Solar Flare Intensity Prediction with Machine Learning Models

Zhenbang Jiao[1], Hu Sun[1], Xiantong Wang[2], Yang Chen[1,3]

[1]Department of Statistics
[2]Climate and Space Sciences and Engineering
[3]Michigan Institute for Data Science
[1,2,3]University of Michigan, Ann Arbor, MI, USA

Key Points:

- We develop deep learning models to predict solar flare intensity values from SHARP parameters in HMI/SDO data set directly instead of flare classes.
- We use information from both flares of all classes and non-flaring time in our model.
- As opposed to solar flare classification, directly predicting solar flare intensity gives more detailed information about each occurring flare of each class.

Corresponding author: Yang Chen, ychenang@umich.edu

Abstract

[enter your Abstract here]

Plain Language Summary

[enter your Plain Language Summary here or delete this section]

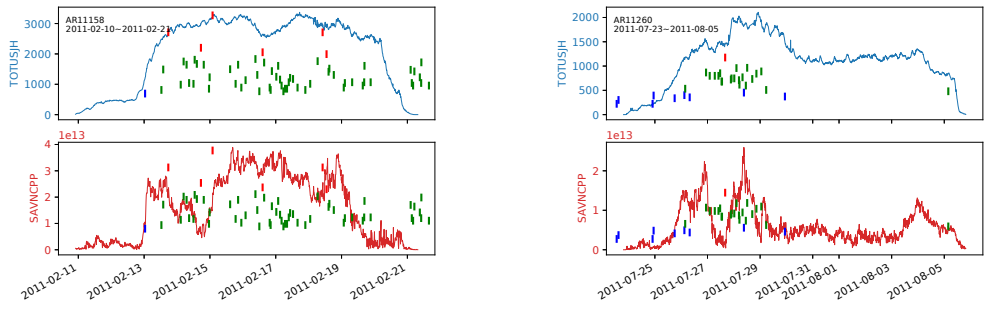
1 Introduction

Space weather involves the dynamical processes of the Sun-Earth system that may affect human life and technology. The most destructive forms of space weather, ranging from electric power disruptions to radiation hazards for astronauts, are due to energetic solar eruptions: producing magnetic storms known as coronal mass ejections (CMEs) and strong solar flares. Predictions of energetic space weather events is critical for safeguarding our technological infrastructure. Extreme space storms – those that could significantly degrade critical infrastructure – could disable large portions of the electrical power grid, resulting in cascading failures that would affect key services such as water supply, health care, and transportation. The threat-assessment report by the Lloyds insurance company (Maynard et al., 2013) concludes that extreme events could cause $2.6 trillion in damage with a recovery time of months. An earlier report by the National Research Council (Baker et al., 2009) arrived at similar conclusions.

The current space weather forecasting based on physical models is far from reliable: the forecasting window is only minutes and the accuracy is low. Previous work has established that solar eruptions are all associated with highly nonpotential magnetic fields that store the necessary free energy. The most energetic flares come from very localized intense kilogauss fields of Active Regions (ARs) (Forbes, 2000; Schrijver, 2009). Among the various data sets that become available for space scientists in recent years, there are the parameters calculated from vector magnetic fields spatially restricted to the near proximity of ARs, called Space-weather HMI Active Region Patches, or SHARPs (Bobra et al., 2014), observed with the Solar Dynamics Observatory (SDO)/Helioseismic and Magnetic Imager (HMI). The HMI/SDO data has magnetic field observations at a 12 minute cadence since its launch on February 11, 2010 till today. How to make the best use of the large amount of data available to provide reliable real-time forecasting of space weather events is one of the major questions for scientists in the field. Recently, data-driven approaches are gaining attention in the space science community with much more data becoming available. Scientists have adopted various machine learning algorithms to perform various space weather prediction tasks, including the solar flare classification using the HMI/SDO SHARP parameters and other data sets, see Leka & Barnes (2018) and Camporeale (2019) for a review and references therein.

Chen et al. (2019) shows that the SHARP parameters from HMI/SDO data could provide useful information for distinguishing strong solar flares of M/X class from weak flares of A/B class. These SHARP parameters are derived from the HMI images based on physically meaningful quantities of the active regions where the flares emerge from, see Bobra et al. (2014) for detailed descriptions of these features. To make the task of binary classification manageable, Chen et al. (2019) only considered the B and M/X class flares, ignoring the more prevalent C class flares. This design is due to the consideration that flare classes are manually discretized based on flare intensity (energy level) thus strong C class flares are essentially indistinguishable from weak M class flares.

Figure takes two active regions (ARs 11158 and 11260, see Fig.1) as an example to show the occurrence of flares of B/C/M/X classes and two of the important SHARP parameters (TOTUSJH and SAVNCPP) across an extended period of time. We can see that many incidences of C class flares accompany a strong flare (of M/X class) and that



(a) AR 11158 (b) AR 11260

Figure 1. Examples of two ARs. Blue curve shows the variation of TOTUSJH across time, red curve shows SAVNCPP. Each small vertical line stands for a recorded flare event. The height of the line is proportional to the *log* scale flare intensity. Red, green and blue represent M/X flare, C flare and B flare respectively.

the SHARP parameters evolve in continuous but locally stochastic ways during the energy buildup and release stages of strong flares. Therefore, it is important to consider the entir22e time series with flares of all classes, especially the highly prevalent C class flares, when training machine learning models as opposed to only the time point where a weak (B class) or strong (M/X class) flare happens as done in Chen et al. (2019).

Handling the sparsely and irregularly observed flare intensity levels, including big gaps between flare events, found in the GOES (Geostationary Operational Environmental Satellites) data set is a unique challenge. We note that the amount of information contained in the observed data is limited, thus the inference objective should be geared towards extracting the maximum amount of *available* information and avoiding over-interpreting the data. Therefore, instead of seeking to model the flare intensity in continuous time, we model aggregated quantities instead, e.g. the maximum flare intensity within a fixed length time window (such as ± 12 hours). In this way, we attach an intensity value to every data point that has a recorded flare in the neighboring ± 12 hours' time window. For the other time points, we define them as being "quiet" locally with an indicator function attached to it. We will explain the details of this data preparation process in Section 2.1.1. In our proposed prediction model, we are able to predict the maximum flare intensity level within a fixed length time window T hours in the future, where T can be specified to a desired value such as 12 or 24 hours, using the time series of SHARP parameters in the past. As a byproduct, we can classify the predicted events into strong or weak flares according to the flare level definitions.

2 Methodology

We provide a detailed description of the data pre-processing pipeline in Section 2.1. The Long-Short Term Memory (LSTM) regression model (Hochreiter & Schmidhuber (1997)) we use is introduced in Section 2.2, including the model structure and the novel loss function. Some sub-models based on the LSTM model are also covered in this section. Finally, Section 2.3 lists evaluation metrics we adopt corresponding to all the models mentioned in Section 2.2.

2.1 Data Pre-processing

In this article, we consider data from 860 HMI Active Region Patches (HARP) provided by the Joint Science Operations Center (JSOC) website as the source of features/predictors. For each HARP, there is a video recording a intensive flaring time interval. The video has 12 minutes interval between adjacent frames, a.k.a. 12 minutes cadence ((Bobra et al., 2014)). The SHARP parameters are calculated over each HARP frame-wisely. Of all the SHARP parameters, we make use of USFLUX, MEANGAM, MEANGBT, MEANGBZ, MEANGBH, MEANJZD, TOTUSJZ, MEANALP, MEANJZH, TOTUSJH, ABSNJZH, SAVNCPP, MEANPOT, TOTPOT, MEANSHR, SHRGT45, SIZE, SIZE_ACR, NACR and NPIX in our study (find definition of the parameters in Table 1 in Chen et al. (2019)).

As for the source of response variables, we use flares recorded in GOES data set ranging from 05/01/2010 to 06/20/2018. There are overall 12012 flares records in GOES data set.

A detailed diagram of how we get machine learning data from the raw data is shown in Fig. 3. Suppose we use m hour of data to predict real-time intensity n hours after. Firstly, we filter out all the HARPs that have more than $5(m+n)$ frames of data available. We take samples every 2 hours (10 frames). Take HARP394 as an example, there are 1334 frames in its video. Available training samples for this HARP are frame $0 \sim$ frame $(5x-1)$, frame $10 \sim$ frame $(5x+9)$, ... , frame $(1330-5y-5x) \sim$ frame $(1330-5y-1)$, 134 samples in total.

2.1.1 Response Variable

Since some of the flares recorded in the GOES data happened in HARPs that are not recorded in the JSOC data, we consider 10,349 out of the total 12,012 flares recorded in the GOES data set during the time range that we consider (see Table 1). In order to make maximum use of the data, we consider not only the class of each flare, but also the exact number of the intensity. Moreover, since the flare intensity is recorded to increase in an exponential manner, we take the $\log_{10}$ transform to keep a linear increase (see Table 2). All flare intensities mentioned later are $\log_{10}$ scale intensities if not specified.

Class/Year	2010	2011	2012	2013	2014	2015	2016	2017	2018	Total
X	0	8	5	12	15	2	0	4	0	46
M	8	84	110	90	169	128	7	37	0	633
C	64	788	906	1105	1231	1194	244	225	11	5768
B	512	519	398	418	94	428	722	606	205	3902

Table 1. The number of X/M/C/B flares recorded in each year.

Flare class	Peak flux range(wtt/m^2)		log_{10} intensity
X	$\geq 10^{-4}$		≥ -4
M	$10^{-5} \sim 10^{-4}$	$\xrightarrow{log_{10}}$	$-5 \sim -4$
C	$10^{-6} \sim 10^{-5}$		$-6 \sim -5$
B	$10^{-7} \sim 10^{-6}$		$-7 \sim -6$

Table 2. Transform from flares class to continuous intensity

Till now, we have over 10000 flare observations with continuous intensities. However, considering purely recorded flares as valid training samples will lead to the following obvious drawbacks:

1. Most of the M and X flares happened accompanied by plenty of C flares. If we simply assign response variable based on flares' peak times, two flares happened adjacent to each other with totally different intensities can have a large amount of training data overlapped. Two observations with similar training data but quite different response variables make no sense and will confuse the model.
2. Even there are over 10000 flare observations, some of them are not recorded in the videos. Besides, some of the videos are not valid since they have quite a few frames missing or they don't have enough frames in the first place. Therefore, the sample size is not enough.
3. The sample size for each class is unbalanced, we have too many B and C flares as opposed to M and X flares, which is not suitable for applying regression techniques (see the distribution of old logic in Fig.2).

In order to overcome those drawbacks, we propose a new way of defining intensities: for each frame, we define its real-time intensity as the maximum flare intensity that happened within 24 hours (12 hours before and 12 hours after). In other words, instead of focusing on each recorded flare in GOES data, we only care about the largest flare happened in each frame's 24-hour sliding window. Apparently, by applying this new mechanism, for example, the response variables of those C flares happened next to strong flares will change to high intensities, hence, balance the distribution of intensities to a large extent (see the distribution of new logic in Fig.2). Plus, the sample size is times larger since we can define a response variable for each frame.

A natural question is how should we deal with the frame where there is no intensive flare recorded on the 24-hour time window. We define one more response variable to denote the flaring or non-flaring of the time window, 1 means there is intensive flares (M/X/C/B) recorded while 0 means no. Now the response variables are more like a set of truncated data with a threshold at -7. Any flare with intensity less than -7 is not recorded. Instead, we use an extra dimension of 0/1 variable to denote this scenario.

To recap, for each sample, we assign it a 2-D response variable, the first dimension Q corresponds to sample's quietness or non-quietness (Boolean, 1 for non-quiet and 0 for quiet) while the second dimension I stands for its real-time intensity (continuous). Specifically, if a sample has $Q = 1$, then we will leave the second dimension of its response variable as N/A (see Table 3).

Label	Response Variable ($[Q, I]$)
M1.5	[1, -4.824]
X1.6	[1, -3.796]
C7.2	[1, -5.143]
Q	[0, N/A]

Table 3. Examples of the definition of response variables, Q of which stands for one quiet sample.

2.1.2 Training/testing Splitting and Normalization

Since all the recorded data ranges from 2010 to 2018, we could simply split the training and testing data by years in order to avoid information leaking. We have roughly 63% of flares happened before 2015 (6536 out of 10349), which is a reasonable ratio of the training and testing sets. Plus, each HARP only has one video, so no HARP will be divided in both the training and testing set. In this article, we split all flares happened before 01/01/2015 into the training set and the rest into the testing set.

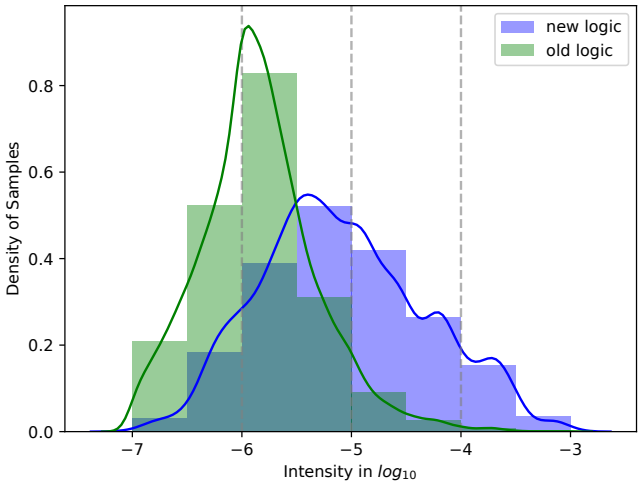


Figure 2. Distribution of non-quiet Samples intensities with new logic and old logic. Old logic is taking only flare intensities recorded on GOES data as response variables. New logic can be seen in Section 2.1.1.

After the training/testing splitting, we normalize all the data by subtracting the mean and dividing the standard deviation of the training data (Hastie et al. (2009), Section 7.10) so that there is no information of testing data being used.

Some of the HARPs have frames missing. Hence, the time interval between two adjacent frames is larger than 12 minutes. For a case like that, we set up a tolerance threshold: if there is only one frame missing, that is to say, the interval is 24 minutes, we ignore this potential problem. However, if the interval is larger than 24 minutes, i.e. there are more than 1 frame missing, we give up all training samples that should have included this frame.

2.2 Model Description

We adopt a modified LSTM regression model to portray the relationship between SHARP data and flares and a novel loss function to evaluate the difference between predicted results and real observations.

2.2.1 Model Structure

The flowchart of the model is shown in Fig.4. For each sample, the input is $5m$ frames of features (see Fig.3), a $1 \times 5m \times 20$ tensor. The output is a 2-D vector (see Table 3).

The model starts with LSTM layers. There are dropout layers (Srivastava et al. (2014)) set between adjacent LSTM layers with dropout ratio = 0.3, number of LSTM layers = 4, number of hidden cells = 30. Take model with 24 hours of training data and 40 samples per batch as an example, the input would be a $40 \times 120 \times 20$ tensor. After the LSTM layers, it turns to $40 \times 120 \times 50$. Then, the tensor will go through the truncation procedure, during which the tensor becomes $40 \times k \times 50$. We keep the last k frames of data, so as to say, the remaining model will only consider the output information of the last k frames.

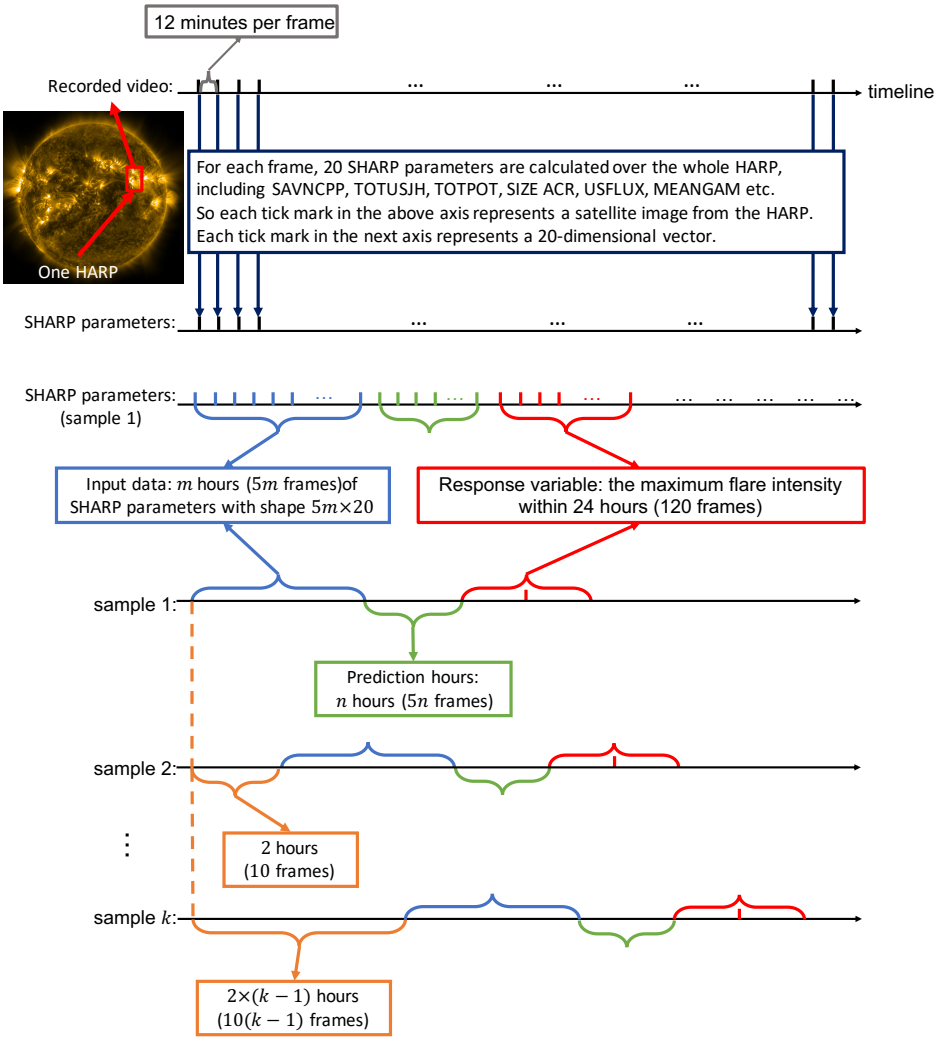


Figure 3. A diagram of how we grab samples. For each frame, 20 SHARP parameters are calculated over the whole HARP. Therefore, we have a $N \times 20$ data matrix for each video, where N is the number of frame. Data in blue braces are input data. Green braces denote the prediction intervals and the response variables are decided based on the maximum flare intensities recorded in red braces. Samples are taken every 10 frames.

Afterwards, we feed it to two separate sub-models for Q and I's training respectively, each of which contains two dense layers. The first dense layer serves the purpose of reducing the second dimension of the tensor to 1, while the second condenses the third dimension to 1. Intuitively, the first dense layer works to combine all the information in all k frames to 1 for each feature and the second combines information of all p features into 1. A Relu function is added between two dense layers to break the linearity. The only difference between these two sub-models is that we further add a Sigmoid function at the end of Q-training model in order to keep its value between $[0, 1]$. Though Q and I are going through two separate pipelines, they are not independent during the training. Later we will introduce the loss function (Section 2.2.2) that consider them jointly.

Notice that, after 4 layers of LSTM, in the particular models shown in this article, k takes the value of 1. In other words, we only take the information of the last cell.

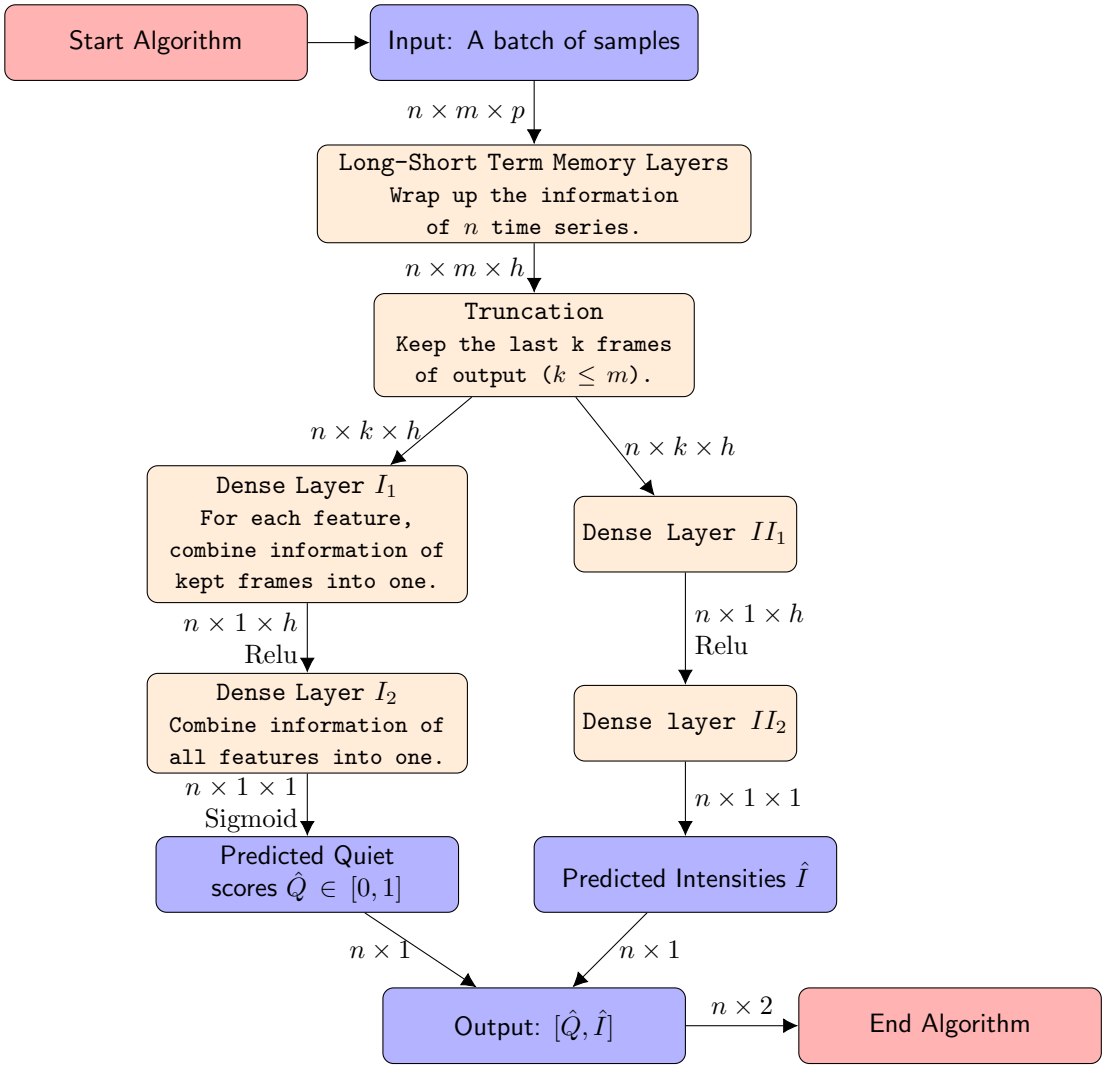


Figure 4. The flowchart of the LSTM regression model. n is the number of samples in one batch. m is the number of frames for each sample (see Fig.3 for details). p is the number of features we take into consideration. h is the hidden size of the LSTM layers. k is the number of frames we keep after going through LSTM layers. See section 2.2.1 for details.

Theoretically, we could take the last few frames into the next step. Considering that LSTM is a sequential model for time series, the output from the last cell has already contained all the information we need for prediction purposes. Plus, we've tried taking more cells output into the next layer and that doesn't offer us a better result.

A flaw of this model is that since we have over 20 times of training samples as the old one, the time cadence of the data is also decreased. Each model will take around 10 minutes to converge and 30 minutes to offer us a satisfying result (on MBP 2.3GHz, i5, 16GB).

2.2.2 Loss Function

For loss function, noted that the response variables contain both Boolean and continuous values, we need some mixture approaches to jointly evaluate the loss of two types

of variables. Another issue is that for those samples with $Q = 0$, there are no exact intensities recorded. We assign N/A to those Is. The loss function should somehow avoid the usage of I for those samples with intensities missing. The flowchart of the loss function is shown in Fig.5.

Since $Q \in [0, 1]$ and $I \in [-7, -3]$, the scale of Q's loss is incomparable to I's loss. We multiply the second dimension loss by a scale parameter r in order to balance the loss from two dimensions. As for w_1, we have more of quiet samples as opposed to non-quiet samples. However, our main focus is on those non-quiet samples when predicting local maximum flares. Still, we multiply the loss of quiet samples with weight $w_1 (< 1)$ in order to attenuate the impact caused by quiet samples when training models. (The values of both r and w_1 are tuned by cross-validation (Hastie et al. (2009)), Section 7.)

As we can see in Fig.2, C flare dominates the data set while the samples on two sides are limited. Noted that we adopt L_2 loss for I's prediction, if we simply treat all the input samples equally, the consequence is that the predicted result will tend to cluster at the central part around -6 to -4.5, which is inconsistent to our original intention that both M and X flares need to stand out of other flares as much as possible. As a result, we add $w_2(.)$ (see Eq 1) which serves to balance the ratio of each class's samples.

$$w_2(I) = |I - \mu| \times \sqrt{2\pi/\sigma^2} \tag{1}$$

where μ and σ^2 are the expectation and variance of the non-quiet training samples' flare intensities.

By conducting this strategy, we manage to combine quiet and non-quiet samples in one model and train them simultaneously.

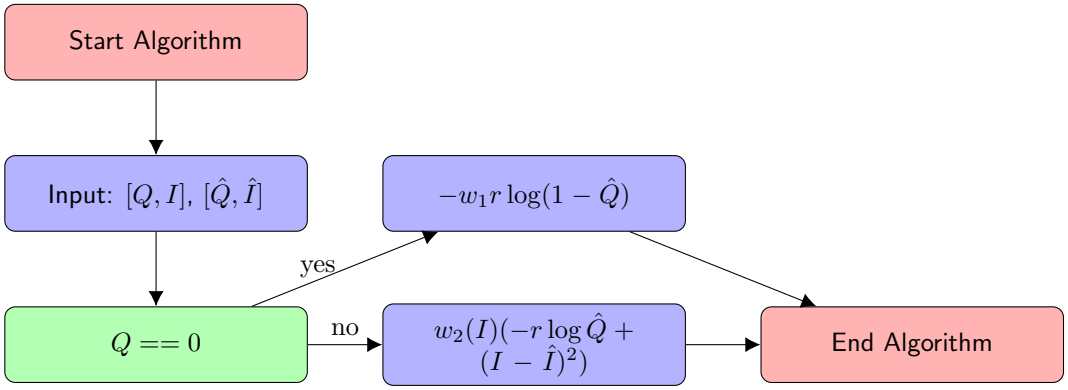


Figure 5. The loss function design. $[Q, I]$ are the real quiet score (Boolean) and intensity (continuous) and $[\hat{Q}, \hat{I}]$ are the predicted quiet score and intensity given by the model. We use binary cross-entropy loss in terms of $\hat{Q}$ and L_2 loss for $\hat{I}$. Further, we introduce three tuning parameters to calibrate the weight of each loss, w_1, w_2 and r. w_1 is the weight for loss caused by quiet samples, while $w_2(.)$ is a function set for non-quiet samples returning weights given specific Is. r is ratio set for the loss generated by Q dimension since we have an unbalanced sample size for quiet and non-quiet samples. See Section 2.2.2 for details.

2.2.3 M/X vs B Classification

In order to give a direct comparison between the old model and new model. We borrow the idea from transfer learning in (Yosinski et al., 2014). We further make use

of the output given by the trained LSTM regression model, $\hat{I}$, and I to decide an optimal threshold between M/X and B flares. Since the lower bound of M is -5 while the upper bound of B is -6, we start with those two bounds and use trisection method to finally end up with a threshold between -5 and -6 with highest training set HSS_2 (see Bobra & Couvidat (2015) for the definition of HSS_2) score. The detailed flowchart can be seen in Fig.6

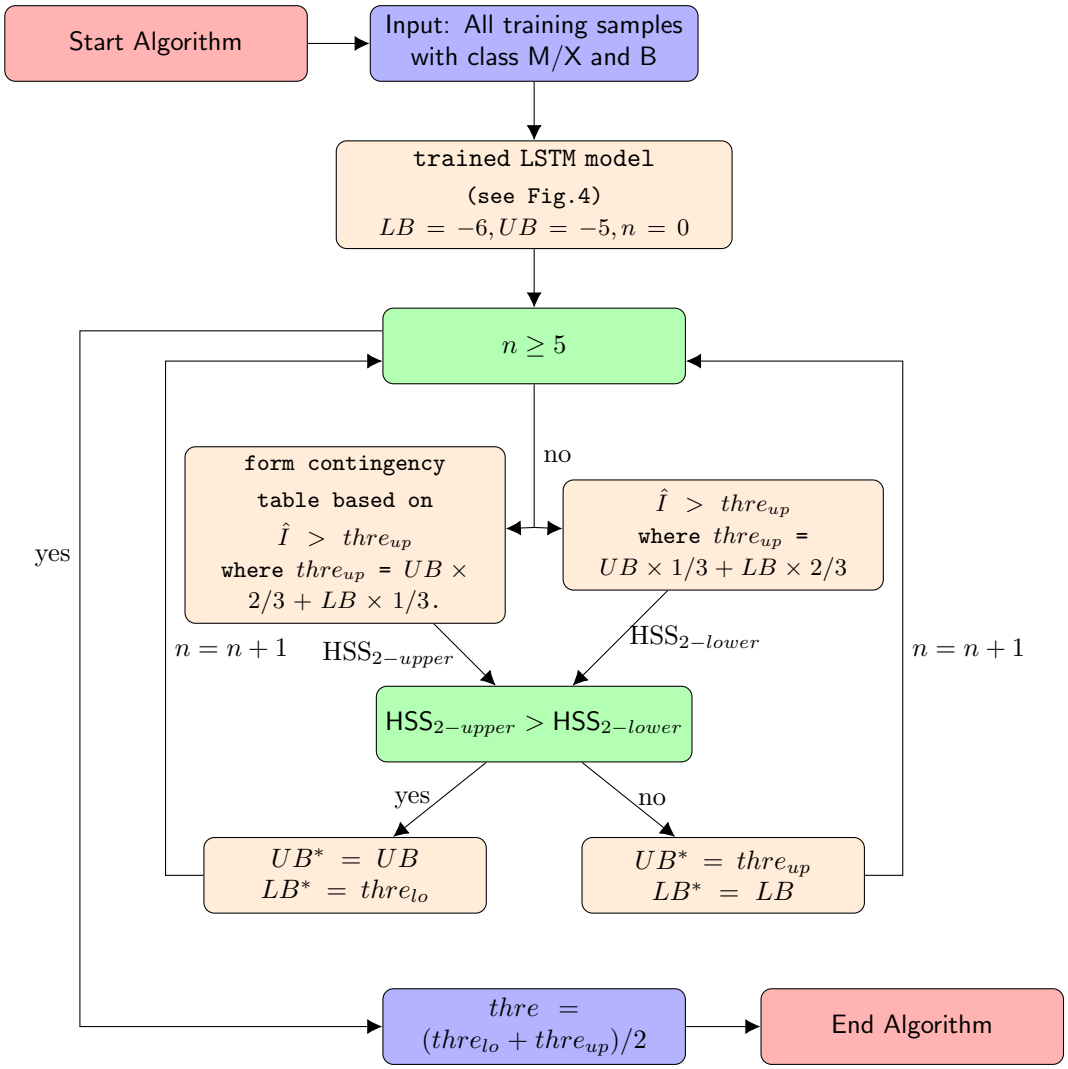


Figure 6. The flow chart of M/X vs B classification. After inputting all training samples with class M/X and B into the trained LSTM algorithm, we use the output $\hat{I}$ together with I to decide an optimal threshold between M/X and B with trisection method. The loop time is set to 5.

2.2.4 M/X vs B/Q Classification

M/X vs B/Q classification adopts the same strategy as M/X vs B classification does on determining the threshold between M/X and B/Q.

2.2.5 M/X vs Others Classification

Different from M/X vs B/Q and M/X vs B, M/X vs Others classification no longer has the sweet $[-6, -5]$ buffering area for us to train a threshold. Once we include C flares into this game, the threshold is naturally set to be -5.

To summarize, both the classifications in Section 2.2.3 and 2.2.4 use trained threshold and Section 2.2.5 directly uses -5 as the threshold between M/X flares and others.

2.3 Evaluation Metrics

For model testing, we consider twofold. On the one hand, we test the model's performance using all the available testing samples which are filtered using the same standard as training samples and calculate the statistics of interest. Results based on the this metric are in the Appendix A, Appendix C and Appendix E. On the other hand, we further filter the test samples with the following requirements:

1. There is a flare happened within this frame. Since for those frames don't, the real-time flare intensities are artificially imposed (see Section 1.1.2).
2. The flare intensity should be equal to the real-time intensity we assign. Some of the samples have actual intensity different from the imposed ones if there is a much more intensive flare happened within 24 hours of this frame.
3. There shouldn't be more than 10 frames in total missing for the input data. In other words, the tolerance level for testing samples is 10 frames (2 hours). For those have missing frames less than 10 frames, we apply hot deck imputation (Andridge & Little (2010)) to fill the missing values.
4. In order to get fair comparisons with the result in the old model. Any overlap of the input data for the same class flares will not be accepted. Specifically, for two observations with the same class of response variables, if they have any input data overlapped, we only consider one of the two. This standard is mainly set for C flares and quiet samples. Most C flares happened continuously and, hence, share a large ratio of overlapped data. Consider them as different samples will flatten the result.

Samples that meet these criteria can be considered as valid flare events and brought into testing set. The theoretical meaning of the prediction result given by the model is not what we are trying to evaluate. However, we can obtain an encouraging prediction result by doing 'goal borrowing', considering the maximum intensity flares within the 24-hour sliding window of the certain frame, the goal of the model, as the real-time intensity, our final goal, of this frame. Hence, we use these models' results as the prediction of the real-time intensity. Section 3, Appendix B, Appendix D and Appendix F are all based on this metric.

2.3.1 Regression

MSE (mean squared error) is used to evaluate the performance of the models. The result will be presented in log_{10} scale.

2.3.2 Classification

We will continue the following 7 metrics, Recall, Precision, F_1, score, HSS_1 (see Bobra & Couvidat (2015) for the definition of HSS_1), HSS_2, TSS, among which HSS_2 and TSS are our main focuses.

3 Results

In this section, we present featured results based on the models mentioned in Section 2, Section 3.1, 3.2, 3.3 and 3.4. Then, Section 3.5 gives some inferences about the truly useful information when doing solar prediction with LSTM. Case studies of intensity prediction are given in Section 3.6.

If the present is specified as time 0, we denote a model as m-n if it uses $[-n, 0]$ hours of data to predict maximum local flare intensities between $[m-12, m+12]$ time window ($m \geq 12$). To clarify, models with the same prediction hours (m), like model 12-06, 12-12, 12-24, are tested based on the same testing data with different truncated lengths for the purpose of having fair comparisons between models.

3.1 the LSTM Regression Model

Each entry in Table 4 is an average MSE of ten models with the same pair of indices (m, n). Also, we specify the MSE of each Class of flare, among which the MSE of M/X flares offers us the most insight. Intuitively, the longer the prediction hour (m) is, the larger MSE we will obtain. There is a sudden increase of M/X MSE when the prediction hour is extended from 12 to 24 hours.

Class	Num of hours before Event - Num of hours of data used							
	12-06	12-12	12-24	24-12	24-24	24-48	36-06	36-24
Average	0.25	0.25	0.24	0.25	0.27	0.28	0.29	0.30
M/X	0.44	0.46	0.48	0.61	0.63	0.69	0.72	0.71
C	0.19	0.20	0.19	0.14	0.19	0.16	0.15	0.15
B	0.25	0.23	0.22	0.29	0.25	0.27	0.26	0.28

Table 4. Mean Square Error (MSE) in log_{10} scale from the LSTM regression model using 20 SHARP parameters.

3.2 M/X vs. B Classification

Table 5 gives us a statistics summary table of M/X vs. B classification. Each entry is an average of ten models' results. Tables shown in Section 3.3 and Section 3.4 follow the same rules. As mentioned in Section 2.2.3, we still have a blank interval, [-6, -5], where there is no flares defined as M/X or B. The input data for this section of analysis are only M/X and B flares as well. This is mainly why we can get seemingly incredibly high scores.

This section is mainly set to make a direct comparison with the same classification problem in Chen et al. (2019). We increase the HSS_2 and TSS scores by roughly 0.1.

3.3 M/X vs. B/Q Classification

The only difference between Table 6 and Table 5 is that the former one includes quiet samples. There is no distinct decrease on both HSS_2 and TSS scores, suggesting that we've done perfectly on classifying quiet samples. The low HSS_1 score is due to the unbalance sample size when including quiet samples.

3.4 M/X vs. Others Classification

We try to classify M and X flares out of all other flares in this section. The result is shown in Table 7. Once we add C flares back to the game, we can hardly get HSS_2 scores greater than 0.5. We manage classify roughly half of the M and X flares out of

Metrics	Num of hours before Event - Num of hours of data used							
	12-06	12-12	12-24	24-12	24-24	24-48	36-06	36-24
Recall	0.89	0.89	0.91	0.80	0.80	0.80	0.74	0.74
Precision	0.92	0.92	0.93	0.89	0.92	0.91	0.94	0.94
F_1 Score	0.91	0.91	0.92	0.85	0.85	0.85	0.82	0.82
HSS_1	0.82	0.81	0.84	0.71	0.72	0.72	0.68	0.69
HSS_2	0.86	0.86	0.88	0.75	0.78	0.76	0.71	0.71
TSS	0.85	0.85	0.88	0.74	0.76	0.75	0.69	0.70

Table 5. Strong and Weak flare classification results from the LSTM regression model using 20 SHARP parameters.

Metrics	Num of hours before Event - Num of hours of data used							
	12-06	12-12	12-24	24-12	24-24	24-48	36-06	36-24
Recall	0.91	0.89	0.90	0.79	0.80	0.80	0.74	0.74
Precision	0.64	0.66	0.66	0.72	0.71	0.68	0.68	0.66
F_1 Score	0.75	0.75	0.76	0.75	0.75	0.73	0.70	0.69
HSS_1	0.39	0.42	0.43	0.48	0.46	0.39	0.34	0.31
HSS_2	0.73	0.74	0.74	0.73	0.72	0.70	0.67	0.66
TSS	0.88	0.86	0.87	0.76	0.77	0.76	0.70	0.70

Table 6. M/X vs. B/Q flare classification results from the LSTM regression model using 20 SHARP parameters.

other flares even when $m = 12$. Almost all of the M and X flares that are misclassified have predicted intensities falling in C flares' intensity range (See Fig.9). Still, similar to what we observe in Table 4, there is a distinct decrease once we increase the prediction hour from 12 to 24.

Metrics	Num of hours before Event - Num of hours of data used							
	12-06	12-12	12-24	24-12	24-24	24-48	36-06	36-24
Recall	0.54	0.49	0.45	0.35	0.34	0.32	0.29	0.32
Precision	0.45	0.47	0.47	0.54	0.52	0.53	0.55	0.56
F_1 Score	0.49	0.48	0.46	0.42	0.41	0.40	0.38	0.40
HSS_1	-0.11	-0.06	-0.05	0.05	0.02	0.03	0.06	0.07
HSS_2	0.47	0.45	0.44	0.39	0.38	0.37	0.35	0.37
TSS	0.51	0.46	0.43	0.33	0.32	0.30	0.28	0.30

Table 7. M/X vs. Others classification results from the LSTM regression model using 20 SHARP parameters.

3.5 Inference: Truly Useful Information

In this section, we use some visualizations of prediction results, combined with the tables mentioned above to find the truly useful information when doing solar flares prediction under the LSTM architecture. Fig.7 and Fig.8 plots the predicted intensity against the real intensity with each point representing a flare event, while Fig.9 offers us the distribution of each class's predicted intensity. Fig.7(b) is plotted based on the training samples. This is also an ideal result we are trying to get for testing, all points lying roughly on the $y = x$ line. While the rest of the five figures in Fig.7 and Fig.8 correspond to 5 models with different length of prediction hours (m) and used data (n).

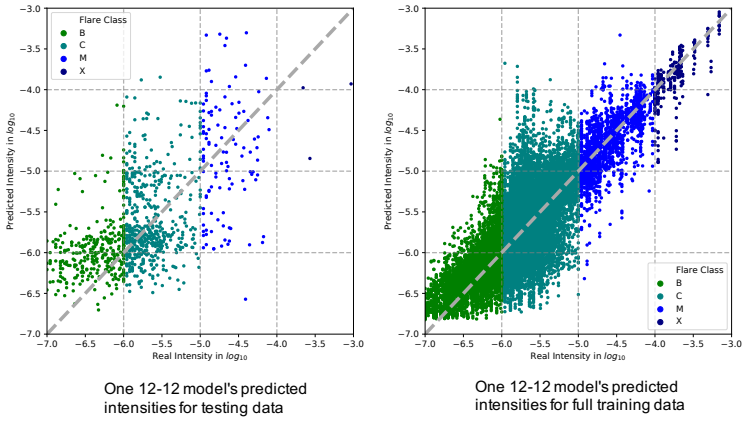


Figure 7. Predicted intensities vs. Real intensities. Each point represents a recorded flare. Its X-axis is the real intensity, Y-axis is the predicted intensity. The gray dashed line $y = x$ shows the ideal positions where every point should locate.

Longer prediction hour (larger m) leads to worse prediction result. Intuitively, predicting events one hour later is certainly easier than predicting ten hours later.

Another finding is that considering more data backwards doesn't ensure a better prediction. The explanation is twofold.

Firstly, the most useful information for predicting the behavior of a HARP limits within 12-24 hours beforehand. So once you have your $m + n \geq 24$, considering more information won't help much. Notice that, even though the TSS and HSS_2 scores decrease as the m increases, they always experience a sharper drop when the prediction hours (m) increases from 12 to 24 in all models. Essentially, since k takes the value of 1 back in the LSTM regression model, we are actually using the output information of the last frame (m hours from the prediction point) to predict the behavior of the prediction point. Worse result indicates that the last frame is less relevant to the prediction point or it is harder for LSTM to build a relationship between these two time points. Thus, the sharp drop from 12 to 24 indicates the solar activities happened within 12-24 hours before the events have a significant influence on the behavior in the prediction point. In other words, the randomness existed within 24 hours before the event happening is very closely related to the behavior of the prediction point but independent to the activities 24 hours before. Hence, the information within 24 hours before the event is incomparably useful and irreplaceable.

Secondly, even though the truly useful information is within 24 hours before the events, considering more information offering us worse result in return is quite counter-intuitive. The reason is due to the limitation of the LSTM model. LSTM is an artificial recurrent neural network (RNN) architecture used for digging out the temporal properties within time series data. The parameter matrices for each gate remain unchanged for all input time series. In other words, LSTM will consider the process indifferently. If the whole time series before the event is not acting homogeneously, adding information 24 hours before can, on the contrary, impair the performance of the prediction. One possible solution is raised in Section 4.

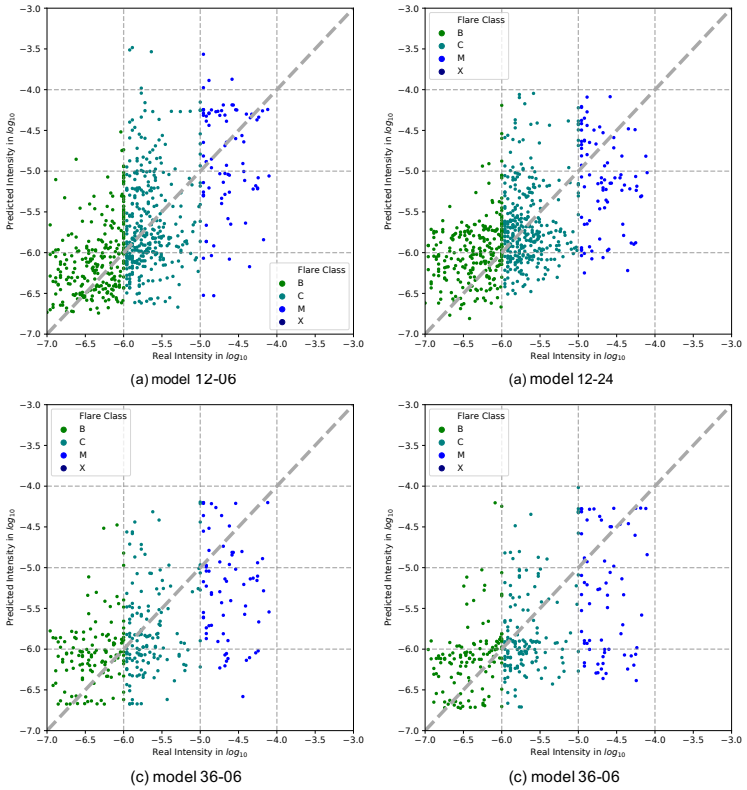
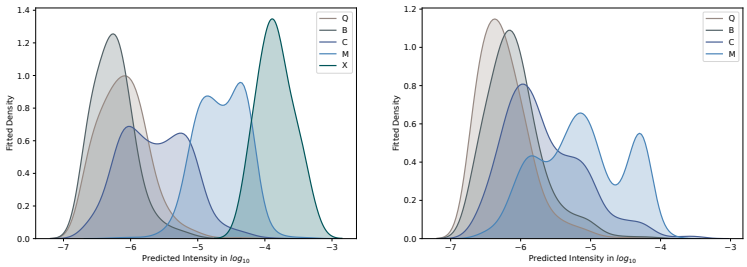


Figure 8. Visualizations for 4 example models. The figures share the same setting as Fig 7. Noted that, in both Fig 8 and 9, there are no X flare plotted. Generally, there is no applicable X flares in testing set for long prediction hours ($m \geq 12$). We have very few X flares. Most of them happened before 2015. For the limited X flares happened after 2015, they either have many frames missing before it happened, or happened only few hours after the video starting. So, for models with long prediction hours we don't have X flares in testing set.

3.6 Case Study

In the case study section, we focus on the prediction of M and X flares mainly for two reasons. Firstly, M and X flares are of primary concern in the flare prediction problem. Secondly, see Table 4, the model can already offer us a decent prediction, i.e. a relatively small MSE, for B and C flares. Also, see Fig.9, not matter for training or testing set, quiet samples' predicted intensities can be restricted below -5. Hence, M and X flares are not only the most important but also the worst predicted flares generating highest MSE.

Fig.10 and Fig.11 show 6 prediction plots, including 4 well and 2 bad-performed examples, each of which corresponds to one HARP. For a successful case, the blue curve in the lower panel of each plot should be as close as possible to the local maximum flare, i.e. local highest round point. Noted that the existence of dimension Q in the response variable is only to compensate for the non-observable flares. Thus, the quiet score $\hat{Q}$ in the upper plot is more than a signal instead of an exact predicted result. As long as the lower plot offers a $\hat{I} \leq -6$, we can still consider the model of having a good prediction result of the quiet time.



(a) The distribution of training data (b) The distribution of testing data

Figure 9. Fitted distribution of predicted intensities based on one 12-06 model. The distribution is fitted using Gaussian kernel with bandwidth = 0.15. X-axis is the values taken by predicted intensities, Y-axis stands for the density of fitted distribution. Ideally, flares with class B, C or M should follow an asymptotically normal distribution. The predicted distribution (a) for training data is close to the ideal setting. While for testing set, the predicted intensities are still having a hard time separate themselves with other flares.

3.6.1 HARP Choosing Metrics

Specifically, 4 example plots in Fig.10 are chosen where at least one of their M and X flares lays near the $y = x$ diagonal line in Fig.8(a). For the 2 cases in Fig.11, we choose two videos where one of their M or X flare has the largest prediction error ($|I-\hat{I}|$) among all M and X flares in the training set and testing set respectively in 12-06 model.

3.6.2 Inference

The two typical cases shown in Fig.11 represent two main situations where M and X are wrongly predicted. Firstly, the model does perceive the increase but not precisely, like Fig.11(a). Models may have increased hours before or after the intensive flares' happening. Secondly, the model doesn't detect the intensive flares totally, like Fig.11(b). However, this scenario only happens when the certain M/X flares lay at the head or tail of the video. Moreover, videos also tend to have a few frames missing at the beginning and the end. This truth inevitably makes us think that it is the potential problem of the data rather than the model that restricts the performance of the prediction.

4 Discussion

We've presented the data pre-processing pipeline to prepare data from SHARP parameters. Then, the LSTM regression-based models with encouraging results on solar flare intensity prediction are presented. The work in this article can be considered as the second step towards early predictions of the intensive solar flare events.

Compared to our previous results in Chen et al. (2019), our current series of models stand out in several aspects.

- The prediction score, TSS and HSS_2 of M/X vs. B is increased by 0.1 when prediction hour is less or equal to 12.
- We predict the exact intensity rather than the class of the flares. We further stratify the strong class to M and X, weak class to C, B and Q.
- We consider more cases, including 06-12, 12-06, 12-12, 12-24, 18-06, 18-12, 18-24, 24-06, 24-12, 24-24, 24-48, 36-06, 36-12, 36-24, 36-48, and drop the cases with trivial number of prediction hours, like 72-12 and 48-12.

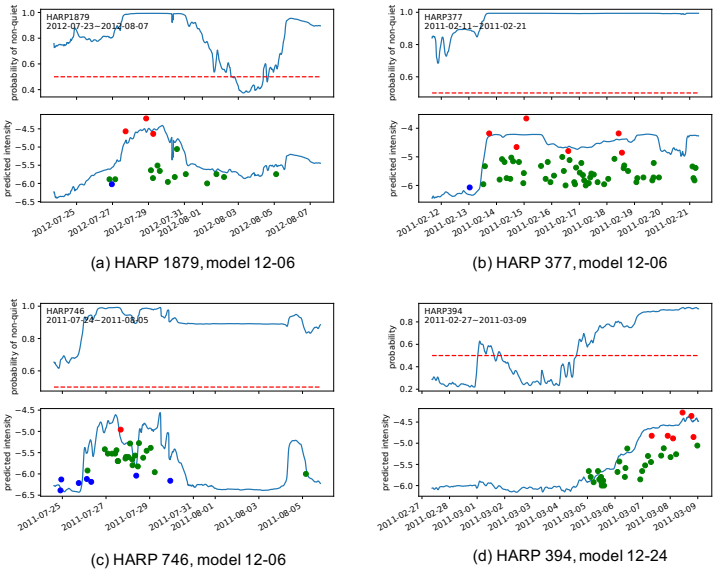


Figure 10. Case Studies: Successful cases. For each plot, the blue curve on the upper panel is the $\hat{Q}$ score. The red dashed line taking the value of 0.5 is the threshold of dividing quiet and non-quiet times. The blue curve on the lower panel is the predicted real-time flare intensity, $\hat{I}$. There is no time shift on each plot. Each red, green or blue round point corresponds to one recorded flare of M/X, C or B class respectively. Unlike Fig. 1, the height of each point is exactly the *log* intensity of the flare it representing.

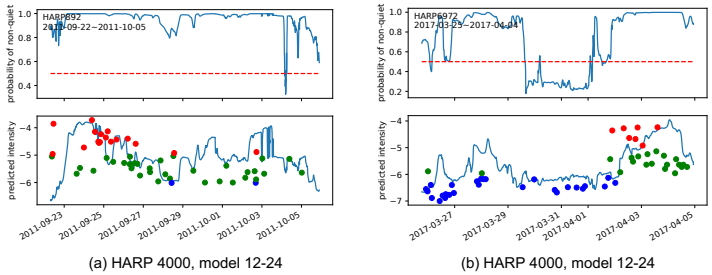


Figure 11. Case Studies: Failed cases. Same setting as Fig. 10.

- We use the same number of flare observations for models with the same prediction hours when training and testing so that we could directly compare the models' results with the same prediction hours but distinct lengths of data used.

However there still exist several potential mishaps that need to be solved.

- The current training testing sets are separated between year 2015. But flares happened after 2015 are not exactly equivalent to flares before 2015 since the sun is experiencing a cycle with up and downs every ten of years.
- Till now, we still consider videos of different HARPs as equivalent, which is certainly not the case. Moreover, there is latent dependence between flares. We only train the SHARP features ignoring the sequence of flares beforehand.

- As mentioned in 3.6.2, we believe the 20 calculated features are not cut finely due to the sphere property of the sun. We are also now doing research on how to find the polarity inversion line of each frame and, hence, calculate the SHARP features along it.

In the future analysis, we use a Bayesian model and Hawkes process to consider the dependence between flares and treat each HARP as an independent sub-model affiliated to a common prior model.

Appendix A Tables of M/X vs. B Confusion Matrices (full)

Model	Contingency Table (mean [min, max])			
	TP	FN	FP	TN
12-06	410.4 [376, 430]	132.6 [113,167]	271.7 [192,366]	3129.3 [3035,3209]
12-12	411.3 [359, 461]	131.7 [82,184]	267.6 [213,309]	3133.4 [3092,3188]
12-24	406.0 [388, 431]	137.0 [112, 155]	255.2 [179,322]	3145.8 [3079,3222]
18-06	401.0 [343, 458]	119.0 [62, 177]	384.7 [169,648]	2735.3 [2472,2951]
18-12	395.4 [367, 424]	124.6 [96, 153]	327.8 [232,445]	2792.2 [2675,2888]
18-24	418.6 [368, 483]	101.4 [37, 152]	375.4 [204,769]	2744.6 [2351,2916]
24-06	263.8 [243,289]	119.2 [94,140]	120.7 [71,171]	1878.3 [1828,1928]
24-12	267.1 [215,295]	115.9 [88,168]	122.2 [73,160]	1876.8 [1839,1926]
24-24	254.9 [220,291]	128.1 [92,163]	115.7 [78,159]	1883.3 [1840,1921]
24-48	254.3 [190,303]	128.7 [80,193]	135.9 [92,232]	1863.1 [1767,1907]
36-06	213.1 [154,253]	86.9 [47,146]	166.9 [74,279]	1457.1 [1345,1550]
36-12	224.6 [182,266]	75.4 [34,118]	218.7 [118,371]	1405.3 [1253,1506]
36-24	211.5 [185,240]	88.5 [60,115]	166.0 [104,334]	1458.0 [1290,1520]
36-48	227.2 [191,251]	72.8 [49,109]	203.5 [114,426]	1420.5 [1198,1510]

Appendix B Tables of M/X vs. B Confusion Matrices

Model	Contingency Table (mean [min, max])			
	TP	FN	FP	TN
12-06	86.2 [83,88]	8.8 [7,12]	7.3 [1,14]	176.7 [170,183]
12-12	84.2 [80,88]	10.8 [7,15]	6.8 [3,10]	177.2 [174,181]
12-24	85.4 [79,88]	9.6 [7,16]	6.4 [4,8]	177.6 [176,180]
18-06	79.5 [74,86]	10.5 [4,16]	7.9 [3,19]	156.1 [145,161]
18-12	79.2 [76,84]	10.8 [6,14]	5.4 [1,12]	158.6 [152,163]
18-24	81.1 [75,88]	8.9 [2,15]	7.9 [1,35]	156.1 [129,163]
24-06	71.7 [66,78]	17.3 [11,23]	4.3 [2,7]	158.7 [156,161]
24-12	70.3 [63,76]	18.7 [13,26]	5.2 [1,9]	157.8 [154,162]
24-24	71.0 [66,76]	18.0 [12,23]	6.8 [3,12]	156.2 [151,160]
24-48	64.4 [60,71]	16.6 [10,21]	6.4 [3,12]	113.6 [108,117]
36-06	57.5 [49,63]	20.5 [15,29]	4.1 [2,9]	89.9 [85,92]
36-12	59.9 [53,67]	18.1 [11,25]	6.8 [2,17]	87.2 [77,92]
36-24	57.6 [53,63]	20.4 [15,25]	4.1 [2,15]	89.9 [79,92]
36-48	59.4 [49,65]	18.6 [13,29]	6.1 [2,14]	87.9 [80,92]

Appendix C Tables of M/X vs. B/Q Confusion Matrices (full)

Model	Contingency Table (mean [min, max])			
	TP	FN	FP	TN
12-06	410.4 [376, 430]	132.6 [113,167]	594.6 [408,845]	14553.4 [14303,14740]
12-12	411.3 [359, 461]	131.7 [82,184]	598.0 [464,697]	14550.0 [14451,14684]
12-24	406.0 [388, 431]	137.0 [112, 155]	562.5 [393,703]	14585.5 [14445,14755]
18-06	401.0 [343, 458]	119.0 [62, 177]	941.9 [400,1680]	13026.1 [12288,13568]
18-12	395.4 [367, 424]	124.6 [96, 153]	779.7 [513,1128]	13188.3 [12840,13455]
18-24	418.6 [368, 483]	101.4 [37, 152]	944.3 [466,2128]	13023.7 [11840,13502]
24-06	263.8 [243,289]	119.2 [94,140]	327.0 [226,443]	8772.0 [8656,8873]
24-12	267.1 [215,295]	115.9 [88,168]	338.0 [216,448]	8761.0 [8651,8883]
24-24	254.9 [220,291]	128.1 [92,163]	321.2 [189,486]	8777.8 [8613,8910]
24-48	254.3 [190,303]	128.7 [80,193]	398.0 [219,790]	8701.0 [8309,8880]
36-06	213.1 [154,253]	86.9 [47,146]	487.2 [152,865]	7066.8 [6689,7402]
36-12	224.6 [182,266]	75.4 [34,118]	642.4 [333,1165]	6911.6 [6389,7221]
36-24	211.5 [185,240]	88.5 [60,115]	516.3 [266,1194]	7037.7 [6360,7288]
36-48	227.2 [191,251]	72.8 [49,109]	624.3 [315,1283]	6929.7 [6271,7239]

Appendix D Tables of M/X vs. B/Q Confusion Matrices

Model	Contingency Table (mean [min, max])			
	TP	FN	FP	TN
12-06	86.2 [83,88]	8.8 [7,12]	49.0 [29,73]	1606.0 [1582,1626]
12-12	84.2 [80,88]	10.8 [7,15]	44.3 [33,55]	1610.7 [1600,1622]
12-24	85.4 [79,88]	9.6 [7,16]	44.6 [35,57]	1610.4 [1598,1620]
18-06	79.5 [74,86]	10.5 [4,16]	63.4 [23,113]	1571.6 [1522,1612]
18-12	79.2 [76,84]	10.8 [6,14]	51.3 [27,78]	1583.7 [1557,1608]
18-24	81.1 [75,88]	8.9 [2,15]	59.0 [21,167]	1576.0 [1468,1614]
24-06	71.7 [66,78]	17.3 [11,23]	25.6 [18,33]	915.4 [908,923]
24-12	70.3 [63,76]	18.7 [13,26]	27.8 [14,40]	913.2 [901,927]
24-24	71.0 [66,76]	18.0 [12,23]	29.8 [20,40]	911.2 [901,921]
24-48	64.4 [60,71]	16.6 [10,21]	32.7 [17,57]	865.3 [841,881]
36-06	57.5 [49,63]	20.5 [15,29]	31.1 [8,80]	840.9 [792,864]
36-12	59.9 [53,67]	18.1 [11,25]	43.0 [19,99]	829.0 [773,853]
36-24	57.6 [53,63]	20.4 [15,25]	33.8 [13,100]	838.2 [772,859]
36-48	59.4 [49,65]	18.6 [13,29]	39.8 [21,78]	832.2 [794,851]

Appendix E Tables of M/X vs. Others Confusion Matrices (full)

Model	Contingency Table (mean [min, max])			
	TP	FN	FP	TN
12-06	191.6 [158,228]	351.4 [315,385]	487.9 [360,611]	17812.1 [17689,17940]
12-12	193.2 [150,238]	349.8 [305,393]	465.8 [370,666]	17834.2 [17634,17930]
12-24	172.8 [137,218]	370.2 [325,406]	424.3 [337,571]	17875.7 [17729,17963]
18-06	149.4 [102,248]	370.6 [272,418]	422.1 [205,612]	16440.9 [16251,16658]
18-12	152.2 [110,201]	367.8 [319,410]	415.4 [256,489]	16447.6 [16374,16607]
18-24	144.8 [109,200]	375.2 [320,411]	392.8 [277,572]	16470.2 [16291,16586]
24-06	99.8 [65,157]	283.2 [226,318]	275.8 [197,368]	10587.2 [10495,10666]
24-12	85.3 [62,104]	297.7 [279,321]	240.7 [153,334]	10622.3 [10529,10710]
24-24	87.3 [60,114]	295.7 [269,323]	251.1 [172,328]	10611.9 [10535,10691]
24-48	88.3 [52,128]	294.7 [255,331]	207.1 [115,298]	10655.9 [10565,10748]
36-06	73.2 [32,106]	226.8 [194,268]	206.2 [126,279]	8745.8 [8673,8826]
36-12	71.3 [39,97]	228.7 [203,261]	239.4 [139,329]	8712.6 [8623,8813]
36-24	66.4 [42,89]	233.6 [211,258]	249.8 [103,394]	8702.2 [8558,8849]
36-48	68.1 [25,100]	231.9 [200,275]	198.9 [112,350]	8753.1 [8602,8840]

Appendix F Tables of M/X vs. Others Confusion Matrices

Model	Contingency Table (mean [min, max])			
	TP	FN	FP	TN
12-06	49.8 [40,57]	45.2 [38,55]	54.7 [37,67]	1998.3 [1986,2016]
12-12	47.1 [38,58]	47.9 [37,57]	53.5 [42,79]	1999.5 [1974,2011]
12-24	41.6 [32,54]	53.4 [41,63]	44.7 [31,64]	2008.3 [1989,2022]
18-06	34.6 [24,51]	55.4 [39,66]	35.5 [24,54]	1856.3 [1838,1868]
18-12	37.3 [29,43]	52.7 [47,61]	31.7 [18,42]	1860.3 [1850,1874]
18-24	35.0 [26,46]	55.0 [44,64]	29.2 [16,41]	1862.8 [1851,1876]
24-06	32.2 [27,40]	48.8 [41,54]	30.7 [20,38]	1137.3 [1130,1148]
24-12	29.4 [24,35]	51.6 [46,57]	26.1 [17,33]	1141.9 [1135,1151]
24-24	28.8 [19,39]	52.2 [42,62]	27.7 [20,33]	1140.3 [1135,1148]
24-48	28.0 [22,38]	53 [43,59]	25.1 [12,32]	1142.9 [1136,1156]
36-06	23.6 [12,33]	54.4 [45,66]	17.0 [10,22]	1025.0 [1020,1032]
36-12	26.9 [13,36]	51.1 [42,65]	19.9 [7,33]	1022.1 [1009,1035]
36-24	25.2 [19,29]	52.8 [49,59]	21.5 [14,40]	1020.5 [1002,1028]
36-48	25.1 [9,35]	52.9 [43,69]	15.9 [9,28]	1026.1 [1014,1033]

Acknowledgments

Enter acknowledgments, including your data availability statement, here.

References

Andridge, R. R., & Little, R. J. A. (2010, Apr). *A review of hot deck imputation for survey non-response.* U.S. National Library of Medicine. Retrieved from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130338/`

Baker, D. N., Balstad, R., Bodeau, J. M., Cameron, E., Fennell, J. F., Fisher, G. M., ... Strachan, Jr., L. (2009). *Severe space weather events–understanding societal*

and economic impacts workshop report. Washington, D.C.: National Academies Press. doi: 10.17226/12643

Bobra, M. G., & Couvidat, S. (2015, Jan). Solar flare prediction usingsdo/hmi vector magnetic field data with a machine-learning algorithm. *The Astrophysical Journal*, *798*(2), 135. Retrieved from `http://dx.doi.org/10.1088/0004-637X/798/2/135` doi: 10.1088/0004-637x/798/2/135

Bobra, M. G., Sun, X., Hoeksema, J. T., Turmon, M., Liu, Y., Hayashi, K., ... Leka, K. D. (2014, apr 01). The Helioseismic and Magnetic Imager (HMI) vector magnetic field pipeline: SHARPs – Space-Weather HMI Active Region Patches. *Solar Phys*, *289*(9), 3549–3578. doi: 10.1007/s11207-014-0529-3

Camporeale, E. (2019, jul). The challenge of machine learning in space weather nowcasting and forecasting. *Space Weather*, *17*. doi: 10.1029/2018sw002061

Chen, Y., Manchester, W. B., Hero, A. O., Toth, G., DuFumier, B., Zhou, T., ... Gombosi, T. I. (2019). Identifying solar flare precursors using time series of sdo/hmi images and sharp parameters. *arXiv preprint arXiv:1904.00125*.

Forbes, T. G. (2000, October). A review on the genesis of coronal mass ejections. *J. Geophys. Res.*, *105*(A10), 23,153–23,165.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction.* Springer. Retrieved from `https://books.google.com/books?id=eBSgoAEACAAJ`

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735-1780.

Leka, K., & Barnes, G. (2018). Solar flare forecasting: Present methods and challenges. In N. Buzulukova (Ed.), *Extreme events in geospace* (pp. 65 – 98). Amsterdam, The Netherlands: Elsevier. doi: 10.1016/B978-0-12-812700-1.00003-0

Maynard, T., Smith, N., & Gonzales, S. (2013, May). *Solar storm risk to the North American electric grid,* (resreport). Lloyd's Insurance Company. Retrieved from `https://www.lloyds.com/news-and-insight/risk-insight/library/natural-environment/solar-storm`

Schrijver, C. J. (2009). Driving major solar flares and eruptions: A review. *Adv. Space Res.*, *43*(5), 739 – 755. doi: 10.1016/j.asr.2008.11.004

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*, 1929-1958. Retrieved from `http://jmlr.org/papers/v15/srivastava14a.html`

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems 27* (pp. 3320–3328). Curran Associates, Inc. Retrieved from `http://papers.nips.cc/paper/5347-how-transferable-are-features-in-deep-neural-networks.pdf`