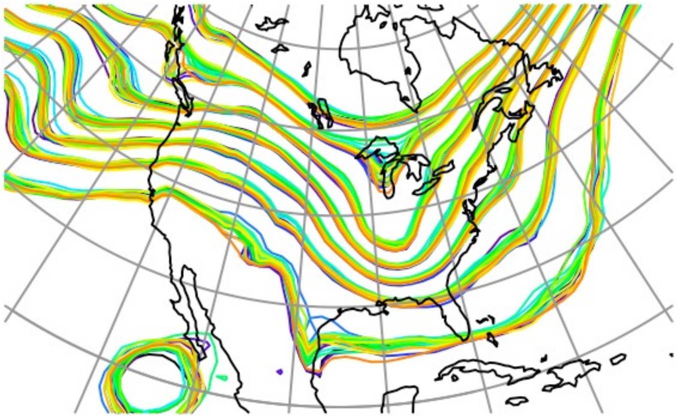
# The CAM6+DART Ensemble Reanalysis Provides a Variety of Datasets for Machine Learning Training and Verification Algorithms



The National Center for Atmospheric Research is sponsored by the National Science Foundation.
Any opinions, findings and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

NCAR | National Center for
UCAR | Atmospheric Research

# People

Kevin Raeder (CISL/DAReS): presenter

Kevin Raeder, Jeff Anderson, Tim Hoar, Moha El Gharamti,
Ben Johnson, Nancy Collins, Jeff Steward (all CISL/DAReS):
Created the reanalysis

Ian Grooms (CU); first query about using this dataset for ML

Katie Dagon, Maria Molina (CGD); exploratory discussions of ML

# Context and Goals

## Wikipedia

High-quality labeled training datasets for supervised and semi-supervised machine learning algorithms are usually difficult and expensive to produce because of the large amount of time needed to label the data.

## Reanalysis

a picture of the state of the atmosphere (or other system) which uses the information in both model hindcasts and observations, taking account of the uncertainties in them.

## Ensemble

The tool we used is the ensemble Kalman filter, which requires an ensemble description of the state of the model’s atmosphere.

## The result

A variety of datasets which may be useful as training and verification data in maching learning contexts. Some of it is (highly?) labeled.

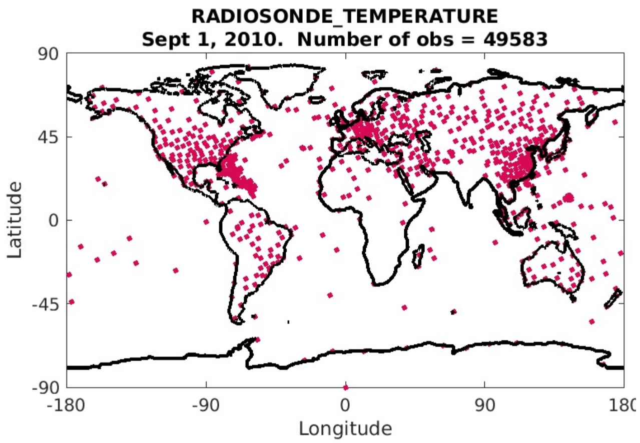
We know data assimilation and these datasets.
We don’t know much machine learning.

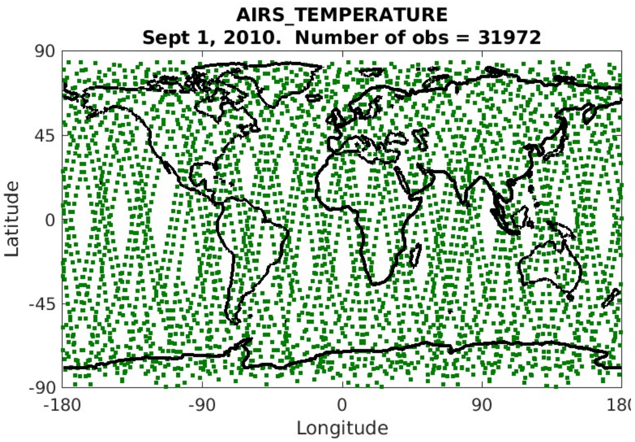
# The Model

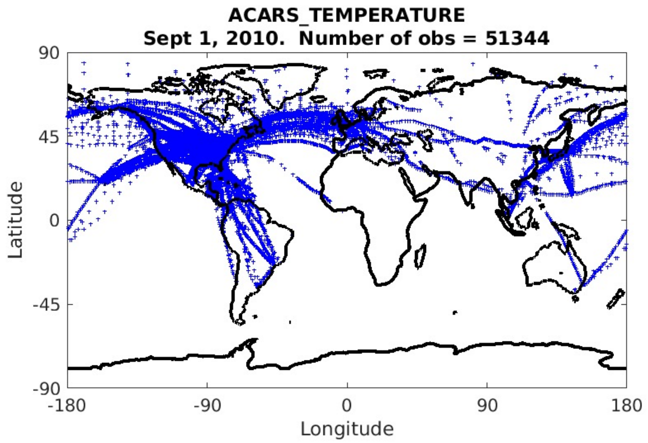
- CESM 2.1 release
- Atmosphere: CAM6.0.34
- 0.9 degree lat. x 1.2 degree longitude, 32 levels
- Land: CLM 5.0 BGC-CROP version, same grid as CAM
- CICE: coverage specified in SST file, the rest prognostic
- MOSART river model
- SST: specified daily 0.25 degree from AVHRR
- Aerosols, greenhouse gases, volcanic forcing: historical when available, moderate climate change scenario otherwise.

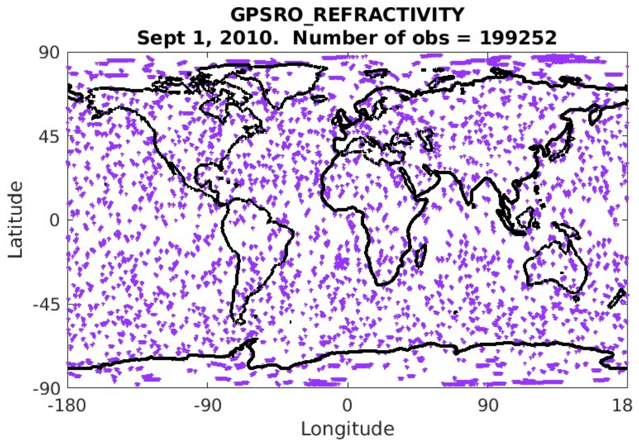
# Several Million Obs/Day

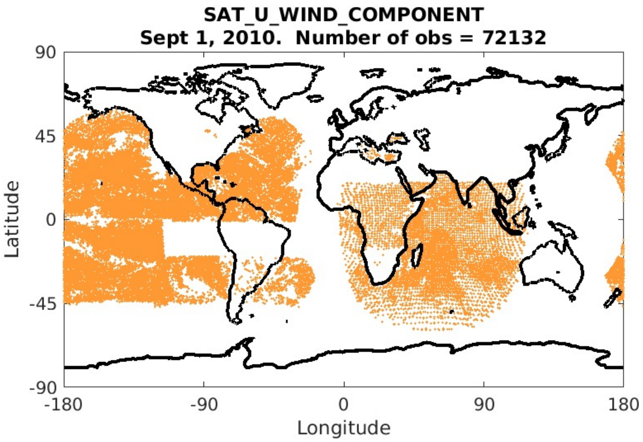
Assimilated into model state: PS, T, U, V, Q, CLDLIQ, CLDICE every 6 hours.

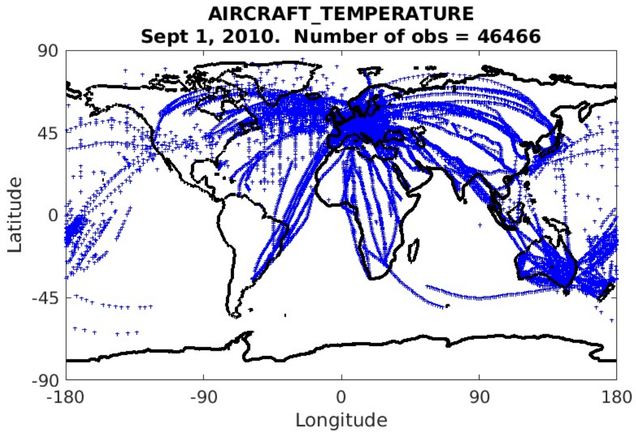

RADIOSONDE_TEMPERATURE
Sept 1, 2010. Number of obs = 49583


AIRS_TEMPERATURE
Sept 1, 2010. Number of obs = 31972


ACARS_TEMPERATURE
Sept 1, 2010. Number of obs = 51344


GPSRO_REFRACTIVITY
Sept 1, 2010. Number of obs = 199252


SAT_U_WIND_COMPONENT
Sept 1, 2010. Number of obs = 72132


AIRCRAFT_TEMPERATURE
Sept 1, 2010. Number of obs = 46466

# What This Yields

80 equally likely CESM states consistent with
✓ the actual weather
✓ CAM6.0.34 physics.

"Consistent" = a balance of the information in the obs and in the model hindcast, which explicitly accounts for the uncertainties in those 2 main sources of information, which are represented by the observational errors and model ensemble spread)

This new reanalysis is similar in quality to the NCEP/NCAR reanalysis but with 80 ensemble members in addition to the mean.

# Unprecedented(?) Data Set

80 member ensemble output

LENS and ensemble DA have shown the value.

4x per day (many data types) for 9 years.

Most are NetCDF, the rest can be converted.

Investigating reformatting to Zarr for use in cloud computing.

Bonnlander, de La Beaujardiere, McGinnis (CISL)

Extensive, free data (https://rda.ucar.edu/datasets/ds345.0 )

# Observations Files

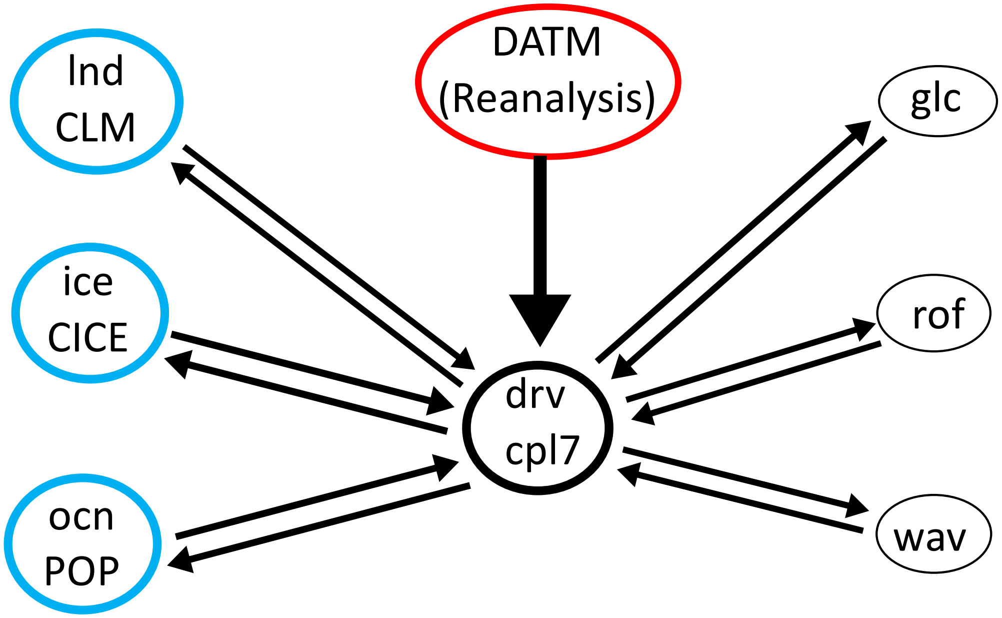
Contain input and output of the assimilation, and a variety of metadata:

- ✓ instrument and quantity observed (input, metadata)
- ✓ actual observation value (input)
- ✓ observation error estimate (input, metadata)
- ✓ ensemble of model estimates of the observation (output)
- ✓ mean and standard deviation of this ensemble (“most likely value”, uncertainty)
- ✓ quality control labels (input and output, metadata)
- ✓ observation locations (4D, non-gridded, time evolving, metadata)
- ✓ up to a million observations in each assimilation time window
- ✓ > 13,000 assimilation windows

Combining obs error with ens spread gives “total spread”;
a better measure of consistency of obs and model.

# “Data Atmosphere” Forcing Files

CESM can be configured to run “surface component”(s) using atmospheric data read from a file, instead of calculated in CAM.



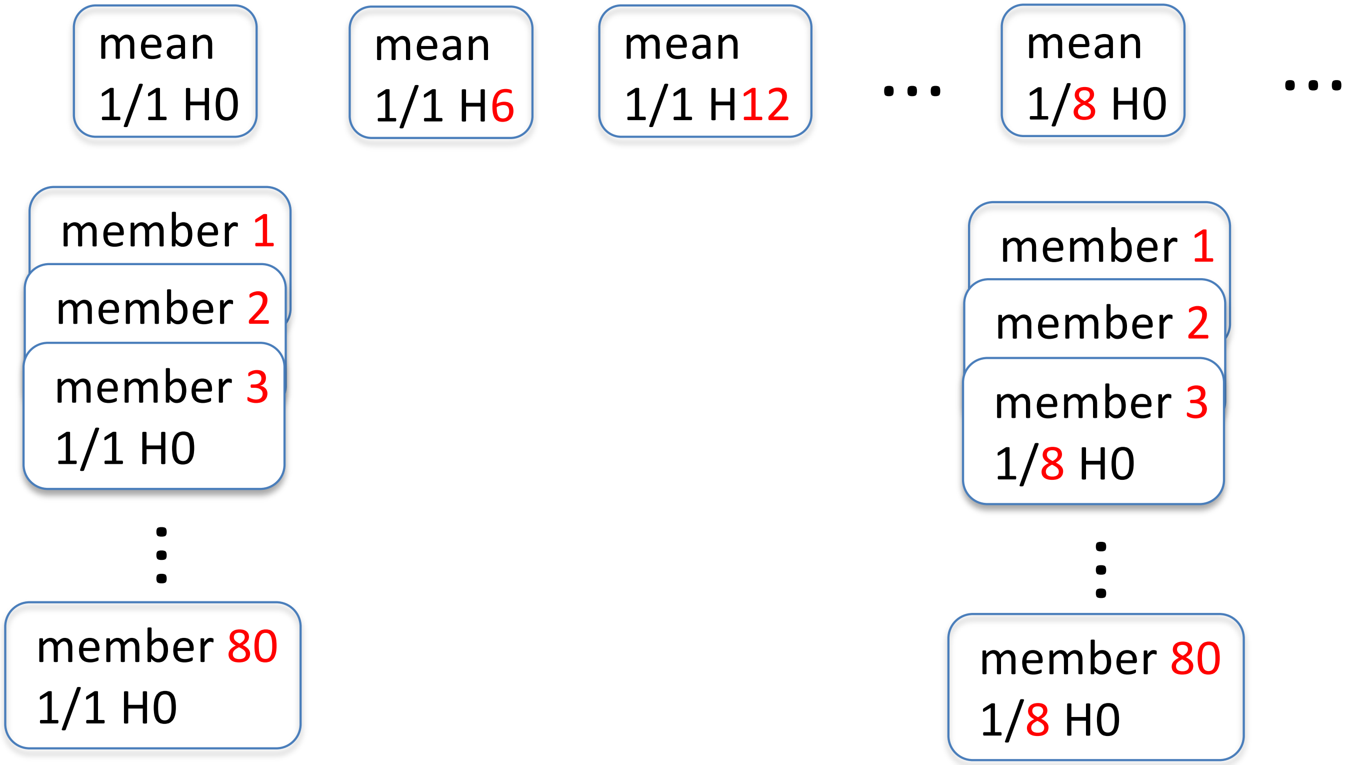
Cpl “history” files:

- Another form of output
- 80 member ensemble
- Cadences ranging from 1-6 hours
- 9 year span
- 2D gridded fields (long., lat.)
- Little metadata
- Very useful
- Expensive to produce more (or to reproduce)

All can use the forcing files for (ensemble) hindcasts.
DART has interfaces to CLM, CICE, and POP, which enables assimilation.

# Atmospheric State Files

Ensemble mean (and st. dev.) of the CAM model state (PS, T, ...) is available every 6 hours.
80 member ensemble is available weekly.
2 different views of CAM's attractor, when constrained by observations.
Is this combination useful or interesting in ML?



mean 1/1 H0

mean 1/1 H6

mean 1/1 H12

...

mean 1/8 H0

...

member 1

member 2

member 3 1/1 H0

⋮

member 80 1/1 H0

member 1

member 2

member 3 1/8 H0

⋮

member 80 1/8 H0

# CLM history files

4x/day frequency
80 member ensemble

Explore the impact of weather observations andor forcing variability on plant/crop growth.
One more step removed from the obs.

Variables archived:

```
'TSA',      'ER',  'EFLX_LH_TOT',     'HR'
'CPHASE',   'GPP', 'GRAINC_TO_FOOD',  'GSSHALN',
'GSSUNLN',  'NPP', 'NPP_NUPTAKE',     'PLANT_NDEMAND',
'QVEGT',    'TLAI'
```

# Discussion

Which data set(s) seem most (or least) useful for ML?

- Observations files
- DATM forcing files
- Atmospheric mean and ensembles
- CLM plant growth

How would they be useful?

Does data need to be reformatted?

Which metadata could qualify as ‘labels’ of the data?

---

*Students: we’d love to work with you!*

https://dart.ucar.edu

dart@ucar.edu

# Discussion (part 2)

How would they be useful?

- Emulators?
  - real obs (+obs error) ->? ensemble of estimated obs
  - real obs (+obs error) ->? DATM forcing files
- Geophysical pattern recognition?
  - Hurricanes, atmospheric rivers, blocking events
  - Patterns within the ensembles
- ...?

Does data need to be reformatted?

- Opinions about OpenML?
- Opinions about Matlab?
  - ensemble methods?
  - deep learning?
  - “tall” framework for datastores?

# Existing Datasets (Wikipedia)

Cloud DataSet

>    El Nino Dataset(1999)  178080 inst of 12 weather attributes

Greenhouse Gas Observing Network

Atmospheric CO2 from Continuous Air Samples at Mauna Loa Observatory

Ionosphere Dataset

>    Ozone Level Detection Dataset (2008) 2536 inst of many features incl. weather @ time

Should I try to put any of these into those lists?