**Replication Data for The Impact of Climate Change on the Economic Costs of Extreme Weather**

*Introduction*

This research is structured using Python projects developed in Visual Studio Code. The Replication Data is organised into dedicated directories for data, code, and results. The “*Data*” directory contains all datasets used in the study. Within this directory, the “*All Data”* subfolder contains the EM-DAT dataset, which provides comprehensive records of global disasters, including event types, economic costs, and human impacts. The “*GDP\_per\_country”* subfolder includes nominal GDP data for all countries.

The “*Code*” directory encompasses the analytical workflows and scripts. The directory contains the following subfolders:

* *Descriptive Analysis*
* *Statistical Analysis*
* *Results*

The “*Descriptive Analysis”* subfolder focuses on exploratory data analysis, with scripts dedicated to summarizing and visualizing variables such as disaster-related costs, disaster frequency, numerical trends over time, and other relevant metrics. This section provides foundational insights into data distributions, temporal patterns, and preliminary correlations.

The “*Statistical Analysis”* subfolder is divided into two components: “*Models Analysis”* and “*Research”*. The “*Models Analysis”* subfolder contains implementations of various forecasting and statistical models, including ARIMA, GARCH, ETS, VAR, XGBoost, Prophet, Dynamic Linear Models, CAT models, and a Long-term trend model. These scripts are designed to predict disaster-related economic costs.

The “*Research”* subfolder addresses sub questions, such as the economic costs of severe weather events and regional or sectoral vulnerabilities to climate risks.

Finally, the “*Results*” directory consolidates the quantitative outcomes of the models. The results include the following error metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Maximum Likelihood Estimates (MLE). The dictionary structure ensures reproducibility and transparency. It enables users to navigate from raw data processing to final analytical conclusions.

*Data source*

The Emergency Events Database (EM-DAT) (Delforge et al., 2025) will be used to evaluate the models. The database provides a central empirical source for assessing the economic impacts of extreme weather worldwide. EM-DAT is maintained by the Center for Research on the Epidemiology of Disasters and contains information on various cost components. The cost components include reconstruction costs, insured damages and total damages. Each cost is reported in thousands of US dollars and is both nominal and adjusted for inflation. Values for “Insured Damages” are usually provided by large reinsurance companies (e.g. Munich Re, Swiss Re), which for consistency also often provide corresponding estimates for “Total Damages”. Despite the extensive scope, EM-DAT's economic figures remain subject to underreporting; losses in countries with limited insurance coverage are often not recorded or are only partially recorded, resulting in significant potential bias.

The EM-DAT data is collected from the following website:

<https://public.emdat.be/data>

To recreate the exact dataset, the classification was set to 'Natural'. The selected regions were Africa, America, Asia, Europe, and Oceania. The time period was set from 1900 to 2025.

References:

Delforge, D., Wathelet, V., Below, R., Sofia, C. L., Tonnelier, M., Van Loenhout, J. A., & Speybroeck, N. (2025). EM-DAT: the Emergency Events Database. *International Journal of Disaster Risk Reduction*, 105509. https://doi.org/10.1016/j.ijdrr.2025.105509

*EM-DAT columns with description*

* DisNo: This column contains a unique identifier for each disaster occasion, formatted to encompass the 12 months, occasion number, and ISO Code.
* Historic: This column indicates whether or not the catastrophe is assessed as an historical event.
* Classification Key: This column presents a specific key describing the class of the catastrophe.
* Disaster Group: This column specifies the overall category of the disaster, such as "Natural."
* Disaster Subgroup: This column offers a more precise categorization in the catastrophe organization, which include "Hydrological" or "Meteorological."
* Disaster Type: This column defines the form of disaster, e.g., "Flood" or "Storm."
* Disaster Subtype: This column offers similarly element at the kind of catastrophe, inclusive of "Coastal Flood" or "Severe Weather."
* External IDs: This column consists of any outside identifiers connected to the disaster occasion, such as move-references to different datasets.
* Event Name: This column gives the name of the catastrophe event, if to be had.
* ISO: This column incorporates the ISO 3166-1 alpha-3 code representing the country wherein the disaster befell.
* Reconstruction Costs ('000 US$): This column lists the prices related to reconstruction efforts, expressed in thousands of US dollars.
* Reconstruction Costs, Adjusted ('000 US$): This column adjusts the reconstruction costs to account for inflation or different monetary elements, expressed in thousands of US dollars.
* Insured Damage ('000 US$): This column shows the insured damages due to the disaster, expressed in thousands of US dollars.
* Insured Damage, Adjusted ('000 US$): This column adjusts the insured damage charges
* Total Damage ('000 US$): This column represents the entire anticipated economic price of the catastrophe, expressed in thousands of US dollars.
* Total Damage, Adjusted ('000 US$): This column adjusts the entire damage costs for inflation or other factors, expressed in thousands of US dollars.
* CPI: This column lists the Consumer Price Index (CPI) cost used to modify monetary figures to account for inflation.
* Admin Units: This column info administrative devices or regions inside the united states in which the disaster befell, if available.
* Entry Date: This column suggests the date while the disaster event was first entered into the database.
* Last Update: This column gives the most current date the document was up to date.

*Data Cleaning and Preprocessing*

The first step became to load the dataset into Python as a dataframe the use of the Pandas library. This allowed for efficient manipulation and exploration of the data. The raw dataset was examined for missing values in key columns, consisting of total damage, and country codes. The following columns have been selected to answer the research question: 'DisNo.', 'Disaster Type', 'Disaster Subtype', 'ISO', 'Country', "Reconstruction Costs ('000 US$)", "Reconstruction Costs, Adjusted ('000 US$)", "Insured Damage ('000 US$)", "Insured Damage, Adjusted ('000 US$)", "Total Damage ('000 US$)", "Total Damage, Adjusted ('000 US$)".

Missing values inside the "Total Damage ('000 US$)" column, that are crucial for analysis, have been dropped to keep away from inaccuracies within the effects. For non-essential columns, missing data was imputed using median values or context-specific rules to maintain consistency while minimizing data loss.

Next, replica rows had been diagnosed and removed to prevent redundancy. As the dataset included numerous disaster events over numerous a long time, entries with overlapping identifiers (e.g., the same disaster kind and year for the same country) had been verified to make certain they did now not represent data duplication or mistakes.

To make certain consistency, the "Year" column become extracted from the particular catastrophe identifier (DisNo) and converted into integer format. This allowed the dataset to be grouped and aggregated by means of year for time-collection evaluation. Additionally, the monetary damage columns, which includes "Total Damage ('000 US$)" and "Reconstruction Costs ('000 US$)," have been converted right into a uniform scale, expressed in millions of US dollars. This step ensured that each one monetary records was comparable and interpretable.

The dataset turned into additionally enriched by using developing extra features to aid in evaluation. For example, inflation-adjusted damage figures were utilized to account for temporal monetary modifications, making sure that monetary values reflected present day requirements. Furthermore, disaster subtypes were grouped into broader categories to lessen dimensionality, enhancing the performance of the analysis.

Finally, the cleaned dataset was exported right into a dependent format for modelling. The time-series information changed into prepared data for modelling (e.g. ARIMA). This included setting the "Year" column as the index to facilitate temporal forecasting. Consistency checks were performed throughout the process to validate the integrity of the data and ensure alignment with the original EM-DAT documentation. By addressing missing values and ensuring consistency across columns, the data cleaning and preprocessing phase established a strong foundation for reliable and accurate analysis. This process minimized biases and errors (Morr et al., 2022), enabling the effective application of statistical models to uncover insights into the economic impacts of extreme weather events.