Predicting Substorms from Magnetometer and Solar Wind Data

Greg Starr
Alan Zhou

Introduction

Magnetic storms and magnetic substorms are natural phenomena where disturbances in the Earth's magnetosphere cause energy to be released from the tail of the magnetosphere into the high latitude ionosphere. Magnetic storms are the rarer event, typically taking about half a day to develop and decaying over a period of a few days. In these events, the decreases in the magnetic field occur world wide and as such are observable from any point on Earth. Substorms, on the other hand, are more frequent and more local events. Across the globe they typically occur multiple times on any given day and are localized to specific regions of the planet, typically observable at the poles and in space.

The two phenomena are certainly related, and during magnetic storms one can often observe especially intense substorms in the polar regions. However, the exact relationship and the factors that cause both kinds of storms is not well understood. There are known measurable quantities that are known to be predictive for magnetic storms and substorms, but there does not exist an effective robust method of actually predicting their occurrences.

While the most visible effect of substorms manifests itself in an increased intensity in polar auroras, the most significant effects come from the disruptions that substorms cause to telecommunication, navigation, and spacecraft.

Because of the unknown nature of the function that leads to substorms, we suspect that a deep learning algorithm might be perfectly suited for substorm prediction.

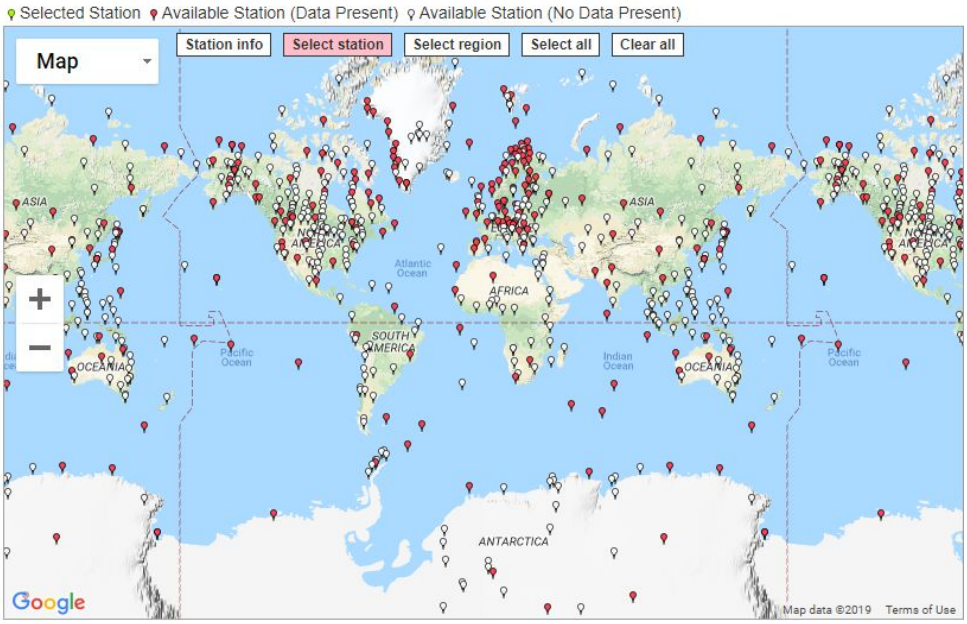
Formal Definition

The problem we are proposing is as follows. Given time series data measured from a set of magnetometers as well as their locations we would like to predict the time and location of future substorms as far out as possible and with as much precision as possible. This definition is intentionally vague because we would like to start with an easier task and increase the difficulty if we have success. To start, we would like to input a block of T minutes of data from all stations which have "good" data and output a binary label which will say whether a substorm will occur in the next hour. Here "good" means that the data has few or no missing values for that interval. Note that different intervals could have a different number of stations with good data. A second iteration could provide two outputs: the binary label for a substorm occurring in the next hour, as well as a 2D location on the earth where the substorm will occur. This is tricky because then we have to decide how to handle the case where multiple substorms occur in the next hour. In order to make it a little harder, we could try binning the prediction interval into smaller intervals, turning it into a multiclass classification problem. Finally we could predict the exact time of the next substorm, making this a regression problem. Each version of this problem has its own advantages and complications, some of which will be described later. One additional

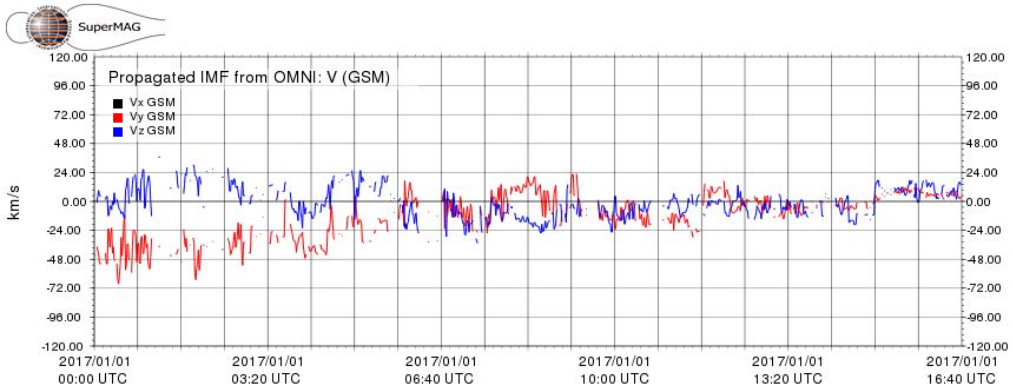
augmentation would be to include other sources of information, time series or otherwise, to make our predictions.

Dataset

The Supermag dataset (http://supermag.jhuapl.edu/) contains time series data from over 300 stations around the globe that measure the magnetic field in the xyz components at one minute intervals. Data exists as far back as 1975, but the further back in time one looks, the fewer stations there are. Most time intervals contain data from only a subset of the full network of stations. The Supermag also provides a list of substorm occurrences including both their time and location. Finally, Supermag also provides historical solar wind data which is known to be a good predictor of substorms. The solar wind data is particularly patchy, with many intervals missing values.



Magnetometer station locations



Patchy solar wind data

Model

Most of the time spent on this project will be figuring out what model architecture is most effective, but here we will discuss some of our initial ideas. So far we have only considered using a CNN. The simplest architecture would be one that only inputs one station at a time and makes a prediction for the local area around the station. Even this simple architecture has merits in that we can run the model on every available station and see where in the world the substorm is most likely. On the other hand, an architecture like this can only use local information to make its predictions. A better architecture might aggregate information from different scales to improve its outputs. This is difficult however because we may have a different number of stations each interval. An interesting paper we found called *Deep Sets* (https://arxiv.org/abs/1703.06114) may be applicable to our model, since we are to some degree interested in permutation invariant knowledge - we care less about which station is telling us where a substorm is happening and more about whether or not a substorm is happening at a certain location. The order of the substations we consider has no meaning in the context of our problem. GaitSet, a model which can identify a person based on a set of silhouettes, is a good example of the successful application of set-based architecture to a complex problem.
Another option is to use a recurrent architecture which may be a good idea because these are often used to model time series'. Whatever architecture we use, we want it to be able to handle variable length inputs (varying number of stations) and possibly variable length outputs.

Technical Difficulties

- Data input: since the data comes from a wide variety of stations across a long time range there are a lot of missing values in the data. We need to find a way to deal with inconsistencies in the data. This could mean allowing variable length input to our model, or filling in missing values. The solar wind data has many intervals of missing values which will be especially challenging to deal with
- Data fusion: since we plan on using data from two different sources, we need to find a way to properly combine the magnetometer data with solar wind data, especially if the predictive power of each data source operates on a different time scale.
- Simultaneous substorm instances: We need to be able to distinguish between single substorm instances and multiple simultaneous substorm instances.

Proposed Work

We have already outlined several of our ideas and most of the work of this project will be figuring out how far we can go with this problem. If we are able to predict substorms with a high degree of accuracy, it would also be interesting to probe the model and see what kinds of features are important for substorm prediction.

Expected Deliverable

We expect to be able to deliver a model that can predict substorm incidence. We hope to be able to predict location as well as time, but the difficulty of that task is as of yet not well known.

A secondary deliverable we hope to provide is a visualization of substorm probability - this will most likely be in the form of a heat map.

References

1. https://en.wikipedia.org/wiki/Substorm
2. https://www-spof.gsfc.nasa.gov/Education/wsubstrm.html
3. https://cdaw.gsfc.nasa.gov/publications/ilws_goa2006/320_Lakhina.pdf
4. Deep Sets - https://arxiv.org/abs/1703.06114
5. GaitSet - https://arxiv.org/pdf/1811.06186.pdf
6. Newell, P. T., K. Liou, J. W. Gjerloev, T. Sotirelis, S. Wing, and E. J. Mitchell (2016), Substorm probabilities are best predicted from solar wind speed, J. Atmos. Sol. Terr. Phys., 146, 28–37, doi:10.1016/j.jastp.2016.04.019.