Modelling of Heterogeneity and Serial Dependencies in Precipitation Data Using Hidden Markov Models

Abstract

This paper explores the application of Hidden Markov Models (HMMs) and Finite mixture models (FMMs) for analyzing precipitation data characterized by unobserved heterogeneity, serial dependencies, and unobserved states. Given the significant role of precipitation in agriculture, water resource management, and disaster risk reduction, the study addresses the challenges posed by the nature of precipitation data. We first developed a simulation framework incorporating autoregressive emissions to model distinct hidden states, and later applied the approach to actual rainfall data from the Bungoma region, Kenya. A Gaussian mixture was applied to model the distinct hidden states and HMM was also applied to model the distinct hidden states by taking into account the serial dependencies. The analysis reveals state-dependent variability in precipitation, with distinct mean and variance parameters across states, and highlights the stability of state transitions. For instance, the actual data analysis revealed that we have three distint states with the means(standard deviation); 58.85(28.03), 151.62(43.50) and 215.95(64.06). Model selection criteria based on the BIC indicate the effectiveness of the HMM approach in capturing the broad dynamics of precipitation patterns, providing valuable insights for enhancing climate change adaptation and flood prediction strategies. The results together with pseudo-residuals underscore the potential of HMMs as robust tools for understanding and forecasting precipitation in the context of global climate variability, such as flood predictions, agricultural planning and climate adaptation strategies.

KEYWORDS: heterogeneity, serial dependency, finite mixture models, hidden markov models, decoding

1 Introduction

Precipitation is part of the global climate system and has a direct influence on agriculture, water resources management, and disaster risk reduction[1]. Description and modeling of rainfall are very important for addressing food security-related challenges, flood prediction, and climate change adaptation. Precipitation data are, however, very variable, seasonally organized, nonlinear, and multimodal, and present severe challenges to traditional statistical modeling techniques[2]. In particular, the modeling of serial dependencies and unobserved states that support rainfall generation is an issue calling for robust tools to deal with this kind of complexity[3].

HMMs are a robust framework to model time series data with stochastic emissions and unobserved states[4]. HMMs are particularly suited for precipitation modeling since they can capture differences between weather states (e.g., wet and dry periods of weather) and take into account temporal dependencies in the data[5]. FMMs, one of the alternatives with widespread usage, cannot take into account the serial dependencies but only try to capture the distributional characteristics of the observations. The method also highly depends on the validity of the distribution assumptions[6]. A comparison of these two methods provides the opportunity to examine their relative strengths and weaknesses within the context of complex environmental time series. Various studies have explored multiple methodologies for modeling precipitation. Gupta et al.(2022)[7] introduced a positively skewed two-parameter family of distributions to describe

precipitation amounts, which was accompanied by nonparametric tests to detect changes induced by factors like cloud seeding. His study provided an innovative approach to parameter estimation from observed precipitation data, highlighting the potential flexibility of distributional techniques.

Similarly, Oduor (2025)[8] explored the application of finite Gaussian mixture models (GMMs) for modeling high-frequency financial data. It highlights the challenges of capturing the complex statistical properties of such data, including volatility clustering, heavy tails, and asymmetry. The study demonstrates that GMMs, which combine multiple Gaussian distributions, provide a flexible framework to accurately represent these characteristics. By fitting the model to real-world high-frequency financial datasets, the authors show its effectiveness in capturing the intricate dynamics of market behavior. The findings suggest that GMMs offer a robust tool for financial data analysis, with potential applications in risk management, pricing, and forecasting. The research underscores the importance of advanced statistical techniques in understanding the complexities of modern financial markets.

Additionally, Oliver et al. (2020)[9] In their 2020 study, presented an advanced HMM designed to simulate sub-daily rainfall time series, addressing critical challenges such as capturing long dry periods, seasonal variations, and extreme rainfall events. The model incorporated innovative features like clone states and temporally non-homogeneous state persistence probabilities, enabling it to better represent the complex temporal structure of rainfall, including diurnal and seasonal patterns. Set within a Bayesian framework, the model provided a robust quantification of parametric and predictive uncertainty, allowing for thorough posterior predictive analyses to validate its performance. Applied to an 8-year hourly rainfall dataset from Exeter, UK, the model demonstrated strong interpretability and accuracy in reproducing key rainfall characteristics, making it a valuable tool for hydrological applications such as flood risk assessment and urban drainage planning. This work highlighted the potential of advanced HMMs in improving the simulation of high-resolution rainfall data, offering significant advancements over traditional approaches.

The emphasis in our current study is to provide an illustration of the appropriateness of HMMs to rainfall time series data.kenyan rainfall data and Simulation on precipitation data that were synthesized to mimic actual characteristics such as seasonality, multimodality, and autocorrelation were used in evaluating the promise of HMMs. The superiority of HMMs over FMMs in capturing sequential dependence in the data, identification of hidden states, and interpretation of the hidden processes underlying the rainfall processes is also demonstrated in this work.

The results of the research are meaningful both theoretically and practically. Theoretically, the research enhances the existing body of knowledge in statistical modeling of time series data. Practically, the outcome has implications in enhancing the precision of rainfall prediction, agricultural planning, and decision-making on climate. In reference to a comparison of HMMs and FMMs, the paper presents detailed information on the applicability of these approaches to modeling rainfall and other intricate environmental processes. Discussion on the simulation approach used to simulate rainfall data, theoretical basis of HMMs and FMMs, and comparative assessment of their performance in extracting prevailing data features is presented in the subsequent sections.

2 Methodology

2.1 Data Exploration

In this study, the simulated and actual precipitation data was explored using various visualization techniques to understand the underlying patterns and characteristics of the data. A time series plot of the simulated data over the entire period was used to evaluate trends and variations in precipitation over time. A histogram of precipitation was created to examine the distribution of precipitation over time. The autocorrelation function (ACF) plot was used to assess serial dependence in the data.

2.2 Finite Mixture Models

In cases of heterogeneity, finite mixture models(FMMs) become a suitable approach, as they allow for the identification of distinct sub-populations within the data and provide a framework to account for the variability unexplained by observed covariates [10]. FMMs are powerful statistical tools designed to handle heterogeneous data by assuming that the observed data arises from a mixture of several distinct sub-populations [10]. These models are particularly useful when a single parametric distribution cannot adequately describe the variability in the data, which is often the case when the

population is composed of several homogeneous sub-groups. Each sub-group, or component, within the mixture model is represented by its own probability distribution, and the overall population distribution is a weighted sum of these component distributions [10]. Mathematically, the density of a g-component mixture model can be expressed as:

$$f(y) = \sum_{i=1}^{g} \pi_i f_i(y) \tag{1}$$

where π_i is the probability of the i-th component, $0 \leq \pi_i \leq 1$, and $\sum_{i=1}^{g} \pi_i = 1$. The component densities $f_i(y)$ depend on parameters that are either known or need to be estimated.
For a g-component mixture of normals with equal variances, the model can be expressed as:

$$Y \mid \mu \sim \mathcal{N}(\mu, \sigma^2) \quad \mu \sim \begin{pmatrix} \mu_1 & \cdots & \mu_g \\ \pi_1 & \cdots & \pi_g \end{pmatrix}. \tag{2}$$

where μ_i is mean for the g^{th} component, and π_i is the probability of an observation being in component g. By fitting an FMM, we can better understand the underlying structure of the data, identify distinct sub-populations, and make more accurate inferences about the relationships between the variables of interest.

Parameter estimation in FMM is done through maximum likelihood estimation (MLE). Maximizing the log-likelihood with respect to all unknown parameters in the model requires numerical iterative procedures. However, since the calculation of the second-order derivative is quite numerically complex, the Expectation-Maximization (EM) algorithm is used as an alternative to the classical Newton-Raphson procedure [11].

It is worth noting that inference for the number of components, g does not follow the standard likelihood theory since the H_0 is on the boundary of parameter space under H_A. To avoid selection of g, we treat it as a parameter in the likelihood and hence estimate it from the data. This is done with the help of the CAMAN package in the R software. To avoid extremely slow convergence due to many support points, the estimation procedure in CAMAN is split into two-phases: VEM and EM [12, 11]. Letting G be the distribution function of the latent variable X, G is referred to as the mixing distribution. The resulting estimate for the distribution of the latent variable X is referred to as a Non-Parametric Maximum Likelihood Estimate (NPMLE). This estimate maximizes the log-likelihood value across the entire class of possible distributions for X [11]. To verify if the estimate obtained is indeed a NPMLE, we use the gradient function, $d(G, x)$. The gradient function is a sample average of the likelihood ratios, defined as:

$$d(G, x) = \frac{1}{N} \sum_{i} \frac{f_i(y_i \mid x)}{f_i(y_i \mid G)} \tag{3}$$

where the numerator is the density for a particular support point x and the denominator is the density of the finite mixture. The mixing distribution G is considered to be NPMLE if and only if the gradient function $d(G, \hat{x})$ never exceeds one in the interval of interest and is equal to one for all the support points x. Furthermore, the gradient function is identically one if and only if $\hat{G}$ is not unique. However, it is worth noting that this approach assumes independence between observations, which in most cases does not apply for time series data where the current observations are influenced by the previous observations, although this will be investigated in our study. Additionally, rainfall patterns frequently involve transitions between distinct weather states, which FMMs are not inherently designed to handle. Thus, one way of taking into account the serial dependence and heterogeneity is to use HMMs as suggested by Cappe and his collaborators [13].

2.3 Hidden Markov Models

HMMs are powerful statistical tools used to model time series data. They are statistical models where the distribution that generates an observation is influenced by the state of an underlying, hidden Markov process. They were first introduced by Baum and his collaborators in the late 1960s[14] and was further proposed for speech recognition by Rabiner and Juang in the 1980s[4]. It operates by defining two interconnected processes: the unobserved parameter process, denoted as $\{C_t : t \in \mathbb{N}\}$, and the state-dependent observation process, represented as $\{X_t : t \in \mathbb{N}\}$. The parameter process $\{C_t\}$

satisfies the Markov property, which suggests that conditioning on the 'history' of the process up to time $t-1$ is the same as conditioning solely on the most recent value C_{t-1}:

$$\Pr(C_t \mid C_{1:t-1}) = \Pr(C_t \mid C_{t-1}), \quad t = 2, 3, \ldots \tag{4}$$

This is regarded as a first relaxation of the independence assumption[9]. The observation process $\{X_t\}$ is such that the distribution of X_t depends only on C_t and is independent of past states or observations:

$$\Pr(X_t \mid X_{1:t-1}, C_{1:t}) = \Pr(X_t \mid C_t), \quad t \in \mathbb{N}. \tag{5}$$

In an m-state HMM, the parameter process $\{C_t\}$ transitions among m discrete states, each associated with a unique probability distribution that governs the observations $\{X_t\}$.

2.3.1 Transition Probability Matrix

The transition probability matrix in a HMM represents the probabilities of transitioning from one hidden state to another in consecutive time steps. The general form of the transition probability matrix, denoted as A, is:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix}, \tag{6}$$

where $a_{ij} = P(C_{s+t} = j | C_s = i)$ is the probability of transitioning from state i at time s to state j at time $s + t$. The rows of the matrix must satisfy the condition $\sum_{j=1}^{N} a_{ij} = 1$ for all i, as they represent probability distributions. When the transition probabilities do not depend on s (position), then the chain is referred to as *homogeneous*, which holds in this study.

A stationary HMM assumes that the unconditional state probabilities remain constant over time. Stationary HMMs are particularly suitable for systems where the underlying dynamics do not vary over time, making them less effective for capturing the dynamic patterns in precipitation [15, 16]. In contrast, a non-stationary HMM allows the unconditional state probability to vary over time, making it suitable for our study with dynamic or time-dependent behavior. Non-stationary HMMs are particularly useful for modeling systems with temporal variations, such as precipitation data [15, 16].

Transition probabilities can be estimated from training data using methods such as maximum likelihood estimation (MLE). If the sequence of hidden states $C = \{C_1, C_2, \ldots, C_T\}$ is known, the transition probabilities are computed as:

$$a_{ij} = \frac{\text{Number of transitions from state } i \text{ to state } j}{\text{Total number of transitions from state } i}. \tag{7}$$

2.3.2 Parameter Estimation

In this study, the data consists simulated data for precipitation, and therefore a Gaussian HMM is used. The parameters of a Gaussian HMM are estimated through Maximum Likelihood Estimation (MLE), which maximizes the likelihood function given the observed sequence of data. For a non-stationary Gaussian HMM with m states, the likelihood function is expressed as:

$$L(\boldsymbol{\theta}, \boldsymbol{\delta}, A \mid x_1, \ldots, x_n) = \sum_{c_1, \ldots, c_n} \delta_{c_1} p_{c_1}(x_1; \theta_{c_1}) \prod_{t=2}^{n} a_{c_{t-1}, c_t} p_{c_t}(x_t; \theta_{c_t}), \tag{8}$$

where $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_m\}$, with $\theta_i = (\mu_i, \sigma_i^2)$, are the parameters of the Gaussian distributions for the m states, $x_1, \ldots, x_n$ are the n observations, and a_{c_{t-1}, c_t} is the transition probability from state c_{t-1} to c_t. The vector $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_m\}$ contains the initial state probabilities, where δ_i represents the initial probability of being in state i. The Gaussian density function for state i is given by:

$$p_i(x; \theta_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x - \mu_i)^2}{2\sigma_i^2}\right). \tag{9}$$

Since in practice the state paths are not known, they are treated as missing data and estimated together with the other parameters using the Baum-Welch algorithm, a version of the Expectation-Maximization (EM) algorithm. The

Baum-Welch algorithm iteratively refines parameter estimates through two key steps: (1) the Expectation (E) step, which computes the expected values for A_{ij} and θ_i based on current parameter estimates, and (2) the Maximization (M) step, which updates the parameters based on the expected values obtained in the E-step to maximize the likelihood. These steps are repeated until convergence. When computing the MLE, the likelihood function might be complex and, as a result, multiple local maxima can occur. This issue is mitigated by testing multiple initial parameter values to obtain a global maximum [15, 16].

To evaluate the likelihood of the observed sequence $x_1, \ldots, x_n$, the forward algorithm is employed. The forward algorithm recursively computes the probability of observing the partial sequence $x_1, \ldots, x_t$ and being in state S_t at time t, denoted as $\alpha_t(i)$. The forward probabilities are initialized as:

$$\alpha_1(i) = \delta_i p_i(x_1; \theta_i), \quad i = 1, \ldots, m, \tag{10}$$

and recursively computed for $t = 2, \ldots, n$ as:

$$\alpha_t(j) = \sum_{i=1}^{m} \alpha_{t-1}(i) a_{ij} p_j(x_t; \theta_j), \quad j = 1, \ldots, m. \tag{11}$$

The likelihood of the entire observed sequence is then obtained by summing the forward probabilities at the final time step:

$$L(x_1, \ldots, x_n) = \sum_{i=1}^{m} \alpha_n(i). \tag{12}$$

The forward algorithm is computationally efficient and avoids the exponential complexity of enumerating all possible state sequences [15].

Complementing the forward algorithm, the backward algorithm is used to compute the probability of observing the future sequence $x_{t+1}, \ldots, x_n$ given that the system is in state $S_t = i$ at time t, denoted as $\beta_t(i)$. The backward probabilities are initialized at the final time step as:

$$\beta_n(i) = 1, \quad i = 1, \ldots, m, \tag{13}$$

and recursively computed backward for $t = n-1, \ldots, 1$ as:

$$\beta_t(i) = \sum_{j=1}^{m} a_{ij} p_j(x_{t+1}; \theta_j) \beta_{t+1}(j), \quad i = 1, \ldots, m. \tag{14}$$

The backward probabilities, along with the forward probabilities, are essential for the Expectation step of the Baum-Welch algorithm, as they enable the computation of posterior probabilities and expected counts of state transitions and emissions.

By combining the forward and backward algorithms, the Baum-Welch algorithm iteratively optimizes the HMM parameters to fit the observed data. Together, these methods provide a robust framework for modeling and analyzing sequential data, such as the precipitation data at hand, where hidden states and temporal dependencies play a significant role [15, 16].

2.3.3 Model Selection

In modeling any data using HMMs, determining the optimal number of states is a critical step. There are two ways to determine the optimal number of states: One might reasonably hypothesize a limited number of states apriori (e.g., dry and rainy season i.e 2 states), or due to the lack of clear prior knowledge about the exact number of states, one could rely on data-driven criteria such as the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for model selection. Since in this study we rely on simulated precipitation data for demontration purposes, we made use of the data driven criteria. Urso, (2024)[17] noted that "in the case of HMM, the problem of model selection (and in particular the choice of the number of states in the Markov chain component model) has yet to be satisfactorily solved." However, in their numerical study, Celeux et al.[18] found that the AIC criterion tends to underpenalize the model's complexity, while the BIC generally performs well when the HMM provides a reasonable representation of the observed process. However, in this research, both AIC and BIC were applied, with a greater reliance on the latter. The selected model was further

diagnosed using ordinary pseudo-residuals. The ordinary pseudo-residuals are based on the conditional distribution given all other observations in the data. They therefore give an indication of the fit of the model to the observed data through discrepancies between the model's fitted values and the observed values. These residuals are extremely useful for evaluating the overall fit of the model and detecting possible outliers that may be influencing the results.

2.4 Hidden States of HMM

Decoding in HMMs is the process of finding the most probable sequence of hidden states given evidence (or observation) and model parameters. Basically, two fundamental approaches to solving this problem exist: local and global decoding.

2.4.1 Local Decoding

Local decoding involves identifying the most likely hidden state for each time point independently by maximizing the conditional probability for that specific time step. For a given time step $t \in \{1, \ldots, T\}$, the objective is to determine the state C_t that satisfies:

$$P(C_t = i \mid X^{(T)} = x^{(T)}), \tag{15}$$

where $X^{(T)} = \{x_1, x_2, \ldots, x_T\}$ represents the sequence of observed data, and C_t denotes the hidden state at time t. To compute $P(C_t = i \mid X^{(T)} = x^{(T)})$, the forward and backward probabilities are used as follows:

$$P(C_t = i \mid X^{(T)} = x^{(T)}) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{j=1}^{m} \alpha_t(j)\beta_t(j)}, \tag{16}$$

where:

- $\alpha_t(i)$, known as the forward probability, represents the likelihood of observing the partial sequence $x_1, x_2, \ldots, x_t$ while being in state i at time t,
- $\beta_t(i)$, the backward probability, captures the likelihood of observing the remaining sequence $x_{t+1}, x_{t+2}, \ldots, x_T$ given that the system is in state i at time t,
- m refers to the total number of possible hidden states in the model.

The numerator, $\alpha_t(i)\beta_t(i)$, corresponds to the joint probability of the system being in state i at time t while generating the entire sequence of observations. The denominator ensures that the probabilities are normalized across all possible states. This approach is computationally efficient and provides valuable information about the likelihood of each state at individual time points [15].

Despite its efficiency, local decoding does have limitations, as it considers each time point independently and does not account for the overall sequence of states. To overcome this, global decoding methods, such as the Viterbi algorithm, were developed to optimize the entire sequence of hidden states simultaneously.

2.4.2 Global Decoding and the Viterbi Algorithm

Unlike local decoding, which independently maximizes the likelihood of states at each time step, global decoding aims to find the single most probable sequence of hidden states $C^{(T)} = \{c_1, c_2, \ldots, c_T\}$ that maximizes the joint conditional probability:

$$\Pr(C^{(T)} = c^{(T)} \mid X^{(T)} = x^{(T)}). \tag{17}$$

The most probable sequence is efficiently computed using the Viterbi algorithm, a dynamic programming approach that recursively determines the optimal path through the state space.

The Viterbi algorithm operates by iteratively calculating the highest probability of observing the sequence $x_1, \ldots, x_t$ and ending in a particular state i at time t. Let $\delta_t(i)$ represent this maximum probability:

$$\delta_t(i) = \max_{c_1, \ldots, c_{t-1}} P(c_1, \ldots, c_{t-1}, c_t = i, x_1, \ldots, x_t). \tag{18}$$

The recursive relationship for $\delta_t(i)$ is given by:

$$\delta_t(j) = \max_{i=1,\ldots,m} [\delta_{t-1}(i)a_{ij}]\, p_j(x_t), \tag{19}$$

where:

- $\delta_{t-1}(i)$: The highest probability of reaching state i at time $t-1$,
- a_{ij}: Transition probability from state i to state j,
- $p_j(x_t)$: Emission probability of observing x_t in state j.

The algorithm consists of three main steps:

1. **Initialization**: At the first time step, the probabilities are initialized as:

$$\delta_1(i) = \delta_i p_i(x_1), \quad \text{for } i = 1, \ldots, m, \tag{20}$$

   where δ_i represents the initial probability of being in state i.

2. **Recursion**: For each subsequent time step $t = 2, \ldots, T$, the probabilities are updated as:

$$\delta_t(j) = \max_{i=1,\ldots,m} [\delta_{t-1}(i)a_{ij}]\, p_j(x_t). \tag{21}$$

3. **Termination and Backtracking**: At the final time step, the most probable ending state is identified as:

$$c_T = \arg\max_{i=1,\ldots,m} \delta_T(i), \tag{22}$$

   and the most likely sequence of states is reconstructed by tracing back through the stored indices.

To address potential issues with numerical underflow during computations, the logarithm of probabilities is often employed. The recursion then becomes:

$$\log \delta_t(j) = \max_{i=1,\ldots,m} [\log \delta_{t-1}(i) + \log a_{ij}] + \log p_j(x_t). \tag{23}$$

By systematically optimizing over all possible state sequences, the Viterbi algorithm identifies the most probable path through the hidden states, making it an essential tool in many applications of HMMs [15, 16].

2.5 Simulation Study

In this study, we simulated precipitation data modeled as a HMM with autoregressive emissions, representing two distinct hidden states: Dry and Wet. The simulation consists of 10 sequences, each containing 100 time steps (T = 100). The initial state probabilities are set to favor the Dry state with a distribution $\pi = (0.7, 0.3)$. The transition probabilities are defined such that there is a 90% probability of remaining in the Dry state and an 85% probability of remaining in the Wet state across time steps, resulting in higher state persistence. The emission parameters are characterized by means of $\mu_1 = 5$ (Dry) and $\mu_2 = 50$ (Wet), with respective standard deviations of $\sigma_1 = 2$ and $\sigma_2 = 10$, ensuring realistic variability in precipitation values. To incorporate temporal dependence, rather than directly sampling emissions, we apply an autoregressive approach where each new precipitation value is influenced by the previous value, specifically: $R_t^{(n)} = 0.8 \cdot R_{t-1}^{(n)} + \mathcal{N}(\mu_{S_t^{(n)}}, \sigma^2_{S_t^{(n)}})$. Finally, we ensure that the simulated precipitation values remain non-negative by truncating any values below zero. The summary of this simulation is displayed in pseudo code 1. The resulting simulated data serves as a basis for further analysis using HMM techniques in our research.

Algorithm 1 : Simulating rainfall using HMM with autoregressive emissions

Set number of states K
Define transition probabilities $\mathbf{P}$
Set initial probabilities π
Define emission parameters for each state: means μ_k and variances σ_k^2
Set sequence length T and number of sequences N
for $n = 1, \ldots, N$ **do**
  Initialize state $S_1^{(n)} \sim$ Categorical(π)
  Initialize precipitation $R_1^{(n)} \sim \mathcal{N}(\mu_{S_1^{(n)}}, \sigma^2_{S_1^{(n)}})$
  for $t = 2, \ldots, T$ **do**
    Sample state $S_t^{(n)} \sim$ Categorical$(\mathbf{P}[S_{t-1}^{(n)}, \cdot])$
    Sample base precipitation $R_t^{(n)} \sim \mathcal{N}(\mu_{S_t^{(n)}}, \sigma^2_{S_t^{(n)}})$
    Update $R_t^{(n)} \leftarrow 0.8 \cdot R_{t-1}^{(n)} + R_t^{(n)}$ {Add autoregressive term}
    Ensure $R_t^{(n)} > 0$ by applying $R_t^{(n)} \leftarrow \max(0, R_t^{(n)})$
  end for
end for
return Precipitation $\{R_t^{(n)}\}$
=0

2.6 Software

The entire analysis was conducted using R version 4.3.3 with the *dthmm*, *BaumWelch* and *Viterbi* functions from the *HiddenMarkov* package. In addition the simulation was conducted using the *simulate hmm* function in the *seqHMM* package.

3 Results: Simulated Data

3.1 Data Exploration

Figure 1 illustrates the time series plot of simulated precipitation over the entire study period. The plot captures the inherent variability in the data, showcasing periods of low precipitation interspersed with intervals of higher precipitation. This variability is further analyzed in Figure 2a, which reveals a multi-modal distribution, suggesting the potential use of a FMM. However, as previously discussed, FMMs assume independence among observations, a key assumption that is contradicted by the autocorrelation plot in Figure 2b. The autocorrelation plot highlights significant serial dependencies, with a strong positive correlation coefficient of approximately 60% for precipitation values separated by one lag. Interestingly, precipitation values separated by 9 lags show a correlation of about 40%, while those 14 lags apart exhibit negative correlations. Positive correlations re-emerge at 19 lags, further emphasizing the cyclical patterns in the data. These findings underscore the limitations of FMMs in this context and demonstrate the necessity of HMMs to capture and account for the observed serial dependencies in precipitation data.

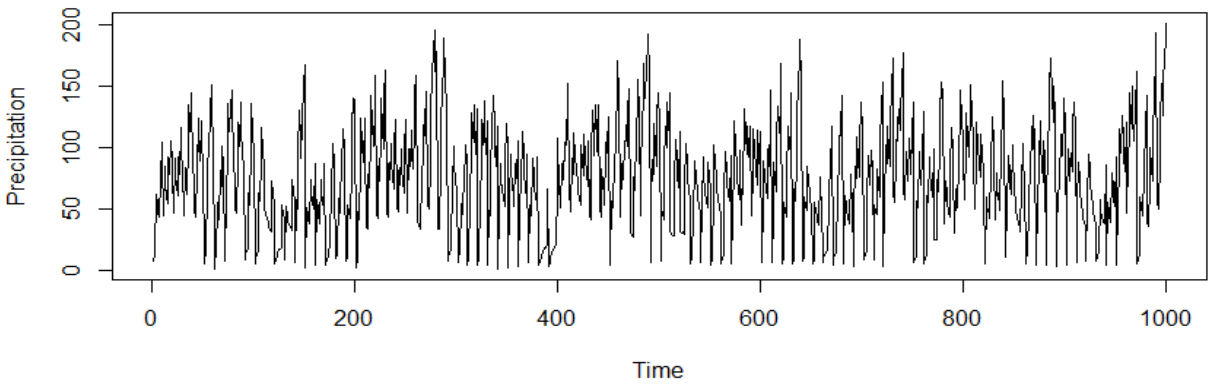
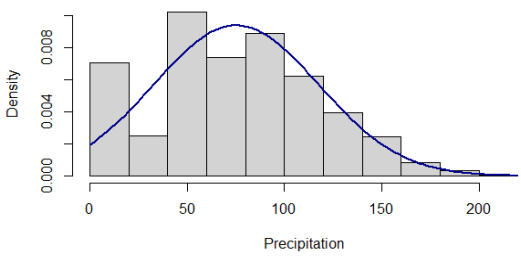
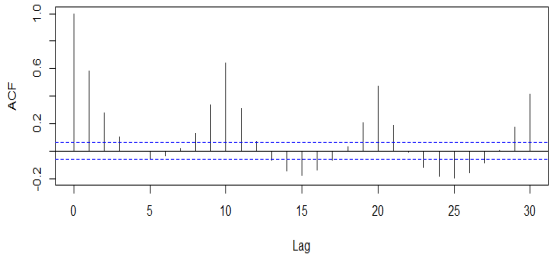


Figure 1: Time series plot of observed data.



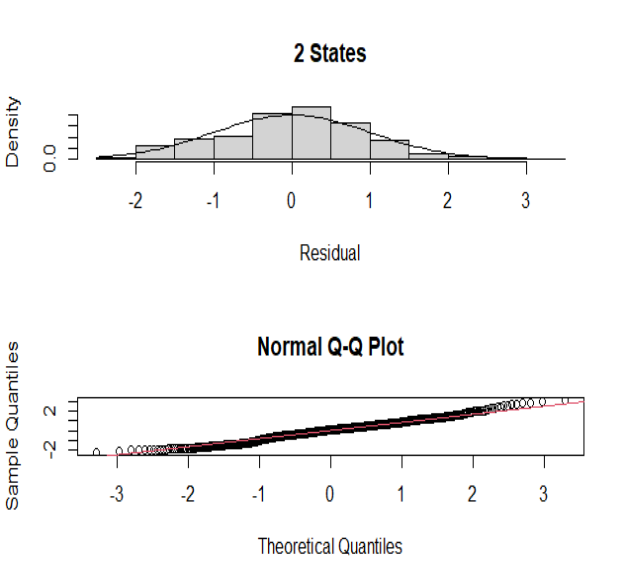
(a) Histogram of Precipitation



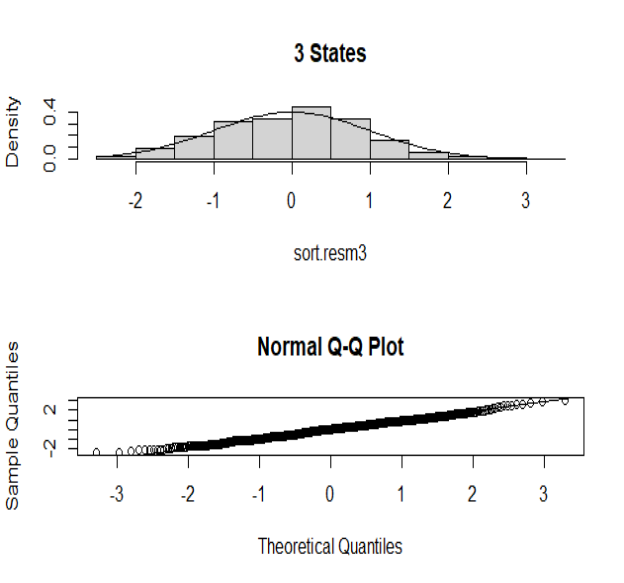
(b) Autocorrelation Function

3.2 Hidden Markov Models

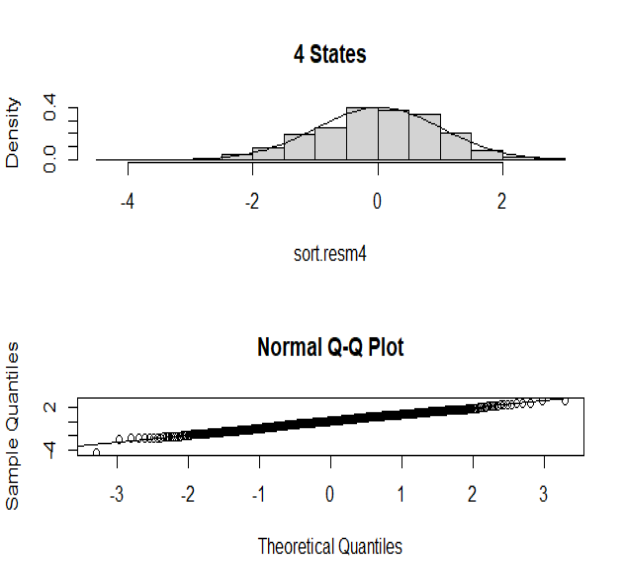
As previously discussed, model selection is a critical step in HMM. To determine the optimal number of states for the given data, we fitted Gaussian HMMs with varying numbers of states ($m = 2, 3, \ldots, 5$). The corresponding values of $-\log L$, AIC, and BIC for each model are presented in Table 1. Notably, both the AIC and BIC criteria selected the model with 4 states. Before interpreting the parameter estimates for this model, it is essential to validate its fit through diagnostic checks. This was achieved using quantile-quantile (Q-Q) and Ordinary pseudo-residual plots, as shown in Figure 3. For comparison, residuals from models with $m = 2, 3,$ and, 5 demonstrated deviations from normality, which indicate a poor fit. The residuals from the selected 4-state model align closely with the standard normal distribution, confirming its suitability for the data. This finding suggests that the model represents a 4-state homogeneous Markov chain, with the transition probability matrix provided in Table 3. Consequently, we can proceed to interpret the parameter estimates derived from the Baum-Welch algorithm for this model. Out of curiosity, we fitted a Gaussian mixture model to observe the difference in these two approaches. The model identified only 3 mixture components, with $-LogL$ and BIC values of 5159.74 and 10354.02, respectively. These values are higher than those of all the HMMs fitted, which is not surprising given the earlier observation that the data points are highly correlated.



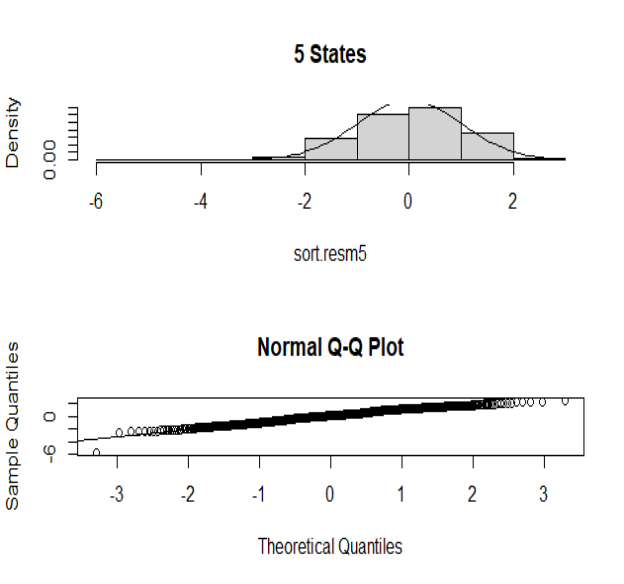
(a) Residuals and Q-Q Plot for 2 States



(b) Residuals and Q-Q Plot for 3 States



(c) Residuals and Q-Q Plot for 4 States



(d) Residuals and Q-Q Plot for 5 States

Figure 3: Ordinary pseudo-residuals and their corresponding Q-Q plots for models with 2, 3, 4, and 5 states. The top row presents the results for 2 and 3 states, separated by a decorative divider from the bottom row, which shows the results for 4 and 5 states.

The initial state probabilities and response parameters, as given in Table 2, provide useful information about the modeled system. The initial probabilities (δ_i) tell us that the process begins with complete certainty in state 3 ($\delta_3 = 1.000$), implying that state 3 has an important role as the initial condition. This may be a beginning stage with moderate

precipitation levels, as indicated by the respective mean value ($\mu_3 = 85.067$) and standard deviation ($\sigma_3 = 17.455$). The response parameters (μ_i and σ_i) have evident differences across the states. State 1 has the lowest mean precipitation ($\mu_1 = 9.100$) and lowest variability ($\sigma_1 = 4.638$), indicating possibly a low-precipitation or dry regime. State 4, on the other hand, has the highest mean precipitation ($\mu_4 = 128.347$) and highest variability ($\sigma_4 = 25.363$), which is the hallmark of heavy precipitation events. The transition matrix (Table 3) provides further information regarding the system's dynamics as well. Transitions are mainly confined to a small number of states, and probabilities of retaining the current state are high. For instance, there is a 55.9% ($P_{11} = 0.559$) chance of retaining state 1, and state 3 has a great chance of remaining stable at 69.9%. Of special note is that state 4 is most likely to persist, with 76.2% chance of self-transitions, meaning likely long periods of intensive rains. These patterns establish a state-dependent variability of rainfall in the system, with higher states (i.e., state 4) mapping onto events and lower states (i.e., state 1) onto dryness. The structure of the transition matrix speaks to relatively stable state processes.

Table 1: Model Selection Criteria for Different Number of States

No. States	k	**-log** L	**AIC**	**BIC**
2	7	5030.343	10074.69	10109.04
3	14	4908.144	9844.29	9912.996
4	23	4603.63	**9253.27**	**9366.15**
5	34	4640.81	9389.628	9556.49

Table 2: Initial State Probabilities and Response Parameters

i	**S1**	**S2**	**S3**	**S4**
δ_i	0.000	0.000	1.000	0.000
μ_i	9.100	48.565	85.067	128.347
σ_i	4.638	10.208	17.455	25.363

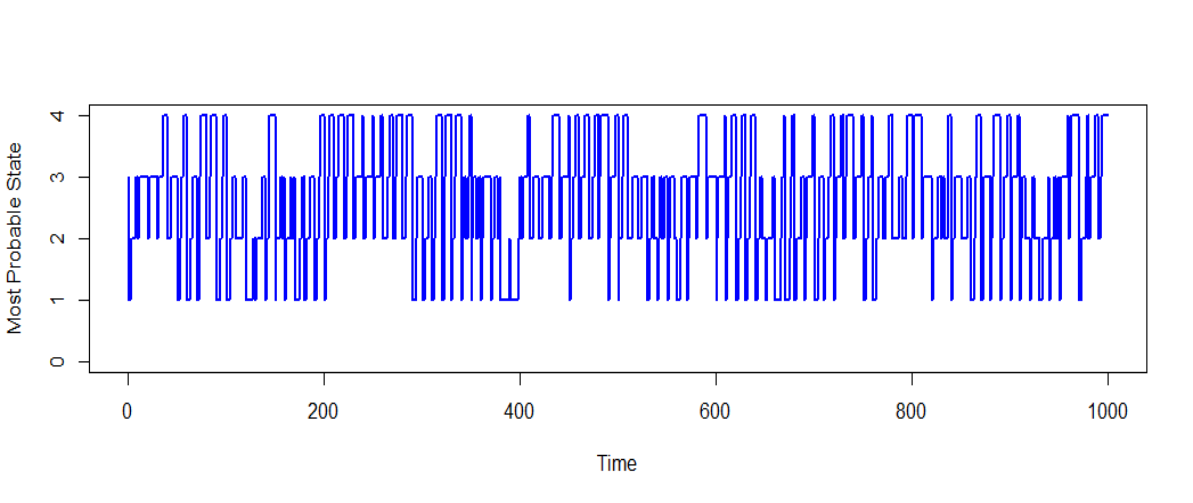
Table 3: Transition Matrix for Hidden Markov Model

State	State 1	State 2	State 3	State 4
1	0.559	0.441	0.000	0.000
2	0.029	0.561	0.411	0.000
3	0.066	0.074	0.699	0.160
4	0.129	0.109	0.000	0.762

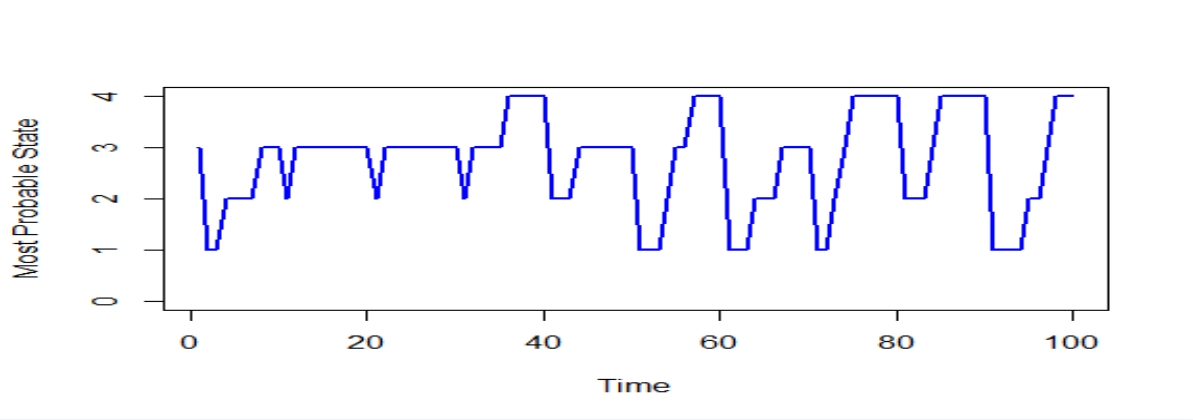
Table 4: Number of corresponding observations in each state.

State	S_1	S_2	S_3	S_4
Obs	95	261	414	230

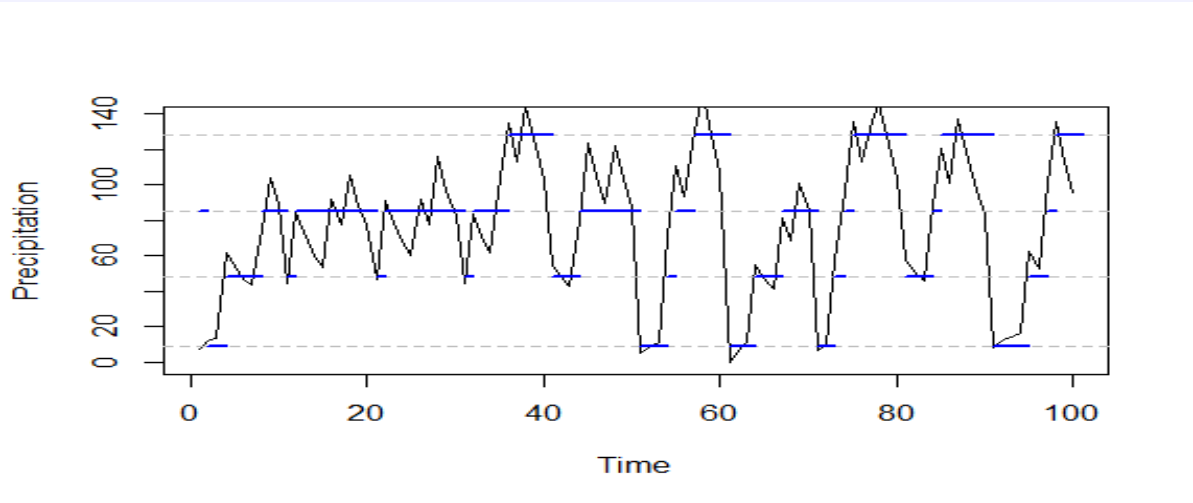
3.3 Most Probable States



(a) Most probable states for the entire time series



(b) Most probable states for the first 100 times.



(c) Combined observed time series and most probable states for each time point.

Figure 4: Visualization of most probable states and their relationship with observed data across different time intervals.

Table 4 shows the number of observations in each state which was obtained via the Viterbi algorithm. State 3 has the highest number of observations while state 1 has the fewest. The most probable states for the entire time series, as shown in Figure 4a, reveal frequent switching across the four identified states, indicating dynamic transitions between varying precipitation conditions. Despite this variability, certain states are relatively longer-lived, signifying periods of more stable weather patterns. This long-term perspective underscores the HMMs ability to capture the temporal structure and variability in precipitation by segmenting the data into distinct states.
Figure 4b focuses on the first 100 time points, offering a closer look at state transitions over a shorter period. Here, distinct clusters of time points dominated by specific states become apparent. This granularity highlights that while some states persist briefly before transitioning, others exhibit longer durations, reflecting short-term fluctuations in precipitation patterns.
Figure 4c overlays the observed precipitation time series with the most probable states, demonstrating a clear correspondence between precipitation levels and state classifications. Higher precipitation values consistently align with states that have higher mean parameters, while lower values are associated with states having smaller means. This alignment validates the model's capacity to appropriately map observed data to the most suitable states based on their emission distributions.

4 Results: Actual Data

To extend this study, we obtained actual rainfall data from the Bungoma Meteorology Office in western Kenya for analysis. The dataset comprises monthly rainfall measurements (in mm) spanning from 1981 to 2015.

4.1 Data Exploration

Figure 5 presents a time series plot illustrating precipitation trends in the region throughout the study period, revealing notable variations between months with low and high rainfall. Meanwhile, Figure 6a shows a skewed distribution, indicating that a single Gaussian distribution may not adequately capture the data, making FMMs a natural consideration. However, Figure 6b highlights significant serial dependencies in the rainfall data. A strong positive correlation of approximately 40% is observed at a lag of one month, with a weaker positive correlation of about 20% at a lag of nine months. A negative correlation emerges at 14 lags, followed by a resurgence of positive correlation at 19 lags, suggesting cyclical precipitation patterns. As demonstrated in the simulation study, these dependencies make HMMs a particularly suitable analytical tool for this dataset.

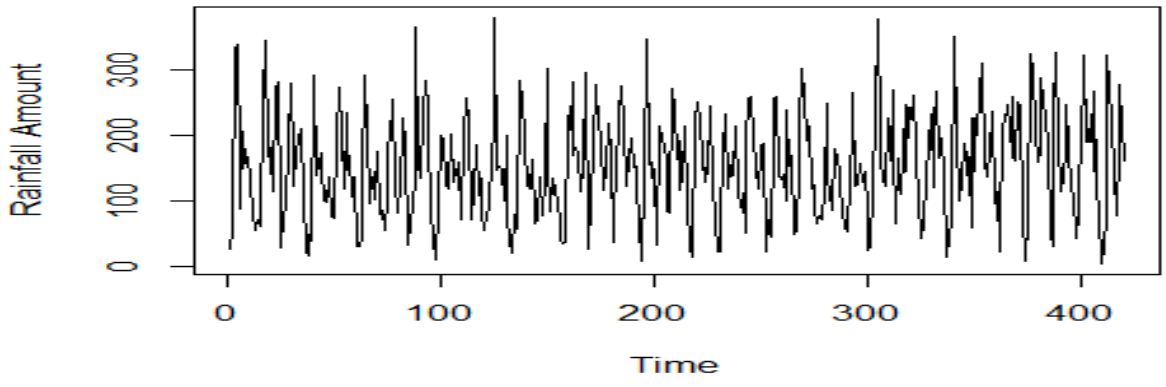
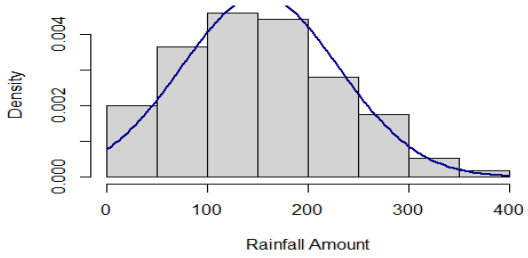
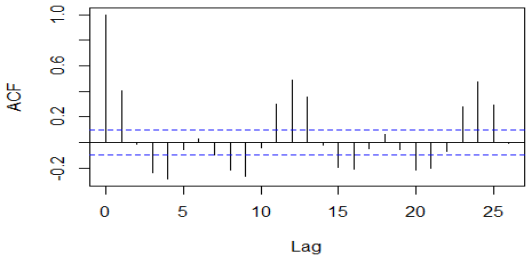


Figure 5: Time series plot of Actual data.



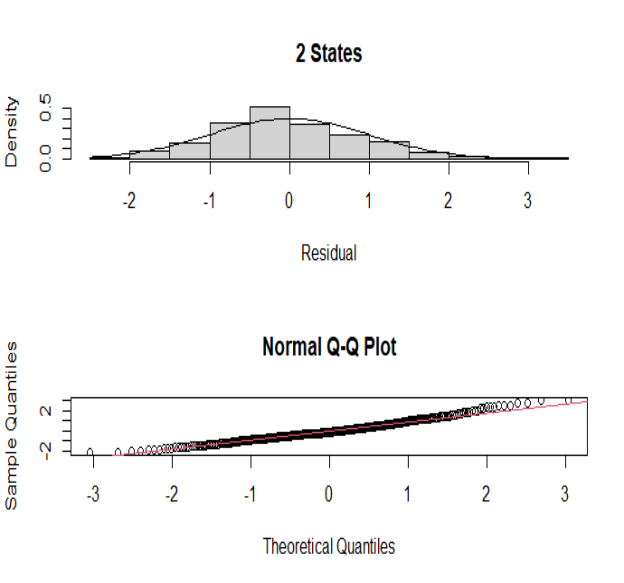
(a) Histogram of Observed Rainfall



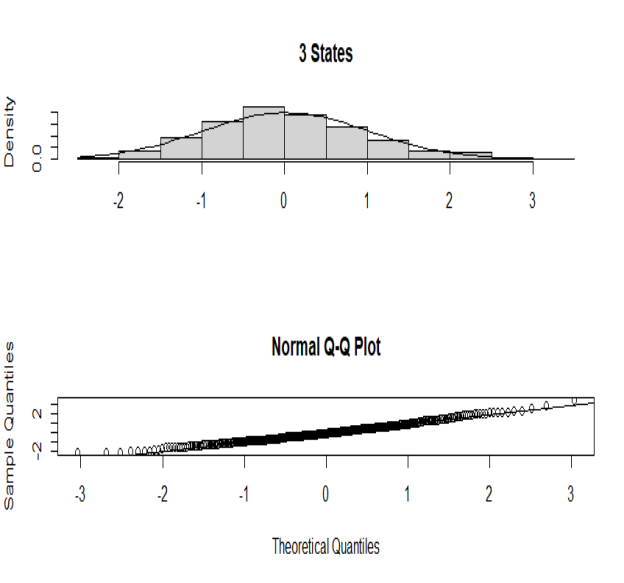
(b) Autocorrelation Function

4.2 Hidden Markov Models

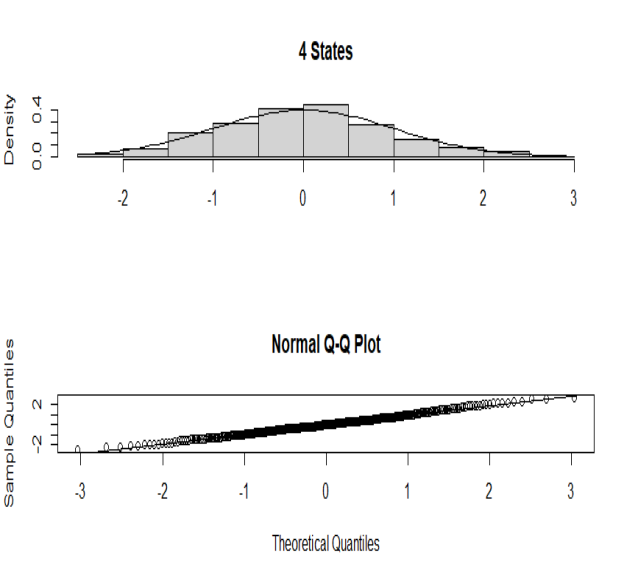
In this case, we fitted HMMs with 2, 3 and 4 states and as observed in table 5, both the AIC and BIC chose the model with 3 states. Diagnostics checks in figure 7 further confirms that indeed a 3-state HMM is sufficient for the data. This aligns with the research done by Marangu et al.[19], where they describe Bungoma region to have 3 rainfall seasons: The March, April and May (MAM) season, July and August (JA) season and October, November and December (OND) season.



(a) Residuals and Q-Q Plot for 2 States



(b) Residuals and Q-Q Plot for 3 States



(c) Residuals and Q-Q Plot for 4 States

Figure 7: Ordinary pseudo-residuals and their corresponding Q-Q plots for models with 2, 3, and 4 states.

Table 5: Model Selection Criteria

No. States	k	**-log** L	**AIC**	**BIC**
2	7	2385.47	4784.93	4813.21
3	14	2351.79	**4731.59**	**4788.15**
4	23	2344.01	4734.02	4826.94

Table 6: Initial State Probabilities and Response Parameters

i	**S1**	**S2**	**S3**
δ_i	1.000	0.000	0.000
μ_i	58.85	151.62	215.95
σ_i	28.03	43.50	64.06

Table 7: Transition Matrix for Hidden Markov Model

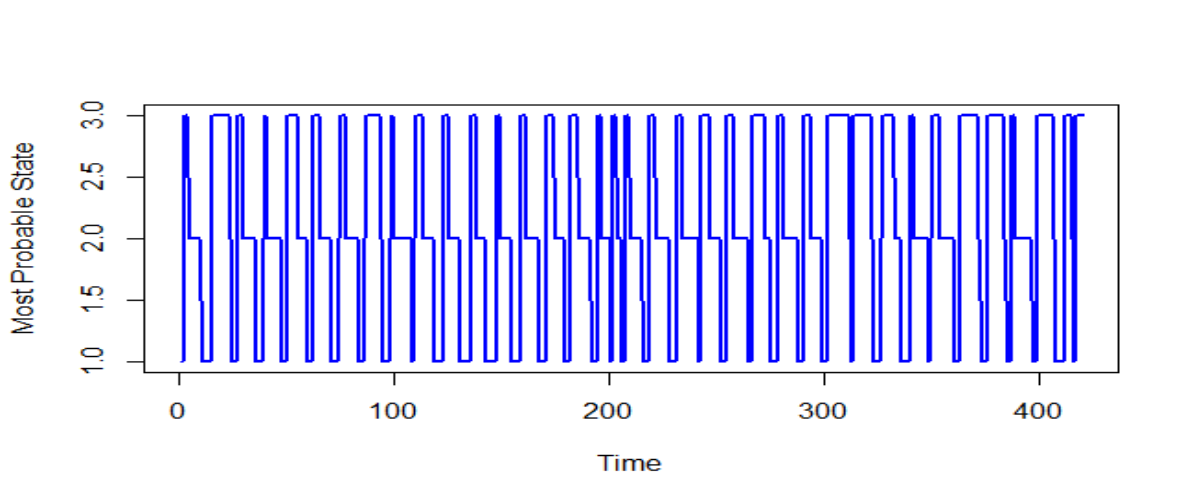
State	State 1	State 2	State 3
1	0.644	0.000	0.356
2	0.235	0.765	0.000
3	0.000	0.240	0.760

Table 8: Number of corresponding observations in each state.

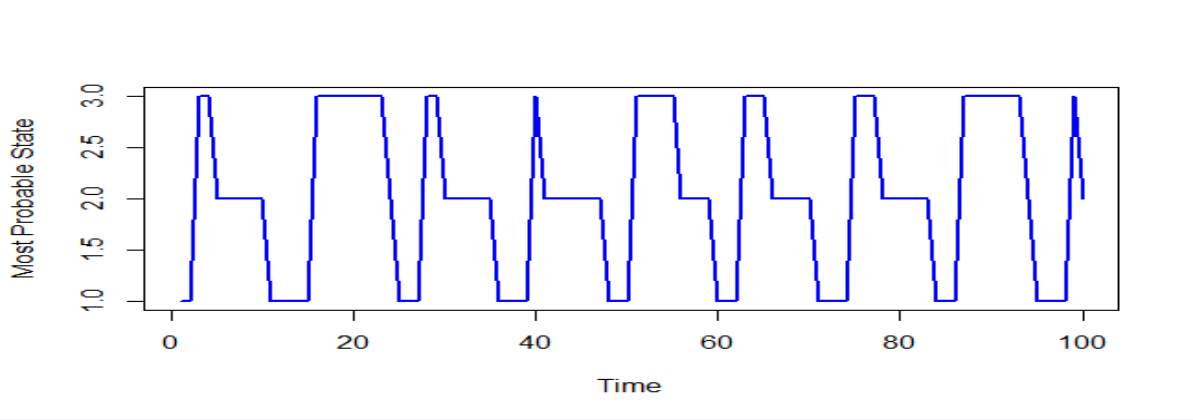
State	S_1	S_2	S_3
Obs	108	164	148

We observe distinct means and standard deviations for the Gaussian emission distribution in each state, with state 3 having the highest values and state 1 the lowest. This suggests that investors might allocate more capital during the third season. However, a key question arises: which months correspond to season 3? This is where the Viterbi algorithm provides clarity, as we will see in the next section. Additionally, the high self-transition probabilities of states 2 and 3 (Table 7) indicate that these seasons tend to persist for extended periods.

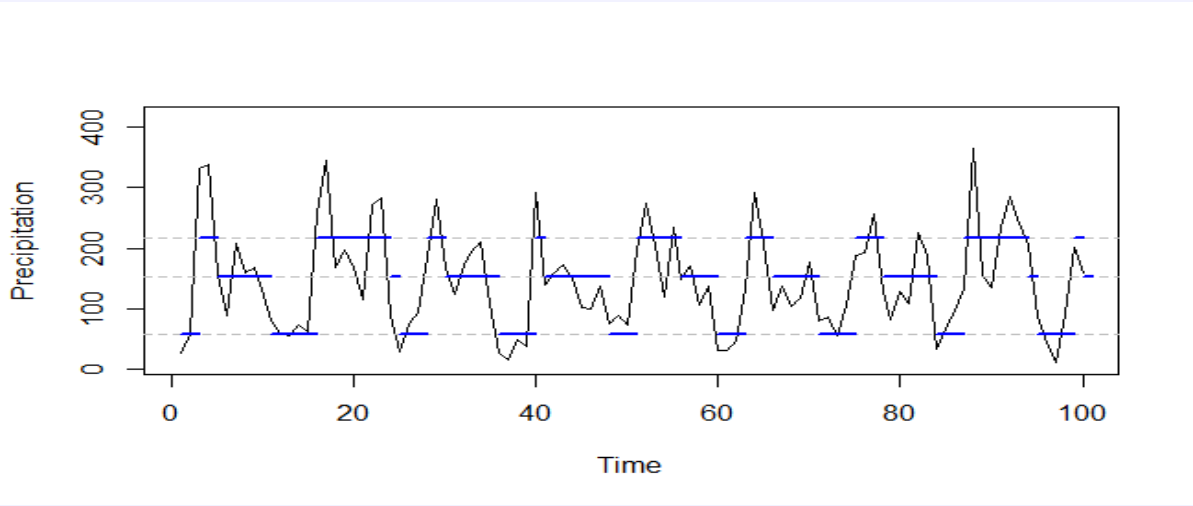
4.3 Most Probable States



(a) Most probable states for the entire time series



(b) Most probable states for the first 100 Months.



(c) Combined observed time series and most probable states for each Month.

Figure 8: Visualization of most probable states and their relationship with observed data across different Months.

Table 8 reveals that season 2 spans the highest number of months, while state 1 has the least. This suggests that, on average, season 2 persists for longer periods, whereas season 1 is relatively short-lived. As previously discussed, investors are particularly interested in identifying the months with the highest rainfall. Although forecasting would improve this insight, the current results already offer valuable information. Figures 8a and 8b illustrate the most probable states for each month. For example, January 1981 corresponds to state 1, which aligns with the region's typical pattern of little to no rainfall during this time. The visibility is made clearer by looking at the combined figure of the observed time series and the corresponding most probable states through global decoding.

5 Discussion

This research seeks to investigate precipitation data modeling using HMMs with an emphasis on heterogeneity, serial dependence and hidden states. While previous research has examined various statistical approaches to rainfall modeling, applying HMMs to closely approximate the broad dynamics of precipitation is still largely unexplored. Our study bridges this gap by presenting a detailed overview of HMMs as a robust method for modeling precipitation variability that is important for agricultural planning and disaster risk management.

Various studies have opened doors to learning about the dynamics of rainfall and its effects. For instance, Amjad et al. (2022)[20] applied FMMs in the analysis of precipitation data in Ireland without highlighting its observation independence assumption which in most cases is not verified in time series data. Similarly, Liu et al. (2020)[21] built autoregressive models for rainfall prediction in South Carolina but ignored the possible influence of latent states on precipitation processes in their work. This work builds on their work by showing the benefit of HMMs in properly capturing the serial dependence present in precipitation data and therefore achieving a more complex representation of precipitation processes.

In our study, we used both simulation-based and actual data methods to create synthetic precipitation data that reflects realistic features such as seasonality, multimodality, and autocorrelation. The HMM analysis revealed four weather states in the simulated data and three in the actual data, while also capturing the dynamics of state transitions. The study further illustrated the importance of model selection using data driven criteria like the AIC and BIC. For instance, although the AIC is underrated for under-penalizing the model's complexity, it identified the same number of states with the BIC in both cases. This could be due to the strong evidence in the data for corresponding number of states. We further demonstrated that the model fit under HMM is verified using the Ordinary pseudo residuals. Deviation from characteristics of a standard Normal distribution would signal that the chosen model does not fit well the data. However, our analysis is limited by the availability of actual data only up to 2015. This constraint may affect the model's applicability to more recent climate patterns, highlighting the need for updated data to improve accuracy and reliability. Furthermore, although our model successfully detects latent states, it is perhaps not complete regarding all exogenous variables affecting rainfall, like geographical variation and the effects of climate change.

Our results show major variation in precipitation levels between the delineated states, with state switching representing higher probabilities of persistence in specific weather conditions. This finding highlights the value of employing HMMs in rainfall forecasting since it improves our capacity to predict precipitation pattern changes and guides agricultural and water resources decision-making. Our research contributions go beyond theoretical gains; they have real-world applications for climate adaptation and disaster risk reduction stakeholders.

The ethical concerns related to our study involve the judicious application of limited data and the recognition of the limitations inherent in modeling complicated environmental processes. Care should be taken to ensure that our results are presented clearly to prevent misinterpretation, especially in policy formulation and resource allocation. FMMs can be applied if the process of generating the current observations does not explicitly depend on the distribution of the previous observation. If serial dependence and heterogeneity are observed, then HMMs offer a better solution.

Subsequent studies can build upon our findings using current precipitation records to compare the performance of HMMs under different geographic environments. A search into merging HMMs with other machine-learning approaches could help increase predictability and allow us to learn more about the forces that underpin variability in rainfall. Through solving these problems, subsequent research can help build on the current understanding of the precipitation processes as well as their significance to resilience and climate-based development.

6 Conclusion

In conclusion, this research highlights the importance of HMMs in modeling precipitation data, addressing the complexities associated with rainfall processes. By capturing heterogeneity and serial dependencies characteristic of precipitation, HMMs provide a robust framework for understanding and predicting rainfall patterns. Our findings demonstrate significant variations in precipitation levels across different states, emphasizing the potential of HMMs to enhance forecasting accuracy and inform decision-making in agriculture and water resource management. Despite the limitations posed by the use of limited actual data, the insights gained from our study pave the way for future research to explore recent precipitation records and further refine HMM applications. This work not only contributes to the theoretical landscape of statistical modeling but also offers practical implications for climate adaptation and disaster risk reduction, underscoring the importance of advanced modeling techniques in addressing pressing environmental challenges.

References

[1] V. Masson-Delmotte, P. Zhai, S. Pirani, C. Connors, S. Péan, N. Berger, Y. Caud, L. Chen, M. Goldfarb, and P. M. Scheel Monteiro, "Ipcc, 2021: Summary for policymakers. in: Climate change 2021: The physical science basis. contribution of working group i to the sixth assessment report of the intergovernmental panel on climate change," 2021.

[2] S. Chakraborty, S. Birmal, P. K. D. Burman, A. Datye, F. A.A., A. G.H., P. Mohan, N. Trivedi, and R. K. Trivedi, "Statistical analysis of the precipitation isotope data with reference to the indian subcontinent," in *Hydrology* (T. V. H. II and P. Rao, eds.), ch. 7, Rijeka: IntechOpen, 2020.

[3] D. L. Finney, J. H. Marsham, D. P. Walker, C. E. Birch, B. J. Woodhams, L. S. Jackson, and S. Hardy, "The effect of westerlies on east african rainfall and the associated role of tropical cyclones and the Madden–Julian oscillation," *Q. J. R. Meteorol. Soc.*, vol. 146, pp. 647–664, Jan. 2020.

[4] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[5] O. A. Kabbaj, L.-M. PÃ©an, J.-B. Masson, B. Marhic, and L. Delahoche, "Occupancy states forecasting with a hidden markov model for incomplete data, exploiting daily periodicity," *Energy and Buildings*, vol. 287, p. 112985, 2023.

[6] A. J. Burgess-Hull, "Applied example,"

[7] N. Gupta and S. R. Chavan, "Characterizing the tail behaviour of daily precipitation probability distributions over India using the obesity index," *International Journal of Climatology*, vol. 42, pp. 2543–2565, Mar. 2022.

[8] O. K. Oduor, "Application of finite univariate gaussian mixture models for high-frequency financial data modelling," *Asian J. Prob. Stat.*, vol. 27, pp. 81–95, Jan. 2025.

[9] O. Stoner and T. Economou, "An advanced hidden markov model for hourly rainfall time series," *Computational Statistics Data Analysis*, vol. 152, p. 107045, 2020.

[10] Y. Yu, "MixR: An R package for finite mixture modeling for both raw and binned data," *J. Open Source Softw.*, vol. 7, p. 4031, Jan. 2022.

[11] G. Verbeke and G. Molenberghs, *Advanced Modelling Techniques, Master in Statistics*. Diepenbeek, Belgium: Hasselt University, 2021.

[12] D. Böhning, P. Schlattmann, and B. Lindsay, "Computer-assisted analysis of mixtures (ca man): statistical algorithms," *Biometrics*, pp. 283–303, 1992.

[13] O. Cappe, E. Moulines, and T. RydÃ©n, *Inference in Hidden Markov Models*. 01 2005.

[14] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state markov chains," *The annals of mathematical statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[15] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[16] W. Zucchini, I. L. MacDonald, and R. Langrock, *Hidden Markov Models for Time Series: An Introduction Using R*. Boca Raton, FL: Chapman and Hall/CRC, 2nd ed., 2017.

[17] F. Urso, A. Abbruzzo, M. Chiodi, and M. F. Cracolici, "Model selection for mixture hidden markov models: an application to clickstream data," *Stat. Pap. (Berl)*, Oct. 2024.

[18] G. Celeux and J.-B. Durand, "Selecting hidden markov model state number with cross-validated likelihood," *Computational Statistics*, vol. 23, pp. 541–564, 2008.

[19] D. M. Marangu and S. W. Wanyonyi, "Establishing the rainfall trend over bungoma region in kenya, during the short and long rain seasons,"

[20] A. Hussein and S. K. Kadhem, "Spatial mixture modeling for analyzing a rainfall pattern: A case study in ireland," *Open Engineering*, vol. 12, no. 1, pp. 204–214, 2022.

[21] H. Liu, D. B. Hitchcock, and S. Z. Samadi, "Spatio-temporal analysis of flood data from south carolina," *Journal of Statistical Distributions and Applications*, vol. 7, pp. 1–19, 2020.