Air Emissions and Disaster Displacement Data Wrangling &Analysis

Compiled by Jia Gu

Project Summary

This project explored the relationship between carbon monoxide (CO) emissions and disaster-induced internal displacement by integrating two complementary datasets: a CSV file from the OECD on national pollutant emissions and a JSON dataset from the IDMC API detailing displacement events. The data were cleaned, standardised, and merged at the country-year level to enable comparative analysis.

The findings suggest that while some countries, such as the United States, exhibit both high CO emissions and high levels of disaster-related displacement, others with significant emissions, including Kazakhstan and Canada, report comparatively lower displacement figures. This points to differing regional experiences of climate-related impacts, potentially influenced by geographic, socioeconomic, and environmental factors.

Further analysis of hazard-specific displacement showed notable variation in the types of disasters affecting different countries. For instance, wildfires and storms were predominant in the United States, storms in Japan, and floods in parts of Europe and Central Asia. These patterns reflect how exposure to hazards is shaped by local climate and topography.

An examination of trends over time in the United States showed a gradual decline in CO emissions, while displacement events continued to occur frequently. This may indicate progress in emissions management alongside ongoing exposure to climate-related risks, reinforcing the importance of resilience planning.

To build on this work, future analysis could incorporate variables such as population size, GDP, and climate risk indices to support normalised comparisons across countries. Adding geospatial detail would allow for sub-national insights, and integrating qualitative or case-based evidence could further enhance the depth and relevance of the findings.

Wrangling Details - OECD Air Emissions Dataset

Origin & Description: The air emissions data were sourced from the OECD Environment Statistics (Air Pollutants Inventories) database. It contains annual national emissions of major air pollutants, including carbon monoxide (CO), nitrogen oxides, particulate matter (PM10, PM2.5), etc., for OECD member countries and partners. The raw dataset was provided as a CSV file with 180,051 rows and 36 columns.

Structure & Audit: The original dataset was structured such that each row represented a recorded observation of a specific pollutant for a given country and year. Key fields included the country code and name, pollutant type, measure type, unit of measurement, year, and the associated emission value. The dataset contained a large number of columns, many of which were either entirely null or held a single repeated value across all rows. Multiple units of measurement were present, including tonnes, kilograms per capita, kilograms per unit of GDP, and index values. Additionally, the dataset included summary entries for regional aggregates, such as "OECD Total" and "OECD Europe", which do not represent individual countries.

Cleaning Steps: The cleaning process focused on retaining only the most relevant data to support merging and subsequent analysis, with a particular emphasis on carbon monoxide (CO), one of the most widely recognised air pollutants.

1. A threshold was set to drop any columns with more than 60% missing values, as such columns provided little analytical value and would hinder efficient processing.
2. To narrow the scope, all non-CO pollutants (e.g., sulphur oxides, nitrogen oxides, volatile organic compounds) were filtered out. Within the CO subset, further filtering was applied to retain only records representing total man-made emissions measured in tonnes, which reflects the absolute emission volumes. This involved selecting entries where the pollutant was “Carbon monoxide” and the unit of measure was “Tonnes”, while excluding entries measured in intensity units such as “Kilogrammes per 1,000 US dollars” or index-based values.
3. During inspection, it was noted that each country and year combination could have multiple emission entries by source category (e.g., “Total mobile sources”, “Total stationary sources”). These sub-categories were aggregated to compute the total CO emissions per country per year by summing across all sources.
4. The emissions values were also associated with a unit multiplier—commonly “Thousands”—indicating the values are expressed in thousands of tonnes. For instance, an entry of 4,825.445 with a “Thousands” multiplier corresponds to 4,825,445 tonnes of CO emissions. To account for this, all values were interpreted accordingly.
5. To maintain data integrity and avoid redundancy, I manually reviewed column values using .value_counts() and removed columns that either contained a single repeated value or duplicated information from other columns. Additionally, metadata fields such as STRUCTURE, ACTION, and other structural identifiers were dropped as they provided no analytical value.

Wrangling Details - IDMC Disaster Displacement Dataset

Origin and Description: The disaster displacement data was sourced from the Internal Displacement Monitoring Centre's Global Internal Displacement Database (GIDD). This database compiles detailed reports of internal displacements caused by both conflict and natural disasters. For the purposes of this project, only disaster-related displacement records were used, covering events from 2008 onward. The dataset was retrieved through an API export in JSON format, resulting in 13,198 disaster-related records as of May 2025. Each entry represents a specific disaster event that triggered new internal displacement in a given country and year.

Structure and Initial Audit: Each record includes information such as the ISO3 country code, country name, year, event description, hazard category and type, and the number of people displaced. The key variable of interest is new_displacement, which quantifies the number of people newly displaced by a specific disaster event. Although the dataset also includes a total_displacement field, this was generally null for disaster events (being more relevant to conflict situations). As part of the cleaning process, this column was dropped due to exceeding the 60% null value threshold.

Minor events reporting very low or zero displacement values were retained, as they do not significantly distort aggregate statistics and may reflect genuine cases with minimal impact.

The dataset includes a structured hazard taxonomy that distinguishes between geophysical disasters and weather-related disasters. Since weather-related disasters are more relevant to air emissions and climate change, the data is filtered to only weather-related disasters. Redundant or low-information fields were removed. For instance, columns like hazard_type and hazard_sub_type, which provided only integer codes, were dropped in favour of more descriptive fields like hazard_type_name.

Cleaning Steps:

1. After loading the JSON into a DataFrame, only the essential fields were retained: ISO3 country code, country name, year, hazard category, hazard type, and new_displacement.
2. The new_displacement field was explicitly converted to an integer type to support accurate aggregation. Country names in the disaster dataset largely aligned with those in the emissions dataset, enabling a smooth merge. In cases of naming discrepancies, such as "Turkey" vs "Türkiye", alignment was achieved through matching ISO3 codes, ensuring consistent country identifiers across both datasets.

Wrangling Details - Merge

Merge Keys: 'ISO3', 'year'

Process:

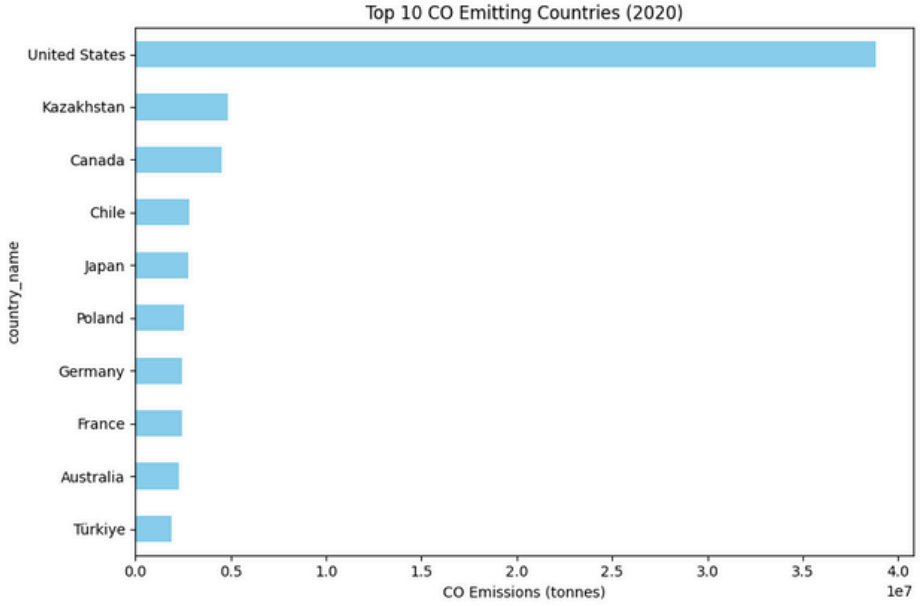
- The filtered emissions data had entries from 1990 to 2022 for 60 countries. The displacement data had entries from 2008 to 2022 for 205 countries.
- The two datasets were merged using the ISO3 country code and year as key identifiers. Matching on ISO3 codes ensured accurate alignment between corresponding countries, while aligning on the year allowed for valid comparisons between emissions and displacement data within the same time period.
- An inner join was used to retain only those country-year combinations that appeared in both datasets. As a result, the merged dataset includes primarily OECD countries and a few additional nations and covers the years 2008 to 2022, aligning with the availability of disaster displacement data.

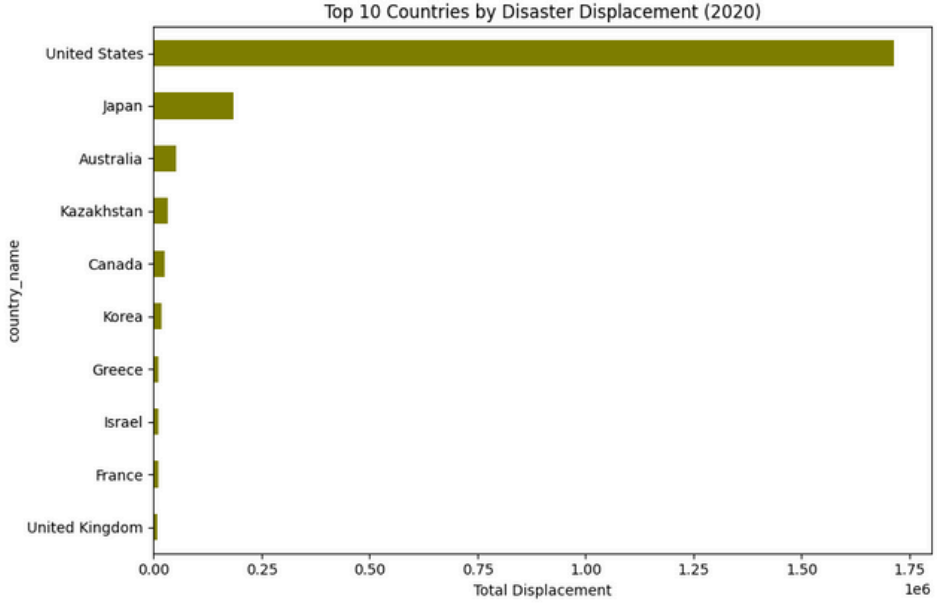
Post-Merge Structure: The final merged dataset, saved as merged_emissions_disasters.csv, includes one or more rows per country-year. The data is maintained at the event level to preserve granularity by hazard type. This means that countries with multiple disaster events in a single year appear multiple times, each row representing a distinct event but sharing the same emission value for that year. For example, the Netherlands appears twice in 2020 due to two recorded wildfire events, each row showing the same 2020 CO emission figure but different event-specific information.

Included fields are: country_name, iso3, year, CO_emissions_tonnes, and event-specific attributes such as event_name, new_displacement, and hazard_type_name. Keeping the dataset in this format allows flexible analysis, enabling aggregation by hazard type, year, or country while maintaining access to detailed emissions data.

Storage and Access:
The combined dataset was stored in CSV format to allow for easy access, analysis, and sharing.
https://drive.google.com/file/d/1deRRYChTK8F79YK2PK-rY5LVetDRFmm0/view?usp=drive_link
This final data store serves as the basis for the analysis and visualisations in the next section.

Are the countries with the highest CO emissions also the most affected by disaster-induced displacement?





To examine this question, the dataset was filtered to focus on the year 2020:

```
df_2020 = merged_df[merged_df['year'] == 2020]
```

The top 10 countries with the highest carbon monoxide emissions were identified using the .groupby() and .nlargest() methods:

```
top10_emitters = df_2020.groupby('country_name')
['CO_emissions_tonnes'].first().nlargest(10)
```

In parallel, the countries with the highest levels of disaster displacement were determined:

```
top10_displacement = df_2020.groupby('country_name')
['displacement'].sum().nlargest(10)
```

The results were visualised using horizontal bar plots, allowing for a direct side-by-side comparison:

```
fig, axes = plt.subplots(1, 2, figsize=(18, 6))
top10_emitters.sort_values().plot(kind='barh', ax=axes[0], color='skyblue')
axes[0].set_title('Top 10 CO Emitting Countries (2020)')
axes[0].set_xlabel('CO Emissions (tonnes)')
top10_displacement.sort_values().plot(kind='barh', ax=axes[1], color='olive')
axes[1].set_title('Top 10 Countries by Disaster Displacement (2020)')
axes[1].set_xlabel('Total Displacement')
plt.tight_layout()
plt.show()
```

Key Insight:

The United States appears in the top 10 for both CO emissions and disaster-induced displacement, but most other top CO emitters—such as Kazakhstan and Canada—do not experience high levels of displacement. This highlights an imbalance where industrial emissions and climate vulnerability do not always overlap geographically.

What types of hazards drive displacement in the most affected countries?

Filter the top 10 countries by displacement:
```
top10_countries = top10_displacement.index
hazard_df = df_2020[df_2020['country_name'].isin(top10_countries)]
```

Group by country and hazard type, summing total displacement:
```
hazard_grouped = hazard_df.groupby(['country_name', 'hazard_type_name'])
['displacement'].sum().reset_index()
```

Pivot for heatmap: countries (rows) vs hazard types (columns):
```
heatmap_data = hazard_grouped.pivot(index='country_name',
columns='hazard_type_name', values='displacement').fillna(0)
```

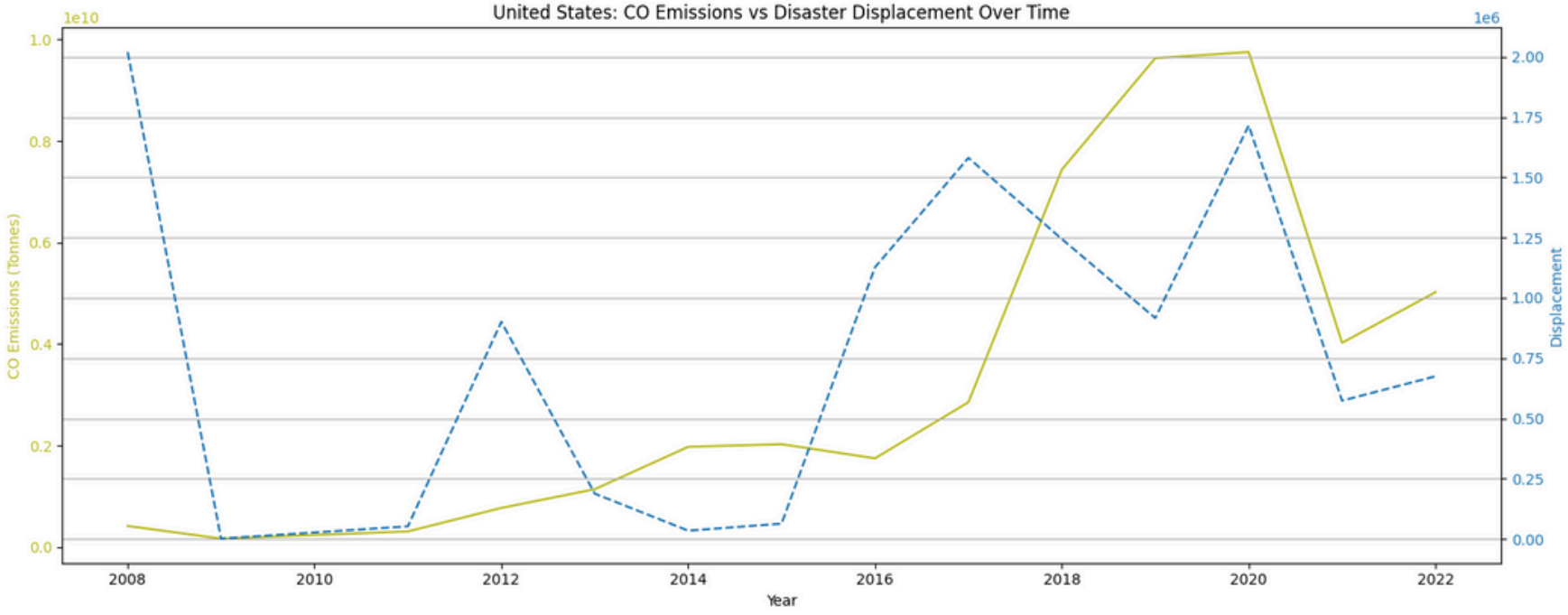
Plot the heatmap to visualise displacement distribution by hazard type for each country:
```
plt.figure(figsize=(8, 8))
sns.heatmap(heatmap_data, annot=True, fmt=".0f", cmap="GnBu",
linewidths=0.5)
plt.title("Displacement by Hazard Type (Top 10 Countries by Total Displacement,
2020)")
plt.ylabel("Country")
plt.xlabel("Hazard Type")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

Key Insight:

Wildfires and storms are the primary drivers of displacement in the United States, which stands out as a clear outlier with significantly higher displacement figures than any other country in the top 10. Japan's displacement is largely driven by storms, while countries like Kazakhstan and France are more affected by floods. These variations reflect how geographic and climatic factors shape each country's disaster profile.

Given the scale of U.S. displacement, it may be useful to analyse trends both including and excluding the U.S. to better observe patterns across the rest of the world. Alternatively, normalising by population or land area can provide more equitable comparisons.

How have emissions and displacement trends evolved over time in the country with the highest total displacement?



The country with the highest cumulative displacement over the available time frame (2008–2022) was identified as follows:
highest_disp_country = merged_df.groupby("country_name")["displacement"].sum().idxmax()

Filter and sort data for that country:
country_df = merged_df[merged_df["country_name"] == highest_disp_country].sort_values("year")

Group yearly totals for emissions and displacement:
trend_df = country_df.groupby("year")[["CO_emissions_tonnes", "displacement"]].sum().reset_index()

A dual-axis line chart was produced to visualise changes in both CO emissions and disaster displacement over time.

Key Insight:
While CO emissions in the U.S. generally increased from 2015 to 2019, disaster displacement spiked sharply in 2020, largely due to wildfire and storm events. However, the trends do not show a direct linear relationship, suggesting other factors like event severity and preparedness play a role.

Appendix

Original Data source:

1. https://data-explorer.oecd.org/vis?df[ds]=DisseminateFinalDMZ&df[id]=DSD_AIR_EMISSIONS%40DF_AIR_EMISSIONS&df[ag]=OECD.ENV.EPI&dq=.A.SOX.T_EM_MM.T&pd=1990%2C2022&to[TIME_PERIOD]=false&vw=ov
2. https://www.internal-displacement.org/database/api-documentation/ [Requires API access from IDMC]