Supplementary Information: SuryaBench

Low Quality AIA Data and QUALITY Keyword

AIA image headers and data may be affected by operational events such as off-pointing, defocusing, or missing data during eclipse seasons. These issues are flagged by a non-zero QUALITY keyword in the image header. It is important to check the QUALITY before detailed analysis of AIA data. Note that the QUALITY keyword is a 32-bit integer with bitwise flags. We present a set of examples in Figure 1.

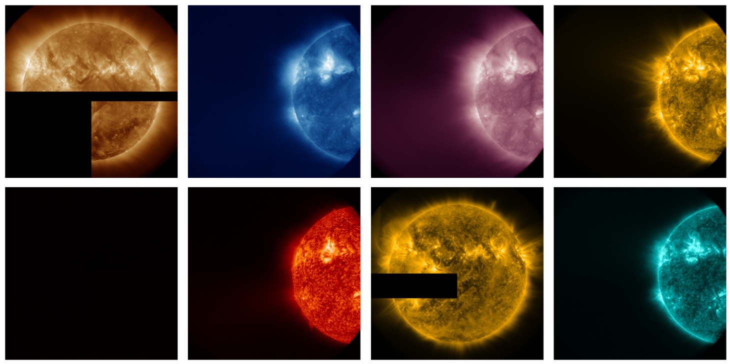

Figure 1. Bad AIA measurements due to a variety of reasons.

Active Region Segmentation

For the active region segmentation task described in Section Active region segmentation, we have created ARPIL dataset. This dataset contains binary maps of detected locations of active region with polarity inversion lines, which are to-be-overlaid on the original pre-processed SDO rasters in SuryaBench dataset. As part of our validation efforts, we have created spatial coverage maps of detected active regions locations for each year between 2010 and 2024 in Figure 2. We also present the temporal evolution of number of pixels covered by the active region locations in Figure 3, where we can observe the solar cycle phases over the lifespan of the dataset.

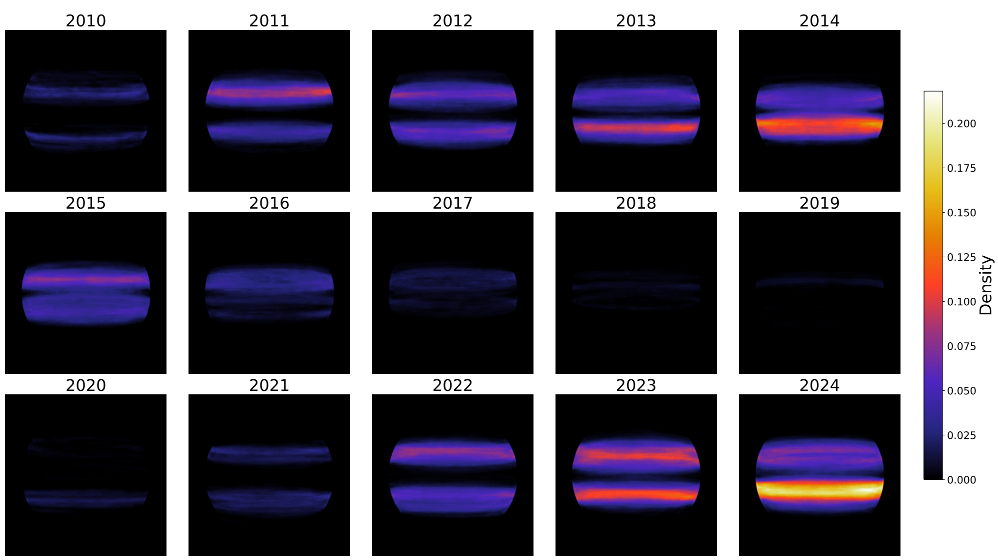

Figure 2. Spatial coverage of active region locations represented as density maps from 2010 to 2024. For each year, binary maps in the ARPIL dataset were summed on a pixel-wise basis. The resulting density values were obtained by normalizing the yearly sums by the total number of maps available for that year.

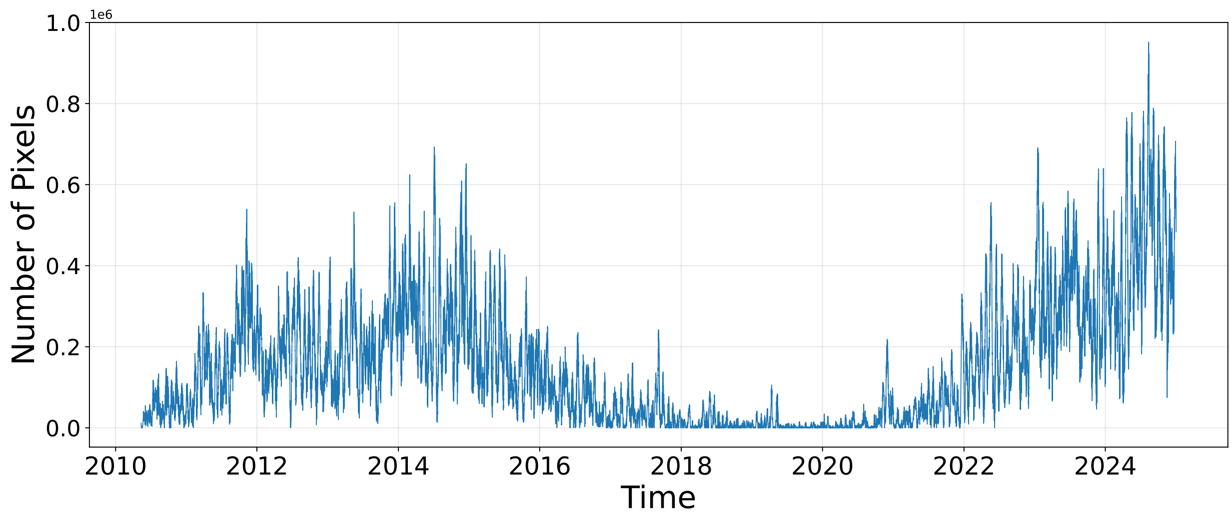

Figure 3. The temporal evolution of area covered by active regions in ARPIL dataset in number of pixels.

Flare Forecasting

For the flare forecasting task described in Section see Flare forecasting section, we have created two labels: maximum flare intensity (X-ray flux) and cumulative flare index for a 24h prediction window after a time point observation. We demonstrate the distribution of ordinal flare GOES subclasses in Figure 4 for maximum intensity labels, meaning we present the counts of each GOES classes (A, B, C, M, and X) and subclasses ([1.0 to 9.9]). Note here that labels are computed using a rolling 24h prediction window. The distribution is strongly skewed toward smaller events: C-class flares dominate, and the number of flares decreases systematically with class (C $\gg$ M $\gg$ X), yet we also note that for A and B classes such relationship does not exist as they are severely underdetected especially in solar minima.

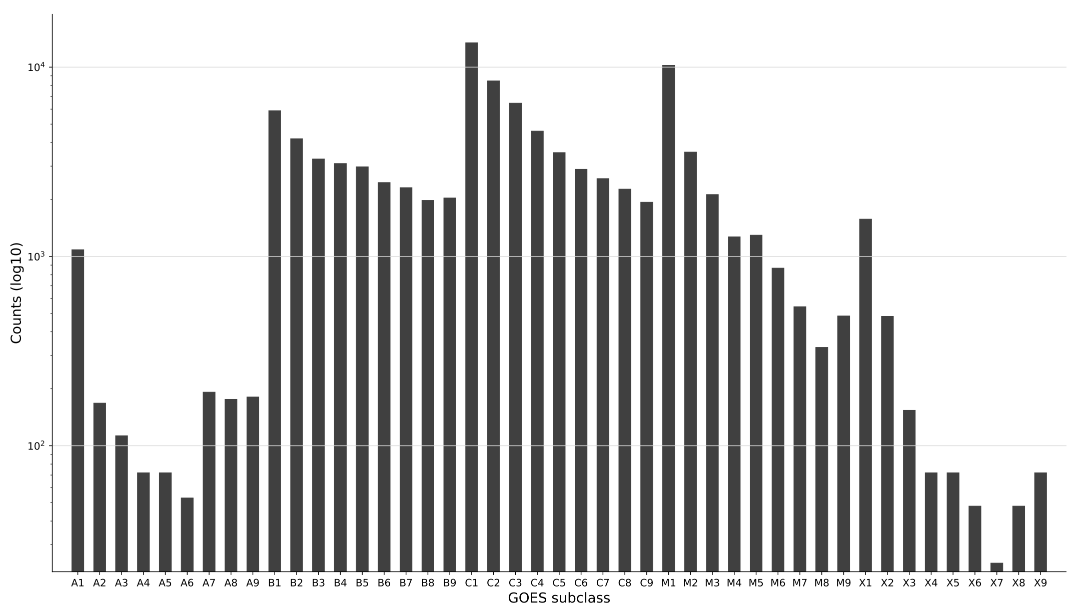

Figure 4. The distribution of total number of flare labels (based on maximum intensity labeling – See L_{max} in Eq. **??**) binned using ordinal GOES class and sub-classes. Each bin covers the interval starting from the bin label to the next one; e.g., C7 shows the total number of L_{max} labels from C7.0 to C7.9. Counts are on logarithmic scale.

Figure 5. The total number of flare label counts for C+, M+, X+ (i.e., $\geq C1.0$, $\geq M1.0$, $\geq X1.0$, respectively) flares aggregated on each month from 2010 to 2024. Counts are on logarithmic scale.

We also show the monthly total number of flare occurrences for C+, M+, and X+ flares in Figure 5 from May 2010 to December 2024. Activity peaks during the maximum of Solar Cycle 24 (2011–2015), reaches a deep minimum around 2018–2020 at the Solar Cycle 24 to 25 transition, and rises again with the onset of Cycle 25 (late 2019/2020 onward). Throughout the record, C-class months outnumber M and X, while X-class months are sporadic.

Dataset Sampling for SuryaBench

To ensure consistent, fair, and reproducible training and evaluation across multiple solar datasets, we adopt a custom, buffered, non-chronological temporal sampling schema. This strategy is inspired by non-chronological leakage-aware propensity modeling, enables balanced sampling across different phases of solar cycles while mitigating temporal data leakage (that may be caused by temporal and spatial autocorrelations) and supporting both in-distribution (ID) and out-of-distribution (OOD) performance evaluation.

For the core SDO imaging dataset and application benchmark datasets where we have consistent coverage, i.e.; active region segmentation (Dataset **DS1**), coronal field extrapolation (Dataset **DS3**), flare forecasting (Dataset **DS4**) and solar wind forecasting (Dataset **DS5**), we adopt the following temporal splitting scheme:

- **Training:** February 15–December 31 of each year from 2010 to 2019.
- **Validation:** January 15–31 of each year from 2010 to 2019.
- **Buffer (Leaky Validation):** January 1–14 and February 1–14 of each year from 2010 to 2019.
- **Testing:** All days from January 1, 2020 to December 31, 2024.

This configuration results in four subsets: *training*, *validation*, *testing*, and *leaky_validation (buffers)*.

This sampling design serves multiple purposes. First, the inclusion of two-week temporal buffers between adjacent segments helps prevent potential data leakage arising from overlapping or closely spaced solar observations. Note that a solar rotation takes about 25-27 days and two-week buffers guarantee a half rotation. These buffer periods act as isolation zones, ensuring that the temporal coherence of solar features does not artificially inflate model performance. Second, the non-chronological validation scheme is deliberately constructed to draw samples from all phases of Solar Cycle 24. By doing so, the validation set represents an in-distribution sample that reflects the full range of solar activity levels, from quiet to active phases, providing a robust assessment of model performance within the same solar cycle. Third, the testing period is defined entirely within Solar Cycle 25 (2020–2024), enabling evaluation on unseen and temporally distant data. This simulates an out-of-distribution testing condition, where models are challenged to generalize across distinct solar cycle characteristics. Finally, the buffered intervals retained as a leaky validation set allow for sensitivity analyses, both optimistic and conservative, on the degree of potential temporal overlap or leakage, thus supporting reproducibility in downstream experiments.

For smaller or specialized datasets (Solar EUV Spectra Modeling and Active Region Emergence Forecasting datasets) with limited temporal or event-based coverage, we adopt modified versions of the above schema while preserving its core principles of buffer isolation and temporal splits.

Solar EUV Spectra Modeling Dataset

The solar EUV spectra modeling dataset (Dataset **DS6**) contains continuous observations from May 13, 2010 to May 26, 2014. We split the datasets using the following schema:

- **Training:** February 15–December 31 of each year from 2010 to 2014.
- **Validation:** January 15–31 of 2012 and 2013.
- **Testing:** January 15–31 of 2011 and 2014.
- **Buffer (Leaky Validation):** January 1–14 and February 1–14 of each year from 2010 to 2014.

This configuration mirrors the generic schema for this limited time frame, while maintaining a reasonable balance between validation and testing sizes. As this dataset spans only partially Solar Cycle 24, no explicit OOD testing is available.

Active Region Emergence Forecasting Dataset

The active region emergence forecasting dataset (Dataset **DS2**) includes samples from November 2010 to February 2023. We use the following splitting schema:

- **Training:** February 15–December 31 of each year from 2010 to 2019.
- **Validation:** January 1–February 14 of each year from 2010 to 2019.
- **Testing:** All days from 2020 to 2023.

This produces a three-way split (*training*, *validation*, *testing*) without explicit buffer zones. Note that leakage is naturally avoided in this dataset as individual instances correspond to separate active regions. Approximately two-thirds of the instances (37/59) belong to the training set, with the remainder divided between validation (9) and testing (13). This configuration is broadly aligned with our generic split.

Notes on Extensions

By defining a common baseline sampling strategy, we aim to promote transparent, longitudinal, and cycle-consistent evaluation practices within the heliophysics and space weather forecasting community. Note that the aforementioned splitting schemata represent our recommended standard splits to facilitate consistent benchmarking and reproducibility across heliophysics datasets. Nevertheless, researchers are encouraged to adapt the approach for datasets originating from alternative heliophysics tasks or instruments. For example, datasets derived from partial field-of-view observations (such as STEREO or different temporal coverage (such as GOES/SUVI) may require altered, custom temporal divisions.

Overview of Baseline Learning Models for Application Benchmark Datasets

This supplementary material provides a detailed examination of the tasks from a machine learning perspective, including task-specific objectives, model architectures, and relevant implementation details.

Solar Flare Forecasting

Solar flare forecasting is framed as a binary classification task where the goal is to predict whether a strong solar flare (i.e., M- or X-class) will occur within a 24-hour window following a given observation time t. The prediction window spans the interval $[t, t+24)$ hours. Within this window, multiple solar flare events may occur. To assign a label to the input at time t, two labeling strategies are used:

- **Maximum Flare Intensity**: We select the flare with the highest intensity in the 24-hour prediction window. If this maximum intensity exceeds a threshold of 10^{-4} W/m^2, the input is labeled as positive (flare will occur); otherwise, it is labeled negative.
- **Cumulative Flare Intensity**: We sum the intensities of all flares that occur in the prediction window. If the cumulative intensity exceeds a threshold of 10, the input is labeled positive; otherwise, negative.

Note that for our experiments, we used the Maximum Flare Intensity label.

Problem Formulation: Given solar observation data at time t, the goal is to predict whether a significant solar flare will occur within the subsequent 24-hour window, i.e., in the interval $[t, t+24)$. This is framed as a binary classification task:

$$f(\mathbf{x}_t) \to \{0, 1\}$$

where $\mathbf{x}_t \in \mathbb{R}^{C\times H\times W}$ (or potentially $\mathbb{R}^{T\times C\times H\times W}$ for temporal stacks) represents the multi-channel input image (or sequence) at time t, and the output is a binary label:

$$y_t = \begin{cases} 1, & \text{if a flare is expected in } [t, t+24) \\ 0, & \text{otherwise} \end{cases}$$

To determine y_t, two flare labeling strategies are considered:

- **Maximum Flare Intensity:** The label is set to 1 if the maximum flare intensity in $[t, t+24)$ exceeds a fixed threshold $\theta_{\text{max}} = 10^{-4}$ W/m^2.
- **Cumulative Flare Intensity:** The label is set to 1 if the sum of all flare intensities in $[t, t+24)$ exceeds a threshold $\theta_{\text{sum}} = 10$.

Evaluated Model Architectures: We evaluate the performance of several standard convolutional neural network (CNN) architectures adapted for binary classification. Each model takes in spatial or spatiotemporal representations of solar magnetic field or other physical parameters (details omitted here) and outputs a binary prediction. Evaluated architectures include the following:

- **AlexNet**: A lightweight CNN with five convolutional layers followed by three fully connected layers. Its shallow depth makes it faster to train and less prone to overfitting in small datasets.
- **MobileNet**: A mobile-optimized architecture using depthwise separable convolutions to reduce computational cost. Useful for efficient forecasting on edge devices.
- **ResNet50**: Residual Network with 50-layer depths, incorporating skip connections to enable better gradient flow and deeper representations.

All models are modified with a final fully connected layer followed by a sigmoid activation to output a probability score for binary classification.

Results and Observations: Table **??** shows the evaluation metrics we used for solar flare forecasting:

- **TSS (True Skill Statistic)**:

$$\text{TSS} = \frac{TP}{TP+FN} - \frac{FP}{FP+TN}$$

Measures the ability to distinguish between flare and non-flare events. Ranges from -1 (inverse prediction) to +1 (perfect prediction), with 0 indicating no skill.

- **HSS (Heidke Skill Score)**:

$$\text{HSS} = \frac{2(TP \cdot TN - FP \cdot FN)}{(TP+FN)(FN+TN)+(TP+FP)(FP+TN)}$$

Evaluates performance relative to random chance, considering both hits and false alarms. Ranges from $-\infty$ to 1.

- **F1 Score**:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP+FP+FN}$$

Harmonic mean of precision and recall, balancing both false positives and false negatives.

Table **??** summarizes the evaluation results. Among the tested architectures, **MobileNet** achieved the best overall performance across TSS, HSS, and F1 metrics, indicating strong recall and balanced precision. **AlexNet** also performed competitively, likely benefiting from its simpler structure and reduced overfitting on limited data. In contrast, deeper models such as **ResNet50** underperformed, possibly due to excessive capacity relative to dataset size or insufficient regularization.

Solar Wind Forecasting

Solar wind forecasting is a critical regression task aimed at predicting the solar wind speed at a given spatial point, specifically within a prediction window of 4 days following an observation time t. Precise forecasting of solar wind speeds is fundamental for mitigating the adverse effects of space weather on satellite communication systems, navigation systems, and electrical grids on Earth.

This dataset comprises scalar measurements of solar wind speeds, recorded hourly from 2010-01-01 through 2023-12-31, resulting in a temporally rich dataset with substantial coverage of solar cycles. The solar wind speed values exhibit significant variability, ranging from 2.4×10^2 km/s to 8.8×10^2 km/s.

Problem Formulation

Formally, given solar observation data (such as AIA and HMI multi-channel solar imaging data) represented by $\mathbf{x}_t$ at observation time t, the task is to predict the scalar solar wind speed at time $t + \Delta t$, where $\Delta t = 4$ days:

$$y_{t+\Delta t} = f(\mathbf{x}_t),$$

where $\mathbf{x}_t \in \mathbb{R}^{C \times H \times W}$ represents the multi-channel, high-resolution input imagery data at time t, and $y_{t+\Delta t} \in \mathbb{R}$ represents the predicted scalar solar wind speed.

Solar EUV spectra prediction

Predicting solar Extreme Ultraviolet (EUV) irradiance accurately is crucial for understanding and forecasting space weather, which directly impacts satellite operations, communication systems, and navigation. This task involves forecasting irradiance across a spectrum of 1343 spectral channels, reflecting complex spatial and temporal patterns captured by solar imagery.

Formally, we frame this as a regression problem: Given multi-channel solar imaging data $\mathbf{x}_t \in \mathbb{R}^{C \times H \times W}$ at time t, the goal is to predict a continuous vector of EUV irradiance values $y_t \in \mathbb{R}^{1343}$:

$$y_t = f(\mathbf{x}_t)$$

Active region emergence forecast

We address the task of **predicting continuum intensity** over a spatially-distributed grid of solar active regions, using historical measurements of solar magnetic flux and acoustic power. This task encapsulates a complex spatiotemporal forecasting problem grounded in heliophysics, where both **temporal dependencies** and **spatial interactions** are crucial for accurate modeling. Understanding and forecasting continuum intensity has strong implications for solar physics and space weather prediction. High-intensity regions on the solar surface are often precursors to flare activity and magnetic storms.

Problem Formulation: Given a tracked solar active region observed over a sequence of $T = 120$ time steps (i.e., 1 day in our dataset), our objective is to predict the *continuum intensity* at a subsequent time point for each spatial grid cell. Formally, let:

- $\mathbf{X} \in \mathbb{R}^{T\times C\times S}$ denote the input tensor for a region, where:
  - $T = 120$ is the number of time steps (~24 hours, sampled at 12-minute cadence),
  - $C = 5$ is the number of physical quantities per cell,
  - $S = 63$ is the number of spatial cells in the tracked region.
- $\mathbf{y} \in \mathbb{R}^{S}$ denote the target continuum intensity per cell.

The model is trained to approximate a function $f_\theta : \mathbb{R}^{T\times C\times S} \rightarrow \mathbb{R}^{S}$ that maps the temporal and spatial patterns of input features to the scalar output intensities.

The five input channels correspond to two types of physical measurements:

- **Mean Unsigned Magnetic Flux (1 channel)**: captures the net strength of local magnetic fields in the region.
- **Doppler Velocity Acoustic Power (4 channels)**: measured across four distinct frequency bands: 2–3, 3–4, 4–5, and 5–6 mHz, these capture multi-scale oscillatory dynamics linked to wave propagation and subsurface flows.

Spatially, each active region is tracked and cropped into a 9×9 grid of tiles. However, we discard the top and bottom rows for normalization reasons (e.g., to suppress edge artifacts), resulting in a 7×9 = 63 spatial cells. For each time step, the full tensor $\mathbf{X}_t \in \mathbb{R}^{C\times S}$ captures these 5-channel inputs over the grid.

Dataset and Temporal Context: The dataset comprises 3,479 unique regions, indexed and temporally aligned. For each region:

- **Input timestamps** span a window of 120 steps (~24 hours),
- **Output timestamp** is a single future point for which continuum intensity is predicted.