of both models slightly increased as the number of suggested clusters increases. As the clusters approached 20, the log-likelihood continued to indicate

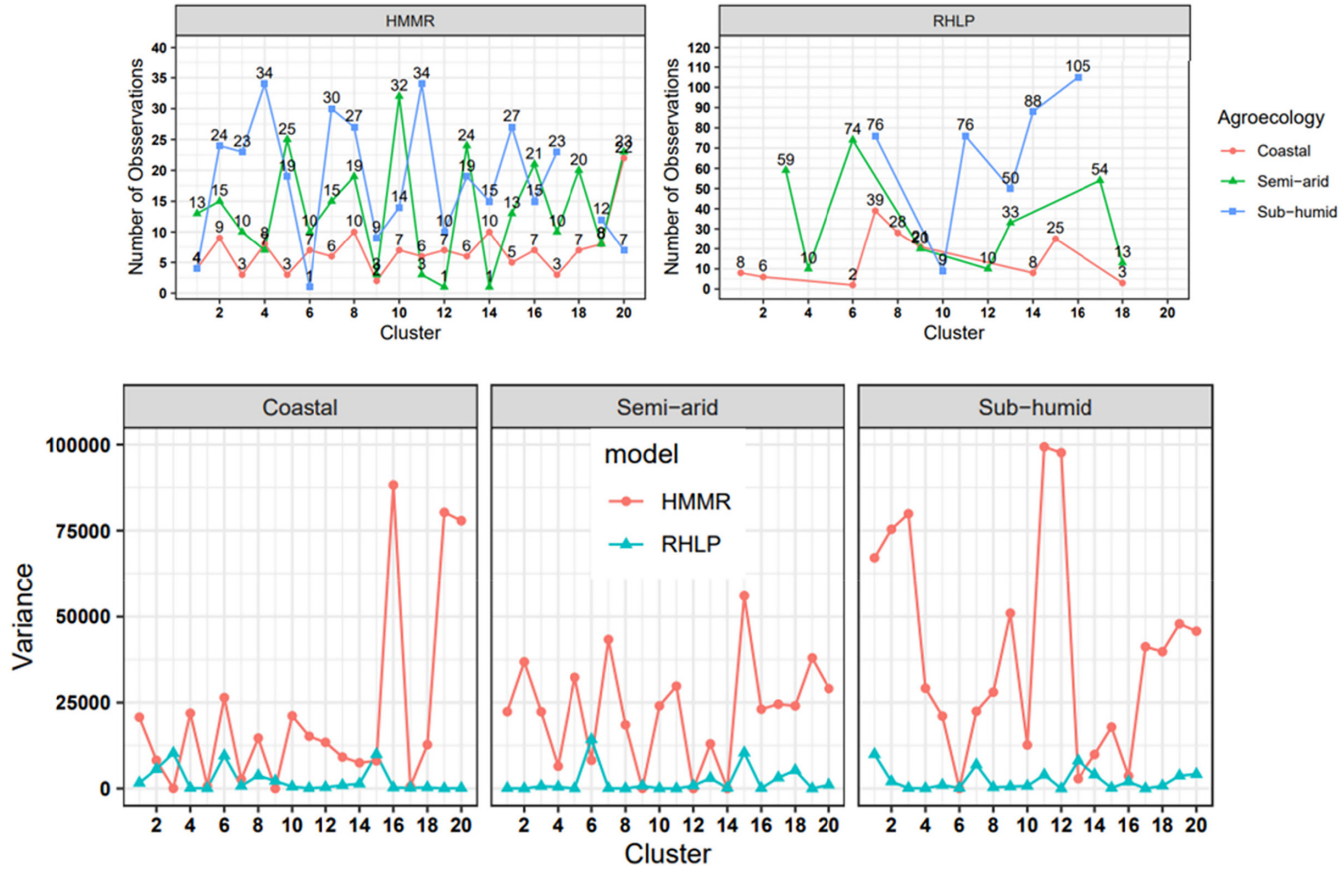

FIGURE 7 Number of observations and estimated variance in each cluster for both HMMR and RHLP models in coastal, semi-arid and sub-humid regions of Southern Africa.

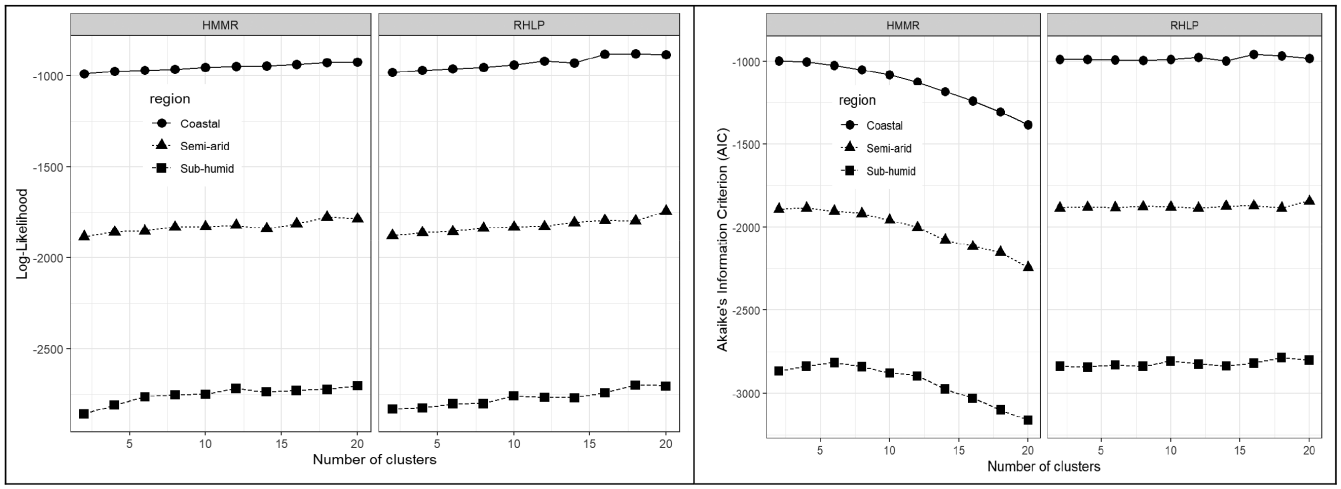

FIGURE 8 HMMR and RHLP Log-Likelihood and Akaike information criteria (AIC) statistics behaviour as the number of regimes (clusters) increases. HMMR, hidden Markov model regression; RHLP, regression with hidden logistic process.

a trend of increment, and this suggested that the rainfall patterns or regime changes in coastal regions cannot be fully explained by less than 20 clusters. In the semi-arid region, the change in log-likelihood was between −2000 and −1500 with a very slight increase in both models. Sub-humid region exhibited a slight increase trend below −2500 log-likelihood which was less than that of the coastal and semi-arid regions. The results suggested that the two suggested models best fit in coastal region rainfall changes, while in semi-arid region there was a decrease in goodness of fit and even worse under the sub-humid regions. In all the three main regions of

Southern Africa a lot is happening to rainfall patterns which makes it unpredictable, as only 20 clusters failed to fully explain the changes.

The AIC exhibited a different trend to that of the log-likelihood with an exponential decrease under the HMMR and a constant trend under the RHLP across all the three regions of Southern Africa as shown in Figure 8. The statistic evaluated the significant differences between the rainfall regimes over time. The HMMR model suggested that in the coastal region the AIC dropped in the negative direction (from −1000 to −1500), in semi-arid region it decreased from −2000 to −2500 and from −3000 to −3500 in the sub-humid region. The RHLP model indicated the unresponsiveness of the AIC as the number of clusters approached 20, with a slight increment in semi-arid and sub-humid regions. As a rule of thumb, the lower the AIC the better the model in explaining rainfall pattern at a given cluster or regime change. The AIC trend under the HMMR indicated that the goodness of the model deteriorates as the clusters continued to increase while the RHLP trend suggested a slight improvement as the number of clusters approached 20 and this indicated that the RHLP model has the potential to accurately trace the rainfall patterns especially in the sub-humid and semi-arid regions of Southern Africa. There is a significant difference in rainfall pattern in the three regions of Southern Africa. The models extracted patterns without being supervised, that is tracing the natural changes in climate leading to unpredictability of rainfall changes.

4 | CONCLUSION

In this study, we presented a combination of commonly used time series approaches (ARIMA and Prophet models) and modern unsupervised statistical approaches (HMMR and RHLP) for the joint segmentation of rainfall trends variability in Southern Africa. The application of the suggested approaches was based on the historical rainfall data measured in various locations of Southern Africa. The ARIMA and Facebook Prophet models indicated a significant increase of forecasted rainfall in sub-humid and coastal areas defined between 800 and 1000 mm. A declining rainfall trend was predicted in semi-arid region with variability between and within years. Similarities in forecast accuracy of the ARIMA and Facebook Prophet pave way for advanced approaches to investigate the rainfall variability between and within seasons. The suggested unsupervised models' regression with hidden logistic process (RHLP) and hidden markov model regression (HMMR) offered a unique clustering approach investigating the rainfall variability within seasons. Historical rainfall trends were segmented into a maximum of 20 clusters as going beyond this gives no further benefit implying cluster predictability especially in coastal areas. In semi-arid region the trend continues to drastically decline. Rainfall as the major indicator of climate change continue to be more complicated to predict because of high variability hence the need for more robust approaches such as unsupervised time series models which can categorize it into different clusters and prediction can be done at cluster level. Much impact of rainfall variability is mainly witnessed in agricultural activities and therefore adaptation to new agricultural practices is critical in Southern Africa and similar environments. It is very critical for meteorological service departments to give farmers accurate forecasted rainfall for seasonal planning purposes. Adoption of new practices such as climate smart agriculture technologies, and new hybrid crop varieties remains an important option for farmers to remain food secure in Southern Africa.

ACKNOWLEDGEMENTS

This study has been embedded into the CGIAR Research Programme MAIZE, Flagship Sustainable intensification of smallholder farming systems. We acknowledge the CGIAR Fund Council and other donors for funding to the CGIAR Research Programme MAIZE. We thank the meteorological stations from the four countries for their contribution in generating rainfall data used in the study.

OPEN RESEARCH BADGES

This article has earned Open Data and Preregistered Research Designs badges. Data and the preregistered design and analysis plan are available at https://zenodo.org/record/8412628 or on request from the corresponding author at lovemore.datascience@gmail.com

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author, lovemore.datascience@gmail.com. The data are not publicly available due to meteorological service departments restrictions and policies.

ORCID

Lovemore Chipindu https://orcid.org/0000-0001-8023-1105

REFERENCES

Adhikari, U., Nejadhashemi, A.P. & Woznicki, S.A. (2015) Climate change and eastern Africa: A review of impact on major crops. *Food and Energy Security*, 4(2), 110–132. Available from: https://doi.org/10.1002/fes3.61

Adimassu, Z. & Kessler, A. (2016) 'Factors affecting farmers' coping and adaptation strategies to perceived trends of declining rainfall and crop productivity in the central rift valley of Ethiopia. *Environmental Systems Research*, 5(1), 1–16. Available from: https://doi.org/10.1186/s40068-016-0065-2

Akaike, H. (2011) Akaike's information criterion. In International Encyclopedia of Statistical Science. Available from: https://doi.org/10.1007/978-3-642-04898-2_110

Akdag, M. & Bozma, G. (2021) STOK AKI Ş MODEL İ VE FACEBOOK PROPHET ALGOR İ TMASI İ LE B İ TCO İ N F İ YATI TAHM İ N İ PREDICTION OF BITCOIN PRICE WITH STOCK TO FLOW. (March).

Arnbjerg-Nielsen, K., Willems, P., Olsson, J., Beecham, S., Pathirana, A., Bülow Gregersen, I. et al. (2013) Impacts of climate change on rainfall extremes and urban drainage systems: a review. *Water Science and Technology*, 68(1), 16–28. Available from: https://doi.org/10.2166/wst.2013.251

Asfaw, A., Simane, B., Hassen, A. & Bantider, A. (2018) Variability and time series trend analysis of rainfall and temperature in northcentral Ethiopia: a case study in Woleka sub-basin. *Weather and Climate Extremes*, 19, 29–41. Available from: https://doi.org/10.1016/j.wace.2017.12.002

Atiqul Haq, S.M., Islam, M.N., Siddhanta, A., Ahmed, K.J. & Chowdhury, M.T.A. (2021) Public perceptions of urban green spaces: Convergences and divergences. *Frontiers in Sustainable Cities*, 3, 755313. Available from: https://doi.org/10.3389/frsc.2021.755313

Babaousmail, H., Hou, R., Ayugi, B., Ojara, M., Ngoma, H., Karim, R. et al. (2021) Evaluation of the performance of cmip6 models in reproducing rainfall patterns over North Africa. *Atmosphere*, 12(4), 1–25. Available from: https://doi.org/10.3390/atmos12040475

Burns, P.J. (2002) Robustness of the Ljung-Box Test and its Rank Equivalent. Available from: https://doi.org/10.2139/ssrn.443560

Chamroukhi, F., Allou Samé, Patrice, Aknin, Gérard Govaert (2011) Model-based clustering with hidden Markov model regression for time series with regime changes. Proceedings of the International Joint Conference on Neural Networks. pp. 2814–2821. Available from: https://doi.org/10.1109/IJCNN.2011.6033590.

Chamroukhi, F., Samé, A., Govaert, G. & Aknin, P. (2009) A regression model with a hidden logistic process for feature extraction from time series. Proceedings of the International Joint Conference on Neural Networks, June. pp. 489–496. https://doi.org/10.1109/IJCNN.2009.5178921.

Chamroukhi, F., Trabelsi, D., Mohammed, S., Oukhellou, L. & Amirat, Y. (2013) Joint segmentation of multivariate time series with hidden process regression for human activity recognition. *Neurocomputing*, 120, 633–644.

Di Luca, A., Pitman, A.J. & de Elía, R. (2020) Decomposing temperature extremes errors in CMIP5 and CMIP6 models. *Geophysical Research Letters*, 47(14), 1–10. Available from: https://doi.org/10.1029/2020GL088031

Fridman, M. and Angeles, L. (2010) Hidden Markov Model. pp. 177–194. Available from: https://doi.org/10.1142/9789814287319_0012.

Guhathakurta, P. & Rajeevan, M. (2008) Trends in the rainfall pattern over India. *International Journal of Climatology*, 28(11), 1453–1469. Available from: https://doi.org/10.1002/joc.1640

Gupta, K. (2011) Hidden Markov Model. p. 1357. Available from: https://doi.org/10.1145/1980022.1980326.

Hossain, M.M., Anwar, A.H.M.F., Garg, N., Prakash, M. & Bari, M. (2021) Monthly Rainfall Prediction for Decadal Timescale using Facebook Prophet at a Catchment Level. (September).

Huang, M., Ji, Q. & Yao, W. (2018) Semiparametric hidden Markov model with non-parametric regression. *Communications in Statistics – Theory and Methods*, 47(21), 5196–5204. Available from: https://doi.org/10.1080/03610926.2017.1388398

Khayyat, M., Laabidi, K., Almalki, N. & al-zahrani, M. (2021) Time series Facebook Prophet model and python for COVID-19 outbreak prediction. *Computers, Materials and Continua*, 67(3), 3781–3793. Available from: https://doi.org/10.32604/cmc.2021.014918

Kim, E., Ha, J., Jeon, Y. & Lee, S. (2004) Ljung-box test in unit root AR-ARCH model. *Communications for Statistical Applications and Methods*, 11(2), 323–327. Available from: https://doi.org/10.5351/ckss.2004.11.2.323

Kyei-Mensah, C., Kyerematen, R. & Adu-Acheampong, S. (2019) Impact of rainfall variability on crop production within the Worobong ecological area of Fanteakwa District, Ghana. *Advances in Agriculture*, 2019, 1–7. Available from: https://doi.org/10.1155/2019/7930127

Lal, R. & Bhat, U.N. (1988) Reduced system algorithms for Markov chains. *Management Science*, 34, 1202–1220. Available from: https://doi.org/10.1287/mnsc.34.10.1202

Landman, W.A. & Beraki, A. (2010) Multi-model forecast skill for mid-summer rainfall over Southern Africa. *International Journal of Climatology*, 32, 303–314. Available from: https://doi.org/10.1002/joc.2273

Landwehr, N., Hall, M. & Frank, E. (2005) Logistic model trees. *Machine Learning*, 59, 161–205. Available from: https://doi.org/10.1007/s10994-005-0466-3

Lei, Q. & Sornette, D. (2023) A stochastic dynamical model of slope creep and failure. *Geophysical Research Letters*, 50(11), 1–11. Available from: https://doi.org/10.1029/2022GL102587

Lounis, M. (2021) Predicting active, death and recovery rates of COVID-19 in Al- geria using Facebook ' Prophet model. Predicting active, death and recovery rates of COVID-19 in Algeria using Facebook' Prophet model, (March). Available from: https://doi.org/10.20944/preprints202103.0019.v1

Luo, J., Zhang, Z., Fu, Y. & Rao, F. (2021) Time series prediction of COVID-19 transmission in America using LSTM and XGBoost algorithms. *Results in Physics*, 27, 104462. Available from: https://doi.org/10.1016/j.rinp.2021.104462

Lyon, B. & Mason, S.J. (2009) The 1997/98 summer rainfall season in southern Africa. Part II: Model simulations and coupled model forecasts. *Journal of Climate*, 22(13), 3802–3818. Available from: https://doi.org/10.1175/2009JCLI2600

Malhi, Y., Franklin, J., Seddon, N., Solan, M., Turner, M.G., Field, C.B. et al. (2020) Climate change and ecosystems: threats, opportunities and solutions. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, 375(1794), 20190104. Available from: https://doi.org/10.1098/rstb.2019.0104

Mkuhlani, S., Mupangwa, W. & Nyagumbo, I. (2018) Maize yields in varying rainfall regimes and cropping systems across Southern Africa: A modelling assessment. In: *University initiatives in climate change mitigation and adaptation*. Berlin: Springer International Publishing, pp. 203–228. Available from: https://doi.org/10.1007/978-3-319-89590-1_12

Mlenga, D.H. (2016) Factors Influencing Adoption of Conservation Agriculture: A Case for Increasing Resilience to Climate Change and Variability in Swaziland Factors Influencing Adoption of Conservation Agriculture: A Case for Increasing Resilience to Climate Change and Va. (January).

Mohan, K. & Fazel, M. (2010) Iterative reweighted least squares for matrix rank minimization. In 2010 48th annual Allerton conference on communication, control, and computing, Allerton 2010. Available from: https://doi.org/10.1109/ALLERTON.2010.5706969

Muktar, A., Elekwachi, W. & Hycienth, N. (2020) Rainfall change detection in Africa using remote sensing and Gis between 1999–2018. *Big Data In Water Resources Engineering (BDWRE)*, 1(2), 52–54. Available from: https://doi.org/10.26480/bdwre.02.2020.52.54

Nyagumbo, I., Mupangwa, W., Chipindu, L., Rusinamhodzi, L. & Craufurd, P. (2020) A regional synthesis of seven-year maize yield responses to conservation agriculture technologies in eastern and southern Africa. *Agriculture, Ecosystems and Environment*, 295, 106898. Available from: https://doi.org/10.1016/j.agee.2020.106898

Ogallo, L. (1979) Rainfall variability in Africa. *Monthly Weather Review*, 107(9), 1128–1132. Available from: https://doi.org/10.1175/1520-0493(1979)107<1133:rvia>2.0.co;2

Prasad, P.V.V. & Staggenborg, S.A. (2011) Growth and production of sorghum and millets. *Soils, Plant Growth and Crop Production*, 2, 1–27.

Precipitation Measurement Missions. (2020) *Climate change: trends and patterns*. Washington, DC: National Aeronautics and Space Administration (NASA). Available from: https://pmm.nasa.gov/science/climate-change

Shah, N.V., Patel, Y.S. & Bhangaonkar, P.D. (2021) Assessing impact of climate change on rainfall patterns of Vadodara District, Gujarat, India. *Journal of Physics: Conference Series*, 1714(1), 12046. Available from: https://doi.org/10.1088/1742-6596/1714/1/012046

Shao, W., Radke, L. F. and Sivrikaya, F. (2021) Adaptive Online Learning for the Autoregressive Integrated Moving Average Models. pp. 1–30.

Shimodaira, H. (2000) Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90, 227–244. Available from: https://doi.org/10.1016/s0378-3758(00)00115-4

Singh, P. (2018) Indian summer monsoon rainfall (ISMR) forecasting using time series data: a fuzzy-entropy-neuro based expert system. *Geoscience Frontiers*, 9(4), 1243–1257. Available from: https://doi.org/10.1016/j.gsf.2017.07.011

Strauch, A.M., MacKenzie, R.A., Giardina, C.P. & Bruland, G.L. (2018) Influence of declining mean annual rainfall on the behavior and yield of sediment and particulate organic carbon from tropical watersheds. *Geomorphology*, 306, 28–39. Available from: https://doi.org/10.1016/j.geomorph.2017.12.030

Su, Y., Gabrielle, B., Beillouin, D. & Makowski, D. (2021) High probability of yield gain through conservation agriculture in dry regions for major staple crops. *Scientific Reports*, 11(1), 1–9. Available from: https://doi.org/10.1038/s41598-021-82375-1

Twenefour, B.K.F., Techie Quaicoe, M. & Baah, E. (2018) Analysis of rainfall pattern in the Western Region of Ghana. *Asian Journal of Probability and Statistics*, 1, 1–12. Available from: https://doi.org/10.9734/ajpas/2018/v1i324538

Wall, P.C., Thierfelder, C., Ngwira, A., Govaerts, B., Nyagumbo, I. & Baudron, F. (2013) Conservation agriculture in eastern and southern africa. In: Jat, R.A., Sahrawat, K.L. & Kassam, A.H. (Eds.) *Conservation agriculture: Global prospects and challenges*. Wallingford, Oxfordshire: CABI, pp. 263–292. Available from: https://doi.org/10.1079/9781780642598.0263

Wang, H., Liu, L., Qian, Z.(.S.)., Wei, H. & Dong, S. (2014) Empirical mode decomposition-autoregressive integrated moving average: hybrid short-term traffic speed prediction model. *Transportation Research Record*, 2460(1), 66–76. Available from: https://doi.org/10.3141/2460-08

WFP (2016) WFPEl Niño situation report1. pp. 1–4. Available from: http://documents.wfp.org/stellent/groups/public/documents/ep/wfp281523.pdf

Wimhurst, J.J. & Greene, J.S. (2021) Updated analysis of gauge-based rainfall patterns over the western tropical Pacific Ocean. *Weather and Climate Extremes*, 32, 100319. Available from: https://doi.org/10.1016/j.wace.2021.100319

Zhang, J., Shang, R., Rittenhouse, C., Witharana, C. & Zhu, Z. (2021) Evaluating the impacts of models, data density and irregularity on reconstructing and forecasting dense Landsat time series. *Science of Remote Sensing*, 4, 100023. Available from: https://doi.org/10.1016/j.srs.2021.100023

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Chipindu, L., Mupangwa, W., Nyagumbo, I. & Zaman-Allah, M. (2023) Unsupervised segmentation and clustering time series approach to Southern Africa rainfall regime changes. *Geoscience Data Journal*, 00, 1–17. Available from: https://doi.org/10.1002/gdj3.228

APPENDIX A

TABLE A1 Regression with hidden logistic process (RHLP) model coefficients in Coastal, sub-humid and semi-arid regions.

	Coastal				Sub-humid				Semi-arid			
Clusters	**Constant**	**X1**	**X2**	**X3**	**Constant**	**X1**	**X2**	**X3**	**Constant**	**X1**	**X2**	**X3**
Beta ($K=1$)	−8.43E+04	8.81E+00	1.44E−02	1.88E−06	−2.37E+05	24.9489	0.040413	5.31E−06	2.30E+04	4.06E+01	−2.15E−02	−2.95E−06
Beta ($K=2$)	−2.25E+05	−1.08E+02	−2.04E−03	5.96E−05	3.54E+07	−53,552.8	27.01569	−0.00454	−1.47E+07	2.25E+04	−1.14E+01	1.94E−03
Beta ($K=3$)	−1.04E+05	4.01E+02	−3.25E−01	7.55E−05	3.69E+03	−5.31963	0.001207	2.86E−07	−5.30E+03	4.55E+00	−6.35E−04	−1.72E−07
Beta ($K=4$)	2.00E+04	−1.06E+01	−5.46E−04	−2.04E−10	1.24E+07	−19,025.8	9.762869	−0.00167	3.26E+07	−1.57E+05	1.35E+02	−3.23E−02
Beta ($K=5$)	−5.16E+05	7.28E+01	1.17E−01	−9.37E−06	1737.6	−1.07842	−1.6E−05	4.37E−07	6.32E+08	−9.68E+05	4.94E+02	−8.40E−02
Beta ($K=6$)	−8.39E+05	−9.44E+01	1.47E−01	6.05E−05	198,580.8	−17.5692	0.10774	−7.6E−05	−5.56E+05	1.81E+02	−8.01E−03	3.03E−05
Beta ($K=7$)	−2.31E+04	2.34E+02	−8.02E−02	−1.66E−05	−2.93E+09	4,430,515	−2234.6	0.37568	−3.10E+05	−6.02E+01	3.87E−03	5.24E−05
Beta ($K=8$)	2.56E+06	−2.35E+03	4.49E−02	2.50E−04	−1.08E+04	395.6592	−0.06329	−6.6E−05	8.50E+05	−7.57E+02	1.55E−01	5.58E−06
Beta ($K=9$)	−3.41E+04	−1.65E+02	7.16E−02	1.02E−05	−1.90E+05	−111.263	0.038178	3.28E−05	6.72E+05	1.80E+03	−2.11E+00	5.19E−04
Beta ($K=10$)	−7.24E+04	−2.87E−02	−1.09E−02	1.57E−05	6.86E+05	178.8255	−0.07791	−9.1E−05	3.86E+05	−2.39E+02	7.15E−02	−2.43E−05
Beta ($K=11$)	4.90E+05	−1.18E+02	−9.89E−02	1.74E−05	3.94E+04	−1.74834	0.010799	−1E−05	−5.23E+05	2.62E+02	1.74E−01	−9.01E−05
Beta ($K=12$)	3.56E+05	−4.65E+02	8.29E−02	3.08E−05	−30,232.8	26.21895	−0.01288	3.3E−06	−5.38E+05	4.91E+02	−1.64E−01	3.07E−05
Beta ($K=13$)	−6.74E+04	−4.37E+02	2.15E−01	1.11E−05	42,129,760	−64,035.1	32.44127	−0.00548	2.88E+06	−2.69E+03	4.90E−01	6.74E−05
Beta ($K=14$)	6.02E+03	3.84E+00	−5.37E−04	−1.31E−06	−23,377.4	24.20527	−0.00489	−9.8E−07	1.12E+05	2.42E+02	−1.91E−01	2.02E−05
Beta ($K=15$)	2.58E+06	−8.44E+02	−5.46E−01	1.59E−04	7.36E+07	−111,622	56.43883	−0.00951	−8.10E+04	2.43E+02	3.62E−02	−6.89E−05
Beta ($K=16$)	−4.79E+09	7.22E+06	−3.63E+03	6.08E−01	2.36E+05	495.132	−0.39017	4.11E−05	−3.92E+05	8.15E+01	2.96E−02	1.40E−05
Beta ($K=17$)	1.57E+06	−4.08E+02	−3.50E−01	8.00E−05	−6.56E+04	194.7008	0.028844	−5.5E−05	9.35E+05	−1.03E+03	2.20E−01	3.64E−05
Beta ($K=18$)	6.91E+04	1.84E+02	−5.32E−02	−2.80E−05	−7.68E+04	15.75576	0.005838	2.8E−06	−3.23E+03	7.93E−02	2.03E−04	5.23E−07
Beta ($K=19$)	1.21E+05	1.30E+01	−3.68E−02	1.68E−07	1,367,539	−1542.92	0.331647	5.51E−05	1.33E+05	1.03E+01	−2.02E−02	−1.06E−05
Beta ($K=20$)	−3.51E+05	−6.87E+02	1.05E+00	−3.15E−04	−385,975.8	619.8269	−0.10796	−5.3E−05	−2.44E+07	3.72E+04	−1.89E+01	3.21E−03