Multimodal Cyclone DataSet

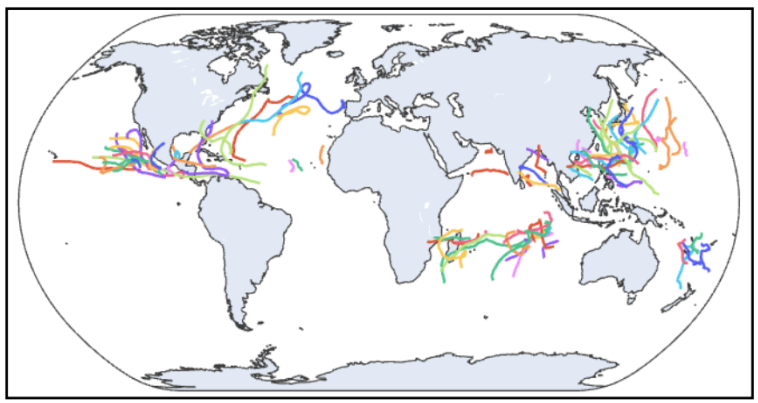
Louis Carron, Rouillé Nathan, Khosrowshahi Sacha, Court Chloé

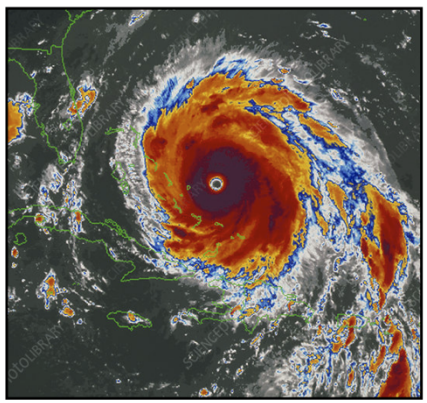
IASD - Université PSL - Paris Dauphine

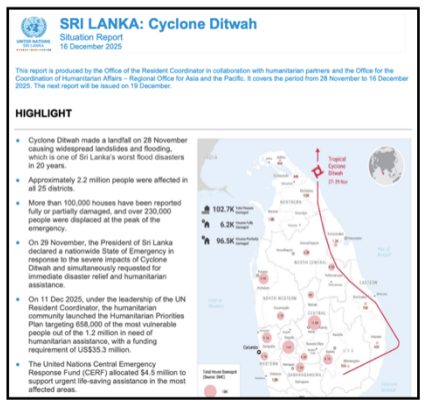
Academic Year 2025-2026

Introduction

Tropical cyclones generate **heterogeneous** data.







Goal : A **multimodal dataset** gathering **physical measurements, satellite images and text reports**.

IBTrACS Dataset

Anchor Point : International Best Track Archive for Climate Stewardship (IBTrACS) Dataset

For each cyclone, observations about its trajectory :

Identifiers	Storm ID (unique), name
Time	timestamp
Location	basin, latitude, longitude

- Direct download format `NetCDF`, from 2022-2025

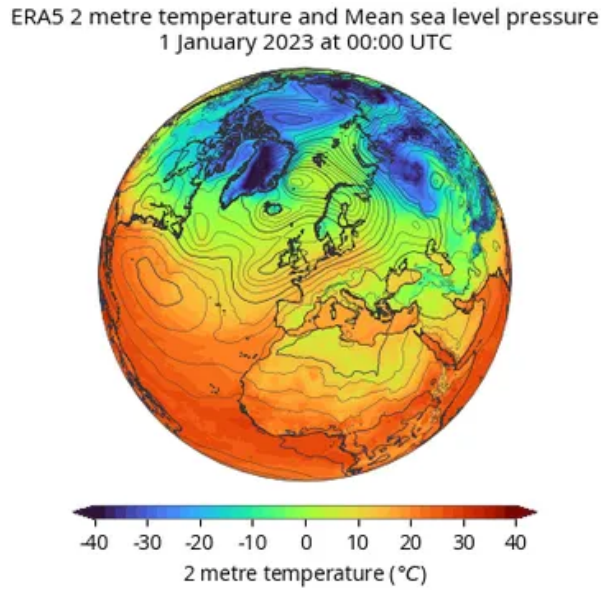
Outline

ERA5 Meteorological Data

ERA5 : a meteorological world **grid**.

API user selection :

- Variables (Temperature, Pressure, Wind Speed)
- Time (years, months, days, hours)
- Area (bounding box)



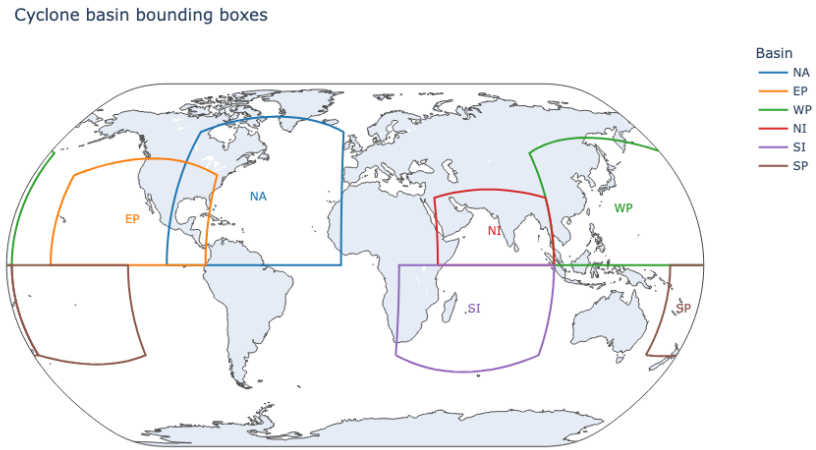
Constraints on each API request:

- Download size limit
- Delay ∼30 s (submission, validation)

ERA5 Downloading & Sampling

Key observations:

- IBTrACS record every **3 hours**.
- Cyclones are **seasonal**.
- A basin is a **bounding box**.



1. **Downloading algorithm:**
   - Loop over years and basins
   - Call the API on the most active months, all days, from 12AM:9PM:3h.
2. **Interpolation:** Spatio-temporal IBTrACS / ERA5 to the nearest
3. **Storage format:** `.csv`

GDELT Overview

- **GDELT** (Global Database of Events, Language and Tone) monitors global news in near real-time.
- Two complementary datasets:
  - **Mentions**: individual news articles mentioning events
  - **GKG**: structured metadata (locations, actors, organizations)
- Data updated every 15 minutes
- We process mentions every **3 hours** and aggregate results **by day**
- Objective: link media coverage to observed cyclone activity

Cyclone Mentions Extraction Pipeline

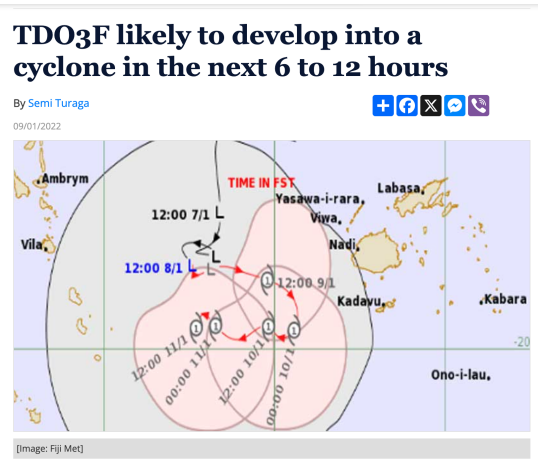
1. **File selection**
   - Retrieve GDELT index
   - Select mentions files for cyclone dates (3-hour intervals)
2. **Text-based filtering**
   - Cyclone name and storm-related keywords
   - Basin-specific geographic keywords
   - URL deduplication to avoid over-counting
3. **Geographic validation (GKG)**
   - If basin unknown, keep articles with locations inside cyclone bounding box

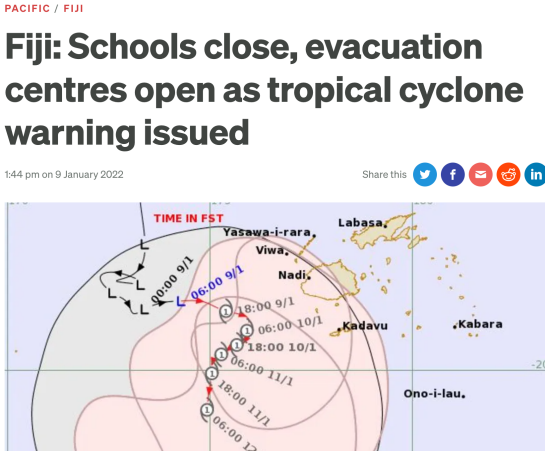
Media Timeline Around Cyclone Cody

- Cyclone: **Cody**
- Region: **South Pacific (Fiji)**
- Media coverage over time:

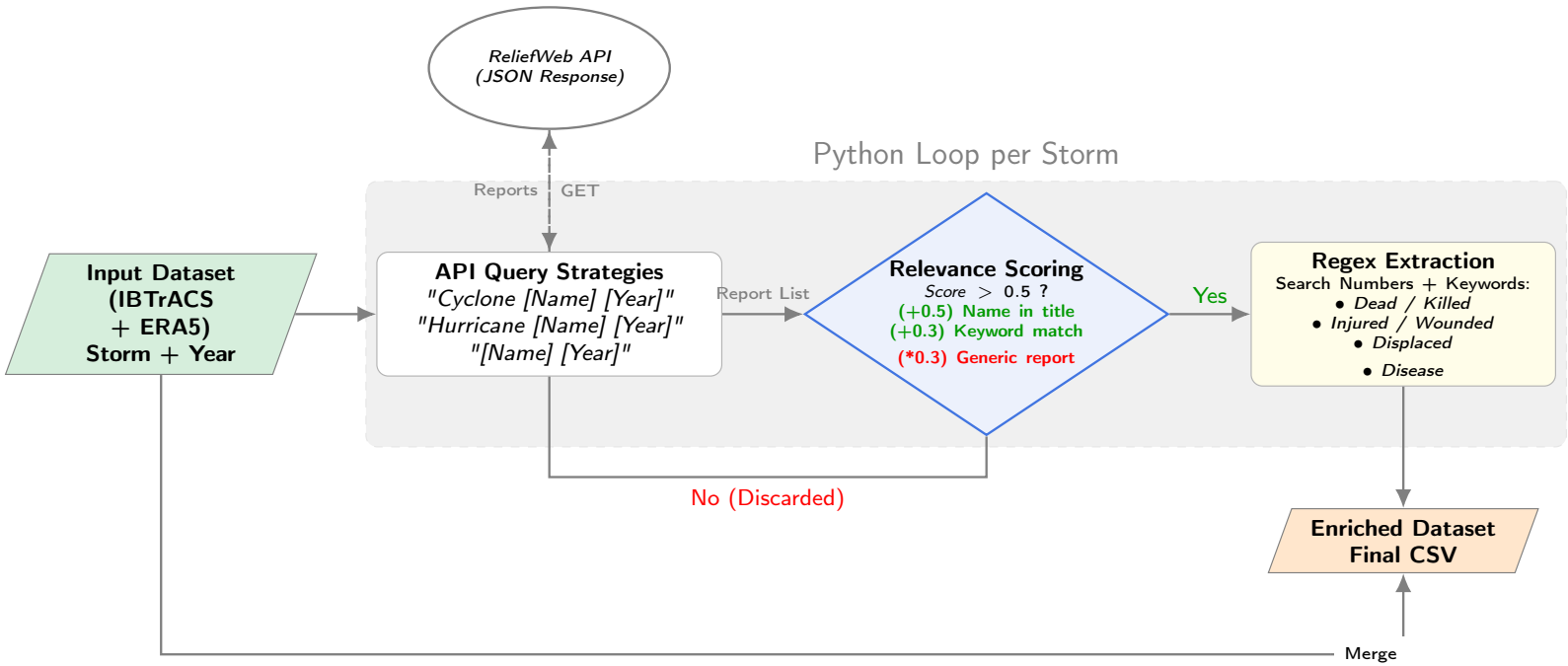
Timeline of Articles

- **Jan 8, 2022 – Forecast**
  - fijivillage.com
- **Jan 9, 2022 – Warning and Preparedness**
  - rnz.co.nz
- **Jan 10, 2022 – Aftermath and Impacts**
  - rnz.co.nz

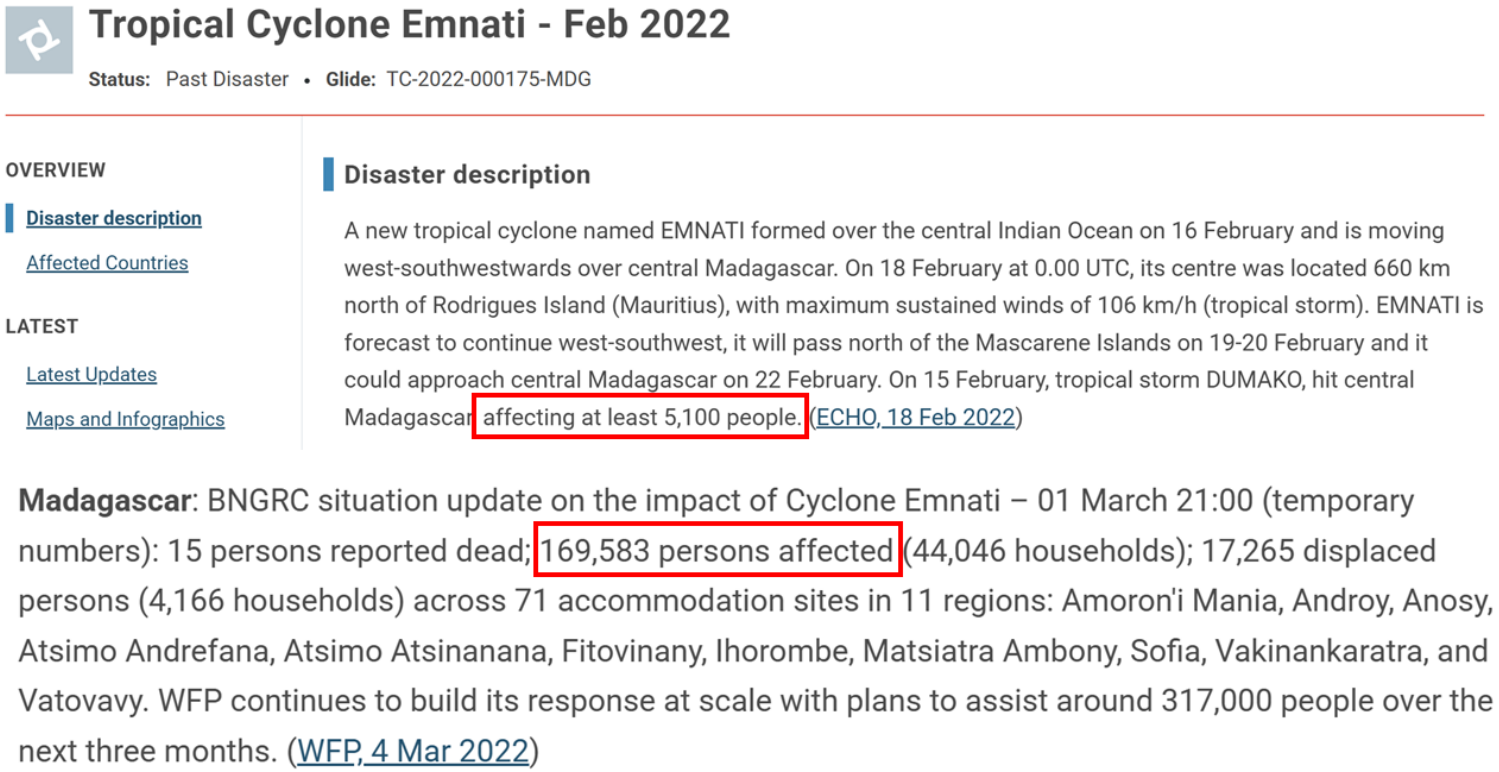




Data Enrichment Pipeline (ReliefWeb)



Limits of RegEx



Tropical Cyclone Emnati - Feb 2022

Status: Past Disaster • **Glide:** TC-2022-000175-MDG

OVERVIEW

Disaster description

Affected Countries

LATEST

Latest Updates

Maps and Infographics

Disaster description

A new tropical cyclone named EMNATI formed over the central Indian Ocean on 16 February and is moving west-southwestwards over central Madagascar. On 18 February at 0.00 UTC, its centre was located 660 km north of Rodrigues Island (Mauritius), with maximum sustained winds of 106 km/h (tropical storm). EMNATI is forecast to continue west-southwest, it will pass north of the Mascarene Islands on 19-20 February and it could approach central Madagascar on 22 February. On 15 February, tropical storm DUMAKO, hit central Madagascar affecting at least 5,100 people. (ECHO, 18 Feb 2022)

Madagascar: BNGRC situation update on the impact of Cyclone Emnati – 01 March 21:00 (temporary numbers): 15 persons reported dead; 169,583 persons affected (44,046 households); 17,265 displaced persons (4,166 households) across 71 accommodation sites in 11 regions: Amoron'i Mania, Androy, Anosy, Atsimo Andrefana, Atsimo Atsinanana, Fitovinany, Ihorombe, Matsiatra Ambony, Sofia, Vakinankaratra, and Vatovavy. WFP continues to build its response at scale with plans to assist around 317,000 people over the next three months. (WFP, 4 Mar 2022)

Figure: Visualizing the boundaries of Regex capabilities

TC-PRIMED: Selection and Multi-Sensor Matching

- **Why TC-PRIMED?**: Global v01r01 "final" dataset providing standardized, storm-centered Infrared (IR) and Passive Microwave (GPROF) imagery.
- **Extraction Strategy**: Used regex to parse timestamps from S3 filenames for instant metadata discovery without network overhead.
- **Temporal Matching**: Aligned satellite overpasses to IBTrACS/ERA5 best-track records using a **1.5-hour tolerance** window.
- **Product Suite**: Extracted IR brightness temperatures alongside 5 GPROF products (Surface/Convective Precipitation, Rain/Cloud/Ice Water Paths) for a complete physical/visual storm profile.

Standardization and Physical Context

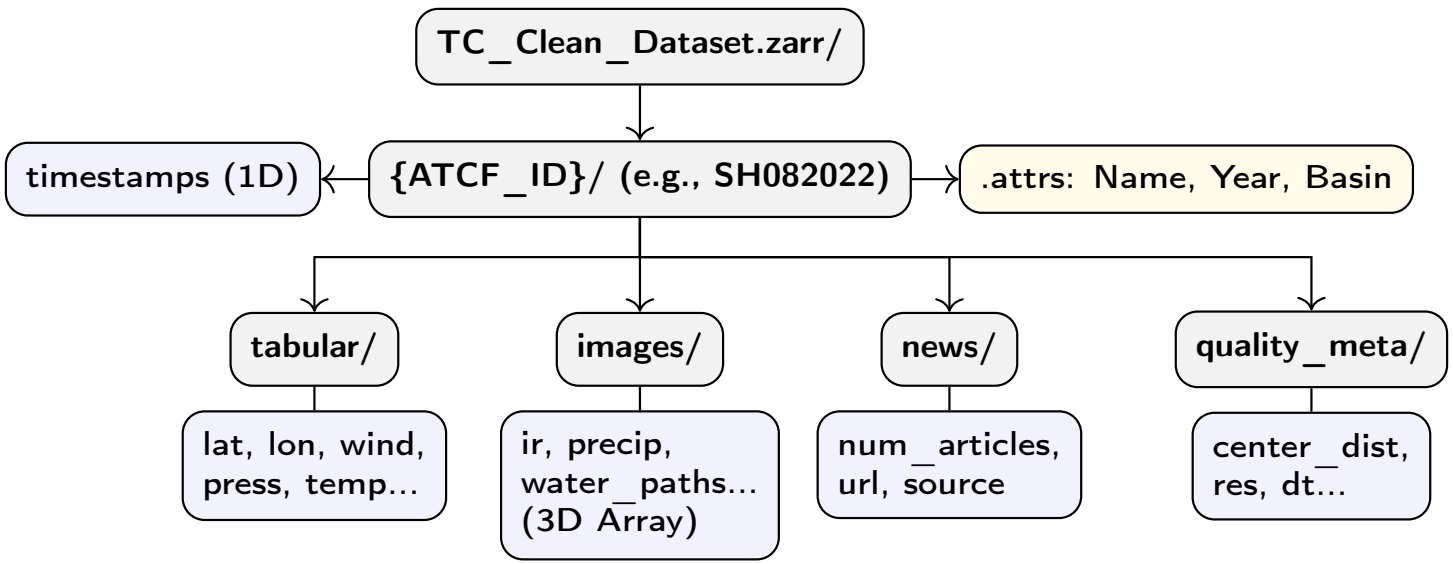
- **Image Processing**: Resized all multi-sensor products to a uniform **224×224** resolution using bilinear interpolation for ML compatibility.
- **Sentinel Masking**: Critical masking of scientific fill-values (-9999.0 for GPROF, 0.0 for IR) to **NaN** *before* resizing to prevent interpolation artifacts.
- **Rich Metadata Extraction**:
  - **Center Distance**: Logged distance (km) between IBTrACS center and satellite-view center (mean $\sim$10-15km) to ensure storm centering.
  - **Time Difference**: Captured exact Δt between observation and best-track.
  - **Pixel Resolution**: Stored post-resize resolution to maintain physical scale awareness (different for IR/GPROF).

Multi-Modal Fusion: Data Integration

- **Broadcasting Strategy**: Daily news was duplicated across sub-daily timestamps, missing days were forward-filled from the previous day to ensure continuity.
- **Synergy**: Every sample aligns **Visual** (Imagery), **Atmospheric** (ERA5), **Intensity** (IBTrACS), **Social** (GDELT), and **Impact** (ReliefWeb) features.

Final Architecture: The Unified Zarr Database

- **Why Zarr?**:
  - **Native Multi-modality**: Natively handles both high-dimensional imagery (2D/3D tensors) and tabular data in a single unified numerical structure.
  - **Tensor Synchronization**: Groups 1D news/ERA5 data and 3D imagery into one container sharing a common temporal index.
  - **ML Efficiency**: Chunked storage allows data loaders to stream samples directly into model tensors without intermediate decoding.



Dataset Quality Assessment

- **ERA5**:
  - **Missing Cyclones**: **7%**
  - **Precision Metrics**:
    - **Mean Center Distance**: **19 km**
- **GDELT**:
  - **Low Media Coverage:** ∼**36%** of cyclones have articles
  - Missing mainly for weak, offshore, or low-media cyclones
  - **Filtering intentionally strict** to avoid false positives
- **ReliefWeb**:
  - **Missing Cyclones**: **48%**
  - Many articles had to be filtered through a scoring function, notable amount left out because of the strict filtering
- **Satellite images**:
  - **Missing Frames**: IR **6.69%** | GPROF **5.58%**.
  - **Precision Metrics**:
    - **Mean Center Distance**: **10.71 km**
    - **Mean Time Bias**: **35.25 min**