Theoretical and Applied Climatology
https://doi.org/10.1007/s00704-019-03016-2

ORIGINAL PAPER

Non-homogeneous hidden Markov model for downscaling of short rains occurrence in Kenya

Aston Matwayi Nyongesa[1] · Gang Zeng[1] · Victor Ongoma[2]

Received: 14 January 2019 / Accepted: 27 September 2019

Abstract

Non-homogeneous hidden Markov model (NHMM) is applied in modeling of daily rainfall occurrences across 16 synoptic stations in Kenya. The time series of the data sets was during the October–December (OND "short rains") season from 1979 to 2005. The tool assumes that the diurnal rainfall events at a network of observing stations are influenced by unobserved states, that is, "weather states." These states' evolution is modeled based on a first-order Markov criterion with state-to-state transition probabilities conditioned on some atmospheric variable indices. The five states are selected using the Bayes information criterion (BIC). To downscale daily rainfall occurrences across 16 stations, a NHMM employed global circulation model (GCM) projection outputs for daily precipitation and sea surface temperatures during the study period. The interannual variability of the mean GCM simulated precipitation and mean historical stations rainfall depicts a weak correlation though significant at 90% confidence level. Thus, it implies that GCM-NHMM simulations do not simulate the rainfall occurrences well. The consecutive wet spell length between the historical rainfall datasets and GCM-NHMM simulated precipitation for 90-day frequencies shows a strong positive correlation significant at 95% confidence level. The findings from this study reveal that the modeling tool is suitable for statistical downscaling of daily rainfall occurrences at multisite stations network. The statistical inference from the model is applicable for drought/flood preparedness, water resource management, and inputs into crop models.

1 Introduction

Climate change poses a significant threat across the globe. The changing climate has led to an intensification of extreme events (Aerenson et al. 2018). Extreme events are associated with loss of lives and vast destruction of property, especially in developing countries. Drought, for instance, is a threat to the agribusiness sector which relies heavily on rainfall. Excessive rainfall, on the other hand, is a challenge too because it leads to flooding, runoff, and general infrastructure destruction. For instance, the drought experienced in January 2014, in Kenya, affected close to 1.6 million of their population (IFRC 2015). The situation was felt in the marginal agricultural and pastoral livelihood regions like the Northern, North Eastern, North Western, South Eastern, and some Coastal areas. The southeast and northwest pastoral regions continued to suffer from food insecurity even during the October to December rains, since La Niña resulted to below-average rains (Funk et al. 2018). To reduce the community's vulnerability to the effects of such extreme events, proper preparedness mechanisms are paramount. Such measures include prior planning of necessary resources for such impending disasters. To realize this, modeling of climatic variables like rainfall, which is essential to rain fed agriculture, needs to be undertaken and the predictions of its likelihood of occurrences quantified (Chen et al. 2013).

To understand the probabilistic structure of precipitation, stochastic models have previously been applied in many studies. Such models have been used to simulate inputs for agricultural crop models, runoff, water resource management, hydrology models, and other fields (Vrugt et al. 2008; Steduto et al. 2009; Maraun et al. 2010). However, previously, the said method did not take into consideration the atmospheric aspects that influence precipitation formation, perhaps due to scanty atmospheric data sets. Thus, such simulations failed

✉ Gang Zeng
zenggang@nuist.edu.cn

[1] Key Laboratory of Meteorological Disaster of Ministry of Education (KLME), Collaborative Innovation Center on Forecast and Evaluation of Meteorological Disasters (CIC-FEMD), Nanjing University of Information Science and Technology, Nanjing, China

[2] School of Geography, Earth Science and Environment, The University of the South Pacific, Laucala Campus Private Bag, Suva, Fiji

Published online: 09 December 2019

to model precipitation adequately. Hence, Gabriel and Neumann (1962) modeled wet and dry daily rainfall instances at a rainfall station in Israel using a homogeneous Markov chain transition matrix. To show seasonal variations, the Markov chain was extended by transforming the transition probabilities using a Fourier series (Stern and Coe 1984). Mechanistic models were also developed by scholars such as Le Cam (1961) who used cluster point criteria to discuss cyclonic rain storm as having "bands." Other researchers, including Waymire and Gupta (1981), studied the point approach method.

Even though these models were used, they failed to incorporate atmospheric parameters. In the recent past, advances in quality data collection and modernization of meteorology have led to a better understanding of Global circulation models (GCMs). Although these models perform better than the stochastic tools, their spatial resolutions are coarse. This implies that simulating a local area with complex terrain and with a smaller spatial scale than grid box using GCM leads to a poor representation of climate (Schubert 1998; McAvaney et al. 2001).

In order to optimize the quality of present atmospheric data to address the challenges above, downscaling is necessary. The statistical downscaling techniques are widely used since it can be applied easily to different GCMs, regions, scales, and are inexpensive (Timbal et al. 2003; Wilby et al. 2004; Wood et al. 2004). Statistical downscaling entails development of the quantitative relationship between a small/local space weather parameters (predictands) with the large-space meteorological variables (predictors), by applying analog methods (circulation typing), regression methods, or soft computing techniques like neural networks (Cannon and Whitfield 2002). Non-homogeneous hidden Markov model (NHMM) allows simulation of rainfall at individual station as well as comparison with historical (observed) rainfall records. To estimate future smaller scale climatology, less than 30 km, the NHMM model employs future GCM projections to run the statistical model.

Statistical downscaling has limitation in that it is inappropriate for use with sparse data and where relationships between predictands and predictors show variations. The common statistical downscaling scheme uses predictand as a function of the predictor. Nonetheless, other relations have been applied. This approach has been used in various studies both in the tropics and mid-latitudes. For instance, Kang and Kim (2010) using a number of dynamical and statistical models assessed *Madden–Julian oscillation* (*MJO*) predictability for boreal winter. To investigate sub-seasonal to interdecadal variability of the Australian monsoon, Robertson et al. (2005) used HMM statistical tool over North Queensland. Applying a hidden Markov model tool, Yoo et al. (2010) studied the Asian summer monsoon's variability in interannual and intraseasonal timescales. Similarly, Guo et al. (2018) used HMM tool to analyze the flood-season rainfall pattern as well as its temporal changes over East China.

To design a statistical downscaling model, four methods were inferred. Firstly, choosing a statistical downscaling method and selection of GCM is done. Then it is preceded by determining a relevant predictor variable in order to base an understanding of local and regional driving factors as well as the appropriate GCM model. To reduce GCM predictors mean and variance biases about observations, data standardization is done on regional climate at a locality using GCMs. The large space atmospheric variable may, for example, represent large-scale circulation flow patterns over a vast region like the tropics while the small scale may represent monthly or daily precipitation from weather observing station. Some studies have utilized products from GCMs to infer various atmospheric parameters. An example is Zeng et al. (2014) that used simulations from phase three of Climate Model Intercomparison Project (CMIP3) to analyze summer precipitation changes in North China and over Yangtze River valley. In a recent study, Ongoma et al. (2018) used CMIP5 simulations to project rainfall and temperature over East Africa. The findings of the study may be used for future planning despite the fact that GCMs do not perform well in reproducing rainfall over the region (Ongoma et al. 2019; Rowell 2019).

A NHMM is a tool for statistical downscaling, which have been used widely in Australia and South America (Zucchini and Guttorp 1991; Hughes and Guttorp 1994; Hughes et al. 1999; Bellone et al. 2000; Charles et al. 2004; Robertson et al. 2004a). However, in Africa, the NHMM has not been applied extensively for such purposes. Through a number of hidden states, NHMM tool relates daily precipitation data in a network of rain gauge stations to global atmospheric patterns. These regimes' evolution is then simulated following a first-order Markov state-to-state transition events set on atmospheric predictor indices. This tool is useful in conceptualizing statistics of day to day rainfall events at station level regarding large scale atmospheric patterns, as well as generating in situ daily rainfall sequence occurrences in a given region for crop models simulation inputs, for water, natural resources management, and environmental protection.

Despite irrigation practices in the region, rain fed agriculture stands in as key economic driver of the agricultural economy in Kenya. Hence, precipitation is paramount. Thus, more accurate weather forecasting and prediction are necessary for water resource management, food security, and environmental protection in the region for sustainable development. Thus, projecting precipitation occurrences during the October–December (OND "short rains") season will be of help in planning especially for the Arid and Semi Arid Land (ASAL) regions where short rains is the main season. The current study aims at downscaling rainfall using NHMMs, examining short rain events of daily rainfall observed at a network of stations

in Kenya, and inferring its variability to large-scale atmospheric systems.

2 Study area, data, and methods

2.1 Study area

The study area is confined within longitude 33° E to 43° E and latitude 5° S to 5° N (Fig. 1). The field of study covers about 582,650 km^2 and is characterized by highly variable topographic features.

The climate of Kenya is a typical equatorial. The short rains pattern is driven by the East African monsoon that brings air with warm and moist conditions. The annual mean monthly temperature ranges between 19 and 24 °C in July (coldest) and March (warmest), respectively (Ongoma et al. 2017). Kenya experiences two major rainfall seasons: "long rains" from March to May (MAM) and "short rains" during October to December (OND) (Ongoma and Chen 2017; Mumo et al. 2018). Annually, mean rainfall is about 2000 mm (Ongoma and Chen 2017).

2.2 Data

2.2.1 Observed data

Observed rainfall datasets were provided by the Kenya Meteorological Department (KMD). Out of the 33 synoptic stations distributed across the country, 16 stations' data for the period 1979–2005 for OND season were used. The 16 meteorological stations were chosen because of their data continuity, homogeneity, and quality. Thus, every 27 years (1979–2005), we had a time series of 92 days. A few missing data (less than 3%) in a station were in-filled by averaging adjacent homogeneous stations during the same period. The study considered the reliable time series based on the standard normal homogeneity test in XLSTAT (Addinsoft 2016).

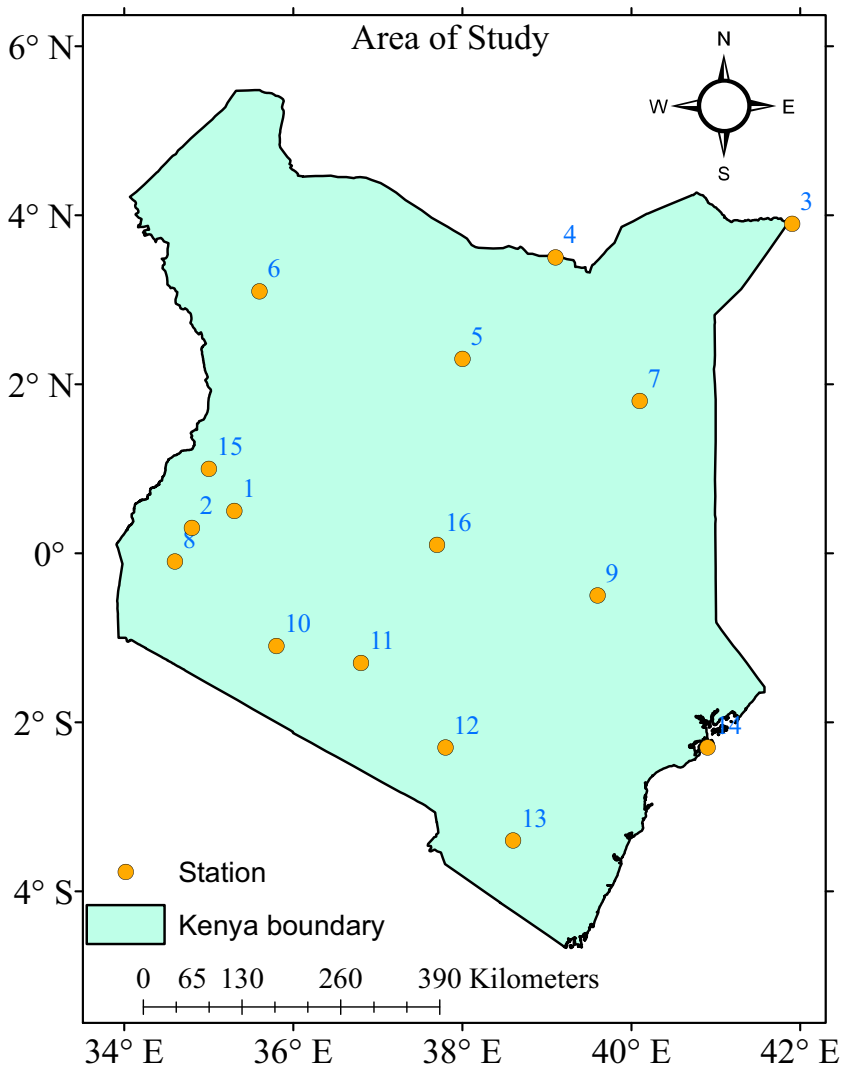

Fig. 1 Map of the study area indicating the spatial distributions of synoptic stations in Kenya. The atmospheric data grid is extended over the entire study domain

2.2.2 Reanalysis data/atmospheric data

The reanalysis products used in our study are the interim European Centre for Medium-Range Weather Forecasts, Re-Analysis data (Dee et al. 2011) and the National Centers for Environmental Prediction-National Center for Atmospheric Research (NCEP/NCAR) Reanalysis I (Kalnay et al. 1996) from 1979 to 2005 in all the cases.

ERA-Interim reanalysis data Large-scale atmospheric variables were derived from the European Centre for Medium Range Weather Forecast (ECMWF) ERA Interim reanalysis data on a 1° × 1° horizontal resolution for the same study period as the observation data to be used for model validation. The variables used include outgoing longwave radiation (OLR), zonal and meridional wind speeds (u and v) at pressure levels (200 and 850 hPa). The predictor domain extended from longitude 10° E to 80° E and latitude 20° S to 20° N to cover the entire research field.

NOAA_ERSST_V4 data Global sea surface temperatures for the same study period (1979–2005; OND) were derived. The version 4 of Extended Reconstructed Sea Surface Temperature (ERSSTv4; Huang et al., 2014) data sets were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their website at http://www.esrl.noaa.gov/psd/.

2.2.3 Global climate model simulations/model data

Ensemble simulation for ECHAM6 Atmospheric Model Intercomparison Project (AMIP) precipitation for the entire study period and domain was used during the analysis. Unlike ECHAM5, the ECHAM6 model is run at a higher vertical resolution which constitutes the upper stratosphere-troposphere with improved precision. The ECHAM6 is forced with the observed historical SSTs.

2.3 Methods

The modeling tool—hidden Markov model (HMM) (Robertson et al. 2004a) enables for modeling daily rainfall amounts as well as the occurrences on many rainfall stations networks. The machine models the observed rainfall by putting in place minute number of hidden rainfall states. The underlined regimes make it possible for the exposition of observed rainfall changes regarding discrete rainfall patterns. Nonetheless, the states (k) in question are not directly visible to the observer hence referred to as "hidden." HMM follows a Markov chain whereby an active "today" state relies solely on the active state "yesterday" basing on transition probabilities. The HMM allows for the simulation of rainfall at individual station in the investigative zones and comparing it with historical (observed) rainfall records for the entire period. From the analyses, important statistical features like wet or dry spell lengths and rainfall probabilities can be evaluated. Thus, the model can be used for generation of large rainfall data sets for input into a water resource simulation, crop management model, and statistical analysis and so forth.

Applying Hughes and Guttorp (1994), hidden Markov model contains two conditional statements. To begin with, we assume more than one rainfall observations R_t during time t are unconstrained of the rest factors in the model till time ; contingent on weather state S_t at time t is expressed as presented in Eq. 1:

$$P(R_t|S_{1:t}, R_{1:t-1}) = P(R_t|\, S_t) \tag{1}$$

The second presumption is that the hidden state process $S_{1:T}$ follows the first-order Markov process.

$$P(S_t|S_{1:t-1}) = P(S_t|S_{t-1}) \tag{2}$$

Also, we assume that the first-order Markov process is time homogeneous; that is, the $k \times k$ probability transition matrix is not altered with a shift in time as in Eq. 2.

States $\boldsymbol{S_1}, \ldots, \boldsymbol{S_T}$ correspond to latent weather states while output vectors $\boldsymbol{R_1}, \ldots, \boldsymbol{R_T}$ are daily precipitation occurrences for the network (Fig. 2).

To make $P(\boldsymbol{R_t}|\, S_t)$ easier, we assume that the observed rainfall in every station during time t is expressed as given in Eq. 3:

$$P(\boldsymbol{R}_t = \boldsymbol{r}|S_t = \mathrm{s}) = \prod_{m=1}^{M} P\left(R_t^m = r|S_t = \mathrm{S}\right) = \prod_{m=1}^{M} p_{smr} \tag{3}$$

where r ∈ {0, 1}, each $p_{smr} \in [0, 1]$ and $p_{sm0} + p_{sm1} = 1$

To connect multisite rainfall occurrence with GCMs details, a NHMM (Hughes and Guttorp 1994) is developed by introducing input parameters to HMM. Under NHMM, the state transition matrix is not stationary. The NHMM tool breaks the temporal and spatial diurnal rainfall via discrete weather states over a gauge station network. In the whole gauge stations, every state forms a set of rainfall distributions and probabilities. Following the same approach as in HMM, we let X_t be N-dimensional column vector of predictors for day t, obtained from meteorological variables. By $X_{1:T}$, we denote the sequence $X_1, \ldots, X_T$. Basing on the homogeneous HMM, we substitute Eq. 2 with Eq. 4:

$$P\left(S_t|S_{1:t-1}, X_{1:T}\right) = P\left(S_t|S_{t-1}, X_T\right) \tag{4}$$

The unobserved state on day t is dependent on the value of unobserved rainfall regime S_{t-1}, on day $t-1$ as well as the predictor column vector X_t for day t. Since X_t varies in time, this leads in transition probabilities connecting states that change due to changes in X, hence an inhomogeneous HMM model. The function of X_t is as in the graphical model (Fig. 3). Here, arrows represent model parameters that are estimated from rainfall data sets.

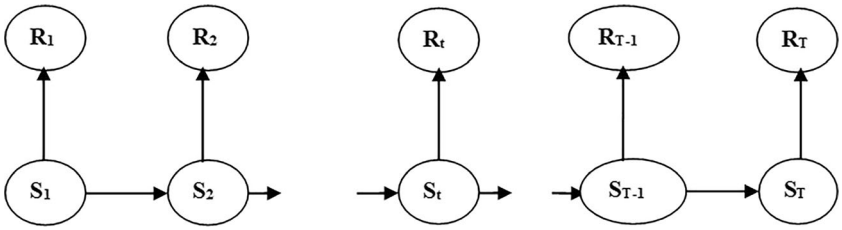

Fig. 2 A graphical model of HMM

States $\boldsymbol{S_1}, \ldots, \boldsymbol{S_T}$ correspond to weather states while output vectors $\boldsymbol{R_1}, \ldots, \boldsymbol{R_T}$ are daily rainfall occurrences for the network; $\boldsymbol{X_1}, \ldots, \boldsymbol{X_T}$ are vectors of atmospheric variables.

Multinomial logistic regression is applied in Eq. 5 to model the hidden state transitions:

$$P\left(\boldsymbol{S}_t = i|S_{t-1} \quad = j, \boldsymbol{X_t}{=}\mathbf{x}\right) = \frac{\exp\left(\sigma_{ji} + \boldsymbol{\rho}_i^{'}\mathbf{X}\right)}{\sum_{k=1}^{K}\exp\left(\sigma_{jk} + \boldsymbol{\rho}_k^{'}\mathbf{X}\right)} \tag{5}$$

The first specific hidden state in the sequence $\boldsymbol{S}_1$, is denoted as in Eq. 6:

$$P\left(\boldsymbol{S}_t = i|S_{t-1} \quad = j, \boldsymbol{X_t}{=}\mathbf{x}\right) = \frac{\exp\left(\lambda_i + \boldsymbol{\rho}_i^{'}\mathbf{X}\right)}{\sum_{k=1}^{K}\exp\left(\lambda_k + \boldsymbol{\rho}_k^{'}\mathbf{X}\right)} \tag{6}$$

The λs and σs comprise real valued parameters whereas ρs are real-valued parameter vectors in N-dimension, where the prime indicates vector transpose. The baseline transition matrix was multiplied by atmospheric predictors like the parameterization derived in previous studies (Hughes and Guttorp 1994; Hughes et al. 1999)

$$P\left(\boldsymbol{S}_t = i|S_{t-1} \quad = j, \boldsymbol{X_t})\alpha = P\left(\boldsymbol{S}_t = i|S_{t-1} \quad = j\right)P(X_t|S_{t-1,} = j, S_t = i)$$

$$= \gamma_{ji}\exp\left[-\frac{1}{2}\left(\mathrm{X}_t{-}\boldsymbol{\mu_{ji}}\right)^{'}\mathbf{V^{-1}}\left(\mathrm{X}_t{-}\boldsymbol{\mu_{ji}}\right)\right]$$

$$\mathrm{A} = \exp\left[\left(\ln\gamma_{ji}{-}\frac{1}{2}\boldsymbol{\mu}_{ji}^{'}\mathbf{V^{-1}}\boldsymbol{\mu_{ji}}\right) + \boldsymbol{\mu}^{'}_{ji}\mathbf{V^{-1}}\mathrm{X}_t\right] \tag{7}$$

The parameter $\boldsymbol{\mu_{ji}}$ is mean of the atmospheric variable linked with transitions from state j to i at day $t-1$ to day$t$. Simplifying and equating$P(X_t|S_t, S_{t-1,})$ to $P(X_t|S_{t,})$, then $\boldsymbol{\mu_{ji}} = \boldsymbol{\mu_i}$, Eq. 6 will be equal to Eq. 5. To elaborate this, we set $\boldsymbol{\rho_i} = \mathbf{V^{-on1}}\boldsymbol{\mu_{ji}}$ and$\boldsymbol{\lambda_i} = \ln\gamma_{ji}{-}\left(\frac{1}{2}\right)\boldsymbol{\mu}_{ji}^{'}\mathbf{V^{-1}}\boldsymbol{\mu_{ji}}$. In order to have statistical identifiability, the transition parameters, σ_{j1}, $\boldsymbol{\rho_i}$, and $\boldsymbol{\lambda_1}$ are set to zero. The maximum likelihood technique is used to learn the parameters that best fit the conditional probability of the observed data sets.

$$L(\boldsymbol{\Theta}) = P(\boldsymbol{R}_{1:T}|\boldsymbol{X}_{1:T}\boldsymbol{\Theta})$$

$$= \sum_{S_{1:T}} P\left(S_1|\boldsymbol{X_{1,}}\right)\prod_{t=2}^{T}P(S_t|S_{t-1},\boldsymbol{X_t})\prod_{t=1}^{T}P(\boldsymbol{R_t}|S_t) \tag{8}$$

The non-homogeneous characteristic allows transition probabilities to change from state to state thus governing external feed (GCM details) which impact the evolution of rainfall features. In our study, ECHAM6 AMIP daily precipitation for OND season has been used in downscaling expected daily rainfall occurrences across stations under investigation (Robertson et al. 2004a, 2009; Verbist et al. 2010; Pineda and Willems 2016). Therefore, to check on OND expected rainfall events in Kenya, we employed GCM-NHMM to conduct probabilistic modeling. This study used principal component analysis (PCA) and the first leading PCs that accounted for more than 10% of ECHAM6 uncoupled precipitation were selected. To take care of sub-seasonal variability and capture the seasonal cycle of precipitation, GCM variance was filtered. After that, we trained the NHMM in cross-validated mode, as described under HMM using the historical station rainfall alongside with the two daily principal components as the inputs, to generate 50 daily rainfall simulations. The 50 simulations were then averaged seasonally at 16 stations.

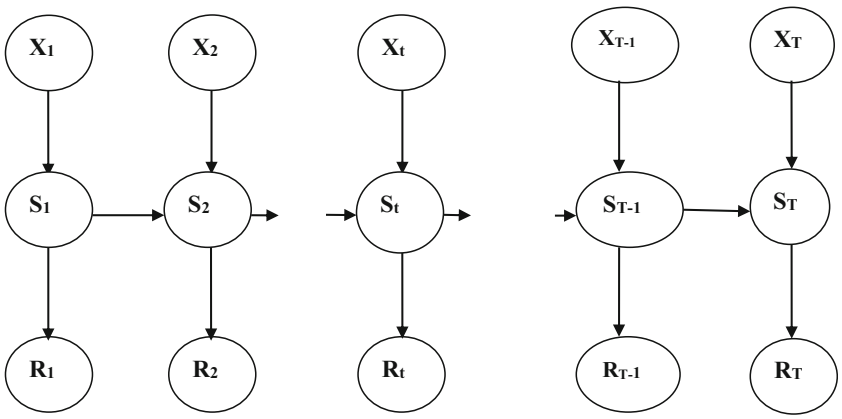

Fig. 3 A graphical model of a NHMM

3 Results and discussion

3.1 Estimation of model parameters

Having chosen the 5-state model, its parameters were estimated over the entire rainfall record. Daily rainfall observations sourced from 16 meteorological stations for OND season for period 1979–2005 are considered in this analysis. The longitudinal, as well as latitudinal stations' information, is utilized during the fitting process. Applying Robertson et al. (2004a), cross-validation was done to ascertain the quality regarding log-likelihood of the fitted number of states. We learn the expected maximization (EM) algorithm by restarting it 20 times from zero initial seeds and utilized the one consisting of highest log likelihood. The out of sample value fork = 1 to 9 was normalized (scaled) and plotted as in Fig. 4. The model used was an independent delta exponential distribution at zero. From the normalized values, both the Bayes information criterion (BIC) and exponential function converges/flatten at (near) state 6. However, state 6 did not improve the model performance; hence we settled for k = 5 in our study. Normalization of the state was accomplished using the following approach; for k states HMM model, the BIC scores was defined as given in Eq. 9

$$BIC_k = 2L\left(\Theta_k^*\right) - plogT \tag{9}$$

where Θ_k^* represents the maximum likelihood parameter vector and is derived by expected maximization upon training k states model; $L\left(\Theta_k^*\right)$ is the model's likelihood evaluated at Θ_k^* as shown in Eq. 8; p is the linear parameter's number used in each k state model whereas T is the sum of the days used in training the model. For more complex model, the term $-plogT$ is used to subject a "penalty." From Kass and Raftery (1995), BIC can be perceived as an example approximation to the real Bayes factor for model sampling though the computation is complex.

As highlighted by Hughes and Guttorp (1994), BIC can provide meaningful information whereby models are backed up by the data though the theoretical aspect of selecting a NHMMs and HMMs model is partially substantiated by Hughes et al. (1999). To derive normalized BIC scores that are of more less the same scale like a standardized log- likelihoods, we substituted BIC_k in equation 9 with $\frac{BIC_k}{2N}$ where N represents total of binary predictions made, that is, $N = 27 \times 92 \times 16$.

The self-transitions (shown in bold face) are quite large showing that the states are persistent, with states 1, 5, and 2–4 being the most and least persistent states respectively (Table 1). There are some rare direct transitions between states 1 and 2, 2 and 4, with states 3 acting as an intermediary. A clear transition direction is lacking through the states. The cells with less than 0.10 probabilities are italicized.

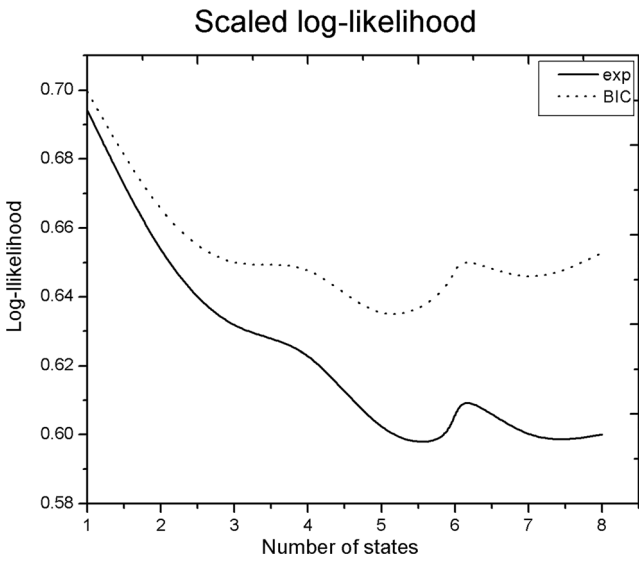

Fig. 4 AHMM Scaled log likelihood

3.2 Representation of HMM state

Having selected a 5-state model, its parameters were estimated from whole 2484 daily rainfall record. The EM algorithm was restarted 50 times, choosing the log likelihood with the highest run. The resultant rainfall parameters regarding amount and probabilities are shown in Fig. 5. In state 1, rainfall is enhanced in the western, central, and coastal regions. Stations in Northwestern, eastern and northeastern regions received depressed rainfall, mostly characterized by light rains (drizzle). State 1 probabilities amount are high along the rift valley stations. Under state 2, rainfall distribution and probabilities show some spatial resemblance. However, the coastal and western regions rains are suppressed. The wettest state, 4, exhibits an almost similar spatial distribution of rainfall occurrence and amount with least values recorded on the northeastern regions. State 2 is the driest state characterized by rainfall events in all stations. However, a comparison between its average amounts and probabilities is lower to state 5, in some stations. The fifth state is perfectly homogeneous in terms of

Table 1 HMM states transition probability matrix

Transition probabilities between HMM hidden states					
	To state				
From state	1	2	3	4	5
1	**0.73**	*0.01*	0.13	*0.06*	0.07
2	*0.02*	**0.82**	*0.08*	*0.01*	*0.07*
3	*0.09*	0.10	**0.74**	*0.03*	*0.04*
4	0.11	*0.02*	*0.09*	**0.70**	*0.08*
5	*0.06*	*0.04*	*0.04*	*0.03*	**0.83**

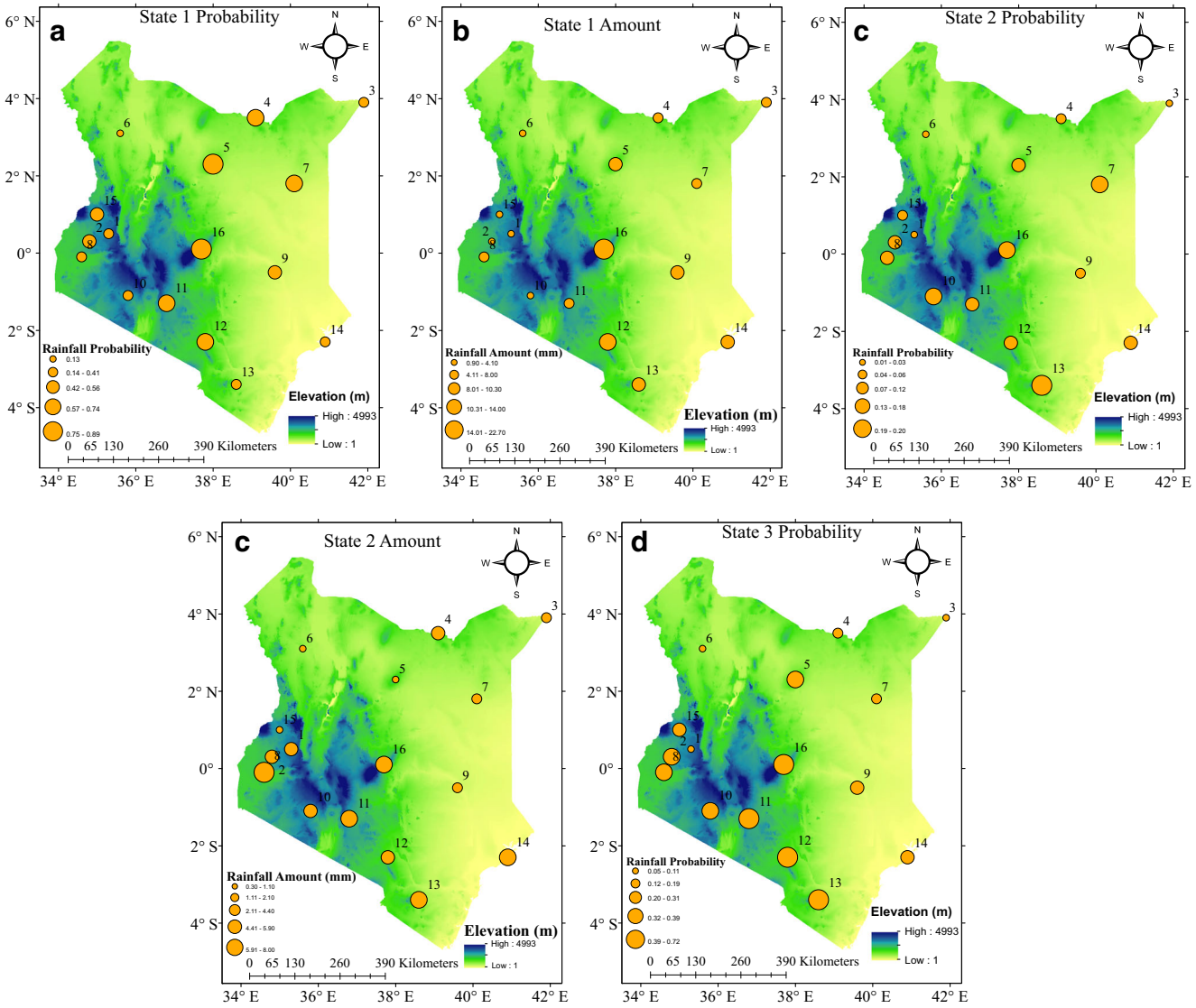

Fig. 5 **a**–**j** Five-state HMM rainfall occurrence probabilities (**a**, **c**, **e**, **g**, **i**) as well as mean amount rainfall (**b**, **d**, **f**, **h**, **j**). The 5 HMM states are derived from the mixed exponential distribution parameters. The background color scale shows the digital elevation on the region

probability and amounts; however, its probabilities are lower than state 3. States 3 and 5 rainfall probabilities show some tendencies of correlation with rainfall intensity; states 1, 2, and 4 do not show such correlations. The state transition matrix is shown in Table 1. The table depicts persistence that highlights the Markov characteristic of HMM. The respective wet and dry states 3 and 4 are most persistent.

3.3 The estimated state sequence

The probable 5-state sequence over 27 fall (OND) period was estimated by using the Viterbi algorithm, a dynamic programming algorithm (Forney 1973). The estimated sequence makes it possible for an interpretation of observed rainfall record regarding these states and atmospheric features that are linked to an individual state. Figure 6 shows the estimated state sequence for 27 seasons. It depicts a marked variability in both subseasonal and interannual variations. The mean interannual variability is shown in Fig. 8. The dry state 4 dominated the 1980s whereas the wet state is well distributed. State 1 has a similar seasonality with state 3, but its frequency is less than state 3. States 2 and 5 dominated depicting a similar seasonal frequency though less than states 3 and 1 but more than state 4.xx

We analyzed the average seasonality of the rainfall state occurrence. The outcomes in Fig. 7 indicate the characteristic trends of the five states. State frequency 1 as well as 3 decreases from early October to mid-November and then gradually peaks in prevalence till late December. States 2 and 5 increase from the beginning of the season and maximize around mid-November followed by significance decline to late December. State 4 showed a less similar characteristic like in states 2 and 5. However, it consisted of bimodal peaks

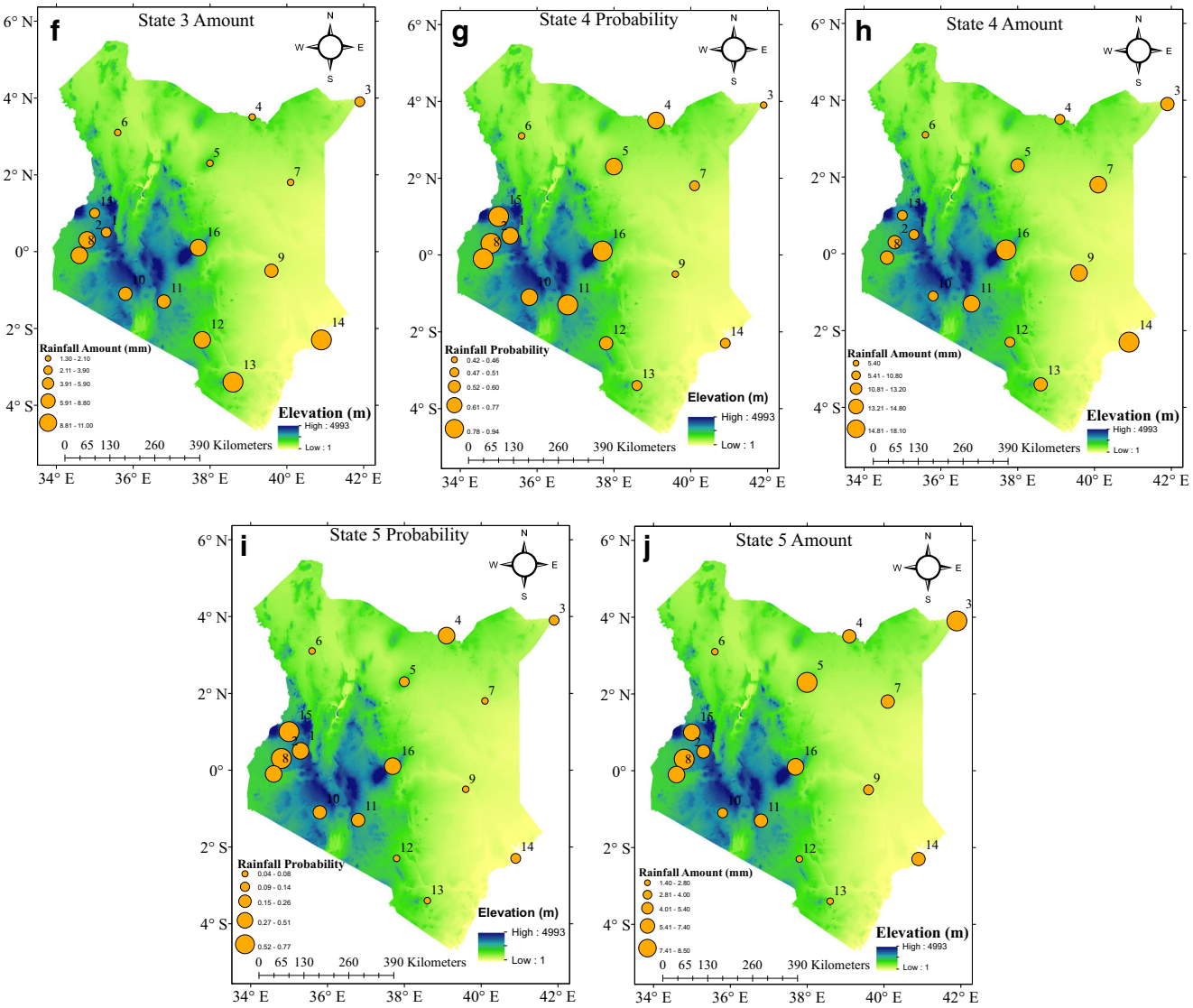

Fig. 5 (continued)

during late October and mid-November. Additionally, it showed stagnancy in seasonal state frequency during the second to third week in October and its frequency, both during the peak and trough were slightly lower than the other states (Fig. 8). The states 1 to 5 minimum-maximum ratios from the 27 years' 10 days moving average are 0.44, 0.49, 0.85, 0.33, and 0.45, respectively. Within the study season, the state frequency varying factor is between 1and 3 and is a direct implication of the non-uniformity in the HMM. The interannual variability for the state sequence was accomplished by cumulating the days during the OND as in the estimated sequence in Fig. 6.

There is evidence of the major interannual variations occurrences, in states 1 and 3, which is consistent with their rainfall and the composites. State 3 occurrence frequencies peaked during 1986 and around 2002, while states 2, 4, and 5 show little interannual changes. El Niño (La Niña years) tend to be linked with more (less) of state 1 as compared to state 3.

3.4 Prevailing synoptic conditions

To check the dynamics of synoptic conditions, composite analysis was employed. Our work slightly differed from previous studies in that the indices used for classifying the El Niño or La Niña events are based on monthly rainfall events for each year for all the states. We considered the top 15% (5 seasons) of the most prevalent standard deviations from the mean to discern the El Niño or La Niña phenomena (Robertson et al. 2004b). The approach was deemed suitable since it captures the El Niño-Southern Oscillation (ENSO) events. Upon forming a composite and respective ENSO event chosen, these phenomena years are normalized in all the states.

The composite analysis was done on ERA interim analysis wind (zonal and meridional) at 200 hPa (Fig. 9) and 850 hPa pressure levels as well as OLR. This was done to

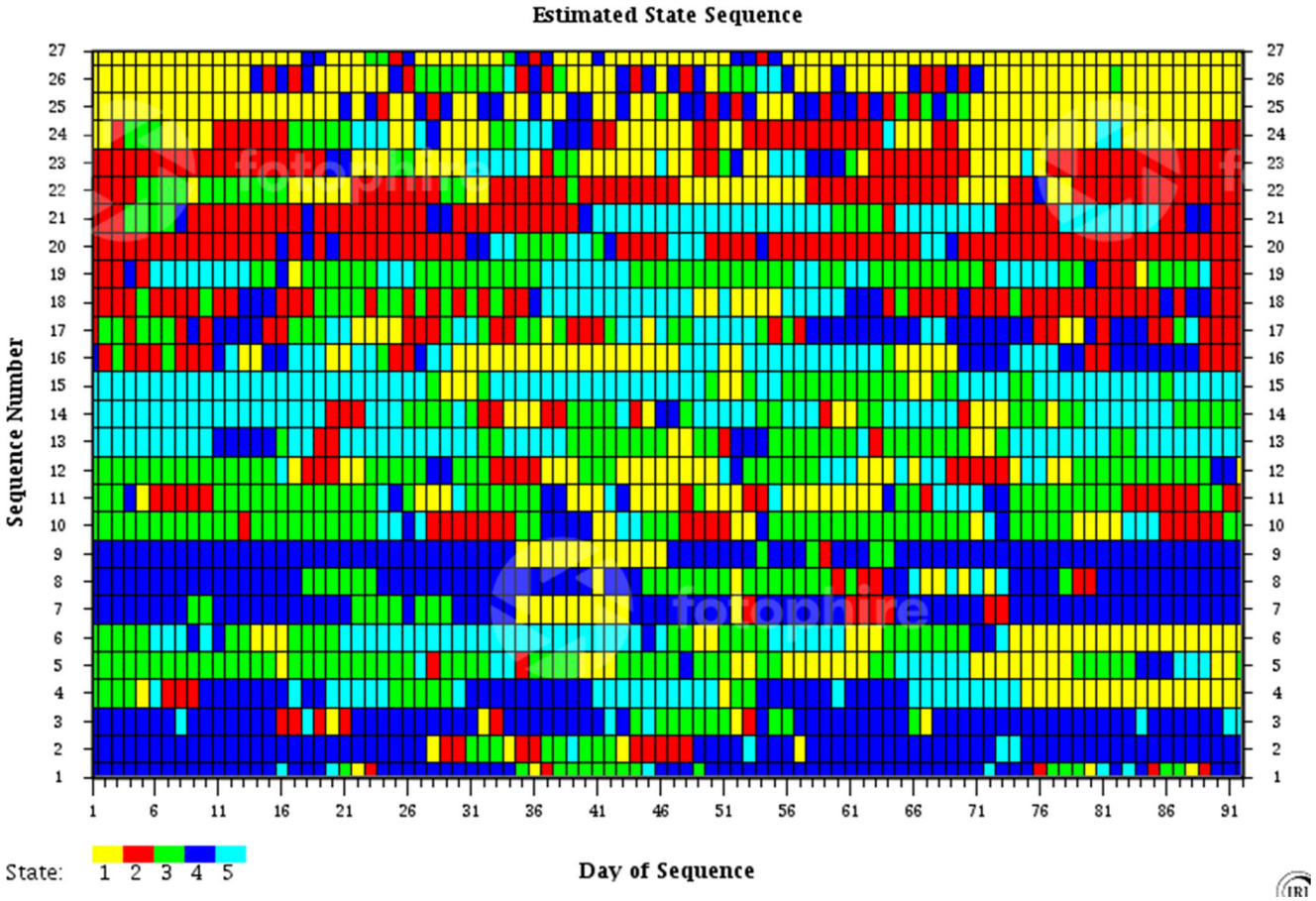

Fig. 6 The estimated HMM state sequence. Colors are running from green (wettest) to dark blue (driest). The numbers of days for particular state 1 to 5 are 367, 541, 606, 577, and 393 respectively

analyze meteorological features linked to the five rainfall states. The wind and OLR anomalies within the tropics for the two pressure levels and the individual rainfall states were plotted during the study/OND season. The mean seasonal anomalies (deviations from the mean) for the three variables were standardized prior to plotting.

The wind at 200 hPa level is paramount weather indicators since they denote regions of convergence and confluence of westerly winds and it determines how the circulation pattern between the surface and upper level are interconnected. Convergence below with divergence above leads to a convectional activity that maintains the frontogenesis activities. Most

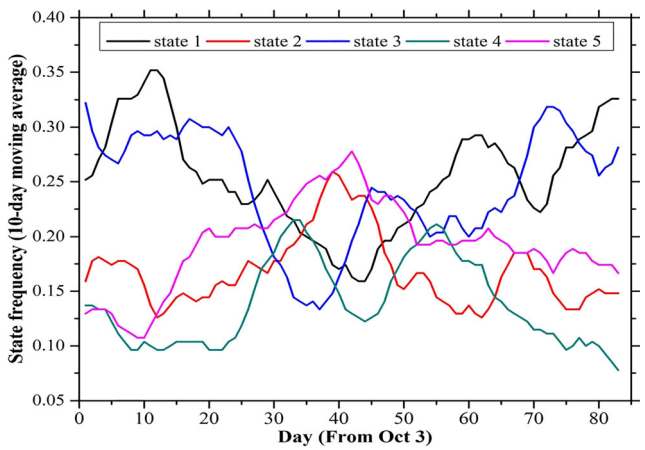

Fig. 7 Seasonal state frequency cycle

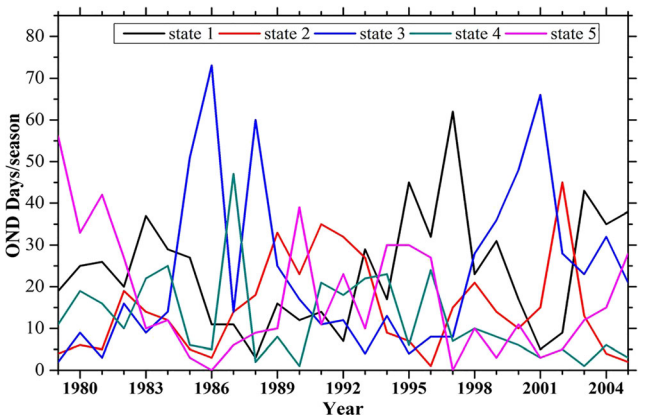

Fig. 8 Interannual state frequency variability in HMM. The interannual variability of state occurrence based on days assigned to the individual state

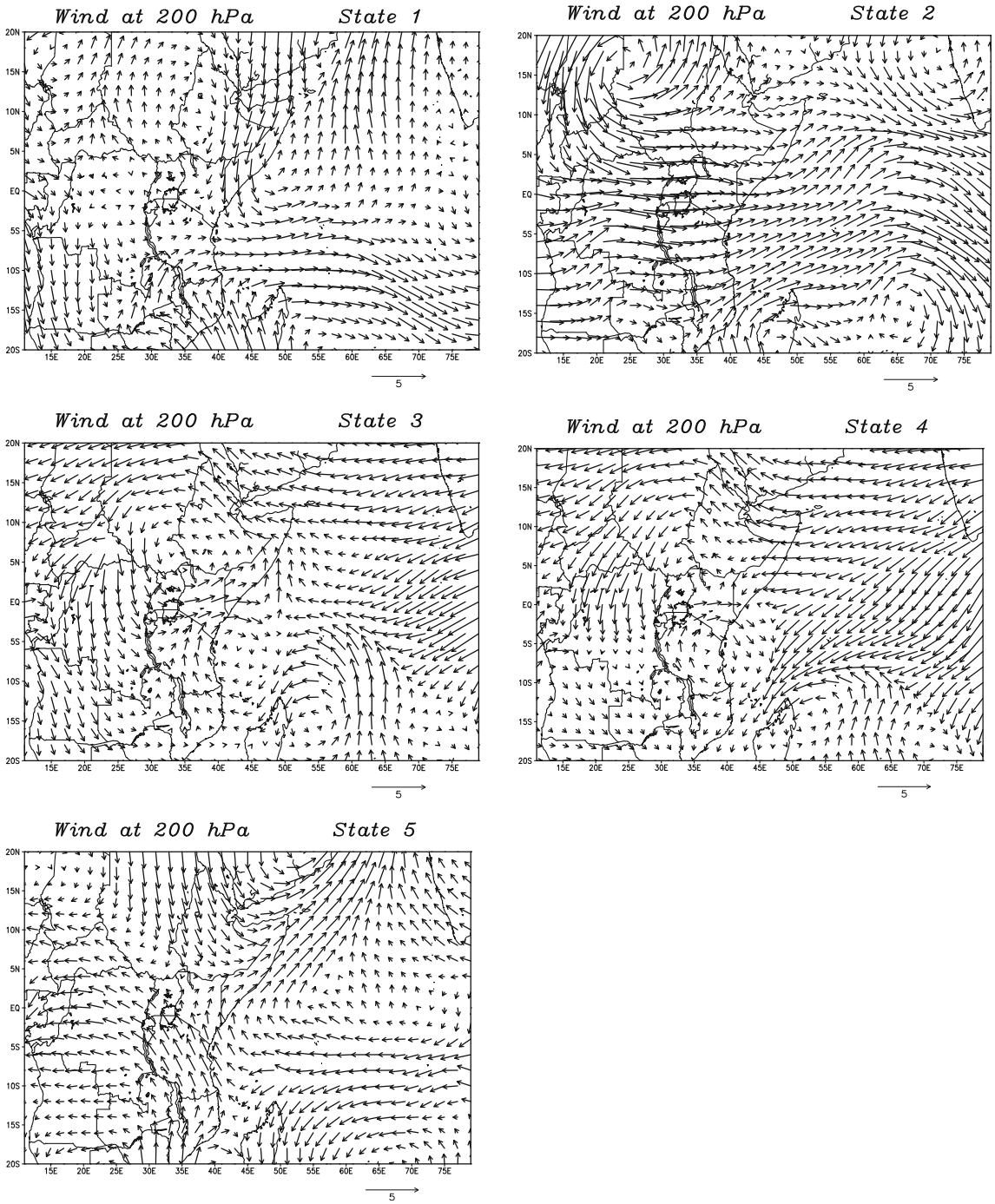

Fig. 9 Composites anomalies concerning the OND climatological mean. Climatological average, for days assigned to every state. The arrows indicate 200 hPa winds. Arrow scale is given below each panel (m/s). The NHMM composite days used are 371, 544, 611, 582, and 376 for states 1, 2, 3, 4, and 5 sequentially

winds at upper level depicted a westerly flow characteristic especially in state 3 (Fig. 9). However, in states 1 and especially state 4 (Fig. 9), the winds are more diffluent. To further probe the synoptic associations, wind composite anomalies at 850 hPa and outgoing longwave radiation were analyzed. State 1 (Fig. 10) shows that the region had a suppressed OLR with weak in directional winds; a small area around the Lake Victoria on the western region experienced an enhanced OLR.

In the wettest state 4 (Fig. 10), the whole country had increased OLR which imply that the convective activities were favored and hence the formation of cumulonimbus clouds rain storms. Besides, state 4 experienced strong maritime easterly winds from the Indian Ocean that aided in the frontogenesis.

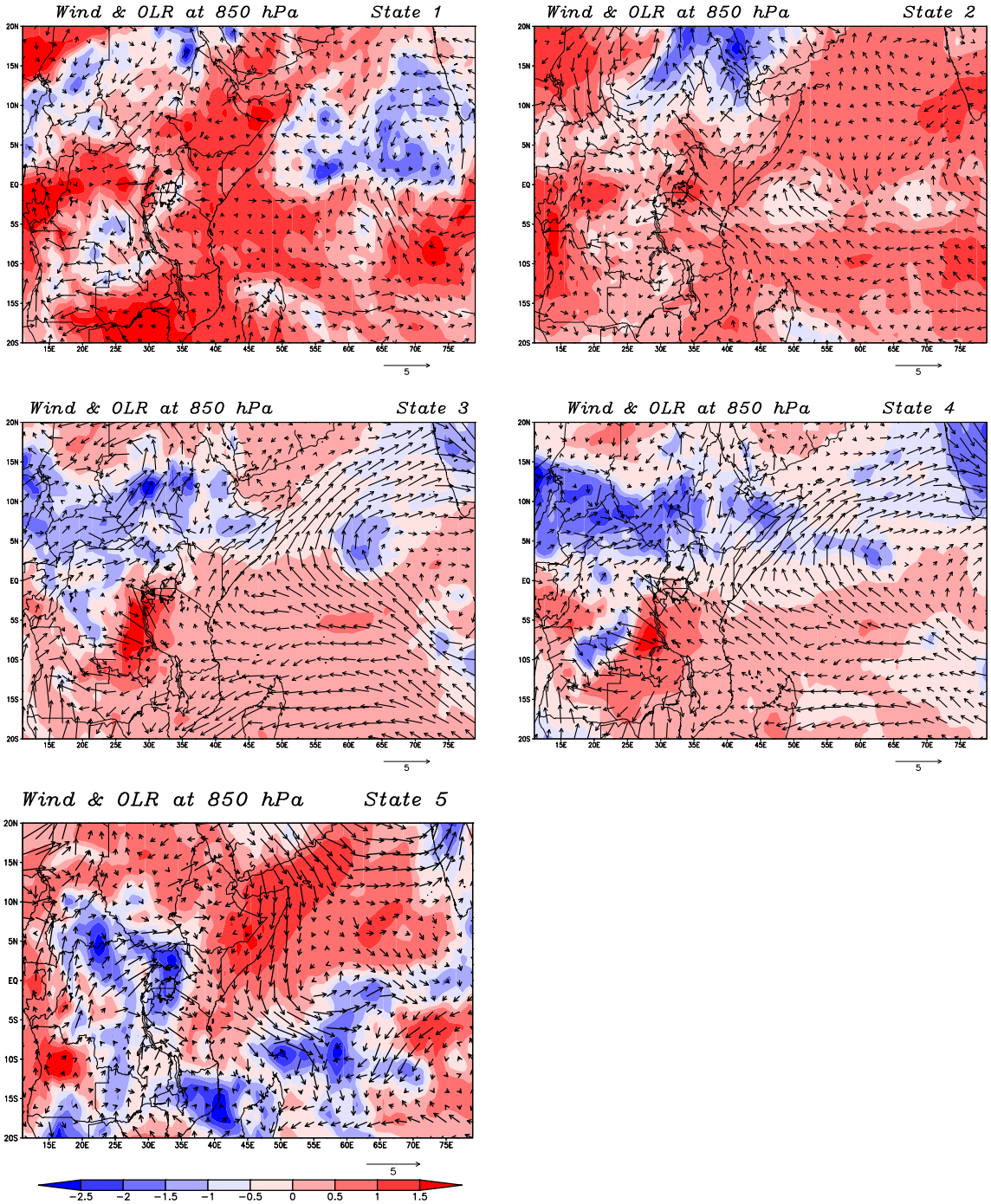

Fig. 10 Composites anomalies concerning October–December climatological mean, over state 1, 2, 3, 4, and 5 assigned days. Arrows indicate 850-hPa winds while shading denotes OLR (W/m^2). Arrow scale is given under each panel (m/s). As shown in Fig. 9, the NHMM composite days used are 371, 544, 611, 582, and 376, respectively

Under state 2 (Fig. 10), the driest state, a less similar condition was noted as in state 4 (Fig. 10). Nonetheless, the winds were very strong, and the convective systems that formed moved further northwards. In state 5 (Fig. 10), negative OLR was recorded in the entire study region. However, convection activity was hampered by the dry continental north easterlies, thus the reduced rainfall events.

3.5 Interannual variability-influence of ENSO

We further did the composite using the seasonal mean SST for an interannual state frequency during the most prevalent years (top 15%) (Fig. 11). The shading shows significance at 90% confidence level. ENSO effects are seen in most states. An ENSO SST anomaly signature is present for years during which states 2 or 4 are highly prevalent, but statistical significance is high only

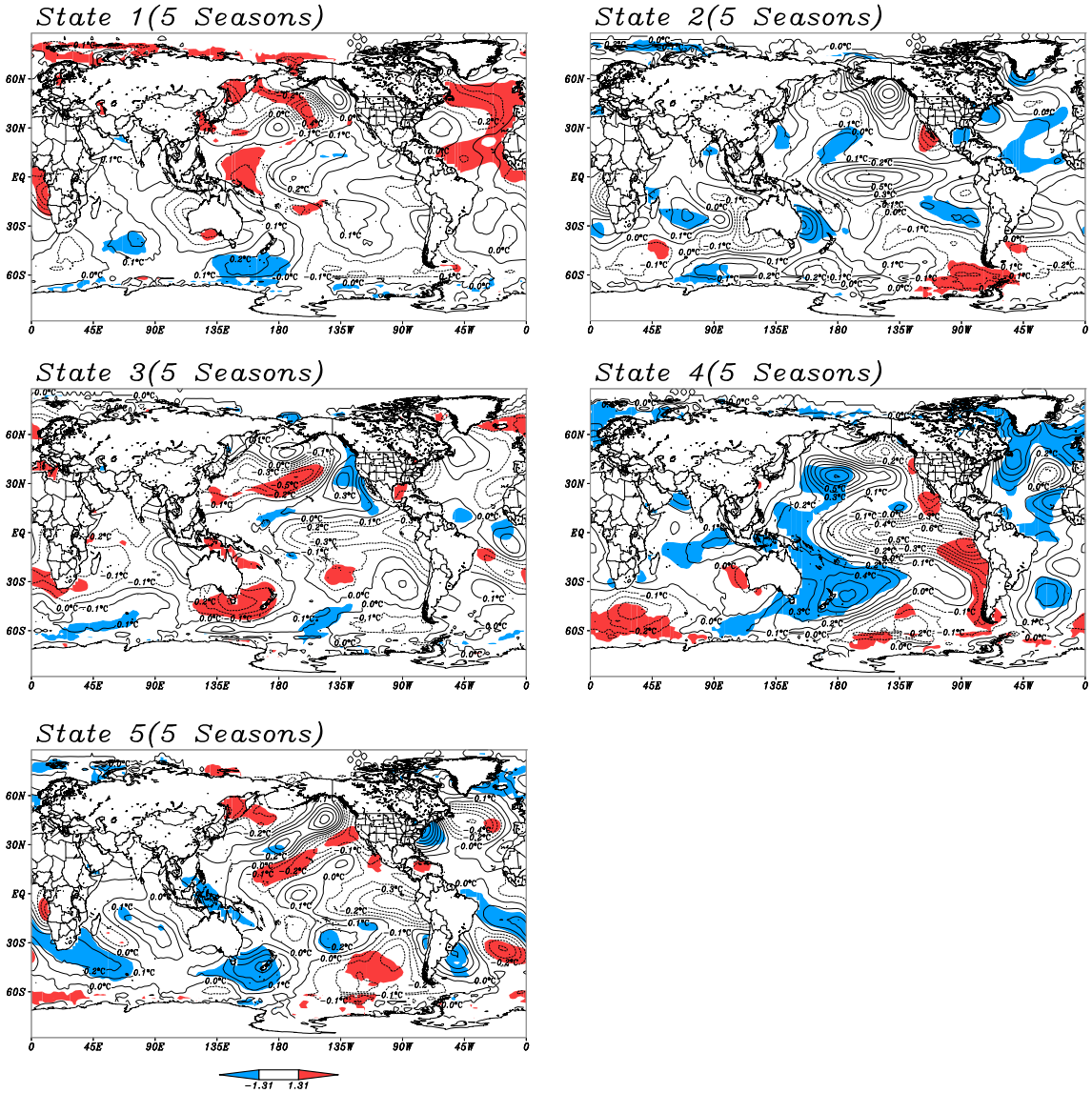

Fig. 11 Anomalous composites of SST for years (OND) during the most prevalent HMM states, for the upper 15% of the interannual distribution of state frequency. The number of seasons in every composite is bracketed. Shading shows 90% statistical significance according to a two-sided Student *t* test. The negative contours are dashed, whereas zero contours are left out, contour interval is 0.2 °C

in state 4 (Indian Niño). During state 4, the positive phase of IOD event, the region of study experiences enhanced precipitation. The opposite holds for state 2 during the negative phase of the Indian Niño event (Fig. 11). State 4 rainfalls are also influenced by El Niño effects whereas in state 2, it is associated with La Niña effects. All the states are associated with Atlantic and Pacific SST anomalies characteristics. States 3 is not linked with appreciable SST anomalies.

3.6 Sub-seasonal and interannual properties of NHMM

To determine wet day length, wet days with more than 1 mm of rainfall were used as the baseline threshold, so any day recording 1 mm of rainfall was considered as a wet spell day. A maximum consecutive wet day length was calculated for the entire season (90 days) to determine the characteristics between the observed and the simulated rainy days. From the analysis, the simulated consecutive spell length peaked till a maximum of 14 consecutive spell length, flattened, then rose up to maximum spell duration of 32 days (Fig. 12). The observed maintained an s-curve and flattened at the maximum spell duration of 58 days. A correlation between the two indicates that both the simulated and the observed have a strong positive correlation of 0.82.

The simulated 90-day period was obtained through averaging NHMM simulated daily rainfall across the 16 stations.

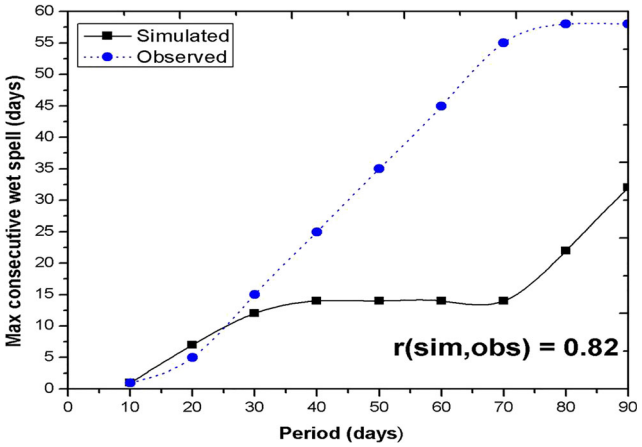

Fig. 12 A maximum consecutive wet spell between the simulated and observed rainfall

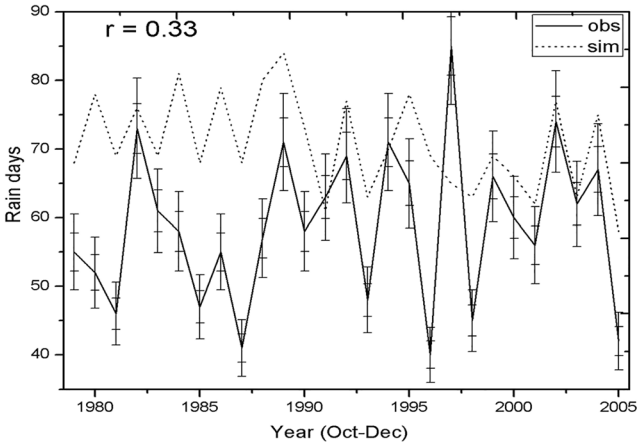

Fig. 14 Interannual variability of NHMM simulated rainfall occurrence versus the observed (solid) averaged over the 16 stations. The average of 50 NHMM simulations (dotted). The per season's rain days number was summed over the entire 16 stations then divided by 16. The error bars show/indicates the whole range of 50 simulations, whereas the inner ticks represent the inter-quartile range

Similarly, a mean of observed daily rainfall from all the stations was derived.

GCM inputs, 50 NHMM runs of daily rainfall occurrence for every station were made. Then, we obtained the average number of rainy days over 16 stations per season. The linear correlation between the simulated and observed rain days was 0.33 significant at a p value of 0.1 (Fig. 13). This implies that NHMM simulations do not recover the mean station quantity as well as the predictive input factor value of the rainfall in this season.

The extremes and the quartiles of simulated distribution were also plotted (Fig. 14). Close to 95% of the years in the observed curve lies within the simulated inter-quartile range, supporting the idea that the variance of the simulated distribution is consistent. During the study years, the 50-member simulated distribution bracketed nearly whole observed ones except one implying it is consistent under the NHMM. This means that under stringent cross-validation, an NHMM is able to generate time series of station's observed mean rainfall occurrence.

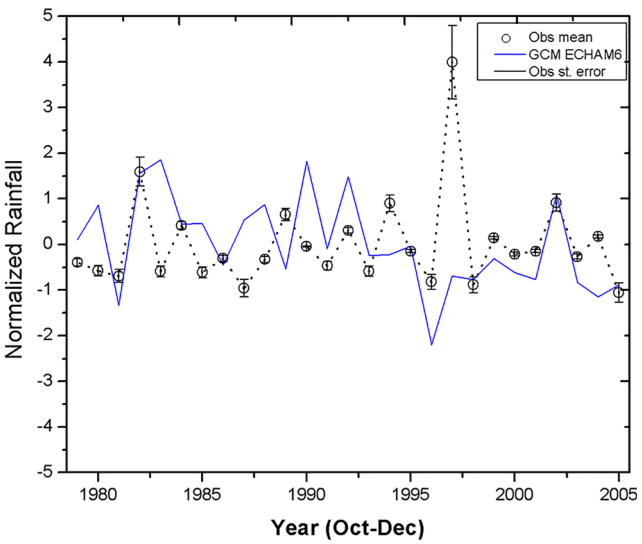

Fig. 13 Interannual variability of candidate GCM predictor variables. Interannual variability of the candidate GCM predictor variables with the observed daily rainfall occurrence, averaged over the 16 stations (circles), the standard error is represented by error bars

4 Conclusion

This study utilized HMMs and NHMM model climatological variations of OND rainfall in Kenya. Historical daily data sets from 16 synoptic stations, reanalysis data, together with the global climate models were employed to accomplish this study. Five states model based on BIC were selected and used for model fitting. States 4 and 2 were the wettest and driest, respectively.

The seasonality and variability exhibited by the estimated state sequence composites like SST and OLR indicates that the weather states, 1, 2, 3, and 5 are linked with large-scale features like the Monsoon, large displacement of Inter Tropical Convergence Zone (ITCZ), Indian Ocean Dipole, ENSO, and North Atlantic Oscillation teleconnections. The dry state 2 is likely to be associated with the inactive Madden-Julian Oscillations phase and or La Niña. Downscaling of daily rainfall occurrences at 16 stations incorporated an NHMM and GCM of daily precipitation with sea surface temperatures during the same study period.

The interannual variability of mean GCM simulated precipitation and average historical stations rainfall shows a weak correlation though significant at 90% confidence interval. Thus, it implies that GCM-NHMM simulations do not recover the rainfall occurrences well. The consecutive wet spell length between the historical rainfall data sets and GCM-NHMM

simulated precipitation for 90-day frequencies shows a strong positive correlation (0.82) significant at a 95% confidence level. Despite the weak association between the GCM-NHMM downscaled rainfalls, the objectives were met.

For sustainable agro-economical, water management and sustenance of human livelihoods, precipitation plays a critical role. Since rainfall maintains the hydrological cycle, proper monitoring and forecasting are inevitable. The non-homogeneous model is a relatively good statistical tool to downscale daily rainfall occurrences. However, some results depict a weak relation between the observed and simulated analyses.

This work recommends filtering of the predictor input variables before the analysis to remove seasonality and trends that may interfere with the results. The performance of other GCM simulations should be tested too since different GCMs perform well in various regions.

Acknowledgments The authors express their appreciation to Ms. Lianyi Guo's thoughtful suggestions and all the data sources. Special thanks go to Nanjing University of Information Science and Technology (NUIST) for promoting a research enabling atmosphere that facilitated completion of this work. Special thanks to NCEP/NCAR and KMD for providing the data used in this study.

Funding information This study was supported by funds from the National Key Research and Development Program of China (Grant No. 2017YF0603804) and the National Natural Science Foundation of China (Grant No. 41575085 and No. 41430528).

References

Addinsoft (2016) XLSTAT, data analysis and statistics with MS excel. Addinsoft, NY, USA. https://www.xlstat.com/en/. Accessed 12 Apr 2018

Aerenson T, Tebaldi C, Sanderson B, Lamarque J-F (2018) Changes in a suite of indicators of extreme temperature and precipitation under 1.5 and 2 degrees warming. Environ Res Lett 13:035009. https://doi.org/10.1088/1748-9326/aaafd6

Bellone E, Hughes JP, Guttorp P (2000) A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. Clim Res 15:1–12. https://doi.org/10.3354/cr015001

Cannon AJ, Whitfield PH (2002) Downscaling recent stream flow conditions in British Columbia, Canada using ensemble neural network models. J Hydrol 259:136–151. https://doi.org/10.1016/S0022-1694(01)00581-9

Charles S, Bates B, Smith I, Hughes P (2004) Statistical downscaling of daily precipitation from observed and modeled atmospheric fields. Hydrol Process 18:1373–1394. https://doi.org/10.1002/hyp.1418

Chen C, Baethgen WE, Robertson A (2013) Contributions of individual variation in temperature, solar radiation and precipitation to crop yield in the North China Plain, 1961–2003. Clim Chang 116:767–788. https://doi.org/10.1007/s10584-012-0509-2

Dee DP, Uppala SM, Simmons AJ, Berrisford P, Poli P, Kobayashi S, Andrae U, Balmaseda MA, Balsamo G, Bauer P, Bechtold P, Beljaars ACM, van de Berg L, Bidlot J, Bormann N, Delsol C, Dragani R, Fuentes M, Geer AJ, Haimberger L, Healy SB, Hersbach H, Hólm EV, Isaksen L, Kållberg P, Köhler M, Matricardi M, McNally AP, Monge-Sanz BM, Morcrette J-J, Park B-K, Peubey C, de Rosnay P, Tavolato C, Thépaut J-N, Vitart F (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q J R Meteorol Soc 137:553–597. https://doi.org/10.1002/qj.828

Forney GD (1973) The Viterbi Algorithm. Proc IEEE 61:268–278

Funk C, Davenport F, Eilerts G, Nourey N, Galu G (2018) Contrasting Kenyan resilience to drought: 2011 and 2017. USAID Special Rep., 20 pp., www.usaid.gov/resilience/contrasting-kenyan-resilience-drought-2011-2017. Accessed 20 Aug 2019

Gabriel KR, Neumann J (1962) A Markov chain model for daily rainfall occurrences at Tel-Aviv. Q J R Meteorol Soc 88:85–90. https://doi.org/10.1002/qj.49708837511

Guo L, Jiang Z, Chen W (2018) Using a hidden Markov model to analyze the flood-season rainfall pattern and its temporal variation over East China. J Meteorol Res 32:410–420. https://doi.org/10.1007/s13351-018-7107-9

Huang, B., Banzon, V. F., Freeman, E., Lawrimore, J., Liu, W., Peterson, T. C.,... & Zhang, H. M. (2014). Extended reconstructed sea surface temperature version 4 (ERSST.v4): Part I. Upgrades and intercomparisons. J Clim 28:911–930. https://doi.org/10.1175/JCLI-D-14-00006.1

Hughes JP, Guttorp P (1994) A class of stochastic models for relating synoptic atmospheric patterns to regional hydrologic phenomena. Water Resour Res 30:1535–1546. https://doi.org/10.1029/93WR02983

Hughes JP, Guttorp P, Charles SP (1999) A non-homogeneous hidden Markov model for precipitation occurrence. J Royal Stat Soc: Series C (Appl Stat) 48:15–30. https://doi.org/10.1111/1467-9876.00136

IFRC Annual Report – 2014. (2015). https://media.ifrc.org/ifrc/wp-content/uploads/sites/5/2017/12/Annual-report-2014.pdf (Accessed 02 January 2018)

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds C, Chelliah M, Ebisuzaki W, Higgins W, Jonowiak J, Mo KC, Ropelewski C, Wang J, Wang J, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year reanalysis project. Bull Am Meteorol Soc 77:437–471. https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2

Kang IS, Kim HM (2010) Assessment of MJO predictability for boreal winter with various statistical and dynamical models. J Clim 23:2368–2378. https://doi.org/10.1175/2010JCLI3288

Kass RE, Raftery AE (1995) Bayes factors. J Am Stat Assoc 90:773–795

Le Cam L (1961) A stochastic theory of precipitation. Fourth Berkeley Symposium on Mathematics, Statistics, and Probability. University of California, Berkeley, California 165–186. https://projecteuclid.org/euclid.bsmsp/1200512811

Maraun D, Wetterhall F, Ireson AM, Chandler RE, Kendon EJ, Widmann M, Brienen S, Rust HW, Sauter T, Themeßl M, Venema VKC, Chun KP, Goodess CM, Jones RG, Onof C, Vrac M, Thiele-Eich I (2010) Precipitation downscaling under climate change: Recent developments to bridge the gap between dynamical models and the end user. Rev Geophys 48:RG3003. https://doi.org/10.1029/2009RG000314

McAvaney B, Covey C, Joussaume S, Kattsov V, Kitoh A, Ogana W, Pitman A, Weaver A, Wood R, Zhao Z-C (2001) Model evaluation. In: Houghton JT, Ding Y, Griggs DJ, Noguer M, van der Linden PJ, Dai X, Maskell K, Johnson CA (eds) Climate Change 2001:The Scientific Basis. Model Evaluation, Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA 881 pp

Mumo L, Yu J, Fang K (2018) Assessing impacts of seasonal climate variability on maize yield in Kenya. Int J Plant Prod 12:297–307. https://doi.org/10.1007/s42106-018-0027-x

Ongoma V, Chen H (2017) Temporal and spatial variability of temperature and precipitation over East Africa from 1951 to 2010. Meteorol

Atmos Phys 129:131–144. https://doi.org/10.1007/s00703-016-0462-0

Ongoma V, Chen H, Gao C, Sagero PO (2017) Variability of temperature properties over Kenya based on observed and reanalyzed datasets. Theor Appl Climatol 133:1175–1190. https://doi.org/10.1007/s00704-017-2246-y

Ongoma V, Chen H, Gao C (2018) Projected change in mean rainfall and temperature over East Africa based on CMIP5 Models. Int J Climatol 38:1375–1392. https://doi.org/10.1002/joc.5252

Ongoma V, Chen H, Gao C (2019) Evaluation of CMIP5 twentieth century rainfall simulation over the equatorial East Africa. Theor Appl Climatol 135:893–910. https://doi.org/10.1007/s00704-018-2392-x

Pineda LE, Willems P (2016) Multisite downscaling of seasonal predictions to daily rainfall characteristics over Pacific–Andean River Basins in Ecuador and Peru Using a nonhomogeneous Hidden Markov model. J Hydrometeorol 17:481–498. https://doi.org/10.1175/JHM-D-15-0040.1

Robertson AW, Kirshner S, Smyth P (2004a) Downscaling of daily rainfall occurrence over Northeast Brazil using a Hidden Markov Model. J Clim 17:4407–4424. https://doi.org/10.1175/JCLI-3216.1

Robertson AW, Lall U, Zebiak SE, Goddard L (2004b) Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. Mon Weather Rev 132:2732–2744. https://doi.org/10.1175/MWR2818.1

Robertson AW, Kirshner S, Smyth P, Charles SP, Bates BC (2005) Subseasonal-to-interdecadal variability of the Australian monsoon over North Queensland. QJR Meteorol Soc 132:511–542. https://doi.org/10.1256/qj.05.75

Robertson AW, Moron V, Swarinoto Y (2009) Seasonal predictability of daily rainfall statistics over Indramayu district, Indonesia. Int J Climatol 29:1449–1462. https://doi.org/10.1002/joc.1816

Rowell DP (2019) An Observational Constraint on CMIP5 Projections of the East African Long Rains and Southern Indian Ocean Warming. Geophys Res Lett 46:6050–6058. https://doi.org/10.1029/2019GL082847

Schubert S (1998) Downscaling local extreme temperature changes in south-eastern Australia from the CSIRO Mark2 GCM. Int J Climatol 18:1419–1438. https://doi.org/10.1002/(SICI)1097-0088(19981115)18:13<1419::AID-JOC314>3.0.CO;2-Z

Steduto P, Hsiao TC, Raes D, Fereres E (2009) AquaCrop—The FAO crop model to simulate yield response to water: I. Concepts and underlying principles. Agron J 101:426–437. https://doi.org/10.2134/agronj2008.0139s

Stern RD, Coe R (1984) A model fitting analysis of daily rainfall data. J R Statist Soc A 147:1–34

Timbal B, Dufour A, McAvaney B (2003) An estimate of future climate change for western France using a statistical downscaling technique. Clim Dyn 20:807–823. https://doi.org/10.1007/s00382-002-0298-9

Verbist K, Robertson AW, Cornelis WM, Gabriels D (2010) Seasonal predictability of daily rainfall characteristics in central northern Chile for dry-land management. J Appl Meteorol Climatol 49:1938–1955. https://doi.org/10.1175/2010JAMC2372.1

Vrugt JA, Ter Braak CJ, Clark MP, Hyman JM, Robinson BA (2008) Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo simulation. Water Resour Res 44:W00B09. https://doi.org/10.1029/2007WR006720

Waymire E, Gupta VK (1981) The mathematical structure of rainfall representations 2. A review of the theory of point processes. Water Resour Res 17:1273–1285. https://doi.org/10.1029/WR017i005p01273

Wilby RL, Charles SP, Zorita E, Timbal B, Whetton P, Mearns LO (2004) Guidelines for use of climate scenarios developed from statistical downscaling methods. Supporting material of the Intergovernmental Panel on Climate Change. http://www.ipcc-data.org/guidelines/dgm_no2_v1_09_2004.pdf (Accessed on 05 June 2017)

Wood AW, Leung LR, Sridhar V, Lettenmaier DP (2004) Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. Clim Chang 62:189–216. https://doi.org/10.1023/B:CLIM.0000013685.99609.9e

Yoo JH, Robertson AW, Kang IS (2010) Analysis of intraseasonal and interannual variability of the Asian summer monsoon using a hidden Markov model. J Clim 23:5498–5516. https://doi.org/10.1175/2010JCLI3473.1

Zeng G, Wang WC, Shen C, Hao Z (2014) Summer precipitation changes over the Yangtze River Valley and North China: simulations from CMIP3 models. Asia Pac J Atmos Sci 50:355–364. https://doi.org/10.1007/s13143-014-0022-9

Zucchini W, Guttorp P (1991) A hidden Markov model for space-time precipitation. Water Resour Res 27:1917–1923. https://doi.org/10.1029/91WR01403