The NCAR AI for Earth System Science Web Portal

David John Gagne
NCAR CISL/RAL

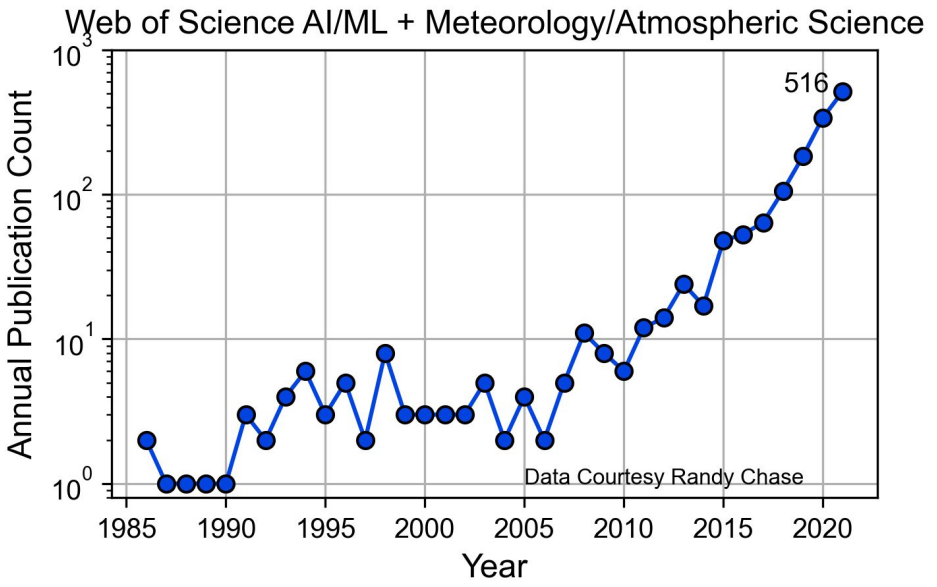
Collaborators: *John Schreck, Charlie Becker, Gabrielle Gantos, Keely Lawrence, Omar Chaarawi, Alma Hodzic, Siyuan Wang, Matt Hayman, Aaron Bansemer*

January 27, 2022

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

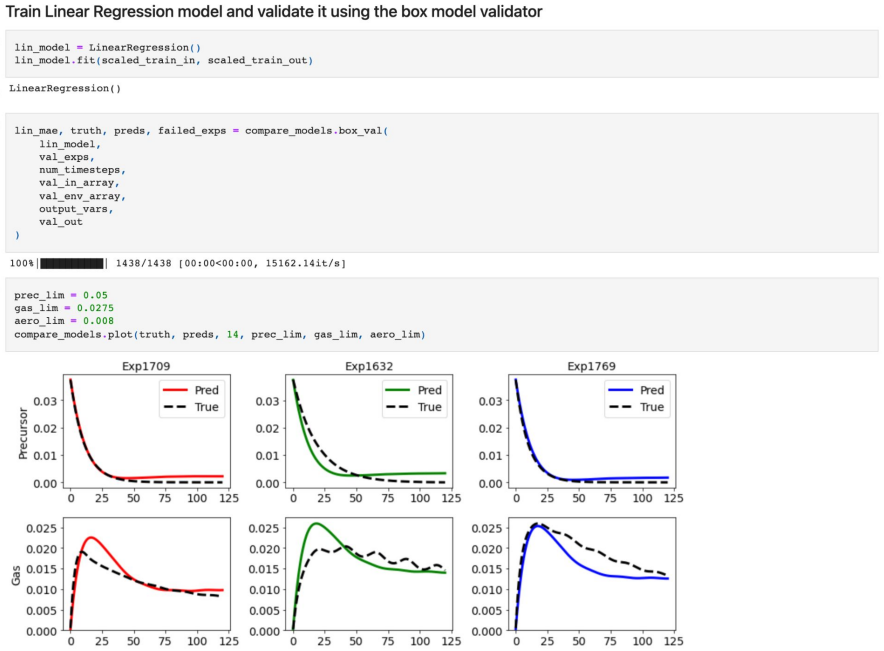
Motivation

- 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

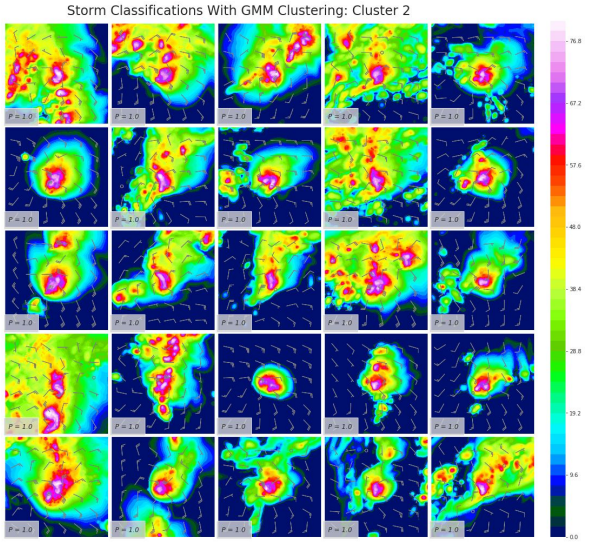


Portal Components

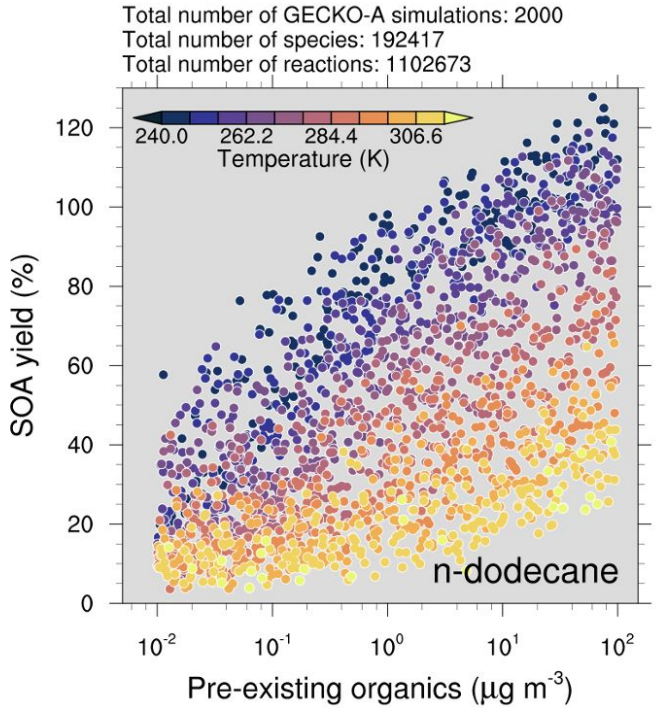
Tutorials



Software



Datasets



Tutorials

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

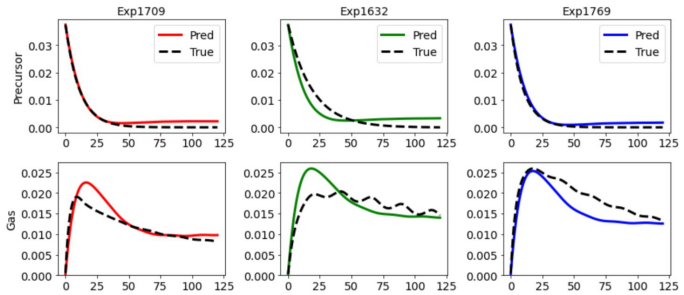
```
In [12]:
lin_model = LinearRegression()
lin_model.fit(scaled_train_in, scaled_train_out)

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)
```



Material courtesy Omar Chaarawi

- 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

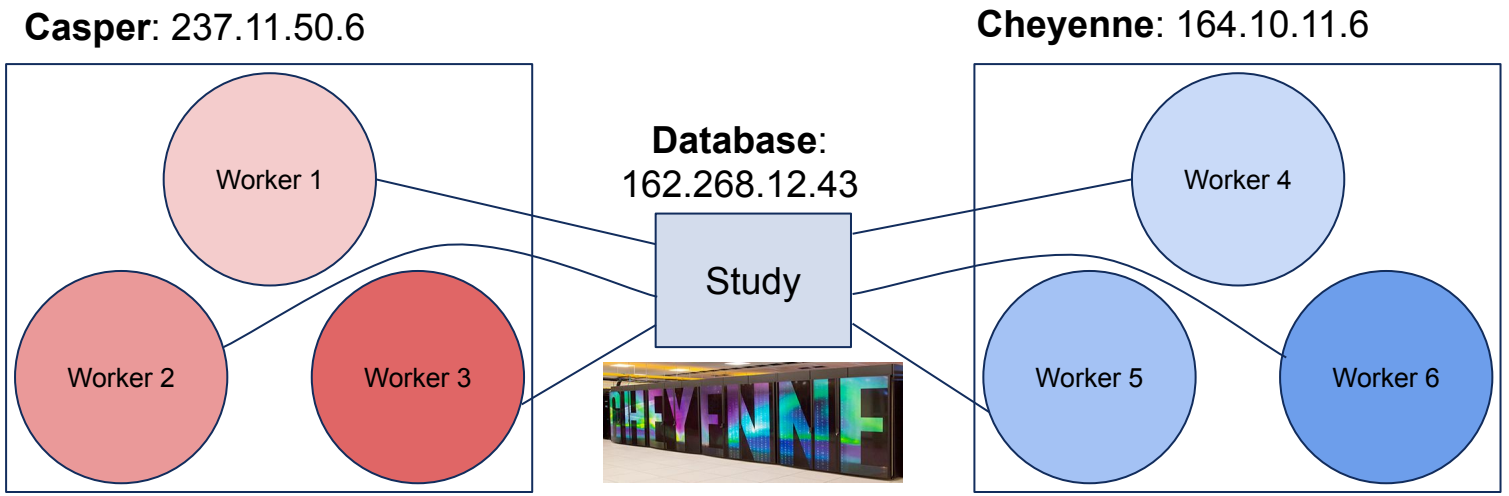


Image courtesy John Schreck

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

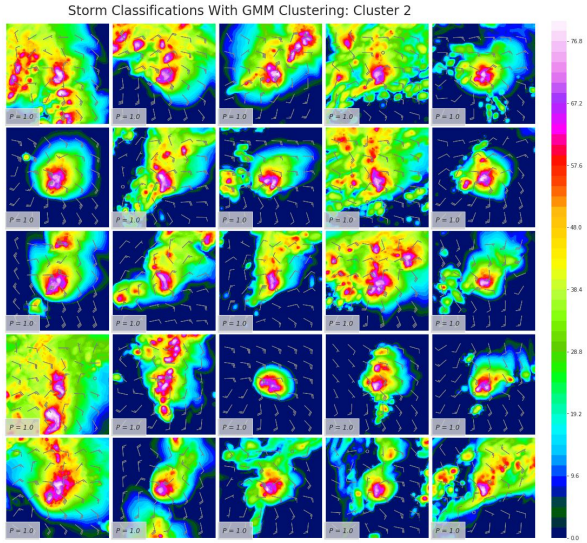
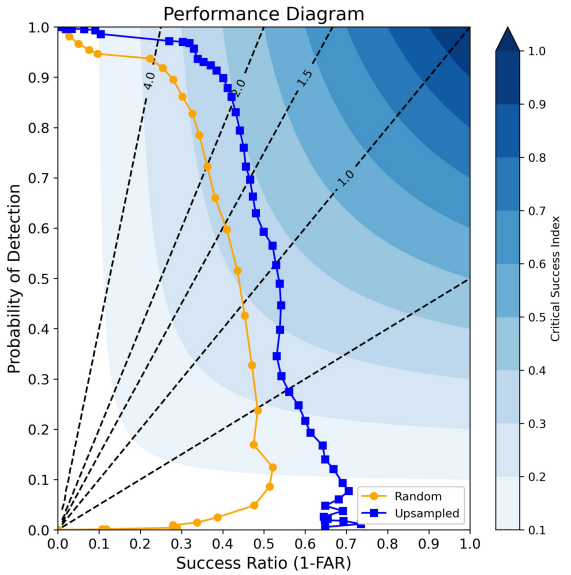
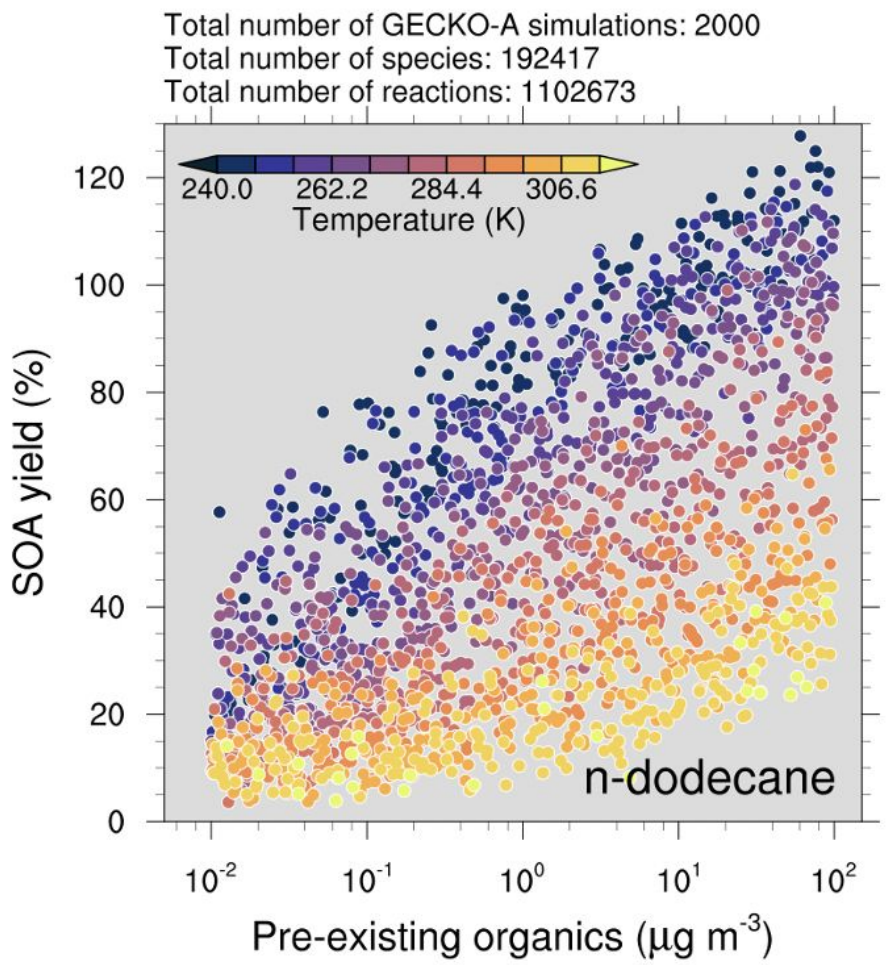


Image courtesy Charlie Becker



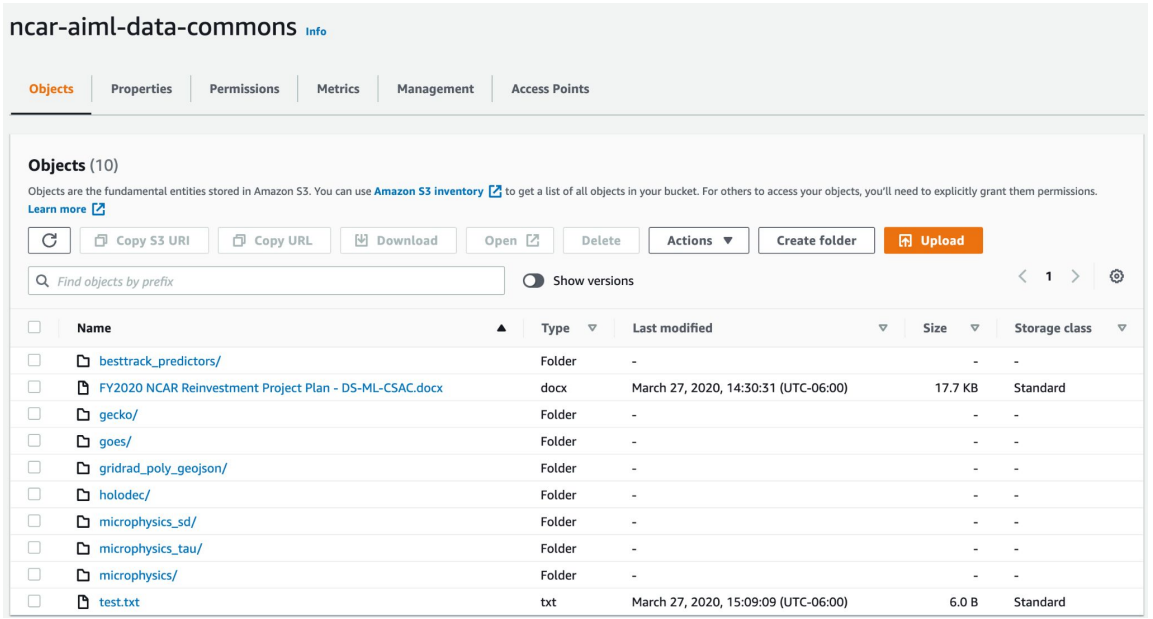
Datasets

GECKO-A Simulations for ML Emulation

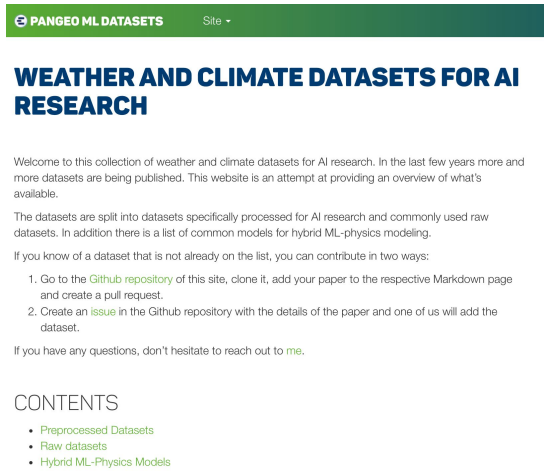


Available at https://doi.org/10.5281/zenodo.5790043

AWS S3 Bucket



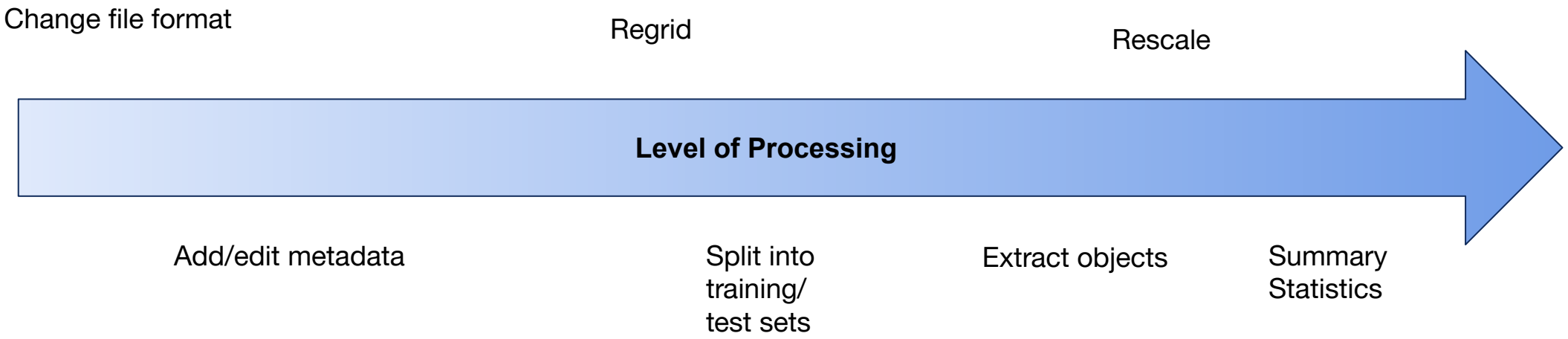
https://ncar-aiml-data-commons.s3.us-west-2.amazonaws.com/



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

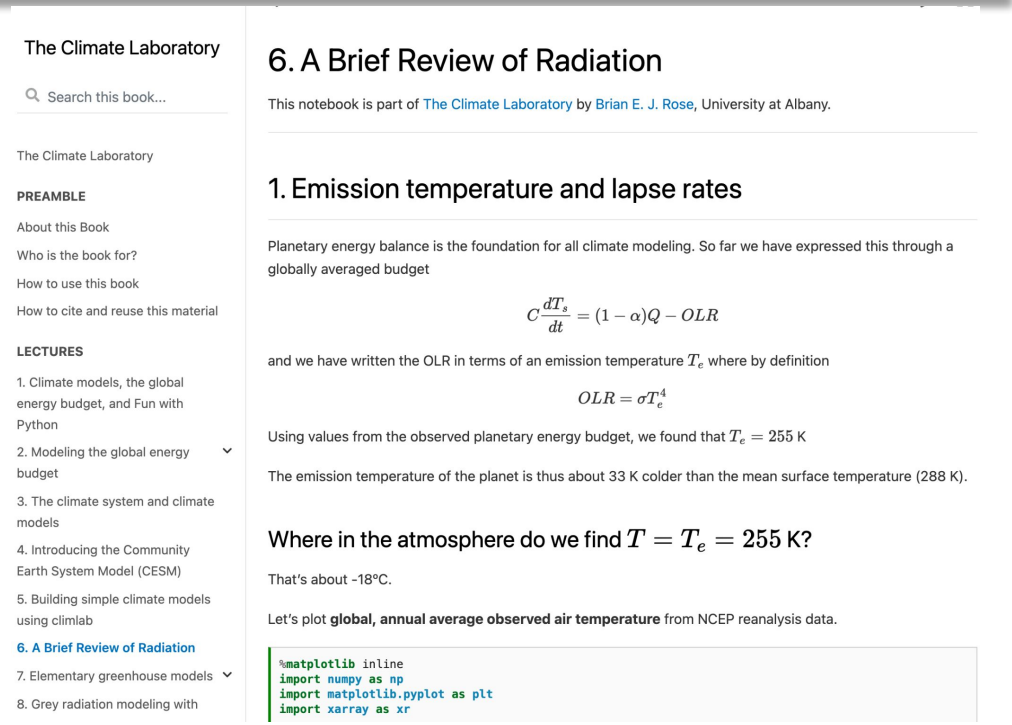


How many intermediate versions of dataset to store versus regenerate on the fly?

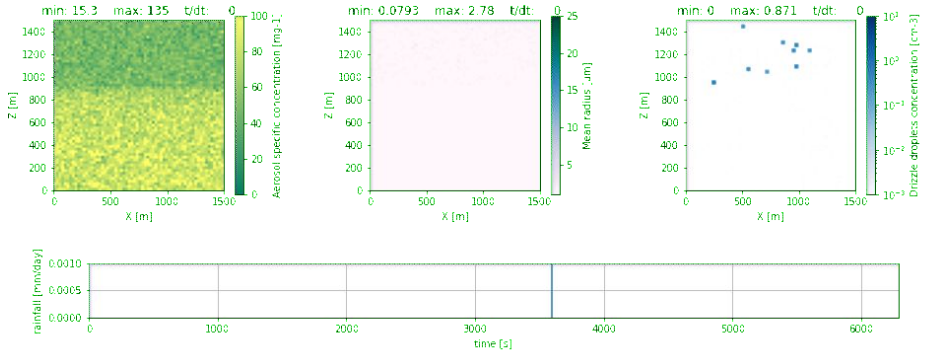
How much metadata to persist through each level of processing?

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

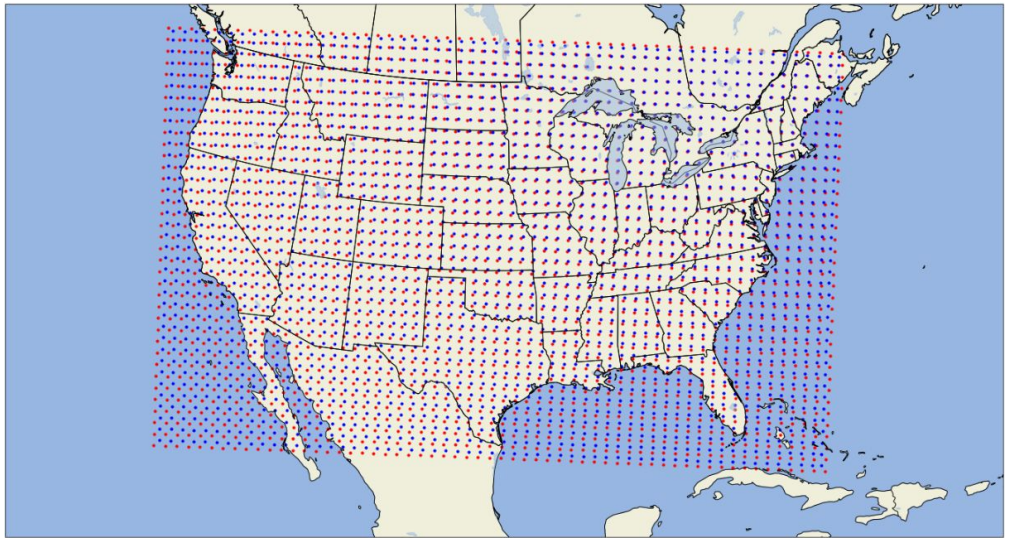


Brian E. J. Rose Climate Laboratory (https://brian-rose.github.io/ClimateLaboratoryBook/home.html)



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

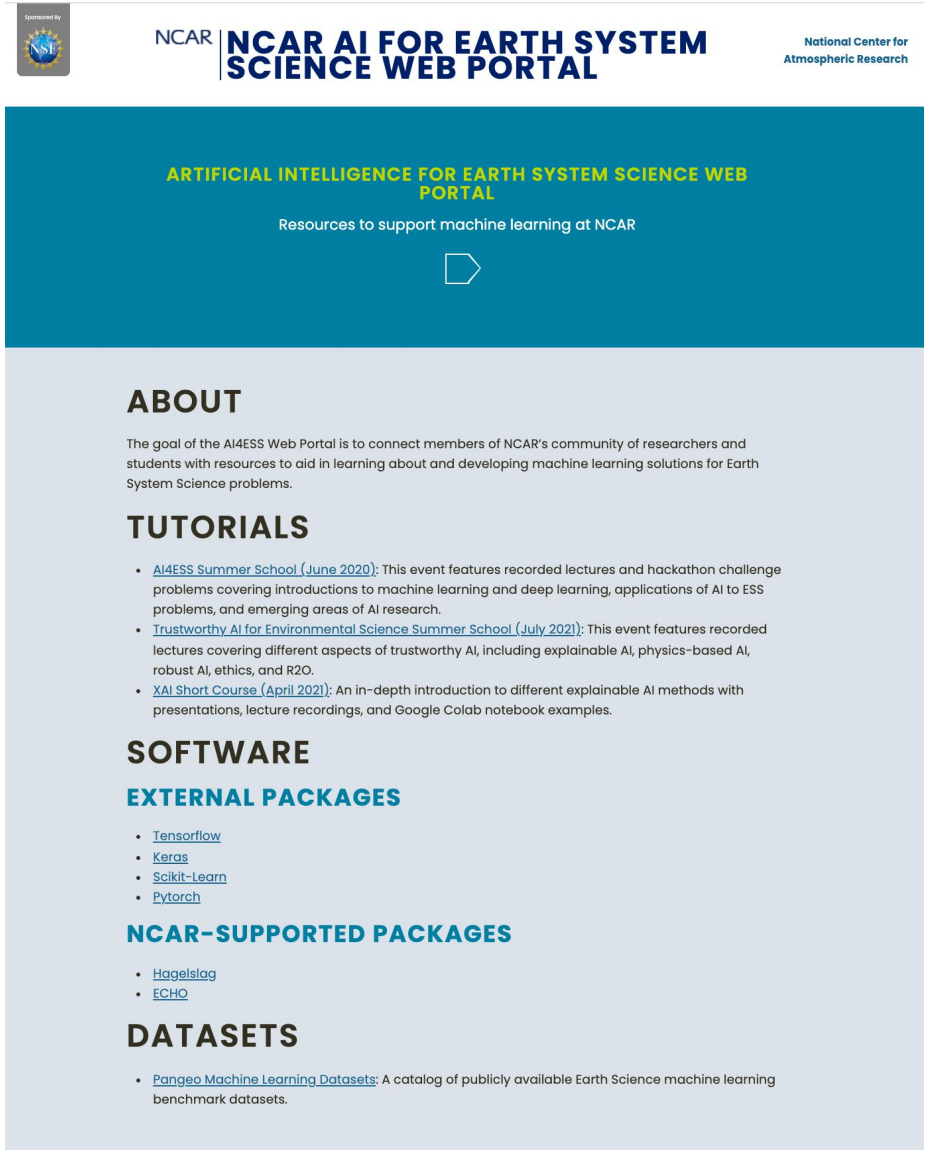
Portal Gaps: Interfaces between ML and Weather/Climate Software

```
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
```

- 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



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