Check for updates

AGU ADVANCING EARTH AND SPACE SCIENCES

AGU Advances

RESEARCH ARTICLE
10.1029/2026AV002524

PROSWIN: Probabilistic Solar Wind Speed Forecasting Using Deep Distributional Regression From Solar Images

Daniel Collin[1,2], **Yuri Shprits**[1,3,4], **Luca Chiarabini**[5], **Stefan J. Hofmeister**[6], **Nadja Klein**[7], and **Guillermo Gallego**[2,8]

[1]Space Physics and Space Weather, GFZ Helmholtz Centre for Geosciences, Potsdam, Germany, [2]Department of Electrical Engineering and Computer Science, Technical University of Berlin, Berlin, Germany, [3]Institute of Physics and Astronomy, University of Potsdam, Potsdam, Germany, [4]Department of Earth, Planetary, and Space Sciences, University of California Los Angeles, Los Angeles, CA, USA, [5]German Aerospace Center, Munich, Germany, [6]Columbia Astrophysics Laboratory, Columbia University, New York, NY, USA, [7]Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, Germany, [8]Einstein Center Digital Future, Berlin, Germany

Peer Review The peer review history for this article is available as a PDF in the Supporting Information.

Key Points:

- Probabilistic solar wind speed forecasting with well-calibrated uncertainties is possible by using distributional regression
- We show advantages of probabilistic over single-value predictions, particularly through risk quantification of fast solar wind conditions
- We provide more informative forecasts than previous models by overcoming a typical trade-off between timeline and high-speed solar wind stream peak performance

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

D. Collin,
collin@gfz.de

Citation:

Collin, D., Shprits, Y., Chiarabini, L., Hofmeister, S. J., Klein, N., & Gallego, G. (2026). PROSWIN: Probabilistic solar wind speed forecasting using deep distributional regression from solar images. *AGU Advances*, *7*, e2026AV002524. https://doi.org/10.1029/2026AV002524

Received 17 APR 2026
Accepted 14 OCT 2026

Author Contributions:

Conceptualization: Daniel Collin, Yuri Shprits, Stefan J. Hofmeister, Nadja Klein, Guillermo Gallego
Data curation: Daniel Collin
Formal analysis: Daniel Collin, Stefan J. Hofmeister, Nadja Klein
Funding acquisition: Yuri Shprits, Guillermo Gallego

© 2026 The Author(s). AGU Advances published by Wiley Periodicals LLC on behalf of American Geophysical Union. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Abstract Accurately predicting fast solar wind conditions is challenging, as uncertainties are large and unquantified by traditional single-value prediction models. In particular, the risks of high-speed solar wind streams (HSSs), which can cause damage to technological infrastructure, cannot be reliably assessed without probabilistic forecasts. We present PROSWIN, a probabilistic machine learning model that forecasts the hourly solar wind speed (SWS) at Earth with a 4-day lead time. The approach combines solar images and magnetograms using a deep neural network coupled to a distributional regression algorithm. Because standard error metrics underweight the relevance of HSS peaks, we further introduce the *prediction score (PS)*, a model-selection metric that jointly rewards timeline and HSS peak accuracy. On 14 years of data, our forecast achieves very well-calibrated uncertainties (<1% average absolute coverage deviation). Using the continuous ranked probability score (CRPS), a metric that assesses distributional accuracy, we obtain a timeline CRPS of 41.0 km/s, an HSS peak CRPS of 45.3 km/s, and a PS of 42.3 km/s. We find that the 171 Å channel is an important complement to the typically used 193 and 211 Å channels and that the PS for model selection improves the applicability of the model. Compared to selected models from the literature, ours is the only one that is accurate for both timeline and HSS peak values, rather than trading one off against the other. These results support the advantages of probabilistic over single-value solar wind models. The introduced methods are also transferable to other forecasting problems.

Plain Language Summary The solar wind is a stream of charged particles emitted by the Sun that can damage technology in space and on the ground. This occurs especially during periods of fast solar wind, making accurate prediction of the solar wind speed (SWS) essential. However, forecasts of fast solar wind conditions are challenging and often have large errors. To quantify the risks of hazardous conditions, single-value predictions, which give only the most likely outcome, are insufficient. Probabilistic predictions are needed to also capture uncertainty. This study presents a probabilistic forecasting model for the SWS that uses machine learning to learn patterns from solar images. We analyze different image wavelengths and their impact on predictions. Using a new evaluation metric that accounts for the errors of fast solar wind peaks, risk estimation is improved, and the model is better calibrated to operational needs than previous methods. Using 14 years of data, we show that predicted uncertainties are highly reliable, deviating by less than 1% from observations. Overall, probabilistic forecasts provide more useful information for operators and enable more comprehensive risk assessment than traditional approaches.

1. Introduction

With the growing number of satellites and space missions, forecasting hazardous space weather effects in real time is becoming increasingly important. However, the complexity of the Sun-Earth interactions is not yet fully understood, which limits the accuracy of deterministic single-value prediction models. Particularly during high-speed solar wind stream (HSS) events, solar wind single-value prediction models frequently make large errors without providing uncertainty information. This is an important limitation, aggravating decision-making and possibly leading to substantial economic and operational costs. To ensure that solar wind forecasts are reliable and actionable, they must therefore be provided in a probabilistic way, that is, accompanied by well-calibrated

Investigation: Daniel Collin, Stefan J. Hofmeister, Nadja Klein
Methodology: Daniel Collin, Luca Chiarabini, Stefan J. Hofmeister, Nadja Klein, Guillermo Gallego
Project administration: Yuri Shprits, Guillermo Gallego
Software: Daniel Collin, Luca Chiarabini
Supervision: Yuri Shprits, Stefan J. Hofmeister, Nadja Klein, Guillermo Gallego
Validation: Daniel Collin, Luca Chiarabini, Stefan J. Hofmeister, Nadja Klein
Visualization: Daniel Collin, Guillermo Gallego
Writing – original draft: Daniel Collin, Nadja Klein
Writing – review & editing: Daniel Collin, Yuri Shprits, Stefan J. Hofmeister, Nadja Klein, Guillermo Gallego

uncertainty estimates (Camporeale, 2019; Gneiting & Katzfuss, 2014). While probabilistic forecasting is standard in other fields such as meteorology, most operational space weather models still provide single-value predictions only.

Among the main drivers of hazardous space weather conditions are HSSs and coronal mass ejections (CMEs). HSSs are plasma streams accelerated along magnetic field lines that are rooted in solar coronal holes and open toward interplanetary space (Krieger et al., 1973), whereas CMEs originate from sudden eruptions of solar filaments, ejecting large and fast plasma clouds into the heliosphere (Gosling, 1997). During the declining phase and the minimum of the solar cycle, HSSs are the dominant cause of geomagnetic storms, while the frequency of CMEs is related to solar activity and thus increases toward the solar maximum (Richardson et al., 2000; Tsurutani et al., 2006). In this study, we focus on the probabilistic prediction of HSSs.

Most operational SWS forecasting models are based on either physical heliospheric simulations (e.g., WSA-ENLIL; Odstrcil, 2003, EUHFORIA; Pomoell & Poedts, 2018) or empirical approaches (e.g., WSA; Arge et al., 2003, ESWF; Milošić et al., 2023; Reiss et al., 2016). At the same time, machine learning approaches (e.g., Bailey et al., 2021; Collin et al., 2025), especially deep learning from solar images (e.g., Brown et al., 2022; Upendran et al., 2020), where important structures such as coronal holes and active regions can be identified well, show the potential to outperform traditional models (Ahn et al., 2025).

However, none of these approaches explicitly predicts forecast distributions, which makes them unable to provide reliable uncertainty estimates. Furthermore, they mainly focus on minimizing errors averaged over the whole data set rather than those of HSS events. Due to the strong imbalance between frequent slow speeds and rare HSS peaks, this strategy generally does not lead to actionable forecasts (Owens & Riley, 2017), but to a systematic underestimation and large unquantified uncertainties for the most important data points (Camporeale, 2025; Collin et al., 2025; Shprits et al., 2019).

There exist several approaches that obtain SWS uncertainties from ensembles (e.g., Bussy-Virat & Ridley, 2014, 2016; Edward-Inatimi et al., 2024, 2026; Issan et al., 2023; Owens & Riley, 2017), whose reliability depends on how the ensemble members are chosen. Ensembles of machine learning models have similarly been used in other applications (e.g., Gu et al., 2019). Probabilistic machine learning models provide an alternative by directly learning the uncertainty from the data as a function of the inputs. For example, Gaussian Process models and variants have been applied to geomagnetic index prediction (e.g., Chakraborty & Morley, 2020; Chandorkar & Camporeale, 2018; Gruet et al., 2018). Also the direct prediction of distributional parameters for space weather forecasts has been explored in the machine learning context (e.g., Bernoux et al., 2022; Hu et al., 2023; Licata et al., 2022; Tahtouh et al., 2025; Tasistro-Hart et al., 2021). Furthermore, methods have been suggested for learning uncertainty of existing single-value predictions (Camporeale & Carč, 2021; Camporeale et al., 2019). However, to the best of our knowledge, no probabilistic machine learning model currently exists for the SWS.

In this work, we introduce PROSWIN, a Probabilistic SWS forecasting approach. It is based on distributional regression (Klein, 2024), which serves as a general framework for training probabilistic models using arbitrary parametric probability distributions. We model the skewed and heavy-tailed SWS distribution with log-normal distributions, which better describe the uncertainty of the SWS than the commonly assumed Gaussian distributions. To learn these distributions directly from solar images and additional physical features, such as the SWS from the previous solar rotation and the state of the solar cycle, we couple distributional regression with deep learning. An overview of PROSWIN is displayed in Figure 1 We optimize the model using a newly proposed metric that balances timeline and HSS peak errors, leading to significant improvements for HSS predictions. Then, we investigate the effect of combining different solar image channels and magnetograms, finding that the information content of 193 and 211 Å is largely redundant, whereas that of 171 Å is complementary. Furthermore, we analyze the reliability of the predicted uncertainties, with a particular focus on HSSs, yielding well-calibrated uncertainties and accurate risk quantification. Finally, we demonstrate that our approach outperforms a selection of other SWS models and that the trained models generalize well to new unseen data. Our contribution also involves the publication of a Python implementation of PROSWIN (Collin et al., 2026a), as well as the publication of our data sets and trained models (Collin et al., 2026b).

The paper is structured as follows: Section 2 describes the data and preprocessing. Section 3 explains the algorithmic approach, including the distributional regression and deep learning frameworks. Section 4 introduces the evaluation procedure, encompassing a new metric. Section 5 studies the impact of different combinations of

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

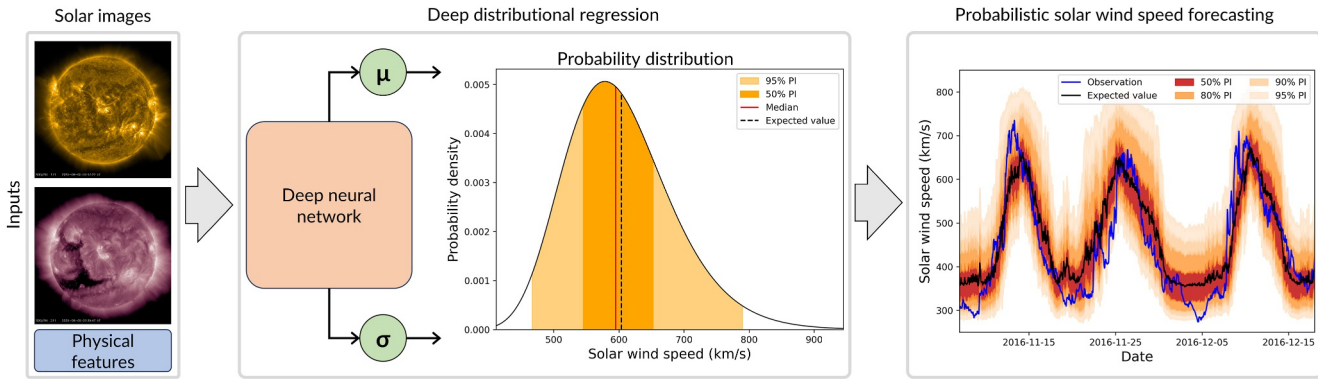


Figure 1. PROSWIN: Solar images and physical features are fed as inputs to a deep neural network, which predicts the distributional parameters of a log-normal distribution for the solar wind speed (SWS). This enables probabilistic SWS forecasting 4 days ahead.

solar images and magnetograms on the model performance. Section 6 analyzes the best-performing model and evaluates the benefits of probabilistic forecasting for SWS prediction. Section 7 compares our probabilistic forecast to other methods. Section 8 tests the model on unseen data, Section 9 discusses the main findings and limitations, and Section 10 draws conclusions for future research.

2. Data Set

Our study combines image data with time series of physical features. The image inputs consist of three solar image channels along with magnetograms. In addition, we incorporate physical features derived from in situ solar wind measurements and indicators of the solar cycle, resulting in a total of 63 features. The prediction target is the hourly average of the SWS, measured 4 days after the input images are recorded. The data set and forecast cadence is 1 hr, allowing to predict SWS variability on the timescale of HSS onsets and peaks, which can produce sharp increases over a few hours. Furthermore, Brown et al. (2022) show that the performance of their solar image-based SWS model improves with finer training data sampling but with diminishing returns, recovering most of the improvement already at hourly resolution.

Our data period spans the period from June 2010 to June 2026. The last 2 years of the data set, from July 2024 onwards, are not taken into account for model development, training, or cross-validation, but are exclusively used for an additional test of the model on previously unseen data (in Section 8). After preprocessing, this yields 123,129 data points for the main data set used throughout this study and an additional 17,425 data points for the test on unseen data. Furthermore, we identify time intervals associated with HSSs and CMEs using the interplanetary coronal mass ejection (ICME) catalog of Richardson and Cane (2004). In this section, we describe the data set in detail, which is also publicly available in Collin et al. (2026b).

2.1. Solar Images

We use solar extreme ultraviolet (EUV) images from the 171, 193, and 211 Å channels of the Atmospheric Imaging Assembly (AIA; Lemen et al., 2012), and line-of-sight magnetograms from the Helioseismic and Magnetic Imager (HMI; Scherrer et al., 2012), both on board the Solar Dynamics Observatory (SDO). The 171 Å channel predominantly captures the emission of Fe IX at about 0.8 MK, highlighting mainly the low solar corona. The 193 Å channel records Fe XII emission in the corona around 1.6 MK, and the 211 Å channel records Fe XIV emission around 2.0 MK. Coronal holes are clearly visible as dark structures in the 193 and 211 Å channels. The 171 Å channel provides complementary information related to the expansion factor of coronal holes. The HMI magnetograms measure the photospheric magnetic field component, providing information on the magnetic polarity distribution and large-scale magnetic structures, such as active regions. AIA data is available with a temporal cadence of 12 s, whereas the magnetograms have a cadence of 12 min. Our approach uses both at an hourly cadence. While AIA images and HMI magnetograms would also be available in near-real time with a latency of less than 1 hour from https://sdo.oma.be/data/aia_quicklook/, and http://hmi.stanford.edu/magnetic/, respectively, we use the definitive level 1 science data from the Joint Science Operations Center at http://jsoc.stanford.edu/.

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We preprocess the images, correcting for position, rotation and scaling of the solar disk, light scattering, and instrument degradation, followed by a value normalization and the selection of a central square whose corners are approximately at the edges of the solar disk. These procedures result in images of 224 × 224 pixels and are adapted from Galvez et al. (2019), Jarolim et al. (2021), Brown et al. (2022), Hofmeister (2024), and Hofmeister et al. (2025). The exact methodology is described in Appendix A.

2.2. Solar Wind Measurements

As SWS data, we use the hourly-averaged plasma bulk speed provided directly by the NASA OMNIWeb database (Papitashvili & King, 2020), recorded in situ at the Lagrange Point 1 (L1). We do not perform any additional averaging ourselves. In the OMNI database, this data set is constructed from high-resolution (1–5 min) upstream measurements that are time-shifted to the Earth's bow shock using a ballistic convection model, in which each individual observation is shifted according to its own concurrent in situ plasma speed before hourly averaging. This time-shifting procedure aligns the data with measurements of other spacecraft closer to Earth and accounts for spacecraft location, Earth's orbital velocity, and phase-front orientation (King & Papitashvili, 2005).

To account for the periodicity of recurrent coronal holes, we also include as physical inputs the hourly averages of the SWS from 26 to 28 days prior to the forecast time, resulting in 49 features. A fraction of values (0.5%) is missing in the SWS time series. Those gaps are filled by linear interpolation. In addition, we include the heliospheric latitude of Earth, that is, the angle between the solar equatorial plane and the current position of Earth, which is derived from the heliographic inertial coordinates provided in the OMNI_M data by the NASA COHOWeb. This feature captures the latitudinal variation of Earth during its revolution around the Sun, relative to the solar latitudes of the solar wind source region (Hofmeister et al., 2018). In total, 50 input features are derived from solar wind measurements.

While OMNI SWS data is typically available with a latency of 20 days, it can at times exceed the 22-day threshold required for our 26–28 day input window. To facilitate a real-time implementation of our approach, OMNI data could be substituted with near-real-time measurements, for example, from ACE (https://sohoftp.nascom.nasa.gov/sdb/goes/ace/monthly/). Although ACE SWS data has a correlation of 0.99 with OMNI SWS data on our data period, minor discrepancies mean real-time performance may not fully match the results reported in this study. The heliospheric latitude of Earth follows a fixed pattern and can be extrapolated if needed in real-time.

2.3. Solar Cycle

SWS models are known to depend on the solar cycle (e.g., Collin et al., 2025). Thus, we include as indicators of the solar cycle the 12 most recent values of the monthly sunspot number from the WDC-SILSO at the Royal Observatory of Belgium (Clette & Lefčvre, 2015) and the number of years that have passed since the start of the current cycle (as a coarse indicator of the current position within the solar cycle). According to the NOAA Space Weather Prediction Center, solar cycle 24 started in December 2008 and solar cycle 25 started in December 2019. We adopt those solar cycle start dates for our study. This yields 13 solar-cycle-related input features. Since the monthly sunspot number is available at the end of each month and we only use the values of past months, all needed features are also available in real-time.

2.4. High-Speed Streams and Coronal Mass Ejections

We focus on forecasting HSSs and therefore need to distinguish between HSSs and CME-related disturbances for the later evaluation of HSS prediction performance. However, CMEs are not removed from the data set. Instead, since our model provides no explicit CME input, we expect CME-driven disturbances to manifest as unexplained variance during training. To distinguish between both events, we apply the procedure introduced in Collin et al. (2025) to filter all SWS enhancements, that is, extended periods of elevated SWS, from the solar wind time series. Then, we use the ICME list of Richardson and Cane (2004) to classify them.

First, we smooth the SWS time series using a Gaussian filter with a standard deviation of 24 hr and identify local enhancements that exceed 390 km/s and have a peak excess, that is, vertical distance between its highest point and its base, of at least 35 km/s. Next, we define the start and end of each enhancement by the points where the smoothed time series crosses the relative height of 0.4 from the peak's base to its maximum. If two events overlap, the longer one is truncated. Enhancements that overlap with an ICME interval or occur within 2 days after an

ICME are classified as a CME-related disturbance. All other events are labeled as HSSs. This produces HSS and CME lists with approximate start and end dates, which are later used for model evaluation.

3. PROSWIN

The following sections describe the probabilistic regression model employed, how to leverage predictive performance for the SWS forecasting by combining the model with deep neural networks, and how to perform forecasting. We call this approach *PROSWIN*. A Python implementation of the method is available in Collin et al. (2026a).

3.1. Distributional Regression

Traditional regression models (e.g., generalized linear or generalized additive models; Nelder & Wedderburn, 1972; Hastie & Tibshirani, 1990) focus on modeling the conditional mean of a target variable, that is, the expected value of the target conditional on a set of input features. This is a relevant limitation, as in many applications, mean-based models can deviate substantially from observations for individual events. For the SWS, whose distribution is strongly influenced by fast speeds, this limitation is particularly relevant. Focusing on the mean typically leads to severe underestimation of fast events and, consequently, high unquantified uncertainty. In contrast, distributional regression methods address this limitation by predicting the parameters of a probability distribution for the target variable instead of the single mean value, thereby modeling the entire conditional distribution (Klein, 2024). Here, we introduce the idea of distributional regression applied to SWS forecasting. More details on the implementation and a more general mathematical definition can be found in Appendix B.

To apply distributional regression, we first assume that the SWS target variable V_{sw} is distributed according to a parametric distribution family. These are probability distributions which are defined by a set of parameters, for example, normal, exponential, or uniform distributions, and the distributional parameters typically represent properties such as the location, scale, or shape of the distribution. Previous mean-based SWS models (e.g., Bailey et al., 2021; Brown et al., 2022; Raju & Das, 2021; Upendran et al., 2020) are trained using a mean squared error loss, which can be interpreted as implicitly assuming that the target follows a normal distribution with constant variance. However, the observed distribution of the SWS is highly skewed with a heavy upper tail. This can be seen in Figure 2a, which shows the empirical distribution of SWS measurements in our data set together with the best-fitting marginal normal and log-normal distributions, respectively. While this perspective is marginal only, that is, not yet accounting for input features, histograms can be helpful for a first visual assessment of which distributional family would be a reasonable choice for the problem. Our choice of a log-normal distribution is motivated by its strict positivity, which assigns zero probability to physically impossible negative SWS values, unlike the normal distribution. Further, the SWS distribution reflects a quiet baseline of slow solar wind punctuated by episodic high-speed enhancements from HSSs and CMEs, which generates an asymmetry that is captured by the skewness of a log-normal distribution. Figure 2a confirms that the normal distribution underestimates the tail behavior, whereas the log-normal distribution reproduces the observed skewness and large positive deviations. This characteristic is important for space weather applications, where the most relevant events are located in the upper tail. Thus, we assume that V_{sw} is log-normally distributed with parameters μ and σ, where μ is the mean and σ the standard deviation of the log-transformed SWS $\log(V_{sw})$.

After choosing a suitable distribution family, the objective within the distributional regression framework is to predict conditional distributions for V_{sw} by modeling the distributional parameters as functions $\mu(x)$ and $\sigma(x)$ of the input features x. This allows the predicted distributions to change depending on the solar conditions represented by the features. Figure 2b illustrates an example prediction characterized by the probability density function (PDF), specifying the likelihood of different SWS values. In contrast, a mean-based single-value prediction model would only provide the expected value. More details on the probabilistic prediction are given in Section 3.3.

A key flexibility of the approach comes through the freedom of choosing specific forms of the predictor functions $\mu(x)$ and $\sigma(x)$. We learn those using a neural network, which is described in detail in Section 3.2. The network's parameters are fitted by minimizing the negative log-likelihood of the observed data. This corresponds to finding functions for $\mu(x)$ and $\sigma(x)$ that assign high probabilities to the observed SWS. This modeling strategy has been popularized as structured additive distributional regression (Klein et al., 2015) or Generalized Additive Models for Location, Scale and Shape (GAMLSS; Rigby & Stasinopoulos, 2005).

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

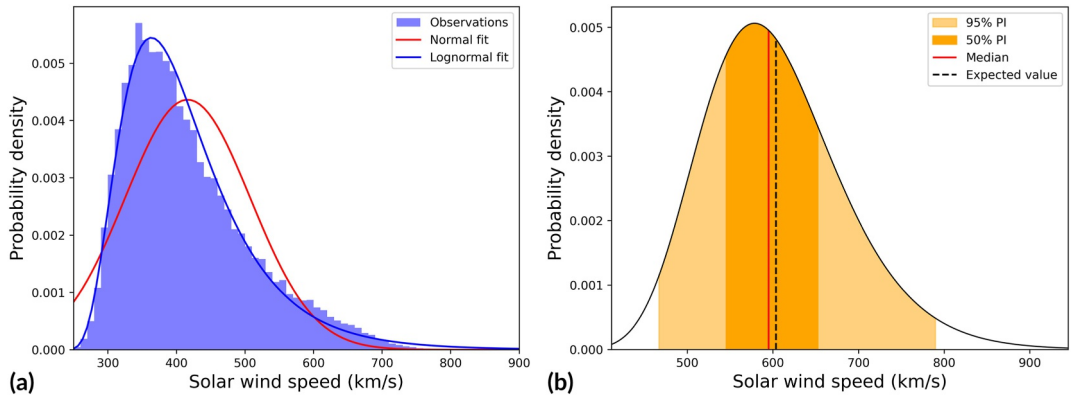


Figure 2. (a) Empirical distribution of solar wind speed observations in the data set, together with the best-fitting normal and log-normal distributions. The log-normal distribution is better suited for the data because of its strict positivity and skewness, particularly modeling the heavy upper tail. (b) Example of a predicted conditional log-normal distribution for a given input. In addition to the expected value, the uncertainty of the prediction is fully quantified.

In summary, we probabilistically model the SWS as $V_{sw} \sim \text{Lognormal}(\mu(x), \sigma(x))$, where $\mu(x)$ and $\sigma(x)$ are represented by neural networks trained using a negative log-likelihood loss. For any model input x, this formulation then yields a full conditional distribution of the SWS.

3.2. Deep Learning of Predictors

Next, we introduce the neural network that is used to predict the distributional parameters μ and σ. The model architecture is illustrated in Figure 3. It is motivated by the findings of Brown et al. (2022), who showed that attention-based neural networks outperform convolutional neural networks for SWS prediction from solar images. We therefore adopt the usage of a Swin Transformer (Z. Liu et al., 2021) as image encoder. It has also been shown to be effective in encoding expansion factor and angular distance maps in the WSA + approach (Mayank et al., 2025). A Swin Transformer is a special type of Vision Transformer, which partitions an image into patches that are flattened, linearly projected, and processed by self-attention layers. While a standard Vision Transformer employs fixed patches, a Swin Transformer uses smaller windows that are shifted between layers, so patches can interact across window boundaries, and a hierarchical structure progressively merges patches. This allows the model to capture long-range spatial dependencies while maintaining computational efficiency.

As the Swin Transformer supports an input of size $224 \times 224 \times 3$, we select up to three SDO channels and stack them to obtain a three-channel image of 224^2 pixels. If fewer than three channels are used, the selected channels are replicated until a three-channel image is obtained. Because the different channels interact with each other through the attention mechanism, the combination of multiple SDO channels can be exploited.

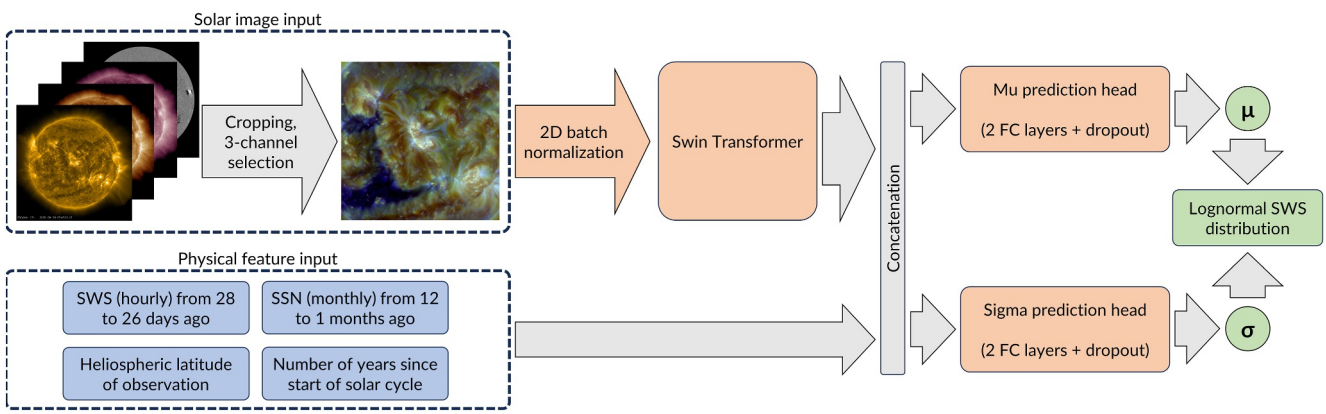


Figure 3. Deep learning model pipeline. Information from images and physical features is combined and used in separate prediction heads to model the distributional parameters of conditional log-normal solar wind speed distributions. Red boxes: trainable functions; blue boxes: input data; green boxes: output data.

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We initialize the Swin Transformer with weights pretrained on the ImageNet-1k data set, a large benchmark data set widely used to learn robust feature representations of images (Deng et al., 2009). We use the tiny version of the Swin Transformer, as larger variants only yielded minor performance improvements while substantially increasing training time in our experiments. Before feeding images into the encoder, we apply a 2D batch normalization layer, which normalizes activations across the mini-batch, reducing internal covariate shift, accelerating convergence, and reducing sensitivity to weight initialization (Ioffe & Szegedy, 2015).

The output of the image encoder is a low-dimensional vector representation, which is concatenated with the physical feature vector. Because we aim to predict both μ and σ, the model contains two prediction heads with identical structure. Each head consists of two fully connected layers, that is, one hidden layer, with dropout applied to the hidden layer. Dropout randomly removes a fraction of neurons during training to reduce overfitting and smooth the output (Srivastava et al., 2014). A LeakyReLU activation function is used in both heads, which allows for a small, non-zero gradient when a neuron is not active, and thereby is more robust during optimization (Maas et al., 2013). The combined feature vector is passed to the prediction heads, producing estimates of μ and σ, which fully specify the conditional distribution. The number of hidden neurons and the dropout rate of both heads are determined by a hyperparameter optimization procedure described in Section 4.3. The neural network has around 27.7 million trainable parameters, depending on the exact hidden layer size. However, 27,519,354 of those are pretrained parameters of the Swin Transformer and only need to be finetuned, whereas around 200,000 parameters need to be trained from scratch.

To improve the numerical stability of the training, we apply some data transformations and train the model in two stages: first the image encoder and the μ-head, then the σ-head, exploiting the fact that μ and σ decouple for normal distributions during training on log-transformed targets. Since the expected value of a log-normal distribution depends on both parameters jointly, this sequential procedure does not imply statistical independence of μ and σ. In the final log-normal output, μ and σ are mathematically covariant. We therefore apply an explicit correction to μ after fitting σ, adjusting it so that the expected value of the resulting log-normal prediction matches the value learned by the μ-head in the first training stage. This correction enforces the physical integrity of the forecast despite the sequential training procedure. In addition, we assign higher training weights to HSS events to account for the class imbalance between fast and average solar wind conditions, and a regularization of the SWS uncertainty to enforce physically realistic boundaries. All technical details on the training process as well as the derivation of the correction and regularization are given in Appendix C and Appendix D, respectively.

3.3. Probabilistic Prediction

After training, the neural network can be used for inference of arbitrary inputs of solar images and physical input features at a particular time point, giving as output distributional parameter estimates $\hat{\mu}$ and $\hat{\sigma}$ and thereby predicting a log-normal SWS distribution. This probability distribution is characterized by the PDF, which represents the relative probabilities of V_{sw} attaining a specific value or falling within a certain range. The PDF can be used to derive any quantity of interest for the predicted distribution, for instance, the expected value, which is the typical output of a single-value prediction model, or the cumulative distribution function (CDF), which specifies the probability that V_{sw} stays below a certain threshold. By subtracting the CDF from 1, we further get the probability that V_{sw} exceeds that threshold. Finally, we get the quantile function, which is the inverse of the CDF. It can be used to construct prediction intervals (PIs) that contain V_{sw} with a specified probability. All additional technical details about the described quantities are given in Appendix E.

4. Evaluation Procedure

4.1. Metrics

We differentiate between model performance averaged over the full data set and the performance for HSSs and their peak values. In the first case, metrics are applied to the aggregated time series and are therefore called *timeline* (i.e., full-series) metrics. Metrics that are applied to the subset of observed HSS peaks and their associated predicted HSS peaks are called *HSS peak value* metrics. Additionally, we combine both performance aspects in a single new metric, which we call *prediction score (PS)*. In the following, we first introduce the timeline metrics we use, then explain the HSS prediction evaluation, and finally introduce the PS.

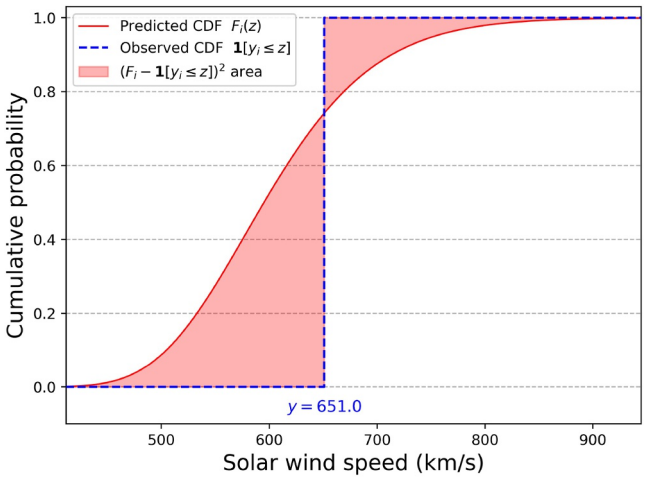


Figure 4. Illustration of continuous ranked probability score computation. The difference between predicted and observed cumulative distribution function is integrated, such that both sharpness and calibration of predictions are evaluated.

4.1.1. Timeline Metrics

To evaluate the predicted distributions, we use the continuous ranked probability score (CRPS; Gneiting & Raftery, 2007), defined as

$$\mathrm{CRPS} = \frac{1}{n}\sum_{i=1}^{n}\int_{-\infty}^{\infty}\left(F_i(z|\mu_i,\sigma_i) - \mathbb{1}\{y_i \le z\}\right)^2 dz.$$

It measures the difference between the CDFs $F_i(z|\,\mu_i,\sigma_i)$ of the predicted SWS distributions, based on the parameter estimations μ_i and σ_i, and the empirical CDFs $\mathbb{1}\{y_i \le z\}$ of the actual SWS observations, which are in fact step functions, because the observed values y_i are attained with 100% probability. These step functions are defined using the indicator function $\mathbb{1}$. As those CDFs are probability distributions defined for all real numbers, we compute their difference by integrating over the real line, with z being the variable of integration which represents all possible SWS values. A visualization of the computation is shown in Figure 4. Then, we compute the average CRPS value over all n data points. The CRPS rewards sharp and well-calibrated distributions by penalizing forecasts that are both inaccurate (far from the truth) and uncertain (too spread out). Lower values indicate better performance.

We further use the Brier score (BS; Brier, 1950) to evaluate predicted probabilities p_i of certain events, defined as

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2.$$

Those events can be, for example, the SWS exceeding a specific threshold or the occurrence of an HSS. The predicted probabilities are compared to the observations o_i, which are 1 if the event occurred, and 0 if not. The BS can take on values between 0, meaning a perfect forecast, and 1, the worst. Both CRPS and BS are commonly used for probabilistic prediction models (e.g., Bernoux et al., 2022; Edward-Inatimi et al., 2024; Klein et al., 2023; Tahtouh et al., 2025).

However, for rare events, for example, the SWS exceeding 700 km/s, the BS can be positively biased, because even underperforming methods can have a good BS. For example, predicting a constant 0% probability is the perfect prediction for most time points due to the rarity of the events, but clearly has no forecast skill. Therefore, the Brier Skill Score (BSS) is often considered as an alternative. It assesses forecast skill relative to a reference model, which we choose as the naive climatological model that always predicts the observed frequency of the event in the data set. That yields the definition

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}}, \quad \mathrm{BS}_{\mathrm{ref}} = \frac{1}{n}\sum_{i=1}^{n}(p_{\mathrm{ref}} - o_i)^2, \quad p_{\mathrm{ref}} = \frac{1}{n}\sum_{i=1}^{n} o_i.$$

The reference model represents a model without predictive skill, which one seeks to improve on. The value $\mathrm{BSS} = 1$ represents the best possible score. $\mathrm{BSS} > 0$ means that the model has an advantage over the reference model, $\mathrm{BSS} < 0$ means that the performance is worse than the reference model.

In addition, we evaluate the accuracy of the expected SWS values by computing the root mean square error (RMSE) between the observations and the expected values of the predicted distributions.

For the CRPS and RMSE metrics, we report the error uncertainty using 95% confidence intervals via a moving block bootstrap (Künsch, 1989) with 1,000 replicates, summarized as a symmetric ± half-width around the single-value estimate. We use a block length of 27 days for timeline metrics, to account for the autocorrelation of the solar wind, and i.i.d. resampling for peak metrics. We do not report confidence intervals for the event detection metrics or the PS, as these are not directly computed from the prediction errors and require more complex resampling schemes.

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

4.1.2. HSS Event and Peak Metrics

To assess the performance of the model for HSSs, we use an event-based approach. We identify SWS enhancements in the time series of predictions of the expected SWS values, following the procedure in Section 2.4, and match them to observed enhancements as described in detail in Collin et al. (2025), and shortly summarized as follows. Two enhancements are associated if they occur within 3 days of each other. If several candidate matches exist, we use the temporally closest pair. Predicted enhancements associated with observed CME disturbances are removed. All other associations are marked as a hit (correct HSS prediction). Predicted enhancements that are not matched to an observation are labeled as false alarms (HSS predicted but not observed), and observed enhancements that are not matched to a prediction are labeled as misses (HSS observed but not predicted). We count the number of hits as true positives (TP), misses as false negatives (FN), and false alarms as false positives (FP). Based on these counts, we compute the verification measures: probability of detection (POD), false alarm ratio (FAR) and threat score (TS) (Woodcock, 1976),

$$\mathrm{POD} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad \mathrm{FAR} = \frac{\mathrm{FP}}{\mathrm{TP} + \mathrm{FP}}, \qquad \mathrm{TS} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}.$$

The POD quantifies the detection ratio among all observed HSSs, the FAR quantifies the ratio of false alarms among all predicted HSSs, and TS balances both aspects, rewarding high detection rates only if they are not at the cost of many false alarms. All three metrics range between 0 and 1, where 1 is the best outcome for POD and TS, and 0 is best for FAR.

To further evaluate the accuracy of HSS peaks, we also compute the CRPS and RMSE between the observed and predicted HSS peak velocities.

4.1.3. Prediction Score

Previous work has shown a systematic trade-off between timeline performance and HSS peak performance in solar wind prediction models (e.g., Collin et al., 2025; Shprits et al., 2019). To account for this trade-off, we introduce the PS, a composite metric that combines the timeline and HSS peak value errors into a single criterion. The PS is based on a weighted harmonic mean of the two error components, which favors models achieving good performance across both aspects over optimizing a single error. The relative importance of the two components can be adjusted through weighting factors depending on the objective of the model and the characteristics of the data set. We define the PS as

$$\mathrm{PS}_w(e_\mathrm{t}, e_\mathrm{p}) = \frac{(w_\mathrm{t} + w_\mathrm{p})}{\frac{w_\mathrm{t}}{e_\mathrm{t}} + \frac{w_\mathrm{p}}{e_\mathrm{p}}},$$

where e_t and e_p are the timeline and peak errors, respectively. We assign double weight to the timeline component, that is, $w_\mathrm{t} = 2$ and $w_\mathrm{p} = 1$, to counteract an imbalance in the number of underlying data points: The timeline error is computed from all 123,129 hourly predictions, while the HSS peak error is computed from only 284 observed peaks. Under equal weighting, each peak prediction would have a disproportionate influence on the score, incentivizing the model to trade general accuracy for isolated peak improvements, for example, via noisier predictions leading to unphysical spikes in the predicted time series. We found empirically that a 2:1 weighting yields the best balance between timeline and HSS peak errors in our experiments.

With our concrete choice of weights, the PS simplifies to

$$\mathrm{PS}(e_\mathrm{t}, e_\mathrm{p}) = \frac{3 e_\mathrm{t} e_\mathrm{p}}{e_\mathrm{t} + 2 e_\mathrm{p}}.$$

4.2. Cross-Validation

We evaluate our model using five-fold cross-validation (CV). First, the data set is divided into five subsets. Then, for each fold, three subsets are used as training data to fit the model, one as validation data to monitor overfitting, and one as test data to evaluate the model on unseen data. The overall model performance is

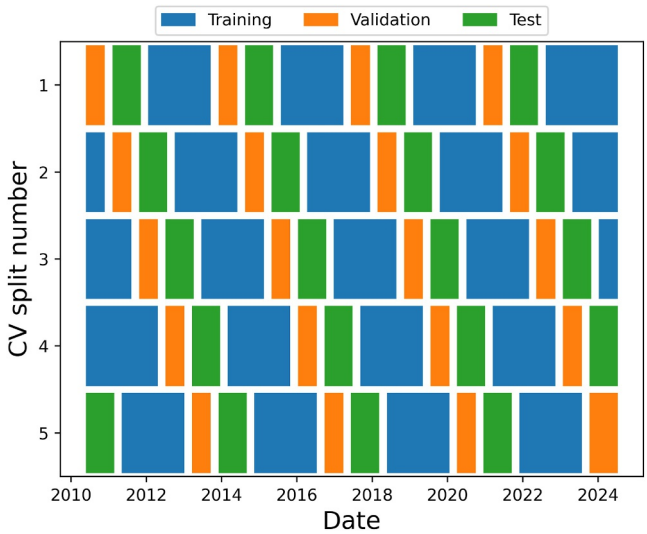


Figure 5. Cross-validation data splits of our data set. Training and evaluation data is distributed over the solar cycle and 90 days are discarded between chunks to avoid data leakage.

assessed by computing the evaluation metrics on the aggregated set of all test data predictions. Each prediction contributing to this aggregated evaluation set is genuinely out-of-sample, as it comes from the one fold in which that point was held out as test data, and its model never observed it during training or validation. All results throughout this manuscript are based on these concatenated out-of-fold predictions from a single five-fold CV run.

The SWS time series exhibits an auto-correlation for up to 4 days and recurring auto-correlation peaks every multiple of 27 days, caused by long-lasting coronal holes that become geoeffective with each solar rotation. To avoid data leakage caused by recurring coronal holes appearing in both training and test data, we discard 90 days of data between training, validation, and test data sets. Furthermore, the state of the solar corona and correspondingly the solar wind parameters vary over the solar cycle (Richardson et al., 2000; Tsurutani et al., 2006). Therefore, we ensure that data subsets are distributed across the solar cycle. To form the five subsets, we divide the data set into 20 continuous chunks of approximately nine months each, and assign to each subset four of those chunks, while ensuring that each data point appears exactly once as test data. The resulting CV splits can be seen in Figure 5.

4.3. Hyperparameter Optimization

To identify optimal hyperparameters, we run a hyperparameter optimization with the tree-structured Parzen estimator (TPE), a sequential, greedy algorithm based on the expected improvement criterion (Bergstra et al., 2011, 2013). For the μ prediction head, we optimize the expected value prediction of the model, consequently minimizing the PS derived from the RMSEs, while for the σ prediction head, we optimize the uncertainty prediction of the model by minimizing the PS derived from the CRPSs. For each head, we tune the learning rate, batch size, hidden layer size, and dropout rate. We perform 100 TPE iterations for the μ head and 150 iterations for the σ head. The resulting best-performing hyperparameters are reported in Appendix F.

Since each hyperparameter trial is evaluated on the same cross-validation splits with hyperparameters selected based on the aggregated cross-validation PS, this introduces a well-known selection bias (Cawley & Talbot, 2010), which can lead to overoptimistic generalization estimates. A fully nested cross-validation, with an outer loop held out of the optimization entirely, would be the standard solution, but our current optimization already requires several weeks of compute time to perform all experiments, making an additional outer loop computationally prohibitive. We instead mitigate this risk through high dropout rates and a modest optimization budget, and further evaluate our models on a new unseen data period excluded from this entire procedure (Section 8) to assess whether generalization is compromised.

5. Study of Solar Image Channels and Magnetograms

In this section, we study which SDO AIA solar image channels, together with HMI magnetograms, are most suitable for SWS forecasting by evaluating a range of input combinations. First, we remove the set of physical input features, which informs the model about the previous solar wind conditions and the state of the solar cycle, and compare models that only use image data as input. These results are investigated in Section 5.1. Then, we use the full set of available input features, which is analyzed in Section 5.2. For each model, we perform the hyperparameter optimization described in the previous section. The optimized hyperparameters are reported in Appendix F. Figure 6 summarizes the CRPS prediction scores, visualizing the trade-off between timeline and HSS peak value errors for both model classes. In the following, we analyze the results.

5.1. Image-Only Model

Table 1 shows the metrics for the models that use only solar images as input. Figure 6a visualizes the CRPS prediction scores. The PS isolines are approximately parallel to the highlighted isoline (dashed curve) of the best model.

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Für Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

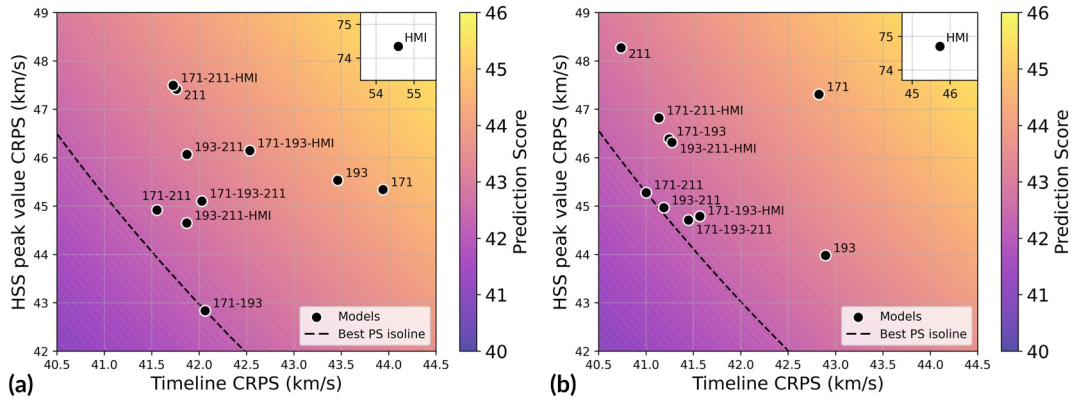


Figure 6. Timeline continuous ranked probability score (CRPS) versus high-speed solar wind stream peak value CRPS of models for different combinations of solar image and magnetogram inputs for the period 2010–2024. The trade-off between both errors is visualized by the prediction score (PS), which combines both metrics. The dashed curve shows the isoline of the best model's PS. The point labels show the used inputs. Panel (a) shows models that only use images as input (see Table 1). Panel (b) shows models that use images and the available physical input features as input (see Table 2).

Among the single-channel inputs, the 211 Å channel input leads to the best SWS prediction performance, due to its advantage for the timeline metrics. Although the HSS peak value errors of the 211 model are slightly worse than those of the 171 and 193 models, it achieves the best PS metrics of all single-channel models and also yields the best HSS event metrics. Especially the FAR is smaller, which is likely due to fewer filament channels being misclassified as coronal holes. The two other AIA channels lead to worse performance in the timeline metrics but facilitate more accurate HSS peak value prediction, although the peak CRPS differences among the single-channel models are small relative to their uncertainty. We conclude that the 211 Å channel contains the most useful information for the general model accuracy, whereas the 171 and 193 Å channels contain better information about HSSs. The HMI model is a clear outlier. All metrics are substantially worse than those of all other models, well beyond the reported uncertainty, indicating that our model cannot extract useful information from the HMI channel.

This can be physically explained since coronal holes are defined by open coronal magnetic topology, visible directly as intensity deficits in coronal EUV emission (our AIA channels), whereas HMI's photospheric line-of-sight field does not directly express this topology without additional coronal extrapolation (e.g., PFSS modeling),

Table 1
Evaluation Metrics of Image-Only Models for Different Combinations of Solar Image and Magnetogram Inputs

	Timeline		HSS peak values		HSS events			Prediction score	
Channel(s)	CRPS	RMSE	CRPS	RMSE	POD	FAR	TS	CRPS	RMSE
171	43.9 (±1.5)	81.0 (±2.7)	45.3 (±5.1)	78.3 (±9.2)	0.65	0.18	0.57	44.4	80.1
193	43.5 (±1.6)	80.0 (±2.8)	45.5 (±4.7)	77.7 (±9.2)	0.66	0.21	0.56	44.1	79.3
211	41.8 (±1.5)	77.0 (±2.7)	47.4 (±4.7)	80.3 (±9.2)	0.66	0.15	**0.59**	43.5	78.0
HMI	54.6 (±2.2)	98.3 (±3.4)	74.3 (±9.3)	119.7 (±11.6)	0.38	0.45	0.29	59.9	104.5
171–193	42.1 (±1.4)	77.6 (±2.5)	**42.8** (±4.3)	**73.5** (±8.5)	**0.67**	0.16	**0.59**	**42.3**	**76.1**
171–211	**41.6** (±1.3)	**76.9** (±2.5)	44.9 (±4.4)	75.7 (±7.6)	0.66	0.17	0.58	42.6	76.5
193–211	41.9 (±1.3)	77.1 (±2.6)	46.1 (±4.5)	79.0 (±9.3)	0.66	0.15	**0.59**	43.2	77.7
171–193–211	42.0 (±1.5)	77.4 (±2.7)	45.1 (±4.7)	76.0 (±8.9)	0.62	**0.14**	0.56	43.0	76.9
171–193-HMI	42.5 (±1.4)	78.4 (±2.5)	46.1 (±5.2)	79.1 (±9.8)	0.57	0.19	0.50	43.7	78.6
171–211-HMI	41.7 (±1.5)	**76.9** (±2.6)	47.5 (±4.7)	80.3 (±9.1)	**0.67**	0.18	0.58	43.5	78.0
193–211-HMI	41.9 (±1.4)	77.3 (±2.5)	44.7 (±4.7)	77.0 (±9.4)	0.65	0.20	0.56	42.8	77.2

Note. CRPS and RMSE values are in km/s. The 95% confidence interval half-width is shown in the second row for each model, where available. Bold numbers indicate the best value per column.

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

which was likely not learned by the model due to a limited data set. Current HMI data products are further affected by known systematic effects (e.g., limb-distance and magnetic-element-size calibration uncertainties, and the lack of far-side field information), so we do not conclude that magnetograms are generally unhelpful for SWS forecasting, only that they add little value in our specific setup.

When combining the 193 Å with the 211 Å channel, the only effect is a small improvement of the HSS peak value metrics, and the PS metrics barely change. This suggests that the information in the 193 Å channel overlaps strongly with that of the 211 Å channel. In contrast, the 171 Å channel, although performing worse when used as a single input, improves both the timeline and the HSS peak value metrics when combined with one of the two other AIA channels. Also, the event detection metrics of the 171–193 model benefit. We conclude that the 171 Å channel provides complementary information to the 193 Å and the 211 Å channels.

Adding a third AIA channel does not further improve the results, besides a small advantage for the addition of the 171 Å channel to the 193–211 model. This confirms the information overlap between the 193 and 211 Å channels and hints again at the complementary information of the 171 Å channel. Also, adding HMI to any two-channel AIA combination does not improve timeline metrics and only yields a slight increase in the HSS peak value metrics for the 193–211 model. This again highlights the limited contribution of HMI for SWS forecasting in our setup.

Overall, the combination of 171 Å with one of the two other AIA channels delivers the best performance, with the 171–193 model slightly outperforming the 171–211 model, but we note that differences of the metrics are within the uncertainty bounds. It achieves a timeline CRPS of 42.1 km/s and a timeline RMSE of 77.6 km/s, and a CRPS of 42.8 km/s and RMSE of 73.5 km/s for the HSS peak values, resulting in the best CRPS PS of 42.3 km/s and RMSE PS of 76.1 km/s. Its POD of 0.67 and FAR of 0.16 yield a TS of 0.59.

Besides the HMI model, differences between channels are generally small, but their ranking is robust against changes in the PS weights. To achieve a good performance, it seems sufficient for the model to identify coronal holes in the images, and the specific channel choice mainly fine-tunes the accuracy rather than leading to major improvements. Due to the negative HMI results, we did not test additional combinations of a single AIA channel with HMI.

In summary, we find that the 193 Å and the 211 Å channels contain largely overlapping information, while 171 Å adds complementary features to them. The HMI channel does not provide useful information in our approach. Therefore, combinations of two AIA channels yield the best performance, with the 171–193 model outperforming the other combinations.

5.2. Full Model

The results of the channel study for the full model, including the physical features, are shown in Table 2. Figure 6b visualizes the corresponding CRPS prediction scores.

All models with physical features improve their timeline metrics as compared to the image-only models. However, for the HSS peak value metrics, there is no clear trend: 5 out of 11 models improve their CRPS and 6 out of 11 improve their RMSE. In combination, 9 of the 11 models improve both prediction scores. For the HSS event detection metrics, we observe a similarly positive trend. The POD increases for most models, accompanied by a slightly higher FAR. In total, 7 out of 11 models improve their TS, three remain unchanged, and only one worsens.

Particularly for the HMI model, the timeline, HSS event, and PS metrics strongly improve, which indicates that the physical features are highly informative. However, since the same features only slightly improve the AIA-based models, it follows that the information content in the physical features strongly overlaps with the information content from the AIA images. This is plausible since the physical features encode the solar wind from recurrent coronal holes and the state of the solar cycle, which can also partially be inferred from the images. Additionally, we conclude that the information provided by the physical features is useful to increase the accuracy of the timeline metrics, but that they do not improve HSS peak predictions.

The relative ordering of the channels is similar to that of the image-only models (Section 5.1). Adding a second channel improves the performance of single-channel models, while a third channel or magnetogram supplies redundant information with the other channels and therefore provides no further advantages. However, including the physical features seems to compensate for some of the differences between the channels, because the model

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Für Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 2
Evaluation Metrics of Full Models (Using All Available Inputs, Including Physical Features) for Different Combinations of Solar Image and Magnetogram Inputs

	Timeline		HSS peak values		HSS events			Prediction score	
Channel(s)	CRPS	RMSE	CRPS	RMSE	POD	FAR	TS	CRPS	RMSE
171	42.8 (±1.5)	78.7 (±2.7)	47.3 (±4.8)	81.2 (±9.7)	**0.69**	0.16	**0.61**	44.2	79.5
193	42.9 (±1.7)	79.1 (±3.1)	**44.0** (±4.7)	75.9 (±10.7)	0.64	0.18	0.57	43.2	78.0
211	**40.7** (±1.4)	**75.0** (±2.5)	48.3 (±5.5)	83.0 (±11.2)	0.66	0.15	0.59	43.0	77.5
HMI	45.7 (±1.8)	84.2 (±3.0)	74.7 (±7.0)	118.2 (±10.9)	0.54	0.18	0.48	52.5	93.2
171–193	41.2 (±1.5)	75.8 (±2.8)	46.4 (±4.8)	79.4 (±8.8)	0.67	**0.13**	**0.61**	42.8	76.9
171–211	41.0 (±1.3)	75.8 (±2.5)	45.3 (±4.6)	77.1 (±9.3)	0.68	0.17	0.60	**42.3**	76.2
193–211	41.2 (±1.4)	76.0 (±2.5)	45.0 (±5.6)	77.8 (±10.8)	0.64	0.16	0.57	42.4	76.6
171–193–211	41.4 (±1.4)	75.9 (±2.7)	44.7 (±4.3)	**75.5** (±9.5)	0.68	0.17	0.60	42.5	**75.7**
171–193–HMI	41.6 (± 1.5)	76.8 (±2.7)	44.8 (±5.0)	78.1 (±10.4)	**0.69**	0.19	0.60	42.6	77.2
171–211–HMI	41.1 (±1.4)	76.1 (±2.4)	46.8 (±5.0)	80.6 (±9.8)	0.68	0.20	0.58	42.9	77.5
193–211–HMI	41.3 (±1.4)	76.4 (±2.6)	46.3 (±5.2)	80.2 (±10.8)	0.66	0.16	0.59	42.8	77.7

Note. CRPS and RMSE values are in km/s. The 95% confidence interval half-width is shown in the second row for each model, where available. Bold numbers indicate the best value per column.

performances are clustered closer together, which can be well seen in Figure 6. The only strong outlier remains the HMI model, although it improves with the physical features.

Notably, the 211 model performs best for the timeline metrics among all models of the study, and the 193 Å model does so for the HSS peak metrics. This is consistent with Brown et al. (2022), who also used the 211 Å channel for their model to optimize the timeline RMSE, and with the ESWF model (Milošić et al., 2023; Reiss et al., 2016), which uses the 193 Å channel for improved HSS peak predictions. However, the 193 and 211 Å models are not well-balanced between the two metrics, which is why the two-channel models achieve better prediction scores.

The best-performing model again uses a combination of 171 Å with another AIA channel, here being the 171–211 model. It has the lowest CRPS PS of 42.3 km/s, with a timeline CRPS of 41.0 km/s and timeline RMSE of 75.8 km/s, and an HSS peak value CRPS of 45.3 km/s and RMSE of 77.1 km/s. It also has the second-best RMSE PS of 76.2 km/s, and its POD of 0.68 and FAR of 0.18 produce a TS of 0.59.

However, the PS differences among the top models (171–211, 193–211, and 171–193–211: 42.3, 42.4, and 42.5 km/s, respectively) are small, and within the uncertainty bounds of timeline and peak metrics. We further verified that the relative ranking within this group is sensitive to the specific choice of PS weighting. While our default weighting ($w_t = 2, w_p = 1$) favors 171–211, an equal weighting ($w_t = w_p = 1$) would instead favor 193–211. We therefore do not interpret 171–211 as a uniquely superior configuration, but select it as a representative choice among several comparable models, based additionally on its favorable general timeline accuracy.

In summary, we find that incorporating the physical features into the model improves its overall accuracy. The advantage is consistent, albeit modest, as their informative value overlaps with the content of the input solar images. We choose the 171–211 model as a representative choice of the best-performing configurations and use it for all further analysis in this work. All models and their predictions are publicly available in Collin et al. (2026b).

6. Analysis of the Proposed Method

In this section, we analyze the prediction capabilities of our best model, which uses the 171 and 211 Å solar image channels and the available physical features as input. We show especially the advantage of PROSWIN over mean-based models. In Section 6.1, we focus on the probabilistic calibration accuracy, in Section 6.2, we analyze HSS predictions, and in Section 6.3, we monitor the performance over the solar cycle. A comparison to other methods is presented separately in Section 7.

6.1. Probabilistic Calibration

In this section, we assess the probabilistic calibration of the model, that is, whether the predicted distributions provide realistic estimates of the uncertainty. Although the CRPS already provides a quantitative measure of probabilistic skill, we focus here on a qualitative assessment and the advantages of probabilistic over single-value predictions.

One of the most important aspects of uncertainty quantification is being able to compute PIs. A PI of level α is expected to contain the observation with probability α if the model is well-calibrated. They can be directly obtained from the quantile function of the predicted distribution, as shown in Section 3.3. Figure 7a shows the predicted expected value and multiple PIs for several HSSs during the declining phase of solar cycle 24. The expected value, which is the only output we would get from a mean-based model, captures the time series and the HSS peaks generally well. Nevertheless, we also observe deviations from the actual SWS, which are only captured by the additional information that the PIs provide. In this example, the 50% PI overlaps with most smaller deviations, and increasing the interval until the 95% PI captures nearly all observations, including the peaks of the HSSs which are underestimated by the expected value alone.

An additional risk quantification is the probability that the SWS exceeds a certain threshold. We obtain this quantity from the predicted CDF, as described in Section 3.3. Figure 7b shows an example of probabilities for exceeding 500, 600, and 700 km/s during a time period near solar minimum. Besides the one HSS that the model clearly misses, it predicts high probabilities of at least around 80% for exceeding 500 km/s for the majority of HSSs, medium probabilities around 30% for exceeding 600 km/s, and low probabilities around 10% for exceeding 700 km/s. The observed HSS peak speeds around 600–650 km/s are slightly higher than one would expect from the predicted probabilities, but generally consistent with the predictions, indicating realistic probabilities. The peaks of the expected values are close to the observed ones, but they provide no uncertainty or risk estimates.

We evaluate these predictions using the BSS, as there is an increasing class imbalance in the data set for higher thresholds. The skill scores are 0.256, 0.154, and −0.095, respectively. This reflects a good skill of the model for the thresholds of 500 and 600 km/s, as it provides a relatively large advantage over naively using the event frequencies in the data set as probabilities. However, we also see a decreasing skill for higher SWS values, becoming slightly negative for the threshold of 700 km/s. Although this indicates a slight underestimation of the probabilities for the most important data points, these make up a fraction of only 0.6% in the data set. 54% of these points correspond to CMEs, which our model cannot predict, and the other 46% concentrate around the peak phase of strong HSSs, with a median distance of 11 hr to the peak. Thus, statistically, these are outlier points and do not affect the capability of the model to predict the occurrence probabilities of HSSs, as illustrated in the next section.

We can assess the quality of the PIs, and thereby the overall calibration of the predictions, by using Figure 7c. It shows for all PI levels between 1% and 99% the fraction of observations falling into the corresponding PI. For a perfectly calibrated model, this fraction of observations is exactly α, that is, the predicted probability of an observation being within the PI. For our model, we observe minimal deviations from the ideal diagonal for the timeline predictions. The average absolute deviation is just 0.9%. For the subset of predicted HSS peak values, the deviation is slightly larger, but with an average absolute deviation of 2.9%, it remains small overall. These results demonstrate that the uncertainty estimates of our model are highly accurate.

Finally, we assess how well the model predicts the underlying distribution of observations. Figure 7d compares the histogram of all observations with the histogram of the single-value predictions, that is, the expected values, and the predicted density of the observations, that is, the averaged predicted distributions. We see that the single-value predictions are clustered around the mean of the observed distribution and underestimate the heavy tails. This is consistent with findings of previous mean-based models (Bailey et al., 2021; Brown et al., 2022; Collin et al., 2025; Raju & Das, 2021; Upendran et al., 2020). In contrast, the probabilistic predictions reproduce the observed distribution almost perfectly. This shows that the model not only predicts realistic uncertainties for individual time steps but also captures the important data points in the upper tail of the distribution.

In summary, we find that from the probabilistic forecasts, we can extract a variety of useful information for operational SWS prediction. The model predicts realistic probabilities for fast solar wind conditions and very accurately calibrated uncertainties, enabling much more comprehensive forecasts than mean-based (i.e., single-value prediction) methods.

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

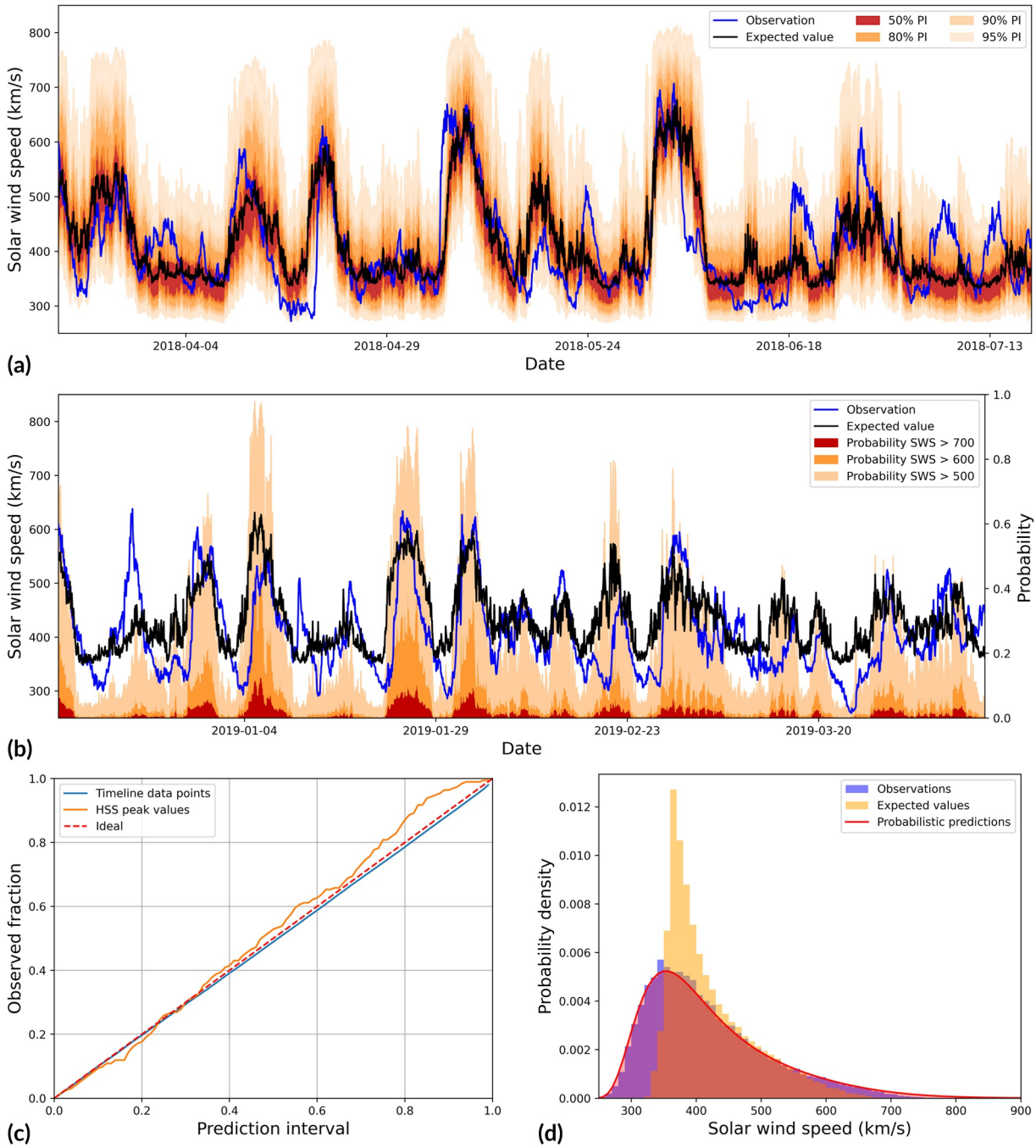


Figure 7. (a) Predicted distributions, shown as expected values with prediction intervals (PIs). All deviations of the expected value from the observations are captured within the predicted uncertainty. (b) Predicted probabilities of the solar wind speed exceeding 500, 600, and 700 km/s, represented by the upper boundary of the corresponding shaded areas. The second *y*-axis gives the scale of that probability. These probabilities provide explicit risk quantification beyond single-value predictions. (c) Calibration curve for predicted distributions of the timeline and high-speed solar wind stream peak values. For PI levels between 0 and 1, the corresponding fraction of observations within each PI is shown. The close agreement with the diagonal indicates very well-calibrated uncertainties. (d) Histograms of observations, expected value single-value predictions, and aggregated predicted probability densities. The predicted distributions closely match the observed distribution, while the single-value predictions are biased toward the mean and underestimate the heavy tail of high speeds.

6.2. High-Speed Streams

In this section, we investigate the capability of the model to forecast the occurrence and peak velocities of HSSs. Similar to the previous section, we derive an HSS occurrence probability from the predicted velocity distributions

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

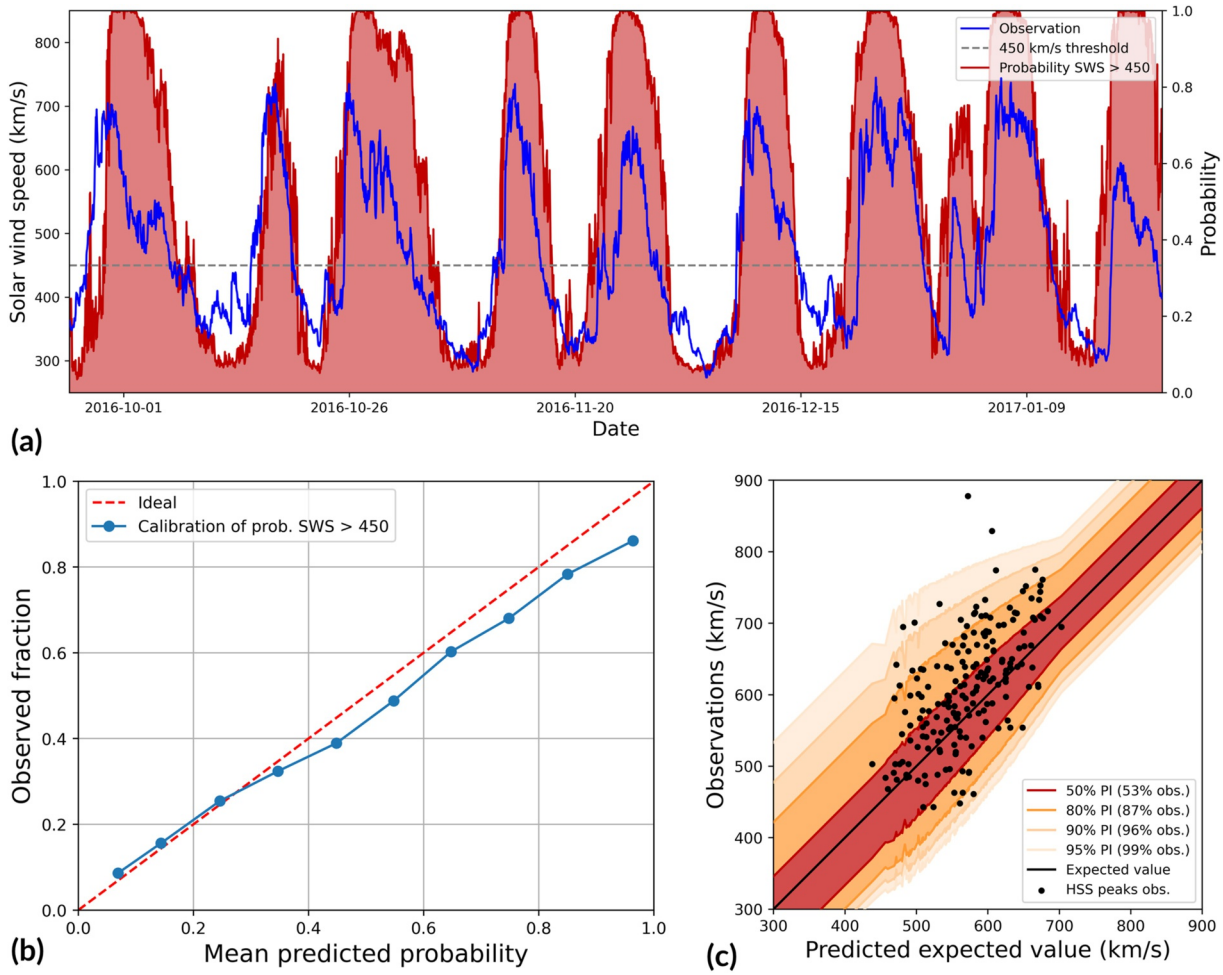


Figure 8. (a) Predicted probability that the solar wind speed (SWS) exceeds 450 km/s, used as a proxy for high-speed solar wind stream (HSS) occurrence. Probabilities above 50% agree well with observed HSSs and provide a probabilistic criterion for the occurrence. (b) Calibration curve for the probability that the SWS exceeds 450 km/s. Predictions are grouped into 10% bins. For each bin, the mean predicted probability and observed event frequency are shown. The close agreement with the diagonal indicates a good calibration for our model. (c) Predicted versus observed HSS peak speeds. The predictions are ordered based on the expected value, and the prediction intervals (PIs) are shown. The actual observations are shown as black dots. The frequency with which observations are contained in a PI is indicated in the legend. The observations scatter around the ideal diagonal line, and, apart from two outliers, the uncertainty is well captured.

by computing the probability that the SWS exceeds 450 km/s Figure 8a illustrates this probability for a sequence of HSSs during the declining phase of solar cycle 24. We see that the probabilities closely follow the observed increase of the SWS and remain high until the end of the event, indicating a useful proxy for the occurrence of HSSs. The BSS is 0.243, proving a high skill of the model for predicting the occurrence probabilities.

If we classify all time steps with an HSS probability exceeding 50% as HSSs and compare them with the (ground-truth) HSS list defined in Section 2.4, we obtain an accuracy of 74%. To further assess the reliability of the predicted probabilities, we compute a calibration curve, shown in Figure 8b. A calibration curve evaluates how well predicted probabilities correspond to the observed event frequencies. In contrast to the calibration curve from Figure 7c, we now evaluate event probabilities instead of PIs. Thus, to construct the curve, we first group the predicted probabilities into bins of 10%. For each bin, we then compute the average predicted event occurrence probability and the corresponding fraction of observed HSS events. Plotting these two quantities against each other yields the calibration curve. We notice a good calibration up to ≈40%. For larger probabilities, we find a small negative deviation, meaning that the model slightly underestimates the likelihood of HSSs. The deviation, however, is at most 10%, which still indicates a good calibration.

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

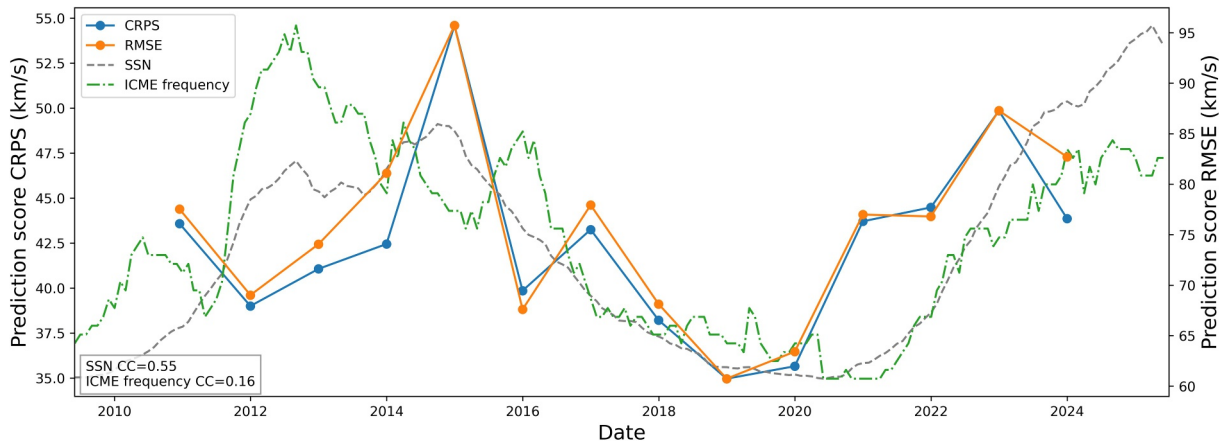


Figure 9. Yearly continuous ranked probability score and root mean square error prediction scores through Solar Cycles 24 and 25. The smoothed monthly sunspot number indicates the phase of the solar cycle. The smoothed monthly interplanetary coronal mass ejection (ICME) frequency indicates the number of coronal mass ejection disturbances of the solar wind speed time series. Both are normalized to the same scale as the prediction scores. The model performance is approximately inversely related to the solar activity, but the correlation with the ICME frequency is low.

We also evaluate the predicted HSS peak values. Figure 8c shows, ordered by expected value, all HSS peak predictions and their uncertainties. The actual corresponding observations are shown as scattered points. For perfectly accurate predictions, all observations would be on the diagonal, that is, equal to the expected values. The presented results show that the model captures the spread of observed peak values well within the predicted uncertainty bounds. We find that the predicted uncertainties are slightly too wide, with 53% of peaks being observed within the 50% PI, 87% within the 80% PI, 96% within the 90% PI, and 99% within the 95% PI. There are two outliers above 800 km/s, which the model does not predict well. One corresponds to a CME–HSS interaction where the CME was not labeled in the ICME list, and the other one is an HSS preceded by multiple other HSSs, which likely precondition interplanetary space for the last HSS, leading to this unusually high speed. We observe a slightly smaller predicted uncertainty for higher expected values, which might be due to the regularization of σ, limiting the uncertainty from above. Further, we find that there are more observations above the expected value than below, indicating a slight underestimation of the fastest HSSs.

In summary, we can confirm that our model predicts the probability of the occurrence of HSSs with good accuracy and reliably quantifies the uncertainties of their peak values.

6.3. Solar Cycle

Finally, we analyze the model performance over the solar cycle. Figure 9 shows the yearly CRPS and RMSE PS values together with the smoothed monthly sunspot number, indicating the phase of the solar cycle. We see that both metrics generally follow the trend of the sunspot number, showing a decrease in accuracy with rising solar activity and an increase in accuracy with declining solar activity. Accordingly, we find the best model performance during the solar minimum in the years 2018–2020 and the worst performance around the solar maximum of Solar Cycle 24 in 2015. Interestingly, the performance remains relatively good during most of the rising phase of Solar Cycle 24, that is, during the years 2010–2013, and only declines strongly for the peak of the solar activity. In contrast, in Solar Cycle 25, the performance starts to decline much earlier, from the start of the rising phase in 2021 onward, consistent with Solar Cycle 25 being a stronger cycle overall, which can be seen by the significantly higher occurrence of sunspots.

To quantify the driver of this solar cycle dependence, we compare the CRPS PS against both the smoothed monthly sunspot number and the monthly ICME frequency according to the Richardson and Cane catalog (Richardson & Cane, 2024), smoothed using the same 13-month rolling mean as for the sunspot number. We find that the PS correlates substantially more with the sunspot number (CC = 0.55) than with ICME frequency (CC = 0.16). Since ICMEs cause short and high-amplitude prediction errors but do not by themselves explain the broader solar cycle-dependent degradation trend, this indicates that unpredicted CME disturbances alone are not the primary driver of the overall performance decline toward solar maximum. Instead, the stronger association

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

with the sunspot number suggests that the global transition of the solar magnetic field from a simple dipolar to a complex multipolar configuration during solar maximum plays a larger role. The increased occurrence of active regions and decreased stability of open field-line coronal holes during this phase likely degrade the predictability of the ambient background solar wind captured by image-based models. We note that this analysis relies on the completeness of the ICME catalog. If the completeness of the catalog itself decreases during high-activity periods, this could partially distort the correlation with the ICME frequency, and we cannot fully exclude a contribution from unidentified CME disturbances to the observed trend.

In summary, we find that our model performs well throughout the solar cycle, but worse at solar maximum than during the declining phase and the solar minimum, driven primarily by the increasing complexity of the solar magnetic field rather than by ICME frequency alone.

7. Comparison to Other Methods

In this section, we compare PROSWIN to a selection of simple statistical baseline models, variations of our own model, three machine learning models from the literature, and a physics-based ensemble approach based on HUXt. We restrict the comparison to machine learning approaches similar to ours and one physical ensemble model. Our analysis should therefore be understood as a comparison of a specific set of approaches, rather than covering all SWS forecasting methodologies. In Section 7.1, we perform our main comparison on data from solar cycle 24, where most models provide predictions. In Section 7.2, a second comparison is done to the ensemble approach of HUXt coupled to WSA on a shorter data period of solar cycle 25.

7.1. Main Comparison on Solar Cycle 24

As baseline models, we use simple probabilistic approaches that predict conditional normal distributions with constant standard deviation. The climatological model estimates the average of the training data as a constant μ prediction. The persistence model uses the SWS observed 27 days before the target time as the μ prediction (Bernoux et al., 2022). In addition, we include a CH baseline model following Collin et al. (2025), which fits a linear regression model to predict μ, using as input features the SWS measured 27 days earlier and the CH area observed 4 days earlier in the region between 45° northern and southern latitude around the solar equator and between 30° western and eastern longitude around the central meridian.

To test the core assumptions of our approach, we further evaluate simplified versions of PROSWIN. One variant assumes log-normal conditional distributions but uses a constant σ, removing the variability of the uncertainty. Another variant assumes normal conditional distributions with constant σ, which additionally removes the log-normal assumption.

For comparison to existing work, we include the published predictions from several models in the literature. We compare to the gradient boosting model of Bailey et al. (2021), which uses magnetic model properties as input. We further include the Swin Transformer model of Brown et al. (2022), which is based on 211 Å AIA images and employs a deep learning architecture similar to our μ prediction pipeline. Their predictions are positively biased due to data leakage between their training and test data. Nevertheless, we include them because of the similarity of the prediction approaches. Additionally, we compare to the polynomial regression model of Collin et al. (2025), which is based mainly on CH area and SWS persistence features. We include both their polynomial output and their second model version that applies a distribution transformation (DT) at post-processing to approximate the observed SWS distribution. Finally, we include predictions from the HUXt model, a reduced-physics approach that produces solar wind flow simulations (Barnard & Owens, 2022). Specifically, we compare to probabilistic hindcasts produced by ensemble simulations from Edward-Inatimi et al. (2024), where HUXt is coupled with the MAS coronal model, and the probabilistic forecasts produced by ensembles from Edward-Inatimi et al. (2026), where HUXt is coupled with the WSA coronal model. MAS + HUXt predictions are concatenated simulations of full Carrington rotations, whereas for WSA + HUXt, we concatenate daily updated simulations of the prediction data closest to our 4-day lead time. Since the latter provides forecasts only for June 2020 until December 2020 and the full year 2023, which do not overlap with the other literature models, we perform a separate comparison to WSA + HUXt in Section 7.2.

Our distributional regression model is, to the best of our knowledge, the first fully probabilistic machine learning model for SWS prediction. As the comparison models from the literature are, besides the HUXt models, mean-

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table 3
Comparison of Our Approach to Other Models, Evaluated on the Intersection of Available Dates

	Timeline		HSS peak values		HSS events			Prediction score	
Model (estimator)	CRPS	RMSE	CRPS	RMSE	POD	FAR	TS	CRPS	RMSE
Climatological	55.0 (±3.0)	98.4 (±5.4)	–	–	–	–	–	–	–
CH baseline	45.2 (±1.8)	80.5 (±3.6)	78.7 (±13.2)	131.4 (±16.7)	0.65	**0.08**	0.61	52.7	92.4
Persistence	56.4 (±2.8)	103.0 (±5.1)	**43.9** (±7.8)	78.8 (±14.9)	0.70	0.26	0.56	51.5	93.4
Collin PR	45.8 (±1.6)	79.1 (±3.8)	68.7 (±10.6)	120.3 (±16.8)	0.72	0.12	**0.66**	51.5	89.3
Collin PR + DT	49.4 (±2.1)	86.6 (±4.1)	53.7 (±8.6)	95.2 (±17.2)	**0.74**	0.14	**0.66**	50.7	89.3
Bailey	46.3 (±1.6)	80.2 (±3.9)	60.6 (±8.9)	106.9 (±16.2)	0.68	0.16	0.60	50.3	87.5
Brown	44.4 (±1.5)	**76.2** (±3.9)	62.8 (±9.2)	110.2 (±16.4)	0.69	0.19	0.59	49.2	84.9
MAS + HUXt	56.9 (±4.1)	97.4 (±6.0)	53.9 (±9.7)	97.3 (±17.5)	0.69	0.22	0.58	55.8	97.3
Ours w/const. σ, log-normal	43.8 (±2.2)	79.7 (±3.8)	59.3 (±5.9)	**78.7** (±12.2)	0.72	0.22	0.60	48.0	**79.4**
Ours w/const. σ, normal	44.6 (±1.9)	79.7 (±3.8)	44.6 (±6.7)	**78.7** (±12.2)	0.72	0.22	0.60	44.6	**79.4**
Ours	**43.4** (±2.2)	79.7 (±3.8)	46.6 (±7.2)	**78.7** (±12.2)	0.72	0.22	0.60	**44.4**	**79.4**

Note. CRPS and RMSE are given in km/s. The 95% confidence interval half-width is shown in the second row for each model, where available. Bold numbers indicate the best values per column. PR = polynomial regression. DT = distribution transformation.

based, a direct comparison is not straightforward. Restricting the comparison to the RMSE of the expected value would neglect the probabilistic information provided by our approach. For this reason, we add a comparison in terms of the CRPS by exploiting the implicit assumption of the mean squared error loss, which was used to train all literature models, namely that predictions follow a normal distribution with constant variance. Based on this assumption, we extend all single-value prediction models to probabilistic ones by transforming their predictions into conditional normal distributions with mean μ given by the predicted value and constant σ. The value of σ is computed as the standard deviation of the SWS over the 11 years of data preceding the comparison period, corresponding approximately to the length of one solar cycle. This ensures that there is no intersection with the test data.

We restrict the evaluation to the intersection of available prediction times, ensuring that all models are assessed on the same data set. Small gaps in the predicted time series are linearly interpolated to facilitate HSS filtering. Most models have few gaps, besides the model of Brown et al., who provide their predictions with frequent gaps of around 4 days. The resulting evaluation period is June 2010 to October 2017, with some remaining larger gaps of multiple months. The results are summarized in Table 3 and Figure 10.

For timeline predictions, our full model outperforms all other models, achieving a CRPS of 43.4 km/s. The improvement over the climatological and persistence baselines is particularly large, that is, around 10 km/s for the CRPS and 20 km/s for the RMSE. We also outperform the Swin Transformer of Brown et al. (2022). Their model, however, has a better RMSE (only 3.5 km/s difference), meaning that the locations of their predicted probability distributions are more accurate. This naturally also leads to an advantage in the CRPS for their model. The fact that our CRPS is still smaller indicates higher probabilistic accuracy, that is, sharper predicted distributions, despite the slightly worse RMSE. MAS + HUXt performs considerably worse than all machine learning models, including the baselines, with a timeline CRPS of 56.9 km/s, comparable to the persistence baseline. The model with the second-best CRPS is the log-normal model with constant σ, which outperforms the normal model with constant σ, confirming that the log-normal assumption is beneficial for probabilistic SWS modeling. The CH baseline model performs surprisingly well (45.2 km/s CRPS and 80.5 km/s RMSE), given its simplicity, and is inferior only to the Swin Transformer models. We note that, given the 95% confidence intervals of the errors reported in Table 3, the differences between the top models with respect to the timeline errors are around the order of the error uncertainty, meaning that these results should be interpreted with caution rather than as absolute rankings.

For HSS peak value predictions, the persistence model performs best, with a CRPS of 43.9 km/s. The model benefits from not having a bias to underestimate the HSS peak values. The second-best model is the normal model with constant σ (44.6 km/s), which slightly outperforms our full model (46.6 km/s), although both achieve the

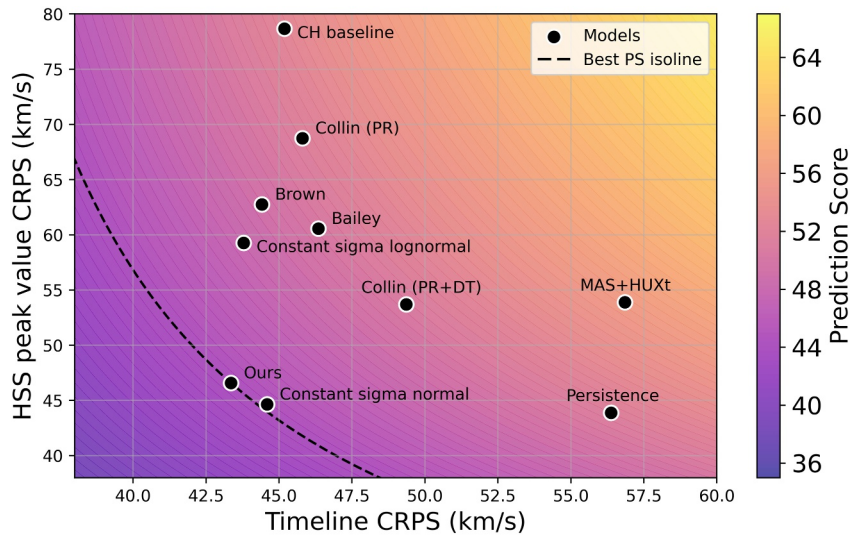


Figure 10. Timeline continuous ranked probability score (CRPS) versus high-speed solar wind stream peak value CRPS of our model compared to other models for the period 2010–2017. The trade-off between both errors is visualized by the prediction score (PS), which combines both metrics. The dashed curve shows the isoline of the best model's PS. PR = polynomial regression. DT = distribution transformation.

same HSS peak RMSE. This suggests that HSS peak values are better represented by normal than by log-normal distributions. The substantially worse performance of the log-normal model with constant σ confirms that the log-normal assumption breaks down for HSS peaks, due to the unphysically long tail of the log-normal distributions. Our full model mitigates this issue through regularization of the predicted σ, restricting it to physically realistic values. Even when ignoring the probabilistic aspect, PROSWIN strongly improves the prediction of the expected value compared to previous models from the literature, achieving an HSS peak RMSE of 78.7 km/s. MAS + HUXt performs relatively well, with an HSS peak CRPS of 53.9 km/s, outperforming the CH baseline and other machine learning models besides our own model variants.

For the HSS event detection metrics, differences between models are small. The DT model of Collin et al. (2025) achieves the best performance with a TS of 0.66, but all models perform comparably well in this category.

For the PS metrics, there is a clear trend. The models from the literature outperform the baseline models, with the exception of MAS + HUXt, while all variants of PROSWIN outperform the literature models. MAS + HUXt achieves the worst PS of all compared models, reflecting its comparatively poor timeline performance despite good HSS peak accuracy. Our full model achieves the best overall performance, with a CRPS PS of 44.4 km/s and an RMSE PS of 79.4 km/s. This demonstrates that our approach balances the competing prediction objectives well and that the combination of log-normal conditional distributions with variable uncertainty provides a benefit for probabilistic SWS forecasts, with a substantial improvement over existing models. The variants of our model with constant σ also outperform all baseline models, showing that the advantage of our approach is not only due to variable uncertainty, but also due to more accurate mean predictions. Notably, the PS advantage of the full model over the normal model with constant σ is relatively small.

The comparison across the three variants of our model also isolates the specific contribution of the distributional regression framework, independent of the model architecture, the choice of physical features and image channels, which are held fixed across all three and whose individual contributions are assessed separately (model architecture in Brown et al. (2022), physical features in Collin et al. (2025), image channels in Section 5). Moving from a constant σ log-normal model to our full model with additionally input-dependent σ leaves the timeline CRPS nearly unchanged (43.8–43.4 km/s) but substantially improves the HSS peak CRPS (59.3–46.6 km/s, a reduction of 21%) and the PS CRPS (48.0–44.4 km/s). This indicates that the primary benefit of explicitly modeling variable uncertainty is not improved average accuracy, which is instead driven mainly by the underlying architecture and input features, but substantially improved calibration and peak-value performance, which is consistent with the goal of probabilistic forecasting that motivates this work.

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In summary, PROSWIN outperforms all models in our comparison, providing the most accurate probabilistic SWS forecasts. Within the group of literature models, our model is the only one without a negative trade-off between timeline and HSS peak performance, thereby achieving a more actionable forecast than the comparison models. The results support our main assumption that the SWS is better modeled by log-normal conditional distributions with variable uncertainty than by normal distributions with constant uncertainty. At the same time, we find that HSS peak values are more accurately described by normal distributions, suggesting to model peaks separately.

7.2. Comparison to HUXt on Solar Cycle 25

Finally, we compare our full model to WSA + HUXt, over the shorter period of 1.5 years in 2020 and 2023 in which WSA + HUXt predictions are available. Although the comparison period is short and can therefore not be viewed as fully representative, this might be the most interesting comparison, as WSA + HUXt yields actual probabilistic forecasts obtained from a physical simulation model that is routinely used for operational forecasting. We find that our model achieves a better timeline accuracy (CRPS of 40.9 (±3.2) versus 60.8 km/s (±5.5)) and substantially better calibrated PIs (mean absolute deviation of 0.02 vs. 0.28) than WSA + HUXt, while WSA + HUXt achieves better HSS peak accuracy (CRPS of 43.4 (±16.6) versus 76.0 km/s (±19.6)). For the HSS event metrics, both models have comparable PODs (ours of 0.58 vs. WSA + HUXt of 0.54), while WSA + HUXt has an exceptionally high FAR of 0.48 compared to our FAR of 0.07, resulting in a TS of 0.36 for WSA + HUXt and a TS of 0.56 for our model. The overall CRPS PS favors our model in this comparison (PS of 48.4 vs. 53.7 km/s). These results highlight two fundamentally different weaknesses of both models. While our model underestimates most HSS during that time, resulting in the high peak error, WSA + HUXt contrarily overpredicts the frequency of HSSs, resulting in the high timeline error and FAR while having a high accuracy for peak value predictions when it correctly predicts a HSS occurrence. However, we emphasize not to over-interpret these results, given the large uncertainties of the peak metrics (±16.6 and ±19.6 km/s) for both models in this short evaluation window.

We summarize the direct comparison between the two HUXt models and our model with the finding that the predicted physical solar wind trajectories of HUXt enable accurate HSS peak predictions, suffering less from the underestimation bias of our model, but with a tendency to overestimate the frequency of HSSs. On the other hand, our probabilistic predictions are on average substantially more accurate and the predicted uncertainties significantly more reliable.

8. Application of PROSWIN to New Unseen Data

As a final experiment, we test whether the models which we have developed and cross-validated on data from June 2010 until June 2024 transfer to new unseen data. For that purpose, we use the data set extension of 2 years from July 2024 until June 2026, which was excluded from all model development, training, hyperparameter optimization, and previous comparisons. This data set encompasses 17,425 data points and we use it as input to produce predictions with the five models trained in the cross-validation run of the 171–211 model analyzed throughout this work. The models are not retrained or fine-tuned. An example of the obtained predictions can be seen in Figure 11 and the evaluation results are shared in Table 4, compared to the climatological and 27-day persistence baselines.

A qualitative assessment of the predicted time series in Figure 11 yields a good performance, as the model predictions capture the majority of SWS enhancements very well, and even underestimated HSS peaks are well within the predicted uncertainty intervals. Quantitatively, all five models outperform both baselines on the timeline metrics and on the combined PS, with the weakest fold still achieving a timeline CRPS of 50.5 km/s and a CRPS PS of 59.2 km/s, compared to 59.8 and 62.1 km/s, respectively, for persistence. The predicted uncertainties remain well calibrated, with a mean absolute deviation of the PIs of 2.4% (range 1.8%–3.3% across folds), just slightly worse than the 0.9% we reported on the data set until 2024, and essentially unchanged from the value of 2.4% obtained for the benchmark comparison period until 2017 (Section 7). Consistent with our previous findings, the persistence baseline again achieves the best HSS peak accuracy, and our model maintains its underestimation bias for HSS peak values, which is more pronounced in this period, with the best model giving an HSS peak CRPS of just 82.2 km/s. The Brier skill scores are positive across all evaluated thresholds, ranging from 0.28 for the threshold of 500 km/s, 0.17 for 600 km/s, to 0.07 for 700 km/s. Overall, these results confirm a high

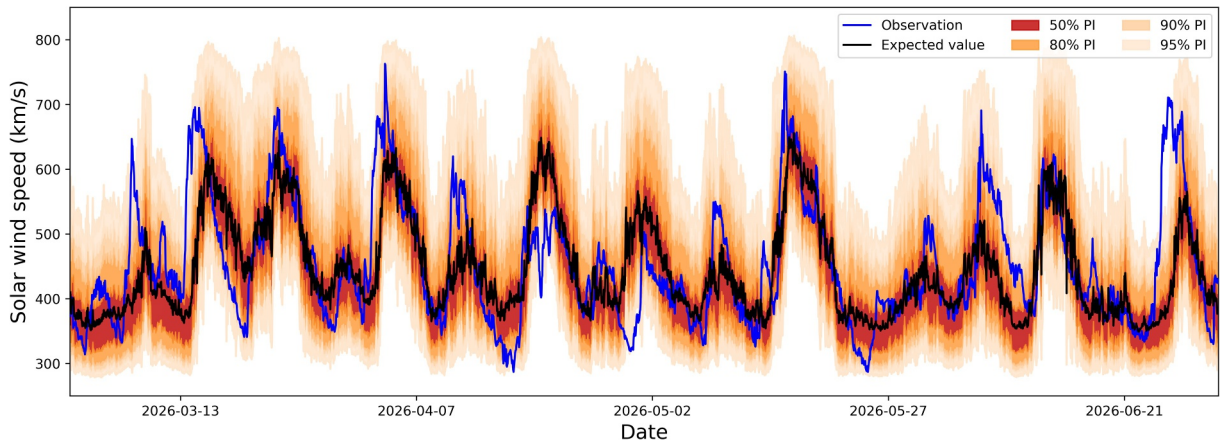


Figure 11. Probabilistic predictions of one of the five cross-validation models on new data. The observations are predicted well, indicating a good generalization of the model to new data.

prediction skill of the probabilistic forecasts on unseen data, besides the known underperformance against the persistence model for the HSS peaks.

The absolute errors in this period are larger than the values reported for our main data set. This is expected, as the evaluation period falls entirely within the maximum phase of solar cycle 25, the regime in which our model performs worst. Even more, given the relationship between the sunspot number and the model performance found in Section 6.3 and the fact that solar cycle 25 is substantially stronger than solar cycle 24, this explains the decrease in performance. For comparison, during the maximum of solar cycle 24 (July 2014 to June 2015), our cross-validated model achieves a CRPS PS of 54.6 km/s and an RMSE PS of 95.7 km/s, compared to 55.5 and 99.4 km/s on the unseen data. The performance in the 2 years from the maximum of solar cycle 25 is therefore comparable to, and only marginally worse than, the cross-validated performance during an equivalent but weaker solar maximum of the previous solar cycle. While this does not rule out a selection bias of the models from the hyperparameter optimization, it indicates that any such bias is small and that the generalization capability of our model is not compromised and estimated realistically by our cross-validation results.

9. Discussion and Limitations

The ultimate goal of any prediction framework should be operational application to provide a basis for real-world decisions by stakeholders. We demonstrate how probabilistic models provide important information to support this decision process, such as PIs and probabilities for hazardous events. Mean-based single-value prediction models do not provide such information, making them less applicable in operational settings. While PROSWIN is not intended as a standalone forecast for decision-making, its probabilistic SWS predictions could serve both as a general indicator of upcoming space weather risk and as input to downstream operational and research models that depend on hourly SWS averages. Examples of downstream models include real-time plasmasphere density models (e.g., PINE-RT; Bianco et al., 2023), operational K_p index forecasting models (e.g., Shprits et al., 2019;

Table 4
Performance on the Unseen Data From July 2024 to June 2026

	Timeline		HSS peak values		HSS events			Prediction score	
Model	CRPS	RMSE	CRPS	RMSE	POD	FAR	TS	CRPS	RMSE
Climatological	62.4	115.8	–	–	–	–	–	–	–
Persistence	59.8	109.9	**67.3**	**120.0**	**0.76**	**0.17**	**0.66**	62.1	113.1
Ours	**47.8**	**89.0**	82.2	129.8	0.68	0.22	0.57	**55.5**	**99.4**
	[46.5–50.5]	[87.2–92.6]	[69.7–91.1]	[117.6–140.3]	[0.60–0.72]	[0.18–0.29]	[0.48–0.62]	[52.7–59.2]	[96.5–104.3]

Note. For our model, the second row shows the range across the five cross-validation fold models. CRPS and RMSE are given in km/s. Bold numbers indicate the best value per column.

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Zhelavskaya et al., 2019), CME arrival time prediction models (e.g., Chierichini et al., 2024; Li et al., 2024; J. Liu et al., 2018; Yang et al., 2023), and radiation belt models (e.g., ORIENT-R; Chu et al., 2021). These models could significantly profit from having access to reliable SWS distributions with a lead time of 4 days, as this enables Monte Carlo sampling from the predicted distribution to generate ensembles of the downstream target application, and represents one necessary component toward extending their effective forecast lead time. However, probabilistic predictions are only useful if they are both accurate and well calibrated, which can be competing objectives (Camporeale & Berger, 2025; Camporeale & Carč, 2021). Developing models that balance accuracy and calibration can be challenging, and it is an important direction to further improve space weather forecasts.

We further note that the hourly cadence of our forecasts inherently limits their applicability to use cases that themselves operate on hourly or coarser timescales, such as the examples listed above. Applications requiring sub-hourly resolution are outside the scope of the present work.

Besides its usage in downstream applications, reliable SWS uncertainty is also relevant for the determination of the geoeffectiveness of space weather events. Alongside the interplanetary magnetic field (IMF) orientation (dominated by the southward component B_z), the SWS is one of the two primary drivers of solar wind-magnetosphere coupling (Newell et al., 2007). This means that a SWS forecast alone is incomplete for geomagnetic disturbance prediction, because a fast HSS with predominantly northward B_z may produce only modest geomagnetic activity, whereas a slower stream with sustained southward B_z can be considerably more geoeffective. However, reliable forecasting of B_z beyond very short lead times remains an open and largely unsolved challenge (Reiss et al., 2021). Given this asymmetry, a well-calibrated SWS forecast still provides meaningful operational value on its own. Since geoeffectiveness scales jointly with speed and IMF orientation, low forecast SWS provides a reliable indicator of low geomagnetic risk, while high forecast SWS indicates a necessary, though not sufficient, precondition for strong geoeffective coupling. Therefore, an accurate probabilistic SWS forecast is an important first step, complementary to future efforts addressing B_z prediction.

Within the described scope, we see the specific contribution of our work as follows. Unlike physics-based numerical models, our approach does not rely on simplifying physical assumptions, avoiding systematic biases associated with such approximations. Our approach further maps directly from solar imagery to Earth-arrival SWS distributions, without requiring a separately derived inner boundary condition, itself a source of error in propagation-based models. We therefore learn the uncertainty directly from the data itself instead of relying on perturbations of boundary conditions leading to correct uncertainties. Our direct comparison to HUXt (Section 7) provides quantitative support for these advantages, showing improved accuracy and substantially better calibration. We therefore position PROSWIN not as a general improvement to solar wind forecasting, but specifically as a computationally efficient, well-calibrated probabilistic alternative for applications operating at hourly to multi-day timescales.

In this work, we demonstrate that utilizing log-normal uncertainties is an effective way to improve probabilistic SWS forecasts, as they are well-suited to model the skewed and heavy-tailed distribution of the ambient SWS. This improvement, however, does not transfer to HSS peak velocities, whose uncertainty is better represented by normal distributions. This finding suggests the use of mixture models, combining one distribution for quiet-to-moderate solar wind conditions and another for upper-tail events. This approach could also help to reduce the underestimation bias of rare hazardous events and would remove the need to clip the predicted variance of log-normal distributions to avoid unphysically high uncertainty intervals.

A fundamental difficulty for most machine learning models in space weather is the strong class imbalance of data sets. While quiet times and moderately disturbed periods appear with high frequency, fast solar wind conditions, which are most important for space weather hazards, occur relatively rarely. Standard training algorithms minimize the overall error of the entire data set and therefore lead to models that perform well for average conditions, but neglect rare conditions in the upper tail of the distribution, systematically underestimating important events. Although we increase the training weights for HSSs to mitigate this issue, a small bias remains in our model. However, we can still capture strong events well by using distributional regression, which models rare events through higher quantiles of the predicted distribution. Statistically, these events are outliers in the data set and are thereby predicted with lower occurrence probabilities. While this does not resolve the underlying issue, it represents one way to address it. Ultimately, however, the goal should be to entirely remove the

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

underestimation bias of fast solar wind conditions, which requires further methodological developments for imbalanced regression scenarios (Camporeale, 2025).

Closely related to this issue is the trade-off between timeline accuracy and the HSS peak accuracy. Forcing a model to predict rare conditions more accurately typically implies stronger deviations from the mean across the entire data set, which increases the timeline error. This leads to competing optimization objectives, which we observe in our experiments and which was also reported in previous studies (e.g., Collin et al., 2025; Shprits et al., 2019). Consequently, it is not sufficient to compute errors averaged over the full data set alone, as it is typically done by solar wind prediction models. To address this, we introduce a new metric that balances these competing objectives by combining timeline and peak errors using a weighted harmonic mean, which we refer to as PS. Operationally, the PS should be interpreted as a model selection criterion rather than as a directly actionable forecast error in the way CRPS or RMSE are. The selected models show good performance across all solar wind conditions, rather than models that optimize one objective at the expense of the other, and we can confirm that our model is significantly better balanced in comparison to previously published SWS models. However, as we developed it for the assessment of HSS predictions, the metric relies on the previous association of predicted and observed HSSs. Further, it does not take into account the detection rate. The metric may be extended to address these weaknesses and to generalize to settings where explicit event detection and association are not required, for example, by using a conditional RMSE or CRPS, as suggested by Camporeale (2025).

A constraint of our work is that we trained our model with definitive level 1 solar images rather than near-real-time data. Although all needed input data is either available in real-time or has an equivalent near-real-time replacement, meaning that an operational implementation would be possible, this means that our work is more a test of potential than an actual assessment of the operational forecast quality.

Relatedly, our hyperparameter optimization procedure, evaluated on the same cross-validation splits used for our reported performance, carries an inherent risk of selection bias (Cawley & Talbot, 2010). While computational constraints prevented a fully nested cross-validation scheme, our evaluation on a new unseen data period (Section 8) showed a model performance consistent with expectations given the solar cycle phase, providing evidence that this bias, if present, is small.

A further consequence of the architecture of PROSWIN is that it predicts marginal distributions independently at each time point, rather than a jointly consistent forecast trajectory, unlike physical propagation models (e.g., HUXt; Barnard & Owens, 2022) or machine learning models that model explicit temporal dependencies (e.g., Gaussian Process approaches; Rasmussen, 2004). We find that prediction residuals are substantially autocorrelated across consecutive hours (the autocorrelation for lags of 1, 12, and 24 hr is 0.97, 0.76, and 0.56, respectively), indicating that independent per-timestep sampling would discard existing temporal structure in the model errors. This suggests that a Gaussian-copula-based trajectory sampling approach, coupling the marginals according to this measured dependence structure (Möller et al., 2013; Pinson & Girard, 2012), could offer a practical route to generating temporally coherent sample trajectories without modifying the predicted marginal distributions or any reported metric. Such a strategy would, however, capture only statistical, pairwise dependence, without guaranteeing that sampled trajectories correspond to physically realistic scenarios. Extensions of our data-driven approach to employ explicit time series models or to complement physical simulation models are natural directions for future work.

Finally, PROSWIN does not explicitly consider CMEs, as no direct information about CME occurrence or properties is provided as input. This restricts the operational applicability, as CMEs are responsible for the most severe geomagnetic storms. We note, however, that this does not necessarily mean CME-related information is entirely unavailable to the model. Active regions, a weak proxy for CME source likelihood, are clearly visible in the AIA channels and HMI magnetograms used as input. As a black-box model, we cannot rule out that some of that signal is implicitly exploited. We tested this by comparing the predicted σ around ICME arrivals against non-ICME periods. We find a small increase in σ of around 3% during ICME disturbance times in comparison to non-ICME periods, providing weak support for the interpretation that CME effects manifest as elevated uncertainty. This is consistent with our channel study (Section 5), where including HMI magnetograms, a higher-fidelity proxy for CME-relevant magnetic complexity, did not improve performance.

Incorporating CME predictions into machine learning SWS predictions remains a major challenge. Coronal mass ejections are primarily detected in coronagraph data, where parameter estimates are highly uncertain. As a result,

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

reliable machine learning CME forecasting tools are rare and have been evaluated only on relatively small data sets (e.g., Chierichini et al., 2024; Li et al., 2024; J. Liu et al., 2018; Yang et al., 2023). This highlights a structural advantage of physics-based propagation models, such as HUXt, in the context of CME forecasting. CMEs can be incorporated comparatively easily as simple parametrized injections. Developing a unified data-driven SWS forecast that combines HSSs and CMEs, therefore, is much harder and remains an open problem for future research.

10. Conclusion

In this study, we show that SWS forecasting can be done in a fully probabilistic way. We developed PROSWIN, a probabilistic machine learning approach to forecast calibrated SWS distributions. Our approach combines a Swin Transformer, which extracts features from multiple SDO solar image channels, with physical input features encoding the solar wind conditions from the previous solar rotation and the current state of the solar cycle. These features are processed by a neural network coupled to a distributional regression framework, which predicts the distributional parameters μ and σ of a log-normal conditional distribution of the SWS 4 days in advance. This allows a quantification of forecast uncertainty and the risk of fast solar wind conditions, which is of high relevance for operational solar wind forecasting and represents a substantial advantage compared to previous mean-based models.

To analyze PROSWIN, we defined a new class of error metrics called PS, which takes into account both timeline and HSS peak performance, and is designed to overcome the limitations of standard error metrics, which largely neglect rare but important conditions. Evaluation of our approach led to the following main results:

1. Our representatively selected best model uses 171 and 211 Å solar image input, although error differences to other models are small. The 171–211 Å model achieves a CRPS PS of 42.3 km/s for the years 2010–2024, based on a timeline CRPS of 41.0 km/s and an HSS peak value CRPS of 45.3 km/s. The expected value of the predicted distributions has a timeline RMSE of 75.8 km/s and an HSS peak value RMSE of 77.1 km/s.
2. PROSWIN outperforms the selected benchmark models for the combined PS metric and is the only model that does not have a negative trade-off between timeline and HSS peak performance. We find that optimizing for the PS leads to a significant improvement of the HSS peak metrics and results in more applicable solar wind prediction models. Interestingly, the model consistently achieving the best HSS peak value performance is a simple 27-day persistence baseline. Additionally, WSA + HUXt achieves a similar HSS peak performance on a shorter, non-overlapping period of solar cycle 25.
3. The predicted uncertainties are very well calibrated. PIs deviate on average by only 0.9% from the observed frequencies, and the aggregated predicted distribution closely matches the observed SWS distribution.
4. The risks of HSSs are accurately quantified. The model predicts HSS occurrence probabilities with errors of at most 10% compared to the observed frequencies. For the peak values, the predicted PIs deviate by only 2.9% on average from the observed frequencies.
5. Our results indicate that a large overlap of the information which is important for SWS prediction is contained in the 193 and 211 Å solar image channels, whereas the 171 Å channel seems to add complementary information to the previous ones. Adding physical features consistently improves model performance, albeit only by a small margin, indicating that a large fraction of the information contained in the used physical features is already encoded in the images.
6. The SWS is better described by log-normal conditional distributions with variable uncertainty than by normal distributions with constant uncertainty. At the same time, HSS peak values are more accurately represented by normal distributions.
7. The model performs best during the declining phase of the solar cycle and during solar minimum, and shows its weakest performance during solar maximum. The correlation of the error is substantially stronger with the sunspot number than with the ICME frequency.
8. Our trained models generalize well to new unseen data, although absolute performance dropped on our test period due to it being selected from around the maximum of a strong solar cycle.

Appendix A: Solar Image Preprocessing

We download 1024 × 1024 pixel images and preprocess the extreme ultraviolet (EUV) images following procedures adapted from Galvez et al. (2019), Jarolim et al. (2021), and Brown et al. (2022). The steps include: (a)

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

updating the satellite pointing information; (b) correcting the image by applying the instrumental point-spread function (Hofmeister, 2024; Hofmeister et al., 2025); (c) registering the images such that the solar disk is centered and rotated with solar north up; (d) scaling the image such that the solar radius corresponds to 976 arcsec; (e) correcting for instrument degradation; (f) normalizing by the exposure time; (g) further downsampling to 512 × 512 pixels by summing in local blocks to emulate lower-resolution observations; (h) cropping a 300 × 300 pixel square with corners that are approximately at the edges of the solar disk; (i) resizing to 224 × 224 pixels using cubic spline interpolation, which is the input size for the image encoder. If an image is missing, we use the closest existing image, up to a gap of 24 hr, and rotate the Sun in this image to the missing time by applying the differential solar rotation rate. For the magnetograms, we apply the same pipeline except for the pointing correction, deconvolution, degradation correction, and exposure normalization steps (steps 1, 2, 5, and 6).

Since the images have values spanning different orders of magnitude, we normalize them as suggested by Jarolim et al. (2021). For the EUV images, we clip the values to a channel-specific upper threshold. The thresholds are 6,457.5 for 171 Å, 7,757.31 for 193 Å, and 6,539.0 for 211 Å. Then, we rescale the values to [0, 1], apply an inverse hyperbolic sine function stretch $\operatorname{asinh}(x/a)/\operatorname{asinh}(1/a)$ with $a = 0.005$, and rescale to $[-1, 1]$. Magnetograms are clipped to the interval $[-100, 100]$ and rescaled to $[-1, 1]$. These steps enhance coronal-hole–specific features, reduce strong variations near active regions and ensure consistent numerical ranges across all image channels for neural network training.

Appendix B: Distributional Regression

To formally introduce distributional regression, let $\mathcal{D}(Y, \theta)$, for the target variable $Y \in \mathbb{R}$, be the distributional model, where $\mathcal{D}$ is a probability distribution with parametric density function $p(Y|\theta)$, defined by the K-dimensional distributional parameter vector $\theta = (\theta_1, \ldots, \theta_K)^\top \in \Theta := \Theta_1 \times \cdots \times \Theta_K \subseteq \mathbb{R}^K$. We then relate each distributional parameter θ_k, to possibly different subsets of the entire input vector $X \in \mathbb{R}^m$. For simplicity, let us write $X = x$ for all distributional parameters. This relation is established via parameter-specific predictors $\eta_k(x)$, such that $\theta_k \equiv \theta_k(x) = h_k\big(\eta_k(x)\big)$. Here, the functions $h_k : \mathbb{R} \mapsto \Theta_k$ are monotonic response functions, ensuring potential parameter space restrictions given through Θ_k. We furthermore denote by $l(\theta; Y_{1:n}) = \sum_{i=1}^{n} \log\{p\big(y_i | \theta(x_i)\big)\}$ the log-likelihood given some observations $Y_{1:n} = \big(y_1, \ldots, y_n\big)^\top$ and associated inputs $\{x_1, \ldots, x_n\}$.

In our model, we assume $Y \sim \text{Lognormal}(\mu(x), \sigma(x))$ within the distributional regression framework, where Y denotes the SWS target variable, and $\mu(x) = h_\mu\big(\eta_\mu(x)\big)$ and $\sigma(x) = h_\sigma\big(\eta_\sigma(x)\big)$ are functions predicting the distributional parameters θ, here μ and σ, based on the input x. To ensure that $\sigma(x)$ is positive for all $x \in \mathbb{R}^m$, we use $h_\sigma(x) = \exp(x)$, while h_μ is the identity. The predictor functions η_μ and η_σ are learned using a neural network. The parameters of the neural network are obtained by minimizing the negative log-likelihood $-l(\mu, \sigma; Y_{1:n})$ of the observations.

Appendix C: Details on Training Process

Before training, we standardize all physical input features by subtracting their mean and scaling them to unit variance. The target values are transformed by removing the location shift ℓ of the empirical log-normal distribution, applying a logarithm, and then standardizing, effectively changing the log-normal distributions to normal distributions. This allows the algorithm to fit normal distribution parameters, which is numerically more stable than using the original target variable directly in combination with a log-normal distribution. As mentioned in Appendix B, we apply an exponential function to the network's output σ, so that the model learns $\log(\sigma)$.

For a normal distribution, the parameters μ and σ decouple due to their orthogonality, which enables sequential training. First, we only train the image encoder together with the μ-head and discard the σ-head. Because μ represents the expected value of the normal distributions, we optimize it using a mean squared error loss function.

To improve the prediction of HSSs, we apply an exponential weighting scheme to the loss. For each HSS event (as defined in Section 2.4), the SWS values are scaled to [0, 1], and the weight function $w(y) = \alpha y^\gamma$ with $\alpha = 5$ and $\gamma = 3$ is applied. This shifts the training focus toward the HSS peaks and helps compensate for the tendency of

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

machine learning models to underpredict extremes when trained on imbalanced data. The weighting is only applied during the first training stage.

In the second training stage, we freeze both the parameters of the image encoder and the trained μ-head, and fit the σ-head alone. We use the negative log-likelihood of the normal distribution as the loss function with the log-transformed standardized SWS values as target variables. To prevent physically unrealistic uncertainty estimates, we impose an upper bound on the 99% quantile of the corresponding log-normal distribution. Since HSSs rarely exceed 800 km/s (Grandin et al., 2019) and there is no information about CMEs in the input data, we set an upper limit of 850 km/s and clamp σ accordingly, with the small margin above 800 km/s left to accommodate the higher speeds occasionally reached by ICMEs, which remain in the target distribution.

Additionally, we need to take into account that the μ-head predicts the conditional mean of a normal distribution, but when we transform back the distributional parameters into those of a log-normal distribution, the expected value also becomes dependent on σ. Thus, we apply a correction to μ, adjusting it such that the expected value of the log-normal distribution remains identical to the one learned during the first training stage. The mathematical details of these transformations and regularizations are given in Appendix D.

We train for up to 200 epochs and evaluate the loss function after each epoch on the validation data set. If the validation loss does not improve by more than 10^{-4} for 20 epochs, we stop training to avoid overfitting. The batch size and learning rate are determined by hyperparameter optimization, as described in Section 4.3.

Appendix D: Parameter Transformations and Regularization

We use a parametrization of the log-normal probability density function (PDF) that includes a location shift parameter ℓ, given by

$$p(y|\mu,\sigma) = \frac{1}{(y-\ell)\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\log(y-\ell)-\mu)^2}{2\sigma^2}\right\} \quad \text{for } y > \ell. \tag{D1}$$

The parameter ℓ is estimated by finding the best-fitting density p of on the training data. Then, the target distribution is transformed to a normal distribution by applying

$$f_{\text{trans}}(y) = \frac{\log(y-\ell)-\bar{m}}{\bar{s}},$$

to the targets, where $\bar{m}$ and $\bar{s}$ are the empirical mean and standard deviation of the log-transformed target $\log(y-\ell)$. In other words, we eliminate the location shift and standardize the targets.

The output of the neural network, which are the mean $\tilde{\mu}$ and standard deviation $\tilde{\sigma}$ of the standardized normal distribution, can be transformed back to the actual log-normal distributional parameters $\hat{\mu}$ and $\hat{\sigma}$ for the SWS prediction by means of

$$\hat{\mu} = \bar{s}\tilde{\mu} + \bar{m}, \qquad \hat{\sigma} = \bar{s}\tilde{\sigma}. \tag{D2}$$

Together with the initially estimated location parameter ℓ, this defines the conditional distribution via the PDF.

In the following, we need to differentiate between the output $\tilde{\mu}$ and $\tilde{\sigma}$ of the neural network, which are the distributional parameters of the normal distributions we fit during the training process, and the backtransformed parameters $\hat{\mu}$ and $\hat{\sigma}$, which are the distributional parameters of the log-normal distributions we use as the actual SWS predictions, as defined in the PDF in Equation D1.

To calculate the upper bounds on $\tilde{\sigma}$ during the training process, we need to find a bound that limits the 99%-quantile $Q_{0.99}$ of the predicted log-normal distribution to the chosen physical bound B, that is,

$$Q_{0.99}\,[Y|\hat{\mu},\hat{\sigma}] = \ell + \exp(\hat{\mu} + \hat{\sigma} z_{0.99}) \overset{!}{\leq} B,$$

where $z_{0.99}$ is the 99%-quantile of the standard normal distribution. Using Equation D2 and solving for $\tilde{\sigma}$ yields

$$\tilde{\sigma} \leq \frac{\log(B-\ell) - \bar{s}\tilde{\mu} - \bar{m}}{\bar{s}z_{0.99}},$$

which we ensure by cropping $\tilde{\sigma}$ in each iteration of training and during inference.

The μ-head predicts $\tilde{\mu}$ as the conditional mean of a normal distribution, which yields the value $\exp(\bar{s}\tilde{\mu} + \bar{m}) + \ell =: E$ when scaled back to the original target distribution. We need to ensure that the expected value $\mathbb{E}[Y]$ of our predicted log-normal distribution is equal to E. However, the expected value of a log-normal distribution is also influenced by $\hat{\sigma}$. Thus, after computing $\tilde{\sigma}$, we correct $\tilde{\mu}$ such that the expected value of our log-normal prediction becomes E, that is,

$$\mathbb{E}\left[Y|\hat{\mu}_{\text{corrected}}, \hat{\sigma}\right] = \ell + \exp\left(\hat{\mu}_{\text{corrected}} + \frac{\hat{\sigma}^2}{2}\right) \overset{!}{=} \exp\left(\bar{s}\tilde{\mu}_{\text{old}} + \bar{m}\right) + \ell.$$

Using again Equation D2 and solving for $\tilde{\mu}_{\text{corrected}}$ yields

$$\tilde{\mu}_{\text{corrected}} = \tilde{\mu}_{\text{old}} - \frac{\bar{s}\tilde{\sigma}^2}{2}.$$

The correction is applied after the cropping of $\tilde{\sigma}$. The combination of both steps ensures that the predicted distributions do not have unphysically large uncertainties and that the expected value remains unchanged during the training phase of the σ-head.

Appendix E: Probabilistic Prediction

After training, the neural network can be used for inference of arbitrary inputs x^*, that is, solar images and physical input features at a particular time point, and the output can be transformed back into log-normal parameter estimates $\hat{\mu}(x^*)$ and $\hat{\sigma}(x^*)$. These yield the prediction of a conditional SWS distribution $V_{\text{sw}} \sim \text{Lognormal}(\hat{\mu}, \hat{\sigma})$, characterized by the PDF

$$p(y|\hat{\mu}, \hat{\sigma}) = \frac{1}{(y-\ell)\hat{\sigma}\sqrt{2\pi}} \exp\left(-\frac{(\log(y-\ell) - \hat{\mu})^2}{2\hat{\sigma}^2}\right),$$

where ℓ is the location shift parameter derived by fitting a PDF to the empirical target distribution of the training data. The PDF represents the relative likelihood of V_{sw} attaining a specific value y. The area under the curve over an interval gives the probability of V_{sw} falling within that range.

We can now easily compute any quantity of interest for the predicted SWS distribution. For instance, the expected value is given by

$$\mathbb{E}[V_{\text{sw}}|\hat{\mu}, \hat{\sigma}] = \exp\left(\hat{\mu} + \frac{\hat{\sigma}^2}{2}\right) + \ell. \tag{E1}$$

The cumulative distribution function (CDF), which specifies the probability that V_{sw} stays below a threshold y, is defined as

$$F(y|\hat{\mu}, \hat{\sigma}) = \Phi\left(\frac{\log(y-\ell) - \hat{\mu}}{\hat{\sigma}}\right),$$

where Φ is the CDF of the standard normal distribution. We can also use the CDF to derive the probability that the SWS exceeds a certain threshold, by computing $1 - F(y|\hat{\mu}, \hat{\sigma})$. Finally, the quantile function, which is the inverse

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

of the CDF, returns the threshold value y for a given probability z, such that the probability of V_{sw} staying below y equals z. It is defined as

$$Q(z|\hat{\mu},\hat{\sigma}) = \exp\left(\hat{\mu} + \hat{\sigma}\Phi^{-1}(z)\right) + \ell, \quad z \in (0,1),$$

where Φ^{-1} is the inverse of the CDF or quantile function of the standard normal distribution. We can use the quantile function to construct PIs that contain V_{sw} with a specified probability level α. The α PI for a prediction is given by

$$[Q(\alpha/2|\hat{\mu},\hat{\sigma}), Q(1-\alpha/2|\hat{\mu},\hat{\sigma})], \quad \alpha \in (0,1).$$

Appendix F: Hyperparameters

In Tables F1 and F2, the optimized hyperparameters for all models presented in Sections 5.1 (image-only models) and 5.2 (full models) are shown, respectively.

Table F1
Optimized Model Hyperparameters Using Only Solar Images as Input, Used for Table 1

	Hidden layer size		Dropout		Batch size		Learning rate	
Channel(s)	μ	σ	μ	σ	μ	σ	μ	σ
171	181	182	0.581	0.714	64	16	5.521×10^{-6}	2.670×10^{-7}
193	159	147	0.659	0.741	32	16	5.659×10^{-6}	2.860×10^{-8}
211	142	198	0.627	0.710	16	16	1.902×10^{-7}	2.402×10^{-8}
HMI	176	100	0.563	0.908	16	16	9.905×10^{-6}	6.477×10^{-9}
171–193	169	57	0.510	0.879	64	16	3.188×10^{-6}	5.888×10^{-9}
171–211	116	151	0.624	0.742	64	16	3.463×10^{-6}	3.200×10^{-7}
193–211	144	151	0.506	0.849	64	16	3.801×10^{-7}	2.858×10^{-7}
171–193–211	141	60	0.622	0.724	16	16	3.324×10^{-7}	4.316×10^{-7}
171–193-HMI	177	115	0.603	0.737	16	32	2.745×10^{-7}	3.098×10^{-7}
171–211-HMI	105	181	0.552	0.848	32	16	3.962×10^{-6}	3.554×10^{-7}
193–211-HMI	191	190	0.592	0.803	32	16	2.975×10^{-6}	8.971×10^{-8}

Table F2
Optimized Model Hyperparameters Using All Available Features as Input Used for Table 2

	Hidden layer size		Dropout		Batch size		Learning rate	
Channel(s)	μ	σ	μ	σ	μ	σ	μ	σ
171	200	135	0.627	0.769	64	16	1.547×10^{-5}	2.457×10^{-7}
193	105	75	0.607	0.734	32	16	7.116×10^{-6}	4.217×10^{-7}
211	128	130	0.731	0.739	64	16	6.870×10^{-7}	7.976×10^{-8}
HMI	166	148	0.685	0.888	16	16	5.631×10^{-5}	1.805×10^{-7}
171–193	131	121	0.530	0.772	16	16	1.218×10^{-6}	1.649×10^{-8}
171–211	168	143	0.666	0.704	32	16	2.863×10^{-6}	3.634×10^{-7}
193–211	182	157	0.558	0.728	64	16	3.639×10^{-7}	2.791×10^{-7}
171–193-211	169	113	0.560	0.922	32	64	9.560×10^{-7}	3.356×10^{-7}

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Table F2
Continued

Channel(s)	Hidden layer size		Dropout		Batch size		Learning rate	
	μ	σ	μ	σ	μ	σ	μ	σ
171–193-HMI	138	162	0.575	0.704	64	16	1.017×10^{-5}	1.893×10^{-7}
171–211-HMI	185	184	0.770	0.702	32	16	5.863×10^{-6}	9.806×10^{-8}
193–211-HMI	182	147	0.559	0.830	32	16	3.480×10^{-7}	3.860×10^{-7}

Conflict of Interest

The authors declare no conflicts of interest relevant to this study.

Availability Statement

Our machine learning dataset (preprocessed images, physical features), trained models, predictions, and all data required for reproducibility are published at https://doi.org/10.5880/GFZ.OJSJ.2026.001 (Collin et al., 2026b). The Python code for training, prediction, and evaluation is available at https://github.com/DanielCollin96/proswin (Collin et al., 2026a). The Python code for downloading and preprocessing solar images and magnetograms is available at https://github.com/DanielCollin96/solar_image_processing (Collin & Hofmeister, 2026).

SDO solar images and magnetograms can be downloaded at http://jsoc.stanford.edu/ajax/exportdata.html. OMNIWeb solar wind measurements are available at https://omniweb.gsfc.nasa.gov/ow.html (Papitashvili & King, 2020), the OMNI_M spacecraft positions at https://omniweb.gsfc.nasa.gov/coho/. WDC-SILSO sunspot numbers from the Royal Observatory of Belgium can be accessed at https://www.sidc.be/SILSO/datafiles (Clette & Lefčvre, 2015). The Richardson & Cane ICME list is available at https://izw1.caltech.edu/ACE/ASC/DATA/level3/icmetable2.htm (Richardson & Cane, 2024).

The predictions from Bailey et al. (2021) can be accessed at https://github.com/helioforecast/Papers/tree/master/Bailey2021_AmbSoWiML/results, the ones from Brown et al. (2022) at https://github.com/eddbrown/solar-swin-transformer-output-data, and the ones from Collin et al. (2025) were published at https://doi.org/10.5880/GFZ.2.7.2024.001 (Collin et al., 2024).

Acknowledgments
We would like to thank Nathaniel Edward-Inatimi for kindly providing us with HUXt prediction data, which made the detailed model comparison possible. AIA and HMI data used here are available by courtesy of the AIA and HMI consortia. We acknowledge use of NASA/GSFC's Space Physics Data Facility's OMNIWeb and COHOWeb service. This work utilized high-performance computing resources made possible by funding from the Ministry of Science, Research and Culture of the State of Brandenburg (MWFK) and are operated by the IT Services and Operations unit of the GFZ Helmholtz Centre for Geosciences. This work is supported by the Helmholtz Association's Initiative and Networking Fund (INF) under the Helmholtz AI platform grant agreement (ID ZT-I-PF-5-1). Daniel Collin acknowledges the support of the Helmholtz International Berlin Research School in Data Science (HEIBRiDS). Yuri Shprits acknowledges support from the German Research Foundation (DFG) through the projects RADIANCE (project number 547667751) and MIPT (project number 462853228). Stefan J. Hofmeister acknowledges support from the National Science Foundation Grant AGS-2229100. Open Access funding enabled and organized by Projekt DEAL.

References

Ahn, S., Son, J., Moon, Y.-J., & Jeong, H.-J. (2025). Comparison of empirical and deep learning models for solar wind speed prediction. *The Astrophysical Journal*, *987*(2), 179. https://doi.org/10.3847/1538-4357/ade0b4

Arge, C. N., Odstrcil, D., Pizzo, V. J., & Mayer, L. R. (2003). Improved method for specifying solar wind speed near the Sun. *AIP Conference Proceedings*, *679*(1), 190–193. https://doi.org/10.1063/1.1618574

Bailey, R. L., Reiss, M. A., Arge, C. N., Möstl, C., Henney, C. J., Owens, M. J., et al. (2021). Using gradient boosting regression to improve ambient solar wind model predictions. *Space Weather*, *19*(5), e2020SW002673. https://doi.org/10.1029/2020SW002673

Barnard, L., & Owens, M. (2022). Huxt—An open source, computationally efficient reduced-physics solar wind model, written in python. *Frontiers in Physics*, *10*, 1005621. https://doi.org/10.3389/fphy.2022.1005621

Bergstra, J., Bardenet, R., Bengio, Y., & Kégl, B. (2011). Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, & K. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 24). Curran Associates, Inc.

Bergstra, J., Yamins, D., & Cox, D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th international conference on machine learning* (Vol. 28, pp. 115–123). PMLR.

Bernoux, G., Brunet, A., Buchlin, E., Janvier, M., & Sicard, A. (2022). Forecasting the geomagnetic activity several days in advance using neural networks driven by solar euv imaging. *Journal of Geophysical Research: Space Physics*, *127*(10), e2022JA030868. https://doi.org/10.1029/2022JA030868

Bianco, S., Haas, B., & Shprits, Y. Y. (2023). PINE-RT: An operational real-time plasmasphere model. *Frontiers in Astronomy and Space Sciences*, *10*, 1116396. https://doi.org/10.3389/fspas.2023.1116396

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*, 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2

Brown, E. J. E., Svoboda, F., Meredith, N. P., Lane, N., & Horne, R. B. (2022). Attention-based machine vision models and techniques for solar wind speed forecasting using solar EUV images. *Space Weather*, *20*(3), e2021SW002976. https://doi.org/10.1029/2021SW002976

Bussy-Virat, C. D., & Ridley, A. J. (2014). Predictions of the solar wind speed by the probability distribution function model. *Space Weather*, *12*(6), 337–353. https://doi.org/10.1002/2014SW001051

Bussy-Virat, C. D., & Ridley, A. J. (2016). Twenty-four hour predictions of the solar wind speed peaks by the probability distribution function model. *Space Weather*, *14*(10), 861–873. https://doi.org/10.1002/2016SW001437

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Camporeale, E. (2019). The challenge of machine learning in space weather: Nowcasting and forecasting. *Space Weather*, *17*(8), 1166–1207. https://doi.org/10.1029/2018SW002061

Camporeale, E. (2025). Paris: Pruning algorithm via the representer theorem for imbalanced scenarios. Retrieved from https://arxiv.org/abs/2512.06950

Camporeale, E., & Berger, T. E. (2025). Verification of the NOAA space weather prediction center solar flare forecast (1998–2024). *Space Weather*, *23*(10), e2025SW004546. https://doi.org/10.1029/2025SW004546

Camporeale, E., & Carč, A. (2021). Accrue: Accurate and reliable uncertainty estimate in deterministic models. *International Journal for Uncertainty Quantification*, *11*(4), 81–94. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2021034623

Camporeale, E., Chu, X., Agapitov, O. V., & Bortnik, J. (2019). On the generation of probabilistic forecasts from deterministic models. *Space Weather*, *17*(3), 455–475. https://doi.org/10.1029/2018SW002026

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*(70), 2079–2107. Retrieved from http://jmlr.org/papers/v11/cawley10a.html

Chakraborty, S., & Morley, S. K. (2020). Probabilistic prediction of geomagnetic storms and the kp index. *Journal of Space Weather and Space Climate*, *10*, 36. https://doi.org/10.1051/swsc/2020037

Chandorkar, M., & Camporeale, E. (2018). Chapter 9—Probabilistic forecasting of geomagnetic indices using Gaussian process models. In E. Camporeale, S. Wing, & J. R. Johnson (Eds.), *Machine learning techniques for space weather* (pp. 237–258). Elsevier. https://doi.org/10.1016/B978-0-12-811788-0.00009-3

Chierichini, S., Liu, J., Korsós, M. B., Del Moro, D., & Erdélyi, R. (2024). CME arrival modeling with machine learning. *The Astrophysical Journal*, *963*(2), 121. https://doi.org/10.3847/1538-4357/ad1cee

Chu, X., Ma, D., Bortnik, J., Tobiska, W. K., Cruz, A., Bouwer, S. D., et al. (2021). Relativistic electron model in the outer radiation belt using a neural network approach. *Space Weather*, *19*(12), e2021SW002808. https://doi.org/10.1029/2021SW002808

Clette, F., & Lefčvre, L. (2015). Silso sunspot number v2.0 [Dataset]. *WDC SILSO—Royal Observatory of Belgium (ROB)(07)*. https://doi.org/10.24414/qnza-ac80

Collin, D., & Hofmeister, S. J. (2026). Solar image processing [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.23158436

Collin, D., Shprits, Y., Chiarabini, L., Hofmeister, S. J., Klein, N., & Gallego, G. (2026a). Proswin: Probabilistic solar wind speed forecasting [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.19630836

Collin, D., Shprits, Y., Chiarabini, L., Hofmeister, S. J., Klein, N., & Gallego, G. (2026b). Solar images and physical features for probabilistic solar wind speed prediction [Dataset]. *GFZ Data Services*. https://doi.org/10.5880/GFZ.OJSJ.2026.001

Collin, D., Shprits, Y., Hofmeister, S. J., Bianco, S., & Gallego, G. (2024). Solar wind speed prediction from coronal holes [Dataset]. *GFZ Data Services*. https://doi.org/10.5880/GFZ.2.7.2024.001

Collin, D., Shprits, Y., Hofmeister, S. J., Bianco, S., & Gallego, G. (2025). Forecasting high-speed solar wind streams from solar images. *Space Weather*, *23*(1), e2024SW004125. https://doi.org/10.1029/2024SW004125

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). https://doi.org/10.1109/CVPR.2009.5206848

Edward-Inatimi, N. O., Owens, M. J., Barnard, L., Turner, H., Marsh, M., Gonzi, S., & Lang, M. (2026). Calibrating probabilistic solar-wind forecasts driven by the Wang-Sheeley-Arge model. *Space Weather*, *24*(3), e2025SW004614. https://doi.org/10.1029/2025SW004614

Edward-Inatimi, N. O., Owens, M. J., Barnard, L., Turner, H., Marsh, M., Gonzi, S., et al. (2024). Adapting ensemble-calibration techniques to probabilistic solar-wind forecasting. *Space Weather*, *22*(12), e2024SW004164. https://doi.org/10.1029/2024SW004164

Galvez, R., Fouhey, D. F., Jin, M., Szenicer, A., Muñoz-Jaramillo, A., Cheung, M. C. M., et al. (2019). A machine-learning data set prepared from the NASA solar dynamics observatory mission. *The Astrophysical Journal Supplement Series*, *242*(1), 7. https://doi.org/10.3847/1538-4365/ab1005

Gneiting, T., & Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, *1*(1), 125–151. https://doi.org/10.1146/annurev-statistics-062713-085831

Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Gosling, J. T. (1997). Coronal mass ejections: An overview. In *Coronal mass ejections* (pp. 9–16). American Geophysical Union (AGU). https://doi.org/10.1029/GM099p0009

Grandin, M., Aikio, A. T., & Kozlovsky, A. (2019). Properties and geoeffectiveness of solar wind high-speed streams and stream interaction regions during solar cycles 23 and 24. *Journal of Geophysical Research: Space Physics*, *124*(6), 3871–3892. https://doi.org/10.1029/2018JA026396

Gruet, M. A., Chandorkar, M., Sicard, A., & Camporeale, E. (2018). Multiple-hour-ahead forecast of the Dst index using a combination of long short-term memory neural network and Gaussian process. *Space Weather*, *16*(11), 1882–1896. https://doi.org/10.1029/2018SW001898

Gu, Y., Wei, H.-L., Boynton, R. J., Walker, S. N., & Balikhin, M. A. (2019). System identification and data-driven forecasting of AE index and prediction uncertainty analysis using a new cloud-NARX model. *Journal of Geophysical Research: Space Physics*, *124*(1), 248–263. https://doi.org/10.1029/2018JA025957

Hastie, T. J., & Tibshirani, R. J. (1990). *Generalized additive models*. Chapman & Hall/CRC.

Hofmeister, S. J. (2024). The basic iterative deconvolution: A fast instrumental point-spread function deconvolution method that corrects for light that is scattered out of the field of view of a detector. *Solar Physics*, *299*(6), 77. https://doi.org/10.1007/s11207-024-02321-y

Hofmeister, S. J., Savin, D. W., & Hahn, M. (2025). Revised point-spread functions for the atmospheric imaging assembly on board the solar dynamics observatory. *The Astrophysical Journal Supplement Series*, *278*(1), 8. https://doi.org/10.3847/1538-4365/adbaed

Hofmeister, S. J., Veronig, A., Temmer, M., Vennerstrom, S., Heber, B., & Vršnak, B. (2018). The dependence of the peak velocity of high-speed solar wind streams as measured in the ecliptic by ACE and the STEREO satellites on the area and co-latitude of their solar source coronal holes. *Journal of Geophysical Research: Space Physics*, *123*(3), 1738–1753. https://doi.org/10.1002/2017JA024586

Hu, A., Camporeale, E., & Swiger, B. (2023). Multi-hour-ahead Dst index prediction using multi-fidelity boosted neural networks. *Space Weather*, *21*(4), e2022SW003286. https://doi.org/10.1029/2022SW003286

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd international conference on international conference on machine learning* (Vol. 37, pp. 448–456). JMLR.org.

Issan, O., Riley, P., Camporeale, E., & Kramer, B. (2023). Bayesian inference and global sensitivity analysis for ambient solar wind prediction. *Space Weather*, *21*(9), e2023SW003555. https://doi.org/10.1029/2023SW003555

Jarolim, R., Veronig, A. M., Hofmeister, S., Heinemann, S. G., Temmer, M., Podladchikova, T., & Dissauer, K. (2021). Multi-channel coronal hole detection with convolutional neural networks. *A&A*, *652*, A13. https://doi.org/10.1051/0004-6361/202140640

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

King, J. H., & Papitashvili, N. E. (2005). Solar wind spatial scales in and comparisons of hourly wind and ace plasma and magnetic field data. *Journal of Geophysical Research*, *110*(A2). https://doi.org/10.1029/2004JA010649
Klein, N. (2024). Distributional regression for data analysis. *Annual Review of Statistics and Its Application*, *11*(1), 321–346. https://doi.org/10.1146/annurev-statistics-040722-053607
Klein, N., Kneib, T., Lang, S., & Sohn, A. (2015). Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Annals of Applied Statistics*, *9*(2), 1024–1052. https://doi.org/10.1214/15-AOAS823
Klein, N., Smith, M. S., & Nott, D. J. (2023). Deep distributional time series models and the probabilistic forecasting of intraday electricity prices. *Journal of Applied Econometrics*, *38*(4), 493–511. https://doi.org/10.1002/jae.2959
Krieger, A. S., Timothy, A. F., & Roelof, E. C. (1973). A coronal hole and its identification as the source of a high velocity solar wind stream. *Solar Physics*, *29*(2), 505–525. https://doi.org/10.1007/BF00150828
Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, *17*(3), 1217–1241. https://doi.org/10.1214/aos/1176347265
Lemen, J. R., Title, A. M., Akin, D. J., Boerner, P. F., Chou, C., Drake, J. F., et al. (2012). The atmospheric imaging assembly (AIA) on the solar dynamics observatory (SDO). *Solar Physics*, *275*(1), 17–40. https://doi.org/10.1007/s11207-011-9776-8
Li, Y., Yang, Y., Shen, F., Tang, B., & Lin, R. (2024). CME arrival time prediction based on coronagraph observations and machine-learning techniques. *The Astrophysical Journal*, *976*(1), 141. https://doi.org/10.3847/1538-4357/ad82e5
Licata, R. J., Mehta, P. M., Weimer, D. R., Tobiska, W. K., & Yoshii, J. (2022). MSIS-UQ: Calibrated and enhanced NRLMSIS 2.0 model with uncertainty quantification. *Space Weather*, *20*(11), e2022SW003267. https://doi.org/10.1029/2022SW003267
Liu, J., Ye, Y., Shen, C., Wang, Y., & Erdélyi, R. (2018). A new tool for CME arrival time prediction using machine learning algorithms: Cat-puma. *The Astrophysical Journal*, *855*(2), 109. https://doi.org/10.3847/1538-4357/aaae69
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., et al. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. https://doi.org/10.48550/arXiv.2103.14030
Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013). Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of ICML* (Vol. 30, p. 3).
Mayank, P., Camporeale, E., Shrivastav, A. K., Berger, T. E., & Arge, C. N. (2025). Neural enhancement of the traditional Wang–Sheeley–Arge solar wind relation. *The Astrophysical Journal Letters*, *994*(1), L5. https://doi.org/10.3847/2041-8213/ae1735
Milošić, D., Temmer, M., Heinemann, S. G., Podladchikova, T., Veronig, A., & Vršnak, B. (2023). Improvements to the empirical solar wind forecast (ESWF) model. *Solar Physics*, *298*(3), 45. https://doi.org/10.1007/s11207-022-02102-5
Möller, A., Lenkoski, A., & Thorarinsdottir, T. L. (2013). Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas. *Quarterly Journal of the Royal Meteorological Society*, *139*(673), 982–991. https://doi.org/10.1002/qj.2009
Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society*, *135*, 370–384.
Newell, P. T., Sotirelis, T., Liou, K., Meng, C.-I., & Rich, F. J. (2007). A nearly universal solar wind-magnetosphere coupling function inferred from 10 magnetospheric state variables. *Journal of Geophysical Research*, *112*(A1). https://doi.org/10.1029/2006JA012015
Odstrcil, D. (2003). Modeling 3-D solar wind structure. *Advances in Space Research*, *32*(4), 497–506. https://doi.org/10.1016/S0273-1177(03)00332-6
Owens, M. J., & Riley, P. (2017). Probabilistic solar wind forecasting using large ensembles of near-sun conditions with a simple one-dimensional "upwind" scheme. *Space Weather*, *15*(11), 1461–1474. https://doi.org/10.1002/2017SW001679
Papitashvili, N. E., & King, J. H. (2020). OMNI hourly data [Dataset]. *NASA Space Physics Data Facility*. https://doi.org/10.48322/1shr-ht18
Pinson, P., & Girard, R. (2012). Evaluating the quality of scenarios of short-term wind power generation. *Applied Energy*, *96*, 12–20. https://doi.org/10.1016/j.apenergy.2011.11.004
Pomoell, J., & Poedts, S. (2018). Euhforia: European heliospheric forecasting information asset. *Journal of Space Weather and Space Climate*, *8*, A35. https://doi.org/10.1051/swsc/2018020
Raju, H., & Das, S. (2021). CNN-based deep learning model for solar wind forecasting. *Solar Physics*, *296*(9), 134. https://doi.org/10.1007/s11207-021-01874-6
Rasmussen, C. E. (2004). Gaussian processes in machine learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced lectures on machine learning: Ml summer schools 2003, Canberra, Australia, February 2–14, 2003, Tubingen, Germany, August 4–16, 2003, revised lectures* (pp. 63–71). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-28650-9_4
Reiss, M. A., Möstl, C., Bailey, R. L., Rüdisser, H. T., Amerstorfer, U. V., Amerstorfer, T., et al. (2021). Machine learning for predicting the Bz magnetic field component from upstream in situ observations of solar coronal mass ejections. *Space Weather*, *19*(12), e2021SW002859. https://doi.org/10.1029/2021SW002859
Reiss, M. A., Temmer, M., Veronig, A. M., Nikolic, L., Vennerstrom, S., Schöngassner, F., & Hofmeister, S. J. (2016). Verification of high-speed solar wind stream forecasts using operational solar wind models. *Space Weather*, *14*(7), 495–510. https://doi.org/10.1002/2016SW001390
Richardson, I., & Cane, H. (2004). Identification of interplanetary coronal mass ejections at 1 au using multiple solar wind plasma composition anomalies. *Journal of Geophysical Research*, *109*(A9). https://doi.org/10.1029/2004JA010598
Richardson, I., & Cane, H. (2024). Near-Earth interplanetary coronal mass ejections since January 1996 [Dataset]. *Harvard Dataverse*. https://doi.org/10.7910/DVN/C2MHTH
Richardson, I., Cliver, E. W., & Cane, H. (2000). Sources of geomagnetic activity over the solar cycle: Relative importance of coronal mass ejections, high-speed streams, and slow solar wind. *Journal of Geophysical Research*, *105*(A8), 18203–18213. https://doi.org/10.1029/1999JA000400
Rigby, R. A., & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape (with discussion). *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *54*(3), 507–554. https://doi.org/10.1111/j.1467-9876.2005.00510.x
Scherrer, P. H., Schou, J., Bush, R. I., Kosovichev, A. G., Bogart, R. S., Hoeksema, J. T., et al. (2012). The helioseismic and magnetic imager (HMI) investigation for the solar dynamics observatory (SDO). *Solar Physics*, *275*(1), 207–227. https://doi.org/10.1007/s11207-011-9834-2
Shprits, Y. Y., Vasile, R., & Zhelavskaya, I. S. (2019). Nowcasting and predicting the Kp index using historical values and real-time observations. *Space Weather*, *17*(8), 1219–1229. https://doi.org/10.1029/2018SW002141
Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.
Tahtouh, M., Bernoux, G., Brunet, A., Standarovski, D., Nguyen, G., & Sicard, A. (2025). Comparison of solar imaging feature extraction methods in the context of space weather prediction with deep learning-based models. *Journal of Geophysical Research: Machine Learning and Computation*, *2*(3), e2024JH000566. https://doi.org/10.1029/2024JH000566
Tasistro-Hart, A., Grayver, A., & Kuvshinov, A. (2021). Probabilistic geomagnetic storm forecasting via deep learning. *Journal of Geophysical Research: Space Physics*, *126*(1), e2020JA028228. https://doi.org/10.1029/2020JA028228

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Tsurutani, B. T., McPherron, R. L., Gonzalez, W. D., Lu, G., Sobral, J. H. A., & Gopalswamy, N. (2006). Introduction to special section on corotating solar wind streams and recurrent geomagnetic activity. *Journal of Geophysical Research*, *111*(A7). https://doi.org/10.1029/2006JA011745

Upendran, V., Cheung, M. C. M., Hanasoge, S., & Krishnamurthi, G. (2020). Solar wind prediction using deep learning. *Space Weather*, *18*(9), e2020SW002478. https://doi.org/10.1029/2020SW002478

Woodcock, F. (1976). The evaluation of yes/no forecasts for scientific and administrative purposes. *Monthly Weather Review*, *104*(10), 1209–1214. https://doi.org/10.1175/1520-0493(1976)104<1209:teoyff>2.0.co;2

Yang, Y., Liu, J. J., Feng, X. S., Chen, P. F., & Zhang, B. (2023). Prediction of the transit time of coronal mass ejections with an ensemble machine-learning method. *The Astrophysical Journal Supplement Series*, *268*(2), 69. https://doi.org/10.3847/1538-4365/acf218

Zhelavskaya, I. S., Vasile, R., Shprits, Y. Y., Stolle, C., & Matzka, J. (2019). Systematic analysis of machine learning and feature selection techniques for prediction of the Kp index. *Space Weather*, *17*(10), 1461–1486. https://doi.org/10.1029/2019SW002271

2576604x, 2026, 8, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2026AV002524 by Karlsruher Institut Fur Technologie, Wiley Online Library on [03/11/2026]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License