Elizabeth Binkina, Victoria Dilday, Haonan Wang, Sandra Garcia Lopez
STAT 431
Trevor Park
13 May 2025

Extreme Weather Trends (2000-2024)

Introduction:

We aim to utilize the dataset titled “Global Environmental Trends, 2000-2024” (Shahim) which provides an overview of important environmental indicators of climate change with data collected over the past 24 years in various countries. The dataset is sourced from OpenML, updated annually by Adil Shahim. It was last updated April 25, 2024. The data includes year, country, average annual temperature in Celsius, annual per capita carbon dioxide emissions, sea level rise in millimeters (mm), country population, and the number of extreme weather events. Our goal is to build a model that captures overall time trends in extreme weather event counts and country-specific differences to understand climate variability globally. We will use a Bayesian hierarchical model using JAGS in R to explore this dataset.

The increasing frequency and severity of extreme weather events continue to raise global concerns regarding climate variability and its societal consequences. Our study uses a 24 year panel dataset of global environmental features to model temporal and spatial patterns in extreme weather events. Using Bayesian hierarchical modeling, we aim to answer these central questions:

1. Is there a global time trend in extreme weather events?
2. How do country-specific patterns differ in their trend and baseline event rates?

We employ a Bayesian hierarchical Poisson regression model to assess both global and country-level changes in extreme weather event counts over time, accounting for country-specific variability.

```
 1st Qu.:10.00     1st Qu.: 576.0   1st Qu.:5.449e+07   1st Qu.:14.85
 Median :23.50     Median : 775.5   Median :1.231e+08   Median :19.80
 Mean   :24.33     Mean   : 937.4   Mean   :2.952e+08   Mean   :31.20
 3rd Qu.:36.00     3rd Qu.:1151.8   3rd Qu.:2.132e+08   3rd Qu.:43.05
 Max.   :59.00     Max.   :2726.0   Max.   :1.426e+09   Max.   :87.20
 Extreme_Weather_Events Forest_Area_pct
 Min.   : 2.00          Min.   : 0.50
 1st Qu.:12.75          1st Qu.:17.23
 Median :18.00          Median :32.00
 Mean   :20.19          Mean   :32.19
 3rd Qu.:25.00          3rd Qu.:48.12
 Max.   :64.00          Max.   :68.50
                         Year                       Country          Avg_Temperature_degC
                            0                             0                             0
CO2_Emissions_tons_per_capita            Sea_Level_Rise_mm                   Rainfall_mm
                            0                             0                             0
                   Population          Renewable_Energy_pct        Extreme_Weather_Events
                            0                             0                             0
              Forest_Area_pct
                            0
[1] 84
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00   12.00   17.00   19.13   24.00   64.00
Mean of Extreme_Weather_Events: 19.13158
Variance of Extreme_Weather_Events: 120.8055
```

Data Description:

Year: *Calendar year of observation*

Country: *Nation or region*

Extreme_Weather_Events: *Count of severe natural disasters*

Temperature: *Average annual surface temperature (°C)*

CO2_Emissions: *Per capita carbon dioxide emissions*

Methane, Nitrous_Oxide: *Greenhouse gases (ppm)*

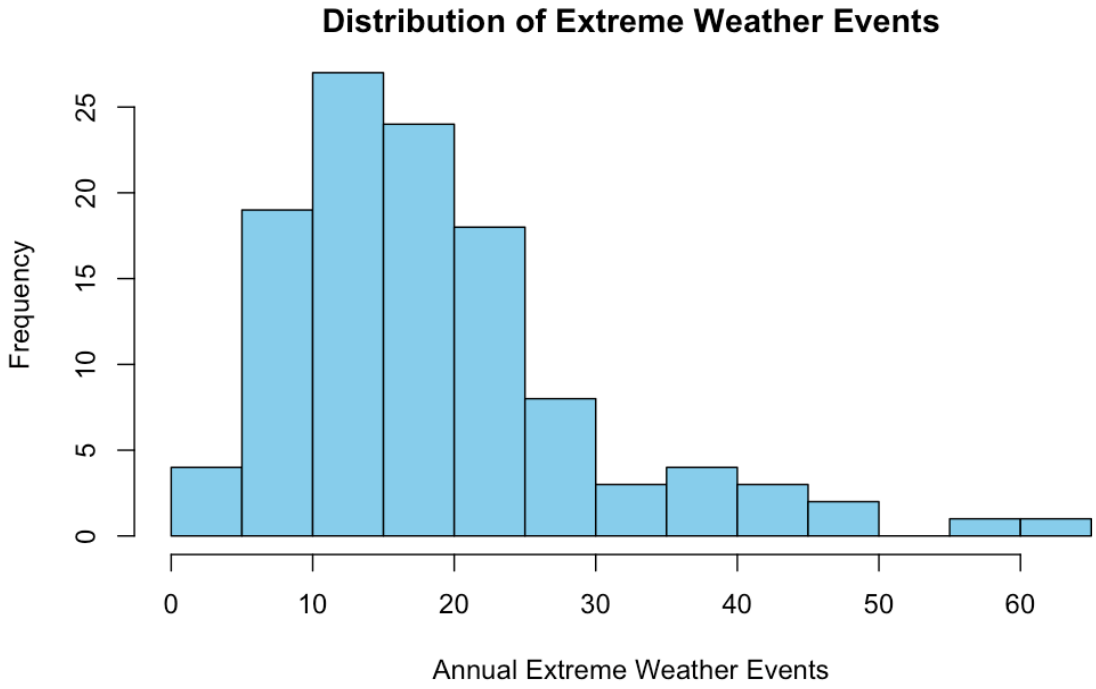
Sea_Level_Rise: *Annual change (mm)*

Population: *Total national population*

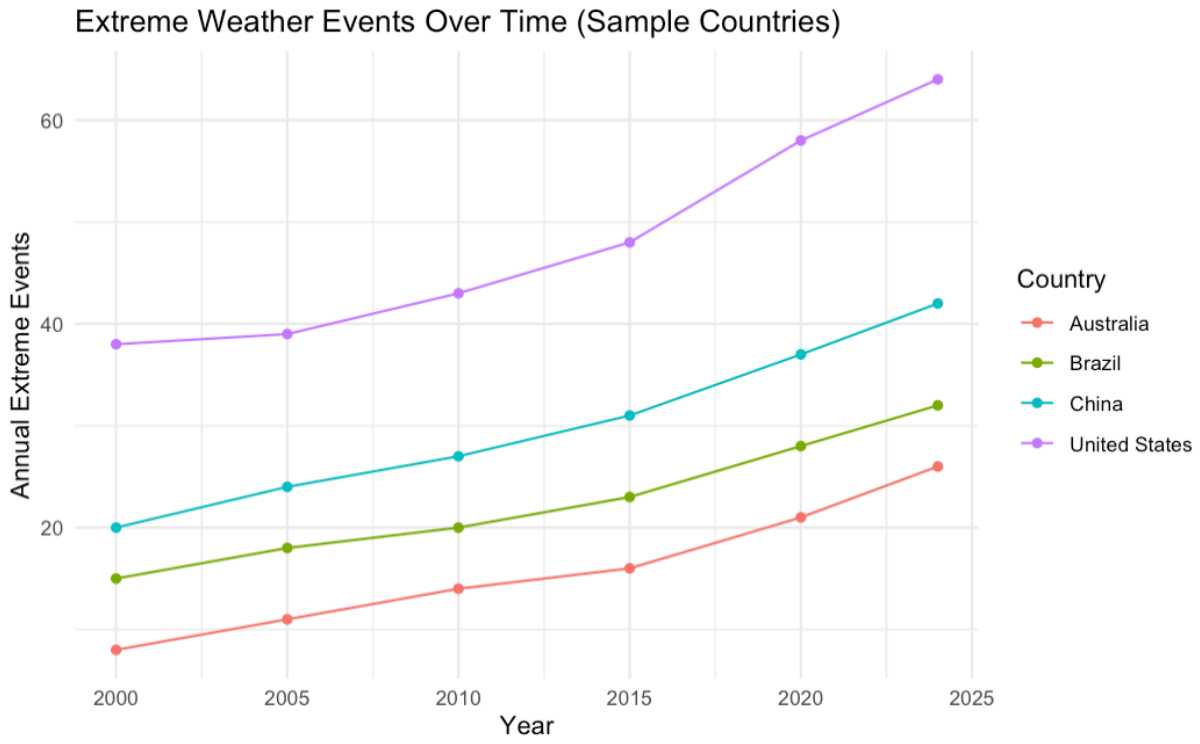
Data Preprocessing:

Duplicates at the (Country, Year) level were averaged, and missing values were handled via complete-case analysis after initial exploration (we discovered no missing data). A new feature logPop = log(Population) was created to serve as an offset in modeling even rates per capita, seen in Appendix A.

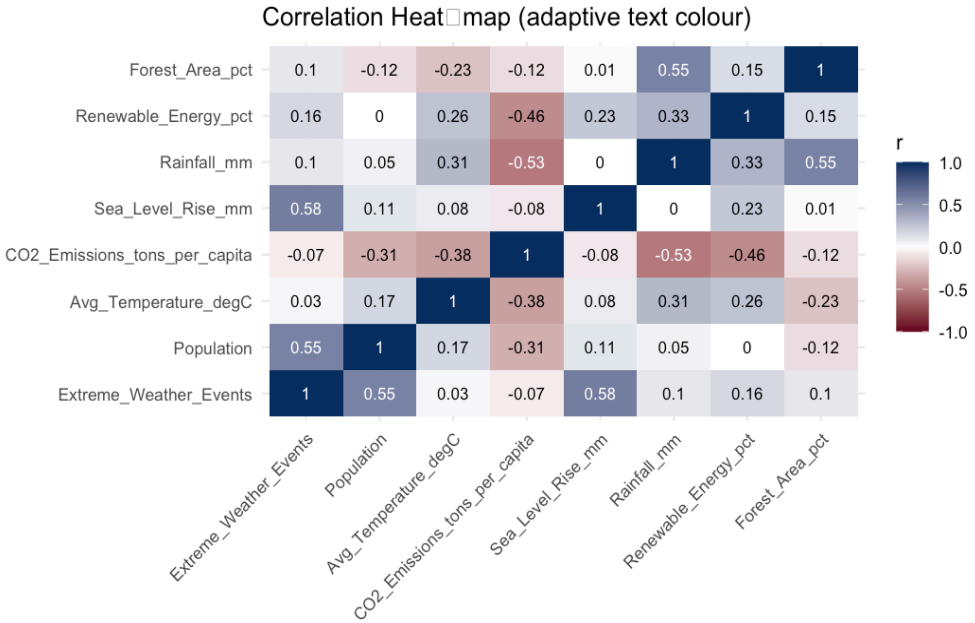
Exploratory Data Analysis:



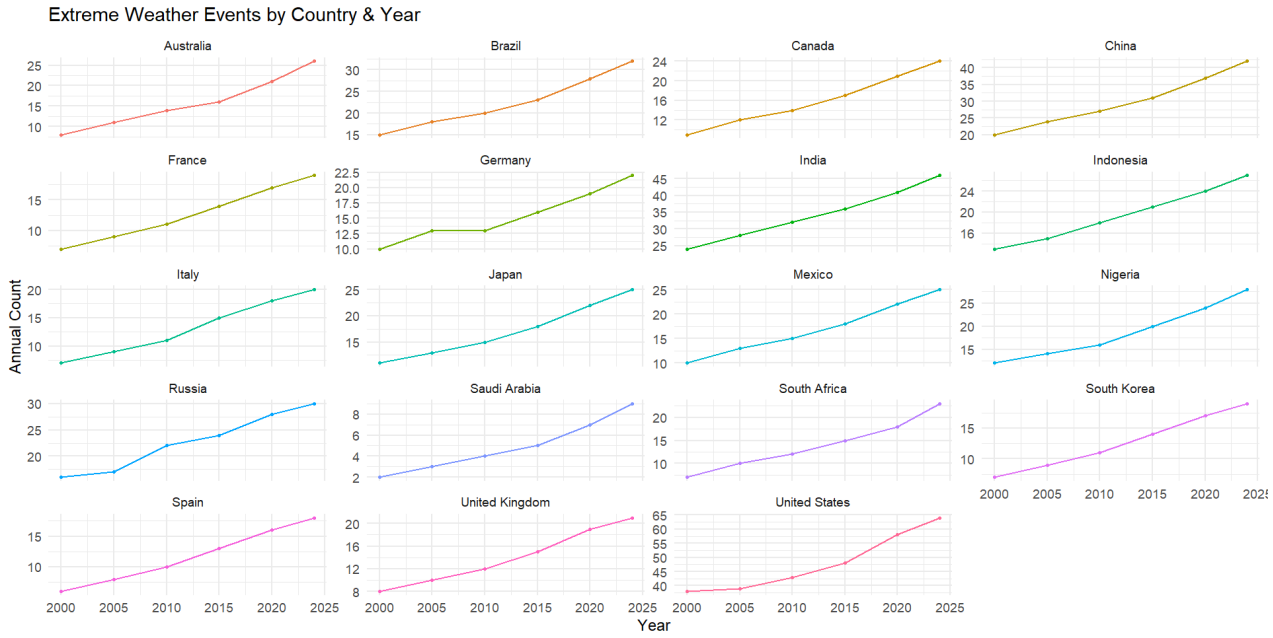
Extreme_Weather_Events are counts, with a right skewed distribution with a long tail. Mean (19.13) is significantly smaller than Variance (120.81), indicating overdispersion problems. These suggest we choose Poisson as the initial model but need posterior check to see if overdispersion is handled.



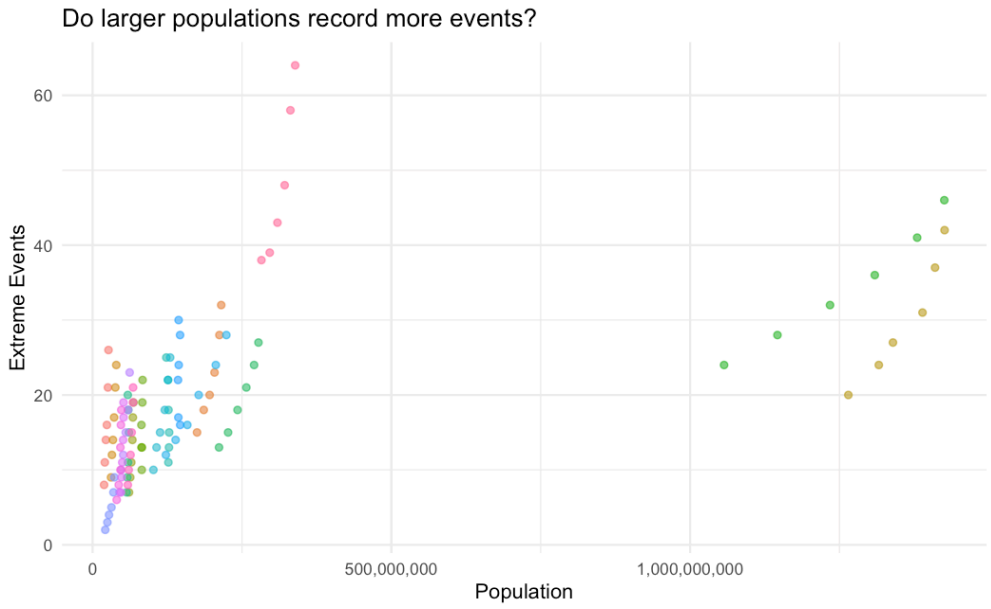
Temporal plots show clear increases in extreme weather frequency for some nations. The United States has the most increase in extreme weather events from 2000-2025, while Brazil, Australia, and China have roughly the same number of increased severe weather events. All 4 countries have different intercepts.



Correlation analysis shows moderate relationships between temperature, CO2, and extreme weather events.



The series of line plots shows trends in annual extreme weather events from 2000 to 2024 for 18 different countries. Many countries show an upward trend, indicating an increase in extreme weather events over the 24-year time period. The United States and China, have the steepest increases and highest counts. Germany, India, and Saudi Arabia also show substantial increases. Some countries like France and Japan show slower but consistent growth. This also indicates that larger countries could potentially have a more increasing rate of extreme weather events every year. Overall, countries have visibly different baselines but broadly parallel trends (similar but not identical) over time, suggesting we use a hierarchical model with random intercepts and slopes for countries, allowing for baseline and trend heterogeneity.



This scatterplot shows the relationship between population size (x-axis, log scale) and number of extreme weather events (y-axis). Each point represents a country-year observation. There is a positive association between countries with larger populations that tend to record more extreme weather events. However, there is variation that exists even among similarly populated countries. This suggests that population size is not the sole driver of event counts, but it's related to event counts which could impact our model. Thus, we added a population offset to ensure rate per capita.

Overall, our exploratory analysis revealed strong evidence of an upward trend in extreme weather events across the majority of countries in our dataset between 2000 and 2024. The time series plot confirms that this increase is consistent across diverse regions, suggesting a potential global pattern. The comparison between countries also shows different intercepts and variation in slopes, suggesting a grouping effect which leads to a hierarchical model with both intercept and slope random. Additionally, the correlation heatmap identified moderate relationships between extreme weather events and climate and global-related covariates such as temperature and greenhouse gas emissions, supporting their inclusion in future modeling. The scatterplot comparing event counts to population size suggests that larger populations tend to experience more extreme events, but with substantial variability, underscoring the need to normalize by population and include country-level effects. These findings motivated the use of a hierarchical modeling approach that can capture both global time trends and country-specific variation in event frequency.

Model Specification:

Likelihood

$$Y_{ij} \mid \mu_{ij} \;\sim\; \text{Poisson}\big(\mu_{ij}\big),$$

Log–mean with population offset

$$\log \mu_{ij} \;=\; \log\big(\text{Pop}_{ij}\big) \;+\; \alpha_i \;+\; \beta_i \,\big(\text{Year}_j - \overline{\text{Year}}\big).$$

Priors (country–level)

$$\alpha_i \sim \mathcal{N}\big(\mu_\alpha,\, \sigma_\alpha^2\big), \qquad \beta_i \sim \mathcal{N}\big(\mu_\beta,\, \sigma_\beta^2\big), \qquad i = 1, \ldots, J.$$

Hyper-priors

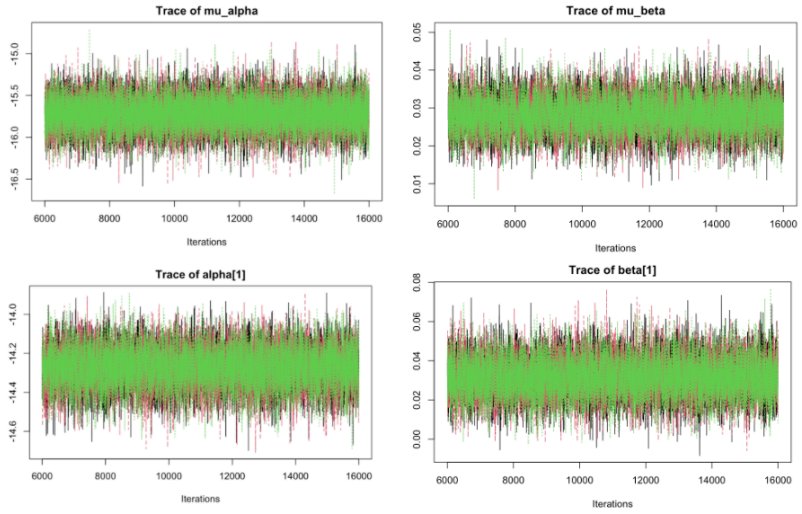
$$\mu_\alpha \sim \mathcal{N}(0,\, 10^6), \qquad \mu_\beta \sim \mathcal{N}(0,\, 10^6),$$

$$\tau_\alpha \;=\; \sigma_\alpha^{-2} \sim \text{Gamma}(0.001,\, 0.001), \qquad \tau_\beta \;=\; \sigma_\beta^{-2} \sim \text{Gamma}(0.001,\, 0.001).$$

Parameter definitions

- Y_{ij} – extreme-weather event **count** for country i in year j
- Pop_{ij} – population of country i in year j (offset)
- μ_{ij} – Poisson mean, $\mathbb{E}[Y_{ij}]$
- α_i – country-specific baseline log-rate (intercept)
- β_i – country-specific yearly trend (slope)
- $\mu_\alpha,\ \mu_\beta$ – global means of α_i and β_i
- $\sigma_\alpha^2,\ \sigma_\beta^2$ – between-country variances
- $\tau_\alpha,\ \tau_\beta$ – precisions $(1/\sigma^2)$ for the normal hierarchies
- J – number of countries in the analysis

JAGS and MCMC Settings:



JAGS string and MCMC settings found in Appendices B and C, respectively.

Initialization: 3 chains with dispersed random inits

Burn-in: 5,000 iterations

Sampling: 10,000 post burn-in iterations.

Diagnostics**:** Trace Plots and Gelman-Rubin statistics (R_hat < 1.1) confirm good mixing and convergence.

Monte Carlo standard error: All time-series SE are less than 1/20 of SD

Posterior Distribution Examination:

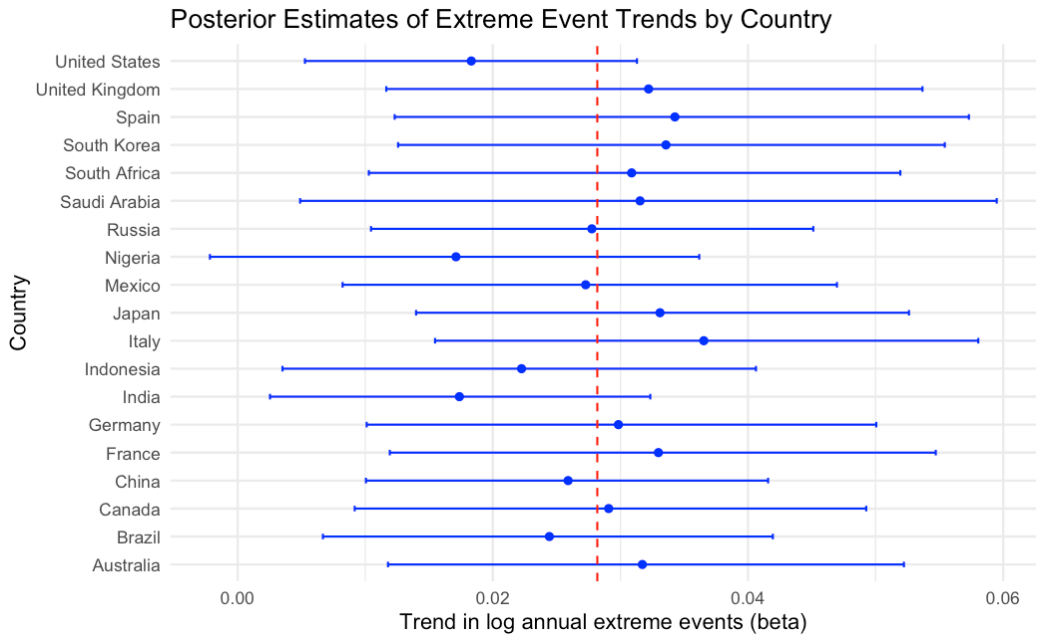
Global Trend Estimate

Posterior mean of global trend: $\mu_\beta = 0.028$

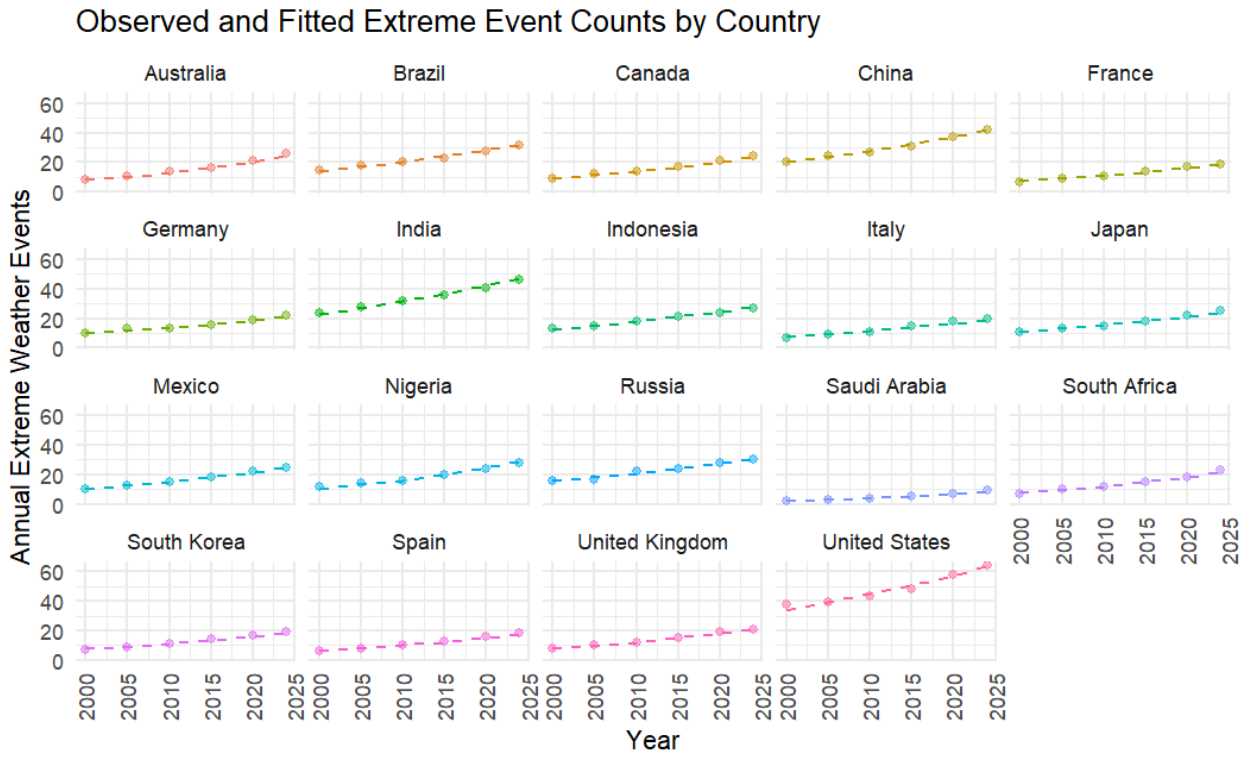
95% Credible Interval: [0.019, 0.038]

Through our data analysis, we found the global trend posterior mean to equal 0.02823 with a 95% credible interval of (0.01881, 0.03758). These indicate a positive relationship between extreme weather events and year. For each year, the globally expected number of extreme weather events has been increasing by 0.02823 on average for the past 24 years, or 2.9% since we are on a log scale. Our 95% credible interval of (0.01881, 0.03758) suggests that the true value of the global increase in extreme

weather events lies somewhere between those two numbers. The interval is relatively narrow and does not contain zero, suggesting statistically significant effects of extreme weather events and year.



The graph of posterior estimates of extreme events analyzes country-specific trends using log annual events, β. The blue dots on each horizontal line represent the country's posterior mean, while the horizontal lines themselves represent the country's 95% credible interval. The red dashed vertical line is the global posterior mean of 0.02823. Looking at the graph, we can see that most countries' posterior means are greater than the global average. The three countries that bring down the average are the United States, Nigeria, and India. The United States, in fact, has the lowest posterior mean and narrowest credible interval. Nigeria has the lowest centered credible interval, and it contains zero indicating a potential deviance from the overall pattern; extreme weather events might not be increasing over time in Nigeria.



Visualizing Fitted Trends Over Time:

The observed and fitted extreme event counts by country uses their posterior means with alpha hat and beta hat, and fits a Poisson rate of lambda hat using these values; calculation can be seen in Appendix D. What we are looking at on the graphs are each countries' lambda hats, or rates of increase. The United States starts the highest and ends the highest, which could explain why the United States' posterior mean was the lowest – there wasn't much of an increase yearly since it was already so high to begin with. India and China are also high near the end of the dataset, both of which had posterior means less than the global posterior mean.

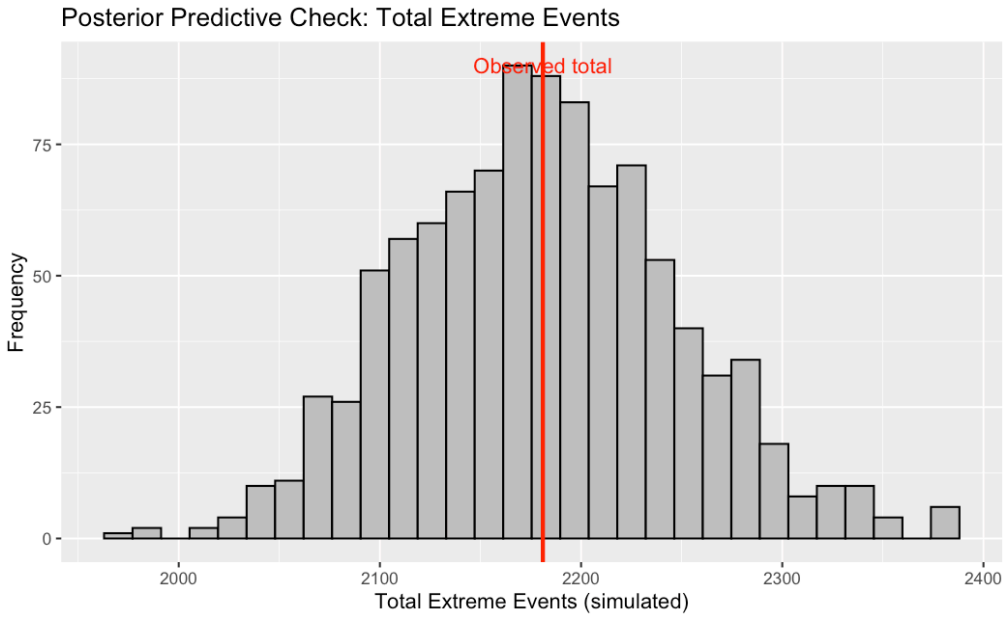
Summary:

Country-level slopes vary, but most show credible intervals overlapping the global trend, suggesting broadly parallel trends with some heterogeneity in magnitude but not direction. Countries also differ substantially in baseline rates, which might be reflecting geographic or socio-political differences in event reporting or exposure.

Model Checks & Diagnostics

Country-wise x^2 PPC p_B: 0.491

Bayesian p-value for total events**:** 0.489



A Bayesian p-value of 0.491 (Appendix G) indicates a decent fit for the model we have used to capture country-specific event counts and 0.489 (Appendix E) indicates a decent fit for overall trends when comparing the observed data to data simulated from the posterior predictive distribution; the simulation can be seen in Appendix E. The histogram plots the frequency of events on the y-axis and the simulated number of events on the x-axis. The red vertical line shows the observed total in the dataset. This graph further indicates a good fit since the observed total lies in the middle of the simulation number where frequency is high.

Sensitivity Check:

	Model	mu_beta	mu_beta_lo	mu_beta_hi	sigma_beta	sigma_beta_lo	sigma_beta_hi	elpd_loo	elpd_se
	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2.5%...1	A	0.02824447	0.01909592	0.03772103	0.01622474	0.01103880	0.02410621	-285.6104	3.521516
2.5%...2	B	0.02910803	0.01070230	0.04729611	0.03804294	0.02734261	0.05403852	-288.7244	3.540350
2.5%...3	C	0.02946962	-0.01751979	0.07524437	0.09996116	0.07476333	0.13714810	-290.5339	3.533575
2.5%...4	D	0.02823668	0.01862457	0.03789834	0.01625862	0.01102716	0.02414165	-285.7272	3.537151

We re-ran the model with four different gamma hyper-priors for the country-level standard deviations. Posterior means range from 0.0282 to 0.0295; all 95 % credible intervals overlap heavily. So a roughly 2.8-2.9% increase is unchanged. Versions A, B and D report similar small spreads (0.016–0.038). Version C, which forces much tighter priors, inflates the estimate to about 0.10; this is an expected prior-driven effect, not a shift in the overall trend. So alterations to hyperpriors barely affect the global trend or predictive quality, which support the robustness of our model. Given our model's posterior predictive check, convergence diagnostics, and sensitivity all show good fit, we believe our model is well-behaved and robust. Codes can be found in appendix F.

Conclusion:

Our final results using a Bayesian hierarchical model with JAGS in R with the extreme weather dataset indicate a global increase in extreme weather events over time. Examining the dataset with country and year chosen as variables, we can see that there is an upward trend in the number of extreme weather events. Looking at the posterior mean, we can see that this increase will most likely not stop, with countries such as the United States, India, and China increasing in extreme weather events. They are all highly populated regions, which coincides with our finding that larger populations experience more extreme weather events.

Countries with larger populations, specially the United States, India, and China, are also some of the largest countries by landmass. The larger the country, the more likely it is to experience extreme weather events as it spans a larger range of climates and a larger area being monitored and recorded for extreme weather events. Another potential explanation for this correlation is that larger populations tend to alter ecosystems more, resulting in increased severe weather. For example, urbanization often comes with more concrete and less vegetation, which severely impacts water infiltration and can contribute to more extreme flooding as water pools have nowhere to go.

Further exploration of this dataset or similar could include exploring the effect of country size on extreme weather events. As discussed earlier, temperature and greenhouse gas emissions have high correlations with extreme weather events, so adding them could result in a more informative model.

Contributions:

Elizabeth Binkina - Wrote the introduction, description of the data, did the exploratory analysis, and model specification analysis, and parts of the report for those sections

Victoria Dilday - wrote posterior distribution examination, model checks & diagnosis, conclusion

Haonan Wang - wrote R and JAGS code for duplicate handling, model specification and MCMC, and made some of the graphs.

Sandra Garcia Lopez - wrote multiple R and JAGS code for alternative data and models that ended up not being used due to being trivial or having errors, editing the report and added to the conclusion

Statement on AI: AI was used to check for grammar and combine repetitive sentences.

Work Cited:

Shamim, Adil. (2025, April 24). Global Environmental Trends 2000-2024 [Dataset]. Kaggle. https://www.kaggle.com/datasets/adilshamim8/temperature?resource=download.

Appendix

Appendix A

```
# ----------2. duplicate-handling ----------
dedup <- climate_data %>%
  group_by(Country, Year) %>%
  summarise(
    # average the reported counts, keep whole number
    Extreme_Weather_Events = round(mean(Extreme_Weather_Events, na.rm = TRUE)),

    # average population
    Population             = mean(Population, na.rm = TRUE),

    # average covariates
    across(where(is.numeric) &
           !matches("Extreme_Weather_Events|Population"),
           ~ mean(.x, na.rm = TRUE)),
    .groups = "drop"
  )

climate_data <- dedup
```

```
  {r model setup}

# ---------- 1. Prepare data for JAGS ----------
# sorted by country and year
climate_data <- climate_data %>% arrange(Country, Year)

# numeric indices for countries and years
country_index <- as.integer(as.factor(climate_data$Country))
year_values   <- climate_data$Year
year_centered <- year_values - mean(year_values)
climate_data <- climate_data %>% mutate(year_ctr = year_centered)

N_country <- length(unique(country_index))
N_obs     <- nrow(climate_data)

climate_data <- climate_data %>% mutate(logPop = log(Population))


# JAGS data
jags_data <- list(
  y          = climate_data$Extreme_Weather_Events,
  logPop     = climate_data$logPop,
  country    = country_index,
  year       = year_centered,
  N_country  = N_country,
  N_obs      = N_obs
)
```

Appendix B

```
# ---------- 2. Define model in JAGS ----------
jags_model_string <- "
model {
  for(i in 1:N_obs) {
    y[i] ~ dpois(lambda[i])
    log(lambda[i]) <- logPop[i] + alpha[country[i]] + beta[country[i]] * year[i]
    loglik[i] <- logdensity.pois(y[i], lambda[i])
  }

  # priors
  for(j in 1:N_country) {
    alpha[j] ~ dnorm(mu_alpha, tau_alpha)    # intercept
    beta[j]  ~ dnorm(mu_beta, tau_beta)      # slope
  }

  # hyper-priors
  mu_alpha ~ dnorm(0.0, 1e-6)
  mu_beta  ~ dnorm(0.0, 1e-6)

  tau_alpha ~ dgamma(0.001, 0.001)
  tau_beta  ~ dgamma(0.001, 0.001)

  # sd
  sigma_alpha <- 1 / sqrt(tau_alpha)
  sigma_beta  <- 1 / sqrt(tau_beta)
}
"
```

Appendix C

```
# ---------- 3. MCMC settings ----------
generate_inits <- function() {
  list(
    alpha = rnorm(N_country, 0, 5),
    beta  = rnorm(N_country, 0, 5),
    mu_alpha = rnorm(1, 0, 5),
    mu_beta  = rnorm(1, 0, 5),
    tau_alpha = rgamma(1, 0.1, 0.1),
    tau_beta  = rgamma(1, 0.1, 0.1)
  )
}

# initial values for each chain
n_chains <- 3
init_list <- list()
for(ch in 1:n_chains) {
  init_list[[ch]] <- generate_inits()
}


# ---------- 4. Gibbs Sampling and Convergence check ----------
jags_model <- jags.model(textConnection(jags_model_string), data=jags_data, inits=init_list,
                         n.chains=n_chains, n.adapt=1000)

# Burn-in
update(jags_model, n.iter=5000)

# Parameters to monitor
params_to_monitor <- c("mu_alpha", "mu_beta", "alpha", "beta", "sigma_alpha", "sigma_beta", "loglik")

# Draw samples
n_iter <- 10000
thin_val <- 1
post_samples <- coda.samples(jags_model, variable.names=params_to_monitor,
                             n.iter=n_iter, thin=thin_val)

# Check convergence
summary(post_samples)
gelman.diag(post_samples, multivariate=FALSE)   # Gelman-Rubin R-hat


# ---------- 5. visualization ----------
# trace plots
traceplot(post_samples[,c("mu_alpha","mu_beta")])

# try a specific country (1)
traceplot(post_samples[,c("alpha[1]","beta[1]")])
```

Appendix D

```r
# posterior mean predictions for each country-year
alpha_post_mean <- apply(post_mat[,grep("^alpha\\[", colnames(post_mat))], 2, mean)
beta_post_mean  <- apply(post_mat[,grep("^beta\\[",  colnames(post_mat))], 2, mean)

climate_data <- climate_data %>%
  mutate(alpha_hat = alpha_post_mean[country_index],
         beta_hat  = beta_post_mean[country_index],
         lambda_hat = exp(logPop + alpha_hat + beta_hat * climate_data$year_ctr))

# Plot observed vs fitted for each country over time
climate_data %>%
  ggplot(aes(x=Year)) +
  geom_line(aes(y=lambda_hat, color=Country), linetype="dashed") +
  geom_point(aes(y=Extreme_Weather_Events, color=Country), alpha=0.5) +
  facet_wrap(~ Country) +
  labs(title="Observed and Fitted Extreme Event Counts by Country",
       y="Annual Extreme Weather Events") +
  theme_minimal() + theme(legend.position="none")
```

Appendix E

```r
# Posterior predictive simulation
set.seed(123)
n_sim <- 1000

obs_discrepancy <- sum(climate_data$Extreme_Weather_Events)
sim_discrepancies <- numeric(n_sim)
posterior_samples_matrix <- as.matrix(post_samples)

for (s in 1:n_sim) {
  idx      <- sample(seq_len(nrow(post_mat)), 1)
  alpha_s  <- post_mat[idx, grep("^alpha\\[", colnames(post_mat))]
  beta_s   <- post_mat[idx, grep("^beta\\[",  colnames(post_mat))]

  lambda_s <- exp(climate_data$logPop +
             alpha_s[country_index] +
             beta_s[country_index] * climate_data$year_ctr)

  y_sim <- rpois(N_obs, lambda_s)
  sim_discrepancies[s] <- sum(y_sim)
}

# Bayesian p-value
p_value <- mean(sim_discrepancies >= obs_discrepancy)
cat("Bayesian p-value for total events:", p_value, "\n")
```

Appendix F

```
# four models to compare
model_A <- jags_model_string                                          # original
model_B <- replace_line(model_A,
            "tau_alpha ~ dgamma\\(0.001, 0.001\\)",
            "tau_alpha ~ dgamma(0.01, 0.01)")
model_B <- replace_line(model_B,
            "tau_beta  ~ dgamma\\(0.001, 0.001\\)",
            "tau_beta  ~ dgamma(0.01, 0.01)")

model_C <- replace_line(model_A,
            "tau_alpha ~ dgamma\\(0.001, 0.001\\)",
            "tau_alpha ~ dgamma(2, 0.1)")
model_C <- replace_line(model_C,
            "tau_beta  ~ dgamma\\(0.001, 0.001\\)",
            "tau_beta  ~ dgamma(2, 0.1)")

model_D <- replace_line(model_A,
            "mu_alpha ~ dnorm\\(0\\.0, 1e-6\\)",
            "mu_alpha ~ dt(0, pow(10,-2), 3)")   # t with df=3, scale=10
model_D <- replace_line(model_D,
            "mu_beta  ~ dnorm\\(0\\.0, 1e-6\\)",
            "mu_beta  ~ dt(0, pow(10,-2), 3)")

models <- list(A = model_A, B = model_B, C = model_C, D = model_D)

# fits one model & returns a named list
fit_one <- function(modtxt, id){
  jm <- jags.model(textConnection(modtxt), data = jags_data,
                   n.chains = 3, n.adapt = 1000,
                   inits = function(){
                     list(alpha=rnorm(N_country,0,5), beta=rnorm(N_country,0,5),
                          mu_alpha=rnorm(1,0,5), mu_beta=rnorm(1,0,5),
                          tau_alpha=rgamma(1,0.1,0.1), tau_beta=rgamma(1,0.1,0.1))
                   })
  update(jm, 3000)
  samp <- coda.samples(jm,
                       c("mu_alpha","mu_beta","sigma_alpha","sigma_beta","loglik"),
                       n.iter = 7000, thin = 1)
  ll_draws <- as.matrix(samp)[, grep("^loglik\\[", colnames(as.matrix(samp)))]
  loo_out  <- loo::loo(ll_draws)
  list(samples = samp,
       loo_elpd = loo_out$estimates["elpd_loo","Estimate"],
       loo_se   = loo_out$estimates["elpd_loo","SE"])
}

sens <- lapply(names(models), \(k) fit_one(models[[k]], k))
names(sens) <- names(models)

sumtab <- purrr::map_dfr(names(sens), function(k){
  M      <- as.matrix(sens[[k]]$samples)
  data.frame(
    Model   = k,
    mu_beta = mean(M[,"mu_beta"]),
    mu_beta_lo = quantile(M[,"mu_beta"], .025),
    mu_beta_hi = quantile(M[,"mu_beta"], .975),
    sigma_beta = mean(M[,"sigma_beta"]),
    sigma_beta_lo = quantile(M[,"sigma_beta"], .025),
    sigma_beta_hi = quantile(M[,"sigma_beta"], .975),
    elpd_loo = sens[[k]]$loo_elpd,
    elpd_se  = sens[[k]]$loo_se
  )
})
print(sumtab, digits = 3)
```

Appendix G

```
a_idx <- grep("^alpha\\[", colnames(post_mat))
b_idx <- grep("^beta\\[",  colnames(post_mat))

S <- nrow(post_mat)
J <- N_country

T_obs  <- numeric(S)
T_rep  <- numeric(S)

for (s in seq_len(S)) {

  a_s <- post_mat[s, a_idx]
  b_s <- post_mat[s, b_idx]

  lam <- exp(climate_data$logPop +
             a_s[country_index] +
             b_s[country_index] * climate_data$year_ctr)

  lam_j <- tapply(lam,  country_index, sum)
  y_j   <- tapply(climate_data$Extreme_Weather_Events,
                  country_index, sum)

  y_sim  <- rpois(N_obs, lam)
  ysim_j <- tapply(y_sim, country_index, sum)

  ## χ^2 discrepancies
  T_obs[s] <- sum((y_j    - lam_j)^2 / lam_j)
  T_rep[s] <- sum((ysim_j - lam_j)^2 / lam_j)
}

bayes_p <- mean(T_rep >= T_obs)
cat("Country-wise χ² PPC  p_B =", round(bayes_p, 3), "\n")
```