

The NCAR AI for Earth System Science Web Portal

David John Gagne
NCAR CISL/RAL

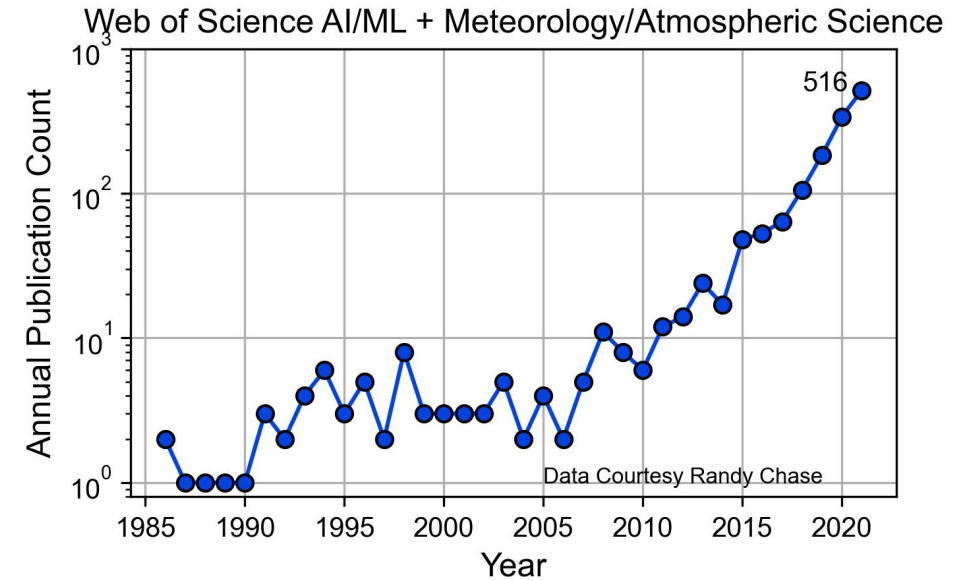
Collaborators: John Schreck, Charlie Becker, Gabrielle Gantos, Keely Lawrence,
Omar Chaarawi, Alma Hodzic, Siyuan Wang, Matt Hayman, Aaron Bansemer



January 27, 2022

Motivation

- Interest in machine learning is growing faster than existing ML experts' capacity to collaborate and mentor people
- How do we best support members of NCAR's community of scientists, faculty, and students who want to incorporate machine learning into their research?
- Goal: develop more online, asynchronous resources to support machine learning for Earth System Science applications
- Goal: Disseminate resources through a central portal website
- Goal: Identify gaps in current publicly available material



Portal Components

Tutorials

Train Linear Regression model and validate it using the box model validator

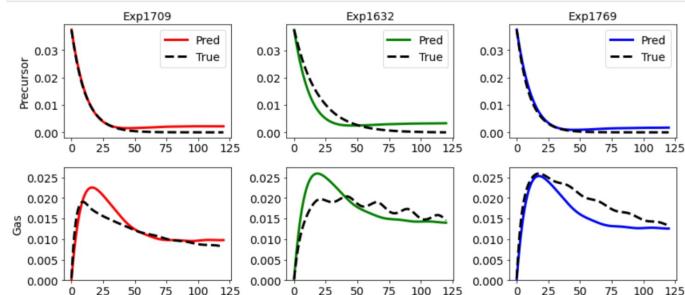
```
In [12]: lin_model = LinearRegression()
lin_model.fit(scaled_train_in, scaled_train_out)
```

```
Out [12]: LinearRegression()
```

```
In [13]: lin_mae, truth, preds, failed_exps = compare_models.box_val(
lin_model,
val_exps,
num_timesteps,
val_in_array,
val_env_array,
output_vars,
val_out
)
```

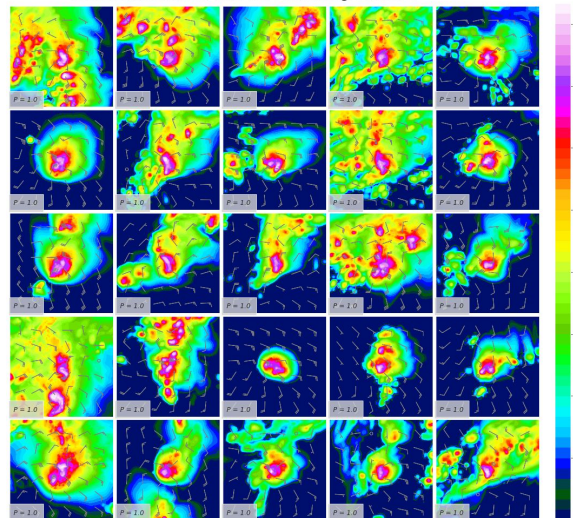
```
100% ██████████ 1438/1438 [00:00<00:00, 15162.14it/s]
```

```
In [14]: prec_lim = 0.05
gas_lim = 0.0275
aero_lim = 0.008
compare_models.plot(truth, preds, 14, prec_lim, gas_lim, aero_lim)
```



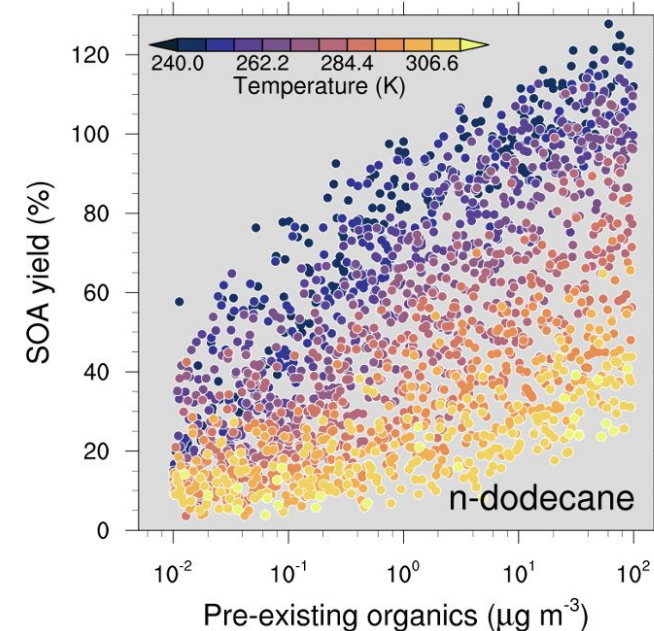
Software

Storm Classifications With GMM Clustering: Cluster 2

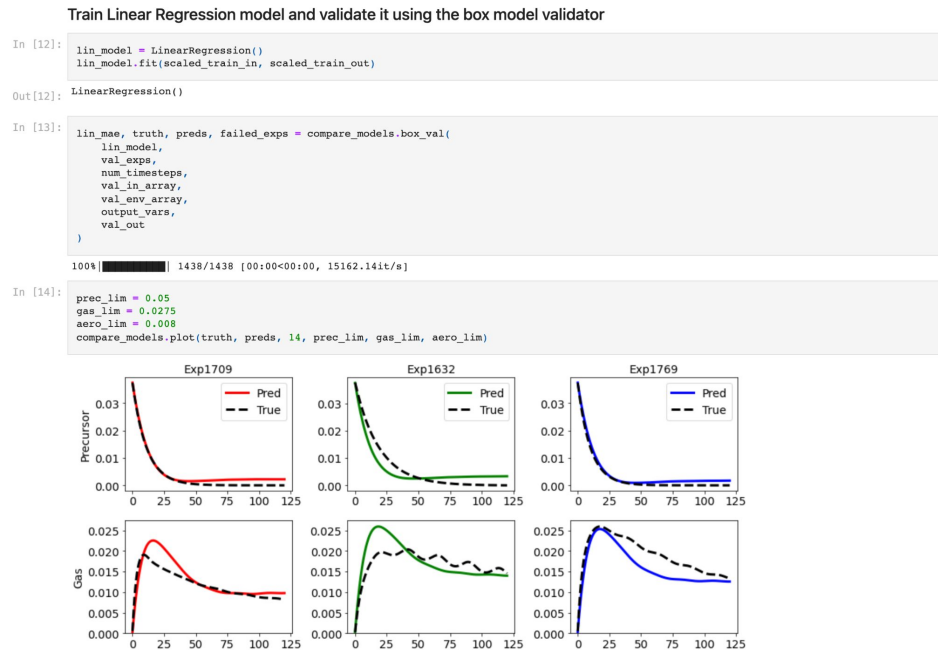


Datasets

Total number of GECKO-A simulations: 2000
Total number of species: 192417
Total number of reactions: 1102673



Tutorials



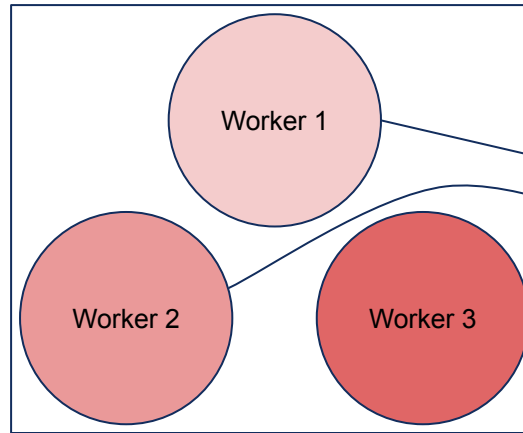
Material courtesy Omar Charawi

- Led and contributed to multiple AI/ML short courses and summer schools
 - AMS Machine Learning in Python for Environmental Science Problems Short Course
 - AI4ESS Summer School
 - Trustworthy AI for ES Summer School
- Courses utilized Jupyter notebooks and cloud-hosted datasets
- Upcoming plans
 - For beginners: asynchronous Jupyter notebook tutorials
 - For advanced users: blog posts on specific technical procedures

Software

Earth Computer Hyperparameter Optimization (ECHO): Enable easy distributed hyperparameter optimization across NCAR HPC resources and GPUs.
Site: <https://github.com/NCAR/echo-opt>

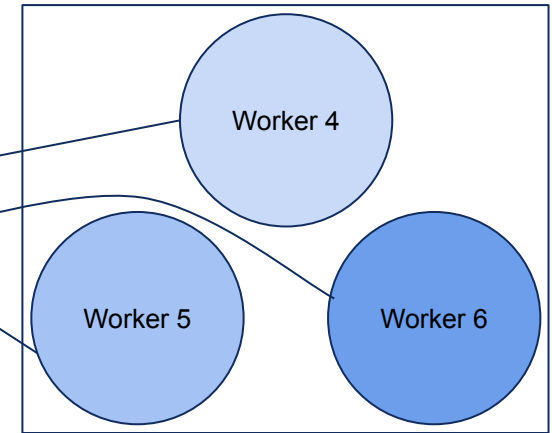
Casper: 237.11.50.6



Database:
162.268.12.43



Cheyenne: 164.10.11.6



Hagelslag: Extract storms from convection-allowing model output, ML post-processing, and probabilistic evaluation
Site: <https://github.com/djgagne/hagelslag>

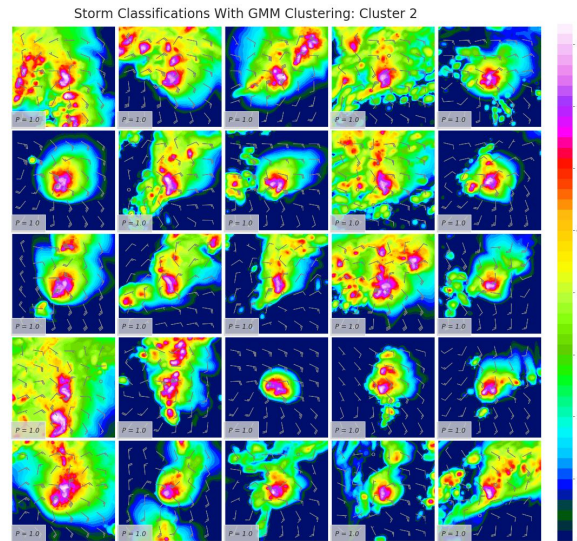
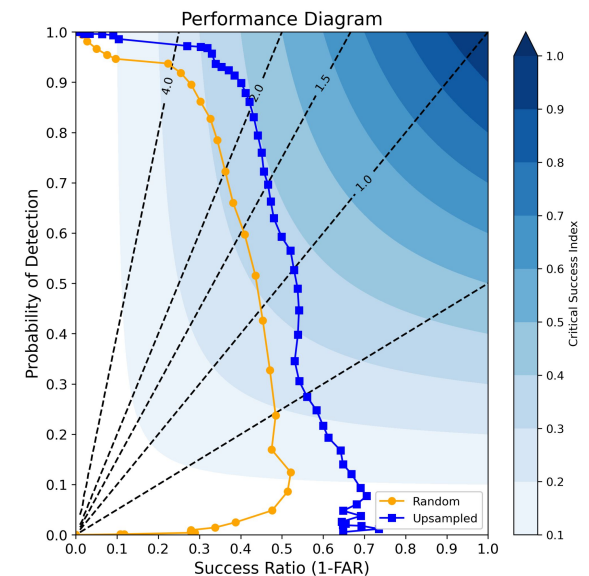


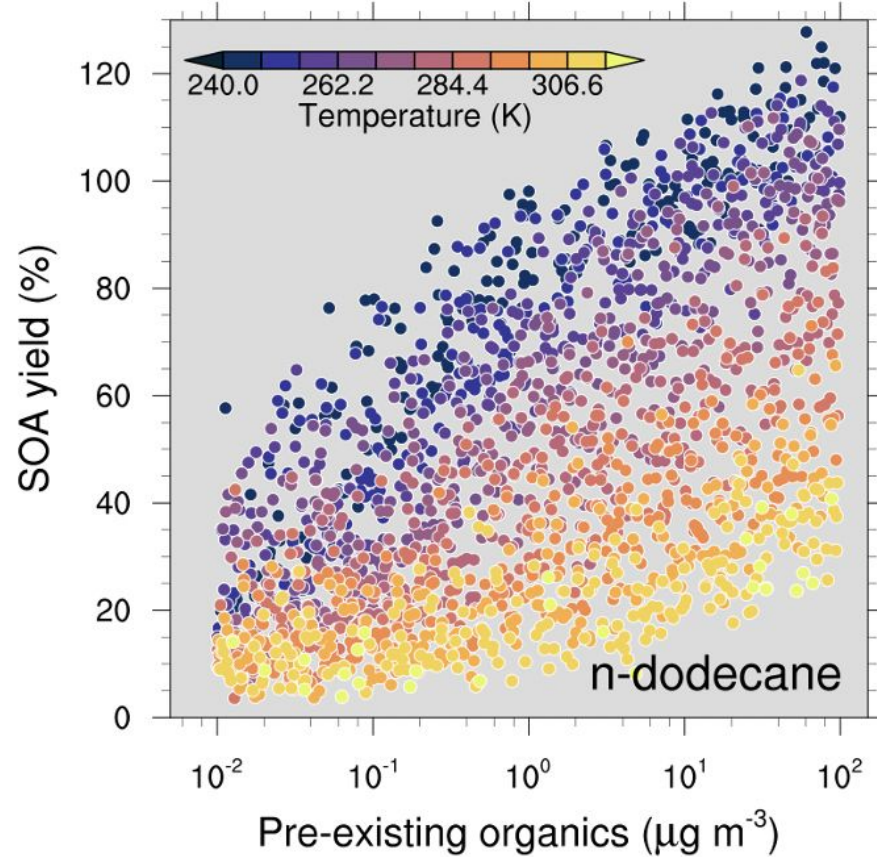
Image courtesy Charlie Becker

Image courtesy John Schreck



GECKO-A Simulations for ML Emulation

Total number of GECKO-A simulations: 2000
Total number of species: 192417
Total number of reactions: 1102673



Available at <https://doi.org/10.5281/zenodo.5790043>

AWS S3 Bucket

ncar-aiml-data-commons [Info](#)

Objects Properties Permissions Metrics Management Access Points

Objects (10)

Objects are the fundamental entities stored in Amazon S3. You can use [Amazon S3 inventory](#) to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. [Learn more](#)

[Refresh](#) [Copy S3 URI](#) [Copy URL](#) [Download](#) [Open](#) [Delete](#) [Actions](#) [Create folder](#) [Upload](#)

Show versions < 1 >

<input type="checkbox"/>	Name	Type	Last modified	Size	Storage class
<input type="checkbox"/>	besttrack_predictors/	Folder	-	-	-
<input type="checkbox"/>	FY2020 NCAR Reinvestment Project Plan - DS-ML-CSAC.docx	docx	March 27, 2020, 14:30:31 (UTC-06:00)	17.7 KB	Standard
<input type="checkbox"/>	gecko/	Folder	-	-	-
<input type="checkbox"/>	goes/	Folder	-	-	-
<input type="checkbox"/>	gridrad_poly_geojson/	Folder	-	-	-
<input type="checkbox"/>	holodec/	Folder	-	-	-
<input type="checkbox"/>	microphysics_sd/	Folder	-	-	-
<input type="checkbox"/>	microphysics_tau/	Folder	-	-	-
<input type="checkbox"/>	microphysics/	Folder	-	-	-
<input type="checkbox"/>	test.txt	txt	March 27, 2020, 15:09:09 (UTC-06:00)	6.0 B	Standard

<https://ncar-aiml-data-commons.s3.us-west-2.amazonaws.com/>

PANGEO ML DATASETS Site

WEATHER AND CLIMATE DATASETS FOR AI RESEARCH

Welcome to this collection of weather and climate datasets for AI research. In the last few years more and more datasets are being published. This website is an attempt at providing an overview of what's available.

The datasets are split into datasets specifically processed for AI research and commonly used raw datasets. In addition there is a list of common models for hybrid ML-physics modeling.

If you know of a dataset that is not already on the list, you can contribute in two ways:

1. Go to the [GitHub repository](#) of this site, clone it, add your paper to the respective Markdown page and create a pull request.
2. Create an [issue](#) in the GitHub repository with the details of the paper and one of us will add the dataset.

If you have any questions, don't hesitate to reach out to [me](#).

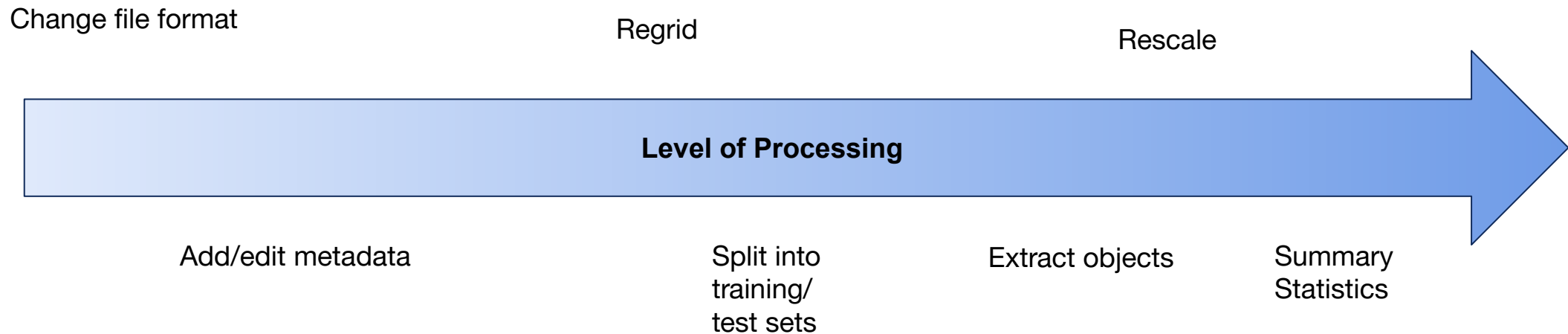
CONTENTS

- [Preprocessed Datasets](#)
- [Raw datasets](#)
- [Hybrid ML-Physics Models](#)

mldata.pangeo.io

Portal Gaps: The AI-Ready Data Processing Dilemma

AI-Ready Data Dilemma: the more processed a dataset is for a machine learning task, the less generally useful it is for other tasks



How many intermediate versions of dataset to store versus regenerate on the fly?
How much metadata to persist through each level of processing?

Portal Gaps: Domain Science Tutorials

- Problem: ML team members not understanding important scientific subtleties of problems and data
- Accessible domain science tutorials far less common or available compared with ML and Python tutorials
- Impact: machine learning models fail to generalize, solve wrong problem, produce unusual failure modes
- Solutions:
 - Include domain science tutorials/blog posts as part of funded collaboration
 - Use Data Science ecosystem tools (e.g., Jupyterbooks, interactive models) to explain Earth Science concepts
 - Train cross-domain students

The Climate Laboratory

Search this book...

The Climate Laboratory

PREAMBLE

About this Book

Who is the book for?

How to use this book

How to cite and reuse this material

LECTURES

1. Climate models, the global energy budget, and Fun with Python

2. Modeling the global energy budget

3. The climate system and climate models

4. Introducing the Community Earth System Model (CESM)

5. Building simple climate models using climlab

6. A Brief Review of Radiation

7. Elementary greenhouse models

8. Grey radiation modeling with

6. A Brief Review of Radiation

This notebook is part of [The Climate Laboratory](#) by [Brian E. J. Rose](#), University at Albany.

1. Emission temperature and lapse rates

Planetary energy balance is the foundation for all climate modeling. So far we have expressed this through a globally averaged budget

$$C \frac{dT_s}{dt} = (1 - \alpha)Q - OLR$$

and we have written the OLR in terms of an emission temperature T_e where by definition

$$OLR = \sigma T_e^4$$

Using values from the observed planetary energy budget, we found that $T_e = 255$ K

The emission temperature of the planet is thus about 33 K colder than the mean surface temperature (288 K).

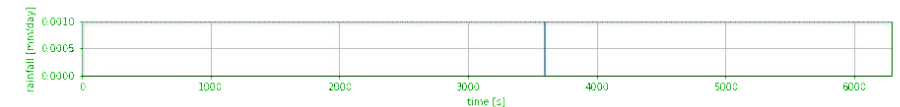
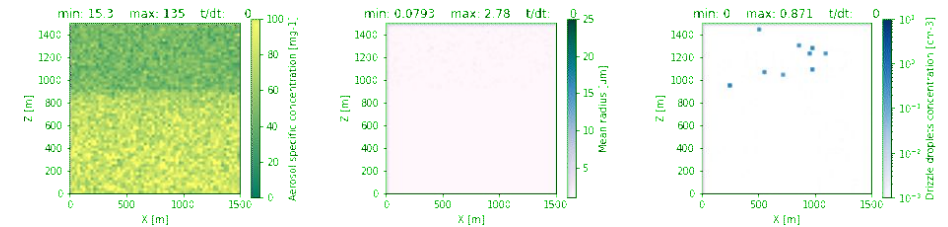
Where in the atmosphere do we find $T = T_e = 255$ K?

That's about -18°C.

Let's plot **global, annual average observed air temperature** from NCEP reanalysis data.

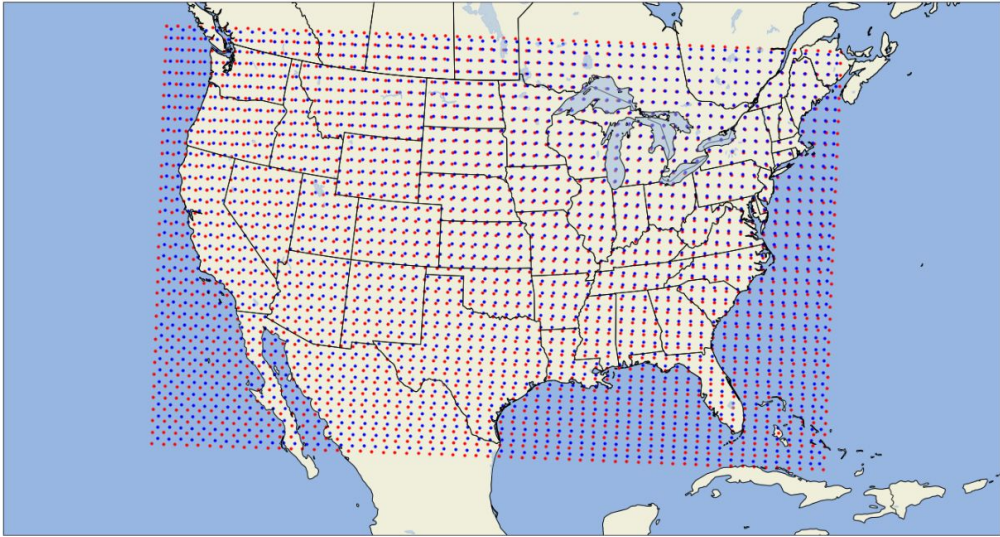
```
%matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
import xarray as xr
```

Brian E. J. Rose Climate Laboratory
(<https://brian-rose.github.io/ClimateLaboratoryBook/home.html>)



<https://github.com/atmos-cloud-sim-uj/PySDM>

Portal Gaps: Metadata



Unable to reproduce original WRF grid because of missing projection parameters

- Problem: Public datasets for machine learning and ESS do not contain all the necessary metadata to analyze the data properly or reproduce parts of analysis pipeline
- Common missing but important metadata:
 - Radius of Earth in WRF map projection info
 - Aggregation procedure (instantaneous, average, averaging window)
 - Time zone
 - Relationships among variables
 - Measured versus derived/diagnosed variables
 - Metadata about machine learning model predictions and XAI results

Portal Gaps: Interfaces between ML and Weather/Climate Software

```
integer, intent(in) :: mgncol
real(r8), dimension(mgncol), intent(in) :: qc, qr, nc, nr, rho, lamc, lamr, lcldm, n0r, pgam, precip_frac
real(r8), intent(in) :: q_small
real(r8), dimension(mgncol), intent(out) :: qc_tend, qr_tend, nc_tend, nr_tend
integer(i8) :: i, j, qr_class, nc_class, nr_class
real(r8), dimension(1, num_inputs) :: nn_inputs, nn_inputs_log_norm
integer, dimension(num_inputs) :: log_inputs
real(r8), dimension(batch_size, 2) :: nz_qr_prob, nz_nc_prob
real(r8), dimension(batch_size, 3) :: nz_nr_prob
real(r8), dimension(batch_size, 1) :: qr_tend_log_norm, nc_tend_log_norm, nr_tend_log_norm
real(r8) :: log_eps = 1.0e-30
do i=1, mgncol
  if ((qc(i) >= q_small) .or. (qr(i) >= q_small)) then
    nn_inputs = reshape((/ qc(i), nc(i), qr(i), nr(i), rho(i), &
      lamc(i), lamr(i), lcldm(i), n0r(i), pgam(i), precip_frac(i) /), (/ 1, num_inputs /))
    log_inputs = (/ 1, 1, 1, 1, 0, 1, 1, 0, 1, 1, 0 /)
    do j=1, num_inputs
      if (log_inputs(j) == 1) then
        nn_inputs_log_norm(1, j) = (log10(max(nn_inputs(1, j), log_eps)) - input_scale_values(j, 1)) / &
          input_scale_values(j, 2)
      else
        nn_inputs_log_norm(1, j) = (nn_inputs(1, j) - input_scale_values(j, 1)) / &
          input_scale_values(j, 2)
      end if
    end do
    ! calculate the qr and qc tendencies
    call neural_net_predict(nn_inputs_log_norm, emulators%qr_classifier, nz_qr_prob)
    qr_class = maxloc(pack(nz_qr_prob, .true.), 1)
    if (qr_class == 1) then
      qr_tend(i) = 0._r8
      qc_tend(i) = 0._r8
    else
      call neural_net_predict(nn_inputs_log_norm, emulators%qr_regressor, qr_tend_log_norm)
      qr_tend(i) = 10 ** (qr_tend_log_norm(1, 1) * output_scale_values(1, 2) + output_scale_values(1, 1))
      qc_tend(i) = -qr_tend(i)
    end if
  end if
end do
```

- Problem: ML packages often use a very different software stack from NWP and Earth System Models as well as forecasting interfaces like AWIPS
- Solutions:
 - Growing ecosystem of interfaces between Fortran and Python ML libraries
 - mlinwrf
 - Fortran-Keras-Bridge
 - SmartSim
 - Interfaces to PyTorch and tensorflow lite
 - Rewriting ESMs in Julia, Jax, etc.
 - Unidata is developing a containerized interface for running ML model pipelines within AWIPS 2 (<https://github.com/Unidata/awips-ml>)

Portal Website (Beta)

<https://ncar.github.io/ai4ess/>

Submit suggestions: <https://github.com/NCAR/ai4ess>



NCAR | **NCAR AI FOR EARTH SYSTEM
SCIENCE WEB PORTAL**

National Center for
Atmospheric Research

ARTIFICIAL INTELLIGENCE FOR EARTH SYSTEM SCIENCE WEB PORTAL

Resources to support machine learning at NCAR



ABOUT

The goal of the AI4ESS Web Portal is to connect members of NCAR's community of researchers and students with resources to aid in learning about and developing machine learning solutions for Earth System Science problems.

TUTORIALS

- [AI4ESS Summer School \(June 2020\)](#): This event features recorded lectures and hackathon challenge problems covering introductions to machine learning and deep learning, applications of AI to ESS problems, and emerging areas of AI research.
- [Trustworthy AI for Environmental Science Summer School \(July 2021\)](#): This event features recorded lectures covering different aspects of trustworthy AI, including explainable AI, physics-based AI, robust AI, ethics, and R2O.
- [XAI Short Course \(April 2021\)](#): An in-depth introduction to different explainable AI methods with presentations, lecture recordings, and Google Colab notebook examples.

SOFTWARE

EXTERNAL PACKAGES

- [Tensorflow](#)
- [Keras](#)
- [Scikit-Learn](#)
- [Pytorch](#)

NCAR-SUPPORTED PACKAGES

- [Hagelslag](#)
- [ECHO](#)

DATASETS

- [Pangeo Machine Learning Datasets](#): A catalog of publicly available Earth Science machine learning benchmark datasets.

Summary

- The NCAR AIML group has developed tutorials, software, and datasets to support the NCAR and broader Earth System Science/machine learning community of practice
- We have also identified gaps in existing resources and tradeoffs of different approaches
- Community feedback and requests welcome!

Trustworthy AI4ES Summer School with Trust-a-Thon

Save the Date:
June 27-July 1, 2022

Contact Me

Email: dgagne@ucar.edu

Twitter: @DJGagneDos