A Multi-Instrument Time-Series Dataset for Pre-Flare Solar Activity and SEP Event Prediction

Pouya Hosseinzadeh, Soukaina Filali Boubrahimi, Shah Muhammad Hamdi

Department of Computer Science, Utah State University, Logan, UT, USA

Abstract

We introduce a harmonized, multi-instrument time-series dataset for studying and predicting Solar Energetic Particle (SEP) events from pre-flare conditions. The dataset combines OMNI solar-wind and interplanetary magnetic-field measurements with GOES X-ray and proton flux observations from 1998–2013. Each flare is represented by a fixed 24-hour pre-flare window at 5-minute cadence and is labeled as SEP or non-SEP through cross-validation among MEMPSEP, GSEP, and NOAA catalogs across four proton energy thresholds. The final product is a reproducible, machine-learning-ready benchmark for SEP classification, statistical analysis, and space-weather forecasting.

Motivation and Gap

SEP events are hazardous for spacecraft operations, astronauts, and communication/navigation systems. However, SEP prediction remains difficult because particle events depend on both flare activity and the surrounding heliospheric conditions before the flare.

Existing GOES, OMNI, and SEP catalog resources are scientifically valuable, but they differ in cadence, format, missing-data structure, and event definitions. This creates a barrier for reproducible machine-learning studies. Our dataset addresses this gap by providing a common event-centered format for pre-flare SEP analysis.

Dataset Construction Workflow

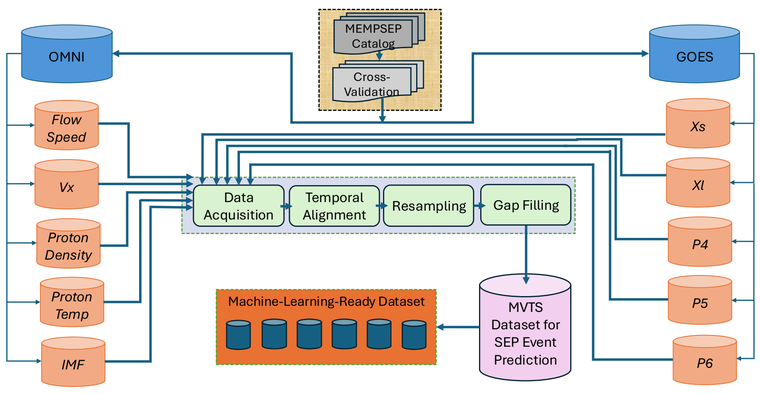

Figure 1. Workflow for event selection, catalog cross-validation, OMNI/GOES data acquisition, temporal alignment, resampling, gap filling, and construction of 24-hour pre-flare windows.

Pre-Flare Event Window

Each flare is represented by a fixed 24-hour observation window immediately preceding the reported flare onset time. At a 5-minute cadence, this produces 288 time steps per event, giving every SEP and non-SEP case the same temporal structure for comparison and machine-learning analysis.

This event-centered design focuses the dataset on pre-flare solar and heliospheric conditions rather than post-flare responses. By excluding information after flare onset, the dataset avoids using signatures that may already be consequences of the eruption, making it better suited for forecasting-oriented studies.

For each event, OMNI and GOES measurements are aligned relative to the flare onset time and stored as synchronized multivariate time series. This allows models to learn from the evolution of solar-wind, magnetic-field, X-ray, and proton-flux behavior before the flare, while preserving a consistent input shape across all energy thresholds and event classes.

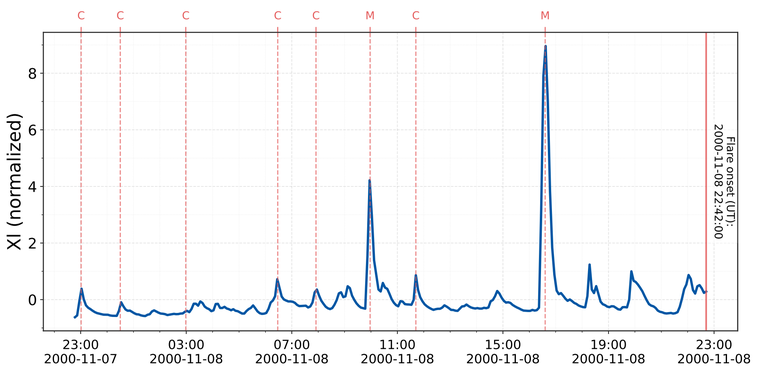

Figure 2. Example GOES X-ray profile during the 24 hours preceding an SEP-associated flare. The fixed pre-flare window captures temporal variability before the reported onset time.

Dataset and Event Distribution

The dataset spans 1998–2013 and includes 168 SEP events at >10 MeV and 17,542 non-SEP events. Labels are provided for four proton energy thresholds: >10, >30, >60, and >100 MeV. Each event contains ten physical variables from OMNI and GOES, including solar-wind velocity, flow speed, proton density, proton temperature, IMF magnitude, X-ray fluxes, and proton flux channels.

Time span	1998–2013
Cadence	5 min
Window length	24 h before flare onset
Samples/event	288
Variables	10 OMNI/GOES measurements
Labels	SEP/NSEP at 4 energy thresholds

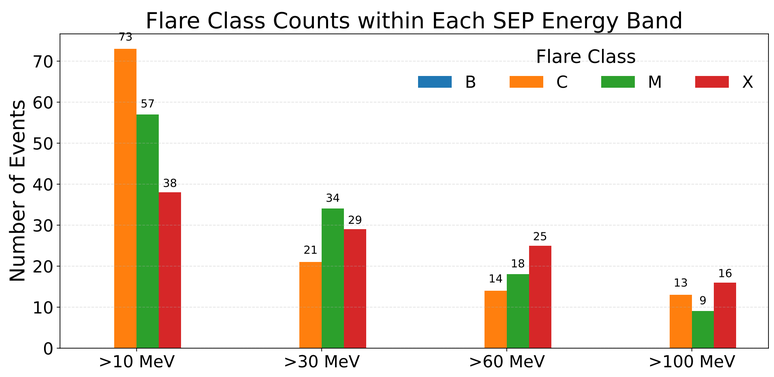

Figure 3. Distribution of B-, C-, M-, and X-class flares for SEP events at >10, >30, >60, and >100 MeV.

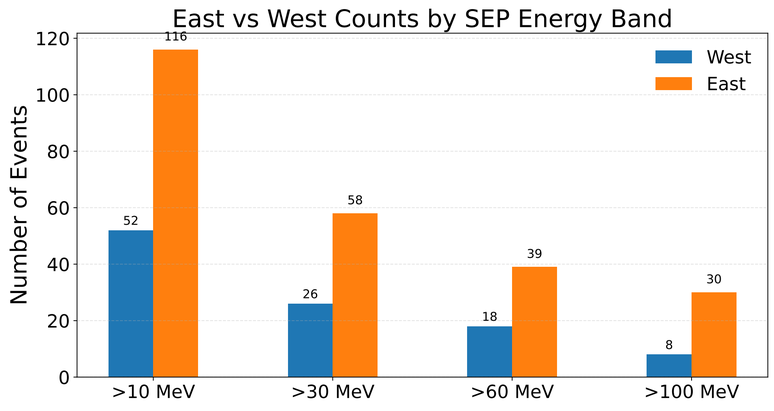

Figure 4. East/west longitudinal distribution of SEP events across proton energy thresholds.

Methodology

Event labeling. Candidate SEP events were identified from MEMPSEP and cross-validated with GSEP and NOAA Solar Proton Event lists. Events confirmed by at least two catalogs and supported by GOES proton flux enhancement were retained as SEP cases; flares without proton enhancement were retained as non-SEP cases.

Time-series harmonization. OMNI and GOES measurements were aligned relative to flare onset and resampled to a uniform 5-minute cadence. Each variable was stored as a per-variable CSV file containing metadata and 288 pre-flare time steps.

Quality control. Missing, sentinel, and nonphysical values were removed or gap-filled using a consistent interpolation and median-filling procedure. This preserves the physical variability of the measurements while producing complete event windows for downstream analysis.

Main Contribution

This work provides a curated, event-aligned benchmark that reduces preprocessing effort and enables direct comparison of SEP forecasting methods.

- Combines OMNI solar-wind/IMF data with GOES X-ray and proton-flux observations.
- Uses fixed 24-hour pre-flare windows at 5-minute cadence for all events.
- Provides catalog-validated SEP/NSEP labels across four proton energy thresholds.
- Delivers a machine-learning-ready format for classification, feature analysis, and benchmarking.

Technical Validation

Validation was performed through catalog agreement, timestamp consistency checks, physical-range screening, missing-value inspection, and exploratory SEP/NSEP comparison. SEP-associated flares show systematic pre-flare differences across solar-wind, IMF, X-ray, and proton-flux variables, especially close to flare onset.

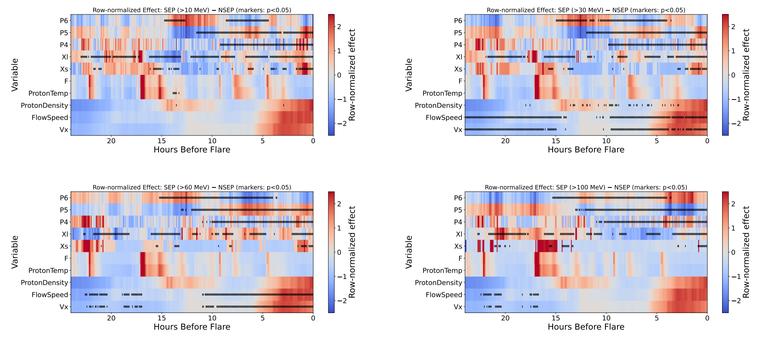

Figure 5. Row-normalized pre-flare profiles for all ten variables, comparing SEP and NSEP events across four proton energy bands. Markers indicate statistically significant group differences.

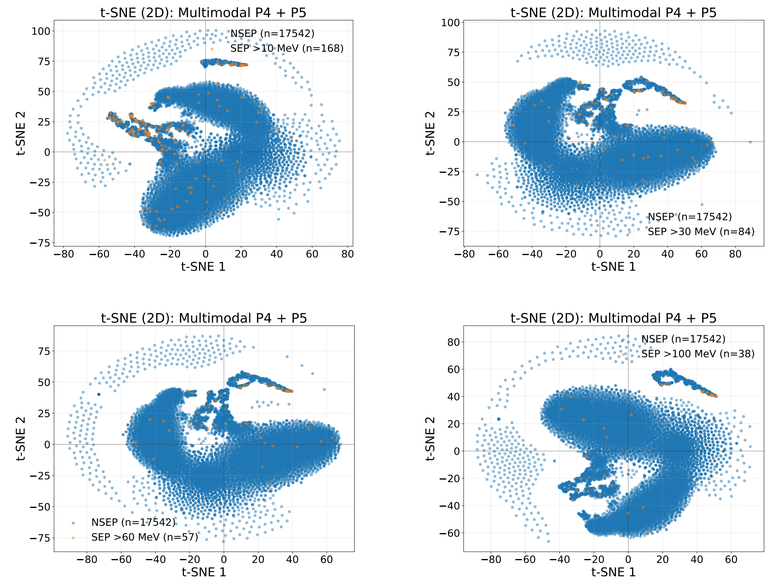

Figure 6. t-SNE projections of event representations based on GOES proton flux variables P4 and P5 for NSEP events and SEP events across four energy thresholds.

Key Findings

- SEP labels are highly imbalanced relative to the non-SEP flare population, motivating threshold-specific evaluation.
- Flare class and east/west longitude distributions vary across proton energy thresholds, so event context should be preserved with the labels.
- Heatmaps and P4/P5 t-SNE projections show pre-flare structure, but also overlapping classes, making this a challenging forecasting benchmark.

Applications

Use cases include SEP/non-SEP classification, energy-threshold prediction, feature analysis, statistical comparison of pre-flare solar-wind and radiative behavior, and benchmarking machine-learning models for space-weather forecasting.

Conclusion

We present a reproducible, multi-instrument, event-centered dataset for SEP prediction. By combining OMNI solar-wind and IMF measurements, GOES X-ray and proton-flux observations, and catalog-validated SEP labels, the dataset provides a consistent 24-hour pre-flare view of solar and heliospheric conditions.

This standardized format supports both physical interpretation and data-driven forecasting, enabling direct comparison of SEP and non-SEP events across shared variables, time windows, and proton energy thresholds. The dataset can serve as a benchmark for developing and evaluating machine-learning models for space-weather forecasting.

Data and Code

Dataset Zenodo record 17595580

Code `github.com/pouyahosseinzadeh/SEP-Multi-Instrument-Dataset`