Application of information theoretical measures for improved machine learning modelling of the electron radiation belt

C. Papadimitriou[1,2], G. Balasis[2], S. Wing[3], I. A. Daglis[1,4]

(1) Department of Physics, National and Kapodistrian University of Athens, Athens, Greece (2) IAASARS, National Observatory of Athens, Athens, Greece
(3) The Johns Hopkins University's Applied Physics Laboratory, Maryland, U.S.A. (4) Hellenic Space Center, Athens, Greece

Abstract

In the past ten years Artificial Neural Networks (ANN) and other machine learning methods have been used in a wide range of models and predictive systems, to capture and even predict the onset and evolution of various types of phenomena. These applications typically require large datasets, composed of many variables and parameters, the number of which can often make the analysis cumbersome and prohibitively time consuming, especially when the interplay of all these parameters is taken into consideration. Thankfully, Information-Theoretical measures can be used to not only reduce the dimensionality of the input space of such a system, but also improve its efficiency. In this work, we present such a case, where differential electron fluxes from the Magnetic Electron Ion Spectrometer (MagEIS) on board the Van Allen Probes satellites are modelled by a simple ANN, using solar wind parameters and geomagnetic activity indices as inputs, and illustrate how the proper use of Information Theory measures can improve the efficiency of the model by minimizing the number of input parameters and shifting them with respect to time, to their proper time-lagged versions.

Datasets

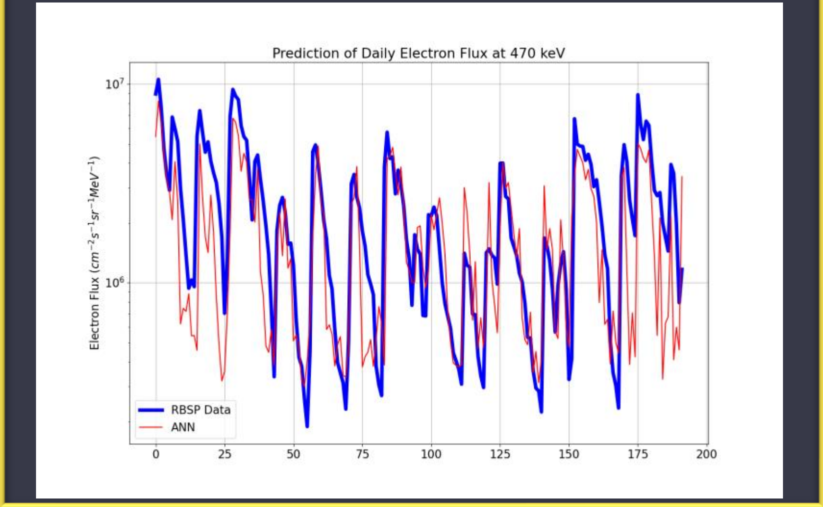
Radiation Belt Data were derived from the MagEIS instrument on board the RBSP-B satellite, from the beginning of 2014 up to the middle of 2019. Data are omnidirectional, differential electron fluxes ($s^{-1}cm^{-2}sr^{-1}MeV^{-1}$). For this demonstration, the dataset was limited to one energy, at 417 keV, L* values from 4.5 to 5 and equatorial pitch angle from 80 to 90 degrees, so these are electrons that remain close to the magnetic equator, near the peak of the outer radiation belt. Finally, to avoid short-term fluctuations, fluxes have been reduced to their daily-averaged values.

Solar Wind Data were retrieved from NASA's OmniWeb and they are daily-averaged parameters of the magnetic field B_total, as well as its components B_X, B_Y, B_Z (all in GSM coordinate system), Solar Wind Proton Density (N), Solar Wind Velocity (V) and Pressure (P).

Geomagnetic Indices Data are comprised of daily averaged values of the typically used Geomagnetic Activity Indices, Kp, Dst, ap, AE, AL, AU and pc (polar cap).

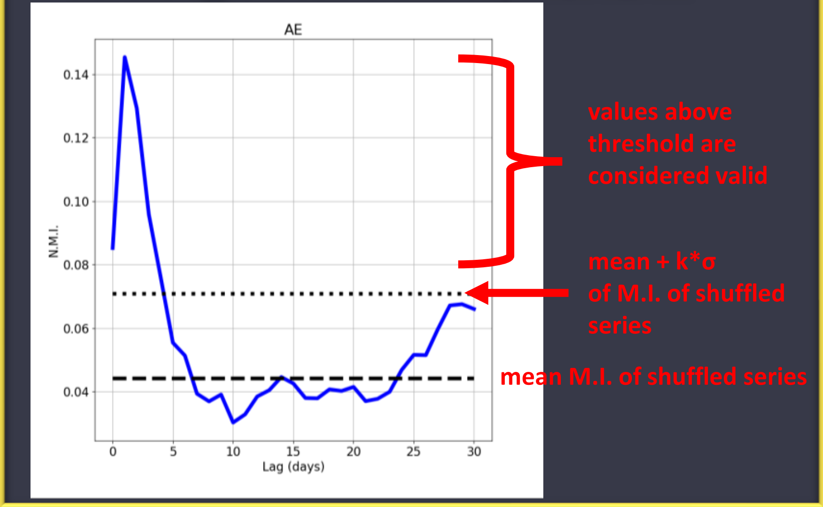
ANN Model

1. Train a simple ANN on the data to predict daily electron fluxes
   a) standard scaling of inputs (mean/std)
   b) 2x100 neurons for the hidden layer (RELU)
   c) 80/20 train/test partition



Information-Theoretic Method

1. Calculate the Mutual Information between each input parameter and its time-lagged versions
2. Use 100 shuffled versions to estimate the threshold of randomness
3. Keep only parameters with M.I. above the threshold and keep only their time-lagged versions for which they exhibit max M.I.



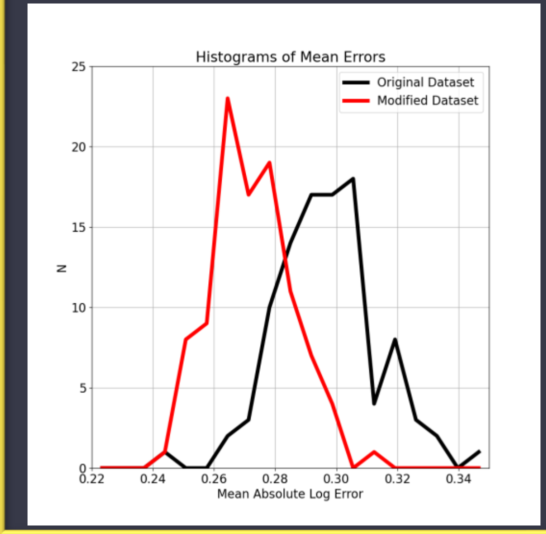
values above threshold are considered valid

mean + k*σ of M.I. of shuffled series

mean M.I. of shuffled series

Final Result

1. Re-train the ANN using the new, reduced dataset and for random 80/20 partitions of the data calculate the errors on the test dataset
   a) Error = Mean[Abs (Log(data) – Log(ANN))]



Error with Original Dataset: 0.30

Error with Modified & Reduced Dataset: 0.27

10% Improvement in score, keeping only 4 out of 14 initial parameters!