SuryaBench: Benchmark Dataset for Advancing Machine Learning in Heliophysics and Space Weather Prediction

Sujit Roy[1,2,*], Dinesha V. Hegde[3,4], Johannes Schmude[5], Amy Lin[1], Vishal Gaur[1], Rohit Lal[1], Kshitiz Mandal[1], Talwinder Singh[6], Andrés Muñoz-Jaramillo[7], Kang Yang[6], Chetraj Pandey[6], Jinsu Hong[6], Berkay Aydin[6], Ryan McGranaghan[8], Spiridon Kasapis[9], Vishal Upendran[10], Shah Bahauddin[11], Daniel da Silva[12], Marcus Freitag[5], Iksha Gurung[1], Nikolai Pogorelov[3,4], Campbell Watson[5], Manil Maskey[2], Juan Bernabe-Moreno[5], and Rahul Ramachandran[2]

[1]Earth System Science Center, University of Alabama in Huntsville, AL, USA
[2]NASA Marshall Space Flight Center, Huntsville, AL, USA
[3]Department of Space Science, The University of Alabama in Huntsville, AL, USA
[4]Center for Space Plasma and Aeronomic Research (CSPAR), The University of Alabama in Huntsville, AL, USA
[5]IBM Research
[6]Georgia State University
[7]Southwest Research Institute
[8]NASA Jet Propulsion Laboratory
[9]Princeton University
[10]SETI Institute
[11]Laboratory for Atmospheric and Space Physics, University of Colorado Boulder
[12]NASA Goddard Space Flight Center
[*]corresponding author(s): Sujit Roy (sujit.roy@nasa.gov)
[†]All authors contributed equally to this work

ABSTRACT

This paper introduces a high resolution, machine learning-ready heliophysics dataset derived from NASA's Solar Dynamics Observatory (SDO), specifically designed to advance machine learning (ML) applications in solar physics and space weather forecasting. The dataset includes processed imagery from the Atmospheric Imaging Assembly (AIA) and Helioseismic and Magnetic Imager (HMI), spanning a solar cycle from May 2010 to July 2024. To ensure suitability for ML tasks, the data has been preprocessed, including correction of spacecraft roll angles, orbital adjustments, exposure normalization, and degradation compensation. We also provide auxiliary application benchmark datasets complementing the core SDO dataset. These provide benchmark applications for central heliophysics and space weather tasks such as active region segmentation, active region emergence forecasting, coronal field extrapolation, solar flare prediction, solar EUV spectra prediction, and solar wind speed estimation. By establishing a unified, standardized data collection, this dataset aims to facilitate benchmarking, enhance reproducibility, and accelerate the development of AI-driven models for critical space weather prediction tasks, bridging gaps between solar physics, machine learning, and operational forecasting.

1 Background & Summary

Advancing heliophysics, the study of the Sun and its influence on the solar system, is crucial given space weather's tangible impacts on critical infrastructure like communications, navigation, and power grids [1]. NASA's Solar Dynamics Observatory (SDO) [2] continually captures extensive ($\sim$1.5 TB/day), high-quality multi-instrument solar data, turning heliophysics into a data-intensive discipline. This vast observational data from SDO offers a unique opportunity to leverage machine learning (ML) techniques to tackle persistent challenges in solar and heliospheric physics [3, 4]. However, leveraging SDO data presents notable challenges, including specialized preprocessing and

domain-aware computational capabilities to homogenize the multi-instrument database [5, 6]. A publicly available SDO-ML-ready dataset exists [5], but its reduced spatial resolution (512×512) limits the full potential of the original SDO observations.

To address these challenges, we introduce a curated, publicly accessible benchmark dataset from SDO, SuryaBench, comprising high-resolution observations of the solar surface and atmosphere, which can be used to study diverse solar and heliospheric phenomena such as flares, coronal holes (CH), active regions (AR), sunspots, solar wind, and coronal loops. To our knowledge, SuryaBench is the largest curated and homogenized dataset to date, and it preserves the full 4096×4096 native spatial resolution of SDO and provides a consistent 12-minute temporal cadence, enabling high-fidelity analysis for data-driven heliophysics research. The dataset is designed to advance data-driven heliophysics research and support operational workflows by offering standardized preprocessing, temporal and spatial homogenization, rich metadata for seamless interoperability, and AI-ready formats, enabling the development and deployment of large-scale machine learning models, specifically self-supervised learning and foundation models [4], and a wide spectrum of heliophysics applications.

SuryaBench is designed to enable the development of advanced, physics-informed models for investigating complex solar phenomena. It features detailed documentation and rich metadata to ensure usability for a wide range of users, including both heliophysics researchers and machine learning practitioners, regardless of their prior domain expertise. To provide a comprehensive and reusable testing environment and facilitate synergistic research, SuryaBench offers standardized application benchmark datasets (hereafter called Datasets, for brevity) for the following six key tasks in heliophysics: (1) solar flare prediction, (2) active region segmentation, (3) active region emergence prediction, (4) coronal magnetic field extrapolation, (5) solar irradiance, and (6) solar wind forecasting. Each application benchmark includes rigorous evaluation protocols and baseline implementations of state-of-the-art machine learning architectures, such as Residual Networks and U-Net. Ultimately, we envision SuryaBench to serve as a robust data resource for diverse, cross-cutting heliophysics tasks with spatio-temporal analysis, multimodal data fusion, and interpretability research. Next, we present a brief overview of selected tasks along with our interdisciplinary motivation for their inclusion. We note that these applications are by no means comprehensive, yet they are relevant to multiple interacting phenomena on the Sun and in the heliosphere, and we envision these to serve as the blueprint for the development of large-scale AI applications using SuryaBench.

At the core of many space weather drivers are active regions (ARs), concentrated areas of magnetic flux that frequently produce solar flares and coronal mass ejections (CMEs), with direct impacts on satellites and terrestrial infrastructure [7]. Within ARs, polarity inversion lines (PILs), which are interfaces where magnetic polarities reverse, are critical sites for energy storage and release, and are strongly associated with solar eruptive activity [8, 9, 10, 11, 12]. Twisted and sheared PIL structures can give rise to current sheets and magnetic reconnection, which are central to flare and CME initiation. We provide two tasks related to ARs in Datasets DS1 and DS2 (Sec. 2.2.1, 2.2.2), where Dataset DS1 is focused on segmentation of ARs with PILs and Dataset DS2 is focused on AR emergence prediction. That said, the emergence and evolution of ARs also significantly affect coronal dynamics, requiring accurate three-dimensional magnetic field modeling. With Dataset DS3 (Sec. 2.2.3), we provide a task on coronal field extrapolation, which supports research on magnetic field extrapolation and AR-induced coronal changes.

Various space weather phenomena have the potential to significantly affect both near-Earth space environments and terrestrial systems. Solar flares, intense eruptions originating in the solar chromosphere and corona, can trigger geomagnetic storms, impacting terrestrial and space-based infrastructure, and posing risks to astronauts [13, 14]. Similarly, the solar wind, which is a continuous outflow of charged particles from the solar corona, modulates Earth's magnetosphere and drives geomagnetic storms with operational implications [1, 15, 16]. We provide two tasks related to space weather forecasting with Datasets DS4 and DS5 (Sec. 2.2.4, 2.2.5), where Dataset DS4 is focused on flare prediction and Dataset DS5 is focused on solar wind prediction. Lastly, solar extreme ultraviolet (EUV) irradiance plays a key role in shaping Earth's ionospheric and thermospheric conditions, influencing satellite drag, communication systems, and GPS accuracy [17, 18, 19, 20]. Dataset DS6 (Sec. 2.2.6) addresses EUV nowcasting and forecasting to support satellite operations and mission planning.

2 Methods

Our benchmark dataset includes a core imaging data collection, which is primarily designated as input, and six application benchmark datasets, intended as labels, covering different solar physics and space weather applications. In the following subsections, we describe the steps and procedures used to create and curate the data along with the details for each of our application datasets.

Table 1. Instrumental properties of SDO/AIA and SDO/HMI. Both instruments take 4096×4096 images. AIA measures photometric intensity in EUV for different wavebands. Its channels, denoted in Å (e.g. 94Å), indicate the wavelength of peak intensity for each pass-band filter. HMI makes spectropolarimetric measurements around magnetically sensitive spectral emission lines. We use inversions using these measurements that estimate the three components of the magnetic field (B_x, B_y, B_z), line-of-sight (LOS) magnetic field (B_{los}), and LOS velocity (V_{los}). Cadence refers to the time interval between two consecutive images. We make the distinction between instrumental cadence (12s to 12m) and the dataset cadence (12m).

Inst.	Resolution (photospheric)	Cadence (instrument)	Cadence (SuryaBench)	Dynamic range	Channels
AIA	1.2" (725km)	12s, 24s	12m	0 to 16,383	94, 131, 171, 193, 211 304, 335, 1600 (in Å)
HMI	1.0" (870km)	45s, 12m	12m	$\sim\pm4,500$ for B $\sim\pm10^4$ for V	B_x, B_y, B_z, B_{los} (in G), v_{los} (in m/s)

2.1 Core SDO Dataset

SDO is a NASA Heliophysics flagship mission launched on February 11, 2010 in geosynchronous orbit. Its main science objectives are to understand how solar magnetism is created, how solar magnetism shapes the extended solar atmosphere that encompasses the entire solar system, and how solar activity affects Space Weather.

SDO has two imaging instruments: 1. The Atmospheric Imaging Assembly (AIA) [21], which measures photometric intensity (per pixel) in the Extreme Ultraviolet (EUV) and UV spectrum. 2. The Helioseismic and Magnetic Imager (HMI)[22], which makes spectropolarimetric measurements used to estimate the surface magnetic field (all three components) and the line-of-sight velocity on the solar surface. Both imagers use 4096×4096 charge-coupled devices (CCDs) to image the solar surface and atmosphere. Table 1 contains details on both AIA and HMI instrumental and data properties.

SDO data is publicly available through the Joint Science Operations Center (JSOC; `http://jsoc.stanford.edu`) as time series of various numerical scalar and raster data products. The data series we have used to create our core dataset are `aia.lev1_euv_12s` for EUV channels, `aia.lev1_uv_24s` for a UV channel, `hmi.M_720s` for a line-of-sight (LOS) magnetogram, and `hmi.B_720s` for vector magnetograms.

2.1.1 AIA Data Acquisition and Processing

The `aia.lev1_euv_12s` series provided by JSOC contains level-1 data. This means that the images still include the roll angle of the satellite, i.e., the solar north-south axis is not aligned with the vertical y-axis, and each channel may have a slightly different pixel scale. To enhance data accessibility, spatial homogenity, and interoperability, we promoted the AIA data from level-1 to level-1.5 . The promotion to level-1.5 involves updating the pointing keywords, removing the roll angle, scaling the image to a common pixel scale of 0.6 arcsec per pixel, and translating the image so that the center of the Sun is located in the center of the image. Besides these steps, exposure time normalization is an extra but necessary step during the promotion because AIA measurements have heterogeneous exposure times ranging from 0.05 to 2.9 seconds. We present an example conversion of AIA data from level-1 to level-1.5 in the two top left panels of Figure 1 using an AIA 171 image instance.

Since the database contains data throughout the SDO lifetime, CCD camera sensor degradation also needs to be taken into account. The table of correction parameters calculated by the AIA science team is made publicly available via JSOC. These parameters are a time series of scalars that can be multiplied by the full disk AIA data to rectify the instrument degradation. However, this method may lead to issues when the corrected values in some of the pixels exceed the instrument saturation value (16,383 for AIA). We post-process and clamp the degradation-corrected image to make sure that none of the pixels reach a value greater than this limit. In Figure 2, we plot the mean pixel intensity values of the full disk images for each of the seven EUV channels of AIA in level 1 (left panel) and level 1.5 with degradation correction (right panel) data. The variation of mean values over the years also reflects the solar cycle, in which the higher mean values indicate stronger solar activity. We notice that the degradation correction restores the higher activity in solar cycle 25 compared with solar cycle 24.

One final step in making sure that the AIA data is ML-ready is to make the solar disk size the same in the whole database for all wavelengths, correcting for the elliptical orbit of the spacecraft. This step makes the solar disk of fixed

radius, 976 arcsecs, in all images. An example of this step is shown in the bottom-left panel of Figure 1.

2.1.2 HMI Data Acquisition and Processing

Though the observable HMI data provided from JSOC are level-1.5, they have a slightly higher resolution of 0.5 arcsec per pixel compared to AIA images. Therefore, the HMI data must be re-projected to be spatially aligned with the level-1.5 AIA images, which have 0.6 arcsec per pixel resolution. This re-projection step involves bilinear interpolation when rescaling the high-resolution images to lower-resolution images. Bilinear interpolation estimates the value of a function $f(x,y)$ at a point (x,y) that lies between four known grid points [23]. In the context of image data, these grid points represent the centers of image pixels, and the function values $f(x,y)$ correspond to pixel intensities. Let the four surrounding grid points be the corners of a rectangle: the bottom-left (x_0,y_0), bottom-right (x_1,y_0), top-left (x_0,y_1), and top-right (x_1,y_1), with corresponding function values $f(x_0,y_0)$, $f(x_1,y_0)$, $f(x_0,y_1)$, and $f(x_1,y_1)$, respectively. If we consider a desired point (x,y) within this rectangle, such that $x_0 \leq x \leq x_1$ and $y_0 \leq y \leq y_1$, the interpolated value is

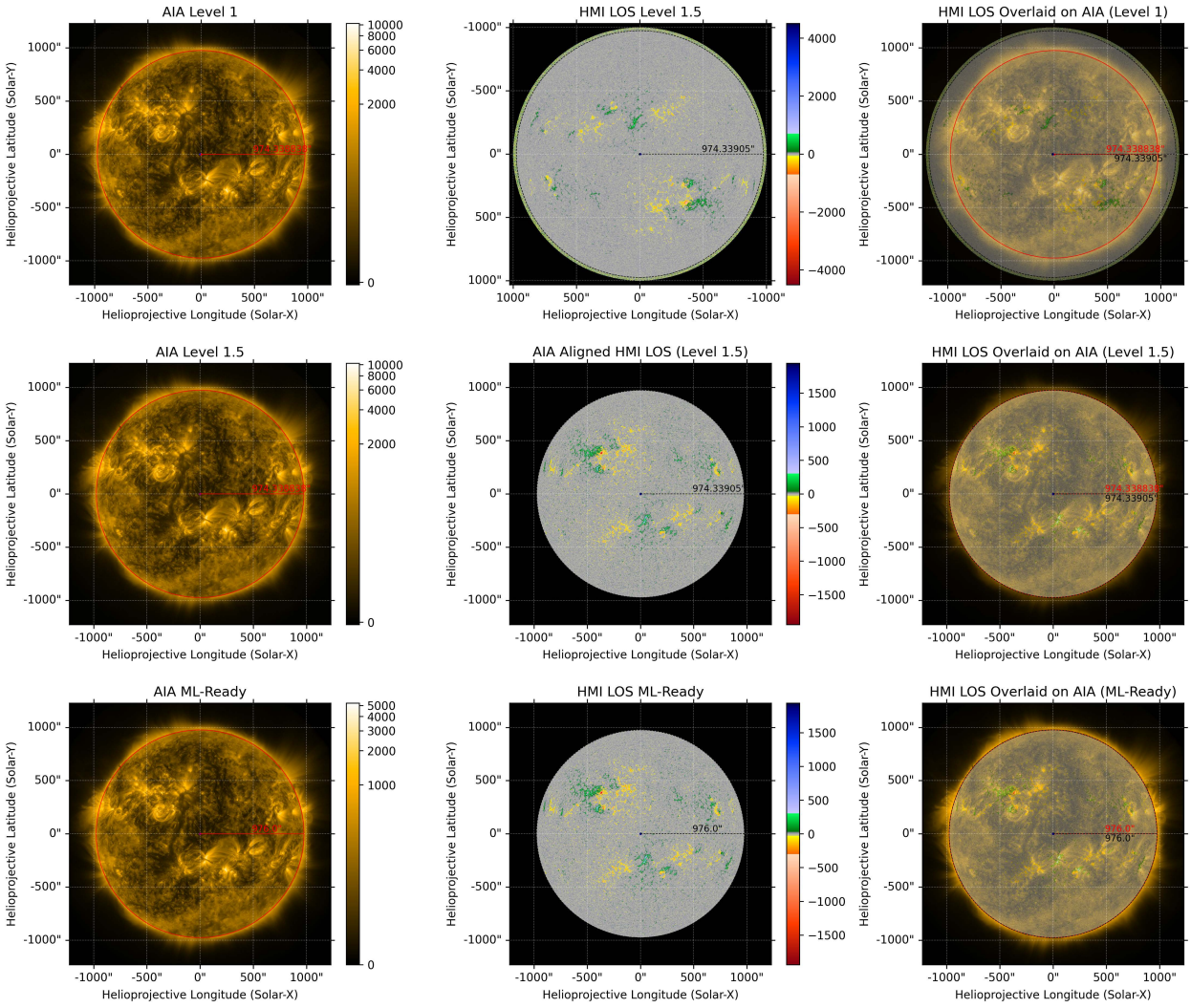


Figure 1. An example of the ML-ready data preparation steps for AIA 171 Å and HMI LOS magnetogram on 2012-01-30 at 22:12 UT. Contours illustrate the image center, solar disk center, disk radius, and solar disk boundary. The top row shows the original AIA Level 1 image, HMI Level 1.5 magnetogram downloaded from JSOC, and HMI overlaid on AIA. The disk centers are misaligned with the image center (unregistered), and one dataset has a 180° roll, with noticeable plate scale differences. The middle row displays the registered AIA Level 1.5 image, HMI aligned with AIA, and HMI overlaid on AIA, showing corrected disk centers and plate scales. The bottom row presents the final ML-ready AIA and HMI images after exposure time normalization and orbital corrections for AIA, with the overlaid image showing proper alignment and a fixed disk radius of 976 arcsecs.

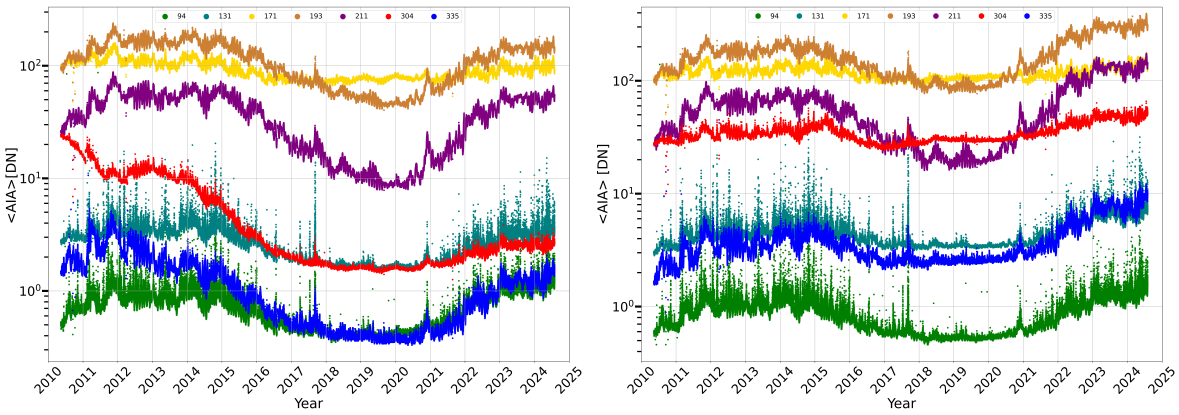


Figure 2. Mean pixel value of full-disk AIA images (extreme ultraviolet channels 94–335Å, DN/sec) over time before (*left panel*) and after (*right panel*) the degradation correction.

given by:

$$f(x,y) = (1-t)(1-u)\,f(x_0,y_0) + t(1-u)\,f(x_1,y_0) + (1-t)u\,f(x_0,y_1) + tu\,f(x_1,y_1).$$

where

$$t = \frac{x-x_0}{x_1-x_0}, \quad u = \frac{y-y_0}{y_1-y_0}$$

represent the normalized distances of the point (x,y) along the x- and y-axes, respectively. The *reproject_to* function available in SunPy [24] was used for this step. An example of the re-projection is shown in the top two panels in the middle column of Figure 1. Finally, similarly to AIA images, the HMI images also needed to be corrected for elliptical orbit variation and fixed to a solar disk size of 976 arcsecs throughout the database. This step is shown with an example in the bottom panel in the middle column of Figure 1. The panels in the right column of this image show that after our processing, the solar disk in the AIA and HMI images are well aligned.

2.1.3 Temporal Alignment of HMI and AIA Data

In our database, we have decided to keep the temporal resolution of the time series at 12 minutes. This is because the low-noise LOS and vector magnetograms are available with this cadence from HMI. For each of the timestamps in `hmi.M_720s` series, we find the corresponding AIA data at that time and add it to our database. If the quality of the AIA data for any of the eight wavelengths is not good at an exact timestamp of HMI, we search for a time within two minutes of the timestamp that contains good quality AIA data in all eight EUV channels and include that data in our database. If we are unable to find good-quality AIA data within the two-minute range, we do not include the data in our database. We consider the quality of the AIA data to be good if the QUALITY flag in the file header is equal to zero. Non-zero values for the QUALITY flag indicate different operational items (e.g., off-pointing or defocusing of the satellite) lowering the quality or significant missing data. In Figure 5 (see the Supplementary Information), we give examples of when the AIA data was found to be of bad quality.

2.2 Application Benchmark Datasets

In this section, we will describe our application benchmark datasets in detail. As shown in Table 2, multiple space weather and heliophysics applications are broadly oriented towards understanding the deposition of energy in the solar surface and atmosphere (DS1 & DS2; Sec. 2.2.1 & 2.2.2), how solar magnetism structures the solar atmosphere (DS3; Sec. 2.2.3), how energy is released in the form of space weather (DS4; Sec. 2.2.4), and how space weather affects the solar magnetosphere (DS5; Sec. 2.2.5) and ionosphere (DS6; Sec. 2.2.6). We note that we created baseline models and evaluated them. Our model architectures, formulated problems, evaluation measures, and results can be found in the Supplementary Information document.

2.2.1 Active region segmentation

Active regions (ARs) are often identified using intensity thresholding on magnetograms, white-light or EUV images [25, 26, 27]. In this work, ARs containing polarity inversion lines (PILs) are identified using full-disk line-of-sight (LoS)

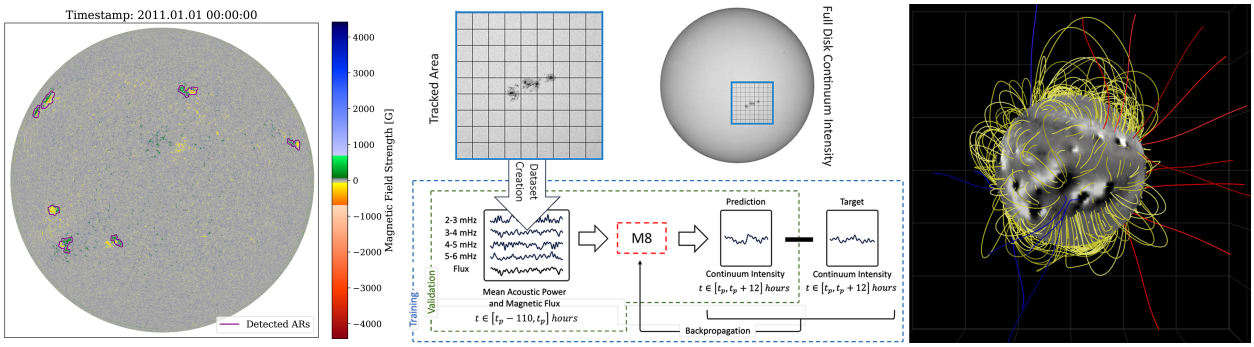


Figure 3. ***Left:*** Example of an AR mask for segmentation. The magnetic field strength from the line-of-sight magnetogram on 2011-01-01 00:00:00 is shown using a blue-to-red color scale. Detected active regions, which have magnetic flux greater than 50 Gauss or less than –50 Gauss and contain polarity inversion lines, are highlighted with purple contours. ***Middle:*** Example of AR intensity time-series for emergence forecast. Tracked regions are split into smaller tiles, and the timeline datasets were created by averaging the values of each tile. The timelines are used as inputs during the training and validation/testing. ***Right:*** Example coronal magnetic field extrapolated from measurements using WSA. The yellow magnetic field lines indicate closed field lines that loop back to the surface of the sun, which the red/blue lines are positive/negative polarity open field lines which extend past the source surface into interplanetary space.

magnetograms from the Solar Dynamics Observatory (SDO). Our previous PIL detection method by [12, 28] relies on AR patches. Extending this, our approach uses full-disk LoS magnetogram rasters with a resolution of 4096×4096. We first generate two binary maps, corresponding to the positive and negative polarity regions, by applying a magnetic field strength threshold of $+50$ and -50 Gauss, respectively. To remove small, noisy patches in positive and negative polarity region maps, we apply a size filter that excludes regions smaller than 100 pixels (approximately 13.3 Mm^2 of photospheric area). Next, we dilate the binary images using a rectangular filter of size 10 pixels. Finally, we identify the intersection of the dilated positive and negative polarity regions, which corresponds to areas containing PILs. Only ARs that include PILs are reported as regions of interest. The eventual AR masks are 2D bitmaps (containing zeros and ones), representing the locations of active regions with PILs, and have a size of 4096×4096. In other words, ARs without the presence of a strong PILs will be omitted from the full-disk mask. In Figure 3-left, the binary mask is overlaid on the original line-of-sight magnetogram from 2011-01-01 00:00:00. The regions outlined in purple indicate the detected active regions.

The ARPIL dataset covers between January 2011 to December 2024. The detection method is applied to each valid LoS magnetogram hourly. In total, we have 119,454 full-disk AR binary masks covering 14 years, each of them with a size of 4096×4096.

2.2.2 Active region emergence forecast

We select 50 ARs that appear on the solar surface within 30 degrees longitude from the central meridian between March 1st, 2010 and June 1st, 2023, persisted for more than 4 days, and reached a total area of 200 millionths of the solar hemisphere. For each one of these 50 ARs, the same five-step pipeline is followed: (1) tracking areas of 512 by 512 pixels of the SDO/HMI magnetic flux, Doppler velocity, and continuum intensity, (2) creating acoustic power maps from the Doppler velocity, (3) downsampling the data to timelines by splitting the tracked region in a 9 by 9 grid, (4) removing the solar sphere geometric effects, and lastly (5) calculating the time evolution of continuum intensity.

Before we create the acoustic maps, we use the dopplergrams series, representing the frames throughout the life of the AR on the solar disk, and create a difference series by subtracting consecutive frames. By working with dopplergram differences, we remove the background solar rotational signal:

$$\Delta V_{\text{dop}}[i,x,y] = V_{\text{dop}}[i+1,x,y] - V_{\text{dop}}[i,x,y], \quad \text{for } i = 1,\ldots,639, \quad (x,y) \in [1,512]^2. \tag{1}$$

Each element of ΔV_{dop} represents the difference between consecutive dopplergrams at each pixel location (x,y). Subsequently, for each pixel, we calculate the Fourier power spectrum of the time-series data in $\Delta V_{\text{dop}}[:,x,y]$. Let $dt = 45$ sec (the sampling interval), $T = 28800$ sec (each .fits file is tracks the active region for 8 hours), and $\mathscr{F}$ denote the real-valued one-sided Fourier transform (*np.fft.rfft* in Python):

$$V_{\text{dop}}^{\text{FFT}}[k,x,y] = \left(\frac{dt^2}{T}\right) \left|\mathscr{F}\{\Delta V_{\text{dop}}[:,x,y]\}[k]\right|^2, \quad \text{for } k = 1,\ldots,320, \quad (x,y) \in [1,512]^2. \tag{2}$$

We calculate the Fourier power spectrum of the time series data. For every timeframe, we calculate the integral along the frequency axis to construct a power map that represents the spectral power in the chosen frequency range at each spatial location on the solar disk. Given a power map series, which represents the temporal evolution of the spatial power distribution for a particular frequency range (e.g., $2-3mHz$) over the solar disk, we divide the solar disk into a grid of smaller, equally-sized tiles. For each of these tiles, we extract the corresponding pixels from all frames of the power map series.

Subsequently, we calculate the mean power within each tile for each frame. The temporal mean power within a tile is calculated by taking the average power over all pixels within the tile for each frame. This procedure results in a one-dimensional time series per pixel, representing the temporal evolution of power within each segment on the solar disk. In parallel, we calculate the total continuum intensity, as well as total magnetic flux for each tile to produce labeled acoustic power time series and magnetic and continuum intensities. The combination of the acoustic power timeseries and continuum intensity time series (as shown in Fig. 3-middle) form the core of this dataset.

Threfore, this dataset includes timeseries for 50 emerging ARs. Each has 6 channels (4 acoustic power channels, magnetic flux, and continuum intensity). Each time series has 240 timestamps. The dynamic range goes between -7.5×10^7 and 5.8×10^7 for the acoustic power channels, -1.4×10^2 to 5.3×10^2 for magnetic flux, and -1.7×10^4 and 4.0×10^3 for continuum intensity. More information about the AR emergence dataset can be found in [29].

2.2.3 Coronal field extrapolation

To model the 3D structure of the coronal magnetic field, it is necessary to first estimate a full coverage (180° latitude, 360° longitude), and subsequently model the transformation of a solar surface boundary condition into a magnetic field that extends into the atmosphere. To do this, we use a coupled simulation known as ADAPT-WSA.

For the first task, we use the Air Force Data Assimilative Photospheric Flux Transport (ADAPT; [30, 31, 32]; [33]), which is a data assimilation model that uses near-side photospheric magnetic field measurements from such instruments as HMI and globally solves a system of magnetic flux transport equations [34]. These equations describe the time-dependent evolution of the photospheric magnetic field, including such effects as differential rotation, and meridional and supergranular flows. ADAPT is an ensemble model which generates 12 realizations (variations) per timestep. Each realization represents processing using different values of unobservable subsurface physics in the ADAPT simulation, which is itself an ensemble Kalman filter.

For the second task, we use the Wang–Sheeley–Arge (WSA; [35, 36, 37]) model, which makes it possible to calculate the global coronal field using a coupled Potential Field Source Surface and Potential Field Current Sheet (PFCS) approaches [38, 39]. In practice, the WSA model solves the equations $\nabla\times\mathbf{B}=\mathbf{0}$ and $\nabla\cdot\mathbf{B}=0$, where $\mathbf{B}$ is the coronal magnetic field, up to a spherical boundary known as "the source surface" and where the field becomes radial, set here at 2.51 $R_\odot$ (solar radii). The coronal field extrapolations are themselves encoded using the spherical harmonics.

This dataset includes an 11-year (full solar cycle) span of ADAPT-WSA runs powered by the HMI magnetogram data at daily cadence. Daily cadence is chosen to provide a sufficient diversity of magnetic topologies in the training dataset and avoid nearly identical training samples. The ADAPT-WSA runs are split into an ensemble composed of 12 realizations per the ADAPT ensemble. The potential field solution at each timestep is encoded in signed spherical harmonic coefficients, normalized using the Schmidt formulation and truncated after the 90th order. An example magnetic field solution is displayed in Fig. 3-right, where the plotted field lines were traced using the spherical harmonics to evaluate $\mathbf{B}$ at each position.

This dataset includes 51,156 sets of harmonic coefficients, each has 2 channels (G and H coefficients), and contains 4,186 harmonic coefficients. The dynamic range goes between $-4.3x10^3$ to $4.3x10^3$.

2.2.4 Flare forecasting

Although the volume of observational data has significantly increased, accurate operational prediction solar flares remains a challenging task. Solar flares are monitored by the Geostationary Operational Environmental Satellites (GOES), measuring the X-ray intensity emitted by the Sun. The National Oceanic and Atmospheric Administration (NOAA) classifies solar flares logarithmically into five major classes –A, B, C, M, and X, based on their peak X-ray intensity in the 1–8Å wavelength range [40]. The strength of a flare within a class is indicated by a numerical suffix ranging from 1.0 to 9.9, which represents the factor by which the event is stronger than the base intensity in that class (e.g., M5.2 is 5.2 times as strong as M1.0). Flares above C-class, particularly M- and X-class flares, are of primary interest due to their significant terrestrial impact, yet the scarcity of stronger events pose a substantial class imbalance challenge.

In this dataset, the input instance at time t_i is associated with a prediction window spanning from t_i to t_i+24h. Each window may contain zero or more solar flares. Note that we use the start time of the flares to determine if they are within

a prediction window. Only flares greater than C-class are considered in this application due to the under-reporting of lower intensity flares. Each input, sampled at an hourly cadence, is labeled in two ways: (1) by the *maximum flare intensity*, as defined in Eq. 2, and (2) by the *cumulative flare intensity*, as defined in Eq. 3. Maximum flare intensity is the label corresponding to the flare with the highest intensity occurring within the prediction window.

$$L_{max}(t_i) = class(\max_{f_j \in F_{t_i}^{C+}} \mathbf{pxf}(f_j)) \tag{3}$$

, where $class(\cdot)$ returns the GOES class corresponding to the peak X-ray flux for the maximum intensity flare. In Figure 4-top, we demonstrate an example prediction window covering four flares shown. This instance is labeled as 'M3.5', which is the flare with the maximum intensity.

Cumulative flare intensity considers the cumulative effect of all the ≥C-class flares in the prediction window. As mentioned earlier, flare sub-classes (e.g., C**5.2**, M**1.0**) are indicated by a numerical suffix ranging between 1.0 and 9.9. To create the the cumulative intensity label, we get the weighted sum of these numerical suffixes/subclass values, as described in Eqs. 3 and 4.

$$S(f_j) = \begin{cases} 0 & \text{if } \mathbf{pxf}(f_j) < 10^{-6} \\ 1 & \text{if } 10^{-6} < \mathbf{pxf}(f_j) \leq 10^{-5} \\ 10 & \text{if } 10^{-5} < \mathbf{pxf}(f_j) \leq 10^{-4} \\ 100 & \text{if } \mathbf{pxf}(f_j) > 10^{-4} \end{cases}, \tag{4}$$

where $S(f_j)$ returns the weight for the flare event f_j. In other words, to differentiate the contribution of C-, M-, and X-class flares, weights of 1, 10, and 100 are applied to their respective subclass values. The weighted sum $L_{cum}(t)$ is then calculated as:

$$L_{cum}(t_i) = \sum_{f_j \in \mathscr{F}_{t_i}^{C+}} S(f_j) \cdot v(f_j), \tag{5}$$

where $v(f)$ denotes the subclass value for f_j. For example, in Figure 4, four subclass values from four flares within the prediction window (with blue background) are considered. The cumulative flare intensity is 50.2, calculated as $2.2\ (\text{C2.2}) + 10 \times 3.5\ (\text{M3.5}) + 7.7\ (\text{C7.7}) + 5.3\ (\text{C5.3})$.

For creating labels for binary classification, we use two thresholds for L_{max} and L_{cum} corresponding to the equivalent strength of an M1.0-class flare. In other words, we create two binary labels checking (1) $L_{max} > M1.0$ and $L_{cum} > 10$. The flare forecasting labels span from May 2010 to December 2024. There are total 128,352 labels in the dataset.

2.2.5 Solar wind forecasting

The solar wind is a stream of charged particles that emanates from the Sun. The interaction of solar wind with Earth's magnetic field results in the formation of near-Earth space weather. Solar wind interactions are known to drive geomagnetic storms, wherein the Earth's magnetic field is perturbed, inducing electrical currents that affect satellites, power grids, oil pipelines, etc., and potentially resulting in economic and livelihood impacts [1, 16]. The solar wind is known to have origins from expansive regions very low in the solar atmosphere, undergoing acceleration as it propagates outwards through the chromosphere, transition region, and corona [15], and then becoming free streaming in the interplanetary medium. In a more global sense, the solar wind shows statistical associations with morphological structures in the solar atmosphere.

For this application benchmark dataset, we use space-based particle data, measured by the Advanced Composition Explorer (ACE; [41]), which are made available through the NASA OMNIWEB database [1]. The OMNI data consist of solar wind speed measurements at the L1 point of the Sun-Earth system in space, and are time-shifted to be at the nose of the Earth's bow shock. Hence, we essentially have a scalar measurement across all time, resulting in a vector of measurements. These data spans 14 years from 2010 to 2023, at a time cadence of 1 hour. The time series of the full dataset is shown in Fig. 4-bottom-left. The solar wind speed, as seen in Figure 4-bottom-right, shows a form of long tailed distribution, with a peak at ≈ 400 km/s. This dataset includes 120,748 measurements of the solar wind speed. The dynamic range goes between 2.4×10^2 to 8.8×10^2 km/s.

[1] https://omniweb.gsfc.nasa.gov/

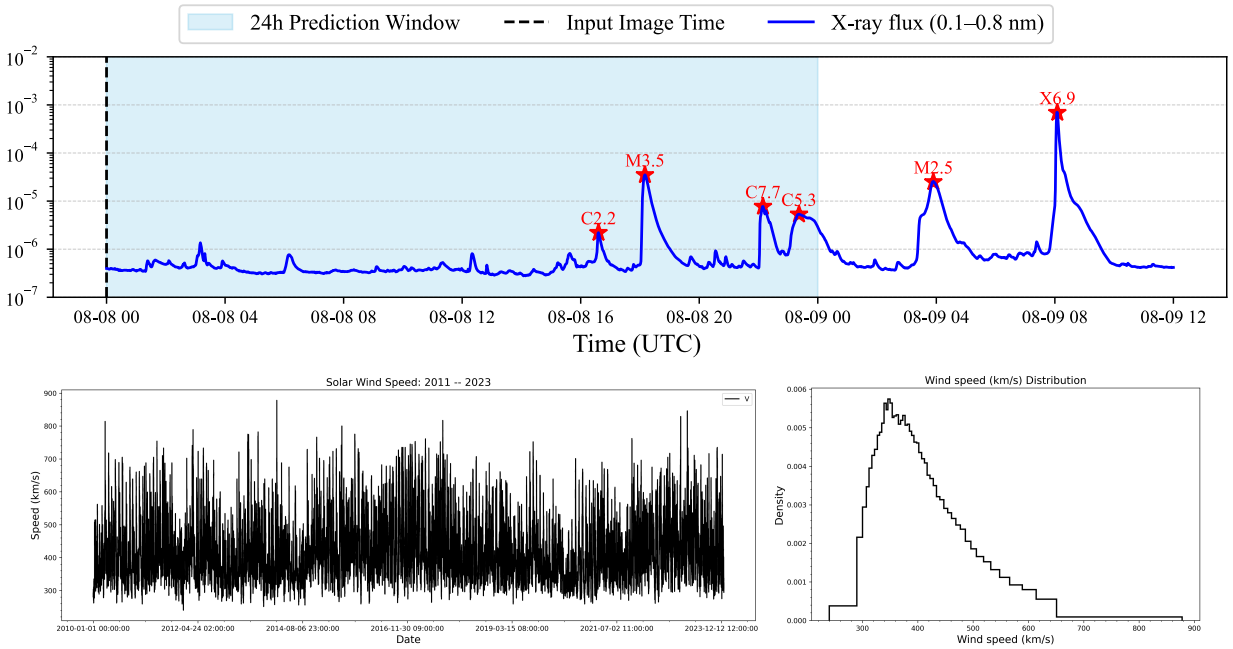


Figure 4. ***Top:*** Labeling inputs with the flare index. The 24-hour prediction window may include multiple flares. Two types of flare indices are defined: the maximum flare intensity and the cumulative flare intensity. Inputs are sampled at an hourly cadence, and the prediction window shifts accordingly based on the input time. ***Bottom-Left:*** The solar wind speed measurements at L1 as a time series. ***Bottom-Right:*** Distribution of solar wind speed measurements.

2.2.6 Solar EUV spectra modeling

The Extreme Ultraviolet Variability Experiment (EVE; [42]) aboard NASA's Solar Dynamics Observatory (SDO) was developed to provide high-resolution, high-cadence measurements of full-disk solar EUV irradiance. In this dataset, we focus on the high energy part of the EUV, captured by a submodule of EVE called MEGS-A, which has a 10-second cadence [43]. We apply metadata-based screening using the quality flags provided in the EVE Level 2 data products [43]. The `SC_FLAGS` byte identifies potential obstructions or pointing issues during observation. Only spectra flagged as `0` (clear, unobstructed) are retained. We compute 1-minute averaged EVE spectra to reduce noise and facilitate temporal alignment with AIA image data. We then match the 12 minute cadence of the SDO/AIA image cubes. Timestamps in which AIA frames affected by saturation (common during large flares), diffraction patterns, and instrument exposure anomalies are flagged and excluded. The resulting dataset contains hundreds of thousands of temporally aligned AIA image cubes and EUV spectra, covering Solar Cycle 24 and parts of Solar Cycle 25, and includes both quiet-Sun and active-region conditions. All irradiance values are first corrected to 1 astronomical unit (AU) to remove the influence of Earth–Sun distance variations.

Event selection and temporal stratification were performed to construct a balanced and representative training set spanning a wide range of solar conditions. For active periods, we rely on the GOES X-ray Sensor (XRS; [44]) flare catalog to identify flaring events. These events are binned according to the integrated soft X-ray flux and stratified into percentiles, rather than raw flare class labels, to ensure a more uniform representation of flare energetics and avoid over-representation of weaker, more frequent flares. For quiet Sun conditions, where changes in irradiance are dominated by solar rotation and large-scale structural evolution, data are sampled at regular 1-day intervals to capture the modulations introduced by active region transit across the solar disk. This dual strategy ensures adequate exposure to both high-energy transient events and slowly varying background structures.

This dataset includes 189,397 EVE spectra, each with 1343 spectral channels. The dynamic range goes between 1.0×10^{-9} to 1.1×10^{-2}.

2.3 Summary

SuryaBench captures diverse solar phenomena across a full solar cycle, with high-resolution multi-instrument observations and rigorous standardization. By integrating data from AIA and HMI with consistent preprocessing into

Table 2. Statistic summary of auxiliary datasets. The shape of all datasets has been cast with the following dimensions: N → Number of datapoints, C → Number of channels, T → Number of timestamps, H → height, and W → width.

	Application	Broader Topic	N	C	T	H	W	Dynamic Range
DS1	AR segmentation	Magnetic energy in solar atmosphere	109,175	1	1	4,096	4,096	0,1
DS2	AR emergence forecasting	Magnetic energy in solar atmosphere	50	6	240	1	1	-1.7×10^4 to 4.0×10^3
DS3	Coronal field extrapolation	Coronal structure & magnetism	51,156	2	1	4,186	1	-4.3×10^3 to 4.3×10^3
DS4	Flare forecasting	Space weather forecasting	128,352	1	1	1	1	0,1
DS5	Solar wind forecasting	Solar forcing of magnetosphere	120,748	1	1	1	1	2.4×10^2 to 8.8×10^2
DS6	EUV forecasting	Solar forcing of ionosphere	189,397	1,343	1	1	1	1.0×10^{-9} to 1.1×10^{-2}

a unified ML-ready format, the dataset provides a high-fidelity view of solar activity. This uniformity enhances data quality and reproducibility, enabling cross-comparison of events (flares, active region evolution, coronal dynamics) and cultivating deeper insight into the dynamics of solar activity and space weather. We believe that the inclusion of curated benchmarks and baseline model results for tasks such as solar flare prediction, coronal field extrapolation, and active region segmentation underscores SuryaBench's value to the machine learning community, and we envision that these application benchmarks will establish clear performance baselines and spurring the development of advanced models. The dataset breadth and ML-focused design bridge the heliophysics and AI, accelerating progress in space weather predictive modeling.

3 Data Records

The SuryaBench datasets contain ML-ready heliophysics data captured from May 13, 2010, to December 31, 2024, with a 12-minute cadence. The datasets (both core and application benchmark datasets) are publicly available on Huggingface as a data collection Suryabench. During this collection interval, there are about 6% data is missing due to either unavailability or poor quality. The processed level-1.5 AIA and HMI data are stored in hourly netCDF files in `float32` format, with data shape of [13, 4096, 4096]. Each netCDF file is about 600 MB, and the total size of the data for training is approximately 360 TB. We have divided the data into training (2010-2018), validation (2019), and test (2020) sets, which include 379,920, 43,680, and 43,800 files, respectively.

4 Technical Validation

We used the SuryaBench datasets to validate against state-of-the-art models commonly used by the machine learning and heliophysics communities. All experiments were performed on 4 Nvidia A100 GPUs with 80 GB of memory. This evaluation helps establish reference points for future research by comparing performance across widely adopted architectures. To create the baseline on the SDO dataset, we framed it as a forecasting problem. Given two input tasks we predicted the next time step. By training on 4 years' worth of data, the modified long-short Spectral Transformer [4] model demonstrated strong performance after training for just 20 epochs on four years of data. For the AIA bands, the model achieved high structural similarity index (SSIM) scores of 0.83 for band 171Å, 0.90 for band 193Å, and 0.86 for band 211Å, while the remaining bands showed SSIM values ranging from 0.4 to 0.65. The corresponding root mean squared error (RMSE) values were 0.11 for band 171Å, 0.095 for band 193Å, and 0.10 for band 211Å. When applied to the HMI channel, the model achieved an SSIM of 0.73 and an RMSE of 0.65. These results indicate the model's ability to accurately reproduce both large-scale and fine-scale solar features, including active regions with higher magnetic field strengths.

The tables below summarize baseline results for two core tasks: solar wind forecasting and binary solar flare prediction. Solar wind forecasting performance is reported using RMSE, MAE, and validation loss for ResNet and U-Net based encoder-decoder models. For flare prediction (classification task), we evaluate models, including AlexNet [45], MobileNet [46], and ResNet [47] variants, using popular forecast skill scores True Skill Statistic (TSS),

Heidke Skill Score (HSS), Composite Skill Score (CSS), and F1-macro, similar to [48]. These baselines provide a standardized performance floor for advancing heliophysics AI.

Table 3. Baseline performance for (a) solar wind prediction and (b) solar flare classification using common deep learning models on test data

Model	RMSE	MAE	Val Loss
UNet	0.1499	0.1116	0.0225
AttentionUNet	0.1449	0.1157	0.0225
ResNet18	0.2108	0.2388	0.0233
ResNet34	0.1462	0.1149	0.0226
ResNet50	0.1445	0.1145	0.0221

(a) Solar Wind Forecasting

Model	TSS	HSS	CSS	F1
AlexNet	0.359	0.354	0.356	0.679
MobileNet	0.326	0.312	0.319	0.662
ResNet18	0.320	0.317	0.318	0.660
ResNet34	0.290	0.289	0.289	0.645
ResNet50	0.261	0.281	0.271	0.627

(b) Solar Flare Prediction

Table 4. Baseline performance for (a) EVE Prediction and (b) AR Emergence Forecasting on test data

Model	RMSE	MAE	Val Loss
UNet	0.0754	0.0558	0.00569
AttentionUNet	0.0754	0.0558	0.00569
ResNet18	0.2108	0.2388	0.02330
ResNet34	0.1462	0.1149	0.02255
ResNet50	0.1445	0.1145	0.02208

(a) EVE Prediction

Model	MSE	RMSE
ST Attention	0.1538	0.3921
ST ResNet	0.1527	0.3908
(LSTM)	0.0140	0.1180

(b) AR Emergence Forecasting (ST: Spatiotemporal)

5 Limitations

Solar data have a few important features that must be taken into consideration:

- ***Solar rotation:*** The Sun takes ≈ 27 days to complete rotation, also referred to as a Carrington rotation. This results in a repeat of magnetic structures every ≈ 27 days [49]. It is preferable to separate the training and testing sets by at least 1/2 Carrington rotation to avoid observing the repeating spatial patterns. The standard practice in heliophysics is to use temporally non-overlapping training-validation-testing splits [50] (e.g., first 5 years for training, next 3 years for validation, and remaining years for testing, or the first 8 months of each year for training (Jan-Aug), the next two months (Sep-Oct) for validation, and the last two (Nov-Dec) for testing.)
- ***Solar Cycle:*** Solar activity undergoes an ≈ 11 year maximum and minimum. This results in a subtle, 11-year variation in the solar activity [51]. Hence, it is ideal to perform data splitting by sampling activity across the whole solar cycle. The standard practice in heliophysics is to maximize dataset coverage to contain at least a full solar cycle (e.g., 2010-2022). This in combination with the split mentioned above, ensures the creation of representative training-validation-test splits.
- ***Ecliptic angle:*** The plane of the Earth's orbit is inclined with respect to the equator. Because of this, as the year progresses, the Earth goes slightly above (below) the north (south) pole. While our data is fixed so that the solar north is always pointing upwards, the solar disk center is almost never on the equator. Any application aiming to use heliographic coordiantes (i.e. latitude and longitude on the solar surface), must account for this perspective effect of the Earth's orbit.
- ***Lack of farside observations:*** While SDO provides one of the most comprehensive datasets, the observations are limited to the visible (Earth) side of the Sun. Many of the applications using this dataset can be directly impacted by events occurring on the farside of the Sun.

6 Code Availability

The SuryaBench datasets are publicly available on Huggingface: https://huggingface.co/collections/nasa-impact/suryabench-68265ce306fc2470c121af7b. Our code for dataset preparation and creation, and baseline model training is publicly available at https://github.com/NASA-IMPACT/SuryaBench.

References

1. Schrijver, C. J., Dobbins, R., Murtagh, W. & Petrinec, S. M. Assessing the impact of space weather on the electric power grid based on insurance claims for industrial electrical equipment. Space Weather. **12**, 487–498, 10.1002/2014SW001066 (2014). 1406.7024.

2. Pesnell, W. D., Thompson, B. J. & Chamberlin, P. C. The Solar Dynamics Observatory (SDO). Sol. Phys. **275**, 3–15, 10.1007/s11207-011-9841-3 (2012).

3. Asensio Ramos, A., Cheung, M. C. M., Chifu, I. & Gafeira, R. Machine learning in solar physics. Living Rev. Sol. Phys. **20**, 4, 10.1007/s41116-023-00038-x (2023). 2306.15308.

4. Roy, S. et al. Ai foundation model for heliophysics: Applications, design, and implementation. arXiv preprint arXiv:2410.10841 (2024).

5. Galvez, R. et al. A machine-learning data set prepared from the nasa solar dynamics observatory mission. The Astrophys. J. Suppl. Ser. **242**, 7 (2019).

6. Poduval, B. et al. Ai-ready data in space science and solar physics: problems, mitigation and action plan. Front. Astron. Space Sci. **10**, 1203598 (2023).

7. van Driel-Gesztelyi, L. & Green, L. M. Evolution of active regions. Living Rev. Sol. Phys. **12**, 1, 10.1007/lrsp-2015-1 (2015). Open Access Review Article.

8. Schrijver, C. J. A characteristic magnetic field pattern associated with all major solar flares. Astrophys. J. Lett. **655**, L117–L120, 10.1086/511857 (2007). Provided by the SAO/NASA Astrophysics Data System.

9. Toriumi, S. & Takasao, S. Numerical simulations of flare-productive active regions: δ-sunspots, sheared polarity inversion lines, energy storage, and predictions. The Astrophys. J. **850**, 39, 10.3847/1538-4357/aa95c2 (2017).

10. Wang, J. et al. Solar flare predictive features derived from polarity inversion line masks in active regions using an unsupervised machine learning algorithm. The Astrophys. J. **892**, 140, 10.3847/1538-4357/ab7b6c (2020).

11. Cicogna, D. et al. Flare-forecasting algorithms based on high-gradient polarity inversion lines in active regions. The Astrophys. J. **915**, 38, 10.3847/1538-4357/abfafb (2021).

12. Ji, A. et al. A systematic magnetic polarity inversion line data set from sdo/hmi magnetograms. The Astrophys. J. Suppl. Ser. **265**, 28 (2023).

13. Zhang, J., Dere, K. P., Howard, R. A., Kundu, M. R. & White, S. M. On the temporal relationship between coronal mass ejections and flares. The Astrophys. J. **559**, 452–462, 10.1086/322405 (2001).

14. Yasyukevich, Y. et al. The 6 september 2017 x-class solar flares and their impacts on the ionosphere, gnss, and hf radio wave propagation. Space Weather. **16**, 1013–1027 (2018).

15. Cranmer, S. R. & Winebarger, A. R. The properties of the solar corona and its connection to the solar wind. Annu. Rev. Astron. Astrophys. **57**, 157–187 (2019).

16. Oughton, E. J., Skelton, A., Horne, R. B., Thomson, A. W. P. & Gaunt, C. T. Quantifying the daily economic impact of extreme space weather due to failure in electricity transmission infrastructure. Space Weather. **15**, 65–83, 10.1002/2016SW001491 (2017).

17. Woods, T. N. et al. Solar EUV Experiment (SEE): Mission overview and first results. J. Geophys. Res. (Space Physics) **110**, A01312, 10.1029/2004JA010765 (2005).

18. Qian, L., Burns, A. G., Chamberlin, P. C. & Solomon, S. C. Variability of thermosphere and ionosphere responses to solar flares. J. Geophys. Res. (Space Physics) **116**, A10309, 10.1029/2011JA016777 (2011).

19. Goncharenko, L. P. et al. A New Model for Ionospheric Total Electron Content: The Impact of Solar Flux Proxies and Indices. J. Geophys. Res. (Space Physics) **126**, e28466, 10.1029/2020JA02846610.1002/essoar.10503730.1 (2021).

20. Buzulukova, N. & Tsurutani, B. Space Weather: From Solar Origins to Risks and Hazards Evolving in Time. Front. Astron. Space Sci. **9**, 429, 10.3389/fspas.2022.1017103 (2022). 2212.11504.

21. Lemen, J. R. et al. The Atmospheric Imaging Assembly (AIA) on the Solar Dynamics Observatory (SDO). Sol. Phys. **275**, 17–40, 10.1007/s11207-011-9776-8 (2012).

22. Scherrer, P. H. et al. The Helioseismic and Magnetic Imager (HMI) Investigation for the Solar Dynamics Observatory (SDO). Sol. Phys. **275**, 207–227, 10.1007/s11207-011-9834-2 (2012).

23. Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. Numerical Recipes in FORTRAN 77: The Art of Scientific Computing (Cambridge University Press, 1992), 2 edn.

24. The SunPy Community et al. The sunpy project: Open source development and status of the version 1.0 core package. The Astrophys. J. **890**, 68–, 10.3847/1538-4357/ab4f7a (2020).

25. Turmon, M., Pap, J. M. & Mukhtar, S. Automated recognition of solar active-region properties. Sol. Phys. **260**, 347–363, 10.1007/s11207-009-9440-8 (2010). Provided by the SAO/NASA Astrophysics Data System.

26. Verbeeck, C., Delouille, V., Mampaey, B. & De Visscher, R. The spoca-suite: Software for extraction, characterization, and tracking of active regions and coronal holes on euv images. Astron. & Astrophys. **561**, A29, 10.1051/0004-6361/201321243 (2014). Provided by the SAO/NASA Astrophysics Data System.

27. Caballero, C. & Aranda, M. C. A comparative study of clustering methods for active region detection in solar euv images. Sol. Phys. **283**, 691–717, 10.1007/s11207-013-0239-2 (2013).

28. Cai, X., Aydin, B., Ji, A., Georgoulis, M. K. & Angryk, R. A framework for detecting polarity inversion lines from longitudinal magnetograms. In 2020 IEEE International Conference on Big Data (Big Data), 4175–4183 (IEEE, 2020).

29. Kasapis, S., Kitiashvili, I. N., Kosovichev, A. G., Stefan, J. T. & Apte, B. Predicting the emergence of solar active regions using machine learning. Proc. Int. Astron. Union **19**, 311–319 (2023).

30. Arge, C. N. et al. Air force data assimilative photospheric flux transport (adapt) model. In AIP conference proceedings, vol. 1216, 343–346 (American Institute of Physics, 2010).

31. Arge, C. N. et al. Improving data drivers for coronal and solar wind models. In 5th international conference of numerical modeling of space plasma flows (astronum 2010), vol. 444, 99 (2011).

32. Arge, C. N. et al. Modeling the corona and solar wind using adapt maps that include far-side observations. In AIP conference proceedings, vol. 1539, 11–14 (American Institute of Physics, 2013).

33. Hickmann, K. S., Godinez, H. C., Henney, C. J. & Arge, C. N. Data assimilation in the adapt photospheric flux transport model. Sol. Phys. **290**, 1105–1118 (2015).

34. Worden, J. & Harvey, J. An evolving synoptic magnetic flux map and implications for the distribution of photospheric magnetic flux. Sol. Phys. **195**, 247–268 (2000).

35. Arge, C. & Pizzo, V. Improvement in the prediction of solar wind conditions using near-real time solar magnetic field updates. J. Geophys. Res. Space Phys. **105**, 10465–10479 (2000).

36. Arge, C. N., Odstrcil, D., Pizzo, V. J. & Mayer, L. R. Improved method for specifying solar wind speed near the sun. In AIP conference proceedings, vol. 679, 190–193 (American Institute of Physics, 2003).

37. McGregor, S., Hughes, W., Arge, C. & Owens, M. Analysis of the magnetic field discontinuity at the potential field source surface and schatten current sheet interface in the wang–sheeley–arge model. J. Geophys. Res. Space Phys. **113** (2008).

38. Schatten, K. H., Wilcox, J. M. & Ness, N. F. A model of interplanetary and coronal magnetic fields. Sol. Phys. **6**, 442–455 (1969).

39. Wang, Y.-M. & Sheeley Jr, N. On potential field models of the solar corona. Astrophys. Journal, Part 1 (ISSN 0004-637X), vol. 392, no. 1, June 10, 1992, p. 310-319. Res. supported by US Navy. **392**, 310–319 (1992).

40. Fletcher, L. et al. An observational overview of solar flares. Space science reviews **159**, 19–106 (2011).

41. Gloeckler, G. M. Ace solar wind ion composition spectrometer (swics) solar wind plasma elemental and isotopic density, speed, thermal speed, charge state, and ratio parameters, level 2 (l2), 1 h data (2023).

42. Woods, T. N. et al. Extreme Ultraviolet Variability Experiment (EVE) on the Solar Dynamics Observatory (SDO): Overview of Science Objectives, Instrument Design, Data Products, and Model Developments. Sol. Phys. **275**, 115–143, 10.1007/s11207-009-9487-6 (2012).

43. Woodraska, D. & Eparvier, F. G. SDO/EVE Level 2B Version 8 Data Product Documentation. `LASP / University of Colorado Boulder Technical Document` (2024). Available at https://lasp.colorado.edu/eve/data_access/eve_data/products/level2b/EVE_L2B_V8_README.pdf.

44. Chamberlin, P. C., Woods, T. N., Eparvier, F. G. & Jones, A. R. Next generation x-ray sensor (XRS) for the NOAA GOES-R satellite series. In Fineschi, S. & Fennelly, J. A. (eds.) Solar Physics and Space Weather Instrumentation III, vol. 7438 of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, 743802, 10.1117/12.826807 (2009).

45. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. Adv. neural information processing systems **25** (2012).

46. Howard, A. G. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861 (2017).

47. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, 770–778 (2016).

48. Pandey, C., Adeyeha, T., Hong, J., Angryk, R. A. & Aydin, B. Advancing Solar Flare Prediction Using Deep Learning with Active Region Patches, 50–65 (Springer Nature Switzerland, 2024).

49. Owens, M. J., Challen, R., Methven, J., Henley, E. & Jackson, D. A 27 day persistence model of near-earth solar wind conditions: A long lead-time forecast and a benchmark for dynamical models. Space Weather. **11**, 225–236 (2013).

50. Pandey, C., Angryk, R. A. & Aydin, B. Solar flare forecasting with deep neural networks using compressed full-disk hmi magnetograms. In 2021 IEEE International Conference on Big Data (Big Data), 1725–1730, 10.1109/bigdata52589.2021.9671322 (IEEE, 2021).

51. Schwenn, R. Solar wind sources and their variations over the solar cycle. Sol. dynamics its effects on heliosphere Earth 51–76 (2007).

7 Author Contributions

Sujit Roy: Conceptualization, Methodology, Data production, Visualization, Writing–original draft, Writing–review & editing, Project administration.
Johannes Schmude: Conceptualization, Methodology, Baseline design, Data Validation.
Vishal Gaur: Methodology, Data production, Visualization, Writing–original draft, Baseline design. Rohit Lal: Methodology, Data production, Visualization, Writing–original draft, Baseline design.
Dinesha V. Hegde: Data preprocessing and production (SDO ML Ready), Visualization, Writing–original draft, Writing–review & editing.
Amy Lin: Data production, Visualization.
Talwinder Singh: Data production (Active region segmentation and flare forecasting), Visualization.
Berkay Aydin: Data production (Active region segmentation and flare forecasting), Writing–editing original draft.
Andrés Muñoz-Jaramillo: Data production coordination for application benchmark datasets, Writing–original draft.
Vishal Upendran: Data production (solar wind), editing - original draft
Daniel da Silva: Data production (Coronal field extrapolation), editing-original draft.
Shah Bahaudding: Data production (Solar EUV spectra).
Spiridon Kasapis: Data production (Active Region Emergence forecast) Kang Yang: Baseline design for AR segmentation.
Chetraj Pandey: Baseline design for flare forecasting.
Iksha Gurung: Data Hosting
Jinsu Hong: Baseline design and experiments for flare forecasting and AR segmentation.
Nikolai Pogorelov: Validation & Review
Manil Maskey: Validation & Review
Rahul Ramachandran: Validation, Review & Project administration.

8 Competing Interests

The authors declare no competing interests.

9 Acknowledgements

This work is supported by NASA Grant 80MSFC22M004. The Authors acknowledge the National Artificial Intelligence Research Resource (NAIRR) Pilot and NVIDIA for providing support under grant no. NAIRR240178.

Supplementary Information

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 5.

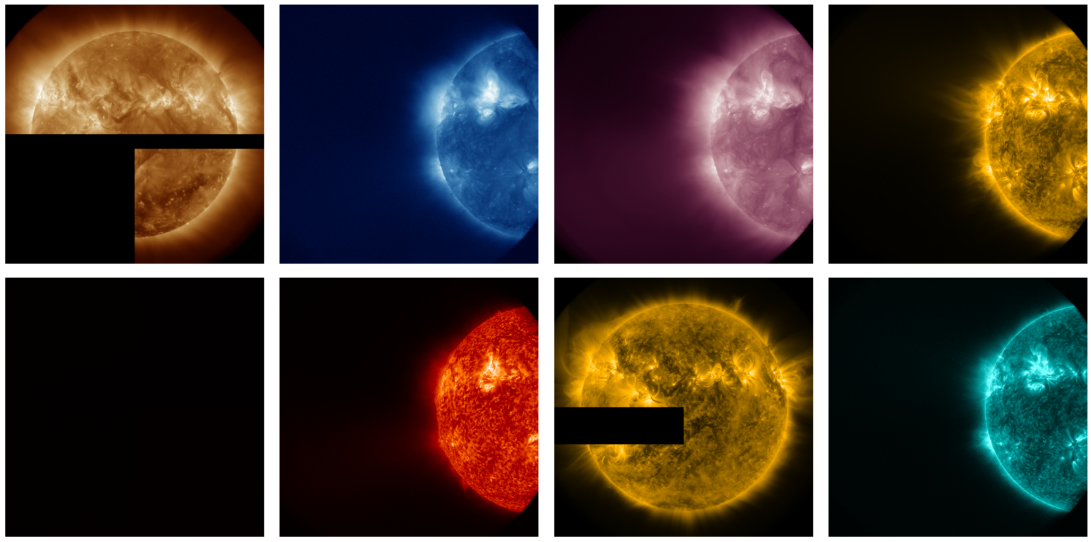


Figure 5. Bad AIA measurements due to a variety of reasons.

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.

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.

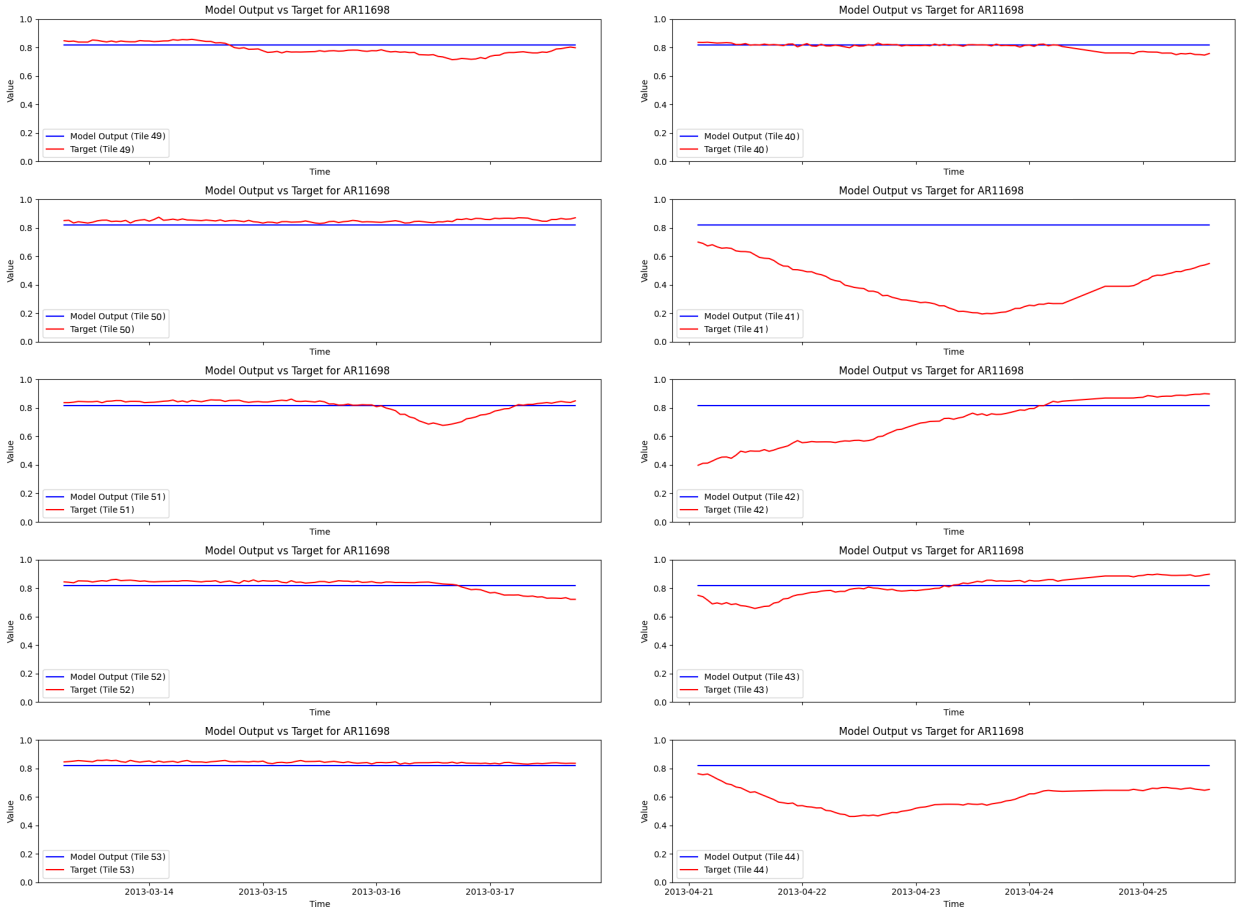


Figure 6. Results for AR emergence forecasting using the HelioFM when validating on AR11698 (left) and AR11726 (right).

- $\mathbf{y} \in \mathbb{R}^S$ denote the target continuum intensity per cell.

The model is trained to approximate a function $f_\theta : \mathbb{R}^{T \times C \times S} \to \mathbb{R}^S$ 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.

Evaluated Model Architectures

We explore two complementary spatiotemporal modeling paradigms: SpatioTemporal Transformer and SpatioTemporal ResNet. The SpatioTemporal Transformer model is a two-stage Transformer that sequentially models temporal and spatial dynamics. Its structure reflects a deliberate architectural bias aligned with the problem's domain priors:

1. **Temporal Attention (per cell)**: For each of the 63 spatial grid cells, we treat the 120-step sequence of 5-channel inputs as a time series. These are projected to a $d = 64$-dimensional embedding and passed through a Transformer encoder with positional encodings, allowing the model to learn temporal dependencies in each spatial location independently.
2. **Spatial Attention (per timestep)**: At each of the 120 timesteps, the spatial pattern of cell embeddings is modeled as a sequence of 63 tokens. A second Transformer block captures interactions and correlations between different locations on the solar disk.
3. **Output Head**: After pooling over time, the spatial embeddings are passed through a linear regressor to predict a scalar continuum intensity for each cell.

This modular design allows the model to explicitly factorize temporal and spatial reasoning, which is beneficial for interpretability and transfer across regions with similar temporal but different spatial dynamics.

As a second baseline, we also implement a 3D convolutional model based on the ResNet-18 video backbone (`r3d_18`), i.e., SpatioTemporal ResNet. Here, the input tensor is reshaped to match the expected input for Conv3D networks:

- $\mathbf{X} \to \mathbb{R}^{B \times C \times T \times H \times W}$ with $H = 1$, $W = 63$, and $C = 5$.

This network is initialized with pretrained weights (optional), and the first convolutional layer is modified to accept 5 input channels instead of 3 (RGB). The final fully connected layer is replaced to regress 63 outputs. This model captures hierarchical spatiotemporal correlations through convolutional filters, offering a computationally efficient and generalizable baseline.

Results and Observations: Models are evaluated using mean squared error (MSE) and mean absolute error (MAE) over all 63 spatial grid cells, averaged across held-out validation regions. Since outputs are per-cell continuous intensities, these metrics offer a direct measure of spatial forecasting accuracy.

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.

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 $\theta_{\max} = 10^{-4}$ 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.
- **ResNet18 / ResNet34 / ResNet50**: Residual Networks with varying depths (18, 34, and 50 layers respectively), 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 Obsersations: Table 5 shows the evaluation metrics we used for solar flare forecasting:

- **TSS (True Skill Statistic)**: Measures the model's ability to distinguish between flare and non-flare events.
- **HSS (Heidke Skill Score)**: Accounts for both hits and false alarms.
- **CSS (Composite Skill Score)**: Provides a balanced measure as the geometric mean of TSS and HSS.
- **F1 Score**: Harmonic mean of precision and recall.

Model	TSS	HSS	CSS	F1
AlexNet	0.359	0.354	0.356	0.679
MobileNet	0.326	0.312	0.319	0.662
ResNet18	0.320	0.317	0.318	0.660
ResNet34	0.290	0.289	0.289	0.645
ResNet50	0.261	0.281	0.271	0.627

Table 5. Solar Flare Forecasting Performance

- **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.

- **CSS (Composite Skill Score)**:

$$CSS = \begin{cases} 0, & \text{if } TSS < 0 \text{ OR } HSS < 0 \\ \sqrt{TSS \times HSS}, & \text{otherwise} \end{cases}$$

  It measures the geometric mean of TSS and HSS when they are positive.

- **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.

The results indicate that AlexNet performs best among the evaluated architectures across all metrics, potentially due to its shallower structure and better generalization on limited data. Deeper architectures such as ResNet50 may suffer from overfitting or excessive capacity relative to the dataset size.

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.

Evaluated Model Architectures: Figure 9 illustrates the architectures utilized for the solar wind forecasting task, emphasizing the distinction between attention-driven and baseline UNet approaches.

We explored and benchmarked several state-of-the-art deep learning architectures detailed below:

- **Attention UNet:** The Attention UNet architecture enhances the traditional UNet through the integration of attention gates, facilitating dynamic suppression of irrelevant spatial features and emphasizing salient regions pertinent to predicting solar wind characteristics. Attention gates, introduced within skip connections between encoder and decoder paths, adaptively weigh encoder outputs based on decoder contexts, significantly improving the discriminative capability of the model. Given the large resolution (4096×4096 pixels) of solar imagery data, adaptive average pooling followed by convolutional layers was strategically employed to condense feature representations, subsequently enabling precise regression to the scalar solar wind speed.

- **Standard UNet:** A standard UNet architecture served as a robust baseline, providing a fundamental encoder-decoder structure with straightforward skip connections. Its primary role was to gauge the incremental benefit derived from attention mechanisms explicitly integrated into the Attention UNet model.

- **ResNet-based Convolutional Models:** As an additional comparative baseline, we employed ResNet architectures (ResNet-18, ResNet-34, and ResNet-50) to leverage deep residual learning's inherent capabilities in capturing complex hierarchical features. These models were initially pre-trained (optional) and specifically adapted for solar data by modifying the first convolutional layer to accept 13 input data representative of solar observational channels (instead of the standard RGB inputs). The final layer of the network was adapted to produce a single scalar output directly.

Notably, we achieved optimal performance metrics and lowest validation loss with the ResNet-50 architecture, likely attributed to its deeper structure and larger parameter count (33 million parameters), facilitating richer representation of complex spatiotemporal solar dynamics.

Results and Observations: Our experiments indicate that incorporating attention mechanisms (Attention UNet) improves the predictive performance over standard UNet, highlighting the importance of adaptive feature weighting in solar wind forecasting. Additionally, deeper architectures such as ResNet-50 outperform shallower networks, emphasizing the complexity and depth required to model solar physics phenomena effectively. These observations underscore a fundamental insight: architectural complexity and context-aware feature selection are critical components in accurately predicting space weather events. Training losses for this task can be found in Figure 8.

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.

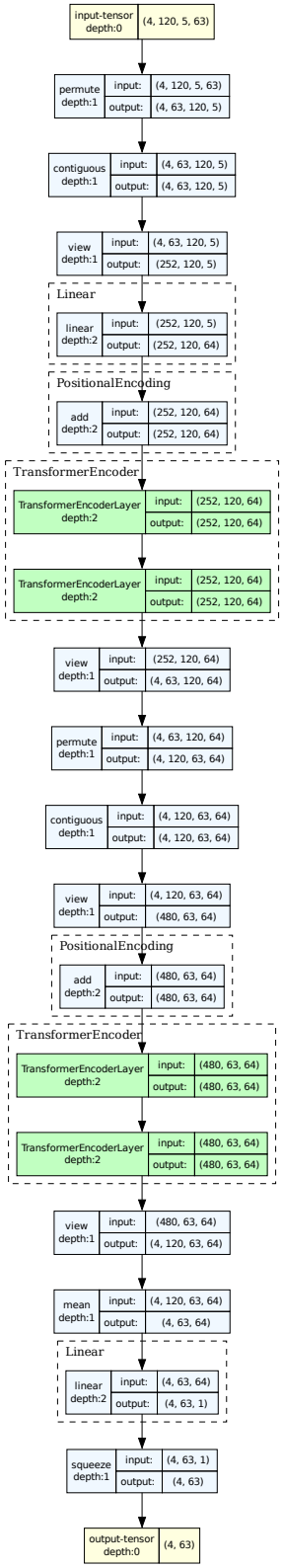
Formally, we frame this as a regression problem: Given multi-channel solar imaging data $\mathbf{x}_t \in \mathbb{R}^{C\times H\times W}$ at time t, the goal is to predict a continuous vector of EUV irradiance values $y_t \in \mathbb{R}^{1343}$:

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

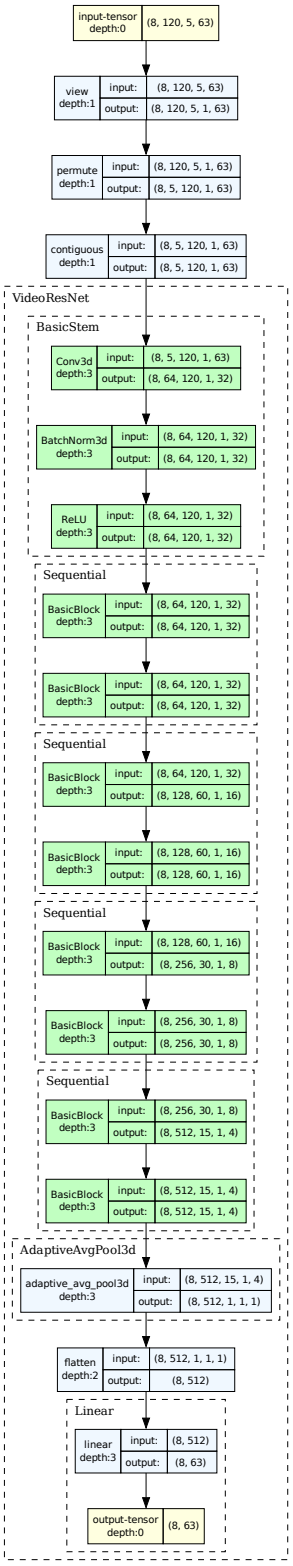
Evaluated Model Architectures: To establish baselines for this task, we evaluated several deep learning architectures commonly used in computer vision and scientific regression tasks. Table 4(a) for main paper presents the performance of these baseline models, comparing both convolutional networks (ResNet variants) and segmentation-inspired architectures (U-Net and Attention U-Net).

Results and Observations: The performance metrics employed include Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and validation loss. These metrics provide complementary views of predictive accuracy and model robustness. U-Net and Attention U-Net significantly outperform traditional convolutional networks, underscoring the efficacy of architectures that inherently model spatial correlations and multi-scale features in predicting complex, high-dimensional irradiance spectra.

This predictive modeling task not only benchmarks model capabilities in handling high-dimensional regression but also advances the applicability of deep learning methods to critical solar physics-driven applications.



(a) SpatioTemporal Attention



(b) SpatioTemporal ResNet

Figure 7. Comparison between the proposed **SpatioTemporalAttention** model and a baseline **SpatioTemporal ResNet**. The SpatioTemporalAttention model explicitly decomposes the modeling task into two sequential Transformer stages: temporal attention applied independently to each spatial grid cell, followed by spatial attention across all cells at each timestep. In contrast, the SpatioTemporal ResNet baseline uses a 3D convolutional backbone adapted from `r3d_18` to learn spatiotemporal features jointly through hierarchical convolutional filters.

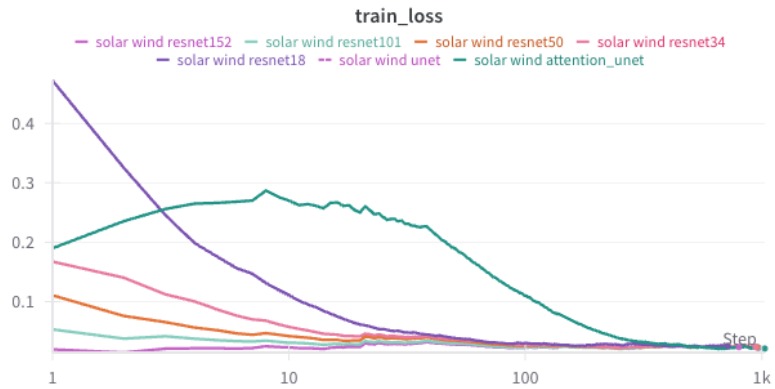
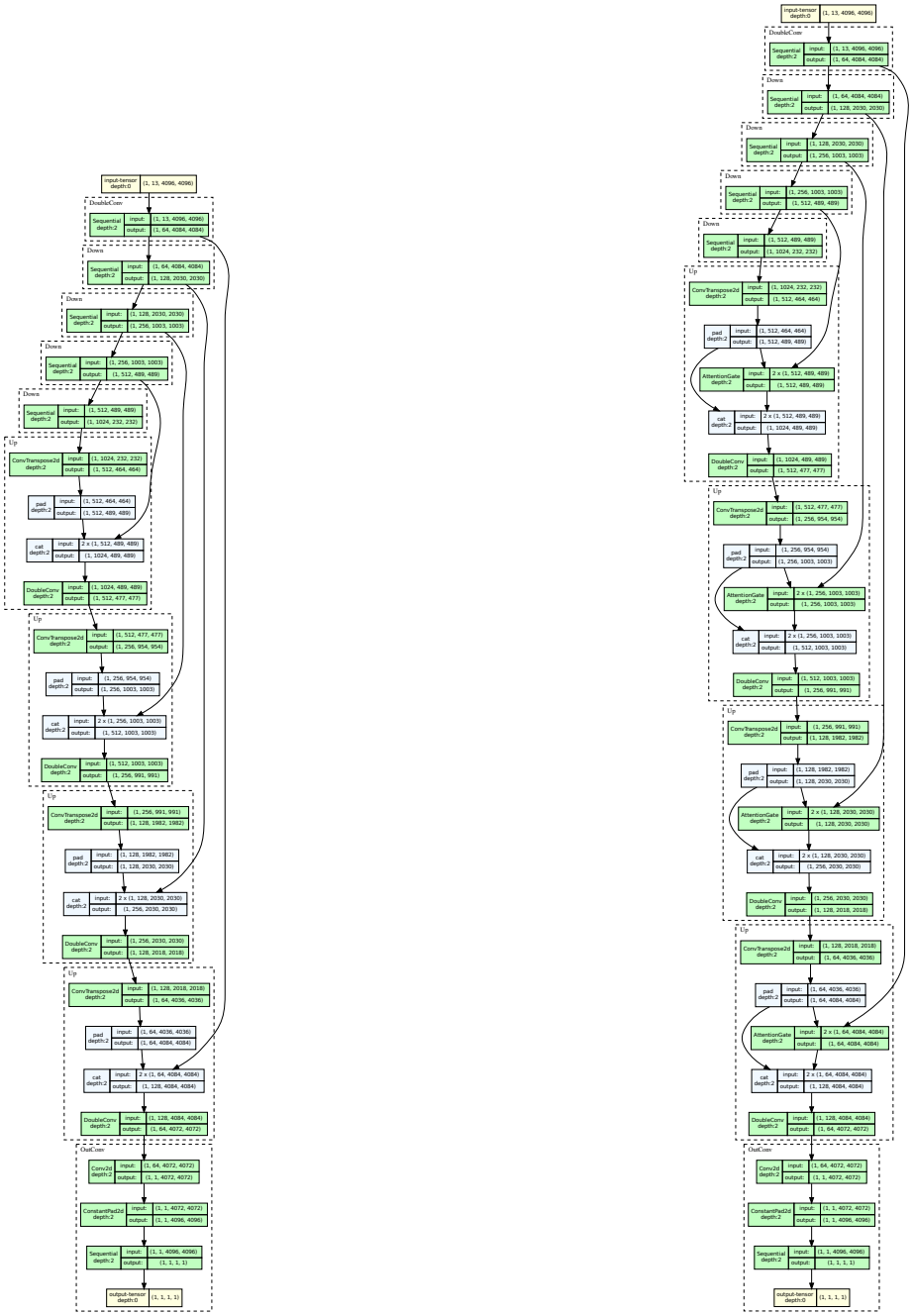


Figure 8. Training loss curves for various deep learning architectures on the solar wind forecasting task. The comparison includes ResNet variants (18, 34, 50, 101, 152), U-Net, and Attention U-Net. Models with deeper architectures (e.g., ResNet152) and attention mechanisms (e.g., Attention U-Net) tend to converge faster and reach lower final training losses, indicating their superior capacity to model the input-output mapping for this regression task. Log-scale on the x-axis highlights early training dynamics.



(a) UNet Model

(b) Attention Unet Model

Figure 9. Detailed architectural diagrams of (a) the UNet model and (b) the Attention UNet model used for predicting solar EUV irradiance and solar wind. The Attention UNet enhances the standard UNet architecture by incorporating attention gates, allowing the network to selectively emphasize relevant spatial features and improve predictive performance in complex regression tasks such as high-dimensional spectral prediction.

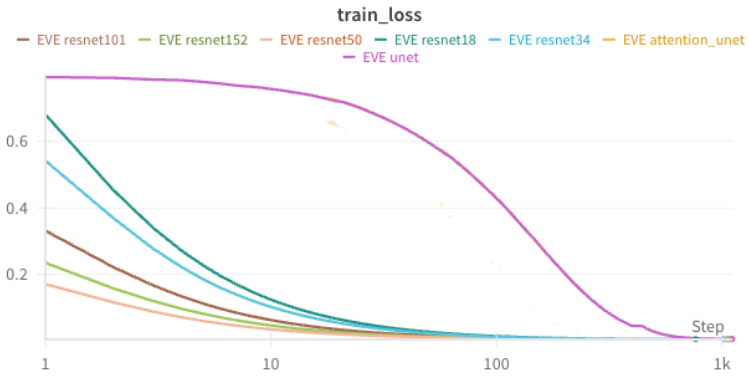


Figure 10. Training loss curves comparing various baseline models (ResNet18, ResNet34, ResNet50, ResNet101, ResNet152, UNet, and Attention UNet) for the Solar EUV irradiance prediction task. UNet and Attention UNet show slower initial convergence compared to ResNet variants but achieve significantly lower final loss values, highlighting their effectiveness in modeling complex spatial patterns inherent to solar EUV spectra.