<u>Project Portfolio</u>

Jonathan Brunton
Email: jahbrunton@gmail.com , **Phone:** +1(309)634-5935

Atmospheric Data Scientist - *Oct. 2023-Present*
Environmental Protection Agency and Oak Ridge Associated Universities
(National Student Services Contract)
Model Evaluation & Outreach/Community Multiscale Air Quality (CMAQ) Model Dev. Team

MSc Environmental Data Science & Machine Learning - *Nov. 2022*
Imperial College London
Notable Awards: Merit Graduate (Reserved for Top 25% of class)

BS Planetary Sciences - *May 2021*
Purdue University (3.70 GPA)
Notable Awards: Presidential Scholarship, Paul & Linda Krishna Scholarship in Earth, Atmospheric, and Planetary Sciences, 6x Dean's List Placement
Extracurricular Activities:
Undergraduate Teaching Assistant (EAPS 102: Earth Science for Elementary Teachers), College of Science Ambassador, Gimlet Leadership Honorary (Secretary), Planetary Science Society of Purdue, Purdue Astronomy Club, Wesley Foundation Service

Portfolio Contents

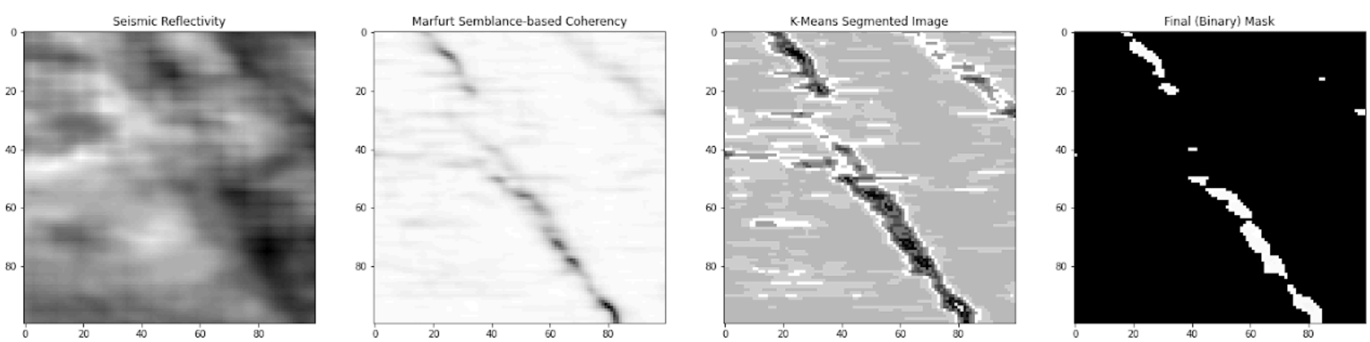
I. Thesis Project

Motivation: Traditional workflows for assessing a subsurface region's ability to store fluid require ample manual interpretation and destructive collection processes in addition to advanced seismic modeling. With next-generation velocity data recovered from Full-Waveform Inversion techniques of a subsurface region, we aim to automate the entirety of this workflow for fault systems specifically by:

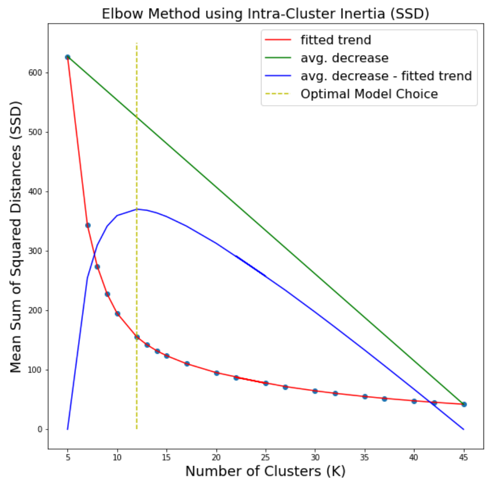
- Extracting meaningful data features for a large, raw seismic reflectivity volume by leveraging advanced numerical methods
- Developing an unsupervised image clustering model to predict the locations of faults within the volume
- Combining and using the predicted volume with velocity data to create a 3D visualization tool and assess predicted fault's ability to store fluid.

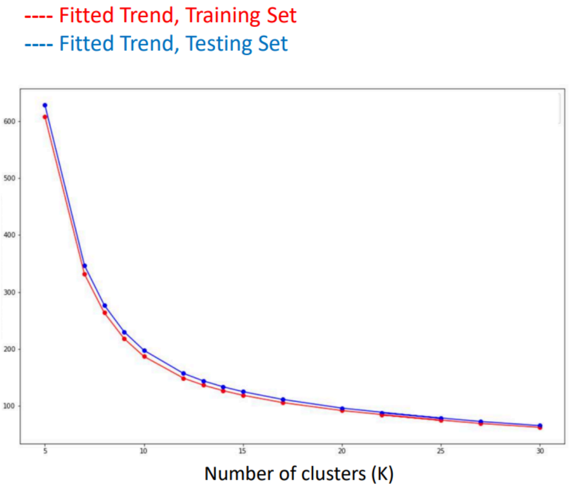
<u>i. Machine Learning Model Development</u>

- Initial steps were taken to clean and transform raw seismic reflectivity image data into a numerically derived coherency volume using 3D interpretation. 2D, 'stackable' images of these transformations were sent to be trained on a K-Means clustering model, which was developed to highlight large, 3D discontinuities within the volume. Only data with viable intra-cluster inertia values were selected for the final predicted fault locations. This process is shown below, left-to-right.



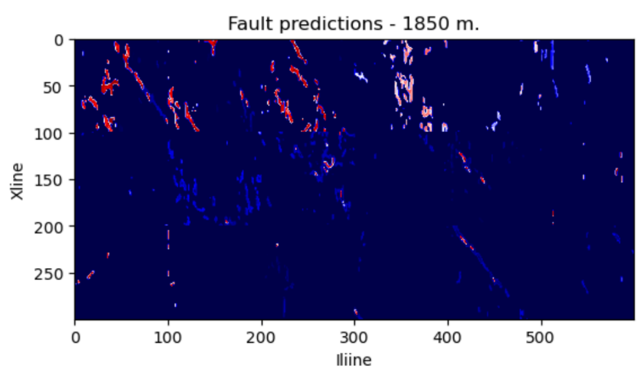
- The clustering model was optimized using a numerical interpretation of the 'Elbow Method' with both average sum of squared distances (SSD) and explained variance as evaluation metrics. Shown below is the numerical interpretation that found the optimal number of image clusters for segmentation in this study, with SSD as an evaluator. Additionally, another elbow graph, with SSD as an evaluator, shows that the fitted trends of the training and testing sets are nearly identical, validating the results of the model from model bias or overfitting.

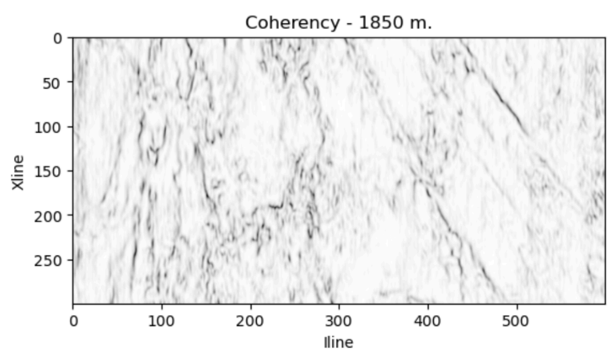




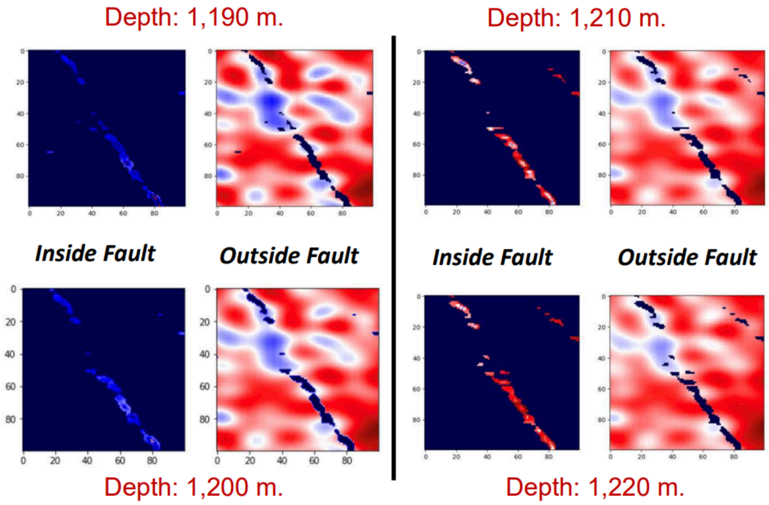
ii. Screening Tool (Visualization & Proposed Research Workflow)

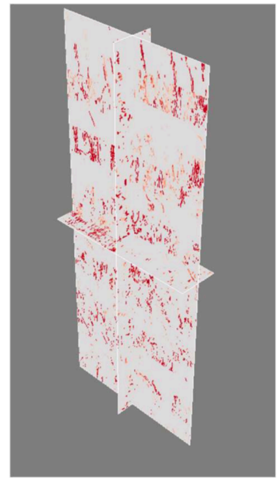
- As briefly aforementioned, this is a novel approach in combining the information from a predicted fault volume with FWI velocity data into a single, interpretable volume. This provides an initial assessment to whether an extracted fault system can store fluid, by highlighting low- or high- velocity anomalies within the image of the system. The relative velocities of some predicted fault locations are shown in the colored image below, while the coherency volume it was derived from is shown in greyscale..





- A major advantage of using a binary mask for predicted fault locations is uncovering the relative velocities from both within and outside of the faults, shown on the left. In addition to the data preprocessing, model training, and prediction workflow contained inside a jupyter notebook, a 3D visualization of the volume's predicted faults/velocities was developed for completing the automation of the manual interpretation of these seismic volumes.





- The complete thesis report including a discussion of results may be found at: ['Developing Screening Tools to Assess CO2 Storage' - Github PDF]

A brief note on results

If afforded more time, there would have been a few avenues to strengthen the power of this screening tool, including:

- An investigation into what other seismic features (numerical interpretations of the raw reflectivity volume) may reveal in their ability to better distinguish faults from the surrounding region and other 3D seismic discontinuities.
- Developing a supervised image segmentation model using labeled faults.
- Optimizing model performance with updated programming methods and migrating model training to high-performance computing environment.
- Developing an ecosystem to host a web-version of the 3D, interactive screening tool with better visualization libraries.

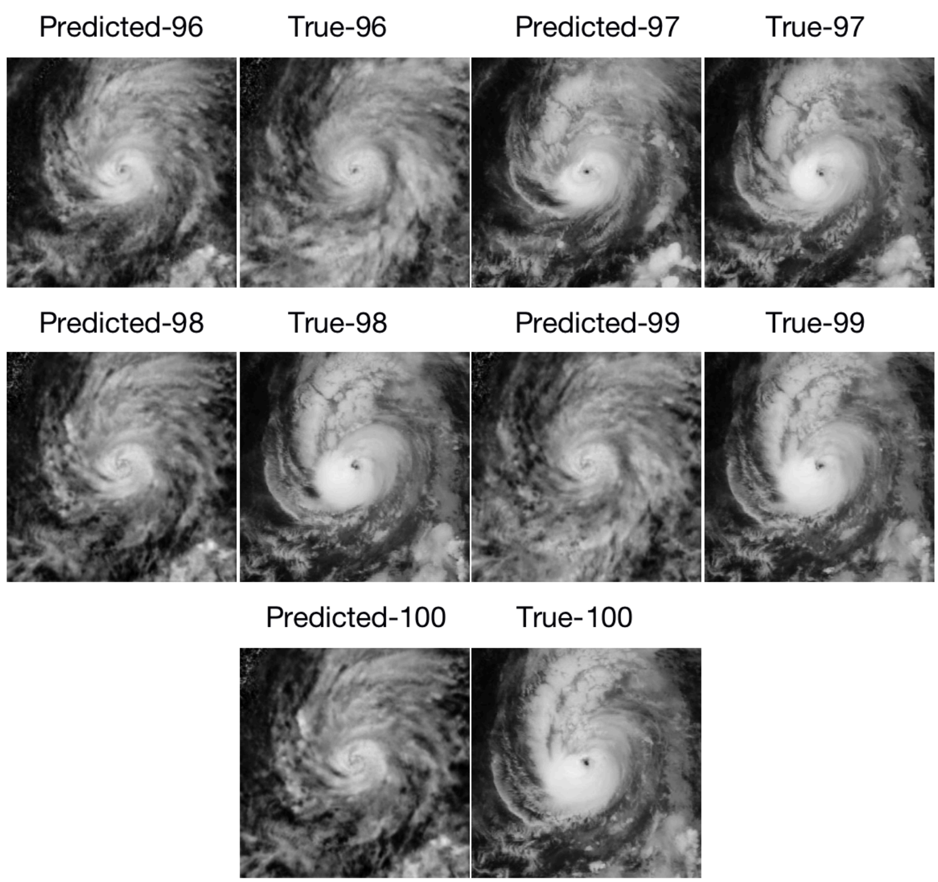
An inactive version of Screening Tool repository and report HERE.

II. Major Master's Projects

*A series of sprint-esque, multi-week project in which 4-7 inter-academic course graduate co-developers develop machine learning applications of environmental datasets, culminating in the 'product release' and presentation to hypothetical clients and interest groups***.**

<u>i. Hurricane Forecasting with Computer Vision</u>

- Motivation: Given a set of satellite images of a hurricane, predict the hurricane's evolution.
- Proposed Solution: A convolutional LSTM and GAN-LSTM were separately trained and tested, both constructed in PyTorch. The former achieved better results in both image prediction (according to a customly developed evaluation metric specific to image pixel similarity) and computational cost. Final results are shown below.

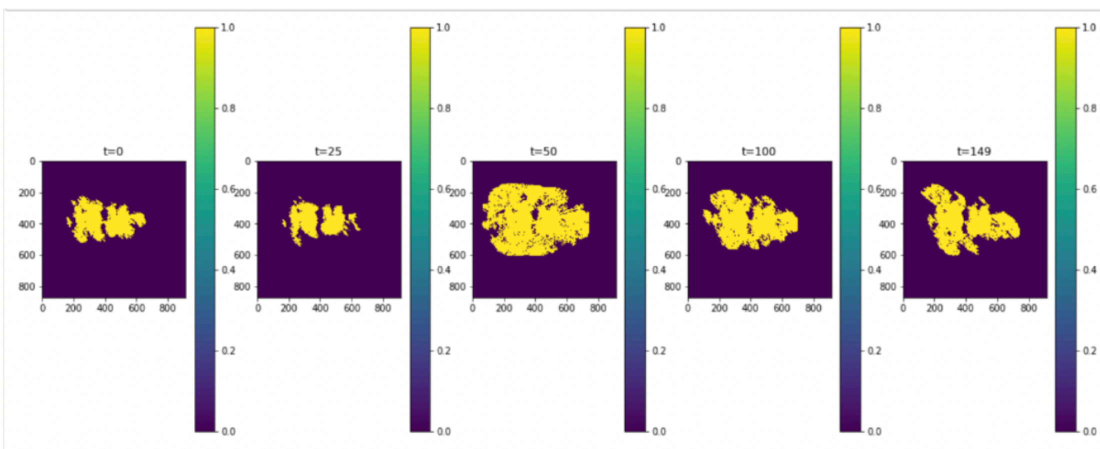


- Project Roles: Lead developer of image preprocessing workflow, co-developer of convolutional LSTM workflow (architecture optimization), documentation and code sustainability maintainer, lead presenter

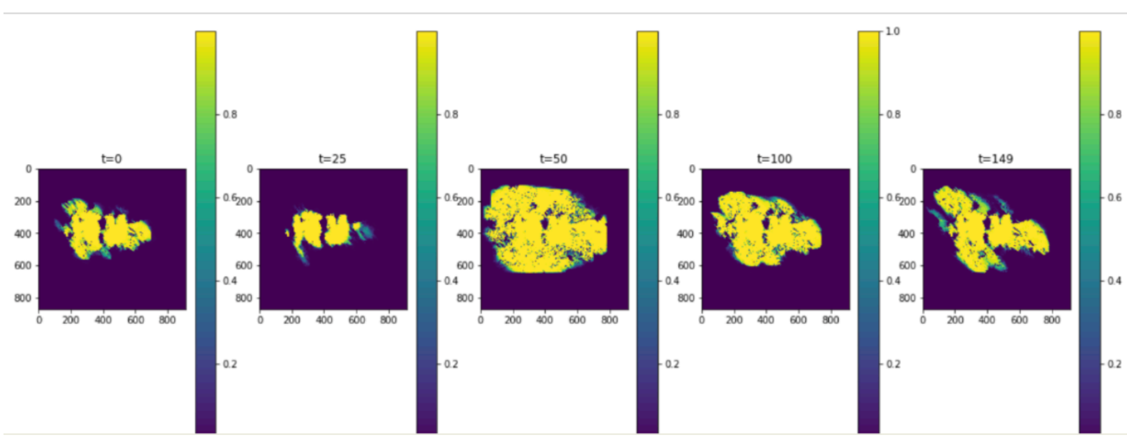
ii. Wildfire forecasting and correction

- Motivation: Using time series of satellite imagery coupled with ground sensor data, develop a neural network to predict the evolution of the wildfire in simulated ‘real-time’.
- Proposed Solution: Principal Components Analysis (PCA) addressed issues of compressing and coupling the data, while an ensemble convolutional autoencoder-random forest regressor predicted wildfire image evolution in physical space. A Kalman filter was developed for the latent space of the autoencoder, correcting predictions in real-time. The images below show a comparison of the wildfire sensor ground truths and model predictions.
- Results:

Daily Ground Truths



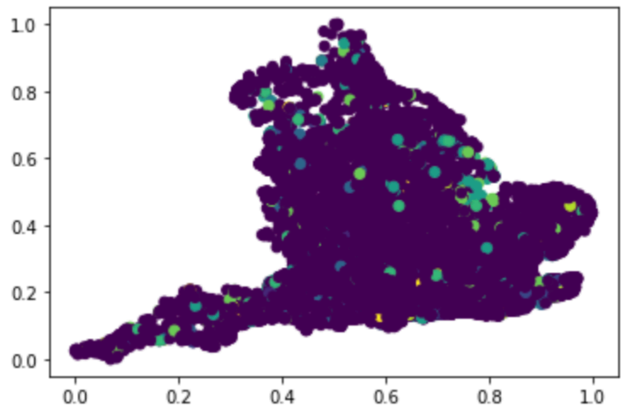
Daily Final Predictions

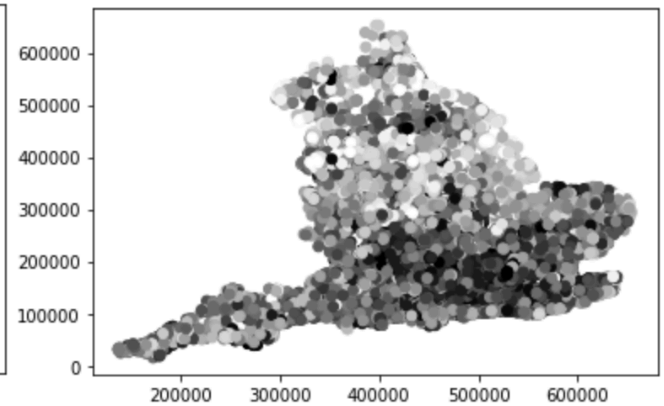


- Project Roles: Lead PCA developer, Co-developer of data assimilation workflow (filter tuning), Python project suite testing, lead presenter

iii. Flood evolution and cost-of-damage prediction

- Problem: Given historical flood data, local rainwater data, and property value estimates, classify the flood probability and predict cost of damages for every postcode in England.
- Proposed Solution: Given the limited amount of data, leverage scientific principles to create new data features from the existing set. The cleaned, manipulated, and integrated dataset is sent to a multi-class model for flood probability. Each of the ten classes outputted by the model is the last input needed for a K-Nearest Neighbors regressor that predicts the corresponding damage cost. Screenshots of results from the classification model are shown in color, with damage cost in greyscale.





- Project Roles: Lead developer of KNN regressor, lead data preprocessing developer, dataset feature creation, documentation and code sustainability maintainer, lead presenter

III. Skills

- Software Development:
  - Python:
    - Expert Numerical processing and applications, data visualization libraries
      - Numpy, Pandas, Scipy, Matplotlib (Pyplot), Plotly, Seaborn
    - Fluent Sci-Kit Learn; basic-to-intermediate machine learning applications
    - Advanced Deep-learning libraries
      - Pytorch, Keras (Tensorflow), OpenCV
    - Advanced Sphinx, Sphinx-Design; for automated HTML deployment of repository code and text (markdown, restructured-text) documentation
    - Advanced PyTest, for automated and customized code testing
    - Experienced Specialized libraries for the environmental data manipulation
      - Geopandas, Xarray, NetCDF4, Segyio, GDAL, Shapely, Rasterio
  - Familiar Trained academically, practiced in research/employment occasionally
    - R (RStudio), C++
  - Experienced/Practicing
    - SQL, primarily through Python's interfacing library mySQL
- Programming Environments
  - Github (Experienced):
    - git, Github CLI
    - Github Actions for basic continuous integration practices and automated documentation creation
    - Github Pages for hosting documentation websites directly from repository
    - Contributor to the Community Multiscale Air Quality (CMAQ) model repository, familiar with integration and repository maintaining best-practices.
  - Jupyter Lab preferred for Python and general software development, familiar with VisualStudio
  - Proficient with command-line programming and shell scripting, and ensuring sustainability across operating systems
    - MacOS, Windows, Linux; Bash, C-Shell
  - Comfortable in virtual and High-Performance Computing environments
    - Most experienced in private, familiar with AWS ecosystem
- Remote Sensing/Environmental Data
  - General Software/Geographic Information Systems (GIS) Experience
    - ERMapper, Petrel, ArcGIS, IDL, ENVI, MASTER, CRISM, Google Earth Engine

- Comfortable in exploiting satellite imagery and various environmental sensor datasets (RADAR, LIDAR, Spectral) for identification of soil, water, greenery, etc. image data.
    - Advanced Model Output Experience:
        - EPA's Community Multiscale Air Quality (CMAQ) Model, NCAR's Weather Research & Forecasting (WRF) Model for air quality and meteorology evaluation, respectively.
    - Advanced Dataset Knowledge:
        - NetCDFs (.nc), SEGY Seismic Volumes (.sgy)
- Machine Learning
    - Familiar and comfortable with using a variety of machine learning and deep-learning algorithms to develop applications for data within and beyond environmental sciences. Experience with regression/classfication and supervised/unsupervised architectures.
        - Linear, Logistic, Forest, CNNs, Convolutional Autoencoders, GANs, LSTMs, FFNs
- Advanced Dataset Knowledge
    - NetCDFs (.nc), Segy Seismic Volumes (.sgy)
- Skilled presentation and instruction of work, scientific principles to large groups (seminars, client groups) and small.
- Interdisplinarian scientific programmer, degreed or employed in each of the fields earth, atmospheric, and planetary sciences.
- 8+ years experience in internet and community-forum browsing for advanced computational and environmental problem solving. 7+ years experience in leveraging Google Scholar to further research.
- Passion for strengthening team interpersonal development and morale
- Basic German language skills and actively learning American Sign Language

IV. Professional and Personal References

- Inquire for details at jahbrunton@gmail.com