**Project Overview**

Climate change and extreme weather events are on the rise, making it critical to understand how growing CO₂ emissions may be linked to natural disaster frequency and severity. This project aims to analyze **50 years of historical disaster data** alongside **CO₂ emissions** and other climate indicators to **forecast** the likelihood and intensity of natural disasters over the **next 50 years**.

**Stakeholder Perspective**

1. **Policy Makers & Governments**: Need accurate long-term forecasts to allocate resources, strengthen infrastructure, and enact preventive measures.
2. **Insurance & Financial Institutions**: Risk assessment models based on historical data and projected future scenarios help set coverage policies and prepare for large-scale claims.
3. **Researchers & NGOs**: Insights into climate change’s role in driving disaster trends inform advocacy and data-driven planning.

**Data Sources**

**1. Historical Weather & Disaster Data**

* **NOAA Storm Events Database**: Offers detailed records of storms, tornadoes, hurricanes, floods, etc. in the United States.
* **EM-DAT (Emergency Events Database)**: A global repository of natural disasters, including severity metrics like economic losses and mortality rates.

**2. CO₂ Emissions Data**

* **EDGAR (Emissions Database for Global Atmospheric Research)**: Provides global emissions data by country and sector.
* **Our World in Data** (CO₂ dataset): Offers a user-friendly interface with historical CO₂ emission records per country.
* **NOAA ESRL or NASA GISS**: Tracks atmospheric CO₂ concentration levels, which can be used in place of or alongside reported emissions data.

**3. Additional Climate Indicators (Optional)**

* **NASA Earth Observing System**: Temperature anomalies, sea surface temperatures, and other long-term climate trends.
* **IPCC Reports**: Summaries of projected climate scenarios (RCP/SSP pathways) that could be used to model future emission or warming scenarios.

**Data Preparation & Feature Engineering**

1. **Data Merging**:
   * Align weather event records with CO₂ data on a yearly or monthly basis (depending on available granularity).
   * Aggregate or interpolate data if sources have different timescales or location references.
2. **Feature Engineering**:
   * **Lag/Shift**: Incorporate lag features to capture delayed impacts of rising emissions on climate (e.g., 5-year lag of CO₂ levels to see delayed disaster effects).
   * **Seasonality Indicators**: For hurricanes or monsoon seasons, add monthly or seasonal markers.
   * **Geospatial Attributes**: If focusing on specific regions, include latitude/longitude or region codes to capture local trends.

**Modeling Approaches**

Since the goal is to **predict both frequency and severity** of future disasters, you may need multiple modeling components or a hybrid approach:

1. **Time Series Forecasting**
   * **ARIMA or SARIMA**: Classic statistical methods for long-term trend forecasting. Good for establishing baseline predictions of event frequency.
   * **Facebook Prophet**: Quick and interpretable time series approach, handles seasonality and trend changes reasonably well.
2. **Machine Learning Regression or Classification**
   * **Random Forest / XGBoost**: Useful for predicting severity (e.g., economic damage or mortality count) by learning patterns from multiple climate and emissions features.
   * **Neural Networks (LSTM, RNN)**: Capture complex temporal dependencies in sequence data. More advanced and flexible for multi-step, long-range forecasting.
3. **Ensemble / Hybrid Methods**
   * Combine a **time series** model for frequency with a **regression** model for severity. For instance, use ARIMA to predict the *number of disasters* in a future year, then feed that into an XGBoost regression to estimate *average damages* or *severity score*.

**Model Training & Evaluation**

1. **Train/Test Split**: Typically, **train** on the first 40 years of data and **validate** on the last 10 years to simulate out-of-sample predictions.
2. **Cross-Validation**: For time series, use a **rolling window** or expanding window approach to validate performance at each fold.
3. **Metrics**:
   * **Frequency Prediction**: Mean Absolute Error (MAE), RMSE, or MAPE for time series counts.
   * **Severity Prediction**: MAE, RMSE for continuous damage values; or precision/recall/F1 if classifying disasters by severity tiers.
4. **Regular Retraining**: Update models periodically with new data to maintain accuracy, especially as emission rates or global climate factors shift.

**Forecasting the Next 50 Years**

* **Project Scenario Runs**:
  + **Scenario A**: Emissions remain at current growth rates (business-as-usual).
  + **Scenario B**: Emissions rise at an accelerated rate.
  + **Scenario C**: Emissions significantly drop due to policy interventions.
* By applying each emissions scenario, you can generate various *what-if* forecasts, demonstrating how policy decisions might influence future disaster risks.

**Deliverables & Applications**

1. **Interactive Dashboard**: A web-based interface that shows historical disaster trends alongside CO₂ emissions, plus forecasted disaster frequency/severity under different emission scenarios.
2. **Technical Whitepaper**: Summarizes data collection, modeling approach, findings, and policy recommendations.
3. **Media Outreach**: Short reports or infographics illustrating the link between CO₂ levels and rising natural disaster frequency.

**Conclusion**

By tying together **50 years of historical disaster data** with **CO₂ emissions** and **climate indicators**, this project will produce valuable insights for governments, insurance companies, and environmental organizations. Advanced modeling techniques—ranging from statistical time series approaches to machine learning ensembles—enable robust forecasts of disaster frequency and severity over the **next five decades**. The ultimate goal is to inform proactive planning, resource allocation, and emissions policies to mitigate the most severe impacts of climate change.

**Damage Cost Estimation**

The DAMAGE\_PROPERTY column contains values like "250K" or "25M", which need to be converted to **numerical values** for proper filtering.

* "K" = **thousands (x1,000)**
* "M" = **millions (x1,000,000)**
* "B" = \*\*billions (x1,000,000,000)`, though likely not present here.

For **insured vs. real costs**, estimates vary:

* Insured losses often cover **40-60%** of total damage.
* A reasonable assumption: **Real cost ≈ 2x insured losses**.
* Using this, we might filter for **reported damage ≥ $12.5 million** to approximate the $25M threshold.
  + March Madness Bracket Predictions
  + <https://www.geeksforgeeks.org/detect-and-recognize-car-license-plate-from-a-video-in-real-time/>
  + <https://www.geeksforgeeks.org/sentiment-analysis-with-an-recurrent-neural-networks-rnn/>
  + Machine Learning Project Recommendation and outline maker
  + <https://www.geeksforgeeks.org/sms-spam-detection-using-tensorflow-in-python/>