# CESM-DART

# Updates on Assimilation for Earth System Models

Helen Kershaw DAReS NSF NCAR





# SIParCS DART NUOPC

# CROCODILE pyDART

# Terminology

**SIParCS** Summer Internships in Parallel Computing

**DART** Data Assimilation Research Testbed

**NUOPC** National Unified Operational Prediction Capability

**CROCODILE** CESM Regional Ocean and Carbon cOnfigurator with Data assimilation and Embedding

**CESM** Community Earth System Model



Group of model forecasts



Group of model forecasts

Measurements





Group of model forecasts

### Measurements

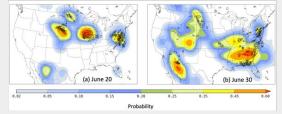




## Improved estimate

### NCAR Real-time ensemble prediction system

Severe weather forecast for two days compared to NWS warnings

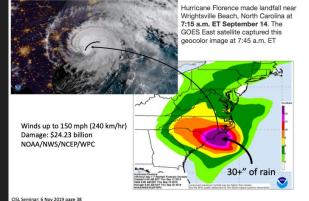


- WRF, 10 member ensemble, GFS for boundary conditions
- Continuous operation from April 2015 to December 2017
- 48 hour forecasts at 3km resolution
- · First continuously cycling ensemble system for CONUS
- CISL Dedicated Queues and Computing Support were Vital

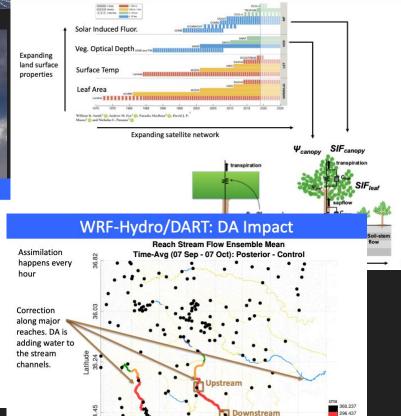
CISL Seminar: 6 Nov 2019 page 28

### WRF-Hydro/DART: Florence 2018

.....



### Advancing models & observations together





Featured project: Computational & Information 51 stems Lab & Research Applications Lab Collaboration

PREDICTING FLOODS AND

**PROTECTING LIVES** 



CL NC State, UC San Diego, MIT & KAUST Collaboration

## UNDERSTANDING GULF OF MEXICO EDDY DYNAMICS





DATA ASSIMILATION FOR THE ENTIRE EARTH SYSTEM Use ensemble DA techniques with USE DATA FROM ANY SOURCE, TEST MANY ALGORITHMS Assimilate any suitable



SUPERCOMPUTERS Complie without MPI for conceptual models or with MPI for GCMs on

LEARN ON LAPTOPS, RUN ON

Documentation Tutorials

MILATION FOR THE

le DA techniques with

models spanning the

EARTH SYSTEM



USE DATA FROM ANY SOURCE, TEST MANY ALGORITHMS Assimilate any suitable observations. Swap out filter and



LEARN ON LAPTOPS, RUN ON SUPERCOMPUTERS

Compile without MPI for conceptual models or with MPI for GCMs on

NEXT-GENERATION SPACE WEATHER PREDICTION

Featured project: University of Michigan, NCAR, NASA & NRL

Collaboration



DATA ASSIMILATION FOR THE



LEARN ON LAPTOPS, RUN ON

# dart.ucar.edu

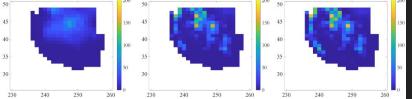
9

## Shout out: Xueli's talk

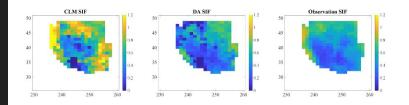
### Monday, June 10th, 2024 CESM Land Agenda 2:08pm- Assimilating SIF and SWE observations into CLM using DART to improve GPP over the high mountains in the Western US Xueli Huo



Despite challenges with snow input and melting, DA successfully corrects Snow



Though the regional average of SIF aligns well with the observation, SIF is overcorrected along the Cascade Range and in Montana



Need to refine the way of leaf nitrogen altered by DA to effectively impact GPP

# Shout out: Moha & Kevin's posters

### The Data Assimilation Research Testbed: Recent Advances and Tools for CESM



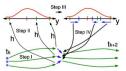
#### Moha Gharamti<sup>1</sup> (DAReS/TDD/CISL; gharamti@ucar.edu), J. Anderson<sup>1</sup>, B. Gaubert<sup>1</sup>, D. Amrhein<sup>1</sup>, A. RafieeiNasab<sup>1</sup>, Y-C. Noh<sup>2</sup> <sup>1</sup>U.S. National Science Foundation National Center for Atmospheric Research (NSF NCAR), Boulder CO, USA <sup>2</sup>Korea Polar Research Institute, Incheon, South Korea

### DART Algorithm and Tools

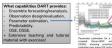
The Data Assimilation Research Testbed (DART) is an ensemble community facility, developed and maintained by the Data Assimilation Research Section at NSF NCAR.

DART's sequential filtering algorithm (Anderson, 2009) solves the following Bayesian problem:

 $p(x_t, \theta | Y_t) \propto p(y_t | x_t, \theta, Y_{t-1}) p(x_t, \theta | x_{t-1})$  $x_t$ : state,  $\theta$ : parameters,  $y_t$ : observations



Step I: A model integration step using 3 members Step II: Forward operator (h) computes expected obs Step III: Ensemble increments in obs space are computed Step IV: Linear repression of obs increments onto the state



#### Additional DA Tools

Inflation: to increase the spread Localization: to limit the impact of observations in space Others: Sampling error correction,

Spread restoration, Hierarchical groups, diagnostics, ... Ensemble covariance inflatio

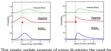
#### References

- Anderson, J., Haar, T., Raeder, K., Liu, H., Collins, N., Torn, R., & Avellana, A. (2009). The data association research testhart & research facility. Buildrin, of the American Meteorological
- Barrison reset. Society, 2013, 123-1234. Anderson, Jeffrey L. (2022). A quantile-conserving ensemble Star framework. Part I: Updating an
- 1(10), 2759-2777. I, Redel, C., Wieringa, M., Ishraque, F., Smith, M., & Kershaw, H. (2024). A Quantile
- EL Gharanti, M. (2223). Hybrid Ensemble-Variational Filter: A Soutially and Temporalis Variet
- e Algorithm to Estimate Relative Weighting, Monthly Weather Review, 245(1), 65-76. awti, M. (2023), A randomized dormant ensemble Kalman filter. Monthly Weather Review. Aust. B. et al. (2024). Neolinear and non-Gaussian ensemble assimilation of MOPITT CO.

### Non-Gaussian DA

Quantile Conserving Ensemble Filtering Framework (Anderson, 2022, 2023, 2024) OCEFF: is a general framework for ensemble DA. It serves as

an alternative to traditional ensemble Kalman filter (EnKF) methods which make implicit Gaussian and linear assumptions violating many Farth system applications

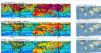


non-Gaussian DA to produce physically meaningful update Left: An EnKF update of ozone using 10 ensemble members Right: QCEFF application using a Gamma distribution for the prior and the likelihood



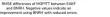
Another example demonstrating the need for nonlinear and non-Gaussian DA update. Displayed are the joint prior (Left) and posterior (Right) distributions of an observer temperature and upobserved ice fraction variables

#### CAM-Chem-DART (Gaubert et al. 2024



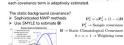
A 1º CAM-Chem (CESM2.2) with 32 vertical layers is coupled to DART to assimilate MOPITT V9I CO retrievals (Left) and in-situ surface CO using NDACC station network (Right). Three DA experiments are compared: Control (no CO), EAKF (Gaussian filter) and the BNRH (a OCEFF variant)

### RMSE differences of MOPITT between EAKF and BNRH. Negative values indicate an



#### Hybrid EnKF & 3D-Var/OI An Adaptive Hybrid Ensemble-Variational Scheme

(Gharamti, 2021) EnKF-OI: is a hybrid filtering scheme combining the ensemble flow-dependent covariances with time-invariant background covariances from 3D-VAR or OI. The weight on





reamflow and Flooding Applicat in DART [Gharamti et al., 2024]

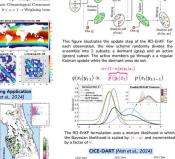


The adaptive hybrid EnKF-OI scheme is tested in a regional flooding domain due to hurricane Ian (Florida, 2022). The static background covariance is estimated using a 40 year retrospective model run



EnKE-OI with different ensemble sizes

cs using all gauge



Extreme Sampling Errors

A Randomized Dormant Ensemble Kalman Filter

(Gharamti, 2023)

RD-EnKF: is a new variant of the EnKF, specifically designed

to tackle extreme sampling errors when the ensemble size is

considerably smaller than the size of the state

Difference in monthly-mean sea ice concentration for the control run (no DA: left), EnKF (middle), and RD-EnKF (right), The RD-EnKF yields larger ensemble spread than the EnKF

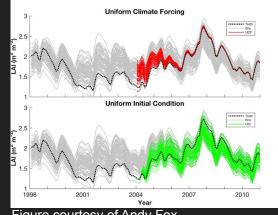
 $p(x_t|y_{1:t-1})$ 

#### Acknowledgements

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility roonsored by the National Science Ecupdation under Cooperative Agreement No. 1852977, Any opinions, findings and conclusions or recommendations expressed in this material do not necessarily reflect the views of NSF

### Essential Tools for Predictability Studies Provided by the Data Assimilation **Research** Testbed

Importance of ensemble forcing to land assimilations and a couple showing what happened to the Reanalysis during COVID (aircraft obs disappeared, both biases increased).



### Figure courtesy of Andy Fox.

# SIParCS DART NUOPC

# CROCODILE pyDART

"Optimizing ensemble data assimilation performance for coupled Earth System models" investigating the use of NUOPC/ESMF for DA, and comparing this to traditional 'offline' modes of Data Assimilation.

Anh Pham Suman Shekhar



"Optimizing ensemble data assimilation performance for coupled Earth System models" investigating the use of NUOPC/ESMF for DA, and comparing this to traditional 'offline' modes of Data Assimilation.

Anh Pham Suman Shekhar Mentors: Dan Amrhein Helen Kershaw



"Optimizing ensemble data assimilation performance for coupled Earth System models" investigating the use of NUOPC/ESMF for DA, and comparing this to traditional 'offline' modes of Data Assimilation.

Anh Pham Suman Shekhar Mentors: Dan Amrhein Helen Kershaw

Thanks to the ESMF team and Alper Altuntas

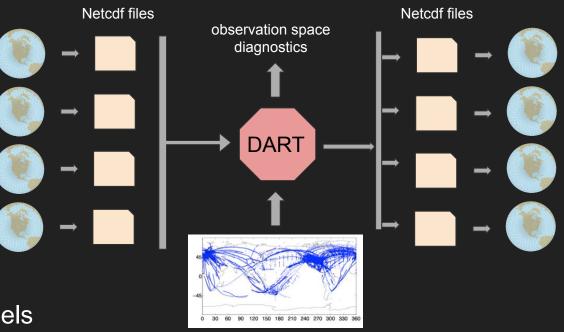


Every assimilation window:

- Advance the ensemble of models
- Every model writes their state to file
- DART reads all model states and observations for the time window
- Assimilation updates the ensemble of states
- DART writes updated state files
- Models restart with updated states

### Lots of data movement <u>"transpose</u>"

## IO: Models -> DART -> Models

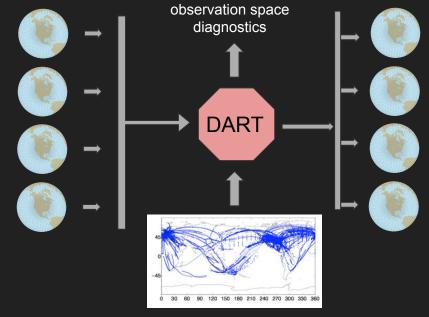


observations

### Every assimilation window:

- Advance the ensemble of models
- DART reads the observations for the time window
- Assimilation updates the ensemble of states
- Models continue with updated states

Lots of data movement "transpose" Not going to disk



observations

# CROCODILE



# 5 year collaboration between NSF NCAR and WHOI

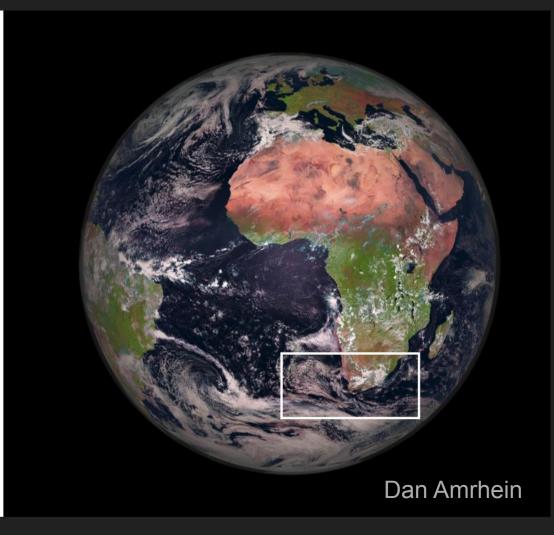
### Motivation

How do we increase usability of infrastructure that translates global dynamics to human / actionable scales?

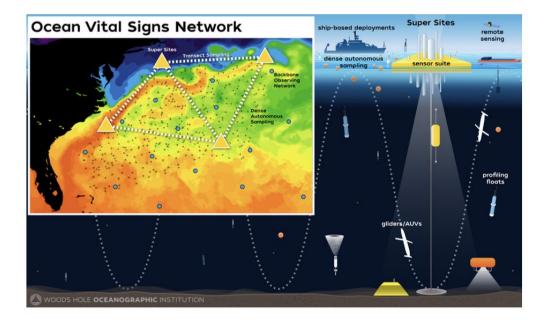
**Regional model configuration** requires setup and tuning.

**Data assimilation** requires years of effort and technical capacity building.

Few have **access** to the computational resources and tools required for configuring and running these systems and then analyzing the relatively large data sets they generate.



### Use case 1: The Ocean Vital Signs Network (OVSN) and ocean CDR



Designing next-generation ocean observing systems



Dan Amrhein

# CROCODILE DART



# 5 year collaboration between NSF NCAR and WHOI

# CROCODILE (py)DART



# 5 year collaboration between NSF NCAR and WHOI

## Data Assimilation Research Testbed

### Observations



## 

### State Space

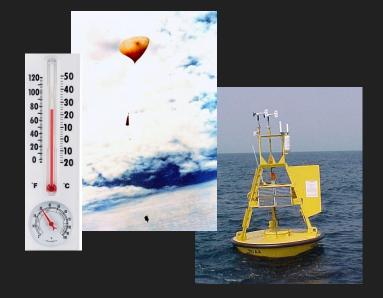
### **Observation Space**

## Data Assimilation Research Testbed

### Observations



State Space



### **Observation Space**

# DART-MOM6

latest

Search docs

#### **GETTING STARTED**

- System requirements
- Fortran90 compiler
- Locating netCDF library
- Downloading DART
- Compiling DART
- Verifying installation

### WHAT IS DATA ASSIMILATION?

Introduction to ensemble data assimilation

The Lorenz 63 model and its relevance to data assimilation

Data assimilation in DART using the Lorenz 63 model

#### WHAT IS DART?

What is DART? The benefits of using DART A brief history of DART

#### A / MOM6

### MOM6

A new ocean component model based on the Modular Ocean Model version 6 (MOM6) has been incorporated into CESM and is anticipated to replace POP2 as the default ocean component in CESM3. An early functional release of the MOM6 ocean component has been made available to users beginning with CESM2.2. Instructions for using MOM6 in CESM are available on the MOM\_interface GitHub Wiki.

This DART-MOM6 interface was developed for MOM6 within the CESM framework.

### MOM6 time

The default in CESM is to run with no leap years. To assimilate real observations, we need to switch to the Gregorian calendar to account for leap years.

./xmlchange CALENDAR=GREGORIAN

To illustrate what happens if you do not set CALENDAR=GREGORIAN, here is an example where the RUN\_STARTDATE is set to 2015-02-01 and MOM6 is run for 10 days.

./xmlchange RUN\_STARTDATE=2015-02-01

The MOM6 restart file has the following meta data, where Time is days from year 1.

### C Edit on GitHub

# **Observation Sequence**

### World Ocean Database example

## DA Input: obs\_seq.out

- The observation types in the file
- Locations of the observations
- Quality control value(s) for each observation
- Observation + any observation specific metadata
- Time and order of observations

obs sequence obs\_kind\_definitions 13 **15 FLOAT SALINITY 16 FLOAT TEMPERATURE** 23 GLIDER SALINITY 24 GLIDER TEMPERATURE 27 MOORING SALINITY **28 MOORING TEMPERATURE** 30 BOTTLE\_SALINITY 31 BOTTLE\_TEMPERATURE 32 CTD SALINITY 33 CTD\_TEMPERATURE 43 XBT\_TEMPERATURE 46 APB\_SALINITY 47 APB\_TEMPERATURE num\_copies: 1 num\_qc: num\_obs: 66643 max\_num\_obs: 66643 WOD observation WOD QC first: 1 last: 66643 OBS 2.704750061035156E-002 -1 -1 obdef loc3d 3.753021682990578 1.345058936474565 10.000000000000000 3 kind 32 43322 151163 2.500000000000000E-007 OBS 2 2.704809951782227E-002 0.000000000000000000E+000 3 obdef loc3d 3.753021682990578 1.345058936474565 15.00000000000000 3 kind 32 43322 151163 2.50000000000000E-007 OBS 3 2.704929924011230E-002 2 4 -1 obdef loc3d 3.753021682990578 1.345058936474565 20.00000000000000 kind 32

# **Observation Sequence**

## Radiance example

## DA Output: obs\_seq.final

- various 'copies' of each observation
  - Observation value
  - Truth
  - Prior and Posterior Forward Operators, mean, standard deviation

obs_sequence		
obs_type_definitions		
216 GOES_16_ABI_	RADTANCE	
num_copies:	4 num_qc:	2
	max_num_obs:	77042
observation		
prior ensemble mean		
prior ensemble spread		
prior ensemble member	1	
GOES QC		
DART quality control		
	st: 77042	
OBS 1		
2.54618100000000		
2.67347519169475		
8.220037240577385E-002		
2.70210881859020		
0.000000000000000E+000		
0.000000000000000E+000		
-1 2	-1	
obdef		
loc3d		
4.102617740631104	0.626809656620	0256 35000.0000000000
kind		
216		
visir		
125.829100000000	68.2444100000000	-888888.00000000
-888888.000000000		
4 16	44	2
-888888.000000000		
1		
43072 152809		
6.250000000000000E-002		
OBS 2		
2.61012400000000		
2.72617267485070		
8.868187079127991E-002		
2.73930031819035		
0.000000000000000000000000000000000000		

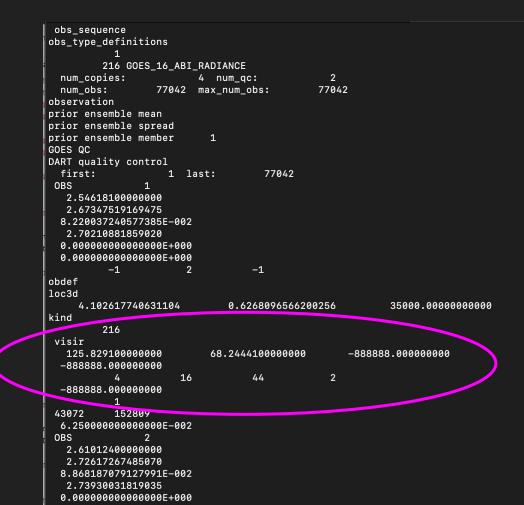
# **Observation Sequence**

## Radiance example

## DA Output: obs\_seq.final

- various 'copies' of each observation
  - Observation value
  - Truth
  - Prior and Posterior Forward Operators, mean, standard deviation

Observations can have additional metadata



# Fortran + Matlab Observation Space Diagnostics

Detailed structure of an obs\_seq file

Creating an obs\_seq file of synthetic observations

Creating an obs\_seq file from real observations

Available observation converter programs

Manipulating obs\_seq files with the obs\_sequence\_tool

The difference between observation TYPE and QUANTITY

Adding support for a new observation TYPE

Radiances

**OBSERVATION CONVERTERS** 

**DART Observations** 

Converter programs

DIAGNOSTICS

Checking your initial assimilation

Computing filter increments

Computing filter increments using a complex model

DART missing data value

plot\_profile.m plots the spatial and temporal average of any specified quantity as a function of height. The

number of observations possible and used are plotted on the same axis.

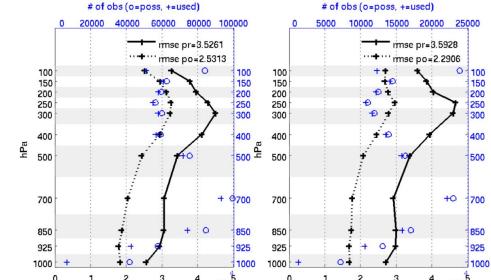
fname = 'POP11/obs\_diag\_output.nc'; copystring = 'rmse'; plotdat = plot\_profile(fname,copystring);



Southern Hemisphere

% netcdf file produced by 'obs diag'

% 'copy' string == quantity of interest



## Fortran + Matlab Observation Space Diagnostics

- 5 !> The programs defines a series of epochs (periods of time) and geographic
- 6 !> regions and accumulates statistics for these epochs and regions.
- 7 !> All 'possible' observation types are treated separately.
- 8 !> The results are written to a netCDF file.
- 9 !> If the rank histogram is requested (and if the data is available),
- 10 !> only the PRIOR rank is calculated.
- 11

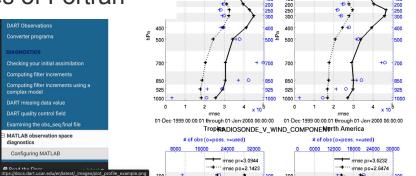
13

12 program obs\_diag

### 4370 lines of Fortran

### Matlab

- not open source
- not freely available



the spatial and temporal average of any specified quantity as a function of height. The tions possible and used are plotted on the same axis.

11/obs\_diag\_output.nc'; e'; \_profile(fname,copystring);

Northern Hemisphere

# of obs (o=poss, +=used)

20000 40000 60000 80000 100000

mse pr=3.5261

+ · · · mse po=2.5313

0 100 100

150 150

; % netcdf file produced by 'obs\_diag' % 'copy' string == quantity of interest ring);

Southern Hemisphere

# of obs (o=poss, +=used)

5000 10000 15000 20000 25000

.... mse po=2.2906

0-100

150

# Fortran + Matlab Observation Space Diagnostics

- !> The programs defines a series of epochs (periods of time) and geographic 5
- 6 !> regions and accumulates statistics for these epochs and regions.
- !> All 'possible' observation types are treated separately.
- 8 !> The results are written to a netCDF file.
- 9 !> If the rank histogram is requested (and if the data is available),
- 10 !> only the PRIOR rank is calculated.
- 11

13

12 program obs diag

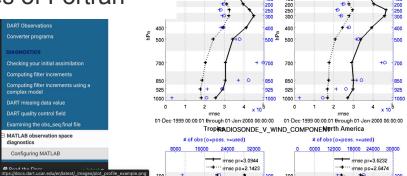
### 4370 lines of Fortran

diagnostics



- not open source
- not freely available





the spatial and temporal average of any specified quantity as a function of height. The tions possible and used are plotted on the same axis.

'11/obs\_diag\_output.nc'; e': \_profile(fname,copystring);

Northern Hemisphere

# of obs (o=poss, +=used)

20000 40000 60000 80000 100000

mse pr=3.5261

+ · · · mse po=2.5313

0 100 100

150 150

200

% netcdf file produced by 'obs\_diag' % 'copy' string == quantity of interest

Southern Hemisphere

# of obs (o=poss, +=used)

Ð-

5000 10000 15000 20000 25000

.... mse po=2.2906

0-100

150

## pyDARTdiags

obs\_ obs\_t

num obser prior prior prior

## obs\_seq.{out/final} -> pandas dataframe

### github.com/NCAR/pyDARTdiags

_sequence type_definitions		🛛 Launcher			×	<b>I</b> o	bs_seq_	_test.ip	ynb >	× 🖪 matt_plot	ipynb	×	+		
22 4 GPSRO_REFRACTIVITY		<b>B</b> + %	Ū	۵ 🕨		C 🕨	Code	le	~						
51 RADIOSONDE_U_WIND_COMPONENT 52 RADIOSONDE_V_WIND_COMPONENT 55 RADIOSONDE_TEMPERATURE 56 RADIOSONDE_SPECIFIC_HUMIDITY 62 AIRCRAFT_U_WIND_COMPONENT 63 AIRCRAFT_V_WIND_COMPONENT 64 AIRCRAFT_TEMPERATURE 66 ACARS_U_WIND_COMPONENT 67 ACARS_V_WIND_COMPONENT 68 ACARS_TEMPERATURE 70 MARINE_SFC_U_WIND_COMPONENT 71 MARINE_SFC_V_WIND_COMPONENT 72 MARINE_SFC_SPECIFIC_HUMIDITY		[1]:	<pre>import sys import os import pandas as pd import numpy as np import plotly.express as px sys.path.append(os.path.abspath("/Users/hkershaw/DART/Projects/Diagnostics/pyDART/ import obs_sequence as dart_os</pre>												
80 SAT_U_WIND_COMPONENT 81 SAT_V_WIND_COMPONENT 83 AIRS_TEMPERATURE		[3]:	obs_seq.df.head()												
84 AIRS_SPECIFIC_HUMIDITY 90 RADIOSONDE_SURFACE_ALTIMETER		[3]:		latitude	ver	tical	vert_un	nit		type	seconds	days	time	ob	
"아프레이어에 가지 않는 것 같아요. 이렇게 가지 않는 것 같아요. 그는 것 같아요. 그는 것 같아요.	2 991025			40.010	239	50.0	pressui (Pa		ACARS	S_TEMPERATURE	75603	153005	2019- 12-01 21:00:03		
rvation r ensemble mean r ensemble spread r ensemble member 1 r ansemble member 2				40.010	239	50.0	pressui (Pa		CARS_U_WIN	ID_COMPONENT	75603	153005	2019- 12-01 21:00:03		

# pyDARTdiags

obs\_seq.{out/final} -> pandas dataframe

- Interactive python Observation Space Diagnostics: language of data science
- User extensible, & allows us to improve obs\_seq format
- Observation preprocessing in be done in python?

### github.com/NCAR/pyDARTdiags

	auno.	cher				×		obs	_seq_te	st.ipyn	ıb	×	🖪 ma	tt_plot	.ipynb
8	+	Х	Ū	Ċ			C	••	Code	~					
	[	<pre>[1]: import sys import os import pandas as pd import numpy as np import plotly.express as px sys.path.append(os.path.abspath("/Users/hkershaw/DART/Project import obs_sequence as dart_os</pre>										Projects			
	[	2]:	: obs_seq = dart_os.obs_sequence('/Users/hkershaw/DART/Projects												rojects/
	[	3]:	<pre>obs_seq.df.head()</pre>												
	[	3]:		latit	ude	ver	rtical	ve	ert_unit					type	seconds
				40	.010	239	950.0	р	oressure (Pa)		ACA	RS_T	EMPERA	ATURE	75603
				40	.010	239	950.0	р	oressure (Pa)	ACAI	RS_U_V	/IND_	СОМРО	NENT	75603

# **Next Steps**

- What diagnostics are people interested in?
- What existing tools can we plug in to?
- Xarray + dask
  - Test case 20 year DART-CAM reanalysis
- pyDARTdiags as a preprocessing tool
- Pipeline of ocean data into CESM-DART

E NCAR / pyDARTdiags			Q   + • 🖸 I1
> Code 💿 Issues 3 🕅 Pull requests 🗔 Discussio	ns 🕞 Actions	Projects	🖽 Wiki 🕕 Security
Collection of software out there kershaw-brown started this conversation in Show and tell	#4		Ec
hkershaw-brown on Apr 23 Maintainer List of software for model, observation comparison:	ed	lited +	Category
MELODIES MONET     https://github.com/NOAA-CSL/MELODIES-MONET     MetPy     https://github.com/Unidata/MetPy			Labels None yet 1 participant
comment	Oldest	lewest Top	Converted from issue This discussion was converted fr issue #3 on April 23, 2024 09:22
hkershaw-brown 2 weeks ago Maintainer Author			Notifications
CUPID: CESM Unified Postprocessing and Diagnostics			∑ Unsubscribe
2024 CESM SEWG Group https://ncar.github.io/CUPiD/			You're receiving notifications because you authored the thread
ADF (CAM)			
https://justin-richling.github.io/ADF-Tutorial/README.html			A Lock conversation → Transfer this discussion
MOM6 tools (dead?)			→ Transfer this discussion ◇ Pin discussion
https://github.com/NCAR/mom6-tools			Pin discussion
IO			<ul> <li>Create issue from discussion</li> </ul>
pyNIO GFDL using this https://www.pyngl.ucar.edu/Nio.shtml			습 Delete discussion
<u>(1)</u> (2)		0 replies	

# **Next Steps**

- What diagnostics are people interested in?
- What existing tools can we plug in to?
- Xarray + dask
  - Test case 20 year DART-CAM reanalysis
- pyDARTdiags as a preprocessing tool
- Pipeline of ocean data into CESM-DART

hkershaw@ucar.edu

