Forecasting low-energy particle flux in LEO : data analysis and first results

Simon Bouriat[1,2,3,4], Mathieu Barthélémy[1,2], Jocelyn Chanussot[3]

1. Univ. Grenoble Alpes, CNRS, IPAG, F38000 Grenoble, France
2. Univ. Grenoble Alpes, CSUG, F38000, Grenoble, France
3. GIPSA-Lab, Grenoble Institute of Technology, Grenoble, France
4. SpaceAble, 89 Boulevard Magenta, Paris, 75010

Introduction

Low-energy particle fluxes: Here defined as **precipitated electrons** (i.e., electrons carried along magnetic field lines from the magnetosphere to the ionosphere) in LEO with **energies < 40 keV**.

Existing models: First models appeared as early as 1987. The current state-of-the-art is here considered to be the last version of OVATION Prime [2] and PrecipNet Deep Learning model [1].

This PhD topic is part of a larger project that aims at creating a rigorous statistical and empirical code to predict the particles' environment in LEO and GEO. The final purpose is to prevent hazards on satellites such as *surface charging*, *internal charging* or *total ionizing dose*, and their consequences such as *electrostatic discharge* or *single-event effects*, by observing the upstream Sun-Earth chain (from near-Sun with SOHO, to L1 with ACE, GEO with GOES and LEO with DMSP & POES).

Here our work is divided into two parts: the data analysis of ACE data gathered from the ACE Science Center (ASC) [3] and first results over McGranghan et al. [1] datasets using Deep Learning. This first step is then a solar-wind-driven model, which is, according to Newell et al. [1], indicated for long-term space weather forecasting and indicated in understanding the solar-wind-magnetosphere coupling. Moreover, our approach is using PyTorch Lightning to be easily extensible to all sources of data thanks to the use of a *Dataloader*.

Objectives

Objective: Forecast low-energy particle fluxes in LEO by preparing the implementation of a Machine Learning (ML) algorithm to and improve existing models.

In this poster:
1. Analysis of ACE's solar wind & IMF parameters (instruments SWEPAM and MAG)
2. First results on forecasting using McGranaghan et al. [1] data.

Data

ACE Solar Wind Data: Level-2 Real-Time Solar Wind (RTSW) data from the Advanced Composition Explorer (ACE) from the ASC [3] - 1998 to 2020.

The following 16-seconds MAG data: (45 365 393 data points)
- IMF X, Y and Z-component, GSE [nT]

The following 64-seconds SWEPAM data: (11,299,710 data points)
- Solar wind proton density [p/cc]
- Solar wind bulk speed [km/s]
- Solar wind ion temperature [K]

AI-Ready data: From McGranaghan et al. (2020) [4], we gathered DMSP Particle Precipitation data that are supposed to be AI-ready. We use this dataset to gain some time and be able to perform tests in parallel.

Methodology

Two works done in parallel are presented here:
1. Full analysis of ACE data from the ASC: The purpose is to present a thorough analysis to expose any biases in the data and to explore the possibilities and limits of a future ML-model.
2. Trials on improving results of McGranaghan et al. 2021 [1] using their AI-Ready data.

What has been done in this research (waiting for publication) in data analysis:
- Histograms analysis
- Handling missing and extreme values
- Plotting correlations, autocorrelations
- 2-D distributions analysis
- Doing a Principal Component Analysis

First results for artificial neural networks by:
- Varying input samples' sizes, from 600k/50k to 100k/10k (respect. train/validation set).
- Testing 16-bit versus 32-bit floating point precision

ACE Data Analysis

Histograms

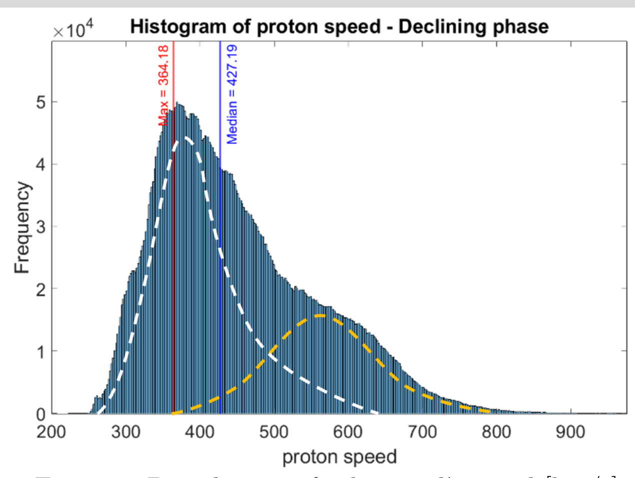

Figure 1. Distribution of solar wind's speed [km/s]

Variables	Mean	Median	0.005th percentile	99.995th percentile	% of missing data
Bx (GSE) [nT]	6,93 x 10^{-2}	8.4 x 10^{-2}	-36.6	25.5	0.128 %
By (GSE) [nT]	2.98 x 10^{-2}	- 9.00 x 10^{-3}	-30.7	38.7	0.128 %
Bz (GSE) [nT]	9.34 x 10^{-3}	2.20 x 10^{-2}	-43.5	32.3	0.128 %
Bt (GSE) [nT]	5.76	5.04	0.32	54.5	0.128 %
Proton Density [p/cc]	5.88	4.54	0.1	80.0	41.59 %
Proton Speed [km/s]	4.30 x 10^{2}	4.08 x 10^{2}	2.38 x 10^{2}	1.03 x 10^{3}	6.80%
Ion Temperature [K]	9.20 x 10^{4}	7.05 x 10^{4}	2.84 x 10^{3}	1.00 x 10^{6}	20.10 %

Table 1. Mean, median, percentage of missing data, 99.995th and 0.005th percentiles for all variables chosen from ACE satellite.

- Distributions are not uniform at all. Rare cases are underrepresented, and algorithms will be unbalanced in their learning. Moreover, gaussian distributions could be overly brought by gaussian noise from measurements and signal-to-noise ratio is difficult to estimate.
- Models will not be reproducible from year to year due to high changes in histograms over time (e.g., figure 1). This high dependence to the solar cycle suggests that one should have at least 11 consecutive years of data to train the algorithm.
- Rounding of ion temperatures values to different orders of magnitude throughout the data, (probably due to a fix number of bits on which measurements are coded) will bias the model by wrongly over-representing or under-representing some values (figure 2).

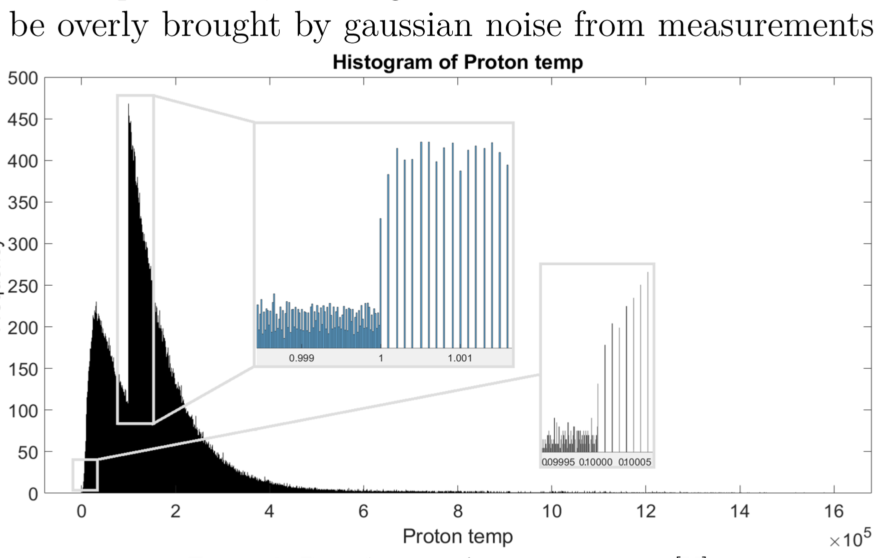

Figure 2. Distribution of ion temperature [K]

Missing values

- There is an extensive number of missing values (e.g., 41.59% in proton density data), that cannot be forgotten and cannot be implemented without processing. The choice of the corresponding processes will be subjective but should consider the distribution of missing values. For instance, large gaps of consecutive missing values cannot be filled in using spline interpolation.

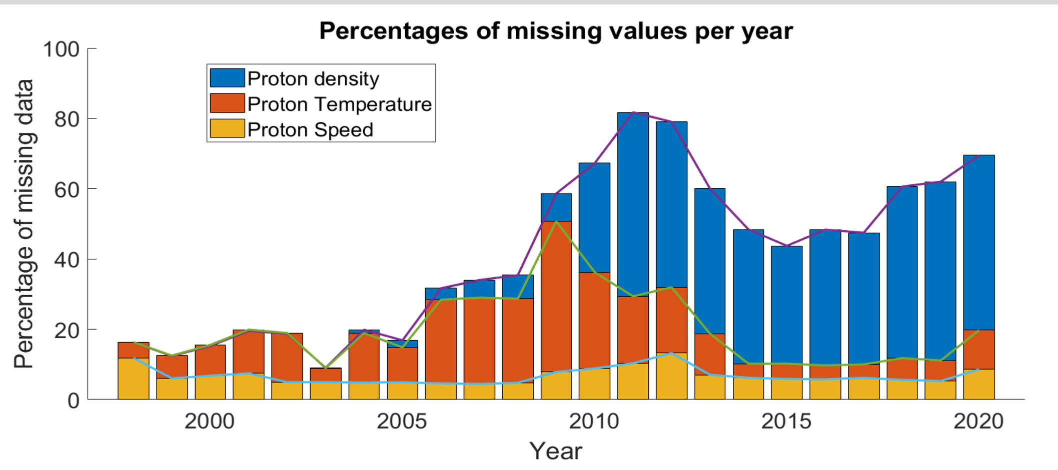

Figure 3. Percentages of missing values for SWEPAM data over the 1998-2020 period.

2-D statistical analysis

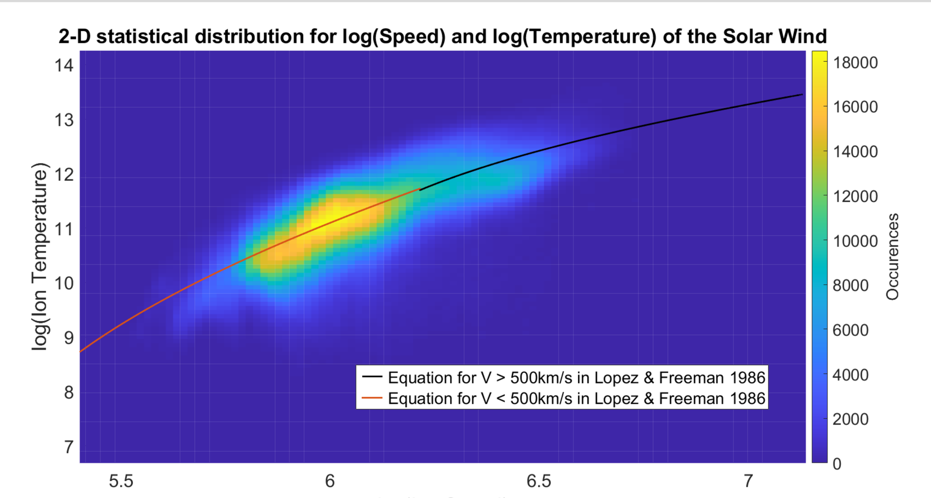

Figure 4. 2-D statistical distribution for log(speed) and log(temperature) of the solar wind.

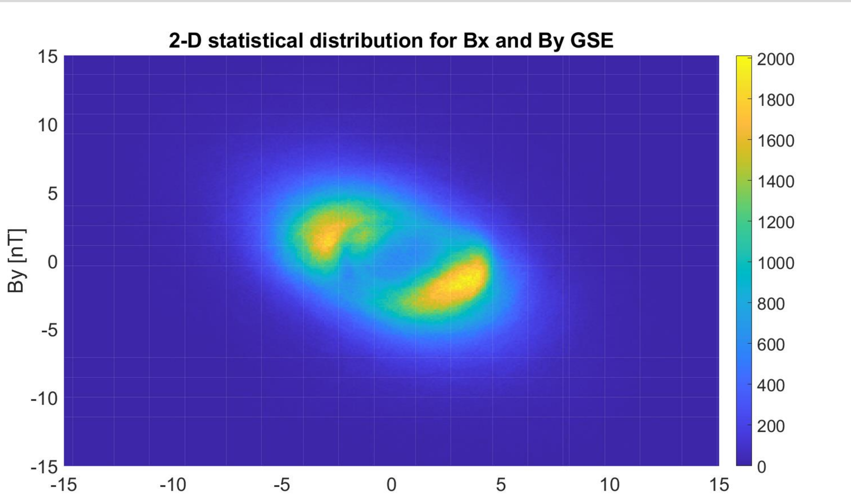

Figure 5. 2-D statistical distribution for the X and Y-components of the IMF [nT]

- Correlation matrix showed a 0.685 correlation coefficient between speed and temperature of the solar wind.
- The two-dimensional statistical distribution for the X and Y-components of the IMF (figure 5) shows the approximate 45° angle between the IMF vector and the radial Sun-Earth direction.

Correlations, Autocorrelations & Principal Component Analysis

- A linear model will not be able to accurately model the data. Our linear analysis (e.g., PCA), struggle to explain the data and their relationships. However, non-linear relationships between data seems to exist.
- Data seem cyclic: apparition of the solar cycle and the synodic rotation period of the Sun when looking at autocorrelations.

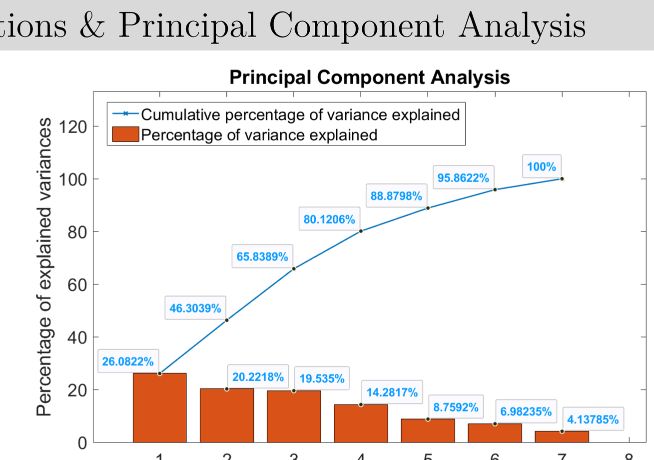

Figure 6. Percentage of explained variance for each principal component from principal component analysis

Deep Learning first results

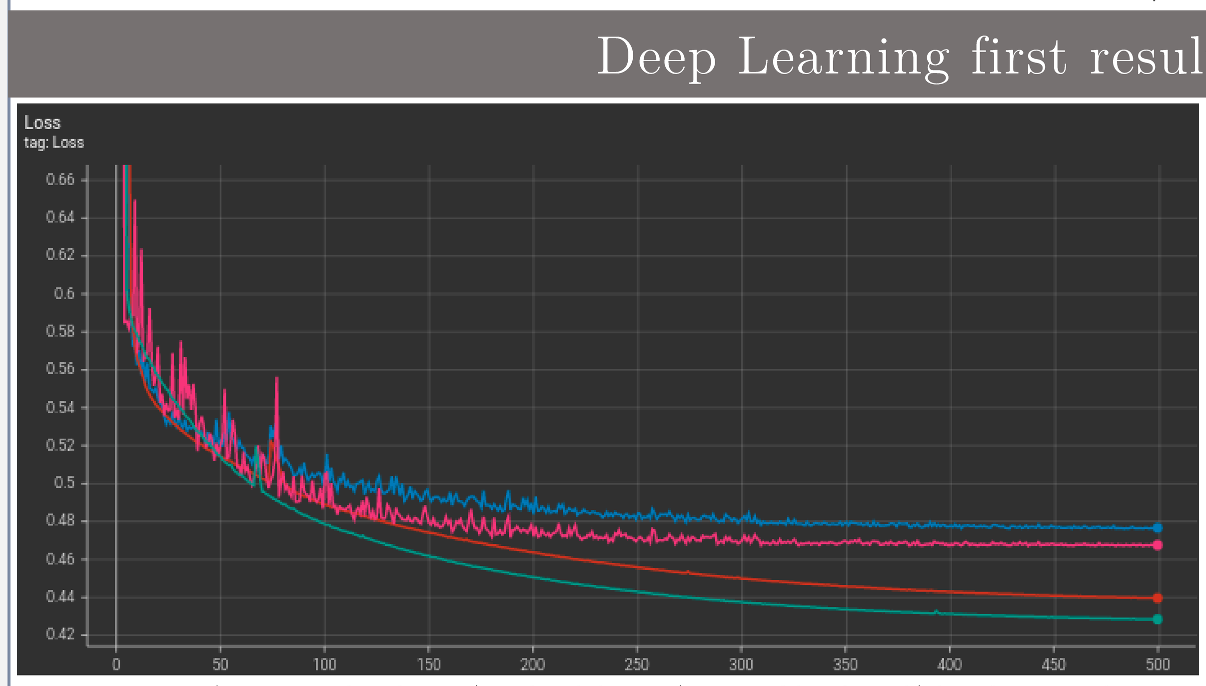

Figure 5. Train (respect. red and green) and Validation (respect. blue & pink) loss functions for two different precisions (respect. 32-bit and 16-bit floating point) with same hyperparameters (right).

Artificial Neural Networks hyper-parameters:
- Batch size: 100
- Learning rate: 0.01
- Epochs: 500
- Hidden layers: 2
- Dimensions: [7000,4]
- No dropout
- No L1/L2 regularization
- Loss function: nn.MSELoss()

Datasets size:
- 600k train samples.
- 50k validation samples.
- 55210 test samples.

Inputs are the data from [4], so-called *AI-Ready data*.
- 72 features: X, Y and Z-component of the IMF, AL, AU, SymH, PC, speed and density of the solar wind for different timestamps (t, t-6hr, t-3hr, t-1hr, t-45mn, t-10mn), $F_{10.7}$ & satellite position ([1][4] for more details).
- 1 label: Logarithm of the electron total energy flux from DMSP satellites multiplied by pi.
- The test set contains only samples corresponding to measurements from DMSP F16 in 2010.

Results

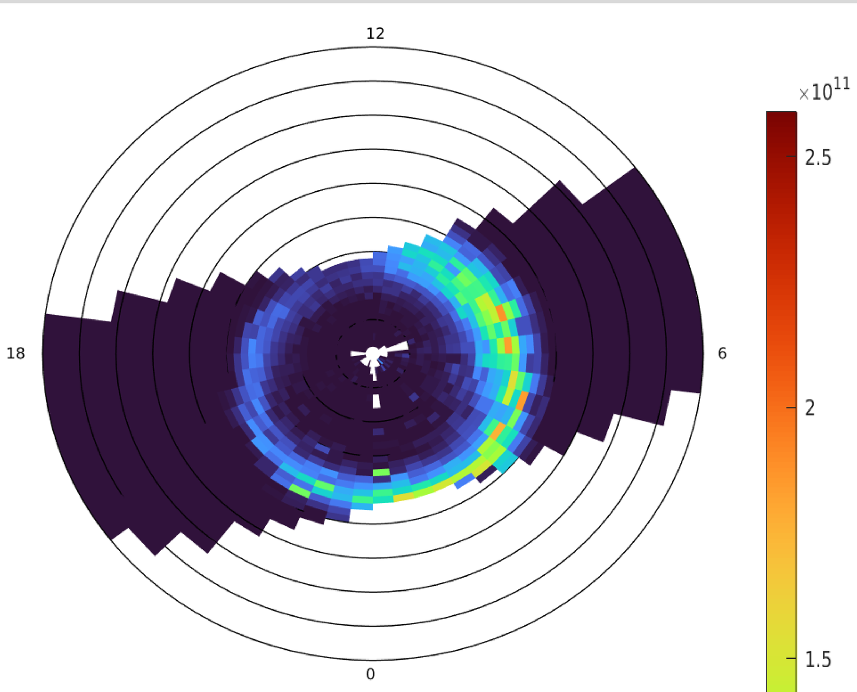

Figure 7. Test data: 2010 measurements from DMSP 16 - MLAT and MLT total electron flux binned median values in AACGM.

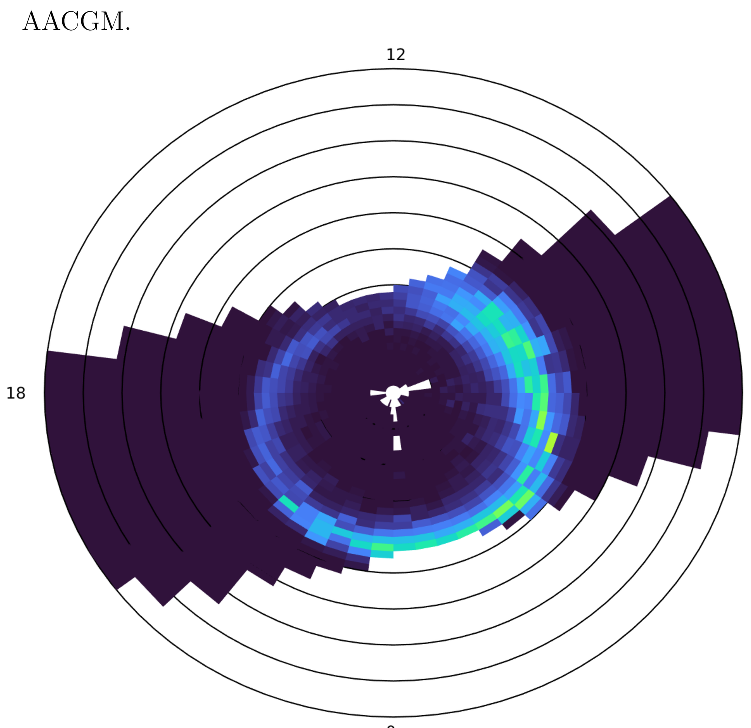

Figure 8. Forecasted value for data shown in figure 6.

- Our ANN is not very reliable on high values / rare cases.
- Our RMSEs for the two cases (16-bit and 32-bit floating points) are respectively 0.6838 and 0.6904, while RMSE of PrecipNet was 0.764 ± 0.011 and RMSE of OVATION was 1.887 [1].
- The values of the validation losses on the 1st epoch were approximately 1.5 but quickly reached approximately 0.6 around the 4th or 5th epoch. This suggests that most of the information contained in the data has been learnt by the ANN by the 10th epoch.
- Last assertion is confirmed by the fact that our results with only 100k train samples and 10k validation samples are the same

Conclusions & Perspectives

- The main conclusion here lies in the input data. Even with a robust scaler, very high-values remain and probably constrain the algorithm from learning properly. Solving this issue will help us identify how much information is contained in the data.
- Another conclusion is that we have a large numerical error as seen in the different results using 16 or 32-bit floating points.
- It is odd that 100k-train samples and 600k-train samples have the exact same results, although the same magnitude was expected. This will be investigated.
- Several tests remain: batch, seed, dropout, MAE, learning rate, sampler and optimizer.
- In the long run, one idea will be to build two 2-block neural networks. One 2-block will be made of one neural network handling rare high-values data and one neural network evaluating the quality of the corresponding prediction. The other 2-block will be the same but with low-values.

References

[1] McGranaghan, R. M., Ziegler, J., Bloch, T., Hatch, S., Camporeale, E., Lynch, K., ... & Skone, S. (2021). Toward a next generation particle precipitation model: Mesoscale prediction through machine learning (a case study and framework for progress). *Space Weather*, e2020SW002684.

[2] Newell, P. T., Liou, K., Zhang, Y., Sotirelis, T., Paxton, L. J., & Mitchell, E. J. (2014). OVATION Prime-2013: Extension of auroral precipitation model to higher disturbance levels. *Space Weather, 12* (6), 368-379.

[3] srl.caltech.edu/ACE/ASC/level2/

[4] Ryan M. McGranaghan, Téo Bloch, Jack Ziegler, Spencer Hatch, Enrico Camporeale, Mathew Owens, ... Susan Skone. (2020). DMSP Particle Precipitation AI-ready Data (Version 1.0.0-alpha) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.4281122

Acknowledgments

I would like to acknowledge the support of my dear colleagues and friends at SpaceAble, as well as Dr. Mathieu Barthélémy and Dr. Jocelyn Chanussot, my supervisors who actively support me and inspire me during this PhD. I would also like to thank the Université Grenobles-Alpes, and my laboratories Gipsa-LAB & IPAG in Grenoble.