Data Mangling with Climate Disasters and Bayesian Modeling: Lessons from Real-World Data

Mlen-Too Wesley

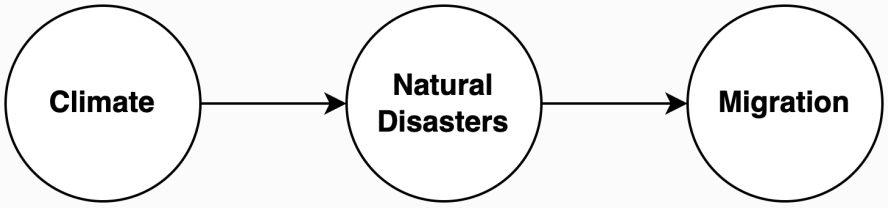
ISYE 6420 - Fall 2024 - Final Project

December 1, 2024

Introduction

I began this project interested in the intersection between climate change and migration, as this intricate relationship has garnered significant attention. Recent events, such as the devastating floods across Nigeria, Mali, and Niger, have displaced millions and exacerbated hunger crises in Africa. These incidents underscore the urgency of understanding how climate-induced disasters influence human mobility.

Climate events take time and as such, my research aimed to measure the causal impact using natural disasters as indicators of extreme events in a quasi-experiment design. I intended utilize Bayesian hierarchical modeling to map the progression from climate change to natural disasters, and subsequently to migration patterns and employ methodologies like regression discontinuity and difference-in-differences to analyze this relationship. I was too ambitious.



Following an extensive phase of data collection and cleaning, seeking datasets that could provide reliable measures of climate change effects, natural disaster occurrences, and migration statistics, the migration data I had was insufficient in scope and granularity to support my project.

Given these constraints, I refocused my analysis to just the relationship between climate change and natural disasters. This adaptation involved developing a simplified model that utilized Bayesian hierarchical modeling techniques to explore and map out the interactions between climatic variables and natural disaster events.

This paper outlines the revised research approach, detailing the methodologies employed, the data utilized, and the insights gained from the analysis. It also discusses the setbacks encountered during the study and how these influenced the project's trajectory. Moreover, it showcases my real life effort working with incomplete data to do Bayesian analysis.

Data Collection

With climate change data, I found more sources that provided maps and satellite imagery than tabular data. However, I eventually stumbled upon the Global Historical Climatology Network (GHCN) managed by the National Oceanic and Atmospheric Administration (NOAA), which has an extensive collection of records globally, that is also accessible online.

I utilized the NOAA Global Summary of the Year (GSOY) data which offers detailed annual climate data. This dataset summarized broader climatic trends and extremes. I actually attempted to use the Global Surface Summary of the Month (GSOM) first, however this was challenging due to its size and complexity. I knew climate and migration was summarized by year, so I went with GSOY.

For natural disaster data, the EM-DAT International Disaster Database by the Centre for Research on the Epidemiology of Disasters (CRED) is a leading source, providing comprehensive historical records of disaster events worldwide, including their impacts on populations and economies. There are some caveats, relating to how past data collection may be unreliable.

Migration data, however, is often more fragmented and challenging to consolidate due to the complexity of migration patterns and the variability in data collection methods across regions.

- The United Nations Department of Economic and Social Affairs (UNDESA) provides international migration stock and flow data, detailing cross-border population movements at periodic intervals, making it one of the most comprehensive sources for global migration statistics.
- The World Bank's Migration and Remittances Data portal offers datasets focused on migration flows and associated financial remittances, which are critical for understanding the economic dynamics of migration. The International Organization for Migration (IOM) manages the Displacement Tracking Matrix (DTM), providing real-time data on internal displacement often linked to natural disasters, conflicts, and other crises.
- The KNOMAD (Global Knowledge Partnership on Migration and Development) platform complements these datasets by offering open-access data on migration trends, policies, and their socio-economic implications.
- The Determinants of International Migration (DEMIG) database, curated by the International Migration Institute, provides detailed data on migration flows, including bilateral migration flow estimates and immigration policies, making it a valuable resource for studying the drivers and patterns of migration over time.

I made a first attempt to use the DEMIG database, but it was very limited, focusing on a handful of countries. So I opted for the UNDESA data, however this was limited to 5-year periods from 1990 to 2020. Unfortunately, this dataset's limitations significantly shaped the study's scope, emphasizing the need for more comprehensive migration datasets in future research. With enough time, I can revisit the data and complete the project as I originally envisioned.

Finally, on a side note, I found a lot of national, sub-national and regional migration data, collected from national census data. However, this was even more difficult to incorporate into my project, and was likely incorporated into the larger DEMIG and UNDESA databases anyways.

The data and source codes are available at: https://github.com/mtwesley/isye6420-final-project

Here's a quick breakdown:

- **demig/demig-total-migration-database_v1-5.xlsx:** Migration data from the DEMIG project
- **emdat/emdat-country-profiles_2024_11_25.xlsx:** Country-level disaster profiles from EM-DAT with disaster frequency, types, and impacts, deaths, economic loss
- **noaa_ncei:** Multiple subfolders and files for organizing large climate datasets from NOAA National Centers for Environmental Information
- **noaa_ncei/gsoy-aggregated-all-countries.csv and noaa_ncei/gsoy-reaggregated-all-countries.csv**: Final merged, processed and reprocessed GSOY datasets, and aggregated at the country level
- **opencage/opencage-ghcnd-prefix-lookup.csv:** Mapping information for OpenCage geolocation services and GHCND station prefixes to geolocating or identifying climate station data
- **opendata/country-codes.csv:** Mapping relationships between country names and codes (ISO Alpha-2, Alpha-3, or M49) to harmonize datasets with different conventions for countries
- **opendata/population.csv**: Population data for countries and regions by year as a demographic indicator
- **undesa/undesa_pd_2020_ims_stock_by_sex_destination_and_origin.xlsx and udesa/undesa_pd_2020_ims_stock_origin_world.xlsx:** Excel files from the UNDESA with migration stock data categorized by destination and origin countries up to the year 2020

Data Preparation

The preparation of the selected data involved extensive cleaning and manipulation using both Bash and R scripts. Such as

```r
input_dir <- "noaa_ncei/gsoy-latest"
output_dir <- "noaa_ncei/gsoy-merged"

if (!dir.exists(input_dir)) stop("Input directory does not exist: ", input_dir)
if (!dir.exists(output_dir)) dir.create(output_dir, showWarnings = FALSE)

file_list <- list.files(input_dir, pattern = "\\.csv$", full.names = TRUE)
get_country_code <- function(filename) substr(basename(filename), 1, 2)
country_files_map <- split(file_list, sapply(file_list, get_country_code))

for (country_code in names(country_files_map)) {
  files <- country_files_map[[country_code]]
  country_data <- data.frame(matrix(ncol = length(variables), nrow = 0))
  colnames(country_data) <- variables
  cat("Processing country:", country_code, "\n")

  for (file_path in files) {
    cat("  Reading file:", file_path, "\n")
    file_data <- read.csv(file_path, stringsAsFactors = FALSE)

    filtered_data <- file_data[, intersect(colnames(file_data), variables), drop = FALSE]
    filtered_data$COUNTRY <- country_code

    for (col in setdiff(variables, colnames(filtered_data))) {
      filtered_data[[col]] <- NA
```

```
    }
    filtered_data <- filtered_data[, variables]

    country_data <- rbind(country_data, filtered_data)
  }

  output_file <- file.path(output_dir, paste0(country_code, ".csv"))
  write.csv(country_data, output_file, row.names = FALSE, na = "")
  cat("  Merged file written for:", country_code, "\n")
}
```

Once merged, data often had to be aggregated or re-aggregated if there were problems or mistakes along the way.

```
input_dir <- "noaa_ncei/gsoy-remerged"
output_dir <- "noaa_ncei/gsoy-reaggregated"

if (!dir.exists(input_dir)) stop("Input directory does not exist: ", input_dir)
if (!dir.exists(output_dir)) dir.create(output_dir, showWarnings = FALSE)

file_list <- list.files(input_dir, pattern = "\\.csv$", full.names = TRUE)

for (file_path in file_list) {
  cat("Processing:", file_path, "\n")

  data <- read.csv(file_path, stringsAsFactors = FALSE)
  non_numeric_cols <- c("COUNTRY", "DATE")
  numeric_cols <- setdiff(colnames(data), non_numeric_cols)

  aggregated_data <- data %>%
    group_by(COUNTRY, DATE) %>%
    summarize(
      across(
        all_of(numeric_cols),
        list(
          MIN = ~ min(., na.rm = TRUE),
          MAX = ~ max(., na.rm = TRUE),
          MEAN = ~ mean(., na.rm = TRUE),
          SD = ~ sd(., na.rm = TRUE)
        ),
        .names = "{.col}_{.fn}"
      ),
      .groups = "drop" # Ensures the result is ungrouped
    ) %>%
    arrange(DATE) %>%
    mutate(across(everything(), ~ ifelse(is.infinite(.), NA, .)))

  output_file <- file.path(output_dir, basename(file_path))
  write.csv(aggregated_data, output_file, row.names = FALSE, na = "")
  cat("  Written:", output_file, "\n")
}
```

The repository contains many other similar scripts used during the data collection and cleaning process. These aided in the process of gathering, merging, and aggregating data for us in the analysis.

Now, with the data on my computer, I explored the variables available to understand how to combine them into one unified dataset and ensuring data integrity, such as handling missing data and standardizing formats across datasets.

For example, the climate data was from weather stations. I could recognize them as two-letter country codes. However, when unifying climate data with natural disaster data, there were many that did not match. I then realized that the coding changes over time. To adjust for this, I had to use an API to map GPS coordinates to current countries, build lookup tables, and remerge and re-aggregate the data. Which took hours.

Here are some examples of the climate data:

row names	alpha2	year	awnd_min	awnd_max	awnd_mean	awnd_sd	cdsd_min	cdsd_max	cdsd_mean	cdsd_sd	cldd_min	cldd_max	cldd_mean	cldd_sd	dp01_min	dp01_max
		1994	3.0	3.0	3.0	NA	3413.4	3413.4	3413.400	NA	3413.4	3413.4	3413.400	NA	256	256
3344	PS	1995	2.5	2.5	2.5	NA	2925.5	3440.0	3182.750	363.80644	2925.5	3440.0	3182.750	363.80644	241	280
3345	PS	1996	2.6	2.6	2.6	NA	2928.8	3466.2	3197.500	379.99918	2928.8	3466.2	3197.500	379.99918	247	274
3346	PS	1997	2.7	2.7	2.7	NA	2954.6	3468.9	3211.750	363.66502	2954.6	3468.9	3211.750	363.66502	215	238
3347	PS	1998	2.6	2.6	2.6	NA	3564.7	3564.7	3564.700	NA	3564.7	3564.7	3564.700	NA	252	252
3348	PS	1999	2.5	2.5	2.5	NA	2698.7	3469.8	3026.367	398.37199	2698.7	3469.8	3026.367	398.37199	276	290
3349	PS	2000	2.4	2.4	2.4	NA	3467.2	3467.2	3467.200	NA	3467.2	3467.2	3467.200	NA	275	275
3350	PS	2001	2.4	2.4	2.4	NA	3611.4	3611.4	3611.400	NA	3611.4	3611.4	3611.400	NA	247	248
3351	PS	2002	2.8	2.8	2.8	NA	3676.9	3676.9	3676.900	NA	3676.9	3676.9	3676.900	NA	243	259
3352	PS	2003	2.6	2.6	2.6	NA	3579.5	3579.5	3579.500	NA	3579.5	3579.5	3579.500	NA	250	278
3353	PS	2004	2.6	2.6	2.6	NA	3677.5	3692.5	3685.000	10.60660	3677.5	3692.5	3685.000	10.60660	240	264
3354	PS	2005	2.8	2.8	2.8	NA	3543.3	3573.5	3558.400	21.35462	3543.3	3573.5	3558.400	21.35462	280	283
3355	PS	2006	NA	NA	NA	NA	3566.5	3619.9	3593.200	37.75950	3566.5	3619.9	3593.200	37.75950	251	269
3356	PS	2007	2.7	2.7	2.7	NA	3518.7	3649.3	3584.000	92.34815	3518.7	3649.3	3584.000	92.34815	268	268
3357	PS	2008	NA	NA	NA	NA	3375.9	3443.0	3409.450	47.44687	3375.9	3443.0	3409.450	47.44687	289	297
3358	PS	2009	2.7	2.7	2.7	NA	3350.3	3407.4	3378.850	40.37580	3350.3	3407.4	3378.850	40.37580	278	282

And the migration data:

	country	m49	year	migrants	alpha2	alpha3
1	Afghanistan	4	1990	7679582	AF	AFG
2	Afghanistan	4	1995	4347049	AF	AFG
3	Afghanistan	4	2000	4750677	AF	AFG
4	Afghanistan	4	2005	4116739	AF	AFG
5	Afghanistan	4	2010	5269518	AF	AFG
6	Afghanistan	4	2015	5400916	AF	AFG
7	Afghanistan	4	2020	5853838	AF	AFG
8	Albania	8	1990	180204	AL	ALB
9	Albania	8	1995	501066	AL	ALB
10	Albania	8	2000	824442	AL	ALB
11	Albania	8	2005	966032	AL	ALB
12	Albania	8	2010	1117940	AL	ALB

And the disaster data:

	year	alpha3	country	region	subregion	storms	floods	deaths	livesAffected	economicDamage	alpha2	m49
1	1900	JAM	Jamaica	Latin America and the Caribbean	Americas	0	1	300	0	0	JM	388
2	1900	USA	United States of America	Northern America	Americas	1	0	6000	0	1098720	US	840
3	1902	MMR	Myanmar	South-eastern Asia	Asia	1	0	600	0	0	MM	104
4	1903	JAM	Jamaica	Latin America and the Caribbean	Americas	1	0	65	0	0	JM	388
5	1903	USA	United States of America	Northern America	Americas	1	2	348	0	16277328	US	840
6	1904	BGD	Bangladesh	Southern Asia	Asia	1	0	0	0	0	BD	50
7	1905	PHL	Philippines	South-eastern Asia	Asia	1	0	240	0	0	PH	608
8	1906	HKG	China, Hong Kong Special Administrative Region	Eastern Asia	Asia	1	0	10000	0	678222	HK	344
9	1906	USA	United States of America	Northern America	Americas	2	0	298	0	0	US	840
10	1909	BGD	Bangladesh	Southern Asia	Asia	2	0	172	0	0	BD	50
11	1909	HTI	Haiti	Latin America and the Caribbean	Americas	1	0	150	0	0	HT	332
12	1909	USA	United States of America	Northern America	Americas	2	1	463	0	0	US	840
13	1910	JPN	Japan	Eastern Asia	Asia	0	1	1379	0	0	JP	392

Overall, the preparation phase was both time-intensive and somewhat incomplete, as I was not able to prepare the migration data in time for submitting the project. However, it was a good experience and has set a solid foundation for subsequent modeling efforts.

```r
# Country codes
country_codes <- read.csv("opendata/country-codes.csv") %>%
  select(alpha2 = ISO3166.1.Alpha.2, alpha3 = ISO3166.1.Alpha.3, m49 = M49)

# Population data
population_data <- read.csv("opendata/population.csv") %>%
  select(alpha3 = `Country.Code`, country = `Country.Name`, year = Year, population = Value)
%>%
  rename_with(tolower)

# Load EM-DAT disaster data
emdat_data <- read_excel("emdat/public_emdat_incl_hist_2024-11-25.xlsx", sheet = "EM-DAT
Data") %>%
  filter(`Disaster Type` %in% c("Flood", "Storm")) %>%
  select(
    year = `Start Year`,
    alpha3 = ISO,
    country = Country,
    region = Subregion,
    subregion = Region,
    disaster = `Disaster Type`,
    deaths = `Total Deaths`,
    livesAffected = `Total Affected`,
    economicDamage = `Total Damage, Adjusted ('000 US$)`
  )

# Focus on floods and storms
disaster_data <- emdat_data %>%
  group_by(year, alpha3, country, region, subregion) %>%
  summarise(
    storms = sum(disaster == "Storm", na.rm = TRUE),
    floods = sum(disaster == "Flood", na.rm = TRUE),
    deaths = sum(deaths, na.rm = TRUE),
    livesAffected = sum(livesAffected, na.rm = TRUE),
    economicDamage = sum(economicDamage, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  left_join(country_codes, by = "alpha3", relationship = "many-to-many")

# Load UNDESA migration data
undesa_data <- read_excel("undesa/undesa_pd_2020_ims_stock_origin_world.xlsx", skip = 3)
%>%
  rename(country = `Region, development group, country or area of origin`,
         m49 = `Location code of origin`)

migration_data <- undesa_data %>%
  pivot_longer(cols = -c(country, m49), names_to = "year", values_to = "migrants") %>%
  group_by(country, m49, year) %>%
  summarise(migrants = sum(migrants, na.rm = TRUE), .groups = "drop") %>%
  left_join(country_codes, by = "m49", relationship = "many-to-many")
```

```r
# Load and clean GSOY climate data
gsoy_data <- read_csv("noaa_ncei/gsoy-reaggregated-all-countries.csv") %>%
  rename(year = DATE, alpha2 = COUNTRY) %>%
  rename_with(tolower)

climate_data <- gsoy_data %>%
  left_join(country_codes, by = "alpha2", relationship = "many-to-many")

# Limit all data to period between 1980–2020
climate_filtered <- climate_data %>%
  filter(year >= 1980 & year <= 2020)

disaster_filtered <- disaster_data %>%
  filter(year >= 1980 & year <= 2020)

migration_filtered <- migration_data %>%
  filter(year >= 1980 & year <= 2020)

# Unify datasets on year and alpha3
climate_variables <- c(
  "year", "alpha3", "alpha2",
  "prcp_mean", "emxp_max", "emnt_min", "emxt_max",
  "tmax_max", "tmin_min", "tavg_mean"
)

disaster_climate_determinants <- climate_filtered %>%
  select(all_of(climate_variables)) %>%
  inner_join(disaster_filtered,
             by = c("year", "alpha3"),
             relationship = "many-to-many") %>%
  select(everything(), -matches("\\.x$|\\.y$"))

# Create a coverage matrix for data availability
country_coverage <- disaster_climate_determinants %>%
  group_by(alpha3) %>%
  summarise(across(where(is.numeric), ~ sum(!is.na(.)) / length(1980:2020)))

country_coverage_averages <- country_coverage %>%
  rowwise() %>%
  mutate(average_coverage = mean(c_across(where(is.numeric)), na.rm = TRUE)) %>%
  select(alpha3, average_coverage)

variable_coverage_averages <- country_coverage %>%
  summarise(across(where(is.numeric), ~ mean(.x, na.rm = TRUE))) %>%
  pivot_longer(everything(), names_to = "variable", values_to = "average_coverage") %>%
  arrange(desc(average_coverage))

# Get the top 80 countries by coverage
top_80_countries <- country_coverage_averages %>%
  arrange(desc(average_coverage)) %>%
  slice_head(n = 80)
```

```r
# Filter the original country_coverage to keep only the top 80
disaster_determinants_80_countries <- disaster_climate_determinants %>%
  filter(alpha3 %in% top_80_countries$alpha3)
```

Model overview and approach

A hierarchical Bayesian model was used to investigate the relationship between climate variables, extreme weather events, and natural disasters, specifically floods and storms. The choice of disasters (floods and storms) was chosen due to its variability across multiple countries and regions, which was validated with the data.

With Bayesian modeling, I could incorporate uncertainty at multiple levels and map out the system hierarchically. The model is designed to capture the interplay between average climatic conditions, extreme weather events, and the occurrence of natural disasters, while allowing for flexibility in the specification of uncertainty through prior distributions.

The modeling framework rests on two primary components:

1. Climate variables on extreme weather events
2. Effect of both climate variables and extreme weather events on disasters such as floods and storms

In future studies, I hope to finally link this with migration.

Observed Data

The observed variables include temperature, precipitation, extreme weather variables, and disaster occurrence counts. These variables form the foundational inputs for the model:

- **Temperature variables**: Mean temperature (t_{avg}), minimum temperature (t_{min}), and maximum temperature (t_{max})
- **Precipitation variable**: Mean precipitation (p)
- **Extreme weather events**: Minimum extreme temperature (e_{mnt}), maximum extreme temperature (e_{mxt}), and maximum extreme precipitation (e_{mxp})
- **Disaster counts**: Flood f and storm s occurrences

These data are preprocessed to ensure completeness and consistency, with missing data excluded to prevent bias in inference. Each variable represents an aggregated yearly value at the country level. In the future, I could also measure grouped effects at regional and subregional levels.

Non-informative priors for predictors and variance terms

Due to lack of domain expertise, the model adopts weakly informative priors for the coefficients linking predictors to outcomes. Priors are necessary in Bayesian analysis to express initial beliefs about parameter values, while allowing flexibility for the data to dominate inference.

The coefficients for the relationship between climate variables and extreme events, as well as between extreme events and disasters, are modeled using normal distributions centered at zero:

$$\beta \sim \mathcal{N}(0, 10)$$

This prior reflects the assumption that, a priori, the relationships between these variables are likely centered around zero with moderate uncertainty. For variance terms related to extreme events and disaster outcomes, we use truncated normal distributions:

$$\sigma \sim \mathcal{N}(0, 5), \text{ truncated to } (0, \infty)$$

The truncation ensures that variance terms are strictly positive, reflecting their role in modeling uncertainty.

Modeling extreme weather events

Extreme weather events are modeled as outcomes of temperature and precipitation variables. The rationale for this modeling choice is that climatic conditions, such as temperature and precipitation, influence the likelihood and severity of extremes.

- Extreme Minimum Temperature (e_{mnt}): The expected value of extreme minimum temperature is modeled as a linear combination of mean temperature (t_{avg}) and minimum temperature (t_{min}) and the observed e_{mnt} is assumed to follow a normal distribution:

$$\mu_{\text{emnt}} = \beta_{\text{tavg}} \cdot t_{\text{avg}} + \beta_{\text{tmin}} \cdot t_{\text{min}}$$
$$e_{\text{mnt}} \sim \mathcal{N}(\mu_{\text{emnt}}, \sigma_{\text{emnt}})$$

- Extreme Maximum Temperature (e_{mxt}): Similarly, the expected value of extreme maximum temperature is influenced by mean temperature (t_{avg}) and maximum temperature(t_{max}):

$$\mu_{\text{emxt}} = \beta_{\text{tavg}} \cdot t_{\text{avg}} + \beta_{\text{tmax}} \cdot t_{\text{max}}$$
$$e_{\text{mxt}} \sim \mathcal{N}(\mu_{\text{emxt}}, \sigma_{\text{emxt}})$$

- Extreme Precipitation (e_{mxp}): The expected value of extreme precipitation is modeled as proportional to mean precipitation (p):

$$\mu_{\text{emxp}} = \beta_{\text{prcp}} \cdot p$$
$$e_{\text{mxp}} \sim \mathcal{N}(\mu_{\text{emxp}}, \sigma_{\text{emxp}})$$

These ensure that extreme weather events are directly tied to climate variables while incorporating uncertainty in their predictions.

Modeling natural disasters

Natural disasters, such as floods (f) and storms (s), are modeled as Poisson processes, where the rate parameters (λ_f and λ_s) are functions of climate variables and extreme events. The Poisson distribution is chosen due to its suitability for modeling count data, such as disaster occurrences, but due to model design, the data is exponentiated prior to being used as a rate.

- **Floods:** The log-rate of floods is expressed as:

$$\log \lambda_f = \beta_{\text{temp-floods}} \cdot (t_{\text{avg}} + e_{\text{mnt}} + e_{\text{mxt}}) + \beta_{\text{prcp-floods}} \cdot p + \beta_{\text{extreme-precip-floods}} \cdot e_{\text{mxp}}$$

- Storms: The log-rate of storms is similarly modeled:

$$\log \lambda_s = \beta_{\text{temp-storms}} \cdot (t_{\text{avg}} + e_{\text{mnt}} + e_{\text{mxt}}) + \beta_{\text{prcp-storms}} \cdot p + \beta_{\text{extreme-precip-storms}} \cdot e_{\text{mxp}}$$

Sampling and inference

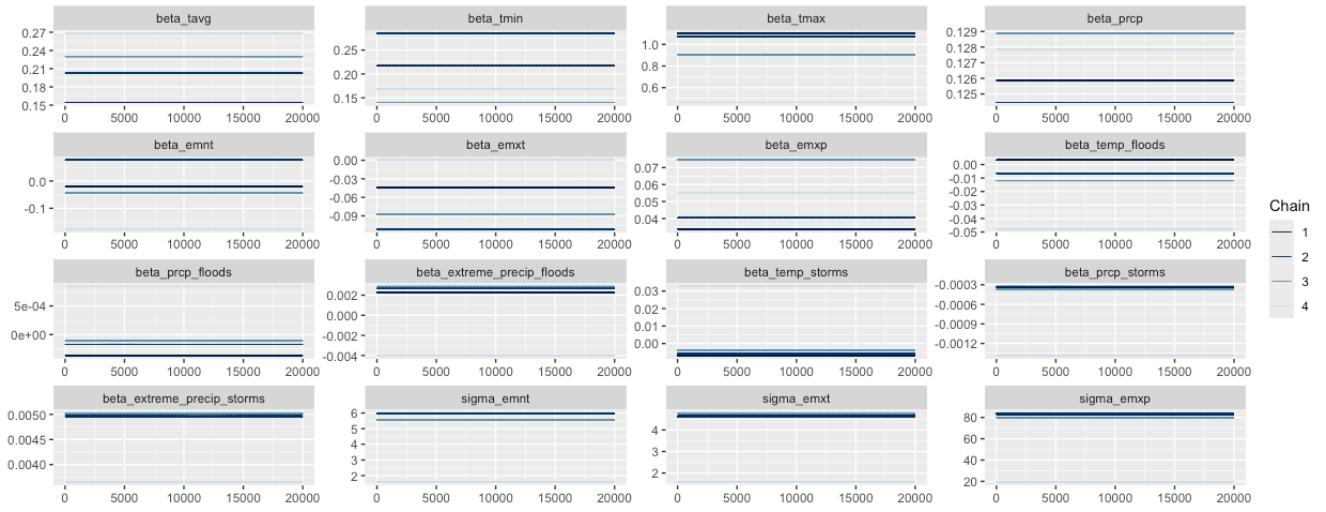
Inference is conducted using Markov Chain Monte Carlo (MCMC) sampling, enabling estimation of posterior distributions for all parameters, including β coefficients and σ variance terms.

The hierarchical structure of the model reflects the natural order of influence, that climate variables affect extreme weather events, which in turn drive disaster outcomes. This approach aligns with theoretical expectations and allows for the incorporation of uncertainty at every stage. The use of Poisson distributions for disasters is particularly suited to count data, while the normal distributions for extreme events offer flexibility in modeling continuous outcomes.

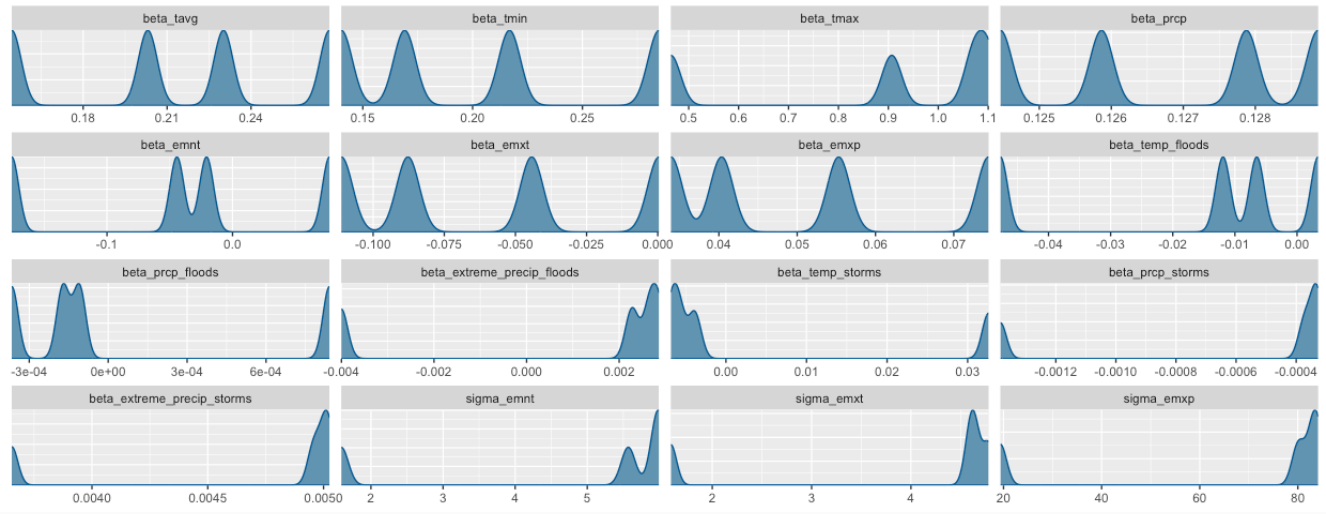
```
library(greta)

determinants <- disaster_determinants_partitioned_subsets[[1]]

# Observed data
tavg <- as_data(determinants$tavg_mean)
tmin <- as_data(determinants$tmin_min)
tmax <- as_data(determinants$tmax_max)

prcp <- as_data(determinants$prcp_mean)

emnt <- as_data(determinants$emnt_min)
emxt <- as_data(determinants$emxt_max)
emxp <- as_data(determinants$emxp_max)

floods <- as_data(determinants$floods)
storms <- as_data(determinants$storms)

# Priors for predictors' effects
beta_tavg <- normal(0, 10)
beta_tmin <- normal(0, 10)
beta_tmax <- normal(0, 10)
beta_prcp <- normal(0, 10)

# Priors for effects on extreme temperatures and precipitation
beta_emnt <- normal(0, 10)
beta_emxt <- normal(0, 10)
beta_emxp <- normal(0, 10)

# Priors for coefficients linking predictors to floods
beta_temp_floods <- normal(0, 10)
beta_prcp_floods <- normal(0, 10)
beta_extreme_precip_floods <- normal(0, 10)

# Priors for coefficients linking predictors to storms
beta_temp_storms <- normal(0, 10)
beta_prcp_storms <- normal(0, 10)
beta_extreme_precip_storms <- normal(0, 10)

# Variance terms for extreme events
sigma_emnt <- normal(0, 5, truncation = c(0, Inf))
sigma_emxt <- normal(0, 5, truncation = c(0, Inf))
sigma_emxp <- normal(0, 5, truncation = c(0, Inf))

# Models for extreme events
emnt_mean <- beta_tavg * tavg + beta_tmin * tmin
```

```
distribution(emnt) <- normal(emnt_mean, sigma_emnt)

emxt_mean <- beta_tavg * tavg + beta_tmax * tmax
distribution(emxt) <- normal(emxt_mean, sigma_emxt)

emxp_mean <- beta_prcp * prcp
distribution(emxp) <- normal(emxp_mean, sigma_emxp)

# Floods modeled as Poisson with exp() rate
floods_rate <- exp(
  beta_temp_floods * (tavg + emnt + emxt) +
    beta_prcp_floods * prcp +
    beta_extreme_precip_floods * emxp
)
distribution(floods) <- poisson(floods_rate)

# Storms modeled as Poisson with exp() rate
storms_rate <- exp(
  beta_temp_storms * (tavg + emnt + emxt) +
    beta_prcp_storms * prcp +
    beta_extreme_precip_storms * emxp
)
distribution(storms) <- poisson(storms_rate)

# Define the model
disaster_model <- model(
  beta_tavg, beta_tmin, beta_tmax, beta_prcp,
  beta_emnt, beta_emxt, beta_emxp,
  beta_temp_floods, beta_prcp_floods, beta_extreme_precip_floods,
  beta_temp_storms, beta_prcp_storms, beta_extreme_precip_storms,
  sigma_emnt, sigma_emxt, sigma_emxp
)

# Sample the posterior
draws <- mcmc(disaster_model, warmup = 4000, n_samples = 10000, chains = 4)

mcmc_trace(draws)
mcmc_dens(draws)

get.stats(draws)
```

Trace plots were pretty horrible.



Density plots were no better:



	variable	mean	median	sd	mad	q5	q95	hdi5	hdi95	rhat	ess_bulk	ess_tail
1	beta_tavg	2.13947e-01	0.216630676	4.13656e-02	4.81426e-02	0.154474926	0.268051994	0.154474896	0.268052059	3.58031	4.35503	10.9591
2	beta_tmin	2.02737e-01	0.192930715	5.47596e-02	5.67896e-02	0.140189137	0.284898663	0.140189115	0.284898683	2.88112	4.58871	24.7431
3	beta_tmax	8.85686e-01	0.988839038	2.54311e-01	1.43800e-01	0.464172273	1.100893271	0.464172208	1.100893293	2.91574	4.56265	32.5982
4	beta_prcp	1.26769e-01	0.126870364	1.71797e-03	2.23104e-03	0.124459623	0.128873674	0.124459614	0.128873773	3.31645	4.42071	11.3343
5	beta_emnt	-4.09033e-02	-0.032493223	9.02989e-02	9.01271e-02	-0.175910897	0.077283651	-0.175910919	0.077283740	3.64204	4.33968	11.5708
6	beta_emxt	-6.06930e-02	-0.066024824	4.27039e-02	4.96033e-02	-0.111196813	0.000474658	-0.111196836	0.000474673	3.47637	4.37273	14.5644
7	beta_emxp	5.09899e-02	0.047824726	1.56162e-02	1.58921e-02	0.033861940	0.074448223	0.033861935	0.074448231	3.68775	4.36456	13.6060
8	beta_temp_floods	-1.56879e-02	-0.009199876	1.93076e-02	1.13511e-02	-0.047749312	0.003397377	-0.047749340	0.003397381	3.27454	4.48449	13.8710
9	beta_prcp_floods	4.83528e-05	-0.000140736	4.67259e-04	1.90475e-04	-0.000366204	0.000841063	-0.000366212	0.000841231	3.34956	4.41599	11.2713
10	beta_extreme_precip_floods	9.46792e-04	0.002464135	2.86565e-03	4.34933e-04	-0.004003483	0.002862227	-0.004003505	0.002862261	3.63353	4.34194	10.9598
11	beta_temp_storms	4.01357e-03	-0.004852968	1.65411e-02	2.21296e-03	-0.006842540	0.032602825	-0.006842565	0.032602829	3.02134	4.51525	23.1269
12	beta_prcp_storms	-6.05338e-04	-0.000354946	4.50156e-04	3.42800e-05	-0.001384523	-0.000326905	-0.001384555	-0.000326897	3.22044	4.45048	10.6925
13	beta_extreme_precip_storms	4.66003e-03	0.004980601	5.81399e-04	5.04919e-05	0.003653808	0.005025376	0.003653574	0.005025389	3.54576	4.35991	11.2712
14	sigma_emnt	4.78177e+00	5.771270902	1.85154e+00	3.16542e-01	1.588358766	5.996174467	1.588358482	5.996174484	3.34787	4.40525	20.5086
15	sigma_emxt	3.90096e+00	4.619608171	1.34024e+00	1.38672e-01	1.582647629	4.781978888	1.582646891	4.781978891	4.50757	4.21528	11.2940
16	sigma_emxp	6.66419e+01	81.535663116	2.73428e+01	3.00812e+00	19.352328336	84.144131937	19.352318449	84.144132322	3.21904	4.45107	11.4940

Alternative model for comparison

Given the poor results, I tested another model. This second model is also designed to explore the relationship between climate variables, extreme events, and natural disasters (floods and storms) by employing a normal distribution for the disaster outcomes (f and s) instead of the Poisson-log-link approach in the first model.

By using normal distributions, this model assumes that disaster occurrences are continuous rather than discrete counts, which would have been useful for other metrics, but it also allows for a simpler probabilistic structure.

Observed Data

As in the first model, the observed data comprises climate variables, extreme events, and disaster outcomes:

- **Climate variables:**
  - t_{avg} (mean temperature)
  - t_{min} (minimum temperature)
  - t_{max} (maximum temperature)
  - p (mean precipitation)
- **Extreme weather events:**
  - e_{mnt} (minimum extreme temperature)
  - e_{mxt} (maximum extreme temperature)
  - e_{mxp} (extreme precipitation)
- **Disaster counts: $**f($floods$)and$s$$** (storms)

The priors remain identical to those in the first model, reflecting the same assumptions about uncertainty and initial beliefs:

- Coefficients for relationships (β) for linking climate variables to extreme events and extreme events to disaster outcomes

$$\beta \sim \mathcal{N}(0, 10)$$

- Variance terms (σ) for extreme events and disasters:

$$\sigma \sim \mathcal{N}(0, 5) \text{ truncated to } (0, \infty)$$

As with the first model, extreme weather events are modeled as functions of climate variables:

1. **Extreme Minimum Temperature** (e_{mnt}):

$$\mu_{\text{emnt}} = \beta_{\text{tavg}} \cdot t_{\text{avg}} + \beta_{\text{tmin}} \cdot t_{\text{min}}$$

$$e_{\text{mnt}} \sim \mathcal{N}(\mu_{\text{emnt}}, \sigma_{\text{emnt}})$$

2. **Extreme Maximum Temperature** (e_{mxt}):

$$\mu_{\text{emxt}} = \beta_{\text{tavg}} \cdot t_{\text{avg}} + \beta_{\text{tmax}} \cdot t_{\text{max}}$$

$$e_{\text{mxt}} \sim \mathcal{N}(\mu_{\text{emxt}}, \sigma_{\text{emxt}})$$

3. **Extreme Precipitation** (e_{mxp}):

$$\mu_{\text{emxp}} = \beta_{\text{prcp}} \cdot p$$

$$e_{\text{mxp}} \sim \mathcal{N}(\mu_{\text{emxp}}, \sigma_{\text{emxp}})$$

Finally, floods (f) and storms (s) are modeled as continuous outcomes, which is distinct from the Poisson process used in the first model:

1. **Floods**:

$$\mu_f = \beta_{\text{temp\_floods}} \cdot (t_{\text{avg}} + e_{\text{mnt}} + e_{\text{mxt}}) + \beta_{\text{prcp\_floods}} \cdot p + \beta_{\text{extreme\_precip\_floods}} \cdot e_{\text{mxp}}$$

$$f \sim \mathcal{N}(\mu_f, \sigma_f)$$

2. **Storms**:

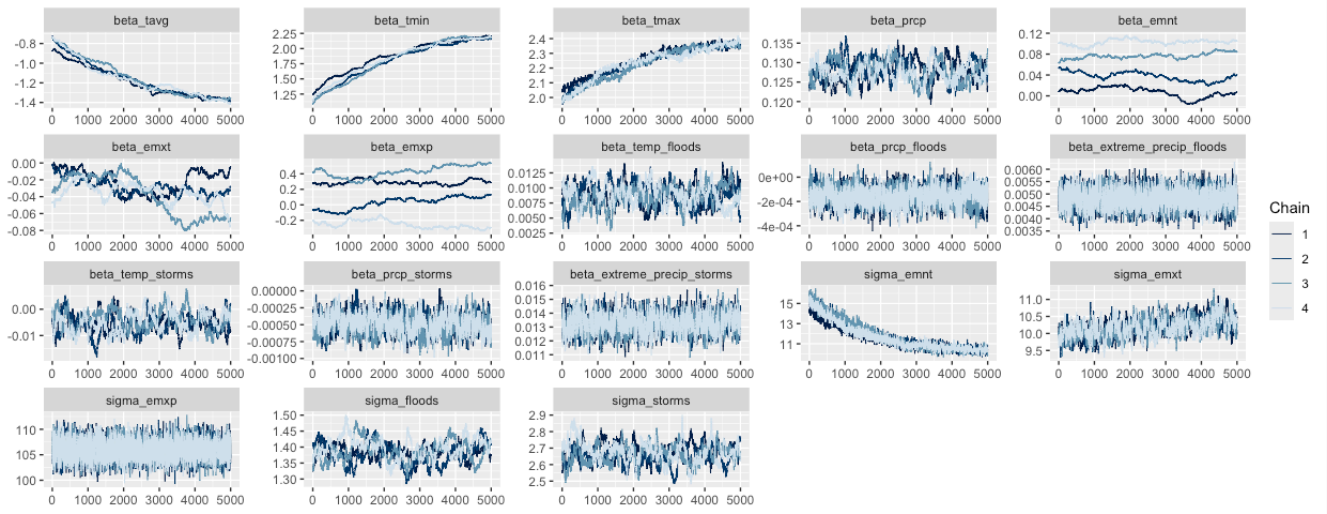
$$\mu_s = \beta_{\text{temp\_storms}} \cdot (t_{\text{avg}} + e_{\text{mnt}} + e_{\text{mxt}}) + \beta_{\text{prcp\_storms}} \cdot p + \beta_{\text{extreme\_precip\_storms}} \cdot e_{\text{mxp}}$$

$$s \sim \mathcal{N}(\mu_s, \sigma_s)$$

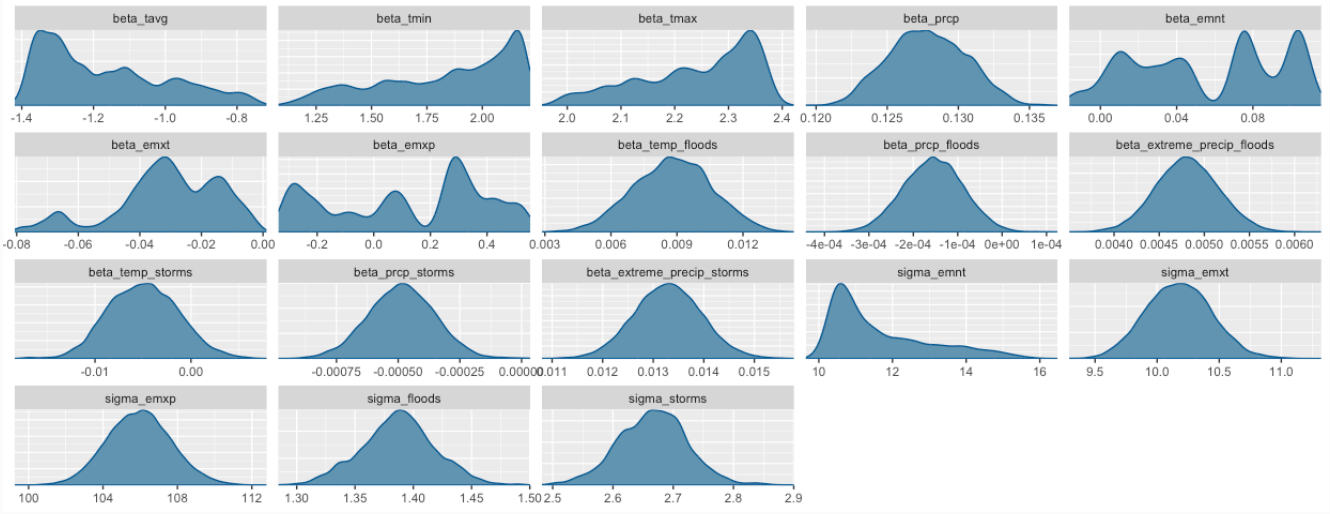
```
library(greta)

determinants <- disaster_determinants_partitioned_subsets[[1]]

# Observed data
tavg <- as_data(determinants$tavg_mean)
tmin <- as_data(determinants$tmin_min)
tmax <- as_data(determinants$tmax_max)

prcp <- as_data(determinants$prcp_mean)

emnt <- as_data(determinants$emnt_min)
emxt <- as_data(determinants$emxt_max)
emxp <- as_data(determinants$emxp_max)

floods <- as_data(determinants$floods)
storms <- as_data(determinants$storms)

# Priors for predictors' effects
beta_tavg <- normal(0, 10)
beta_tmin <- normal(0, 10)
beta_tmax <- normal(0, 10)
beta_prcp <- normal(0, 10)

# Priors for effects on extreme temperatures and precipitation
beta_emnt <- normal(0, 10)
beta_emxt <- normal(0, 10)
beta_emxp <- normal(0, 10)

# Priors for coefficients linking predictors to floods
beta_temp_floods <- normal(0, 10)
beta_prcp_floods <- normal(0, 10)
beta_extreme_precip_floods <- normal(0, 10)

# Priors for coefficients linking predictors to storms
beta_temp_storms <- normal(0, 10)
beta_prcp_storms <- normal(0, 10)
beta_extreme_precip_storms <- normal(0, 10)
```

```
# Variance terms for extreme events and disaster outcomes
sigma_emnt <- normal(0, 5, truncation = c(0, Inf))
sigma_emxt <- normal(0, 5, truncation = c(0, Inf))
sigma_emxp <- normal(0, 5, truncation = c(0, Inf))
sigma_floods <- normal(0, 5, truncation = c(0, Inf))
sigma_storms <- normal(0, 5, truncation = c(0, Inf))

# Models for extreme events
emnt_mean <- beta_tavg * tavg + beta_tmin * tmin
distribution(emnt) <- normal(emnt_mean, sigma_emnt)

emxt_mean <- beta_tavg * tavg + beta_tmax * tmax
distribution(emxt) <- normal(emxt_mean, sigma_emxt)

emxp_mean <- beta_prcp * prcp
distribution(emxp) <- normal(emxp_mean, sigma_emxp)

# Model for floods
floods_mean <- beta_temp_floods * (tavg + emnt + emxt) +
  beta_prcp_floods * prcp +
  beta_extreme_precip_floods * emxp
distribution(floods) <- normal(floods_mean, sigma_floods)

# Model for storms
storms_mean <- beta_temp_storms * (tavg + emnt + emxt) +
  beta_prcp_storms * prcp +
  beta_extreme_precip_storms * emxp
distribution(storms) <- normal(storms_mean, sigma_storms)

# Define the model
disaster_model <- model(
  beta_tavg, beta_tmin, beta_tmax, beta_prcp,
  beta_emnt, beta_emxt, beta_emxp,
  beta_temp_floods, beta_prcp_floods, beta_extreme_precip_floods,
  beta_temp_storms, beta_prcp_storms, beta_extreme_precip_storms,
  sigma_emnt, sigma_emxt, sigma_emxp,
  sigma_floods, sigma_storms
)

# Sample the posterior
draws <- mcmc(disaster_model, warmup = 2000, n_samples = 5000, chains = 4)

mcmc_trace(draws)
mcmc_dens(draws)

get.stats(draws)
```

With the following trace:



And density plot:



And results:

variable	mean	median	sd	mad	q5	q95	hdi5	hdi95	rhat	ess_bulk	ess_tail
beta_tavg	-1.16796e+00	-1.21895e+00	1.74076e-01	1.75888e-01	-1.37085e+00	-8.36787e-01	-1.39185e+00	-8.20576e-01	1.74515	6.06796	39.7856
beta_tmin	1.83942e+00	1.91428e+00	2.98068e-01	3.44693e-01	1.28493e+00	2.17949e+00	1.28279e+00	2.20942e+00	1.79716	5.88699	37.1393
beta_tmax	2.24002e+00	2.26735e+00	1.06506e-01	1.09168e-01	2.03560e+00	2.36790e+00	2.02764e+00	2.38642e+00	1.73414	6.09646	40.8489
beta_prcp	1.27824e-01	1.27769e-01	2.70342e-03	2.87461e-03	1.23409e-01	1.32271e-01	1.22780e-01	1.32965e-01	1.05303	92.77072	325.8063
beta_emnt	5.59284e-02	5.85832e-02	3.73772e-02	4.99228e-02	4.56213e-04	1.06995e-01	-3.41122e-04	1.12809e-01	3.36681	4.45250	17.6973
beta_emxt	-3.10695e-02	-3.02588e-02	1.69643e-02	1.60877e-02	-6.70333e-02	-7.11611e-03	-6.81346e-02	-2.90629e-03	1.53168	7.14886	18.7571
beta_emxp	1.23546e-01	1.85838e-01	2.67961e-01	3.12396e-01	-2.97596e-01	5.08999e-01	-3.08794e-01	5.21222e-01	3.59757	4.36652	11.5297
beta_temp_floods	8.80471e-03	8.82780e-03	1.77298e-03	1.82380e-03	5.85280e-03	1.17231e-02	5.38366e-03	1.22510e-02	1.04935	114.42229	264.3864
beta_prcp_floods	-1.55369e-04	-1.54682e-04	6.76772e-05	6.83521e-05	-2.67112e-04	-4.47782e-05	-2.90662e-04	-2.79243e-05	1.01463	366.30278	1366.6892
beta_extreme_precip_floods	4.81224e-03	4.80874e-03	3.50886e-04	3.50201e-04	4.24080e-03	5.39910e-03	4.11742e-03	5.49648e-03	1.00339	2358.83877	6654.7845
beta_temp_storms	-4.94489e-03	-4.97717e-03	3.63511e-03	3.79465e-03	-1.06266e-02	1.03833e-03	-1.19725e-02	2.02748e-03	1.04489	120.76263	225.0092
beta_prcp_storms	-4.85940e-04	-4.85444e-04	1.36031e-04	1.35887e-04	-7.08263e-04	-2.64575e-04	-7.54656e-04	-2.23336e-04	1.02017	285.01496	926.2904
beta_extreme_precip_storms	1.32877e-02	1.32894e-02	6.68258e-04	6.67160e-04	1.21938e-02	1.43885e-02	1.19256e-02	1.45436e-02	1.00458	1120.47783	2246.9003
sigma_emnt	1.16550e+01	1.11490e+01	1.35676e+00	1.02378e+00	1.02329e+01	1.44827e+01	9.94205e+00	1.45658e+01	1.70975	6.17292	43.6094
sigma_emxt	1.01758e+01	1.01739e+01	2.65431e-01	2.69072e-01	9.74359e+00	1.06160e+01	9.64866e+00	1.06714e+01	1.17053	16.25290	167.5812
sigma_emxp	1.05951e+02	1.05956e+02	1.72783e+00	1.73752e+00	1.03145e+02	1.08832e+02	1.02618e+02	1.09378e+02	1.00025	4993.73389	10051.3785
sigma_floods	1.38715e+00	1.38757e+00	3.17830e-02	2.96981e-02	1.33327e+00	1.44017e+00	1.32307e+00	1.44843e+00	1.04793	57.40238	119.2452
sigma_storms	2.66624e+00	2.66739e+00	5.77422e-02	5.71188e-02	2.57076e+00	2.75977e+00	2.54355e+00	2.77379e+00	1.05259	129.12769	213.2052

The Poisson model is more aligned with the nature of disaster data, as floods and storms are discrete events that cannot be negative or fractional. The inclusion of a log-link transformation ensures that the rate parameter is positive, making it more interpretable in terms of multiplicative relationships. However, this approach introduces computational complexity, especially for hierarchical models.

The second model is not realist as it assumes disaster occurrences are continuous, but it is simpler to implement and slightly more computationally efficient. It may not accurately capture the discrete nature of the data, particularly for countries with low disaster counts, but it may be useful for future analysis with more levels if migration was added.

Conclusions

Throughout the process, several setbacks were encountered, primarily concerning data availability and model complexity. The initial lack of suitable migration data compelled a focus shift solely to the relationship between climate change and natural disasters. Additionally, the high complexity of the chosen models led to computational difficulties and extended processing times, impacting my project timeline.

The final models demonstrated that some relationship does exist between climate variables and the incidence of natural disasters. None of the models, however, were significant enough. More design and testing is necessary, and will be worked on in the future. Overall, these results were promising, even if the absence of migration data in the analysis was a notable limitation. The findings provide a basis for understanding how climate extremes can be modeled as an influencer of disaster occurrences.

References

Centre for Research on the Epidemiology of Disasters (CRED). EM-DAT: The International Disaster Database. Retrieved from https://www.emdat.be

National Oceanic and Atmospheric Administration (NOAA). Global Surface Summary of the Month (GSOM). NOAA National Centers for Environmental Information. Retrieved from https://www.ncei.noaa.gov

United Nations Department of Economic and Social Affairs (UNDESA). International Migration Database. Retrieved from https://www.un.org/development/desa/pd/content/international-migration-database

Beck, H. E., Zimmermann, N. E., McVicar, T. R., Vergopolan, N., Berg, A., & Wood, E. F. (2018). Present and future Köppen-Geiger climate classification maps at 1-km resolution. Scientific Data, 5(1), 180214. Retrieved from https://doi.org/10.1038/sdata.2018.214

World Bank Climate Change Knowledge Portal (CCKP). Retrieved from https://climateknowledgeportal.worldbank.org/

NOAA Global Historical Climatology Network (GHCN). Retrieved from https://www.ncei.noaa.gov/products/land-based-station/global-historical-climatology-network-monthly

ECA&D. European Climate Assessment & Dataset. Retrieved from https://www.ecad.eu/

Climate TRACE. Retrieved from https://www.climatetrace.org/