**PHYS 677**

**Final Report**

**Forecast Substorms Using Machine Learning Approaches**

**Yiwen Zhu**

**2021 May 12th**

**1. Introduction**

**1. Background**

**A. Story about Substorm and Space Weather**

On April 5th, 2020, a communication satellite called Galaxy 15 encountered a severe anomaly and ceased responding to ground commands. It caused loss of communication over a vast range of area, more troubling is that during its anomaly, the operation team on the ground lost control of the satellite, so in hours, the satellite was drifting around in the space, posing great danger to other satellites and man-made spacecraft in the area [Loto'aniu et al., 2015]. Some news reports called it a ‘zombie satellite’. Fortunately, the team gained control of Galaxy 15 again, but the event keeps worrying people about how dangerous it can be when this kind of event happened in space. The investigation found that the anomaly of Galaxy 15 is highly possibly the result of a space phenomenon called the ‘Substorm’ and the Galaxy 15 was at a ready position to receive the full impact of this substorm which led to its dysfunctionality. This Galaxy 15 is a typical example of how the phenomenon in space can affect our space operations, and the substorm is very common and crucial among these phenomena, that is why we choose substorms as the concerned topic for this study.

The study of these phenomena is called the ‘Space Weather’, it is analogous to the weather in the atmosphere. When severe weather events such as a hurricane or tornado happened, the result is catastrophic, and the same goes for the events in space. When the ‘weather’ in space is bad, it can be extremely harmful to human activities in space, as in the case of the Galaxy 15, where a ‘hurricane’ -- substorm stopped a satellite from normal functioning and threatened other satellites in the area. Not only can the space weather harms human activities in space but it can also threaten the activities on Earth. In 1859, a super violent geomagnetic storm called the Carrington Event happened, producing an auroral that is visible globally. It acted like a very strong electromagnetic pulse (EMP) and had caused a long-time disconnection of telegram between Europe and North America.

In 2020 alone, NASA’s fund for the space physics study is roughly 700 million dollars. The study of space weather is becoming more and more important for people’s everyday life and space research. Substorm plays a very essential part in it, because it is very common, and modern space physics research believes that substorm plays the main role of transporting energy and particles from space to the near-earth region. These facts motivate us to see the possibility of forecasting the substorm.

**B. Literation Review and the Need for Machine Learning**

Just like meteorologists seek to forecast severe weather events, the ultimate question in space weather study is the forecast of destructive events, in this study we are particularly interested in forecasting substorms. There have been some efforts about forecast substorms using statistical approaches, these efforts are well represented by the work done by Lyons, et al, [Lyons et al., 1997], where they tried to find the correlations between the happening of substorms with other observational signatures, such as the turning of the interplanetary magnetic field. But these efforts were not satisfying, in Lyons’s work, the best prediction rate is about 21%. It shows us that by simply finding the correlation between the observations and the onset of the substorms may not work very well, possibly because the simple linear trend in these observations can not reveal the physics behind the generation of substorms.

The recent process of machine learning, however, gives us hope. The idea is, for a physics system such as substorms that we can not give a full explanation of the mechanism, machine learning can perform the deep learning about the complex correlations between the observed data and the happening of substorms. Rather than simply checking the turning of the interplanetary magnetic field to identify a substorm as in Lyons’ work, we use machine learning to learn the tread and vary of various observational quantities. On the one hand, if the onset of substorm indeed is correlated or caused by these quantities, the proper machine learning model should be able to produce a good prediction of the onset of substorms, on the other hand, these predictions, in turn, can give us information and intuitive about what quantities are the drive of the substorms.

**C. Detailed Introduction to Substorms**

Substorm is defined to be the disturbance in Earth’s magnetosphere which is the Earth’s magnetic dipole field region, as illustrated in Figure 1. Substorms are thought to be mainly driven by the Sun’s activities such as Coronal Mass Ejection (CME). These activities change the status of the solar wind -- the charged energetic particles from the Sun’s atmosphere. The change of solar wind interacts with the Earth’s magnetosphere and eventually caused substorms.

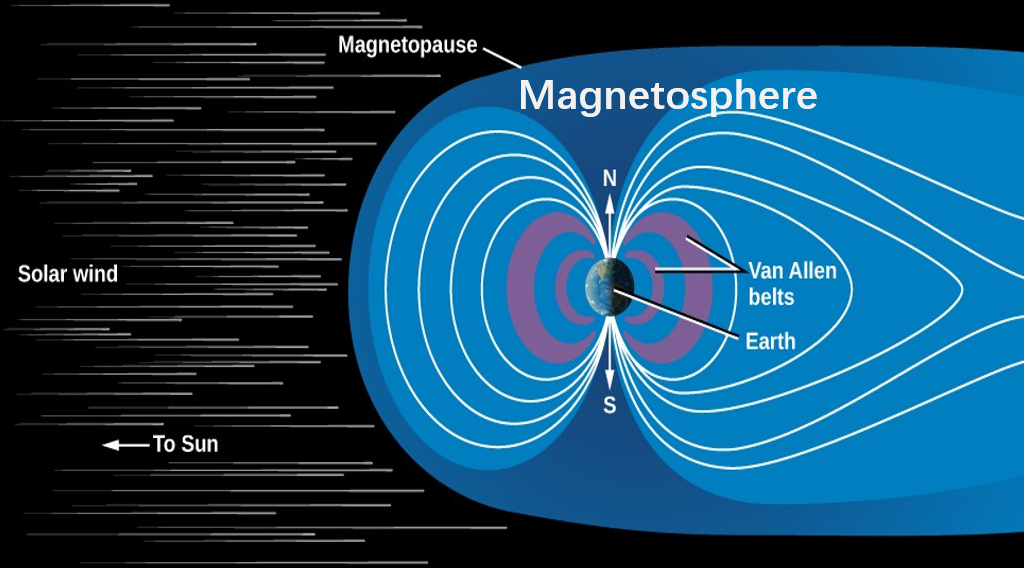


Figure1. Illustration of Earth’s magnetosphere, the blue region is the magnetosphere, left to it is Sun. The solar wind coming from the Sun's atmosphere impinges on the Earth’s magnetosphere and drives the substorms in it.

Observationally, the onset substorms are seen as the brightening of the auroral and electrojet. Substorms can lead to many consequences both in the magnetosphere and in the atmosphere. During a severe substorm, the magnetoelectric condition in space is very chaotic. A substorm can change the magnetic and electric field in space, disturb the current in the ring current and Von Allen Radiation Belt region, and even affect the geomagnetic conditions on the ground, cutting off the electric power and telegram communications. Recent studies have shown that substorms play important roles in transporting the energy and particles from the tail of the magnetosphere thus are critical for explaining the dynamics in the magnetosphere. Substorm is a very frequent event in the magnetosphere, happens several times per day on average.

Since we aim to forecast the substorm in this study, further introduction about the identification of the onset of a substorm and potential drivers for it is needed.

Quantitatively, the substorm is identified by the Auroral Electrojet or AE index. AE index is derived by the geomagnetic variations in the horizontal magnetic field component observed at the ground observatories. The onset of substorms is identified by the sudden jump of the AE index. In practice, the lower value of the AE index, or AL index, is more often used.

Since the substorm is highly correlated to the Sun’s activities, the solar wind conditions are the natural candidates for drivers of the substorms, we will primarily use solar wind quantities such as the interplanetary magnetic field and solar wind velocity for forecasting substorms.

**2. Objectives and Significance**

The main objective of this study is to build a reliable and robust model for forecasting substorms, for such a forecasting model to be practically valuable, the anticipated prediction rate should be over 70%, and it should be applicable for substorms in all conditions over a long period. After building up such a model, the physics behind it is also highly interesting, we will analyze what factors are most important for the forecast, which gives us insight into the main driver for the onset of substorms. Also, at Rice, we have a numerical model called the Rice Convective Model (RCM) which does the simulation in the magnetosphere so that we can make a comparison between our results from the machine learning with the numerical model.

The significance of the study comes into two parts, firstly the unprecedented accurate forecasting system for substorms could be a big triumph as a practically usable model, for the first time we can make a forecast for such an important phenomenon in space. Secondly, it shows the great potential for introducing machine learning techniques into the space weather study, giving a new way of building the space weather model, which is very different and innovative compared to traditional space weather research.

**3. Pipeline for the Study**

The pipeline consists of six steps, shown in Figure 2. The first step is to collect the database, then identify the substorms and the non-substorm cases, next we do the data processing and visualization, finally, we build and compile the model, the test data will give us the prediction accuracy. As introduced, we will also determine which factors are important for forecasting and compare our result with the numerical model RCM.

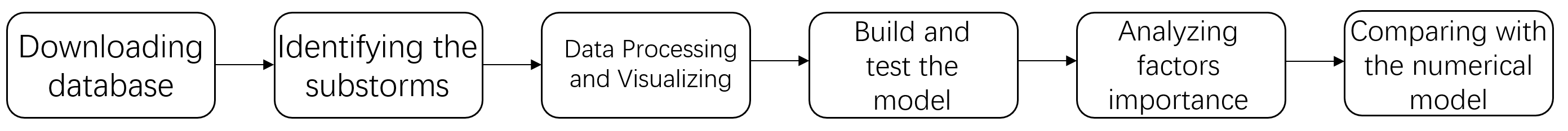


Figure 2. The pipeline for the study

**2.Database**

**1. Databases Description**

The forecasting objective can be described as a binary classification problem, therefore we need solar wind quantities to make the forecast and the SML index to label an event as being a substorm or not.

**A. SML Data**

To make the forecast for substorms, first, we need to identify the substorm events. The established method of identifying the onset of substorms is by the AL index as introduced, however, it has been pointed out that for substorm identification, using the SML index is more suitable and accurate [Hajra et al., 2016].

SML index is an advanced version of AL index, the name can be seen as the ‘Supermag’s version of AL index’. Supermag is a website that gives the collection of AL index [Supermag], it is better than AL index in the sense that it has a more comprehensive set of ground stations than what is used to give AL index. SML index is constructed by over 300 ground-based stations, covering a more latitudinal area, making it ideal for studying substorms.

The derivation of SML is the following: from each station, the magnetometer records the magnetic field near the ground, the north component of the field will be then used with the baseline removed, the field data is then corrected by each station’s geomagnetic latitude:



then average all stations we get the SML index.

The availability of SML data on the Supermag website starts from 1970, the finest provided solution is 1 minute. The data can be readily downloaded from the Supermag website in txt format. In this study, we are interested in substorm cases in the solar cycle 23 from 1996 August to 2008 December, using the 1-minute solution the total data size is roughly 170 MB.

**B. Solar Wind Conditions Data**

In addition to identifying the substorms, we also need the solar wind quantities data as the input to make the prediction. The solar wind conditions data is categorized on the OMNIweb website [OMNIweb], which is a coordinated NASA website collecting specific NASA missions’ data.

The missions that observe the solar wind conditions include Wind, GOES, Geotail, etc, their data is categorized on OMNIweb. The solar wind conditions quantities we are interested in in this study are the three components of the interplanetary magnetic field (IMF), namely IMF Bx, By, and Bz, and the solar wind pressure, velocity, and number density. Thus there are six solar wind quantities that we will employ to forecast substorms.

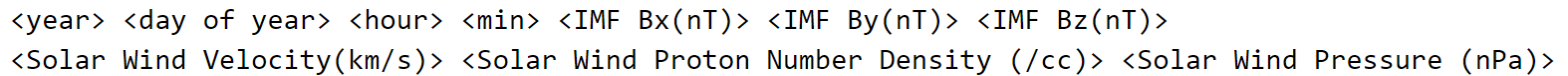
The IMF is observed by these missions’ onboard magnetometer, the solar wind pressure, velocity, and number density are calculated as the moments by the particle flux from the missions’ fluxgate facility.

The availability of these particular solar wind data on OMNIweb starts from 1995, the solution is 1-minute. Collecting the same period from 1996 August to 2008 December with 1-minute resolution can we get the database for solar wind conditions, the filesize is roughly 400 Megabytes.

The data is easily appropriately formatted, for the SML index the format is:



The solar wind conditions’ data format is:



Without the further operation, these data can be directly loaded. More steps about data processing and data visualization will be discussed in the Methods section after introducing what method will be used for identifying substorms.

**3. Methods**

**1. Finding substorms**

It is commonly acknowledged that the onset of the substorms can be identified by the sudden decrease of the AE index, and as introduced we will adopt the SML index to find the onset of substorms. According to Ohtani and Gjerloev [Ohtani and Gjerloev, 2020], the substorm events can be defined as the following as in Figure 3:

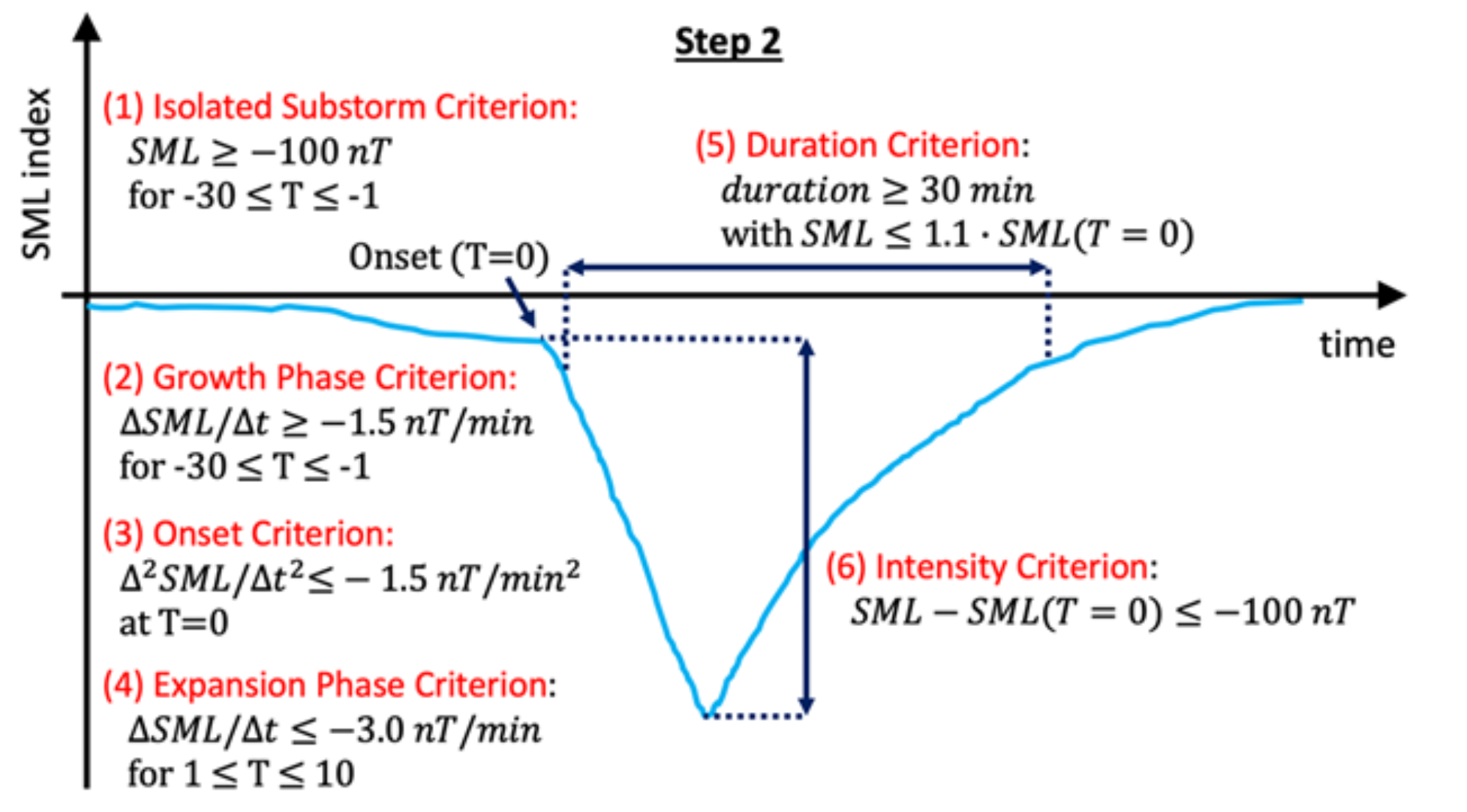


Figure 3. Identification of substorms using the SML index.

Firstly, for a substorm event to be isolated, the SML index must keep above -100 nT for at least 30 minutes before the onset of it, and during these 30 minutes, due to the requirement of the decrease of SML, the average SML decreasing rate should be over -1.5 nT/minutes. At the onset point, the decreasing rate should also be over -1.5 nT/minutes. After the onset is what is been called the expansion phase where we require the SML decreasing rate to be over -3 nT/minutes for 10 minutes. The shortest duration of each event should be over 30 minutes with the SML at the end less than 1.1 times the SML at the beginning. Lastly, the intensity of the substorm event should satisfy the condition that the difference between the peak SML value and the SML value at the beginning be over 100 nT.

Following these criteria, by searching the substorm events in 12 years in Solar Cycle 23, we have found 12204 cases substorm events, or roughly 3 cases per day on average.

**2. Visualizing Data**

After getting the substorms list, performing a case study to visualize the data is beneficial for later data processing and model building work. We select the substorm event on December 15th, 1996. The onset time of the substorm is 03:15 UT (Universal time), Figure 4 shows the six solar wind quantities along with SML value in the period starting from 3 hours before the onset and ending at the 3 hours after the onset.

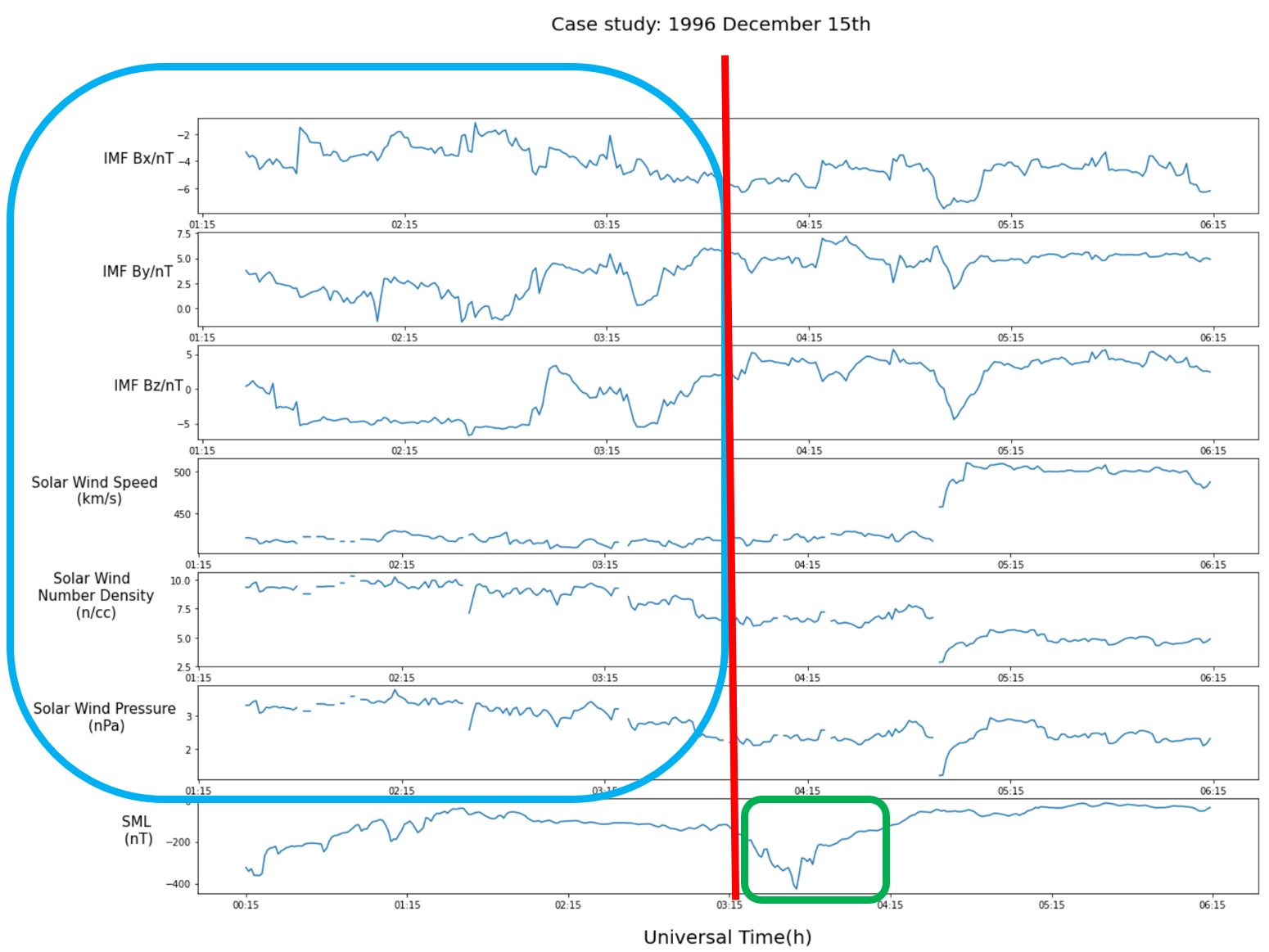


Figure 4. Case study of a substorm event on December 15th, 1996. The top to bottom panels shows the solar wind quantities (IMF Bx, By, Bz in nT, solar wind speed in km/s, solar wind number density in 1/cc, and solar wind pressure in nPa) and SML index in nT.

As the green box in the bottom panel shows, the SML sees a sudden jump after the onset of the substorms, which validates that this is a substorm event. The blue box shows the six solar wind quantities 3 hours before the onset, this gives us an intuitive about what correlation would be between these quantities and the substorm.

To statistically confirm what this correlation might be, we further conduct an epoch analysis over all the substorm events. As Figure 5 shows, for each substorms events in Solar Cycle 23, we choose the onset time as the epoch time and select the time series that start from 1 hour before the onset and end at 2 hours after the onset, then average all the substorms events we get the epoch analysis figure. From Figure 5 we can see before the onset, each solar wind quantities have certain features, for example, the solar wind velocity, pressure, and number density all see a slight amount of increase, but the most significant feature is that the IMF Bz is decreasing by roughly 1 nT. These features are correlated with the onset of substorms and can be identified by the neuron network.

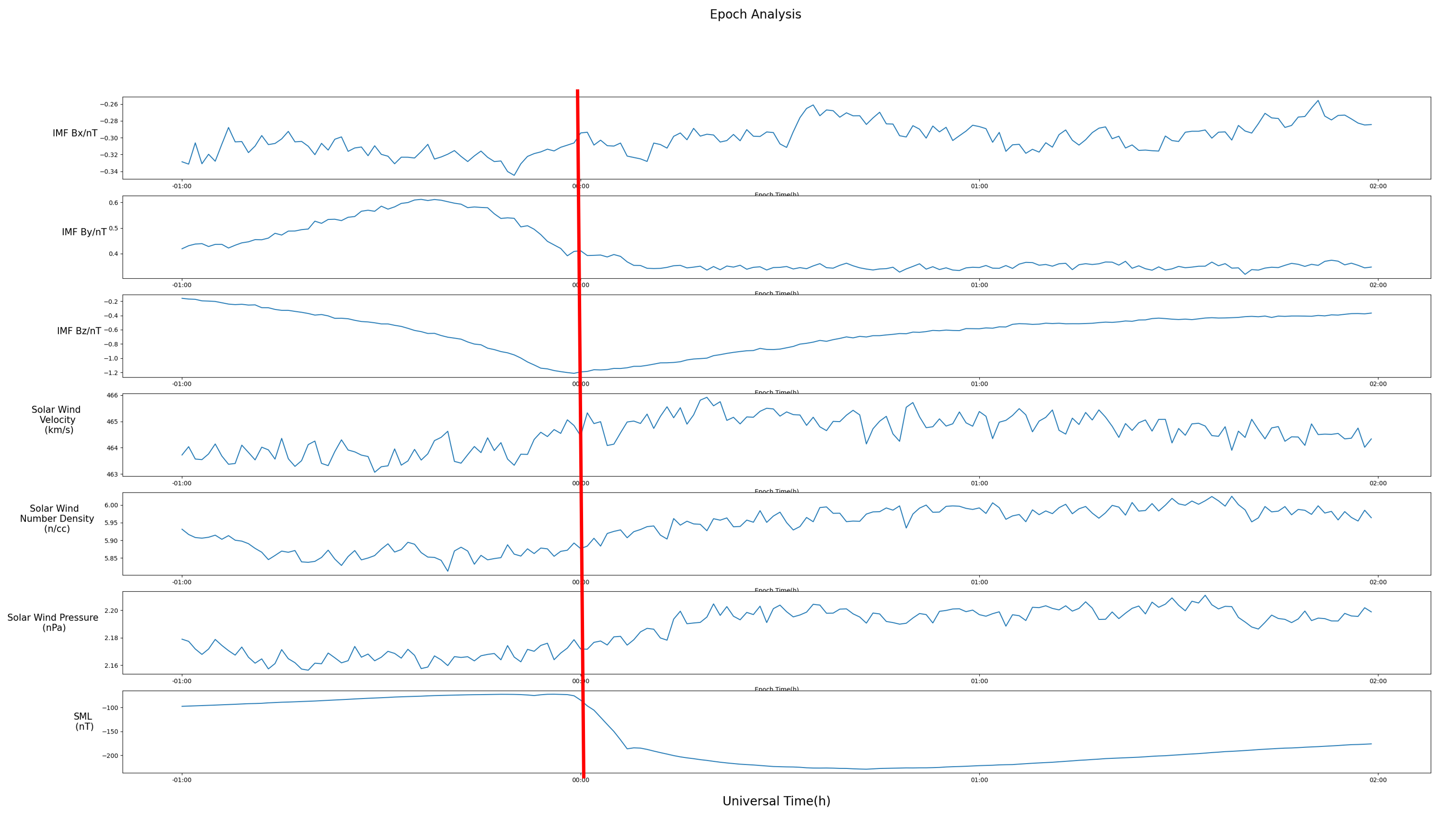


Figure 5. Epoch analysis, the selected period is 1 hour before the onset and 2 hours after the onset. Top to bottom panels are IMF Bx, By, Bz in nT, solar wind speed in km/s, solar wind number density in 1/cc, solar wind pressure in nPa and SML index in nT.

**3. Processing Data**

This section describes how the final data and label is prepared to be input into the machine learning model.

In addition to the substorm events that are already identified, non-substorms events are also needed to let the model learn how to distinguish between these two kinds of events. We randomly picked the non-substorm events over the entire 12 years period, and combine them with the substorm cases, and then shuffle the order, so that we have the collection of cases containing substorms and non-substorms, the order is random.

We then do the comparison between the substorm cases and non-substorm cases to show their distinguishment, as shown in Figure 6, for each substorm or non-substorm event we select time series 1 hour before the onset and average all the substorms or the non-substorms events. We can see there are clear distinguishments between the solar wind quantities between the two kinds of cases, the most notable difference is for the substorm cases, the IMF Bz undergoes a substantial decrease while for the non-substorm cases there is no such trend.

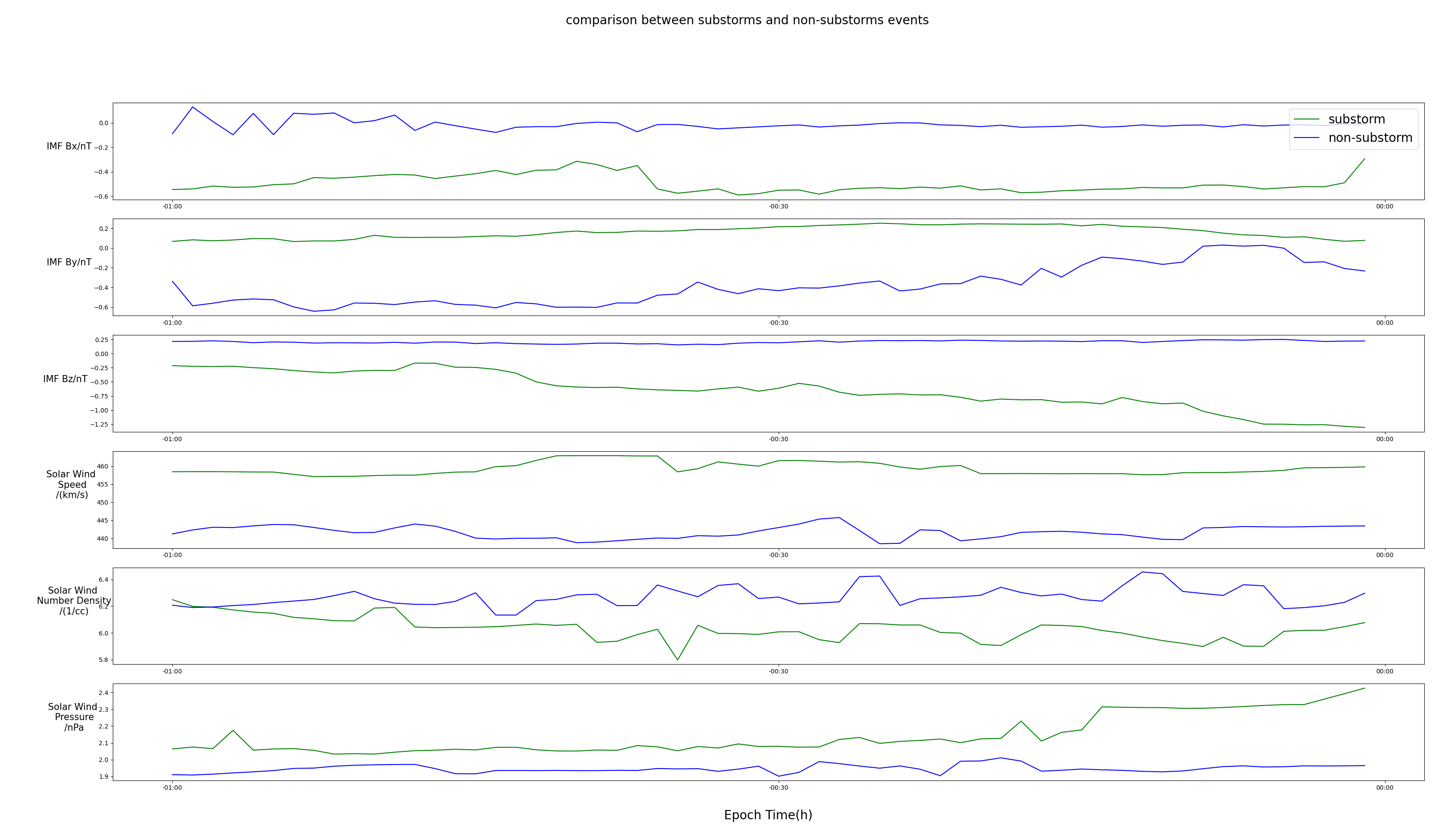


Figure 6. The comparison between the substorms cases and the non-substorm cases. All the substorms or the non-substorms cases are averaged. Top to bottom panels are IMF Bx, By, Bz in nT, solar wind speed in km/s, solar wind number density in 1/cc, solar wind pressure in nPa and SML index in nT.

The last step of data processing is to interpolate the matrix since, in our solar wind quantities data, there are NaN values, indicating no satellite observation at these times. If at some time the data is NaN, we will use the linear interpolation between the times before and after it to get the linear interpolated data.

The input data set is then prepared. As illustrated in Figure 7, Our input should be a 3-dimensional tensor, it consists of the time series of 6 solar wind quantities for all the cases. The time series duration we select is 60 minutes before the onset. Thus the tensor’s shape is 6 quantities by 60 minutes and by the number of cases we have.

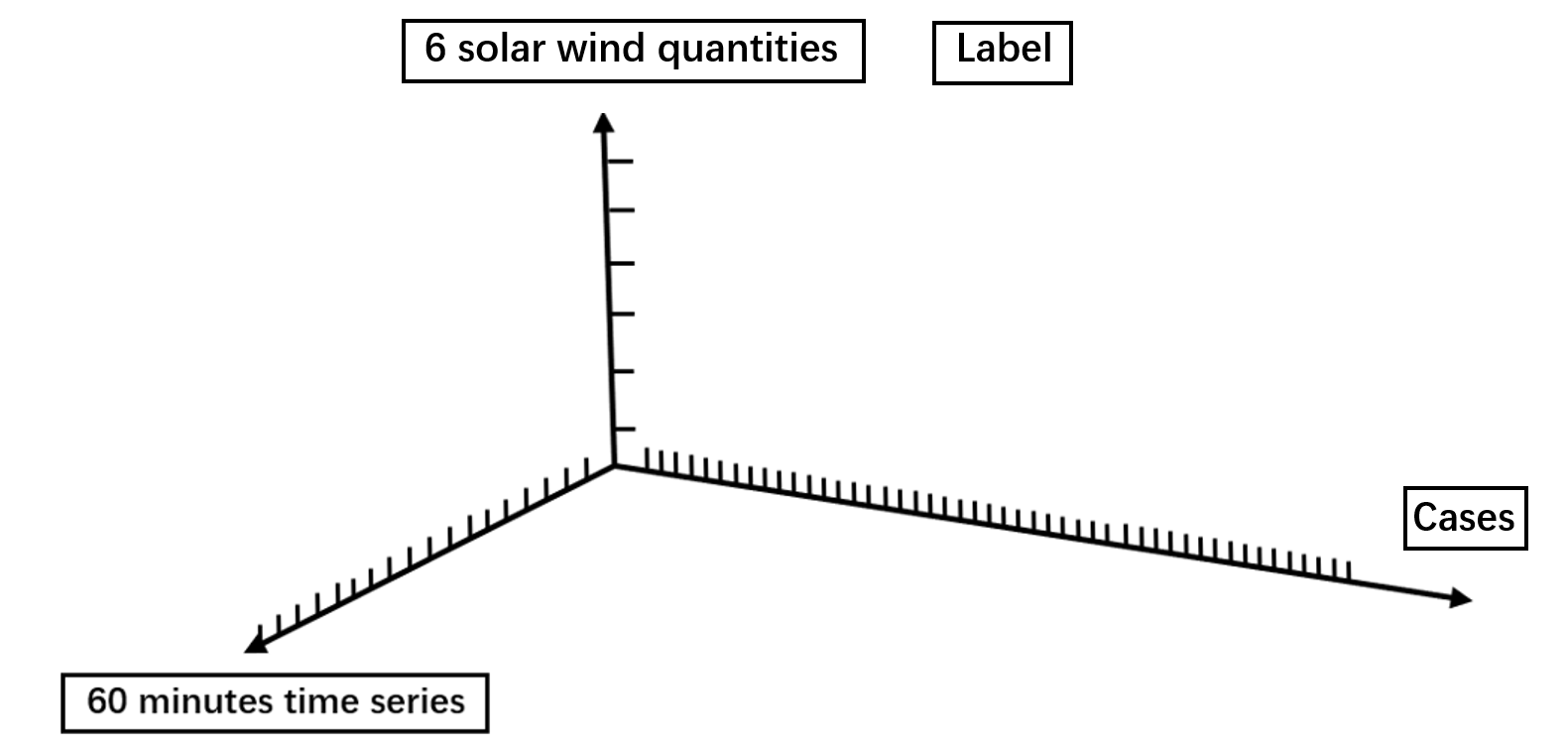


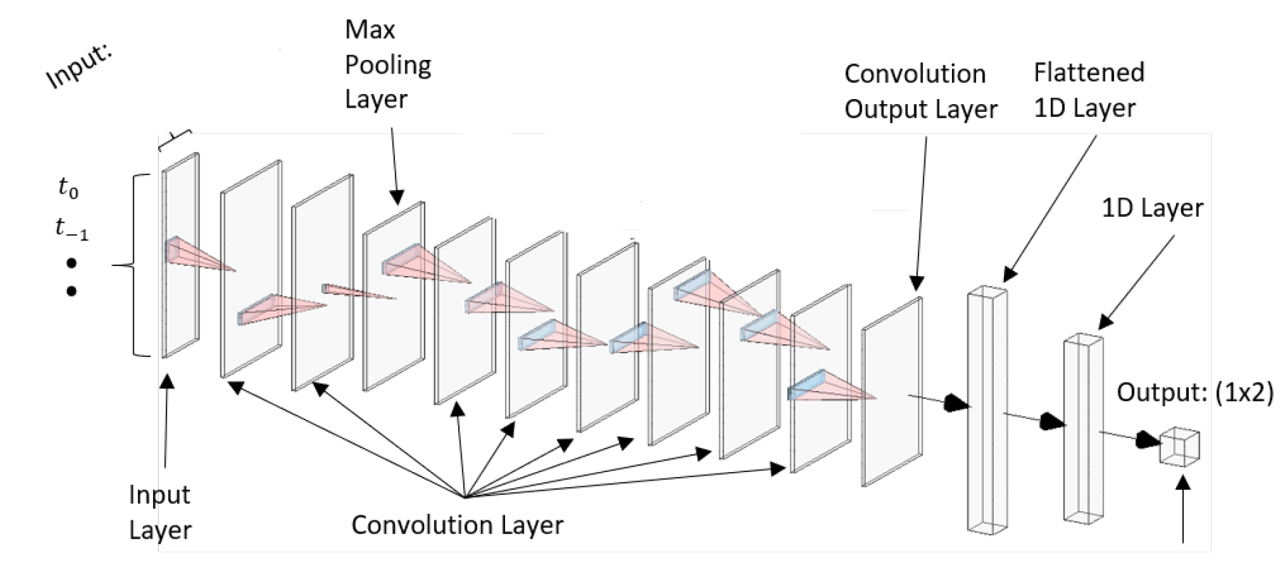
Figure 7. Illustration of the input matrix.

As a standard procedure for neuron network building, we will split our data into three groups: the training group, the validation group, and the test group, the first two will be used to compile the model and the test group will give the accuracy of the prediction for this study. We set 60% of the data to be the training data, 15% to be the validation data, and 25% to be the test data.

**4. Models**

**1. Convolutional Neuron Network (CNN)**

Convolutional Neuron Network (CNN) is arguably one of the most established machine learning models and has been extensively applied in many fields. Traditionally it is been used to identify the images, the current implementation is not limited to the images but applicable to very general matrix input. There have been some efforts of applying CNN in classification problems for time series-like input and they work reasonably well [Zhao et al., 2017]. In this study, we will first use CNN as an initial attempt for applying neuron networks to our time series tensor input.



Figute8. Illustration of CNN structure.

The CNN’s structure is illustrated in Figure 8, it consists of the input layer, convolution layer, flattened layer, and the output. The core idea is that in the convolutional layer, CNN uses the convolutional kernel to ‘scan’ through the input matrix to learn the information embedded in the input matrix, after flattened and selected activation function, it gives the output. In this study, the input tensor is the time series of 6 solar wind quantities that is introduced above, the input size is thus 120 by 6. The convolution layer is formed by 3 Conv2D layers and 2 Maxpooling layers, after it is the flatten layer and two dense layers, in the final dense layer, the selected activation function is ‘Sigmoid’. The optimizer we use is the ‘Adam’ with the learning rate to be 0.001.

The model summary is the following:

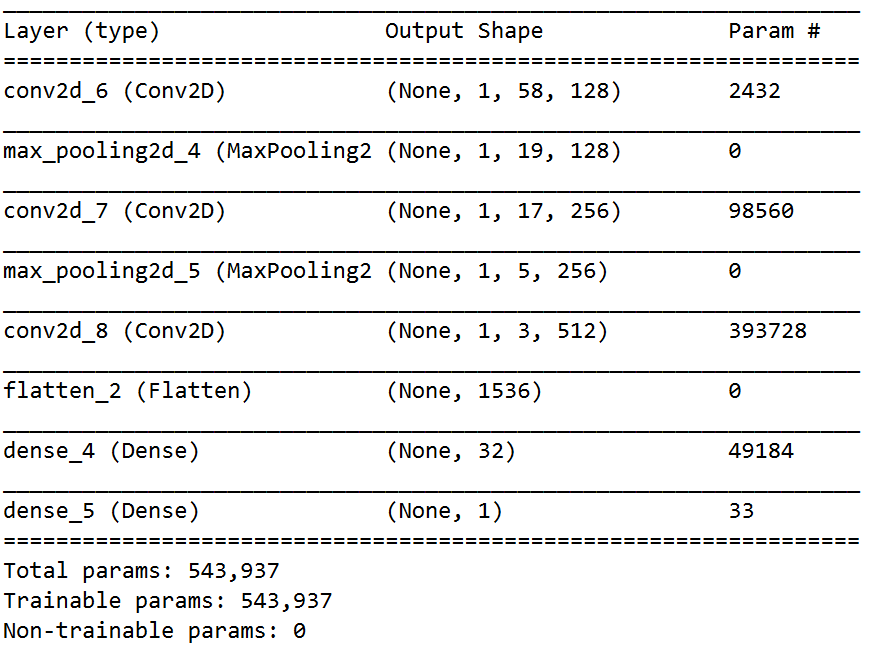


Figure 9. CNN model summary

There are 543937 trainable parameters, enough for the neuron network to learn the physics information.

**2. Long Short-Term Memory(LSTM)**

Another promising model is the long short-term memory (LSTM) model, which is an advanced version of the Recurrent Neuron Network (RNN), having the superb capability of memorizing the time order information of the input. Since our input tensor is time-series data, the LSTM is better than the CNN in the sense that it can memory the time information of the data, while CNN can only identify the local features of the data.

In our implementation of the LSTM, we add one embedding layer, one LSTM layer, and one dense layer. The optimizer we use is the ‘Adam’ with the learning rate to be 0.001.

The model summary is the following:

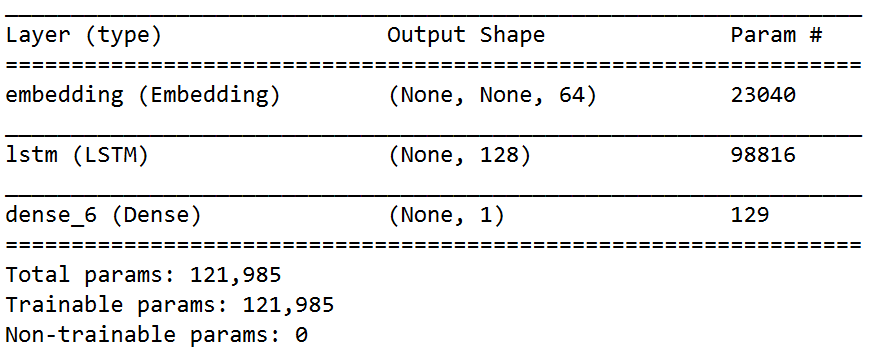


Figure 10. LSTM model summary

There are 121985 trainable parameters, enough for the neuron network to learn the physics information.

**5. Results Analysis**

**1. Test Accuracy**

For the CNN model, we first set epoch times to be 10, Figure 11 shows the accuracy and loss versus epoch times for training data and test data.

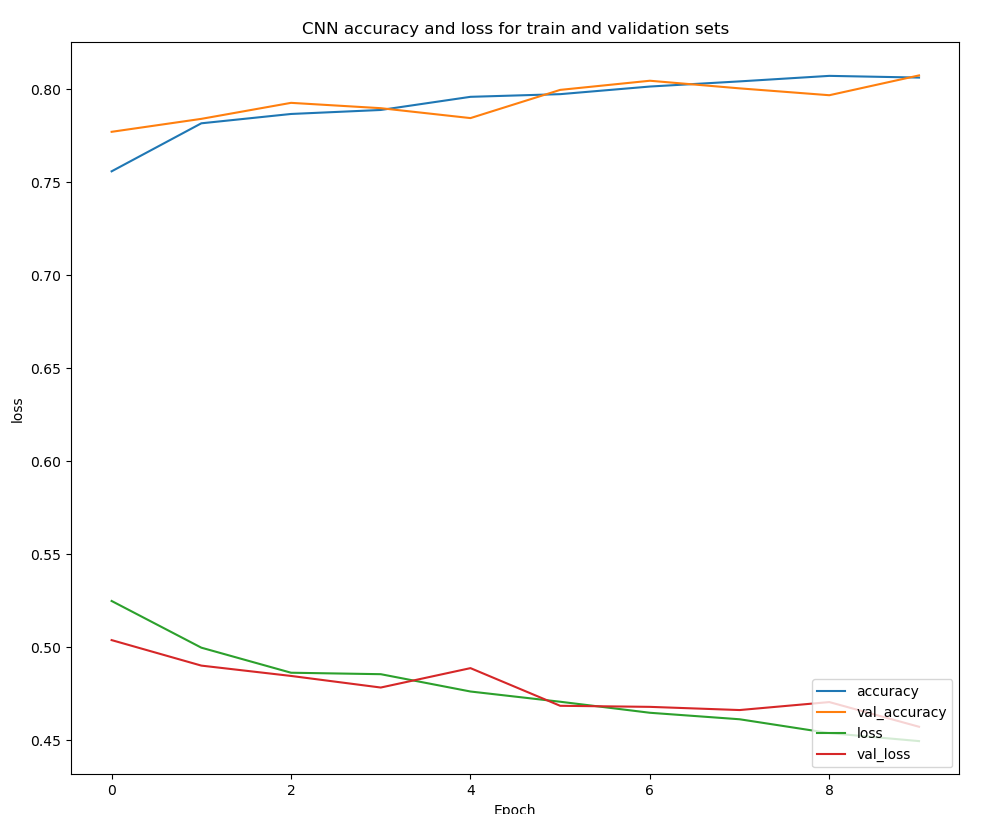


Figure 11. CNN model compiling result. The train and validation set accuracy and loss versus epoch time

The validation loss starts to increase at epoch time 3, which means the overfitting appears at epoch 3, after that, the model is learning the artificial features, not the physics ones. So we should train the model with epoch time equals to 3. Then we input the test set, the prediction accuracy is 78.96%.

For the LSTM model, again we first set the epoch times to be 10, Figure 12 shows the accuracy and loss versus epoch times for training data and test data.

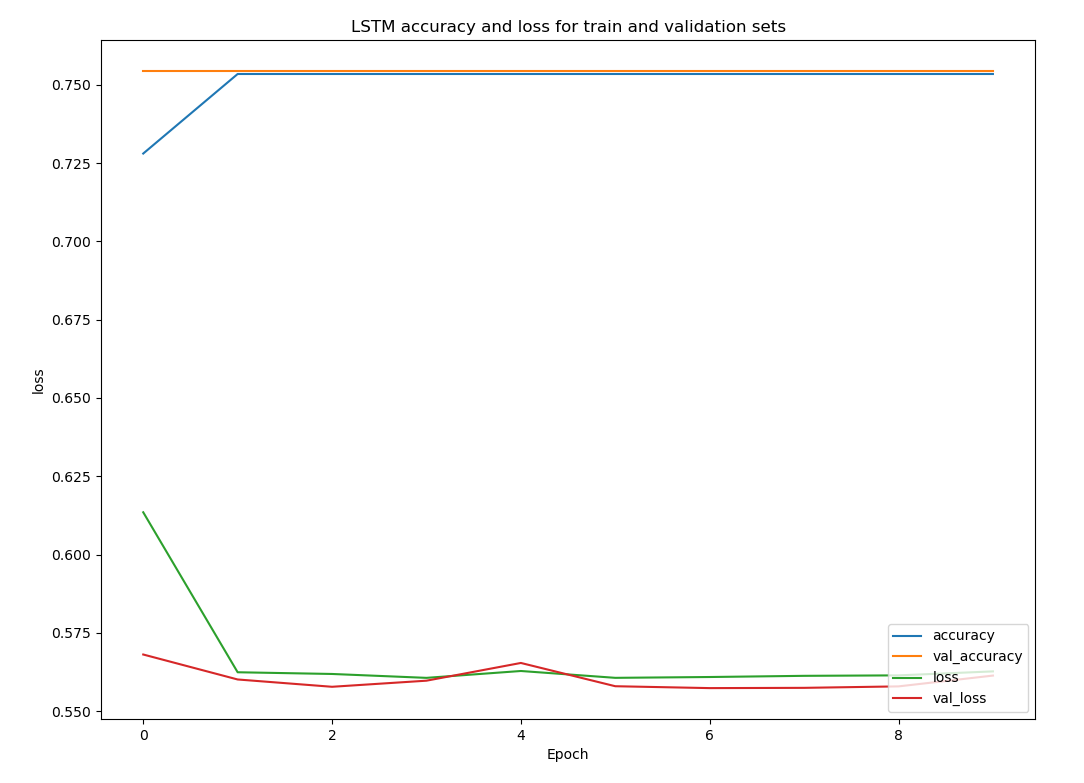


Figure 12. LSTM model compiling result. The train and validation set accuracy and loss versus epoch time

The overfitting appears at epoch time 2 when the validation loss is starting to increase, thus we compile the model using the epoch time equals to 2 and input the test data, the predicted accuracy is 73.90 %.

**2. Permutation Importance Analysis**

Apart from the machine learning aspect, in this study, we are also extremely interested in the physics of the substorms, especially the physics interpretation of our result. One important insight we want to extract from this forecast study is which factors are important for forecasting substorms. This would serve as a strong hint for the physical driver for substorms. To achieve this, we do the permutation importance analysis.

The logic of the permutation importance analysis is that we add some random noise to one factor such as the Bx in this study, then run the model and give the new prediction. If the new prediction deviates significantly from the original one, this means this particular factor is important in our model for prediction.

The result of the permutation importance analysis is as the following:

IMF Bx: 0.132621, IMF By: 0.102764, IMF Bz: 0.964451, solar wind velocity: 0.003472 solar wind number density: -0.001389, solar wind pressure: 0.204139.

We can see that IMF Bz is the most important factor. This is anticipated when we do the data visualization in the Method section where we see from the epoch analysis that IMF Bz is indeed an obvious and significant feature for substorms cases.

**3. Comparison with Numerical Model (RCM)**

One innovation in this study is to compare our result from the neuron network to the numerical model that we have at Rice -- the RCM model. RCM, or the Rice Convective Model, is a numerical model, it solves the physics partial differential equations that describe the plasma dynamics in the magnetosphere, so as to provide reasonable simulation in the magnetosphere. The innovation of this comparison lies in that the results from the neuron network and the RCM come from two entirely different approaches, one from extracting the features of solar wind quantities before the onset of substorms, the other one is from the governing physics equations of substorms. The consistency of these two approaches would strongly validate the conclusion that we draw and provide insight about how to combine the data analysis method and the simulation in the space weather study.

The attached movie S1 shows a simulation about the substorm event -- Galaxy 15 that we introduced in the Introduction section. Figure 13 shows a frame of the movie, the color bar shows the pv\_gamma or the entropy. The blue structure near X=-20 Re is a depleted bubble. This low entropy bubble in the tail of the magnetosphere can trigger a substorm. Figure 14 shows that the Bz is the magnetosphere is decreasing just as the same in the result of the neuron network where we conclude the decreasing Bz is crucial evidence of the happening of a substorm. This agreement firmly supports the scientific validity and reproducibility of our results.

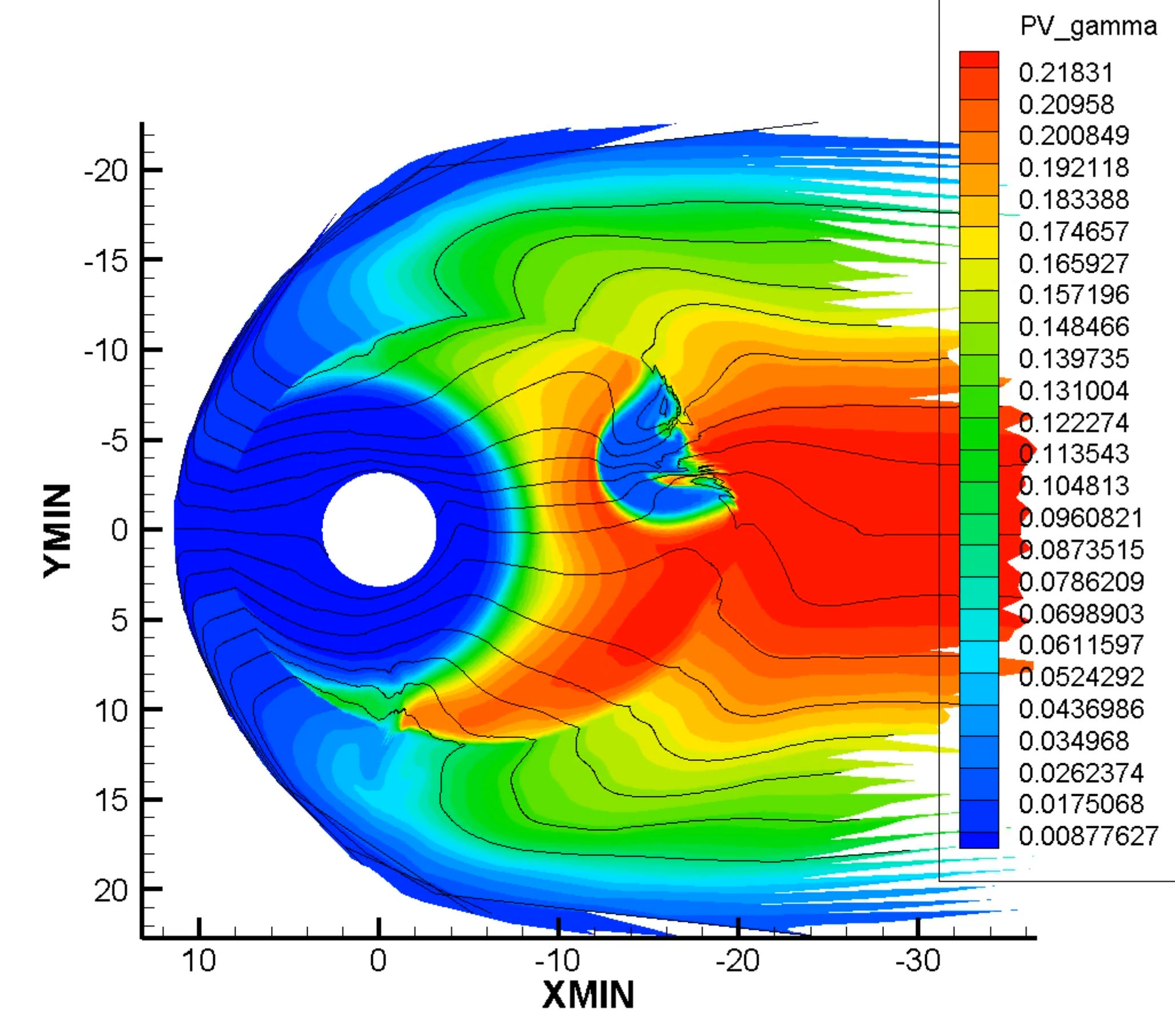


Figure 13. A frame of RCM simulation, the color bar shows the pv\_gamma or the entropy. The blue structure near X=-20 Re is the depleted bubble

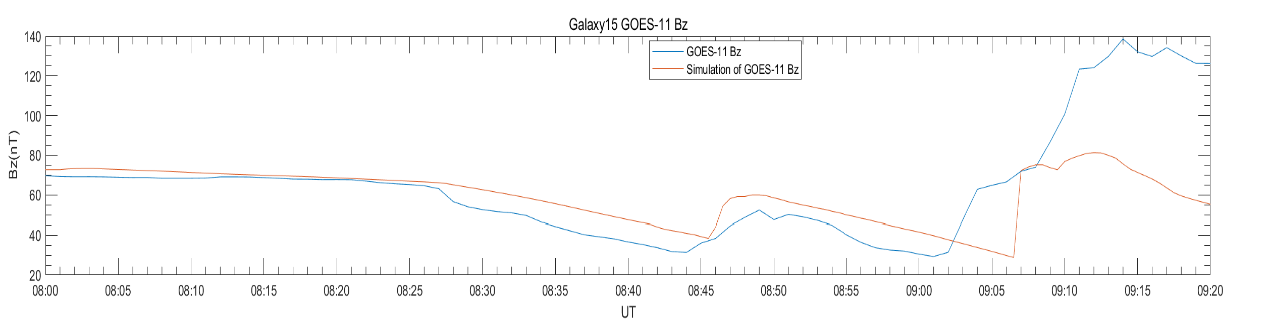


Figure 14. Bz decrease in Galaxy 15 substorm event, the orange line is the simulation result, the blue line is the observation.

**6. Conclusions**

**1. The Unprecedented High Prediction Accuracy**

The test accuracy for CNN is 78.96%, which is much higher than reported by Lyons et al. [Lyons et al., 1997] where they used the simple indicator such as the IMF Bz turning to predict the substorms and ended up getting poor prediction accuracy, whereas, in our scheme, the neuron network learns enough information about the pattern and tendency of the six solar wind quantities to get a much more reliable result. It hints that rather than having the simple driver, the onset of substorms has complicated reasons that involve the comprehensive evolution of solar wind conditions, further simulation and theoretical study could confirm this and hopefully can give the full explanation for substorms.

Our test set cases are randomly drawn from 12 years, under all space weather conditions such as different solar conditions and geomagnetic conditions, the fact that it still delivers a good prediction accuracy proves the robustness of our model. The approach and the model used in this study could serve as the beginning of a practically usable model for forecasting substorms, whose value in avoiding severe space weather and understanding the space is boundless.

**2. CNN and LSTM Comparison**

The CNN gives an accuracy of 78.96%, the LSTM gives an accuracy of 73.90 %. The two prediction accuracies are close, and both converge very fast, meaning that both neuron networks learn the physics information quickly. This also suggests that the utmost accuracy that we can get from these neuron networks is about 80%.

The number of trainable parameters for LSTM is 121985, much less than the CNN, but the LSTM model still gives a close accuracy as the CNN, which may be due to the LSTM’s memory feature, which makes it better for time series problem in this study, and in the physics side, this also means that the time dependence of these solar wind quantities is important for the onset of substorms.

**3. Consistency Between the Neuron Network and Numerical Model**

The results we get from the neuron network and the numerical model correspond well: both showed that the decrease of Bz is highly related to the happening of substorms. This correspondence verifies our model building and gives us insight that the Bz is suggestive of the onset of substorms.

**4. Significance, Innovation, and Reproducibility**

The significance of this study is that we have built a reliable and robust model for forecasting the substorms. The unprecedented high accuracy of the model is surprisingly exciting, providing ways for building a real space weather forecast system.

The approach that we employed is seldomly seen in the traditional paradigm of space weather study. Previously the space weather study emphasizes using simulation or observation data to understand the physics of events in space, our success in applying deep learning techniques to forecast substorms shows the other way around: by understanding and learning the data structure can we extract the physics, which is a meaningful and promising approach that complements the traditional paradigm we have today.

The techniques and approaches used in this study are appropriate and justified: the 6 solar wind quantities that we choose to make the forecast do correlate with the onset of substorms, this is justified by the epoch analysis, the permutation importance analysis, and the RCM simulation; We use the time series that are 60 minutes before the onset, this is an optimal choice of the amount of time, taking both the average duration of substorms and prediction effect into account; The two neuron networks are also proper for building this forecast model since they give close and high accuracies with a quick converging speed.

This is study is also scientifically reproducible: In building the neuron network, two different models give us roughly the same accuracy; the permutation importance analysis verifies the conclusion that we have by data visualization; the neuron network model and numerical model lead to the same conclusion, these all demonstrate that this study has valid scientific reproducibility.

**Reference**

Forsyth, C., Rae, I., Murphy, K., Freeman, M., Huang, C., Spence, H., Boyd, A., Coxon, J., Jackman, C., Kalmoni, N. and Watt, C., 2016. What effect do substorms have on the content of the radiation belts?. Journal of Geophysical Research: Space Physics, 121(7), pp.6292-6306.

Hajra, R., Tsurutani, B., Echer, E., Gonzalez, W. and Gjerloev, J., 2016. Supersubstorms (SML < −2500 nT): Magnetic storm and solar cycle dependences. Journal of Geophysical Research: Space Physics, 121(8), pp.7805-7816.

Loto'aniu, T., Singer, H., Rodriguez, J., Green, J., Denig, W., Biesecker, D. and Angelopoulos, V., 2015. Space weather conditions during the Galaxy 15 spacecraft anomaly. Space Weather, 13(8), pp.484-502.

Lyons, L., Blanchard, G., Samson, J., Lepping, R., Yamamoto, T. and Moretto, T., 1997. Coordinated observations demonstrating external substorm triggering. Journal of Geophysical Research: Space Physics, 102(A12), pp.27039-27051.

Ohtani, S. and Gjerloev, J., 2020. Is the Substorm Current Wedge an Ensemble of Wedgelets?: Revisit to Midlatitude Positive Bays. Journal of Geophysical Research: Space Physics, 125(9).

Supermag, Supermag.jhuapl.edu. 2021. SuperMAG: About SuperMAG. [online] Avail-

able at:<https://supermag.jhuapl.edu/info/>

OMNIweb, OMNIweb.gsfc.nasa.gov. 2021. SPDF - About OMNIWeb Data. [online]

Available at: <https://omniweb.gsfc.nasa.gov/html/ow_data.html>

Movie S1, accessible at https://drive.google.com/file/d/1s6aKx2k17YuhHK6RtJG3QX6C--QC5Q9a/view?usp=sharing.