Climate Change: Use of Non-Homogeneous Poisson Processes for Climate Data in the Presence of a Change-Point

Viktoriia Kharchenko

March 2024

1 Introduction

Climate change is a phenomenon that touches every corner of the globe, altering weather patterns and the rhythm of natural systems [8], [2]. Monitoring these shifts is vital, and it involves a close study of climate variables like precipitation, temperature, and sea levels. The focus is particularly on instances when these variables notably differ from their average, such as unusually high or low rainfall or temperature spikes or dips. Historical records reveal that the earth's average temperature has risen, especially pronounced since the mid-20th century [2].

This project report presents a review and extension of a study that employed non-homogeneous Poisson processes (NHPP) to analyze climate data [4]. The original study utilizes NHPP to examine climate data from Kazakhstan and Uzbekistan[12], and the USA [1], focusing on variations in yearly average precipitation, average temperature, and yearly average maximum temperature observed since the late 19th century. My replicated research analyzes the climate data from Kazakhstan and Uzbekistan[12] as detailed in the paper, and additionally broadens its scope by incorporating data on Ukraine's average air temperature data from 1891 to 2022 collected by the Boris Sreznevsky Central Geophysical Observatory [3] - a dataset not previously examined. The study aimed to understand the frequency of years where climate variables surpassed certain established thresholds which were determined based on the overall average measurements for each climate variable. The research explored two versions of the NHPP model: one without considering change points and one that includes a change point to account for shifts in climate patterns. In light of this, the original study estimates the model parameters under a Bayesian approach using standard Markov chain Monte Carlo (MCMC) methods, notably Gibbs sampling [10], using the OpenBugs software [11] to simulate the MCMC samples.

Building upon the foundation laid by the original study, this project replicates the analysis by implementing MCMC using Metropolis-Hastings algorithm. Instead of using OpenBugs, this project takes a more hands-on approach to estimating model parameters. The Metropolis-Hastings algorithm [5] is a cornerstone of the MCMC method and is widely used in statistical computations. By recreating the study's methodology, this project not only

verifies the original findings but also deepens my understanding of MCMC techniques and their applications.

2 Methodology

2.1 Non-homogeneous Poisson Models

A Poisson process is a stochastic process that models a series of events occurring in a fixed period, where these events happen with a known mean rate and independently of the time since the last event. The classical Poisson process assumes homogeneity, meaning the expected number of events occurring is proportional to the length of the time interval, with the rate (λ) being constant over time:

$$P(N(t) = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!} \tag{1}$$

However, this assumption does not hold in many practical situations where the rate at which events occur could vary over time, leading to the concept of the Non-Homogeneous Poisson Process (NHPP).

In an NHPP, the rate function $\lambda(t)$ varies with time, offering a more flexible and realistic model for data with time-dependent patterns. It allows the intensity of events to change, reflecting periods of higher or lower activity, which is often the case with climate data. For example, the rate of extreme weather events could vary seasonally or show trends over the years, influenced by broader climatic shifts.

Studying climate data with NHPP is particularly relevant because climate variables do not change uniformly. NHPPs are adept at capturing the variability in the occurrence of climate-related events, such as exceeding a certain threshold of temperature or rainfall. This approach is crucial for understanding patterns such as the increasing frequency of extreme weather events, which can be linked to global climate change.

In this study, a non-homogeneous Poisson process (NHPP) is employed to assess the likelihood that a climate metric, such as precipitation or temperature, surpasses a certain threshold multiple times within a specified time frame. This threshold is the overall mean of the observed climate data over the period $[0, T]$, with $T >= 0$ representing the end of the observation period. Let M_t represent the count of occurrences where the climate variable exceeds this average threshold within the interval $[0, t)$, where t ranges from 0 to T. The process $M = M_t : t >= 0$ is modeled as an NHPP characterized by a rate function $\lambda(t)$ that is strictly positive and a mean function $m(t)$, defined as the integral of $\lambda(s)$ from 0 to t.

$$m(t) = \int_0^t \lambda(s)ds, \quad t \geq 0. \tag{2}$$

Analyzing climate data with non-homogeneous Poisson processes (NHPP), various parametric forms can be adopted for the rate function, which is often inspired by reliability theory. These functions are parameterized by a vector, typically denoted as θ, which influences both the rate and mean functions of the process, symbolized as $\lambda(t|\theta)$ and $m(t|\theta)$, respectively.

2.2 Parametric Forms of NHPP Models

Two common NHPP formulations were used in the paper: NHPP-1 and NHPP-2. NHPP-1 defines the mean value function as $m(t) = \alpha F(t)$, where $F(t)$ is a specified cumulative distribution function, and α is an unknown parameter to be estimated. This form was utilized in software reliability studies to represent the accumulation of software failures over time[6]. On the other hand, NHPP-2 takes a different approach by setting $m(t) = -log(1 - F(t))$ commonly applied in software reliability applications [9].

In the climate data analysis presented, five parametric models were considered to understand the patterns and frequency of climatic events exceeding specific thresholds. These models are chosen for their capability to capture various trends and behaviors in the event data over time.

1. Power Law Process (PLP): This model is rooted in the idea that the frequency of events can accelerate over time according to a power law, which is a common pattern in reliability and climate data: $F(t) = \exp\left(-\left(\frac{t}{\sigma}\right)^{\alpha}\right), \quad t > 0.$

2. Musa–Okumoto Process (MOP): Originating from software reliability, this model assumes that the occurrence rate of events decreases over time, following a Lomax or Pareto type II distribution: $F(t) = 1 - (1 - \frac{t}{\alpha})^{-\beta}, \quad t > 0.$

3. Goel–Okumoto Process (GOP): Another model from the field of software reliability, the GOP describes situations where events happen at a rate that diminishes exponentially over time: $F(t) = 1 - \exp(-\beta t), \quad t > 0.$

4. Generalized Goel–Okumoto Process (GGOP): As an extension of the GOP, this model allows for a more flexible Weibull distribution, accommodating various patterns of event rates: $F(t) = 1 - \exp(-\beta t^{\gamma}), \quad t > 0.$

5. Exponentiated-Weibull Process (GPLP): This model generalizes the PLP by incorporating an exponentiated-Weibull distribution, offering a richer behavior description, which includes increasing, decreasing, and bathtub-shaped event rates: $F(t) = 1 - \exp\left[-(\frac{t}{\sigma})^{\alpha}\right]^{\beta}, \quad t > 0$.

Each of these models uses a different formulation for the mean value function (NHPP-2 for PLP, MOP, and GPLP models, and NHPP-1 for GOP and GGOP) which in turn influences the corresponding mean function, $m(t)$, determining the event's intensity over time.

$$
\begin{aligned}
m_{PLP}(t|\theta) &= \left(\frac{t}{\sigma}\right)^{\alpha}, && \text{where } \theta = (\alpha, \sigma); \alpha, \sigma > 0, \\
m_{MOP}(t|\theta) &= \beta \log(1 + \frac{t}{\alpha}), && \text{where } \theta = (\alpha, \beta); \alpha, \beta > 0, \\
m_{GOP}(t|\theta) &= \alpha[1 - \exp(-\beta t)], && \text{where } \theta = (\alpha, \beta); \alpha, \beta > 0, \\
m_{GGOP}(t|\theta) &= \alpha[1 - \exp(-\beta t^{\gamma})], && \text{where } \theta = (\alpha, \beta, \gamma); \alpha, \beta, \gamma > 0, \\
m_{GPLP}(t|\theta) &= -\log[1 - F_{EW}(t)], && \text{where } \theta = (\alpha, \beta, \sigma); \alpha, \beta, \sigma > 0,
\end{aligned}
\tag{3}
$$

With the cumulative function for exponentiated-Weibull:$F_{EW}(t) = 1 - \exp[-(t/\sigma)^{\alpha}]^{\beta}$.

$$
\begin{aligned}
\lambda_{PLP}(t|\theta) &= (\alpha/\sigma)(t/\sigma)^{\alpha-1}, && \text{where } \theta = (\alpha, \sigma); \alpha, \sigma > 0, \\
\lambda_{MOP}(t|\theta) &= \beta/(t+\alpha), && \text{where } \theta = (\alpha, \beta); \alpha, \beta > 0, \\
\lambda_{GOP}(t|\theta) &= \alpha\beta \exp(-\beta t), && \text{where } \theta = (\alpha, \beta); \alpha, \beta > 0, \\
\lambda_{GGOP}(t|\theta) &= \alpha\beta\gamma t^{\gamma-1} \exp(-\beta t^{\gamma}), && \text{where } \theta = (\alpha, \beta, \gamma); \alpha, \beta, \gamma > 0, \\
\lambda_{GPLP}(t|\theta) &= G(t)/[1 - F_{EW}(t)], && \text{where } \theta = (\alpha, \beta, \sigma); \alpha, \beta, \sigma > 0,
\end{aligned} \tag{4}
$$

With the function G(t): $G(t) = \alpha\beta\sigma^{-1}[1 - \exp(-(t/\sigma)^{\alpha})]^{\beta-1} \exp[-(t/\sigma)^{\alpha}](t/\sigma)^{\alpha-1}$.

In the scope of this project, the Power Law Process (PLP) model was replicated as it demonstrated the most promising results among the various models tested.

2.3 Power Law Process

The PLP model is a specific case of the Non-Homogeneous Poisson Process (NHPP) where the mean value function $m(t|\theta)$ and the intensity or rate function $\lambda(t|\theta)$ are derived from the Weibull distribution. Weibull distribution is a generalization of the exponential distribution. The Weibull cumulative distribution function (CDF) is $F(t; \sigma, \alpha) = 1 - \exp[-(\frac{t}{\sigma})^{\alpha}]$, it gives the probability that the random variable T, which is the time until failure, is less than or equal to a particular time t. The probability density function (PDF) describes the likelihood of failure at exactly time t: $f(t; \sigma, \alpha) = \frac{\alpha}{\sigma}\left(\frac{t}{\sigma}\right)^{\alpha-1} \exp[-\left(\frac{t}{\sigma}\right)^{\alpha}]$. The Weibull distribution is chosen due to its flexibility in modeling various types of failure rates, which in this context translates to the occurrence rates of climate events exceeding the threshold.

The hazard function, also known as the hazard rate, is a fundamental concept in survival analysis and reliability engineering. It quantifies the instantaneous rate of failure or occurrence of an event at a certain time, assuming that the event has not yet occurred up until that time. It can be defined through the PDF $f(t)$ and the survival function $S(t)$: $h(t) = \frac{f(t)}{S(t)}$. where $S(t) = P(T > t)$ is the probability that the event has not occurred by time t,$f(t)$ gives the unconditional probability of the event occurring at exactly time t. In the case of Weibull distribution: $S(t) = 1 - F(t) = \exp\left[-\left(\frac{t}{\sigma}\right)^{\alpha}\right]$. Then the hazard rate is given by: $h(t) = \frac{f(t)}{S(t)} = \frac{\frac{\alpha}{\sigma}\left(\frac{t}{\sigma}\right)^{\alpha-1} e^{-\left(\frac{t}{\sigma}\right)^{\alpha}}}{e^{-\left(\frac{t}{\sigma}\right)^{\alpha}}} = \frac{\alpha}{\sigma}\left(\frac{t}{\sigma}\right)^{\alpha-1}$.

In the context of climate science, if we are modeling the time until a certain extreme weather event occurs, the hazard function could represent the risk of that event occurring at any given time, given that it hasn't happened yet. It can represent the instantaneous rate $\lambda(\alpha, \sigma)$ of event occurrences in a process where past events do not influence future occurrences.

$$
\lambda_{\text{PLP}}(t|\theta) = (\frac{\alpha}{\sigma})(\frac{t}{\sigma})^{\alpha-1}. \tag{5}
$$

For the PLP model, the mean value function is given by:

$$
m_{\text{PLP}}(t|\theta) = \int_0^t \frac{\alpha}{\sigma}\left(\frac{u}{\sigma}\right)^{\alpha-1} du = \left(\frac{t}{\sigma}\right)^{\alpha} \tag{6}
$$

Here, α is the shape parameter and σ is the scale parameter of the Weibull distribution.

Applying the PLP model to climate data, the parameters α and σ were estimated using the Markov Chain Monte Carlo (MCMC) method.

2.4 Hon-homogeneous Poisson Process with a Change Point

The concept of a non-homogeneous Poisson process (NHPP) with a change point is used to model situations where the rate of occurrence of events changes at some unknown point in time. This is particularly useful in analyzing climate data, where a change point could indicate a shift in climate patterns due to various factors, such as environmental policies or significant natural events. For the model without a change point, the rate function is consistent over time. However, when a change point τ is present, the rate function has two different sets of parameters: one set for the time before the change point (θ_1) and another for the time after the change point θ_2. The model assumes that at time $t = \tau$, the rate function shifts from $\lambda(t|\theta_1)$ to $\lambda(t|\theta_2)$.

$$\lambda(t|\theta) = \begin{cases} \lambda(t|\theta_1) & \text{if } 0 \leq t \leq \tau, \\ \lambda(t|\theta_2) & \text{if } t > \tau, \end{cases} \tag{7}$$

When there is a change point in the process, the mean value function, which represents the expected number of events by time t, is a combination of the mean value functions before and after the change point.

$$m(t|\theta)_j = \begin{cases} m(t|\theta_1) & \text{if } 0 \leq t \leq \tau, \\ m(\tau|\theta_1) + m(t|\theta_2) - m(\tau|\theta_2) & \text{if } t > \tau. \end{cases} \tag{8}$$

In Equation 8 $m(\tau|\theta_1)$ represents the expected number of events up to the change point, while $m(t|\theta_2) - m(\tau|\theta_2)$ - the additional expected number of events from the change point τ to the current time t under the new rate parameter set θ_2, we subtract $m(\tau|\theta_2)$ since want to count events starting from τ rather than from time 0.

For the power law process (PLP) with a change point, the rate function is given by two different power-law expressions based on the parameters before and after the change point:

$$\lambda(t|\theta) = \begin{cases} \frac{\alpha_1}{\sigma_1}\left(\frac{t}{\sigma_1}\right)^{\alpha_1 - 1} & \text{if } 0 \leq t \leq \tau, \\ \frac{\alpha_2}{\sigma_2}\left(\frac{t}{\sigma_2}\right)^{\alpha_2 - 1} & \text{if } t \geq \tau. \end{cases} \tag{9}$$

The mean value function for the PLP model with a change point follows similarly:

$$m(t|\theta) = \begin{cases} \left(\frac{t}{\sigma_1}\right)^{\alpha_1} & \text{if } 0 \leq t \leq \tau, \\ \left(\frac{t}{\sigma_1}\right)^{\alpha_1} + \left(\frac{t}{\sigma_2}\right)^{\alpha_2} - \left(\frac{\tau}{\sigma_2}\right)^{\alpha_2} & \text{if } t \geq \tau. \end{cases} \tag{10}$$

2.5 Likelihood Function Without the Presence of Change-Points

In constructing the likelihood function for the model, the authors considered the years when a climate variable exceeds a pre-established threshold.

To quantify this, the count of years within the total observation period, $[0, T]$, where these exceedances have occurred, was considered. The observation period begins at time 0 and extends to time T. We denote the years when these exceedances occur as $\{t_1, t_2, ..., t_n\}$, with each t_i representing a specific point in time within our observation period where an exceedance was recorded ($0 < t_1 < t_2 < ... < t_n < T$).

The dataset D_T, containing the observed values, can be expressed as $\{n; t_1, ..., t_n; T\}$. This set includes the total count of exceedance events n, the specific times t_i when each event was observed, and the total length of the observation window T. This dataset forms the basis of the likelihood function, which we will use to estimate the parameters of the model and understand the underlying patterns of climate exceedances.

The likelihood function should quantify how likely the observed data are, given certain values of the model parameters. It can be defined by calculating the probability of observing zero events in the interval $(0, t1)$, one event in $(t1, t1 + \delta t1)$, no events in $(t1 + \delta t1, t2)$, one event in $(t2, \delta t2)$, and so on. The probability of observing zero events by time t is then given by the NHPP: $P(N(t) = 0) = e^{-\int_0^t \lambda(s)\, ds}$. As $\delta \to 0$: the expectation of observing exactly one event in a very small interval approximates to $\lambda(t)\delta$ because the expression $e^{-\lambda(t)\delta}$ approaches 1, simplifying the formula. In the limit as $\delta \to 0$, for a very small interval around each t_i, the likelihood of observing an event is very small, but because we did observe events at these times t_i, we plug these observed times into the rate function to calculate the likelihood of the observed pattern of events. This way the instantaneous rates $\lambda(t_i; \theta)$ contribute to the probability of the observed data when considering the exact times at which the events occurred. Combining the likelihood of the individual event times with the probability of no further events gives us the likelihood function for the NHPP without change points:

$$L(\theta|D_T) = \lambda(t_1)e^{-\int_0^{t_1} \lambda(t)\, dt} \cdot \lambda(t_2)e^{-\int_{t_1}^{t_2} \lambda(t)\, dt} \cdots \lambda(t_n)e^{-\int_{t_{n-1}}^{T} \lambda(t)\, dt}$$

This is equivalent to the defined in Equation 11 likelihood formula.

$$L(\theta|D_T) = \left[\prod_{i=1}^{n} \lambda(t_i|\theta)\right] e^{-\int_0^T \lambda(t|\theta)dt} = \left[\prod_{i=1}^{n} \lambda(t_i|\theta)\right] e^{-m(T|\theta)} \tag{11}$$

For computational stability and convenience, I worked with the log-likelihood:

$$\log L(\theta \mid D_T) = \sum_{i=1}^{n} \log \lambda(t_i \mid \theta) - m(T \mid \theta) \tag{12}$$

2.6 Likelihood Function in the Presence of a Change-Point

The likelihood function for a non-homogeneous Poisson process (NHPP) with a change point involves partitioning the observation period at the change point τ. This partition distinguishes between the events occurring before and after the change point, each with different intensity functions.

For n total exceedances observed within the time interval $[0, T]$, we have the occurrence times given by $D_T = t_1, ..., t_N$. With a change-point τ, we can specify N_τ as the number of

events occurring before the change-point. Thus, the times $t_1, ..., t_{N_\tau}$ are associated with the first segment of the process (from 0 to τ) and times $t_{N_\tau+1}, ..., t_n$ with the second segment (from τ to T).

The likelihood function $L(\theta|D_T)$ for this NHPP with one change-point can be expressed as the product of two terms: one for each segment of the process. It accounts for the different rates at which events occur before and after the change point. Mathematically, the likelihood is given by:

$$L(\theta|D_T) = \left[\prod_{i=1}^{N_\tau} \lambda(t_i|\theta_1)\right] \exp\left[-m(\tau|\theta_1)\right] \times \left[\prod_{i=N_\tau+1}^{n} \lambda(t_i|\theta_2)\right] \exp\left[-m(T|\theta_2) + m(\tau|\theta_2)\right] \quad (13)$$

The log-likelihood function $L(\theta|D)$ is the sum of the logs of the intensity functions minus the mean value functions evaluated at the change point and the end of the observation period:

$$\log L(\theta|D_T) = \sum_{i=1}^{N_\tau} \log \lambda(t_i|\theta_1) - m(\tau|\theta_1) + \sum_{i=N_\tau+1}^{n} \log \lambda(t_i|\theta_2) - \left[m(T|\theta_2) - m(\tau|\theta_2)\right] \quad (14)$$

2.7 Markov chain Monte Carlo Simulation

If we return to the defined earlier PLP model, we are interested in estimating α and σ parameters, which determine the shape and scale of the process. The goal is to update our beliefs about these parameters based on combining the prior distribution with the observed data via the likelihood function. Marginal posterior densities allow us to understand the probability distribution of each parameter independently, integrating the effects of other parameters. This is particularly useful in complex models where parameters are interdependent, and we're interested in the distribution of a single parameter while accounting for the uncertainty in others.

However, it is often difficult, or even impossible, to derive analytical expressions for these marginal posterior densities. The reason is the complexity of the likelihood function associated with the model. For many models, especially those that are non-linear or involve many parameters, this function can be so complex that it's intractable to compute exact solutions. Instead, we need to use numerical methods to approximate these densities. One common strategy is Markov chain Monte Carlo (MCMC) simulation, which can sample from the posterior distribution even when it can not be calculated directly.

In the original study, the authors utilized MCMC method based on Gibbs sampling using OpenBUGS software to estimate the parameters of the five defined in 3 and 4 models: the Power Law Process (PLP), Musa-Okumoto process (MOP), Goel-Okumoto process (GOP), Generalized Goel-Okumoto process (GGOP), and the Exponentiated-Weibull process (GPLP). OpenBUGS is a powerful tool for performing Bayesian analysis using MCMC methods. It allows users to specify models directly in terms of probability distributions and to automate the process of drawing samples from the posterior distribution of the parameters.

Below is a brief description of the Gibbs sampling algorithm:

1. Suppose $\pi(\theta \mid y)$ is a joint posterior distribution, where $\theta = (\theta_1, \ldots, \theta_k)$, from which we want to draw inferences.

2. For each parameter θ_i, we update its value by sampling from its conditional distribution given the current values of all other parameters. This update is performed in a sequential manner, one parameter at a time:

   - Generate $\theta_1^{(1)}$ from $\pi(\theta_1 \mid y, \theta_2^{(0)}, \ldots, \theta_k^{(0)})$, which is the conditional distribution of θ_1 given the data y and the current values of the other parameters.
   - For the second parameter, generate $\theta_2^{(1)}$ from $\pi(\theta_2 \mid y, \theta_1^{(1)}, \theta_3^{(0)}, \ldots, \theta_k^{(0)})$.
   - Continue this process for all parameters up to θ_k.

3. After updating the value of the last parameter θ_k, one full iteration is completed. The new set of parameters $\Theta^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}, \ldots, \theta_k^{(1)})$ is used as the starting point for the next iteration.

4. This iterative process is repeated for a large number of iterations. As the number of iterations increases, the distribution of the parameter vector Θ converges to the joint posterior distribution $\pi(\theta \mid y)$.

In contrast to the original study where Gibbs sampling was utilized for parameter estimation, my project adopted a different approach. I implemented the Metropolis-Hastings algorithm, a widely recognized MCMC method, from scratch within the R programming environment. To align with the original findings and enable a comparative analysis, the prior distributions defined by the authors were adopted in my study.

Below is a simplified overview of how MCMC Metropolis-Hastings algorithm works:

1. Start with an arbitrary point x_0 from the sample space as the initial state of the Markov chain.

2. Propose a new state x' based on a proposal distribution $q(x'|x)$, which suggests a new sample given the current sample x.

3. Calculate the acceptance probability $A(x, x')$, which determines the likelihood of moving to the proposed state x' from the current state x. This probability is calculated as:

$$A(x, x') = \min\left(1, \frac{p(x')q(x|x')}{p(x)q(x'|x)}\right)$$

   where $p(x)$ is the probability density of state x under the target distribution, and $q(x'|x)$ is the probability density of proposing state x' given the current state x under the proposal distribution.

4. Decide whether to accept the proposed state x' by generating a random number u from a uniform distribution over [0, 1] and comparing it to $A(x, x')$. If $u \leq A(x, x')$, the proposed state x' is accepted, and the Markov chain moves to x'; otherwise, it remains at x.

5. Repeat steps 2-4 for a large number of iterations to generate a sequence of samples.

The uniform distribution $U[0,1]$ is used because it is a way to make a fair decision about whether to accept the new state. The uniform distribution gives us a random number between 0 and 1 with equal probability for all outcomes in that range. This way If $A(x,x')$ is high (close to 1), there's a high chance of accepting x', suggesting that x' is a good move that increases (or at least doesn't decrease) the probability of observing the data. If $A(x,x')$ is low, there's only a small chance of moving to x', reflecting that it might not be a suitable state.

3 Project Implementation

3.1 Modeling Climate Event Occurrences

To prepare the climate data for a statistical analysis based on the NHPP , we need to set a threshold, which is the overall mean of the climate data, and then identify which years (time points $\{t_1, ..., t_n\}$) had average precipitation levels above the overall mean. The example code of preprocessing Kazakhstan data is provided in Appendix A.

To accurately capture the dynamic nature of climate data, I investigate two versions of the PLP model, with and without the presence of a change point. The following R code snippets are designed to implement these two approaches.

```r
lambda <- function(t, alpha, sigma) {
  (alpha / sigma) * (t / sigma)^(alpha - 1)
}

m_t <- function(t, alpha, sigma) {
  (t / sigma)^alpha
}
```

Code 1: R code defining rate and mean function of PLP model without change-points

The first function, lambda, calculates the rate or intensity function $\lambda(t|\theta)$ for given values of time t, shape parameter α, and scale parameter σ. This function represents the instantaneous rate at which climate events exceeding the predefined threshold occur at time t. The second function, m_t, computes the mean value function $m(t|\theta)$, which estimates the cumulative number of climate events expected to exceed the threshold up to time t.

The given below R code defines the intensity function and the mean value function for a NHPP with a change point τ describe in Section 2.4. These functions account for different parameter values before and after the change point.

```r
lambda <- function(t, theta1, theta2, tau) {
  ifelse(t < tau,
         (theta1[1] / theta1[2]) * (t / theta1[2])^(theta1[1] - 1),
         (theta2[1] / theta2[2]) * (t / theta2[2])^(theta2[1] - 1))
}

m_t <- function(t, theta1, theta2, tau) {
```

```
  ifelse(t < tau,
         (t / theta1[2])^theta1[1],
         (tau / theta1[2])^theta1[1] + ((t / theta2[2])^theta2[1] -
   (tau / theta2[2])^theta2[1]))
}

theta1 <- c(mean_alpha1, mean_sigma1)
theta2 <- c(mean_alpha2, mean_sigma2)
tau <- mean_tau
```

Code 2: R code defining rate and mean function of PLP model with a change point

3.2 MCMC implementation

The code below represents my implementation of a Markov Chain Monte Carlo (MCMC) simulation using the Metropolis-Hastings algorithm described in 2.7, designed for estimating the parameters of a model based on observed climate data.

```
# Log-likelihood function
likelihood_function <- function(alpha, sigma, t_i, T) {
  log_lambda_t_i <- log((alpha / sigma) * (t_i / sigma)^(alpha - 1))
  m_T <- (T / sigma)^alpha
  log_likelihood <- sum(log_lambda_t_i) - m_T
  return(log_likelihood)
}

mcmc_sampling <- function(t_i, T, n_iter, burn_in, thinning = 100) {

  alpha <- alpha_prior()
  sigma <- sigma_prior()

  # Calculate the number of samples to keep after burn-in and
   thinning
  n_samples <- ceiling((n_iter - burn_in) / thinning)

  # Storage for thinned samples after burn-in
  samples <- matrix(nrow = n_samples, ncol = 2)
  colnames(samples) <- c("alpha", "sigma")

  sample_count <- 0
  stored_samples <- 0

  for (i in 1:n_iter) {
    # Metropolis step for alpha
    alpha_proposed <- alpha_prior()
    likelihood_current <- likelihood_function(alpha, sigma, t_i, T)
    likelihood_proposed <- likelihood_function(alpha_proposed, sigma
   , t_i, T)
```

```r

    if (log(runif(1)) < (likelihood_proposed - likelihood_current))
   {
      alpha <- alpha_proposed
    }

    # Metropolis step for sigma
    sigma_proposed <- sigma_prior()
    likelihood_current <- likelihood_function(alpha, sigma, t_i, T)
    likelihood_proposed <- likelihood_function(alpha, sigma_proposed
   , t_i, T)

    if (log(runif(1)) < (likelihood_proposed - likelihood_current))
   {
      sigma <- sigma_proposed
    }

    sample_count <- sample_count + 1

    # Store thinned samples after burn-in
    if (i > burn_in && sample_count %% thinning == 0) {
      stored_samples <- stored_samples + 1
      samples[stored_samples, ] <- c(alpha, sigma)
    }
  }

  return(samples)
}
```

Code 3: R code for MCMC Sampling Without Change Point

The *likelihood_function* calculates the log-likelihood of observing the data given a set of parameters (*alpha* and *sigma*) for the model. This function is fundamental in the MCMC Metropolis-Hastings algorithm, as it's used to compute the acceptance probability of new parameter values. The log-likelihood is used for numerical stability, especially when dealing with products of probabilities.

The *mcmc_sampling* function is the core of the MCMC simulation. It aims to sample from the posterior distribution of the model parameters given observed data (t_i, and T), where t_i represents the times when the observed data exceed the overall mean value of the dataset, and T is the total observation period.

For each parameter (*alpha* and *sigma*), prior distributions are specified through the *alpha_prior* and *sigma_prior* functions. Then within the MCMC loop, new values for the parameters (*alpha_proposed* and *sigma_proposed*) are generated from the prior distributions as proposals. The acceptance of these proposals is based on the ratio of the likelihoods of the proposed and current parameters. If the proposed set is accepted based on the acceptance probability, it replaces the current set of parameters; otherwise, the algorithm retains the current parameters.

To reduce autocorrelation in the sampled sequences, the algorithm includes a thinning

procedure, which only retains every 100th sample. Moreover, a burn-in period is included to allow the algorithm to reach a state where the samples are representative of the target posterior distribution. Samples collected during this period are discarded.

The code provided below is an R implementation of the MCMC Metropolis-Hastings algorithm for a model that accounts for a change point in the data.

```
likelihood_function_with_changepoint <- function(alpha1, sigma1,
    alpha2, sigma2, t_i, T, tau) {
  # Points before the change-point
  t_i_before <- t_i[t_i < tau]
  log_lambda_t_i_before <- log((alpha1 / sigma1) * (t_i_before /
    sigma1)^(alpha1 - 1))
  m_tau <- (tau / sigma1)^alpha1

  # Points after the change-point
  t_i_after <- t_i[t_i >= tau]
  log_lambda_t_i_after <- log((alpha2 / sigma2) * (t_i_after /
    sigma2)^(alpha2 - 1))
  m_T_minus_m_tau <- (T / sigma2)^alpha2 - (tau / sigma2)^alpha2

  # Compute log-likelihood
  log_likelihood <- sum(log_lambda_t_i_before) - m_tau + sum(log_
    lambda_t_i_after) - m_T_minus_m_tau
  return(log_likelihood)
}

mcmc_sampling_with_changepoint <- function(t_i, T, n_iter = 200000,
    burn_in = 11000, thinning = 100) {

  alpha1 <- alpha_prior()
  sigma1 <- sigma_prior()
  alpha2 <- alpha_prior()
  sigma2 <- sigma_prior()
  tau <- tau_prior(T)

  # Calculate the number of samples to keep after burn-in and
    thinning
  n_samples <- ceiling((n_iter - burn_in) / thinning)

  # Storage for samples
  samples <- matrix(nrow = n_samples, ncol = 5)
  colnames(samples) <- c("alpha1", "sigma1", "alpha2", "sigma2", "
    tau")

  sample_count <- 0
  stored_samples <- 0

  for (i in 1:n_iter) {
```

```
36
37        # Metropolis step for alpha1
38      alpha1_proposed <- alpha_prior()
39      likelihood_current <-likelihood_function_with_changepoint(alpha1
        , sigma1, alpha2, sigma2, t_i, T, tau)
40      likelihood_proposed <- likelihood_function_with_changepoint(
        alpha1_proposed, sigma1, alpha2, sigma2, t_i, T, tau)
41
42      if (log(runif(1)) < (likelihood_proposed - likelihood_current))
        {
43        alpha1 <- alpha1_proposed
44      }
45
46      # Metropolis step for sigma1
47      sigma1_proposed <- sigma_prior()
48      likelihood_current <- likelihood_function_with_changepoint(
        alpha1, sigma1, alpha2, sigma2, t_i, T, tau)
49      likelihood_proposed <- likelihood_function_with_changepoint(
        alpha1, sigma1_proposed, alpha2, sigma2, t_i, T, tau)
50
51      if (log(runif(1)) < (likelihood_proposed - likelihood_current))
        {
52
53        sigma1 <- sigma1_proposed
54      }
55      # Metropolis step for alpha2
56      alpha2_proposed <- alpha_prior()
57      likelihood_current <- likelihood_function_with_changepoint(
        alpha1, sigma1, alpha2, sigma2, t_i, T, tau)
58      likelihood_proposed <- likelihood_function_with_changepoint(
        alpha1, sigma1, alpha2_proposed, sigma2, t_i, T, tau)
59
60      if (log(runif(1)) < (likelihood_proposed - likelihood_current))
        {
61
62        alpha2 <- alpha2_proposed
63      }
64
65      # Metropolis step for sigma2
66      sigma2_proposed <- sigma_prior()
67      likelihood_current <- likelihood_function_with_changepoint(
        alpha1, sigma1, alpha2, sigma2, t_i, T, tau)
68      likelihood_proposed <- likelihood_function_with_changepoint(
        alpha1, sigma1, alpha2, sigma2_proposed, t_i, T, tau)
69
70      if (log(runif(1)) < (likelihood_proposed - likelihood_current))
        {
```

```
71
72        sigma2 <- sigma2_proposed
73      }
74
75      # Metropolis step for tau
76      tau_proposed <- tau_prior(T)
77      likelihood_current <- likelihood_function_with_changepoint(
      alpha1, sigma1, alpha2, sigma2, t_i, T, tau)
78      likelihood_proposed <- likelihood_function_with_changepoint(
      alpha1, sigma1, alpha2, sigma2, t_i, T, tau_proposed)
79
80      if (log(runif(1)) < (likelihood_proposed - likelihood_current))
      {
81        tau <- tau_proposed
82      }
83
84      sample_count <- sample_count + 1
85
86      # Store thinned samples after burn-in
87      if (i > burn_in && sample_count %% thinning == 0) {
88        stored_samples <- stored_samples + 1
89        samples[stored_samples, ] <- c(alpha1, sigma1, alpha2, sigma2,
      tau)
90      }
91    }
92
93    return(samples)
94  }
```

Code 4: R code for MCMC Sampling With a Change Point

The likelihood function takes into account the presence of a change point at time *tau*. It does so by separating the observed events into two groups: those that occurred before tau (the change point) and those that occurred after. The likelihood of these events is then calculated based on two sets of parameters: $alpha1$ and $sigma1$ for events before the change point, and $alpha2$ and $sigma2$ for events after the change point. This division allows the algorithm to appropriately account for the potential shift in rates at the change point.

In the MCMC simulation for each parameter within the model, a sequence of 200,000 values was generated. To ensure the removal of initial bias (burn-in) the initial 11,000 values of this sequence were discarded. For the priors of the model parameters, uniform distributions denoted as $U(a, b)$ and Gamma distributions denoted as $Gamma(c, d)$ were utilized, where the hyperparameters a, b, c, and d were predetermined constants.

4 Results

This study analyzes yearly rainfall data, measured in millimeters, spanning from 1879 to 2002 (covering a total of 124 years), and annual average maximum temperatures, measured in

degrees Celsius, gathered from 1915 to 2003 (across 88 years) from a meteorological station in Almaty, Kazakhstan. Additionally, it includes yearly highest temperature averages observed from 1894 to 2003 (a period of 110 years) in Tashkent, Uzbekistan, and Ukraine's average air temperature data (in Celsius) from 1891 to 2022 (132 years) collected in Kyiv.

In the original study, the authors defined the prior distributions for parameter estimation using non-informative priors for each no-change-point model's parameters. The chosen prior distributions are detailed below.

For the Power Law Process (PLP): $\alpha \sim U(0, 5)$ $\sigma \sim U(0, 10000)$

For the Musa–Okumoto process (MO) and Goel–Okumoto process (GO):

$\alpha \sim \text{Gamma}(0.01, 0.01)$ $\beta \sim \text{Gamma}(0.01, 0.01)$

For the generalized Goel–Okumoto (GGO): $\alpha \sim \text{Gamma}(0.01, 0.01)$

$\beta \sim \text{Gamma}(0.01, 0.01)$ $\gamma \sim U(0, 5)$

And for the Generalized Power Law Process (GPLP): $\alpha \sim \text{Gamma}(0.1, 0.1)$

$\sigma \sim \text{Gamma}(0.1, 0.1)$ $\beta \sim U(0, 100)$

In my project, to maintain consistency with the original findings and to enable a direct comparison, I adopted the same prior distributions for implementing the Metropolis-Hastings algorithm.

The subsequent sections provide a detailed examination of the findings previously reported in the original study alongside the results obtained through my implementation of the MCMC algorithm.

4.1 Kazakhstan Climate Data

In the initial analysis of Almaty climatic data, yearly precipitation averages and maximum temperature averages over two different periods were observed. The Figure 1 and Figure 2 reveal an upward trend in both precipitation and temperature data, notably post-1920 and post-1950, respectively. This uptrend surpasses the long-term average, suggesting potential change-points that indicate the shifts in climatic behavior.

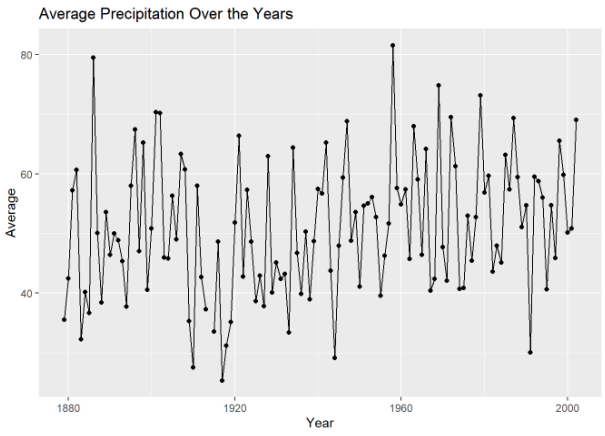

Figure 1: Yearly precipitation averages in Almaty.

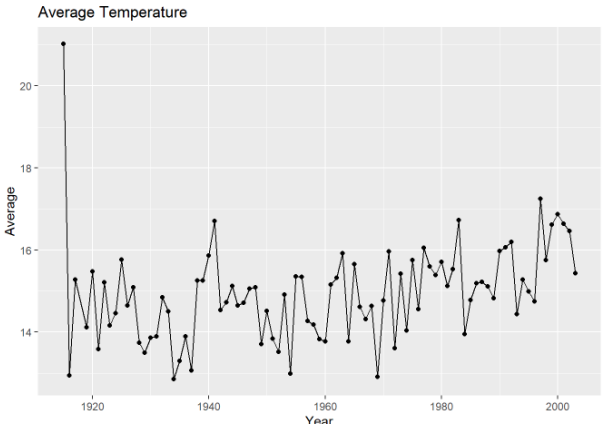

Figure 2: Yearly maximum temperature averages in Almaty.

4.1.1 Yearly Rainfall Averages

The precipitation data from Almaty averaged 50.88 mm over 124 years. Therefore, for this analysis, a threshold average at 51 mm is set up, which results in a total of 57 years within the observed period where precipitation exceeded this threshold.

The statistical summary provided in Table 1 includes the posterior means, standard deviations, and the 95% credible intervals obtained in the original study for the parameters of the models (estimated parameter distributions received using MCMC with Gibbs sampling) applied to the annual precipitation data in Kazakhstan, utilizing a threshold of 51.

Table 1: Posterior summaries (Kazakhstan precipitation averages data)

Model	Par.	Post. Mean	Std. Dev.	95% Cred. Int.	
				Lower	Upper
PLP	α	1.376	0.173	1.069	1.718
	σ	6.795	2.542	2.832	12.35
GPLP	α	1.153	0.202	0.784	1.565
	β	2.465	3.411	0.263	11.19
	σ	5.862	3.044	1.305	12.79
MOP	α	201.6	91.03	73.71	417.6
	β	115.7	44.20	53.66	224.2
GOP	α	239.9	101.8	109.4	495.2
	β	0.003	0.001	0.001	0.005
GGOP	α	168.2	81.94	73.06	378.2
	β	0.002	0.001	0.001	0.002
	γ	1.470	0.193	1.116	1.844

In Bayesian statistics, the mean of the posterior distribution is often considered the estimated parameter value because it is the expected value of the parameter given the data. This is rooted in the concept of minimizing the expected loss under the squared error loss function, which is commonly used in statistical estimation.

The plot on Figure 3 compares empirical precipitation data with the theoretical estimates from five different proposed rate functions (4), assuming that there are no change-points present in the data. Basically, on these plots the authors illustrate the accumulated count of precipitation exceedances which means they track the number of years when precipitation went above the usual average—indicating unusually wet years. The proposed rate functions is calculated based on the estimated mean parameter values shown in Table 1. One can visually assess that the best-fitting model that closely aligned with the observed cumulative exceedances in this study, was the PLP model.

Figure 4 in turn illustrates the accumulated average precipitation exceedances which is a result of my implementation of the Metropolis-Hastings algorithm, where the PLP model was employed using parameter distributions derived from my MCMC simulations. For a detailed exploration of the plotting code and an explanation of its components, please refer to the Appendix A section of this document.

The parameter statistics calculated from my implementation is detailed in Table 2, while histograms on Figure 5 show the distribution of the parameters alpha and sigma from the

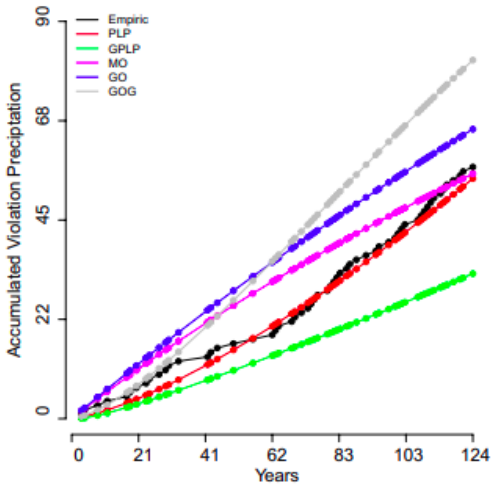

Figure 3: Accumulated precipitation exceedances (empirical and fitted m(t)) using the Kazakhstan precipitation data. The plot was derived from [4]

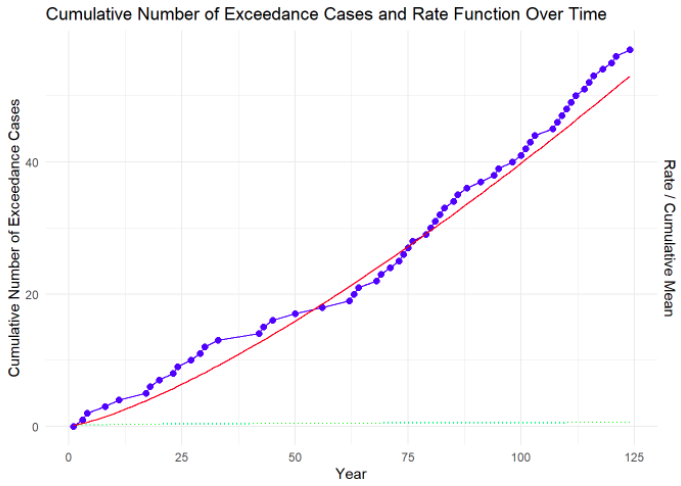

Figure 4: Accumulated precipitation exceedances (empirical and fitted m(t) with parameters estimated with Metropolis-Hastings algorithm) using the Kazakhstan precipitation data.

MCMC Metropolis-Hastings algorithm. When compared to the findings of the original study, my results exhibit a high degree of similarity, which demonstrates that my implementation reliably replicates the analysis performed in the original study.

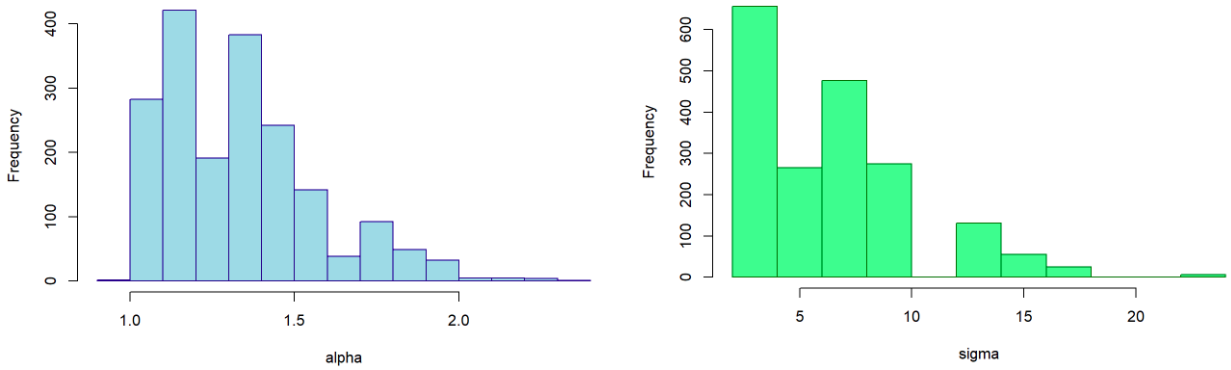

Figure 5: Distribution of alpha and sigma.

Table 2: Posterior summaries (Kazakhstan precipitation averages data)

Model	Par.	Post. Mean	Std. Dev.	95% Cred. Int.	
				Lower	Upper
PLP	α	1.33	0.231	1.05	1.8997
	σ	6.32	3.496	2.978	15.38

Estimates with the Presence of a Change-Point. In a scenario with a potential change-point, the authors incorporated only PLP model, which had previously been identified as the most accurate model without considering change-points. To inform the priors for this updated model, they utilized insights gained from the Bayesian estimates previously calculated under the PLP model without a change-point. This methodological step ensured that their priors were not chosen arbitrarily but were grounded in empirical evidence. They assumed the following prior distributions: $\alpha_j \sim \text{Gamma}(1.3, 1)$, $\sigma_j \sim \text{U}(0, 10)$, $\tau \sim \text{U}(1, 124), j = 1, 2$.

Table 3 presents the statistical summaries of the posterior distributions of the PLP model parameters incorporating a change point derived by utilizing Gibbs sampling to generate posterior distributions.

Table 3: Posterior summaries (Kazakhstan precipitation averages data)

Model	Par.	Post. Mean	Std. Dev.	95% Cred. Int.	
				Lower	Upper
PLP change-point	α_1	1.354	0.821	0.498	3.611
	α_2	1.307	0.208	0.799	1.604
	σ_1	5.082	2.217	1.293	9.396
	σ_2	6.012	2.653	0.593	9.861
	τ	42.63	36.01	1.174	120.0

Upon analysis, it can be observed that the model identified the change-point $\tau = 42.63$, that is $\tau = 43$, which corresponds to a year 1921, marking it as a potential moment when climatic behavior underwent some changes.

While running the Metropolis-Hastings algorithm to analyze the data, slight variations in parameter estimation were observed compared to the original study's outcomes. The identified change-point $\tau = 45.25$, which corresponds to a year 1923. The specific statistics for these parameters are documented in Table 4. Furthermore, to visually represent the distribution of parameters as derived from the MCMC Metropolis-Hastings algorithm, histograms are provided in Figures 6 and 7 offering a clear depiction of the parameter distribution.

Table 4: Posterior summaries (Kazakhstan precipitation averages data)

Model	Par.	Post. Mean	Std. Dev.	95% Cred. Int.	
				Lower	Upper
PLP change-point	α_1	1.247	0.507	0.482	2.525
	α_2	1.339	0.181	0.949	1.617
	σ_1	5.204	2.213	1.462	9.584
	σ_2	6.31	2.35	1.515	9.843
	τ	45.252	34.826	1.794	120.786

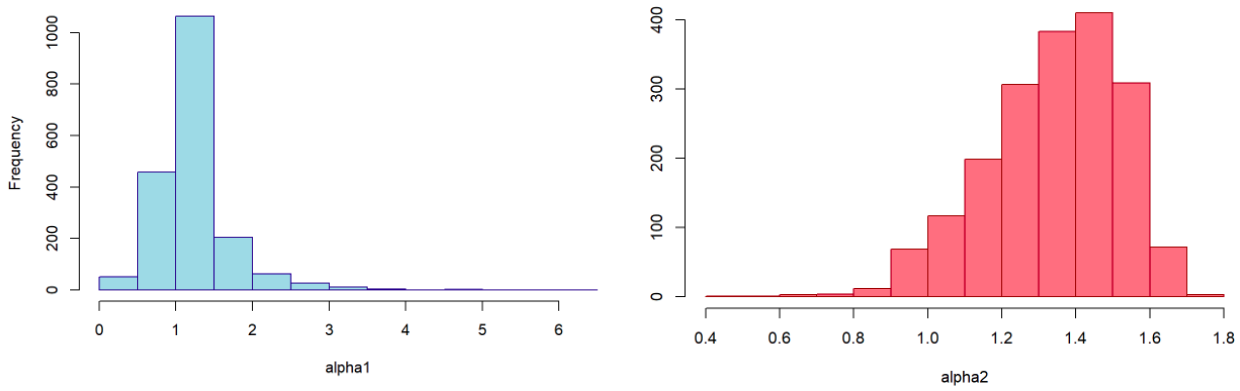

Figure 6: Distribution of α_1 and α_2.

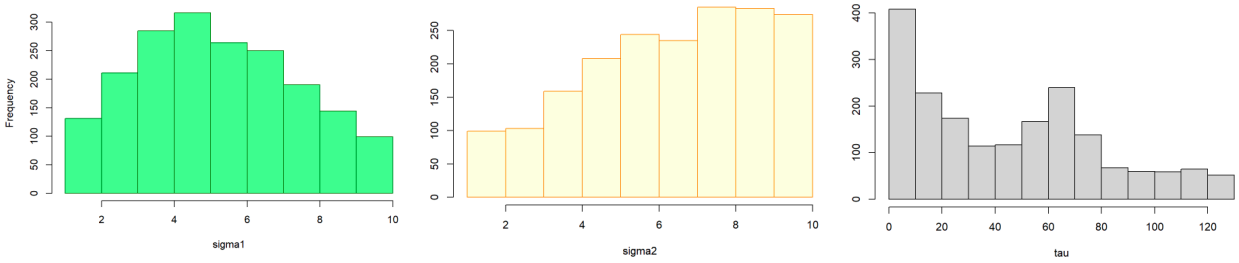

Figure 7: Distribution of σ_1, σ_2, and τ.

Figure 8 in turn showcases accumulated exceedances calculated using the estimates from the PLP model with the change-point, provided in the original study (left), alongside the estimates using my implementation of the Metropolis-Hastings algorithm (right). The right plot also emphasizes the identified change-point $\tau = 45$.

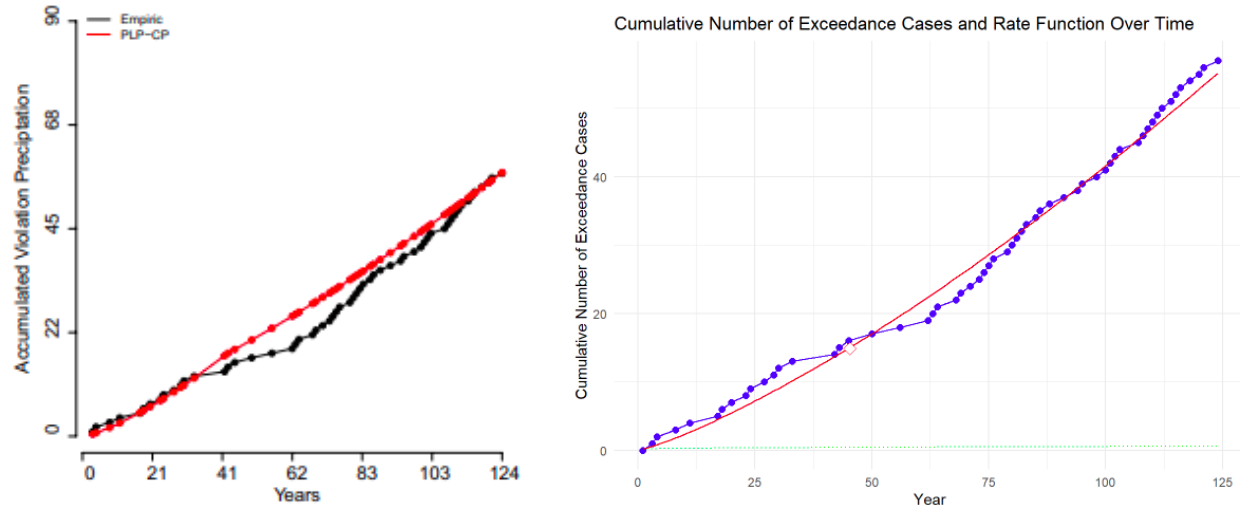

Figure 8: Accumulated precipitation exceedances (empirical and fitted m(t) with PLP model with a change point) using the Kazakhstan precipitation data.

The mean value function of the replicated model is:

$$m(t|\theta) = \begin{cases} (\frac{t}{5.204})^{1.247} & \text{if } t \leq 45, \\ (\frac{45}{5.204})^{1.247} + (\frac{t}{6.31})^{1.34} - (\frac{45}{6.31})^{1.34} & \text{if } t > 45. \end{cases}$$

$$\lambda(t|\theta) = \begin{cases} \frac{1.247}{5.204} \left(\frac{t}{5.204}\right)^{1.247-1} & \text{if } t \leq 45, \\ \frac{1.34}{6.31} \left(\frac{t}{6.31}\right)^{1.34-1} & \text{if } t > 45. \end{cases}$$

4.1.2 Yearly Maximum Temperature Averages

For the temperature dataset spanning 88 years, an overall average temperature of 14.939 was identified. Consequently, a threshold of 15 degrees Celsius was set for the Kazakhstan temperature data analysis. Within this time frame, exceedances were observed in 43 out of 88 years, indicating years where temperatures surpassed the set threshold. This analysis adopts the same NHPP model with specified intensity functions and utilizes identical prior distributions and MCMC simulation methodology as those applied to the Kazakhstan precipitation data.

As Figure 9 displays the accumulation of exceedance instances where maximum temperature averages surpassed the threshold of 15, alongside the mean value functions predicted by the five proposed models, in scenarios where no change-points were considered. From this visualization, the PLP model emerges as the most suitable fit for the data, mirroring findings from the precipitation data analysis. Additionally, the right side of Figure 9 showcases the mean function estimated using the PLP model derived from the implemented Metropolis-Hastings MCMC simulations.

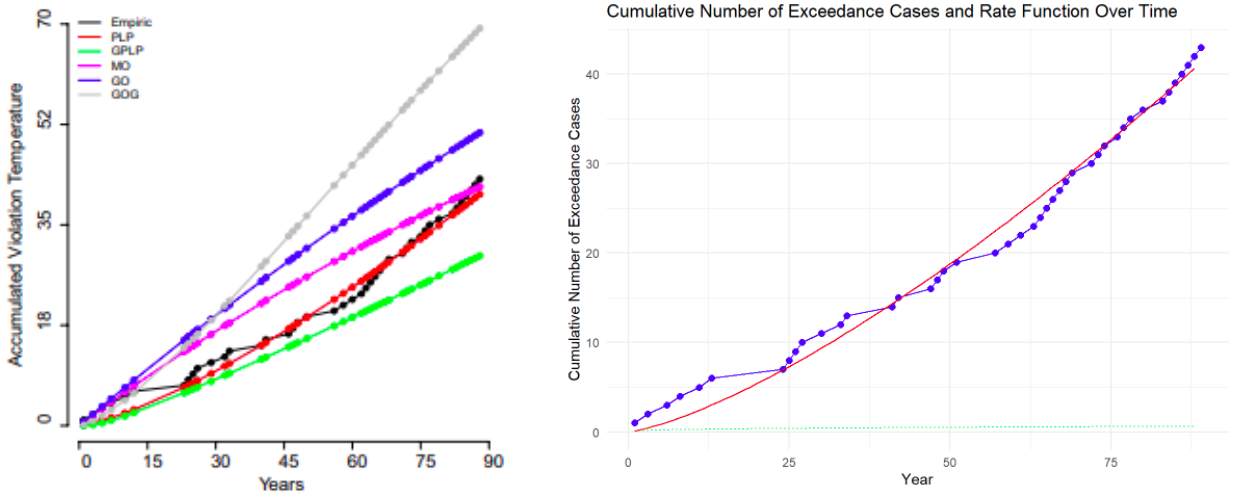

Figure 9: Accumulated temperature exceedances (empirical and fitted m(t)) using the Kazakhstan temperature data.

The posterior statistics for the replicated PLP model parameters are shown in Table 5. From the visual comparison of the plots, it can be observed that the estimated model closely aligns with the PLP model derived in the original study. The parameter estimates,

including the posterior means, standard deviations, and the 95% credible intervals for the replicated model are similar to those obtained in the original study, underscoring the accuracy and reliability of the Metropolis-Hastings MCMC simulations used in the replication. The detailed statistics for each model's parameters mentioned in the original paper are available for review in Appendix B. The rate function is given below:

$$\lambda(t|\theta) = \frac{1.375}{5.98}\left(\frac{t}{5.98}\right)^{1.375-1} \tag{15}$$

Table 5: Posterior summaries (Kazakhstan temperature data)

Model	Par.	Post. Mean	Std. Dev.	95% Cred. Int.	
				Lower	Upper
PLP	α	1.375	0.22	1.135	2.12
	σ	5.98	2.945	3.145	16.43

On Figure 10, we can observe the visual juxtaposition of the empirical mean value function $m(t)$ against the one fitted in the original study using the PLP model that incorporates a change-point. MCMC sampling was conducted using the following prior distributions: $\alpha_j \sim Gamma(1.3, 1)$, $\sigma_j \sim U(0, 10)$ and $\tau \sim U(1, 88)$; $j = 1, 2$. To the right of Figure 10, we can see the cumulative count of exceedances for the mean value function derived employing the Metropolis-Hastings MCMC simulations.

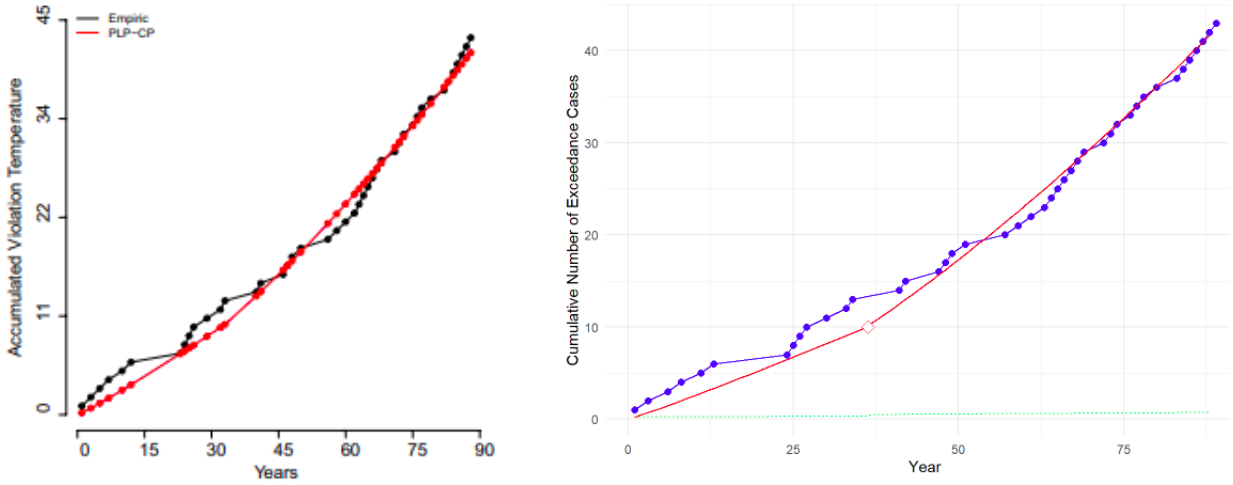

Figure 10: Accumulated precipitation exceedances (empirical and fitted m(t) with PLP model with a change point) using the Kazakhstan temperature data.

When analyzing the PLP model's behavior with a change-point using the Metropolis-Hastings algorithm for MCMC simulation, the posterior summaries can be found in Table 6 providing insights into parameter distributions after the change-point has been accounted for. In my study, the simulations identified the change-point at $\tau = 36$, aligning with the year 1951. This contrasts slightly with the original findings, which placed the change-point

earlier at year 1948. The statistical figures from both the original study are available for review in Appendix B. The mean and rate functions of the replicated model are:

$$m(t|\theta) = \begin{cases} (\frac{t}{4.28})^{1.077} & \text{if } t \leq 36, \\ (\frac{36}{4.28})^{1.077} + (\frac{t}{6.29})^{1.43} - (\frac{36}{6.29})^{1.43} & \text{if } t > 36. \end{cases}$$

$$\lambda(t|\theta) = \begin{cases} \frac{1.077}{4.28} \left(\frac{t}{4.28}\right)^{1.077-1} & \text{if } t \leq 36, \\ \frac{1.43}{6.29} \left(\frac{t}{6.29}\right)^{1.43-1} & \text{if } t > 36. \end{cases} \tag{16}$$

Table 6: Posterior summaries (Kazakhstan temperature data)

Model	Par.	Post. Mean	Std. Dev.	95% Cred. Int.	
				Lower	Upper
PLP change-point	α_1	1.077	0.336	0.486	1.794
	α_2	1.43	0.21	0.992	1.781
	σ_1	4.28	2.094	1.195	8.924
	σ_2	6.29	2.419	1.389	9.795
	τ	36.315	25.25	2.22	85.189

Additionally, Figure 11 gives a visual comparison of the rate function $\lambda(t_i|\theta)$ for the PLP models with (given in Equation 16) and without (Equation 15) the presence of a change point.

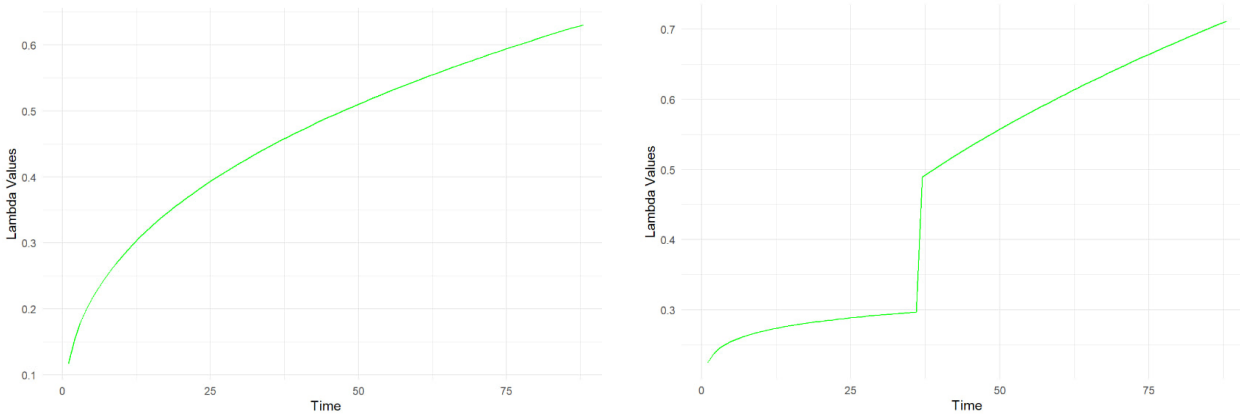

Figure 11: Rate function $\lambda(t_i|\theta)$ with (right) and without (left) the presence of a change point

4.2 Uzbekistan Maximum Temperature Data

Turning to the climate data from Tashkent, Uzbekistan, the maximum temperature averages recorded at a local meteorological station over a period from 1894 to 2003 were examined. The time series depicted in Figure 12 reflects this data set. Over this 110-year span, the overall mean temperature reached 20.7395°C. For this analysis, a threshold of 21°C was used

to determine exceedances, leading to 50 instances where the annual maximum temperature average surpassed this set point.

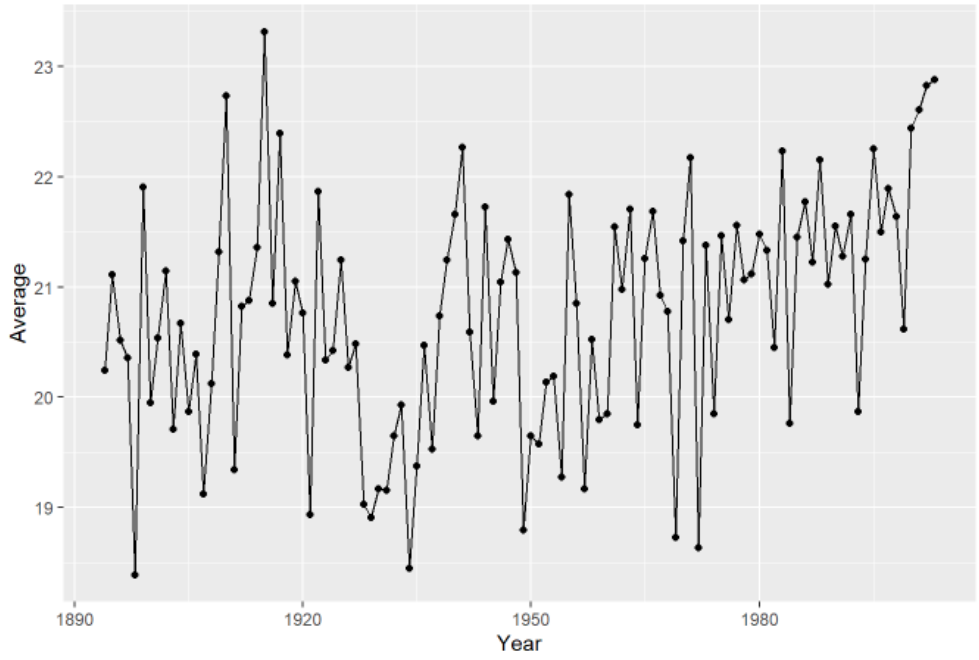

Figure 12: Yearly average of the maximum temperature averages in Tashkent, Uzbekistan

The results of the PLP model parameters estimated using the Metropolis-Hastings MCMC algorithm are documented in Table 7. In Figure 13, the left figure displays the mean value function estimations derived from the various rate functions proposed in the study (Equation 3). These visualizations demonstrate that the PLP model once again emerges as the most suitable fit. The right-side plot shows the mean function as determined by the PLP model, with the parameter estimates taken from the posterior mean values listed in Table 7. The mean and rate functions of the replicated model are:

$$m(t) = \left(\frac{t}{12.45}\right)^{1.76}$$

$$\lambda(t) = \frac{1.76}{12.45}\left(\frac{t}{12.45}\right)^{1.76-1}$$

Table 7: Posterior summaries (Uzbekistan temperature data)

Model	Par.	Post. Mean	Std. Dev.	95% Cred. Int.	
				Lower	Upper
PLP	α	1.762	0.183	1.417	2.129
	σ	12.456	2.847	7.165	17.908

In Figure 14, we observe comparative visualizations of the PLP models, both with the inclusion of a change-point. The representations include results from Gibbs sampling

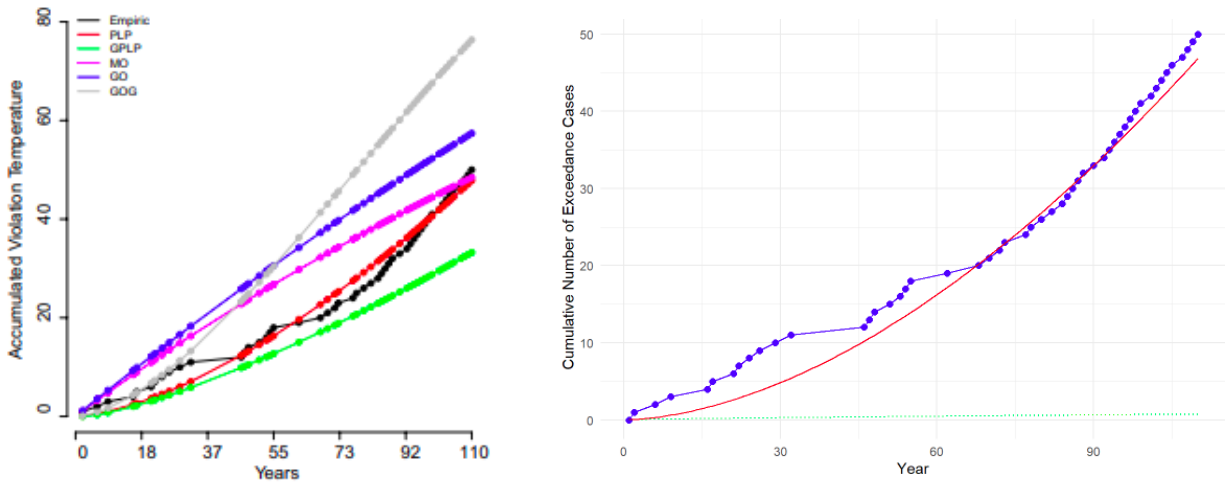

Figure 13: Accumulated temperature exceedances (empirical and fitted m(t)) using the Uzbekistan temperature data.

(left) and those obtained from my implementation using the Metropolis-Hastings algorithm (right), considering the following prior distributions: $\alpha \sim Gamma(1.5, 1)$, $\sigma_j \sim U(0, 30)$, and $\tau \sim U(1, 110)$, $j = 1, 2$.

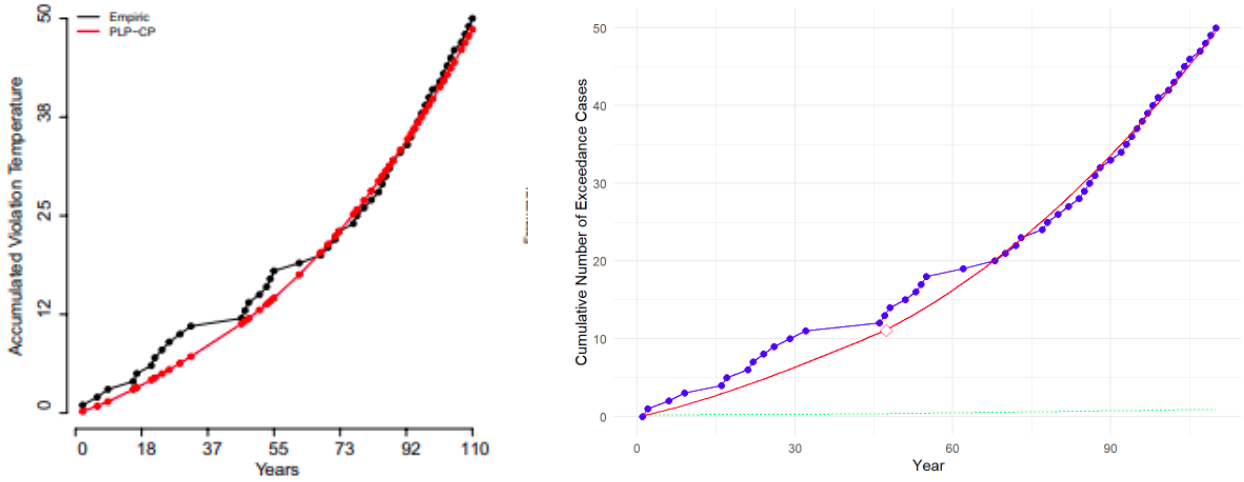

Figure 14: Accumulated precipitation exceedances (empirical and fitted m(t) with PLP model with a change point) using the Uzbekistan temperature data.

The posterior statistics for the PLP model parameters is shown in Table 8 (according to the MCMC simulations I conducted). The table suggests a Bayesian estimate for the change point (reflected by the posterior mean) of 47.37103 ($\tau = 47,$) translating to the year 1940. It is noteworthy that this estimation exhibits discrepancies from the original study's change-point estimation, which was identified four years earlier, in 1936 (Appendix B).

Table 8: Posterior summaries (Uzbekistan temperature averages data)

Model	Par.	Post. Mean	Std. Dev.	95% Cred. Int.	
				Lower	Upper
PLP change-point	α_1	1.25	0.429	0.564	2.144
	α_2	2.11	0.48	1.199	2.875
	σ_1	6.94	4.225	1.766	18.0
	σ_2	17.97	7.615	3.353	29.467
	τ	47.37	26.686	3.89	101.96

The mean and rate functions are given below:

$$m(t|\theta) = \begin{cases} (\frac{t}{6.94})^{1.25} & \text{if } t \leq 47, \\ (\frac{47}{6.94})^{1.25} + (\frac{t}{17.97})^{2.11} - (\frac{47}{17.97})^{2.11} & \text{if } t > 47. \end{cases}$$

$$\lambda(t|\theta) = \begin{cases} \frac{1.25}{6.94}\left(\frac{t}{6.94}\right)^{1.25-1} & \text{if } t \leq 47, \\ \frac{2.11}{17.97}\left(\frac{t}{17.97}\right)^{2.11-1} & \text{if } t > 47. \end{cases}$$

4.3 Ukraine Annual Average Temperature Data

The climate data from Ukraine contains the average air temperature records collected by the Boris Sreznevsky Central Geophysical Observatory [3] spanning a period from 1891 to 2022. The corresponding time series, illustrated in Figure 15, portrays this dataset. Throughout this 132-year interval, the overall mean temperature is 9 degrees Celsius, therefore a threshold of 9 was employed to identify exceeding instances, resulting in 39 occasions where the annual mean temperature surpassed this threshold. Notably, Ukraine data includes many recent years, up to 2022, unlike some earlier datasets. The temperature plot in Figure 15 makes it clear that the temperature rises have become much more noticeable since around 1995. There is a clear difference in the average temperatures when comparing the years before and after 1995, pointing to a significant shift in climate in recent times.

The parameter estimates for the PLP model, obtained through the Metropolis-Hastings MCMC method, are presented in Table 9. The resulting mean and rate functions of this modeled data are displayed below:

$$m(t) = \left(\frac{t}{22.377}\right)^{2.03}$$

$$\lambda(t) = \frac{2.03}{22.377}\left(\frac{t}{22.377}\right)^{2.03-1}$$

The posterior parameter statistics for the PLP model with a change point are detailed in Table 10. The posterior mean of τ is 103.8545 (thus, $\tau = 104$), which corresponds to the year 1994. The following prior distributions were used: $\alpha \sim Gamma(2,1)$, $\sigma_j \sim U(0,40)$, and $\tau \sim U(1,132)$, $j = 1,2$.

The visual representation of the distribution of model parameters derived from the MCMC Metropolis-Hastings algorithm is provided in Figures 16 and 17.

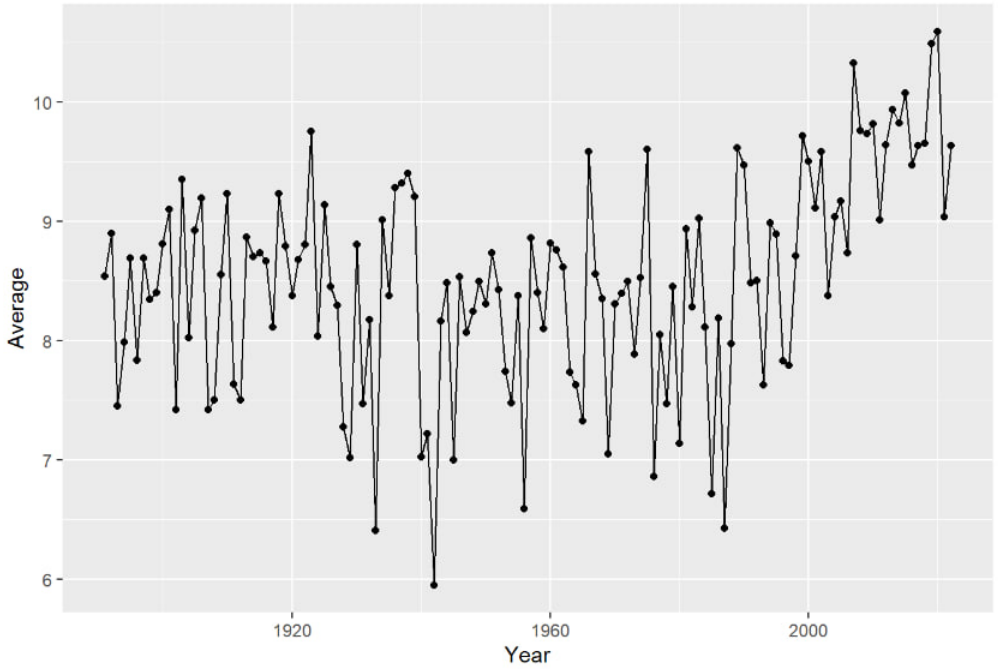

Figure 15: Yearly average temperature in Kyiv, Ukraine

Table 9: Posterior summaries (Ukraine temperature data)

Model	Par.	Post. Mean	Std. Dev.	95% Cred. Int.	
				Lower	Upper
PLP	α	2.03	0.289	1.54	2.671
	σ	22.377	5.797	13.067	36.067

Table 10: Posterior summaries (Ukraine temperature averages data)

Model	Par.	Post. Mean	Std. Dev.	95% Cred. Int.	
				Lower	Upper
PLP change-point	α_1	1.212	0.285	0.76	1.892
	α_2	2.296	0.565	1.151	3.178
	σ_1	10.84	5.302	2.634	22.603
	σ_2	23.44	11.219	2.385	39.391
	τ	103.85	13.22	51.855	115.837

The mean and rate functions for the model with a change point are given below:

$$m(t|\theta) = \begin{cases} (\frac{t}{10.84})^{1.212} & \text{if } t \leq 104, \\ (\frac{104}{10.84})^{1.212} + (\frac{t}{23.44})^{2.296} - (\frac{104}{23.44})^{2.296} & \text{if } t > 104. \end{cases}$$

$$\lambda(t|\theta) = \begin{cases} \frac{1.212}{10.84} \left(\frac{t}{10.84}\right)^{1.212-1} & \text{if } t \leq 104, \\ \frac{2.296}{23.44} \left(\frac{t}{23.44}\right)^{2.296-1} & \text{if } t > 104. \end{cases}$$

Figure 18 displays a comparison between observed temperature data (cumulative count of

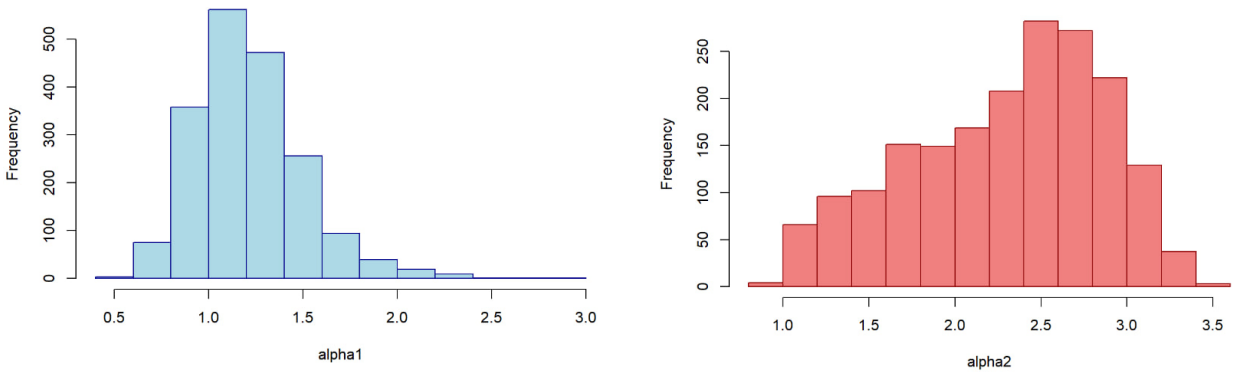

Figure 16: Distribution of α_1 and α_2.

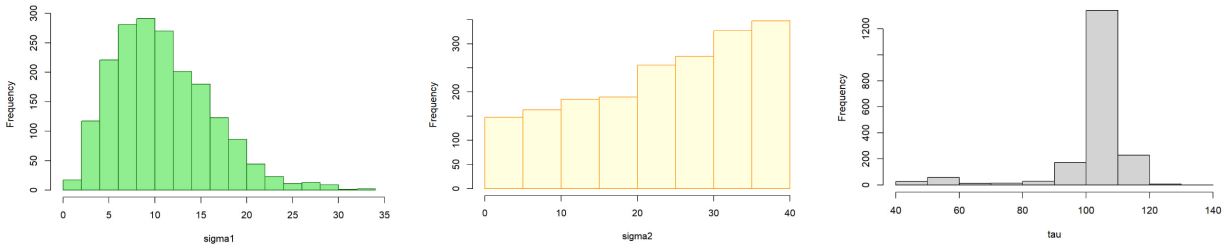

Figure 17: Distribution of σ_1, σ_2, and τ.

temperature exceedances) and estimated mean value function $m(t|\theta)$ under the assumption that the data does not exhibit any change points. Additionally, the figure presents estimates generated by my application of the Metropolis-Hastings algorithm for a model that incorporates a change point.

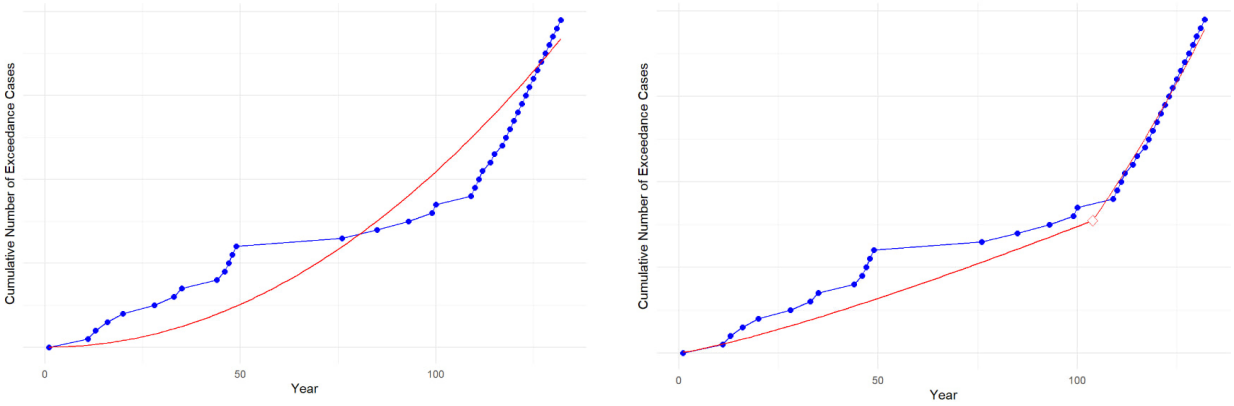

Figure 18: Accumulated average temperature exceedances (empirical and fitted $m(t)$) with PLP model with (right) and without (left) the presence of a change point

The plot in Figure 18 seems to illustrate that the estimated mean value function $m(t)$ does not align as closely with the observed temperature data compared to previous datasets. This less accurate fit could be attributed to a more pronounced increase in observed temperature values after the identified change-point. As a result, there's a smaller overall number of

exceedances, possibly because the threshold set for determining exceedances was too high relative to the actual temperature trends.

Additionally, the plot that incorporates a model with a change-point exhibits a noticeably better fit. This improvement indicates that accounting for the change-point — a point in time after which the rate of temperature increase has changed — allows the model to adapt to the shift in the distribution of temperatures. By doing so, the model more accurately captures the pattern of exceedances, reflecting the reality of the temperature shifts over the years, particularly after the change-point.

Figure 19 displays a plot that visually emphasizes the mean values of observed data before and after the year 1994, the year identified as the estimated change point. It can be observed that the post-1994 period exhibits a significant increase in the mean temperature compared to the pre-1994 period, this can indicate an accelerating warming trend, which the change point model is designed to detect and represent.

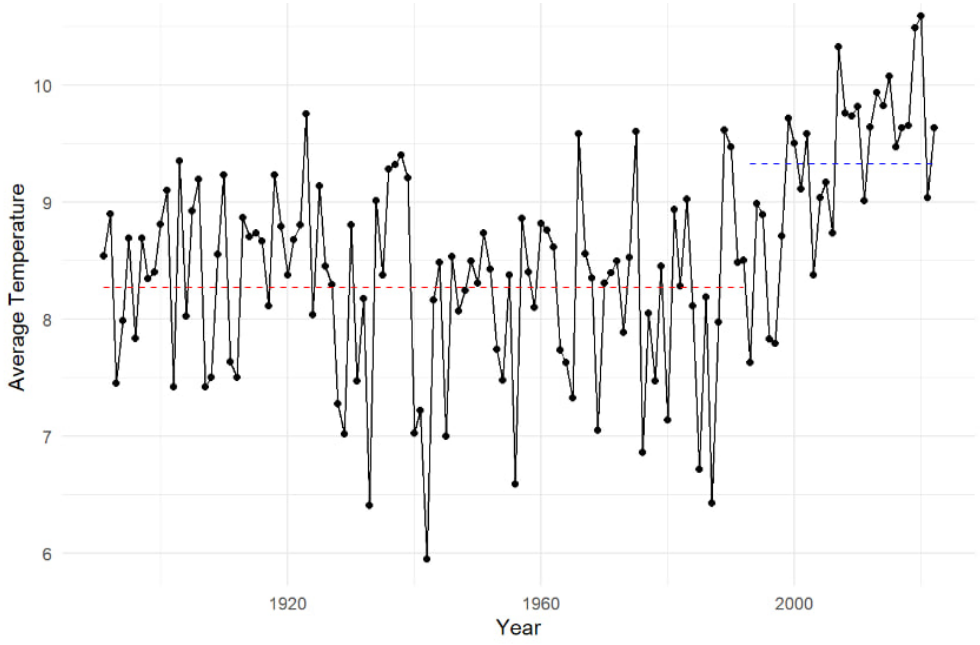

Figure 19: Yearly average temperature in Kyiv, Ukraine

5 Possible Extensions

Despite the successes of the NHPP methodology, this study did not extend to the detection of multiple change-points within the climate records. The rationale behind this limitation is grounded in the nature of climatic changes, which typically unfold over extended periods. Consequently, the likelihood of observing multiple significant shifts within the time span covered by this study is low. Future research, however, could explore this avenue, especially if it involves a wider dataset that spans a greater temporal scope.

Another potential extension of the current methodology is the application of Hawkes processes [7], which was proposed by the authors of the study as a promising area. This

process particularly suited for scenarios where events, such as climate exceedances, tend to occur in closely-knit bursts or clusters.

In addition to these methods, spatial Poisson processes could offer another layer of analysis by considering the geographical distribution of climate events. This approach could illustrate regional variations in climate shifts and could be particularly valuable in understanding localized effects in the context of global climate patterns.

6 Conclusions

This work builds upon the findings of a paper titled "Climate Change: Use of Non Homogeneous Poisson Processes for Climate Data in Presence of a Change-Point" [4], where a comprehensive analysis of climate data using non-homogeneous Poisson processes (NHPP) was conducted to understand and quantify the instances where climate measurements crossed certain thresholds representing the long-term averages for various climate indicators. These thresholds provide a baseline for detecting significant deviations that may signal changing climate patterns.

Central to this analysis were two different NHPP models: one that assumes a stable climate without any abrupt changes, and another that considers a particular change point. These models made it possible to test whether there were significant changes in the behavior of the climate during the specified time frames.

The study looked closely at annual averages of precipitation, as well as average and maximum temperatures in selected regions around the world over a long period of time. In the original study, intensity functions for NHPP were modeled using five time-dependent parametric forms driven by a set of unknown parameters: the power law process (PLP), the Musa–Okumoto process (MOP), the Goel–Okumoto process (GOP), a generalized form of Goel–Okumoto (GGOP); and the exponentiated-Weibull (GPLP), which generalizes the PLP model. Parameter estimation within these models was conducted under a Bayesian framework. Instead of the Gibbs sampling used in the original study.

Building on this foundation, the current project aims to replicate the paper's conclusions by focusing on the PLP model, which demonstrated a superior fit to empirical observations in all examined datasets. The parameter estimation for the PLP model was estimated through the Metropolis-Hastings algorithm as part of the Markov chain Monte Carlo (MCMC) methods.

The findings from the data analysis revealed notable shifts in climate patterns. For instance, in Kazakhstan, the data indicated a slight rise in average rainfall post-1923 and an increase in average maximum temperatures post-1950. Similarly, Uzbekistan's data suggested an uptick in average maximum temperatures after 1940. The replication of the original study's findings through my implementation, however, did present slight discrepancies. The change-points detected in my analysis varied by up to four years compared to those reported in the study. Despite these minor variations, the overall trend of increasing average temperatures and precipitation levels remained consistent with the original findings. The data from Ukraine further reinforces these patterns. With a marked change-point occurring in 1994, the Ukraine dataset underscores a significant shift in climate trends, particularly after the turn of the millennium.

This project sets the stage for future explorations that may extend to more complex models, including multiple change-points or even spatial Poisson processes, to capture a finer granularity of climate dynamics.

In conclusion, the study replicates the original results using an alternative MCMC method and validates the robustness of the findings and the versatility of Bayesian approaches in addressing complex environmental data sets.

A Appendix

Code 5 considers precipitation values collected and cleans the data, excluding anomalies like the placeholder value of -999.0, which likely represents missing or unrecorded data. It calculates the average precipitation for each year and identifies which years had average precipitation levels above the overall mean. This preprocessing approach was uniformly applied across datasets from Kazakhstan, Uzbekistan, and Ukraine.

```
excel_file <- "Precip_v1.xls"
data <- read_excel(excel_file)
almaty_data <- subset(data, data["Name"] == "Almaty")

# Replace -999.0 with NA
data_cleaned <- mutate(almaty_data, across(I:XII, ~ifelse(. ==
    -999.0, NA, .)))
sorted_data <- arrange(data_cleaned, Years)

sorted_data$Average <- rowMeans(sorted_data[, c("I", "II", "III", "
    IV", "V", "VI", "VII", "VIII", "IX", "X", "XI", "XII")], na.rm =
    TRUE)

df <- select(sorted_data, Years, Average)

ggplot(df, aes(x = Years, y = Average)) +
  geom_line() +
  geom_point() +
  labs(x = "Year", y = "Average") +
  ggtitle("Average Precipitation Over the Years")

# Calculate the mean of the average values
mean_average <- mean(df$Average, na.rm = TRUE)

# Identify years where the average exceeds the mean
years_above_mean <- df$Years[!is.na(df$Average) & df$Average > round
    (mean_average)]

start_year = min(df$Years)
period_numbers_above_mean <- years_above_mean - start_year + 1
```

Code 5: R code for data preprocessing and identification of years exceeding average precipitation levels

In Code 6, the Power Law Process (PLP) model was applied to the data using the Metropolis-Hastings algorithm within a Markov Chain Monte Carlo (MCMC) simulation framework. The parameter estimates obtained from these simulations, mean_alpha and mean_sigma, were used to calculate the rate function $\lambda(t)$ and the cumulative mean value function $m(t)$ for the PLP model over the entire observation period T, which spanned 124 years.

```r

samples <- mcmc_sampling(period_numbers_above_mean, 124)
alpha <- mean(samples[, 1])
sigma <- mean(samples[, 2])

t_values <- seq(1, T, by = 1)

# Calculate lambda(t) and m(t) for each t
lambda_values <- lambda(t_values, alpha, sigma)
m_t_values <- m_t(t_values, alpha, sigma)

sim_df <- data.frame(t = t_values, Lambda = lambda_values, M_t =
   cumsum(lambda_values))

df$YearInterval <- df$Years - min(df$Years) + 1

exceedances_df <- df %>%
  mutate(Exceedance = ifelse(is.na(Average), 0, ifelse(Average >
   threshold, 1, 0)),
         CumulativeCases = cumsum(Exceedance))  # Cumulative sum of
   exceedance cases

exceedances_df_filtered <- filter(exceedances_df, !is.na(
   CumulativeCases))

exceedances_df <- exceedances_df %>%
  mutate(Increment = c(1, diff(CumulativeCases))) %>%
  filter(Increment > 0)

# Plot for cumulative number of exceedance cases
ggplot() +
  geom_line(data = exceedances_df, aes(x = YearInterval, y =
   CumulativeCases), color = "blue") +
  geom_point(data = exceedances_df, aes(x = YearInterval, y =
   CumulativeCases), color = "blue", size = 2) +
  # Plot for rate function and cumulative mean value function
  geom_line(data = sim_df, aes(x = t, y = M_t), color = "red") +
  labs(x = "Year", y = "Cumulative Number of Exceedance Cases / Rate
    Function",
       title = "Cumulative Number of Exceedance Cases and Rate
   Function Over Time") +
  theme_minimal() +
  scale_y_continuous(name = "Cumulative Number of Exceedance Cases",
                     sec.axis = sec_axis(~., name = "Rate /
   Cumulative Mean", labels = NULL))
```

Code 6: R code for estimation cumulative mean functions

B Appendix

Table 11: Posterior summaries (Kazakhstan maximum temperature averages data)

Model	Par.	Post. Mean	Std. Dev.	95% Cred. Int.	
				Lower	Upper
PLP	α	1.342	0.193	1.001	1.771
	σ	5.597	2.236	2.022	10.66
GPLP	α	1.129	0.267	0.671	1.711
	β	2.686	6.493	0.139	14.40
	σ	4.251	2.927	0.413	11.18
MOP	α	151.3	82.39	46.13	369.9
	β	90.90	40.74	37.74	199.8
GOP	α	191.9	91.55	78.85	419.2
	β	0.004	0.002	0.001	0.008
GGOP	α	160.0	103.2	60.00	453.9
	β	0.002	0.001	0.001	0.004
	γ	1.383	0.199	1.001	1.767
PLP change-point	α_1	1.090	0.386	0.455	2.038
	α_2	1.407	0.217	0.927	1.740
	σ_1	3.884	2.121	0.697	9.031
	σ_2	6.240	2.542	0.997	9.904
	τ	32.56	24.70	1.895	83.74

Table 12: Posterior summaries (Uzbekistan maximum temperature averages data)

Model	Par.	Post. Mean	Std. Dev.	95% Cred. Int.	
				Lower	Upper
PLP	α	1.550	0.199	1.195	1.969
	σ	9.073	3.013	4.256	15.64
GPLP	α	1.354	0.268	0.903	2.002
	β	1.627	2.673	0.140	7.330
	σ	8.18	4.137	2.033	18.70
MOP	α	209.3	96.61	72.46	449.3
	β	114.7	45.80	51.62	230.3
GOP	α	247.1	113.3	107.6	555.2
	β	0.002	0.001	0.001	0.005
GGOP	α	146.0	66.91	66.72	315.4
	β	0.001	0.001	0.000	0.002
	γ	1.661	0.236	1.218	2.255
PLP change-point	α_1	1.264	0.412	0.571	2.289
	α_2	2.103	0.458	1.207	2.831
	σ_1	6.752	4.229	1.607	17.35
	σ_2	18.12	7.486	3.685	29.46
	τ	43.20	25.88	2.462	98.94

References

[1] National Centers for Environmental Information (NCEI), NOAA. https://www.ncei.noaa.gov/access/monitoring/climate-at-a-glance/.

[2] Nasa climate evidence. https://climate.nasa.gov/evidence/.

[3] Boris sreznevsky central geophysical observatory. http://cgo-sreznevskyi.kyiv.ua/uk/.

[4] J.A. Achcar and R.P. de Oliveira. Climate change: Use of non-homogeneous poisson processes for climate data in presence of a change-point. *Environmental Modeling Assessment*, 27:385–398, 2022. doi: 10.1007/s10666-021-09797-z. URL https://doi.org/10.1007/s10666-021-09797-z.

[5] S. Chib and E. Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[6] J. E. R. Cid and J. A. Achcar. Bayesian inference for nonhomogeneous poisson processes in software reliability models assuming nonmonotonic intensity functions. *Computational Statistics & Data Analysis*, 32(2):147–159, 1999.

[7] P. J. Laub, T. Taimre, and P. K. Pollett. Hawkes processes, 2015.

[8] V. Masson-Delmotte, P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfeld. *IPCC, 2018: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty.* Intergovernmental Panel on Climate Change, 2019.

[9] R. P. Oliveira, J. A. Achcar, J. Mazucheli, and W. Bertoli. A new class of bivariate lindley distributions based on stress and shock models and some of their reliability properties. *Reliability Engineering & System Safety*, 211:107528, 2021.

[10] A. F. Smith and G. O. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(1):3–23, 1993.

[11] D. J. Spiegelhalter, A. Thomas, N. Best, and D. Lunn. *WinBUGS version 1.4 user manual*. MRC Biostatistics Unit, Cambridge, 2003. URL: http://www.mrc-bsu.cam.ac.uk/bugs.

[12] M. Williams and V. Konovalov. Central asia temperature and precipitation data, 1879–2003. Boulder, Colorado: USA National Snow and Ice Data Center, 2008. URL https://doi.org/10.7265/N5NK3BZ8. [Accessed in 10/20/2020].