Using an LSTM and Classification Methods to Determine Risk of dB/dt Threshold Crossings as Proxy for Geomagnetically Induced Currents

Michael Coughlan[1], Amy Keesee[1,2], Victor Pinto[2], Hyunju Connor[3], Jeremiah Johnson[4]

[1]Department of Physics and Astronomy - University of New Hampshire, [2]Institute for the Study of Earth, Oceans, and Space - University of New Hampshire
[3]Department of Physics and Geophysical Institute - University of Alaska Fairbanks, [4]Department of Electrical and Computer Engineering - University of New Hampshire

MAGICIAN
Machine Learning Algorithms for Geomagnetically Induced Currents in Alaska and New Hampshire

Introduction

- The interaction between the solar wind and the Magnetosphere can produce Geomagnetically Induced Currents (GIC's) on the ground, which are capable of causing power outages and damage to crucial infrastructure.
- The ability to predict when and where these events may occur could allow us to avoid the worst of this damage.
- The use of physics-informed machine learning models can offer a computationally inexpensive method of predicting GIC events using horizontal dB/dt as a proxy, though most models thus far have fallen short of consistently accurate predictions. dB/dt was defined as:

$$\frac{dB_H}{dt} = \sqrt{\frac{dE}{dt}^2 + \frac{dN}{dt}^2}$$

With N and E the North and East components of the magnetic field respectively.

- Here, a Long-Short Term Memory (LSTM) model was used to determine the risk of dB/dt going over thresholds of 9, 18, 42, 66, and 90 nT/min for the Ottawa (OTT) ground magnetometer station.
- Three storms were chosen for testing and removed from the training set: March 30, 2001 (~ -211nT), December 14, 2006 (~ -437nT), & August 05, 2011 (~ -126nT).
- The storms were chosen for several reasons; they represent different storm intensities, they occurred at different points in the solar cycle, and there are minimal gaps in the data that needed to be interpolated over.

Method/Data

- The model was trained exclusively on storm time data as defined by a SYM-H value of -50 nT or less for a minimum period of 2 hours.
- The storm data was extracted from a combined data frame of OMNI data and Supermag data from the Ottawa (OTT), mid-latitude station.
- The input features included solar wind speed (VT, Vx, Vy, Vz), IMF_GSE (BT, Bx, By, Bz), proton density, dynamic pressure, solar wind electric field, SYM-H, horizontal magnetic field (N,E), and ground magnetometer sin(MLT) and cos(MLT).
- The LSTM layer utilized 30 minutes of time history to determine if the dB/dt value would go above a series of thresholds, between 30 and 60 minutes into the future.
- The machine learning model consisted of a single LSTM layer with 'RELU' activation and a Dense output layer using 'softmax' activation, implemented using TensorFlow with the Keras backend.
- The 'softmax' activation layer allows the model to interpret the inputs to the layer as discrete probability distributions, allowing us to interpret the outputs of a node as the probability of that node occurring. In this case the output of the node is the probability that dB/dt will cross the given threshold.
- The softmax activation can be described:

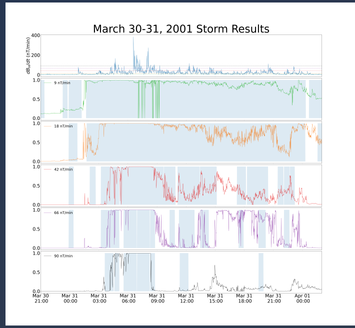
$$softmx(x_i) = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$$

- The Probability of Detection (POD), Probability of False Detection (POFD), Frequency Bias (FB), and Heidke Skill Score (HSS) were calculated. These metrics rely on the comparison of whether the actual and predicted values crossed certain thresholds within a defined time period.
- The metrics above utilize a comparison of actual and predicted threshold crossings, where: A is a True Positive, where both the actual and predicted cross the threshold, B is a False Positive, the actual does not cross but the predicted does, C is a False Negative, the actual crosses but the predicted does not, & D is a True Negative, neither actual nor predicted cross the threshold.
- To determine threshold crossings in the actual data, we calculated the maximum value in the 30 minute prediction window and compared that value to the thresholds. Because the softmax activation function outputs a probability, predicted values greater than or equal to 0.5 were considered positive predictions.
- The metrics, as well as the the Precision and Recall metrics used for plotting the Precision-Recall curves, are defined as:

$$POD = \frac{A}{A+C} \qquad FB = \frac{A+B}{A+C} \qquad Precision = \frac{A}{A+B}$$

$$POFD = \frac{B}{B+D} \qquad HSS = \frac{2(AD-BC)}{(A+C)(C+D)+(A+B)(B+D)} \qquad Recall = \frac{A}{A+C}$$

Plots



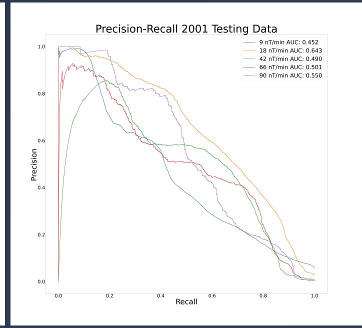
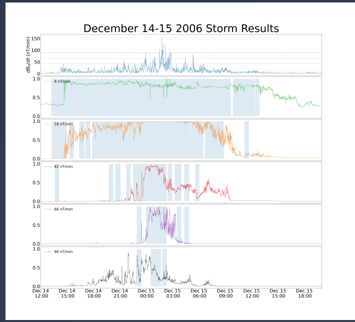


Figure 1a: Model results for the March 2001 storm. First panel on left side is the true dB/dt values with horizontal lines indicating each threshold. The subsequent panels are the results for each threshold with the blue shaded area being a threshold crossing of the real data, and the other colors indicate the model's predicted probability of crossing the respective thresholds 30-60 minutes into the future. Figure 1b: The Precision-Recall curves for the 5 different thresholds., for the full testing data from January 2001 – June 2001.



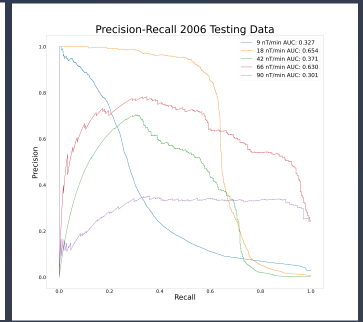
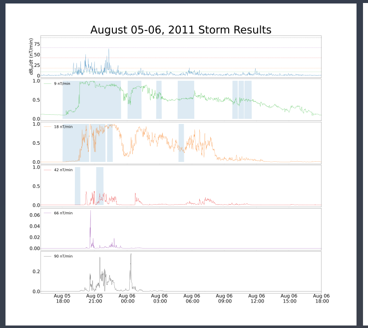


Figure 1a: Model results for the March 2001 storm. First panel on left side is the true dB/dt values with horizontal lines indicating each threshold. The subsequent panels are the results for each threshold with the blue shaded area being a threshold crossing of the real data, and the other colors indicate the model's predicted probability of crossing the respective thresholds 30-60 minutes into the future. Figure 1b: The Precision-Recall curves for the 5 different thresholds., for the full testing data from September 2006 – March 2007.



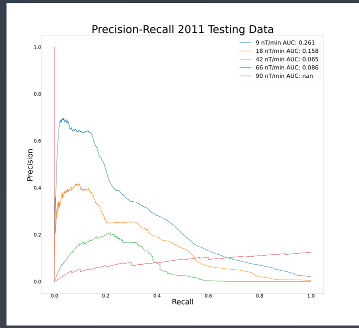


Figure 1a: Model results for the March 2001 storm. First panel on left side is the true dB/dt values with horizontal lines indicating each threshold. The subsequent panels are the results for each threshold with the blue shaded area being a threshold crossing of the real data, and the other colors indicate the model's predicted probability of crossing the respective thresholds 30-60 minutes into the future. Figure 1b: The Precision-Recall curves for the 5 different thresholds., for the full testing data from May 2011 – November 2011.

Metrics/Precision-Recall

- The Precision-Recall curves are a way of evaluating the skill of a model when dealing with an imbalanced dataset. Table 1 shows the percent of positive classes for each threshold in the testing set. For these degrees of imbalance, a model of no skill would have an Area Under the Curve (AUC) of near 0.
- None of the models exhibit the near perfect score of 1, but all of the models show some skill.

Thresholds	2011 Data	2006 Data	2001 Data
9	2.187	2.498	5.860
18	0.488	0.509	2.492
42	0.081	0.164	0.759
66	0.010	0.078	0.441
90	0	0.043	0.268

Table 1: Percent of positive classes for each testing set

- The metric scores in Table 2 were calculated using a greater or less than value of 0.5 for the predicted probability of a threshold crossing. With a predicted probability of 0.5 or greater being considered a 1 and less than 0.5 being a 0.
- Two different models were implemented to optimize metric scores. The 9, 18, and 90 nT/min thresholds used a model that included an additional 30 minutes of the recovery phase of the storm in the training data, while the models for the 42 and 66 nT/min thresholds produced better scores without the extra recovery time.
- All of the models achieved low POFD scores, indicating very few false alarms predicted.
- Several of the models were able achieve high POD scores, with the lowest threshold being above 0.7 and getting slightly worse as the thresholds are increased and the number of crossings is decreased.
- This is born out more in the FB scores, few of which are close to the perfect score of 1. A score below 1 indicates the model is predicting fewer crossings than the real data, which will artificially make the POFD score lower, and a score above 1 means the model is predicting more crossings, inflating the POD scores.
- The August 2011 storm scored the lowest across the board for the HSS scores. This is most likely a product of it being the least intense storm of the three examined. The December 2006 storm preforms the best across the board with the exception of the 90 nT/min threshold where the March 2001 storm scores higher. No HSS scores exceed a score of 0.8, meaning there is more work to be done to improve the models.
- It is important to point out that all of the models miss the initial spike in dB/dt. If this cannot be resolved, it could negate much of the utility of this type of model.

Threshold	POD2011	POD2006	POD2001	POFD2011	POFD2006	POFD2001	FB2011	FB2006	FB2001	HSS2011	HSS2006	HSS2001
9	0.706	0.936	0.901	0.298	0.356	0.190	1.212	1.066	0.922	0.391	0.615	0.536
18	0.460	0.857	0.831	0.219	0.127	0.277	1.509	0.948	0.891	0.201	0.722	0.462
42	0.000	0.327	0.377	0.001	0.006	0.102	0.028	0.343	0.444	-0.002	0.409	0.239
66	NaN	0.445	0.552	0.000	0.018	0.325	NaN	0.576	1.133	NaN	0.522	0.220
90	NaN	0.137	0.578	0.000	0.034	0.028	NaN	0.595	0.687	NaN	0.128	0.622

Table 2: The metric scores for the thresholds of the three storms examined. A perfect score for the POD, FB and HSS is 1 and a POFD is 0

Acknowledgements

This work was supported by NSF EPSCoR Award OIA-1920965

Additional MAGICIAN team posters: Connor (2) & Keesee (57)