Contributing to Geoscientific Python Through MetPy and Project Pythia

Ana Castaneda Montoya

Drew Camron, Thomas Martin, Ryan May

Summer Internship Program, NSF Unidata

July 31, 2024

This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

Introduction

During my internship in the summer of 2024 I had the privilege of contributing to some of NSF Unidata's Python projects.

PROJECT PYTHIA

- Pythia Foundations
- Hackathon
- Machine Learning Cookbook

MetPy

- New function for splitting xarray datasets for ML
- WPC Surface Bulletins in the Southern Hemisphere
- New PlotSurfaceAnalysis class

unidata

This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

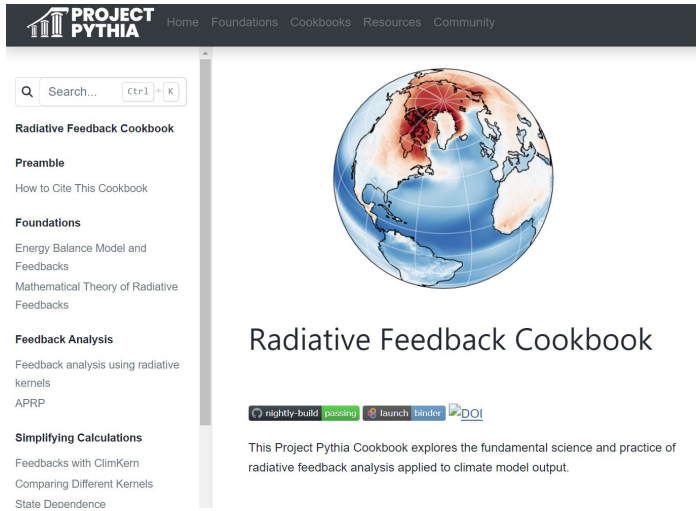
Project Pythia

Project Pythia is a Python-centered educational and training resource for the geoscience community.

- My involvement with Project Pythia started in Pythia Foundations

- Then, I participated in the Project Pythia hackathon

Cookbook:



This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

Machine Learning Cookbook

- Introduction to running machine learning algorithms and complements NSF Unidata's CyberTraining project.

- Presents a basic regression problem with Linear Regression and Decision Tree Regressor models.

- It introduces learner to the scikit-learn API through the following sections:
    - Exploratory Data Analysis
    - Dataset Splitting
    - Dataset Scaling
    - Training and Evaluation of Model

This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

PROJECT PYTHIA Home Foundations Cookbooks Resources Community

Search... Ctrl + K

Precipitation Machine Learning Cookbook

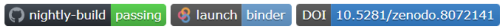
Preamble

How to Cite This Cookbook

Introduction

Regression ML Analysis - Atmospheric Data Quick Start

Precipitation Machine Learning Cookbook



This Project Pythia Cookbook covers an extremely basic precipitation classification project. This notebook will introduce learners to the scikit-learn API, basic exploratory data analysis (EDA), and evaluations. It is meant to be a very early and basic introduction to these concepts, it is not meant to be an in-depth intorduction to machine learning. It could be the first introduction to machine learning for learners familiar with weather data.

Regression ML Analysis - Atmospheric Data Quick Start

Overview

In this cookbook we will go through a simple example showing the Supervised Machine Learning Regression Framework, using the scikit-learn ecosystem.

1. Reading Data and Exploratory Data Analysis
2. Splitting Dataset and Scaling Data
3. Training, Testing, and Validating Model

This is a MVP (or minimum viable product) for ML modeling analysis. There are many more things we could test and add to this. The dataset itself is also very small for fast loading and general speed. This CookBook is meant to be a companion to Unidata's Cybertraining project.

Cookbook:

This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

MetPy

MetPy is an open-source collection of tools in Python for reading, visualizing, and performing calculations with weather data.

First MetPy task: New `train_test_split()` function for xarray

I wrote a new function to create training/testing datasets. Unlike its scikit-learn counterpart, this function:

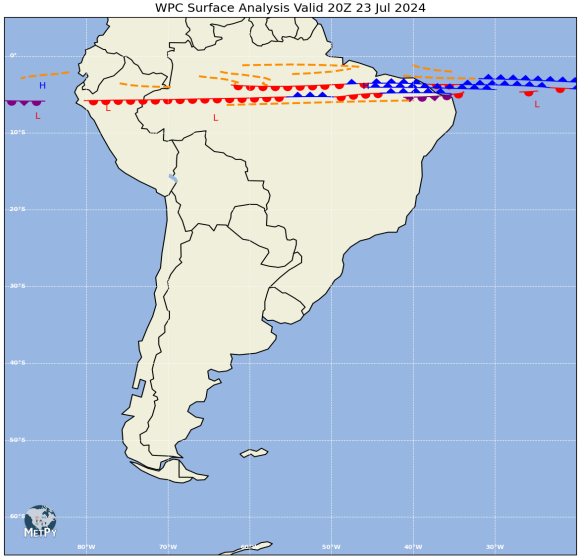
- Splits the data using the time variable. Specially useful for Climate and Weather data where we deal with time series.
- Creates an optional validation dataset
- Handles xarray datasets

This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

WPC Surface Bulletin for Southern Hemisphere Issue

- MetPy's `parse_wpc_surface_bulletin()` function is used to parse the Weather Prediction Center's (WPC) Surface Bulletins into a Pandas Dataframe

- GitHub issue #3535 pointed out that bulletins were being parsed incorrectly for the Southern Hemisphere

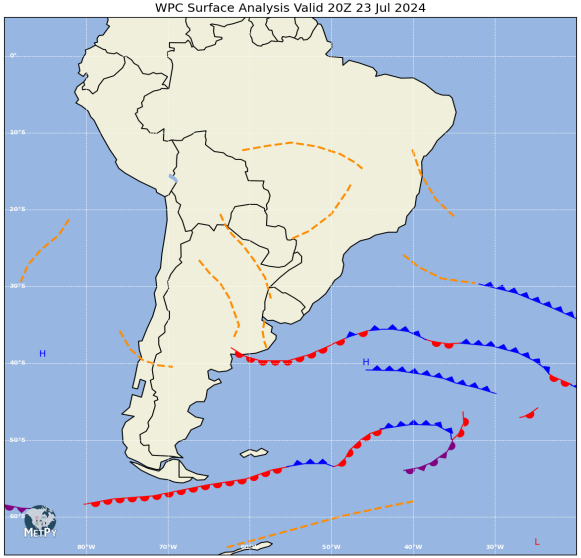


Before

This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

WPC Surface Bulletin for Southern Hemisphere Issue

- My modifications to the function resolved the issue.
- I also added new tests to ensure its correct performance.



After

This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

Plotting WPC Surface Bulletins Made Easy

- After the WPC Surface Bulletins are parsed, we can plot Surface Weather Maps
- I wrote and implemented a new class called `PlotSurfaceAnalysis()` that simplifies the process to a few lines of code
- The plots are fully customizable. Customizable traits include: colors, labels, label fontsize, markersize, linewidths, linestyles, and label offset.

This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

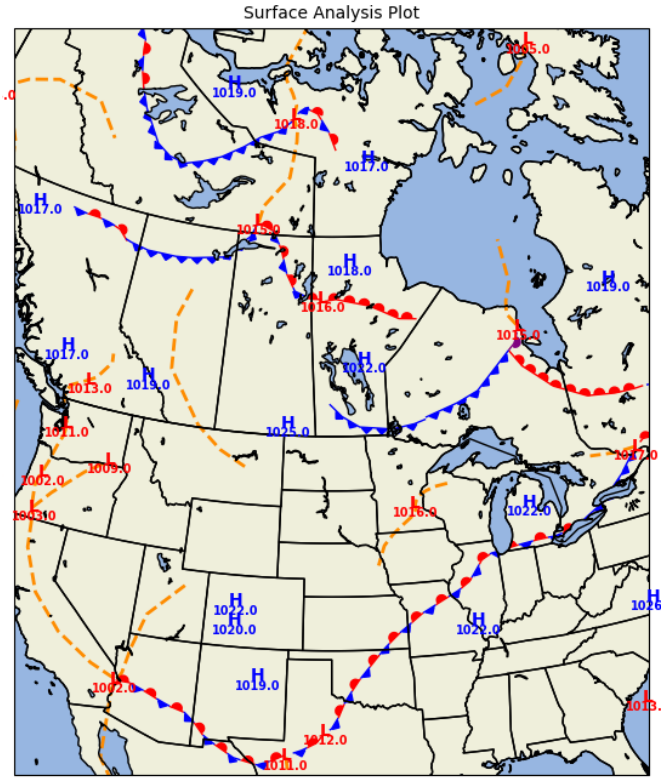
Plotting WPC Surface Bulletins Made Easy

```
df = parse_wpc_surface_bulletin('WPC_file.txt')
```

```
ps = PlotSurfaceAnalysis()
ps.geometry = df['geometry']
ps.feature = df['feature']
ps.strength = df['strength']
```

```
panel = MapPanel()
panel.area = [-120, -80, 30, 70]
panel.projection = 'lcc'
panel.layers = ['lakes', 'land', 'ocean',
                'states', 'coastline', 'borders']
panel.plots = [ps]

pc = PanelContainer()
pc.size = (12,8)
pc.panels = [panel]
pc.show()
```

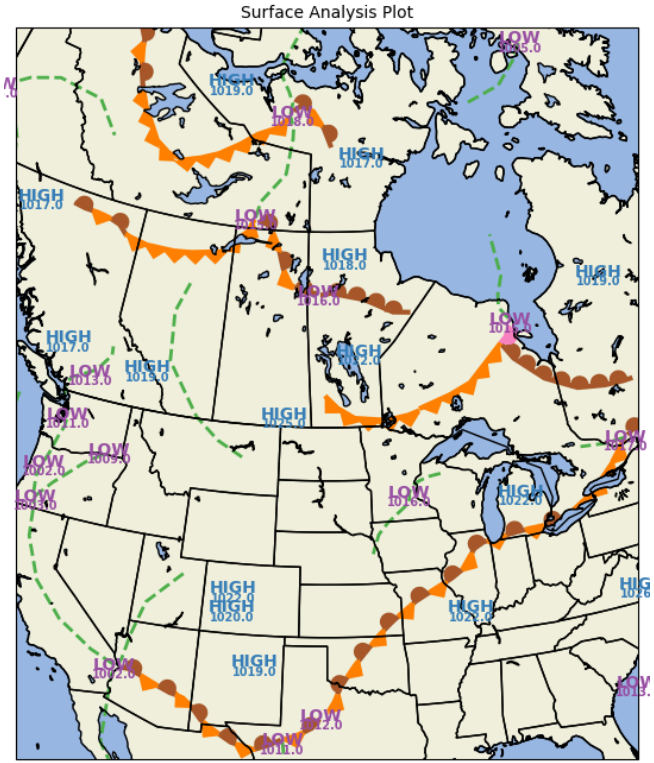


This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

Plotting WPC Surface Bulletins Made Easy

The plot to the right shows the same map as the previous slide with some customization: colors, labels, marker size and linewidth of fronts.

Complete plotting example:



This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

In Conclusion, Open-Source Projects…

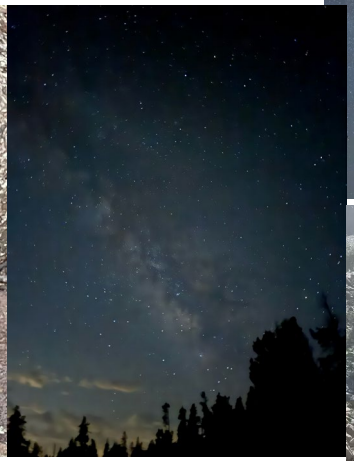
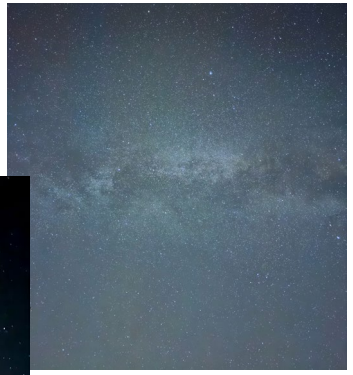
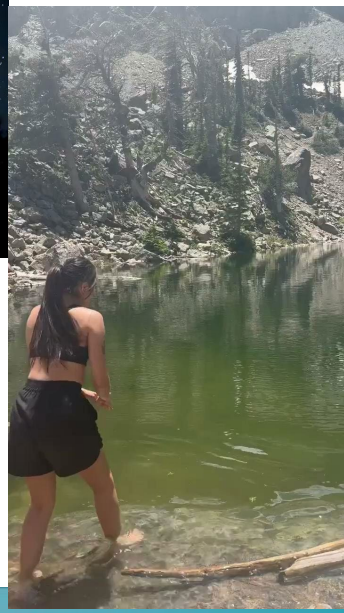
- Are incredibly valuable to the scientific community
- Are easier to contribute to than it seems…
- Have many different ways to get involved

This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

Some pictures!





This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.

Thank you!

NSF Unidata is one of the University Corporation for Atmospheric Research (UCAR)'s Community Programs (UCP), and is funded primarily by the U. S. National Science Foundation (Grant AGS-2403649).

This material is based upon work supported by NSF Unidata under award #2403649 from the U. S. National Science Foundation. Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of the NSF.