Scientific African 29 (2025) e02924

Contents lists available at ScienceDirect

Scientific African

journal homepage: www.elsevier.com/locate/sciaf

Predicting precipitation using dynamic distributed lag models in arid and sub-humid regions of South Africa

Lyson Chaka [a,*], Mohamed A.M. Abd Elbasit [b], Simbarashe Jombo [b,c]

[a] Department of Statistics, University of Pretoria, Pretoria, South Africa
[b] Arid Region Water Research Centre, Sol Plaatje University, Kimberley, South Africa
[c] Department of Physical and Earth Sciences, Sol Plaatje University, Kimberley, South Africa

A R T I C L E I N F O

Edited by Editor: DR B Gyampoh

Keywords:
Precipitation
Transfer functions
ARIMAX models
Dynamic linear regression models
Anomalies
Time series forecasting

A B S T R A C T

Ocean characteristics have contributed to a series of unusual rainfall patterns and floods, leading to severe land degradation, loss of life and infrastructure in various regions. Modelling and prediction of precipitation using in-situ data and oceanographic variables is possible. There are limited studies to substantiate this approach in less-developed countries. This study aims to model and predict precipitation in the arid, semi-arid and sub-humid regions of South Africa using dynamic linear regression (DLR) models, with sea surface temperature (SST) anomalies, evaporation-precipitation differences, longwave radiation (*lwRad*), net surface heat flux and relative humidity as input variables. The prediction accuracy of the autoregressive integrated moving average model with extra data (ARIMAX) and dynamic distributed lag (DDL) models was compared on the mean monthly rainfall data for the period 2008 to 2022. The results highlighted that the DDL models predict better than the other ARIMAX models, with SST anomalies and *lwRad* having a significant contribution (p-values < 0.05). These models had the smallest root mean squared error (RMSE) values for the arid (8.27 mm), semi-arid (19.15 mm) and the sub-humid (26.77 mm) regions, indicating that DDL models are suitable tools for the prediction of precipitation in these regions. However, additional oceanographic predictors such as sea surface salinity, ocean heat content, and upper-ocean current patterns may further enhance precipitation prediction accuracy, particularly in regions with strong ocean-atmosphere coupling, such as coastal or monsoon-influenced areas.

Introduction

Precipitation is an important factor that supports livelihoods through agricultural productivity in arid and sub-humid regions [1]. Crops and livestock depend on regular rainfall to grow and thrive [2]. However, excessive precipitation can have negative impacts, such as flooding and landslides [3,4], damage to homes, transportation routes and other structures. In recent years, the coastal regions of Southern Africa have experienced a series of extreme weather events in the form of heavy rainfall, flooding, and cyclones that have caused loss of life and significant damage to infrastructure and natural landscapes [5]. The effects of persistent extreme precipitation, changes in rainfall patterns, and their impact on climate change have become alarming concerns in most parts of the world [6]. This triggered more research interest in determining the changes in hydrologic and climatic time series [7].

* Corresponding author.
E-mail addresses: lyson.chaka@up.ac.za (L. Chaka), mohamed.ahmed@spu.ac.za (M.A.M. Abd Elbasit), simba.jombo@spu.ac.za (S. Jombo).

https://doi.org/10.1016/j.sciaf.2025.e02924
Received 16 February 2025; Received in revised form 17 August 2025; Accepted 26 August 2025
Available online 27 August 2025

In recent years, the prediction of monthly and seasonal rainfall has relied on three main methodological approaches. These approaches are dynamical models, statistical models, and hybrid dynamical-statistical frameworks [8]. Dynamical models, such as coupled ocean-atmosphere systems, simulate the physical processes governing climate variability, while statistical models establish empirical relationships between historical climate patterns and future conditions [9]. Hybrid models attempt to integrate the strengths of both the ocean-atmosphere systems and the statistical models for improved forecast skill [10]. More recently, machine learning has emerged as a powerful tool in climate prediction due to its ability to handle nonlinear relationships and high-dimensional data, offering promising improvements in precipitation forecasting accuracy [11]. Despite these advancements, the skill of current prediction systems in capturing South African precipitation remains limited, particularly on monthly timescales, due to the region's complex climatic drivers and high variability [12,13]. This study aims to contribute to addressing this gap by exploring machine learning techniques for monthly rainfall prediction in South Africa.

Accurate precipitation predictions are essential for planning and decision-making related to agriculture (e.g. [14]), hydrology (e.g. [1]), spatial planning (e.g. [15]), natural disaster risk reduction (e.g. [5]), and the development of climate models (e.g. [3,4]). Sophisticated weather models incorporate data from radar, satellites, and ground-based instruments to produce forecasts that are much more accurate than those made in the past [16]. Initially, weather forecasting was costly and relied on knowledge of weather patterns, observations of the sky and the surrounding environment to make predictions [17]. However, the use of traditional methods with precipitation data alone might be unreliable due to several reasons, including data limitations and algorithmic inconsistencies [18].

Numerous studies apply machine learning (ML) and statistical time series analysis methods to predict rainfall. These models include the antileakage least-squares spectral analysis (ALLSA) [19], vector autoregressive model [20], multi-layer perceptron (MLP) neural network (NN) [21] and the autoregressive integrated moving average (ARIMA) models [22]. The findings in these studies provide significant insights into trend-season estimates, monthly and regional precipitation forecasts. For example, the study by Ghaderpour et al. [19] compared the ALLSA method against ML and time series forecasting methods for monthly precipitation. The findings indicated that the ALLSA method outperformed the ML and time series methods. While various machine learning and statistical models have been used in rainfall forecasting, this study adopts a Dynamic Distributed Lag (DDL) framework due to its proven utility in modelling lagged climatic effects and maintaining interpretability in regional hydrological applications.

In studies involving climatic data, the challenge of missing data is a common feature [3], and the Prediction and Research Moored Array in the Tropical Atlantic (PIRATA) and Research Moored Array for African-Asian- Australian Monsoon Analysis and Prediction (RAMA) meteorological and oceanic data are no exceptions [23,24]. However, this challenge can be mitigated by applying approaches such as imputation [25], the use of robust statistical models for missing data in regression analysis [26], and the Bayesian Dynamic Linear (BDL) regression model that considers regression coefficient relationships as dynamic and capable of changing over time.

Climate change has an adverse impact on the arid and semi-arid regions of South Africa [27] that are characterised by low rainfall, high temperatures, and sparse vegetation [28]. The impacts include an increase in temperatures and rainfall variability [1]. The uncertainty regarding the expected amount of precipitation is a persistent challenge for farmers, policymakers, and planners [29], making it difficult for these groups to effectively plan and prepare for future agricultural activities. Therefore, rainfall prediction is crucial for the generation of information needed for future planning and the sustainability of agricultural systems in these regions. This study used PIRATA and RAMA observations to generate precipitation prediction models for the arid, semi-arid and sub-humid regions of South Africa. The specific objectives of the study were to (i) fit the dynamic linear regression (DLR) model for precipitation (mm/month) in each of the three regions based on sea surface temperature anomalies (*anom*), evaporation minus precipitation (*emp*), longwave radiation (*lwRad*), net surface heat flux (*qnet*) and relative humidity (*rh*), and (ii) compare the prediction accuracy of the

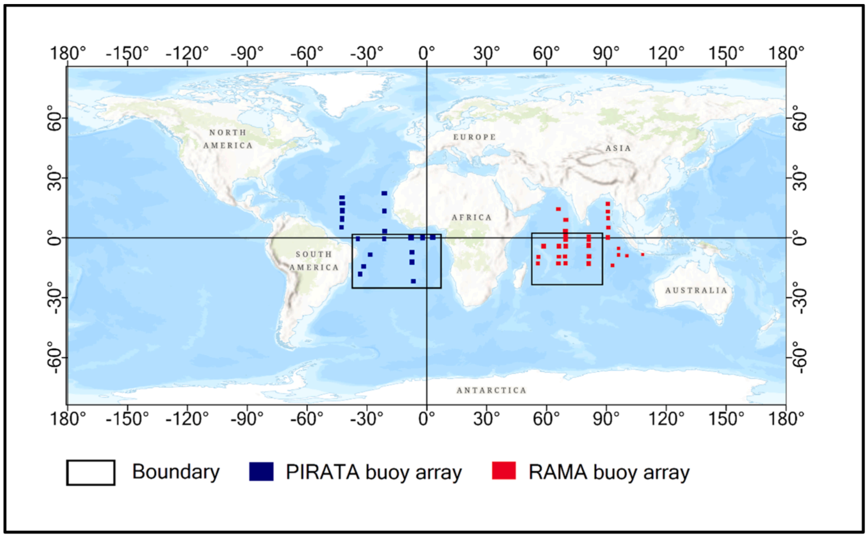

Fig. 1. The Prediction and Research Moored Array in the Tropical Atlantic (PIRATA) and Research Moored Array for African-Asian-Australian Monsoon Analysis and Prediction (RAMA) locations.

fitted models when predicting monthly average precipitation in these regions.

Methodology

The data generated from the oceanographic multi-satellite weather stations is useful for the prediction of inland rainfall, which directly affects agricultural and other economic activities [30]. Once established, precipitation forecasts can be generated for the interpretation of rainfall patterns and their effects in the targeted regions.

Data

The use of PIRATA and RAMA meteorological and oceanographic moored buoy networks provides continuous, basin-scale time series data in both the Atlantic and Indian Oceans for climate study and forecasting [31]. Oceanographic factors such as *anom, emp, lwRad, qnet* and *rh* from the PIRATA and RAMA stations (see location plots on Fig. 1) were explored for the prediction of precipitation. These factors were selected for their relevance to ocean-atmosphere interactions and climate variability, as they influence surface temperature, convection, moisture balance and energy exchange [32,33]. This is attributed to their ability to affect the Earth's climate and weather patterns. A study by Fontaine et al. [32] established that there is a synchronous relationship between rainfall anomalies over some parts of North Africa and the thermal contrasts (sea surface temperature differences) between the Mediterranean Sea and the Indian Ocean. At this moment, climate services to forecast precipitation using statistical methods in arid and semi-arid regions of South Africa are still limited.

Previous studies have used the PIRATA and RAMA global moored buoys' observations to validate the accuracy of wind data [33, 34], compare with satellite-derived precipitation datasets [35] and validate Sea surface salinity (SSS) measurements retrieved from SMAP radiometer [36]. Similarly, studies have used observations from PIRATA and RAMA global moored buoys to monitor and forecast precipitation. For example, the study by Prakash et al. [37], which used multi-satellite data from the PIRATA and RAMA ocean buoys between 1998 and 2010, indicated that the ocean buoys offer precise measurements of precipitation, enhanced by 5 to 9 % over oceanic and mountainous regions. Similarly, Bowman et al. [38] compared various satellite-based algorithms used for estimating rain rate with data from ocean buoy rain gauges, including the PIRATA buoy network, to validate the accuracy of the estimations. The results from the study indicated that there is little difference between the biases of the satellite rain algorithms with the data obtained from the PIRATA stations. Wu & Wang [35] used the PIRATA and RAMA data to evaluate the accuracy of the satellite precipitation data. Additionally, the study took note of the biases of the various observational systems used. The results highlighted the differences between the satellite datasets and data from the ocean buoys. Despite the successes in showing differences between data from the ocean buoys and satellite data, it is unclear whether data from PIRATA and RAMA can be used to predict precipitation in arid and sub-humid regions.

The oceanographic and surface meteorological data collected from 2008 to 2022 at some of the PIRATA and the RAMA locations, operated by the Tropical Atmosphere Ocean (TAO) Project Office of the National Oceanic and Atmospheric Administration (NOAA/ PMEL) in America, provide the oceanographic variables. The datasets were publicly available from the NOAA Pacific Marine Environmental Laboratory (PMEL) at https://www.pmel.noaa.gov/tao/drupal/disdel/. The datasets were obtained from mooring stations in the tropical oceans south of the equator, primarily between 0°S to 15°S and spanning longitudes of 10°W to 35°W (PIRATA, Atlantic Ocean) and 55°E to 80°E (RAMA, Indian Ocean). These datasets were investigated for their influence on the precipitation variability in three identified regions of South Africa. Fig. 1 shows the PIRATA and RAMA locations used in this study.

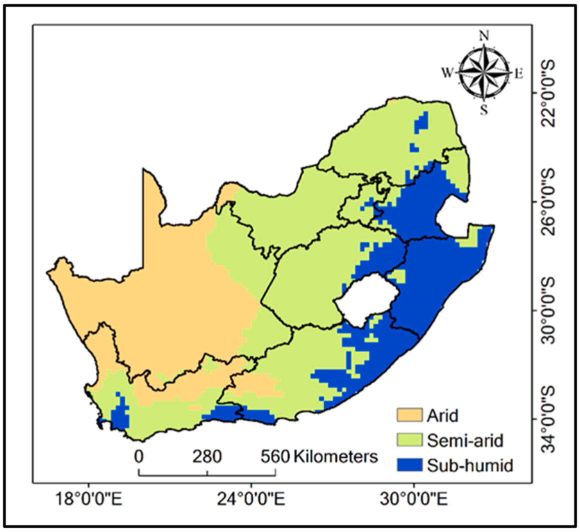

Fig. 2. Aridity zones of South Africa.

The oceanographic variables considered for this study include the Atlantic and Indian oceans sea surface temperature anomalies, the evaporation-precipitation differences, *lwRad*, net surface heat flux and the relative humidity. The rainfall (mm/month) data were retrieved from a long-term satellite climatic project using the Tropical Rainfall Measuring Mission (TRMM), for the period 2008 to 2022. The data were averaged at the regional level for each climatic zone - arid, semi-arid, and sub-humid, using a spatial resolution of 25 km × 25 km. Models were calibrated separately for each zone to account for climatic variability. Fig. 2 shows the arid, semi-arid and sub-humid regions of South Africa that receive different amounts of annual precipitation.

Analytical methods

To determine which variables to include in the models, we applied two diagnostic procedures to justify the inclusion of *anom, emp, lwRad, qnet* and *rh* as predictors in the DDL models. First, multicollinearity was assessed using the Variance Inflation Factor (VIF), with all values below 2.2, indicating acceptable levels. Second, Granger causality tests (lag = 6) were performed to evaluate the temporal predictive relevance of each predictor for three precipitation regimes (arid, semi-arid and sub-humid precipitation) over time. The results showed that all predictors significantly Granger-caused each dependent variable (all p-values < 0.05), confirming that they contribute valuable forecasting information. Together, these tests provide statistical justification for the inclusion of the chosen predictors in the subsequent forecasting models.

To explore the influence of oceanographic variables (*anom, emp, lwRad, qnet*, and *rh*) on inland precipitation, DLR models were used. We selected distributed dynamic linear regression (DDLR) to model monthly regional rainfall due to its balance of interpretability, computational efficiency and ability to account for both autocorrelation and delayed effects of predictors. DDL models are particularly suited to hydrological and climate time series, where environmental drivers exert influence over varying temporal lags. This approach also supports transparent interpretation of lag-specific relationships, aiding scientific communication and policy relevance. Unlike complex nonlinear and other black-box models such as long short-term memory (LSTM) networks or support vector regression (SVR), DDL provides a transparent framework in which the time-varying effects of predictors can be directly interpreted. This feature is especially valuable in climate data, where understanding the short-term and long-range dependence features of the errors, evolving relationships between covariates, and the target variable, is essential [39]. Moreover, DDL's linear structure allows for efficient computation, significantly lower computational overhead and simpler tuning and straightforward distribution across multiple nodes, making it well-suited to our oceanographic data from different environments [39,40]. The proposed models form part of the resources needed for water management and other economic activities that rely on weather forecasts.

Prediction models

We considered DLR models that predict regional precipitation using the oceanographic factors as exogenous variables. The autoregressive integrated moving average (ARIMA) model with extra data (X), commonly known as the ARIMAX model, is a special case of the dynamic linear regression models, which is a class of linear regression models that essentially generate better forecasts than the traditional regression models. The ARIMAX model is expressed as a combination of a structural equation and a disturbance time series model given as:

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \eta_t, \tag{1}$$

$$\mathit{\Phi}(B)\Phi(B)\Delta^d\Delta_S^D\eta_t = \theta(B)\Theta(B)a_t, \tag{2}$$

where model (1) is a structural equation that predicts the response, y_t, using the current set of predictors, $\boldsymbol{X}_t$, and the noise process, η_t, which describes the unexplained part of y_t; and model (2) is a noise (or disturbance) model with white-noise error terms, i.e., $a_t \sim i.i.d.\ N(0,\ \sigma^2)$. The ARIMAX model fuses a linear regression model (1) of order k with the $ARIMA(p,d,q)\ \times (P,D,Q)_S$ noise or disturbance model (2), where:

$$\mathit{\Phi}(B) = 1 - \mathit{\Phi}_1B - \mathit{\Phi}_2B^2 - \dots - \mathit{\Phi}_pB^p;$$

$$\Phi(B) = 1 - \Phi_1B^S - \Phi_2B^{2S} - \dots - \Phi_PB^{PS};$$

$$\theta(B) = 1 - \theta_1B - \theta_2B^2 - \dots - \theta_qB^q$$

$$\Theta(B) = 1 - \Theta_1B^S - \Theta_2B^{2S} - \dots - \Theta_QB^{QS},$$

are the non-seasonal, $\mathit{\Phi}(B)$, and seasonal, $\Phi(B)$, autoregressive components; non-seasonal, $\theta(B)$, and seasonal $\Theta(B)$, moving average components that are expressed as backshift operator functions of the autocorrelation and moving-average parameters. The noise process can be expanded and expressed as

$$\eta_t = \sum_{i=1}^{p}\rho_i\eta_{t-i} + \sum_{j=1}^{q}\theta_ja_{t-j} + a_t, \tag{3}$$

where some of the coefficients ρ_i and θ_j are allowed to be zero [41]. Thus, the ARIMAX model, written as a single model,

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \sum_{i=1}^{p}\rho_i\eta_{t-i} + \sum_{j=1}^{q}\theta_j a_{t-j} + a_t, \tag{4}$$

where the ρ_i's and some of the θ_j's are constrained to zero results in the basic DLR model (i.e., the ARIMAX model), whereas if the ρ_i's and some of the θ_j's are allowed to be non-zero, and other DLR forecasts emerge [41].

Considering the intrinsic temporal component of the response and predictor variables involved, DLR models are widely known procedures for modelling the linear dependence between a stochastic response, Y_t, and a set of k covariates, $\boldsymbol{X}_t = \left(X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(k)}\right)$, a collection of competing candidates for predictors in times less than t [42]. Thus, the ARIMAX model (1) is modified and expressed as

$$y_t = \boldsymbol{X}_{t-i_p}\boldsymbol{\beta} + \eta_t, \tag{5}$$

where $i \geq 0$, $p = 1, 2, \ldots, k$, and $\eta_t \sim ARIMA(p, d, q) \times (P, D, Q)_S$.

Other variations of DLR models would include the autoregressive (AR) structure or lagged effects of the response (y_{t-i}) in the ARIMAX model to predict the current value y_t. In general, different types of DLR models exist from the numerous forms of parameter structure (parameter or transfer functions), of which the ARIMAX models are special cases, used to determine the effect of the covariate on the response [43,44]. The parameter structure that incorporates the lagged version of both the response and the predictor variables is called the dynamic distributed lag (DDL). A DDL model with p lags of the dependent variable and q lags of the independent variable, also known as the Autoregressive Distributed Lag (ARDL) model, is expressed as

$$y_t = \beta_0 + \sum_{i=1}^{p}\phi_i y_{t-i} + \sum_{j=0}^{q}\beta_j x_{t-j} + \eta_t, \tag{6}$$

where $\phi_i y_{t-i}$ are the weighted lagged dependent variable terms, and the other variables are as defined in Eq. (5). This study explores the linear dependence between the response (regional precipitation) and the covariate (oceanographic factors) through DDL models that involve some autoregressive structure (time lags ranging from 1 to 6 months) of both the response, Y_t, and the predictors, $\boldsymbol{X}_t$.

A comparison of all the possible models for predicting precipitation in the three regions was performed. The DLR models were generated by the automatic forecasting function, *auto.arima()*, implemented in the Comprehensive R Archive Network (CRAN) package, *forecast*, which uses the Hyndman-Khandakar algorithm [45]. The R package, *dynlm* [46], was used as an additional tool to fit the DLR models and provide a means to identify other parameters for possible inclusion in the exogenous set of the DLR models. The best DLR model to estimate rainfall in each region was determined based on Akaike's information criterion (AIC) approach in sync with the error measures, Mean Absolute Error (MAE), and the Root Mean Squared Error (RMSE).

Model fit tests

The Ljung-Box test [47] was used to test for the existence of significant autocorrelations in the residuals of the fitted disturbance $ARIMA(p, d, q) \times (P, D, Q)_S$ model based on the following hypotheses:

H_0 : The residuals from the $ARIMA\ (p,\ d,\ q) \times (P, D, Q)_S$ model have no autocorrelation. against

H_1 : The residuals from the $ARIMA\ (p,\ d,\ q) \times (P, D, Q)_S$ model have autocorrelation.

Under H_0, the test statistic,

$$Q = n(n+2)\sum_{k=1}^{v}\frac{\widehat{\rho}_k^2}{n-k}, \tag{7}$$

where n is the sample size, $\widehat{\rho}_k$ is the sample autocorrelation at lag k, and v is the number of lags under consideration, follows the Chi-Square $\left(\chi_v^2\right)$ distribution, with v degrees of freedom. We reject H_0 when Q exceeds the critical value $\chi_v^2(1-\alpha)$, or equivalently, when the p-value is less than α level of significance.

Cross-Correlation function (CCF)

The CCF of the response variable (precipitation) and the lagged predictor variables was examined to identify the candidate predictors for the DLR model. A general guide that was used is that a correlation greater than or equal to 3.0 is considered significant at any particular time lag of the predictor variable.

Validation approach

To evaluate the performance of the selected time series models, a resampling procedure known as the rolling-origin cross-validation (ROCV) method was proposed [48]. The ROCV technique is particularly appropriate for time series data with serial autocorrelation, as it accounts for the temporal dependencies inherent in such data [49]. This addresses a key limitation of traditional cross-validation methods, which assume that all observations are independent and identically distributed (i.i.d.). Given the temporal dependency inherent in time series data, the ROCV technique was employed to preserve the serial autocorrelation structure. The dataset was partitioned into training and testing sets in an approximately 70:30 ratio, respectively. Model performance was assessed using RMSE, MAE, and the Symmetric Mean Absolute Percentage Error (sMAPE), a scale-independent measure. These metrics were computed on the test dataset covering the period from October 2018 through September 2022, derived through rolling-origin cross-validation to ensure temporal integrity of the evaluation. The sMAPE between forecast values, $\hat{y}_t$, and actual values, y_t, over n time periods is calculated as:

$$sMAPE = \frac{100\%}{n} \times \sum_{t=1}^{n} \frac{|y_t - \hat{y}_t|}{(|y_t| + |\hat{y}_t|)/2}. \tag{8}$$

sMAPE values typically range between 0 % and 200 %, where a lower value indicates better predictive performance of the model. For example, sMAPE values between 0 % and 100 % are commonly considered as practical forecasting.

Results

Fig. 3 displays the rainfall distribution in the arid, semi-arid and sub-humid regions of South Africa. The semi-arid and sub-humid regions portray almost similar rainfall patterns, although some records of rainfall amount above 150 mm were observed in certain months in the sub-humid region (Fig. 3).

Stationarity is an important aspect of time series analysis, especially when modelling autocorrelation in hydrology and climate data [7]. In simple terms, a time series is considered stationary when it has a constant mean and an autocorrelation function that does not depend on the time lag [50]. It is, therefore, crucial to test for stationarity in both the response and each of the predictor variables before fitting a DLR model to the time series data. Table 1 displays the results from the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) unit root test for stationarity of the arid, semi-arid and sub-humid precipitation data, and the associated oceanographic variables (*anom, emp, lwRad, qnet* and *rh*).

In all the time series stationarity tests, the KPSS tests produced test-statistic values that do not exceed the critical values. Therefore, stationarity was confirmed in the time series data sets.

Correlations with precipitation

To identify some relatively simple regression structures for modelling precipitation in the three regions, the CCF, which is a set of correlations between the response (precipitation, y_t) and the lagged predictors $x_{t\pm k}$, for $k = 0, \pm 1, \pm 2, \ldots$, was considered. We focus on leading x-variables (i.e., with k negative in one or more x_{t+k},) since we use these past observations to predict future values of the response variable, y_t. Table 2 displays the correlation coefficients between precipitation and lagged oceanographic factors.

The variable '*anom*' had some significant correlation values with the precipitation in the arid region (at lags $k = -2, -3, -4, -8$,

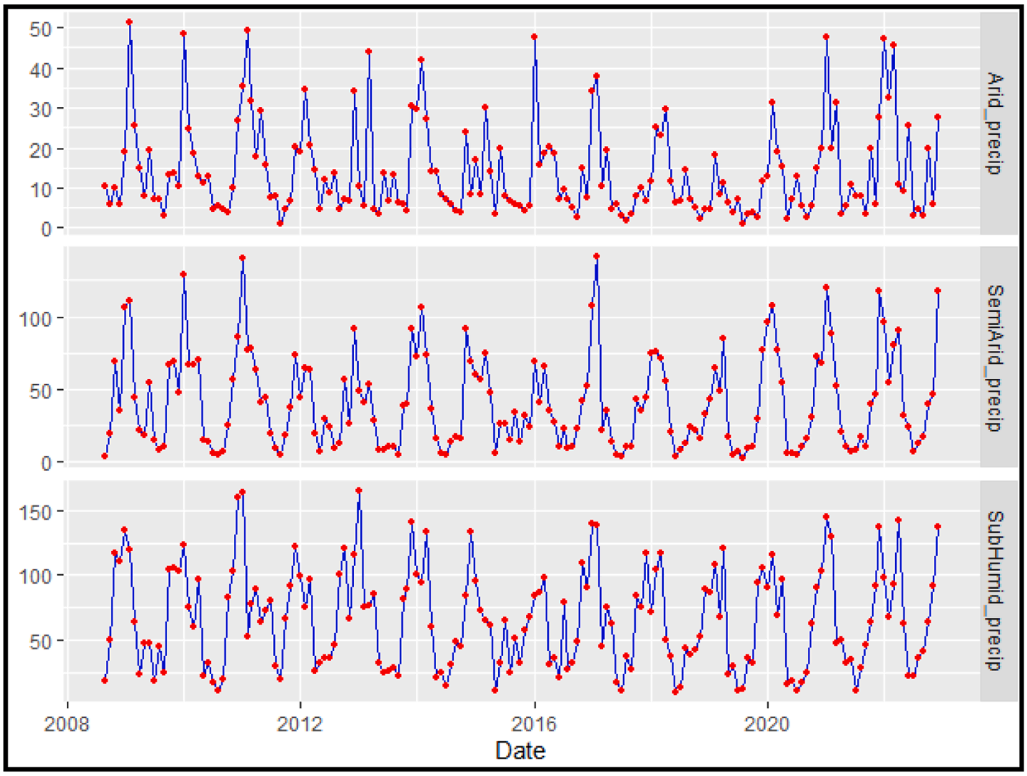

Fig. 3. Precipitation time series in arid, semi-arid and subhumid regions.

Table 1
Kwiatkowski–Phillips–Schmidt–Shin (KPSS) unit root test for stationarity.

Variable	Test statistic	Critical values	
		$\alpha = 0.05$	$\alpha = 0.01$
Arid precipitation	0.12	0.46	0.74
Semi-arid precipitation	0.05	0.46	0.74
Sub-humid precipitation	0.07	0.46	0.74
anom	0.04	0.46	0.74
emp	0.45	0.46	0.74
lwRad	0.22	0.46	0.74
qnet	0.05	0.46	0.74
rh	0.19	0.46	0.74

-9), semi-arid region (at lags $k = -1, \ldots, -4, -7, -8, -9$) and the sub-humid region (at lags $k = -1, -2, -3, -7, -8$). The variable '*emp*' had non-significant correlation values with the precipitation in the arid and semi-arid regions, and two significant correlation values in the sub-humid region (at lags $k = -6$, and -7). Several significant correlation values between the response and the variable *lwRad* were observed, with the highest $r = -0.64$ at lag $k = -3$ in the semi-arid and ($r = -0.67$) at lag $k = -2$ in the subhumid region. Moderate to low correlation values between the response and *rh* were detected, with the max (-0.50) at lag $k = -7$ in the semi-arid region.

Models of predicting precipitation for the arid region

Table 3 displays the estimated coefficients, model fit and error measures for the ARIMAX model, with *ARIMA(2,1,1)(0,0,1)[12]* errors in the noise model, and the DDL model for the prediction of precipitation in the arid region.

The Ljung-Box test for the autocorrelation of the error terms for the ARIMAX model had a p-value of 0.3324 for the test statistic $Q = 24.29$, whereas for the DLR model, the Breusch-Godfrey (BG) test for serial correlation had a p-value of 0.08 for the test statistic $Q = 34.53$ on 24 degrees of freedom. Since the p-values for the two tests are both greater than the significance level ($\alpha = 0.05$), a good model fit is confirmed. Based on AIC and error measures displayed in Table 3, the ARIMAX model had a better model fit (lower AIC, MAE and RMSE values) than the DLR model.

The DLR model in Table 3 suggests that the current sea surface temperature anomalies [L(anom, 0:5)0] and lagged longwave radiation [L(lwrad, 0:5)3] have a significant influence on the rainfall received by the arid region. Combining the significant effects from the DLR and the ARIMAX models in Table 3, the updated DDL model (see Eqs. (9) and (10), in the Supplementary Document), with additional coefficients in the covariate and ARIMA(2,0,0) errors for the prediction of precipitation in the arid region, was fitted.

The previous month's precipitation (y_{t-1}), current and previous months' anomalies, current and lagged *lwRad* observations contribute significantly (p-values < 0.05) to the prediction of the current precipitation in the arid region. The Ljung-Box test for the autocorrelation of the error terms from the DDL model (10) produced a p-value of 0.59 for the test statistic $Q = 19.93$ on 22 degrees of freedom. Since the p-value (0.59) is greater than the significance level ($\alpha = 0.05$), the fitted model fits the data well and compares better (lower $AIC = 1123.83$ and $RMSE = 8.26$) than the first two summarised in Table 3. All the oceanographic variables in the covariate, except for relative humidity (*rh*), have a significant influence (p-values < 0.05) on the prediction of precipitation received in

Table 2
Cross-Correlation Function (CCF) of precipitation and the predictors.

Arid	Lag (k)												
Variable	0	−1	−2	−3	−4	−5	−6	−7	−8	−9	−10	−11	−12
anom	−0.12	0.22	0.45	0.55	0.47	0.25	−0.04	−0.23	−0.30	−0.36	−0.31	−0.22	−0.07
emp	−0.14	−0.22	−0.24	−0.21	−0.10	0.08	0.21	0.22	0.24	0.21	−0.05	−0.09	−0.17
lwRad	0.27	−0.05	−0.30	−0.54	−0.54	−0.34	−0.06	0.20	0.33	0.42	0.38	0.35	0.22
rh	0.23	0.37	0.40	0.33	0.28	0.07	−0.13	−0.33	−0.37	−0.21	−0.05	0.13	0.23
Semi-Arid													
anom	0.16	0.47	0.62	0.59	0.36	0.03	−0.27	−0.48	−0.53	−0.49	−0.33	−0.08	0.17
emp	−0.25	−0.28	−0.24	−0.19	−0.01	0.18	0.28	0.31	0.25	0.15	−0.07	−0.20	−0.27
lwRad	−0.01	−0.37	−0.58	−0.64	−0.52	−0.24	0.12	0.40	0.57	0.60	0.50	0.28	−0.01
rh	0.36	0.45	0.40	0.25	0.06	−0.20	−0.39	−0.50	−0.42	−0.24	−0.01	0.24	0.37
Sub-humid													
anom	0.29	0.53	0.59	0.46	0.18	−0.13	−0.37	−0.53	−0.54	−0.47	−0.27	0.04	0.35
emp	−0.29	−0.29	−0.20	−0.09	−0.06	0.23	0.33	0.33	0.15	0.09	−0.15	−0.25	−0.29
lwRad	−0.19	−0.51	−0.67	−0.59	−0.40	−0.09	0.24	0.48	0.61	0.61	0.47	0.15	−0.18
rh	0.44	0.47	0.40	0.21	−0.06	−0.26	−0.42	−0.41	−0.29	−012	0.11	0.33	0.45

Correlation is considered significant if $|r| \geq 0.3$.

Table 3
Models for the prediction of precipitation in the arid region.

ARIMAX Model				DDL Model			
Coeff	Estimate	S.E	P-value	Coeff	Estimate	S.E	P-value
ma1	−0.94	0.07	<0.001*	Intercept	−29.76	30.73	0.335
sma1	0.18	0.10	0.008*	L(arid, 1)	−0.86	0.08	0.074
anom	0.25	0.87	0.773	L(anom, 0:5)0	−2.72	1.12	0.016*
emp	2.73	15.72	0.862	L(anom, 0:5)1	3.08	1.37	0.256
lwRad	0.45	0.11	<0.001*	L(lwRad, 0:5)0	0.23	0.12	0.056
qnet	0.09	0.02	<0.001*	L(lwRad, 0:5)3	−0.35	0.12	0.005*
				L(lwRad, 0:5)4	−0.29	0.18	0.104
				L(lwRad, 0:5)5	0.26	0.14	0.057
				L(rh, 0:5)0	0.54	0.28	0.058
Model Fit		**Error Measures**		**Model Fit**		**Error Measures**	
AIC	1241.89	MAE	7.23	AIC	1283.44	MAE	13.03
BIC	1263.72	RMSE	9.46	BIC	1288.06	RMSE	15.99

'*' Significant at $\alpha = 0.05$. L(*variable, n*) represents the variable at time lag *n*.
Note: 'ma1'is the first-order MA, 'smar1' is the seasonal moving average term of order 1.

the arid region.

Fig. 4 shows the 12-step-ahead monthly precipitation forecasts around the forecast curve. There is a consistent pattern in the forecasts for the observed rainfall amounts. The correlation between the observed rainfall values in the arid region and the values from the fitted model was 0.70, which suggests that the updated model better aligns with observed values. This is an indication that the estimated model is performing well in predicting precipitation in the region.

The forecast errors ($MAE = 6.18$, $RMSE = 8.26$) obtained from the updated model for the prediction of precipitation in the arid region were comparable to the hold-out cross-validation error measures ($MAE = 2.51$, $RMSE = 10.50$). The forecast errors are slightly above those of the training set, as expected. Thus, the model predicts well on unseen data.

Models for the prediction of precipitation in the semi-arid region

The coefficients, model fit, and error measures of the estimated ARIMAX model, with *ARIMA(0,0,0)(2,1,0)[12]* errors in the noise model, and the DDL models for predicting rainfall in the semi-arid region are summarised in Table 4.

Based on the AIC and error measures displayed in Table 4, the DDL model, with time lags (1, 2, 5, 6, 12) on semi-arid region precipitation (*sarid*), time lags 2 on the *anom* and *lwrad*, predicts semi-arid rainfall better than the ARIMAX model. In both models, *anom* and *emp* seem to have no significant influence on the rainfall amount in this region (p-values > α (0.05)).

The significant coefficients in the DDL model (Table 4) were used to update the ARIMAX model since, by the time of writing, the

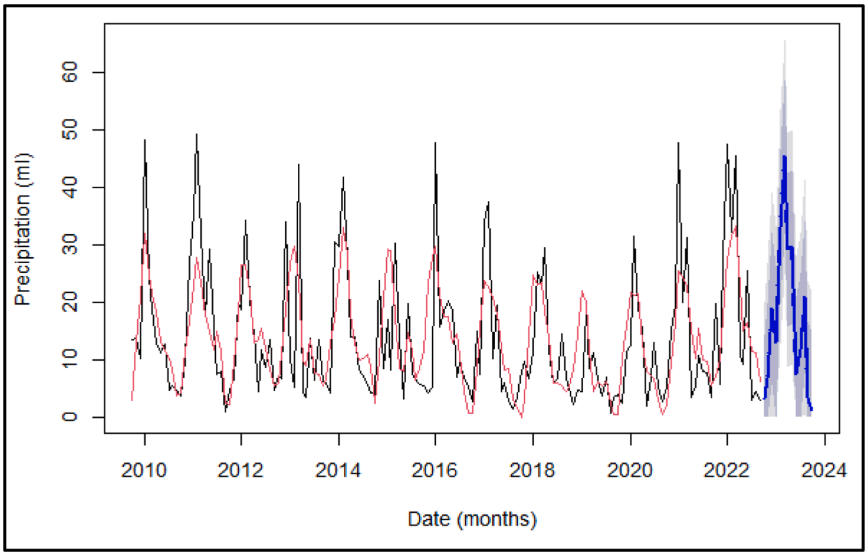

Fig. 4. Forecast plot of the precipitation in the arid region. The black and red curves represent the observed and the fitted precipitation values, respectively. The blue curve represents the 12-month-ahead forecasts, and the grey-shaded areas are the 80 % and 95 % confidence intervals of the forecasts.

Table 4
Models for the prediction of precipitation in the semi-arid region.

ARIMAX Model				DDL Model			
Coeff	Estimate	S.E	P-value	Coeff	Estimate	S.E	P-value
sar1	−0.52	0.07	< 0.001*	Intercept	51.82	16.71	0.002*
sar2	−0.51	0.07	< 0.001*	L(sarid, 1:2)1	0.21	0.08	0.008*
anom	−0.86	1.64	0.602	L(sarid, 1:2)2	0.11	0.07	0.122
emp	43.69	29.16	0.132	L(sarid, 5:6)5	−0.17	0.07	0.024*
				L(sarid, 5:6)6	−0.09	0.08	0.289
				L(sarid, 12)	0.19	0.08	0.021*
				L(anom, 2)	2.76	1.88	0.144
				L(lwRad, 2)	−0.59	0.26	0.026*
Model Fit		Error Measures		Model Fit		Error Measures	
AIC	1412.13	MAE	14.14	AIC	1395.45	MAE	13.46
BIC	1427.37	RMSE	20.29	BIC	1422.90	RMSE	17.03

'*' Significant at $\alpha = 0.05$. L(*variable, n*) represents the variable at time lag *n*.
Note: 'sar1' is the seasonal autoregressive term of order 1,.
'sar2' is the seasonal autoregressive term of order 2.

dynlm procedure could not provide a prediction procedure [46]. The updated model, with *ARIMA(0,0,2)(2,0,0)[12]* errors (see Eqs. 11 and 12, Supplementary document), was fitted for the prediction of the current precipitation (y_t) in the semi-arid region. The moving average error terms and the seasonal autoregressive (lag2) error terms in model (12), the previous semi-arid precipitation observations (lags 1 and 2), anom (lag 2) and lwRad (lag 2) in model (11) had significant influence (p-values < 0.05) in the prediction of precipitation in the semi-arid region. The rest of the terms had insignificant effects (p-values > 0.05). The Ljung-Box test for the autocorrelation of the error terms from the fitted model (11) produced a p-value of 0.86 for the test statistic $Q = 13.44$ on 20 degrees of freedom. Since the p-value (0.86) is greater than the significance level ($\alpha = 0.05$), the fitted model fits the data well and compares better (lower AIC = 1390.37 and RMSE = 19.15) than the first two, summarised in Table 4.

Fig. 5 portrays the 12-step-ahead monthly precipitation forecasts (blue curve) and the 80 % and 95 % confidence intervals (grey and blue shaded areas, respectively) around the forecast curve of the fitted model (11). There is a relatively consistent pattern of the forecasts with respect to the original observations. The correlation between the observed rainfall values and the fitted values from the semi-arid rainfall data was 0.82, indicating a strong positive linear relationship. This suggests the model effectively captures the overall trend and variability in the semi-arid precipitation data. This further confirms that the proposed model (11) is performing and predicting the precipitation in this region very well.

The forecast errors ($MAE = 14.22$, $RMSE = 18.85$) obtained from the fitted model for the semi-arid region precipitation are comparable to the hold-out cross-validation error measures ($MAE = 17.71$, $RMSE = 22.88$), which are slightly above those of the training set, as expected. Thus, the fitted model predicts well on unseen data.

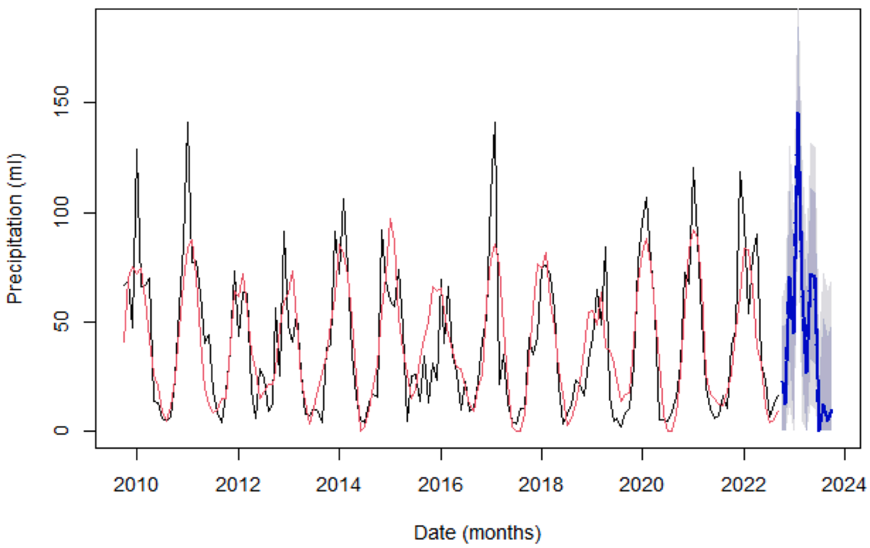

Fig. 5. Forecast plot of precipitation in the semi-arid region. The black and red curves represent the observed and the fitted precipitation values, respectively. The blue curve represents the 12-month-ahead forecasts, and the grey-shaded areas are the 80 % and 95 % confidence intervals of the forecasts.

Models for the prediction of precipitation in the sub-humid region

Similarly, the DLR models for estimating and predicting rainfall in the sub-humid region were compared. Table 5 displays the coefficients, standard errors and model fit measures for the estimated ARIMAX model, with *ARIMA(0,0,0)(2,1,1)[12]* errors, and the DDL model in the noise model for rainfall prediction in the sub-humid region. The ARIMAX model consists of the sea surface anomalies (*anom*) as the only exogenous input variable whose disturbance model is made up of the seasonal autocorrelation (lag 2) and seasonal moving averages (lag 1) terms, and with white-noise errors. On the other hand, the DDL model showed significant prediction effects from the previous longwave radiation (*lwRad, 2*) observation (lag 2) and the past observations (lags 1 and 5) of precipitation (*shumid*) in the sub-humid region.

The Ljung-Box test for the autocorrelation of the error terms for the ARIMAX model had a p-value of 0.60 for the test statistic $Q = 18.74$, whereas the Breusch-Godfrey (BG) test for serial correlation had a p-value of 0.83 for the test statistic $Q = 17.47$ for the DLR model. Since the p-values for the two tests are greater than the significance level ($\alpha = 0.05$), a good model fit is confirmed. Although the AIC and error measures displayed in Table 5 are a bit too high, the ARIMAX model looks better (lower AIC, MAE and RMSE values) than the DDL model in this case. With non-significant seasonal autoregressive and significant moving average effects of the error terms, the sea surface temperature anomalies seem to have a significant negative influence on the rainfall amount in this region.

Using the significant effects of the DLR models summarised in Table 5, the DDL, with *ARIMA(0,0,1)* errors, for precipitation prediction in the sub-humid region was fitted (see Eqs. 13 and 14, Supplementary document). All the parameters of the fitted model (13–14) had a significant influence (p-values < 0.05) in the prediction of precipitation in the sub-humid region. The Ljung-Box test for the autocorrelation of the *ARIMA(0,0,1)* error terms from the updated model (13) produced a p-value of 0.51 for the test statistic $Q = 22.19$ on 23 degrees of freedom. Since the p-value (0.51) is greater than the significance level ($\alpha = 0.05$), the fitted model fits the data well and shows a slight improvement (lower $AIC = 1467.23$) from the competing models displayed in Table 5. The updated model (13–14) relies on the previous observations of sea surface temperature anomalies (lag 5), longwave radiation (lag 2) and sub-humid rainfall (lags 1 and 5) for the prediction of the current precipitation in the region.

Fig. 6 portrays the 12-step-ahead monthly precipitation forecasts and their respective 80 % and 95 % confidence intervals generated by the DDL model in the sub-humid region. Relative to the original monthly rainfall observations in this region, there is consistency in the forecasts generated by the fitted model, as suggested by a high positive correlation of 0.78. This confirms that the proposed model is performing well in the prediction of monthly rainfall in this region.

The forecast errors ($MAE = 20.53$, $RMSE = 25.77$) obtained from the updated DDL model (10) for the prediction of precipitation in the sub-humid region were comparable to the hold-out cross-validation error measures ($MAE = 21.53$, $RMSE = 26.77$), which are slightly above those of the training set, as expected. This confirms that the fitted model predicts well on unseen data.

To assess forecast performance, we employed rolling-origin cross-validation for each regional rainfall model (semi-arid, arid, sub-humid). Model accuracy was evaluated using the error-based metrics, RMSE, MAE, and sMAPE. Results are summarised in Table 6.

The sub-humid model had the lowest sMAPE (38.11 %), suggesting the most consistent relative accuracy. The arid model had the lowest absolute errors (RMSE = 8.77, MAE = 6.82), reflecting tight forecast accuracy on a smaller scale. The semi-arid model performed moderately, with mid-range error metrics but higher sMAPE (64.74 %). Among the three models, the sub-humid region model demonstrated the lowest relative error (sMAPE = 38.11 %), suggesting higher proportional accuracy despite a higher RMSE due to greater rainfall variability. The arid model achieved the lowest absolute error (RMSE = 8.77, MAE = 6.82), while the semi-arid model showed intermediate performance, with RMSE = 22.31 and sMAPE = 64.74 %. These results highlight regional variability in model predictability and the influence of seasonal dynamics on forecast precision.

Discussion

Regional rainfall prediction has long faced significant limitations due to data availability and model resolution, challenges that

Table 5
Models for the prediction of precipitation in the sub-humid region.

ARIMAX model				DDL model			
Coeff	Estimate	S.E	P-value	Coeff	Estimate	S.E	P-value
sa1	−0.17	0.23	0.464	Intercept	143.65	14.89	< 0.001*
sar2	−0.09	0.19	0.623	L(shumid, 1)	0.27	0.07	< 0.001*
sma1	−0.68	0.24	0.004*	L(shumid, 5)	−0.18	0.07	0.007*
xreg	−4.69	2.29	0.040*	L(anom, 5)	−0.53	1.76	0.763
				L(lwRad, 2)	−1.63	0.29	< 0.001*
Model Fit		**Error Measures**		**Model Fit**		**Error Measures**	
AIC	1469.61	MAE	18.19	AIC	1532.10	MAE	25.07
BIC	1484.86	RMSE	24.19	BIC	1550.66	RMSE	33.16

'*' Significant at $\alpha = 0.05$. L(*variable, n*) represents the variable at time lag *n*.
Note: 'sar1' is the seasonal autoregressive term of order 1,.
'sar2' is the seasonal autoregressive term of order 2.
'smar1' is the seasonal moving average term of order 1.

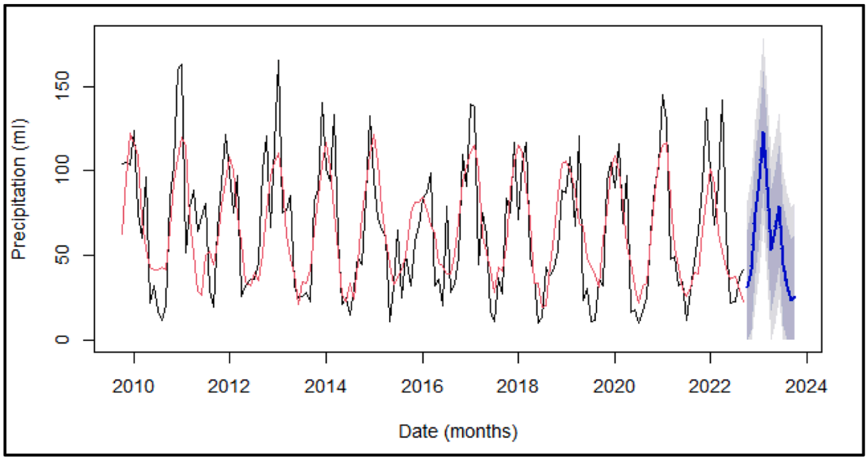

Fig. 6. Forecast plot of precipitation in the sub-humid region. The black and red curves represent the observed and the fitted precipitation values, respectively. The blue curve represents the 12-month-ahead forecasts, and the grey-shaded areas are the 80 % and 95 % confidence intervals of the forecasts.

Table 6
Models for the prediction of precipitation in the South African regions.

Region	RMSE	MAE	sMAPE (%)
Semi-arid	22.31	17.35	64.74
Arid	8.77	6.82	60.27
Sub-humid	26.78	20.50	38.11

climate change has exacerbated by introducing greater variability and extremes. Recent advancements in modelling techniques, such as ML, present new opportunities for improving accuracy and understanding the impacts on agriculture and water resources. Recent advances in modelling techniques, such as ML, present new opportunities to improve accuracy and understanding of the impacts on agriculture and water resources [51]. Analysis of rainfall distribution across the arid, semi-arid, and sub-humid regions reveals clear climatic gradients, with the arid zone consistently receiving the lowest monthly precipitation. Rainfall distribution patterns align with findings from previous studies [52,53], showing similar seasonal trends in semi-arid and sub-humid regions. However, occasional monthly totals above 150 mm in the sub-humid zone, as also reported by previous research studies [54,55], point to localised climatic events influencing precipitation variability. Although rare, these sudden heavy rainfall events show how unpredictable sub-humid climates can be, and emphasise the need for location-specific monitoring to detect such anomalies.

Comparison of dynamic linear regression (DLR) models

The cross-correlation analysis revealed significant relationships between precipitation and lagged oceanographic variables, with notable variations across different climatic regions. In particular, the variable ‘anom’ exhibited strong correlations ($|r| \geq 0.3$) in multiple lags in arid, semi-arid, and sub-humid regions, suggesting its potential as a key predictor of precipitation variability. This aligns with the findings of Aryee et al. [56], who demonstrated that sea surface temperature (SST) anomalies significantly influence rainfall patterns in tropical and subtropical regions. The highest negative correlations for longwave radiation (lwRad) were $r = -0.64$ and -0.67 in semi-arid and sub-humid regions, respectively. This indicated that lwRad from the air plays a critical role in modulating convective activity, consistent with studies linking reduced lwRad to enhanced cloud cover and precipitation [57,58]. Meanwhile, ‘emp’ showed limited predictive power, except in the sub-humid region, possibly due to regional differences in moisture flux dynamics. The moderate correlations with ‘rh’ further support the influence of atmospheric moisture on precipitation, particularly in semi-arid zones, as observed in similar climatic studies [59,60]. These findings highlight the importance of region-specific lagged predictors in precipitation modelling, offering valuable information to improve seasonal forecasting in vulnerable ecosystems.

Model accuracy in predicting monthly average precipitation

The ARIMAX and DDL models demonstrated strong predictive performance for precipitation in the arid region, with the ARIMAX model outperforming the DDL model. In particular, lwRad and qnet were significant predictors ($p < 0.001$), aligning with studies that highlight radiative forcing as a key driver of convective rainfall in arid zones [61]. The DDL model further revealed that lagged SST anomalies (anom) and lwRad (lags 3 and 5) significantly influenced precipitation, consistent with findings by Ogwang et al. [62], who identified SST anomalies as critical for modulating moisture convergence in drylands. The high correlation ($r = 0.70$) between observed and fitted values, along with low forecast errors (MAE = 6.18, RMSE = 8.26), underscores the reliability of the model for seasonal forecasting in water-scarce regions.

In the semi-arid region, the DDL model outperformed the ARIMAX model, with lagged precipitation (lags 1, 2, 5, 6, 12) and lwRad (lag 2) being significant predictors ($p < 0.05$). The strong seasonal autoregressive effects (SAR1, SAR2) suggest persistent intra-annual rainfall patterns, corroborating research by Nkiaka et al. [63] on the role of land-atmosphere feedbacks in semi-arid climates. However, unlike the arid region, the anomalies of the SST (anom) did not show a significant influence, possibly due to the greater dependence on localised moisture recycling in the region [64]. The high correlation ($r = 0.82$) between observed and predicted values indicates robust model performance, although the higher sMAPE (64.74 %) suggests challenges in capturing extreme rainfall variability, a known limitation in semi-arid forecasting [65].

For the sub-humid region, the ARIMAX model performed better than the DDL model, with SST anomalies (anom) exerting a significant negative influence ($p = 0.04$). This aligns with studies linking warmer SSTs to suppressed convection in humid tropics [66,67]. The DDL model highlighted the importance of lagged lwRad (lag 2) and precipitation (lags 1 and 5), suggesting strong moisture retention and memory effects in sub-humid ecosystems [60]. Despite higher absolute errors (RMSE = 25.77), the model achieved the lowest sMAPE (38.11 %), indicating superior relative accuracy, which is a critical metric for high-rainfall regions where absolute errors scale with precipitation magnitude [68]. The correlation ($r = 0.78$) further validates the utility of the model for operational forecasting in sub-humid zones.

Importance of the study

The research findings emphasise the urgent need for South Africa to develop effective adaptation strategies in response to projected changes in precipitation patterns. Priority should be given to strategies that improve water resource management, promote climate-resilient agricultural practices, strengthen public health infrastructure, and build climate-resilient infrastructure [69]. Following the implementation of these measures, South Africa can proactively address the challenges posed by changing precipitation patterns and ensure the sustainability and resilience of its socio-economic and ecological systems.

Additionally, the study highlights the significant influence of oceanographic factors on the monthly average precipitation in the arid, semi-arid, and sub-humid regions of South Africa [70]. Integrating and understanding these oceanographic factors within climate models and forecasting systems can greatly enhance the precision of precipitation predictions and provide valuable information for adaptive decision-making. These findings have crucial implications for multiple sectors within the country, including agriculture and socio-economic growth. Leveraging the improved understanding of oceanographic influences on precipitation, policymakers and stakeholders can devise targeted strategies to mitigate food insecurity, improve agricultural practices, and promote economic development. South Africa can advance sustainable development practices and effectively address the challenges posed by evolving precipitation dynamics by aligning adaptation efforts with the insights derived from this study.

Furthermore, the findings of this study offer concrete pathways to strengthen climate adaptation efforts in South Africa. This research provides direct support for farmers in key agricultural regions like the Free State, enabling them to optimise planting cycles and irrigation strategies in anticipation of dry spells by converting monthly rainfall forecasts into actionable tools. In flood-vulnerable urban centres such as eThekwini, these predictions could improve preparedness for extreme rainfall events through targeted infrastructure planning. At the policy level, more reliable seasonal projections are expected to improve the implementation of the National Water and Sanitation Master Plan through data-driven water allocation decisions. The potential of the prediction model extends beyond South Africa, with applications possible in the climate-variable regions of southern Africa. However, realising this potential requires ongoing refinement to minimise predictive uncertainties and greater collaboration between researchers and local stakeholders. Nowhere is this need more pressing than in the arid and semi-arid zones of South Africa, where changing precipitation patterns increasingly jeopardise water security and agricultural stability. This work ultimately demonstrates how climate science, when properly integrated with governance systems and community knowledge, can deliver practical solutions to pressing environmental challenges.

Limitations of the study

Although the dynamic linear regression (DLR) model framework demonstrated the ability to predict monthly average precipitation in arid, semi-arid, and sub-humid regions based on oceanographic variables, several limitations were encountered in this study. The PIRATA and RAMA ocean buoys were geographically distant from the weather stations where the precipitation data were collected, introducing potential spatial disparities and temporal differences that could have affected the accuracy of the results. Moreover, the findings were specific to the selected regions of South Africa and may not be directly applicable to other geographic areas with different climatic conditions. The generalisability of the results should be approached with caution when applying them to different regions. Additionally, the accuracy of predictions using the DLR model is based on assumptions made within the model, such as the regression coefficients assumed to be constant over time and the linearity of the relationships between predictor variables and precipitation. Recognising the potential limitations and uncertainties associated with these assumptions is crucial. When interpreting the study's results and considering their applicability to real-world scenarios, it is important to acknowledge these limitations. More research and validation efforts are necessary to improve the accuracy and robustness of precipitation predictions in arid, semi-arid, and sub-humid regions using the proposed approach.

Conclusion

In summary, this study demonstrates the effectiveness of DDL models as powerful tools in improving the accuracy of rainfall forecasting in Southern Africa, where climate variability poses significant challenges to sustainable water and agricultural resource management. Although traditional forecasting methods face constraints due to data and software limitations, this research illustrates how DDL models, when tailored to the unique climatic conditions of the region and enriched with relevant oceanographic and exogenous variables, can significantly improve precipitation prediction. Despite some functional limitations in existing software, statistical software such as R provides viable solutions for model refinement and integration of critical parameters. These innovations present a strategic pathway to more reliable and actionable forecasts to support regional climate resilience. From the synthesised data and modelling approaches, several key findings have emerged.

1. DDL models, when regionally adapted, offer substantial improvements in rainfall prediction accuracy across diverse climatic zones in Southern Africa.
2. The incorporation of oceanographic factors into DDL and DLR models significantly enhances their predictive capacity, underscoring the importance of environmental context in model design.
3. Software limitations do not pose insurmountable barriers, alternative methods available in R and other software can effectively integrate vital data inputs.
4. Enhanced rainfall forecasts are achievable by refining the composition of exogenous variables and combining multiple predictive parameters through robust DDL and ARIMA modelling techniques.

Advancements in modelling support more informed decision-making for water resource planning and agricultural management, ultimately strengthening resilience to climate change in the region. For the academic community, this study provides a replicable and data-driven modelling framework that advances the use of machine learning in environmental forecasting, offering new pathways for research in climate science, hydrology, and sustainable development. This study contributes a valuable modelling framework that blends advanced machine learning with environmental data, offering a scalable and adaptable tool for meteorological forecasting and climate-informed policymaking in Southern Africa.

Credit author statement

Lyson Chaka: Conceptualization, Methodology, Data Analysist, Writing - Original draft preparation. **Mohamed A. M. Abd Elbasit**: Data collection, Conceptualization, Writing - Original draft preparation. **Simbarashe Jombo**: Visualization, Writing - Original draft preparation.

Funding source declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

Authors have no competing interests to declare.

Acknowledgements

We gratefully acknowledge the National Aeronautics and Space Administration (NASA), the Distributed Active Archive Center (DAAC), and the Goddard Earth Sciences Data and Information Services Center (GES DISC) for providing access to Earth science data. We also thank the Tropical Atmosphere Ocean (TAO) Project Office of the National Oceanic and Atmospheric Administration (NOAA/PMEL) for making the PIRATA and RAMA oceanographic datasets available. We also thank the Water Research Commission (WRC) and the National Research Foundation (NRF) of South Africa for their support.

References

[1] C. Nhemachena, L. Nhamo, G. Matchaya, C.R. Nhemachena, B. Muchara, S.T. Karuaihe, S. Mpandeli, Climate change impacts on water and agriculture sectors in Southern Africa: threats and opportunities for sustainable development, Water (Basel) 12 (10) (2020) 2673, https://doi.org/10.3390/w12102673.

[2] S.A. Lamega, M. Komainda, M.P. Hoffmann, K.K. Ayisi, J.J.O. Odhiambo, J. Isselstein, It depends on the rain: smallholder farmers' perceptions on the seasonality of feed gaps and how it affects livestock in semi-arid and arid regions in Southern Africa, Clim Risk Manag 34 (2021) 100362, https://doi.org/10.1016/j.crm.2021.100362.

[3] T. Gao, H.J. Wang, T. Zhou, Changes of extreme precipitation and nonlinear influence of climate variables over monsoon region in China, Atmos Res 197 (2017) 379–389, https://doi.org/10.1016/j.atmosres.2017.07.017.

[4] G. Liu, A. Xiang, Z. Wan, Y. Zhou, J. Wu, Y. Wang, S. Lin, Variations of extreme precipitation events with sub-daily data: a case study in the Ganjiang River basin, Natural Hazards Earth System Sci 23 (3) (2023) 1139–1155, https://doi.org/10.5194/nhess-23-1139-2023.

[5] F.M. Mashao, M.C. Mothapo, R.B. Munyai, J.M. Letsoalo, I.L. Mbokodo, T.P. Muofhe, W. Matsane, H. Chikoore, Extreme rainfall and flood risk prediction over the East Coast of South Africa, Water (Basel) 15 (1) (2022) 50, https://doi.org/10.3390/w15010050.

[6] H. Du, L.V. Alexander, M.G. Donat, T. Lippmann, A. Srivastava, J. Salinger, A. Kruger, G. Choi, H.S. He, F. Fujibe, M. Rusticucci, B. Nandintsetseg, R. Manzanas, S. Rehman, F. Abbas, P. Zhai, I. Yabi, M.C. Stambaugh, S. Wang, Z. Wu, Precipitation from persistent extremes is increasing in most regions and globally, Geophys Res Lett 46 (11) (2019) 6041–6049, https://doi.org/10.1029/2019GL081898.

[7] F. Sun, M.L. Roderick, G.D. Farquhar, Rainfall statistics, stationarity, and climate change, Proceed National Acad Sci 115 (10) (2018) 2305–2310, https://doi.org/10.1073/pnas.1705349115.

[8] Y. Zhu, J. Sun, J. Ma, A hybrid statistical-dynamical prediction model for summer precipitation in northwestern China based on NCEP CFSv2, Atmos Res 283 (2023) 106567.

[9] R. Farneti, Modelling interdecadal climate variability and the role of the ocean, Wiley Interdiscipl Rev: Climate Change 8 (1) (2017) e441.

[10] L.J. Slater, L. Arnal, M.A. Boucher, A.Y.Y. Chang, S. Moulds, C. Murphy, G. Nearing, G. Shalev, C. Shen, L. Speight, G. Villarini, Hybrid forecasting: blending climate predictions with AI models, Hydrol Earth Syst Sci 27 (9) (2023) 1865–1889.

[11] O.A. Wani, S.S. Mahdi, M. Yeasin, S.S. Kumar, A.S. Gagnon, F. Danish, N. Al-Ansari, S. El-Hendawy, M.A Mattar, Predicting rainfall using machine learning, deep learning, and time series models across an altitudinal gradient in the North-Western Himalayas, Sci Rep 14 (1) (2024) 27876.

[12] D. Weldon, C.J.C. Reason, Variability of rainfall characteristics over the South Coast region of South Africa, Theor Appl Climatol 115 (2014) 177–185.

[13] C.M. McBride, A.C. Kruger, L. Dyson, Changes in extreme daily rainfall characteristics in South Africa: 1921–2020, Weather Clim Extrem 38 (2022) 100517.

[14] M. Baljon, S.K. Sharma, Rainfall prediction rate in Saudi Arabia using improved machine learning techniques, Water (Basel) 15 (4) (2023) 826, https://doi.org/10.3390/w15040826.

[15] E.T. Busayo, A.M. Kalumba, I.R. Orimoloye, Spatial planning and climate change adaptation assessment: perspectives from Mdantsane Township dwellers in South Africa, Habitat Int 90 (2019) 101978, https://doi.org/10.1016/j.habitatint.2019.04.005.

[16] G. Hu, S.L. Dance, R.N. Bannister, H.G. Chipilski, O. Guillet, B. Macpherson, M. Weissmann, N. Yussouf, Progress, challenges, and future steps in data assimilation for convection-permitting numerical weather prediction: report on the virtual meeting held on 10 and 12 November 2021, Atmospheric Sci Letters 24 (1) (2023), https://doi.org/10.1002/asl.1130.

[17] K. Song, G. Yang, Q. Wang, C. Xu, J. Liu, W. Liu, C. Shi, Y. Wang, G. Zhang, X. Yu, Z. Gu, W. Zhang, Deep learning prediction of incoming rainfalls: an operational service for the City of Beijing China, in: 2019 International Conference on Data Mining Workshops (ICDMW), 2019, pp. 180–185, https://doi.org/10.1109/ICDMW.2019.00036.

[18] L. Xu, N. Chen, C. Yang, H. Yu, Z. Chen, Quantifying the uncertainty of precipitation forecasting using probabilistic deep learning, Hydrol Earth Syst Sci 26 (11) (2022) 2923–2938, https://doi.org/10.5194/hess-26-2923-2022.

[19] E. Ghaderpour, H. Dadkhah, H. Dabiri, F. Bozzano, G. Scarascia Mugnozza, P. Mazzanti, Precipitation time series analysis and forecasting for Italian regions, Eng Proc 39 (1) (2023) 23, https://doi.org/10.3390/engproc2023039023.

[20] N. Koyama, M. Sakai, T. Yamada, Study on a water-level-forecast method based on a time series analysis of urban River basins—A case Study of Shibuya River Basin in Tokyo, Water (Basel) 15 (1) (2022) 161, https://doi.org/10.3390/w15010161.

[21] T. Kashiwao, K. Nakayama, S. Ando, K. Ikeda, M. Lee, A. Bahadori, A neural network-based local rainfall prediction system using meteorological data on the internet: a case study using data from the Japan Meteorological Agency, Appl Soft Comput 56 (2017) 317–330, https://doi.org/10.1016/j.asoc.2017.03.015.

[22] Y. Lai, D.A. Dzombak, Use of the Autoregressive integrated Moving Average (ARIMA) model to forecast near-term regional temperature and precipitation, Weather Forecast 35 (3) (2020) 959–976, https://doi.org/10.1175/WAF-D-19-0158.1.

[23] B. Bourlès, R. Lumpkin, M.J. McPhaden, F. Hernandez, P. Nobre, E. Campos, L. Yu, S. Planton, A. Busalacchi, A.D. Moura, J. Servain, J. Trotte, THE PIRATA PROGRAM: history, accomplishments, and future directions, Bull Am Meteorol Soc 89 (8) (2008) 1111–1126, https://doi.org/10.1175/2008BAMS2462.1.

[24] M.J. McPhaden, M.J. McPhaden, M.J. McPhaden, M.J. McPhaden, M.J. McPhaden, M.J. McPhaden, M.J. McPhaden, M.J. McPhaden, M.J. McPhaden, M. J. McPhaden, M.J. McPhaden, M.J. McPhaden, The global tropical moored buoy array, in: Proceedings of OceanObs'09: Sustained Ocean Observations and Information for Society, 2010, pp. 668–682, https://doi.org/10.5270/OceanObs09.cwp.61.

[25] J.F. Hair, W.C. Black, B.J. Babin, R.E. Anderson, Multivariate Data Analysis, 8th edition, Cengage Learning EMEA, 2019.

[26] T. Dimri, S. Ahmad, M. Sharif, Time series analysis of climate variables using seasonal ARIMA approach, J Earth Syst Sci 129 (1) (2020) 149, https://doi.org/10.1007/s12040-020-01408-x.

[27] S. Vetter, Drought, change and resilience in South Africa's arid and semi-arid rangelands, S Afr J Sci 105 (1/2) (2009) 29, https://doi.org/10.4102/sajs.v105i1/2.35.

[28] C.J. Harmse, H. Gerber, A. van Niekerk, Evaluating several vegetation indices derived from sentinel-2 imagery for quantifying localized overgrazing in a semi-arid region of South Africa, Remote Sens (Basel) 14 (7) (2022) 1720, https://doi.org/10.3390/rs14071720.

[29] M. Faramarzi, K.C. Abbaspour, S. Ashraf Vaghefi, M.R. Farzaneh, A.J.B. Zehnder, R. Srinivasan, H. Yang, Modeling impacts of climate change on freshwater availability in Africa, J Hydrol 480 (2013) 85–101, https://doi.org/10.1016/j.jhydrol.2012.12.016.

[30] Dabrowski, J.J., Rahman, A., Li, M., Shao, Q., Bakar, S., Powell, A., & Henderson, B. (2022). Quality control in weather monitoring with dynamic linear models. https://doi.org/10.48550/ARXIV.2211.04528.

[31] L.R. Centurioni, J. Turton, R. Lumpkin, L. Braasch, G. Brassington, Y. Chao, E. Charpentier, Z. Chen, G. Corlett, K. Dohan, C. Donlon, C. Gallage, V. Hormann, A. Ignatov, B. Ingleby, R. Jensen, B.A. Kelly-Gerreyn, I.M. Koszalka, X. Lin, D. Zhang, Global in situ observations of essential climate and ocean variables at the air–Sea interface, Front Mar Sci 6 (2019) 419, https://doi.org/10.3389/fmars.2019.00419.

[32] B. Fontaine, M. Gaetani, A. Ullmann, P. Roucou, Time evolution of observed July–September sea surface temperature-Sahel climate teleconnection with removed quasi-global effect (1900–2008), J Geophys Res 116 (D4) (2011) D04105, https://doi.org/10.1029/2010JD014843.

[33] J. Yang, J. Zhang, Comparison of oceansat-2 scatterometer wind data with global moored buoys and ASCAT observation, Adv Meteorol 2019 (2019) 1–9, https://doi.org/10.1155/2019/1651267.

[34] J. Swain, P.A. Umesh, A.S.N. Murty, Demonstration of an efficient interpolation technique of inverse time and distance for Oceansat-2 wind measurements at 6-hourly intervals, Int J Ocean Climate Syst 8 (3) (2017) 101–112, https://doi.org/10.1177/1759313117736596.

[35] Q. Wu, Y. Wang, Comparison of oceanic multisatellite precipitation data from tropical rainfall measurement mission and global precipitation measurement mission datasets with rain gauge data from ocean buoys, J Atmos Ocean Technol 36 (5) (2019) 903–920, https://doi.org/10.1175/JTECH-D-18-0152.1.

[36] W. Tang, A. Fore, S. Yueh, T. Lee, A. Hayashi, A. Sanchez-Franks, J. Martinez, B. King, D. Baranowski, Validating SMAP SSS with in situ measurements, Remote Sens Environ 200 (2017) 326–340, https://doi.org/10.1016/j.rse.2017.08.021.

[37] S. Prakash, C. Mahesh, R.M. Gairola, Comparison of TRMM Multi-satellite Precipitation Analysis (TMPA)-3B43 version 6 and 7 products with rain gauge data from ocean buoys, Remote Sens Letters 4 (7) (2013) 677–685, https://doi.org/10.1080/2150704X.2013.783248.

[38] K.P. Bowman, C.R. Homeyer, D.G. Stone, A comparison of oceanic precipitation estimates in the Tropics and Subtropics, J Appl Meteorol Climatol 48 (7) (2009) 1335–1344, https://doi.org/10.1175/2009JAMC2149.1.

[39] Goodwin, T., Quiroz, M., & Kohn, R. (2024). Dynamic linear regression models for forecasting time series with semi long memory errors (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2408.09096.

[40] Zhu, X., Li, F., & Wang, H. (2019). Least squares approximation for a distributed system. https://doi.org/10.48550/ARXIV.1908.04904.

[41] Stata Corporation, Arima—ARIMA, ARMAX, and other dynamic regression models. http://www.stata.com/manuals13/tsarima.pdf, 2013.

[42] A.X. Ezquerro, G. Aneiros, M. Oviedo de la Fuente, Automatic covariates selection in dynamic regression models with application to COVID-19 evolution, in: Proceedings of V XoveTIC Conference. XoveTIC 2022, Kalpa Publications in Computing 14, 2023, pp. 136–138.

[43] H. Demirhan, dLagM: an R package for distributed lag models and ARDL bounds testing, PLoS One 15 (2) (2020) e0228812, https://doi.org/10.1371/journal.pone.0228812.

[44] R. Ren, All the confusion about ARIMA, ARIMAX, transfer function, dynamic regression models, Research Blog, https://ruqinren.wordpress.com/2020/02/21/all-the-confusion-about-arima-arimax-transfer-function-dynamic-regression-models/, 2020.

[45] R.J. Hyndman, Y. Khandakar, Automatic Time Series Forecasting: the **forecast** package for *R*, J Stat Softw 27 (3) (2008), https://doi.org/10.18637/jss.v027.i03.

[46] A. Zeileis, Dynlm: dynamic Linear Regression. R package version 0.3-6. https://cran.r-project.org/web/packages/dynlm/dynlm.pdf, 2019.
[47] G.M. Ljung, G.E.P. Box, On a measure of lack of fit in time series models, Biometrika 65 (2) (1978) 297–303, https://doi.org/10.1093/biomet/65.2.297.
[48] L.J. Tashman, Out-of-sample tests of forecasting accuracy: an analysis and review, Int J Forecast 16 (4) (2000) 437–450, https://doi.org/10.1016/S0169-2070(00)00065-0.
[49] Q. Li, T. Bedi, C.U. Lehmann, G. Xiao, Y. Xie, Evaluating short-term forecasting of COVID-19 cases among different epidemiological models under a bayesian framework, Gigascience 10 (2) (2021) giab009, https://doi.org/10.1093/gigascience/giab009.
[50] R.H. Shumway, D.S. Stoffer, Time Series Analysis and Its applications: With R examples, 4th edition, Springer, 2017.
[51] O.S. Ojo, S.T. Ogunjo, Machine learning models for prediction of rainfall over Nigeria, Sci Afr 16 (2022) e01246, https://doi.org/10.1016/j.sciaf.2022.e01246.
[52] A.B.M. Incoom, K.A. Adjei, S.N. Odai, Rainfall variabilities and droughts in the Savannah zone of Ghana from 1960 to 2015, Sci Afr 10 (2020) e00571, https://doi.org/10.1016/j.sciaf.2020.e00571.
[53] A. Apraku, J.F. Morton, B.A. Gyampoh, Climate change and small-scale agriculture in Africa: does indigenous knowledge matter? Insights from Kenya and South Africa, Sci Afr 12 (2021) e00821, https://doi.org/10.1016/j.sciaf.2021.e00821.
[54] Department of Water and Sanitation (DWS) (2024). National State of Water Report 2023. Integrated Water Resource Studies Report Number WII/IWRS/NSoW 2023 PRETORIA, South Africa.
[55] World Meteorological Organization, State of the Climate in Africa 2023 (WMO-No. 1360), WMO, Geneva, 2024. https://library.wmo.int/idurl/4/69000.
[56] J.N.A. Aryee, F.O.T. Afrifa, K.H. Agyapong, N.G. Frimpong, K.T. Quagraine, P. Davies, Quantifying climatic heavy-precipitation-induced floods in West Africa using multiple precipitation indices, Sci Afr 25 (2024) e02309, https://doi.org/10.1016/j.sciaf.2024.e02309.
[57] Q. Ma, Q. You, Y. Ma, Y. Cao, J. Zhang, M. Niu, Y. Zhang, Changes in cloud amount over the Tibetan Plateau and impacts of large-scale circulation, Atmos Res 249 (2021) 105332, https://doi.org/10.1016/j.atmosres.2020.105332.
[58] J.M.L. Júnior, J.C. Ceballos, S.M.S. da Costa, F.L.L. de Mesquita, H.S. de Jesus, A.R. Gonçalves, A physical parameterization for cloudy-sky downward longwave radiation: validation for tropical and subtropical regions in Brazil, J Atmos Sol Terr Phys (2025) 106512, https://doi.org/10.1016/j.jastp.2025.106512.
[59] K. Rishmawi, S.D. Prince, Y. Xue, Vegetation responses to climate variability in the northern arid to sub-humid zones of Sub-Saharan Africa, Remote Sens (Basel) 8 (11) (2016) 910, https://doi.org/10.3390/rs8110910.
[60] I. Zouitane, M. Ferioun, K. Bouamri, K. Derraz, S.I. Koraichi, M. Idbella, N. El Ghachtouli, Comparative assessment of physicochemical properties, functional diversity and enzymatic activities in three Opuntia ficus-indica soils across diverse climatic regions in Morocco, Sci Afr 28 (2025) e02663, https://doi.org/10.1016/j.sciaf.2025.e02663.
[61] C. Munday, R. Washington, Controls on the diversity in climate model projections of early summer drying over Southern Africa, J Clim 32 (12) (2019) 3707–3725, https://doi.org/10.1175/JCLI-D-18-0463.1.
[62] B.A. Ogwang, H. Chen, X. Li, C. Gao, The influence of topography on East African October to December climate: sensitivity experiments with RegCM4, Adv Meteorol 2014 (1) (2014) 143917.
[63] E. Nkiaka, R.G. Bryant, M. Okumah, F.F. Gomo, Water security in sub-Saharan Africa: understanding the status of sustainable development goal 6, Wiley Interdiscipl Rev: Water 8 (6) (2021) e1552, https://doi.org/10.1002/wat2.1552.
[64] R.J. Van Der Ent, H.H. Savenije, Oceanic sources of continental precipitation and the correlation with sea surface temperature, Water Resour Res 49 (7) (2013) 3993–4004, https://doi.org/10.1002/wrcr.20296.
[65] T. Vischel, G. Panthou, P. Peyrillé, R. Roehrig, G. Quantin, T. Lebel, C. Wilcox, F. Beucher, M. Budiarti, Precipitation extremes in the West African Sahel: recent evolution and physical mechanisms. Tropical Extremes, Elsevier, 2019, pp. 95–138.
[66] S.E. Nicholson, Climate and climatic variability of rainfall over eastern Africa, Reviews of Geophysics 55 (3) (2017) 590–635, https://doi.org/10.1002/2016RG000544.
[67] R.E. Said, H.E. Hasieb, M.A. Khaled, E. Mohammed-AbdAllah, A.G. Osman, Regional variability in remotely sensed data with respect to the distribution of some snapper fishes (Family: lutjanidae) between the Red Sea and the Arabian Gulf, Sci Afr 23 (2024) e02104, https://doi.org/10.1016/j.sciaf.2024.e02104.
[68] C.J. Willmott, K. Matsuura, Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance, Clim Res 30 (1) (2005) 79–82, https://doi.org/10.3354/cr030079.
[69] World Health Organization, Operational Framework for Building Climate Resilient Health Systems, World Health Organization, 2015.
[70] J. Huang, M. Ji, Y. Xie, S. Wang, Y. He, J. Ran, Global semi-arid climate change over last 60 years, Clim Dyn 46 (3–4) (2016) 1131–1150, https://doi.org/10.1007/s00382-015-2636-8.