Take Home Instructions

Objective

The objective of this assignment is to showcase your ability to problem solve and construct a prototypical solution for a potentially unfamiliar problem. We will provide data for this task; however, the core implementations and solutions are up to you.

Background

Machine learning (ML) has become the state-of-the-art approach for predicting thermospheric density due to its ability to capture complex, nonlinear interactions between diverse drivers of thermospheric dynamics that are difficult to model using traditional physics or empirical approaches alone.

The thermosphere's behavior is influenced by many factors, such as solar radiation, geomagnetic activity, and lower atmospheric waves, whose effects are highly coupled and nonlinear. ML models handle these complexities by learning patterns directly from large volumes of observational data, often integrating inputs from multiple sources like satellites, ground-based sensors, and historical records. This data-driven approach allows ML models to more accurately represent the thermosphere's response to a wide range of conditions, including extreme events like geomagnetic storms, where physics-based models typically falter due to reliance on simplified assumptions or linear approximations. These models also reduce uncertainties inherent in empirical approaches by optimizing model parameters through data-driven estimation rather than relying on empirical formulas or heuristic tuning. By leveraging these advantages, ML-based models have set a new benchmark in thermospheric density prediction, significantly outperforming traditional models and enabling better forecasting of space weather effects on satellite operations and space missions.

Materials

You have been provided with the following data in the shared google folder:

1. Space weather indices for 2002 to 2007 - *Kp_ap_Ap_SN_F10_2002_2007.txt*
    a. See here for more information on the file format and contents
2. Geomagnetic index for 2002 to 2007 - *DST_2002_2007.txt*
    a. See here for more information
3. Solar indices - *SOLFSMY.txt*
    a. See here for more information

4. CHAMP density measurements for 2002 to 2007 - *champ-2002-2007.tar.gz*
5. High resolution OMNI data for 2002 to 2007 - *omni.zip*
    a. See here for more information on the file format and contents
    b. Each *.fmt* file provides metadata about the lst file format
    c. Each *.lst* file contains the actual data itself

Tasks

1. **Feature Selection**
    a. Using the provided data sources, devise data cleaning and preprocessing steps to be used when performing Task 2.
    b. Identify any inconsistencies in temporal resolution between the data sources and create an alignment strategy.

2. **Model Training**
    a. Create a model that is capable of predicting neutral density given a `(year, month, day, hour, minute, second, latitude, longitude, local_solar_time, altitude)` as input.
        i. The model ***must also*** be capable of accepting additional inputs from the provided data sources (e.g. solar indices, geomagnetic indices and so on). Feel free to use any combination of additional inputs that you see fit.
        ii. The model should be optimized using ground truth density values taken from CHAMP measurements.
        iii. Some useful resources to get you started
            1. Paper 1
            2. Paper 2
            3. Paper 3
            4. Paper 4
            5. Paper 5

3. **Analysis of Results**
    a. For the period of October 28-31, 2003, generate density values at a 1 hour cadence on the grid
        i. Latitude = -90:90:10 (min:max:stepsize)
        ii. Longitude = -180:180:10
        iii. Altitude = [350, 450, 550, 650] km
    b. Plot 2D contour plots of predicted density values (for each altitude and timestep) where Longitude is your X-dimension and Latitude is your Y-dimension.
    c. Plot 2D contour plots of predicted density values (for each altitude and timestep) where Local Solar Time is your X-dimension and Latitude is your Y-dimension.
    d. Comment on the different structures you see *over time* in parts **b** and **c** and postulate why they arise.

Tip: If you maintain a google account, you can leverage Google Colab to *train the model*. It's free (with some limitations) and allows you to utilize powerful compute nodes.

Stretch Goal

1. **Model Improvement**
    a. Identify ***additional and relevant*** sources of data that may improve your model predictions and incorporate them into your model.
    b. Analyze whether your new data sources have improved model performance during G3 and G4 geomagnetic storms.

Deliverables

While you are free to send us your solutions in whatever format that is convenient, we recommend the following:

A. A single (or many) Jupyter notebook(s)
B. A collection of Python files + PDF summary of your results
C. Some combination of the two options
    a. E.g. Write the required functionality in Python modules and train the model on Google collab (in a notebook) using those modules.

The PDF/notebook should include:

1. A brief comment on why you chose your particular model architecture.
2. A brief comment on which additional space weather inputs you chose for your model and why you chose them.
3. Comments/notes regarding any engineering challenges you encountered during the implementation of the model and how you addressed them.
4. Plots of relevant metrics collected during model training (loss values, mean squared error, etc.)

Timeline and Next Steps

We recommend that you wrap this up within 7 days. If you are unable to finish all the tasks, please send us whatever you have completed. We aim to understand how you make design decisions and your thought process.

Once you’re done, respond to our email with the notebooks/code/PDF’s and we’ll get back to you after our team has had a chance to review your work.