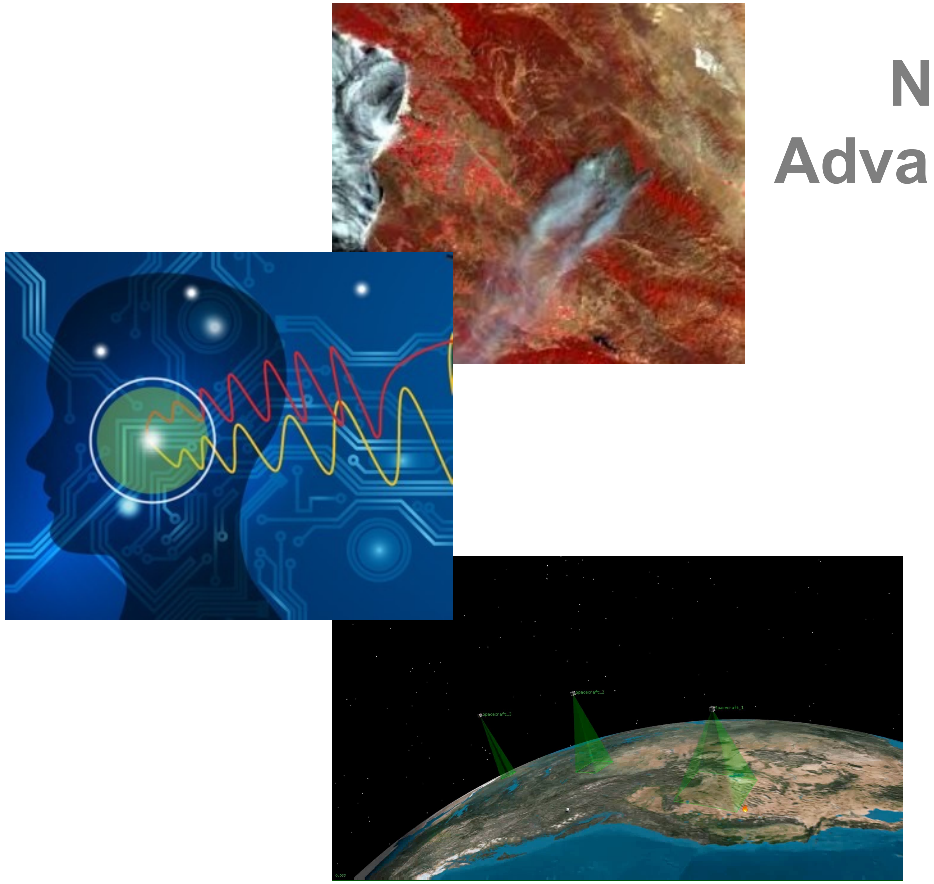


NASA Earth Science Technology Office (ESTO)
Advanced Information Systems Technology (AIST)

Analytics Collaborative Frameworks (ACF)

Annual Technical Reviews

Jacqueline Le Moigne

January 22 & February 5, 2021

Advanced Information Systems Technology (AIST) Program Management Team

“Investment in information systems that NASA Earth Science will need in the 5 to 10-year timeframe”

Jacqueline Le Moigne, Program Manager

Mike Seablom, Senior Strategist

Marge Cole, Outreach and Validation

<u>Associates:</u>
Ian Brosnan, Transitions/Infusions
Laura Rogers, Biodiversity & Ocean
Nikunj Oza, AI & Knowledge Systems
Ben Smith, Autonomy

Jackie Ferguson, Resources Analyst

Bob Connerton, Advisor

Paul Padgett, Communications

NOS and ACF for Science Data Intelligence

Optimize measurement acquisition using many diverse observing capabilities, collaborating across multiple dimensions and creating a unified architecture

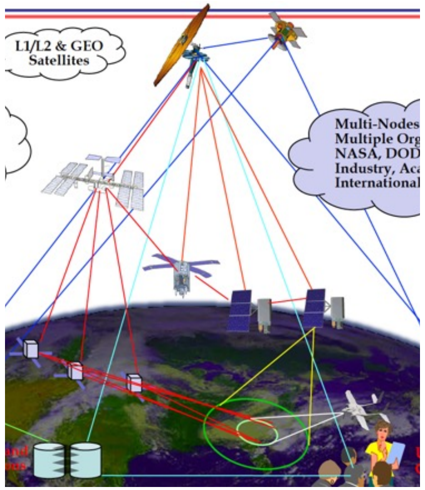
Enhance and enable focused Science investigations by facilitating access, integration and understanding of disparate datasets using pioneering visualization and analytics tools as well as relevant computing environments

Assimilate Observations

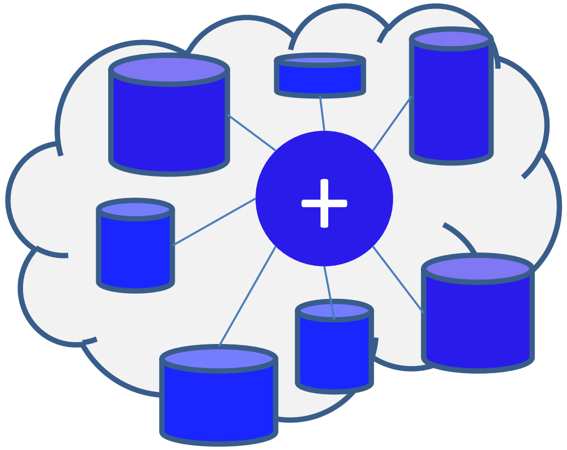
New Observing Strategies (NOS)

Acquire coordinated observations

Track dynamic and spatially distributed phenomena



Analytic Collaborative Frameworks (ACF)



Assimilate many various data into models and analytic workflows.

What additional observations are needed?

Observation Requests

Example: NOS Testbed Demonstration planned for Spring 2021 targeting Mid-West Floods with LIS Models as well as Space and ground observations

Example: OceanWorks, ACF for Ocean Science https://oceanworks.jpl.nasa.gov

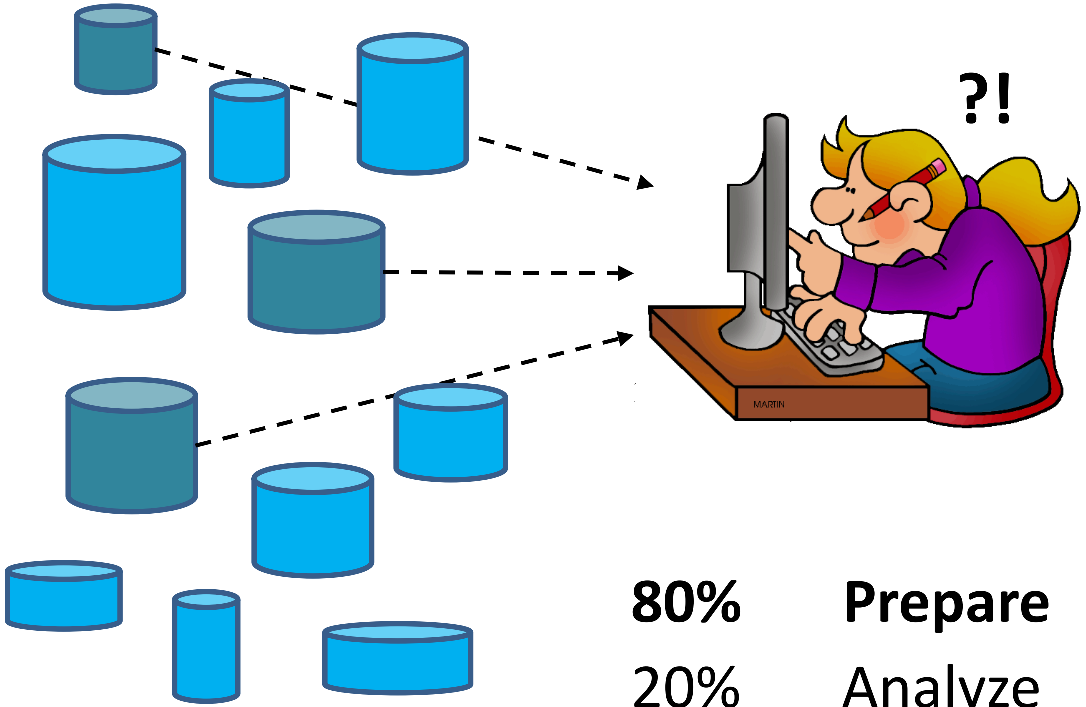
NOS+ACF acquires and integrates complementary and coincident data to build a more complete and in-depth picture of science phenomena

From Archives to Analytic Centers: *Focus on the Science User*

Data Archives

Focus on data capture, storage, and management

Each user has to find, download, integrate, and analyze

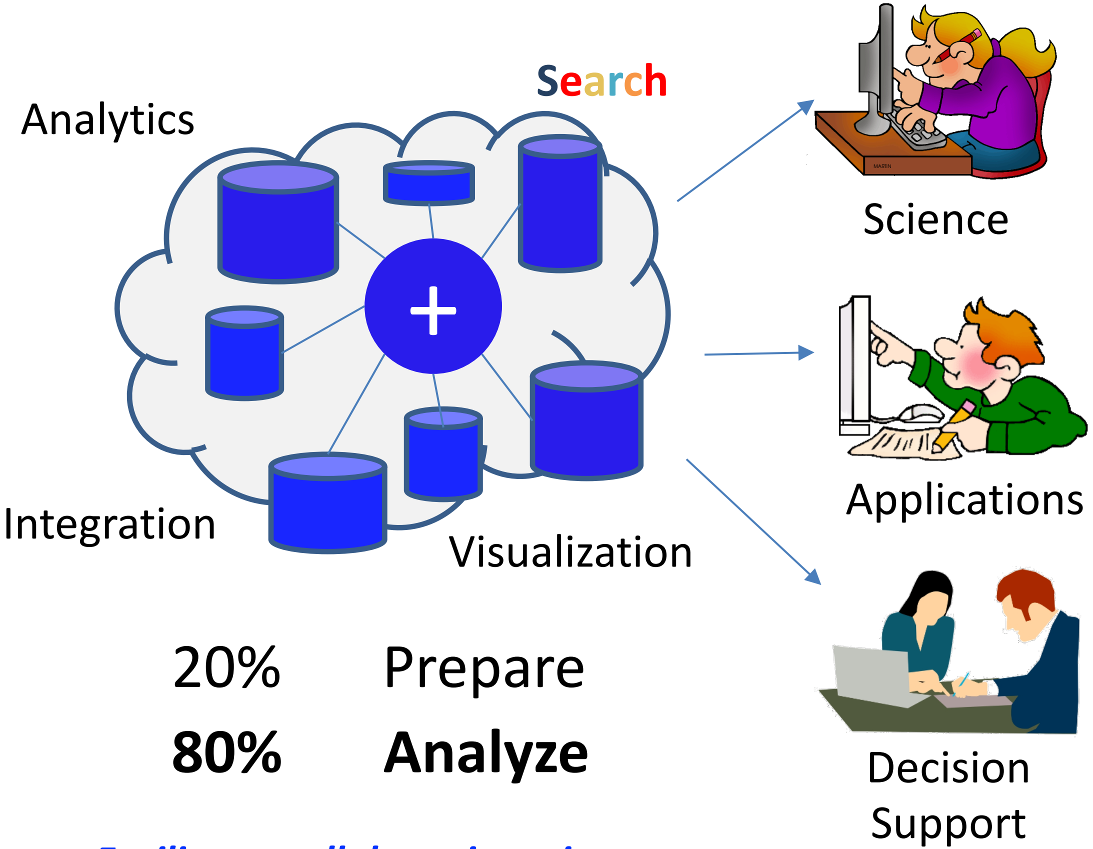


80%	**Prepare**
20%	Analyze

Analytic Centers

Focus on the science user

Integrated data analytics & tools tailored for a science discipline

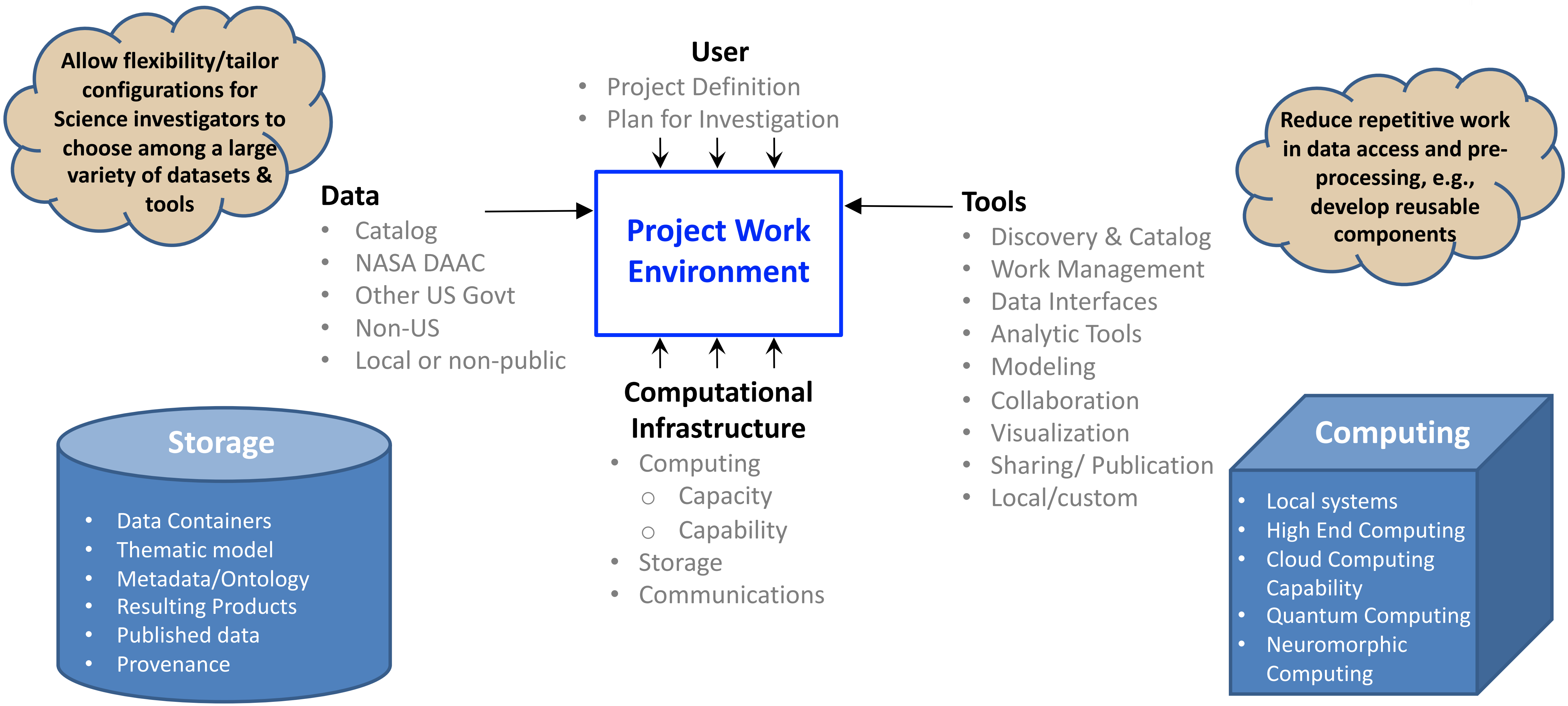


20%	Prepare
80%	**Analyze**

Facilitates collaborative science across multiple missions and data sets

Analytic Collaborative Frameworks (ACF)

Focus is on the Science User



Allow flexibility/tailor configurations for Science investigators to choose among a large variety of datasets & tools

User

- Project Definition
- Plan for Investigation

Reduce repetitive work in data access and pre-processing, e.g., develop reusable components

Data

- Catalog
- NASA DAAC
- Other US Govt
- Non-US
- Local or non-public

Project Work Environment

Tools

- Discovery & Catalog
- Work Management
- Data Interfaces
- Analytic Tools
- Modeling
- Collaboration
- Visualization
- Sharing/ Publication
- Local/custom

Computational Infrastructure

- Computing
  - Capacity
  - Capability
- Storage
- Communications

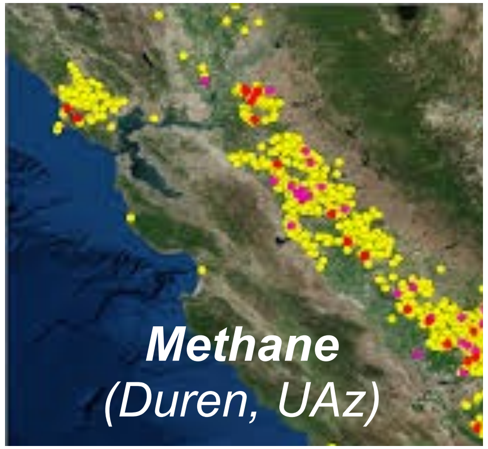
Storage

- Data Containers
- Thematic model
- Metadata/Ontology
- Resulting Products
- Published data
- Provenance

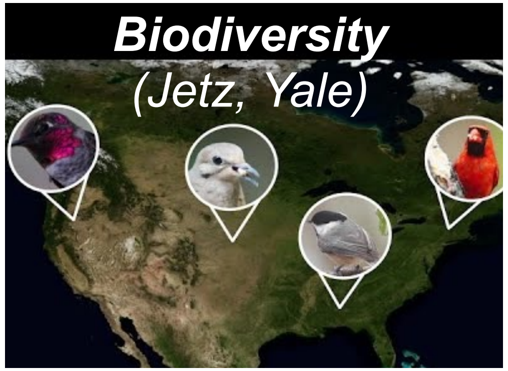
Computing

- Local systems
- High End Computing
- Cloud Computing Capability
- Quantum Computing
- Neuromorphic Computing

Analytic Collaborative Frameworks (ACF) support several Earth Science Disciplines



Methane
(Duren, UAz)



Biodiversity
(Jetz, Yale)

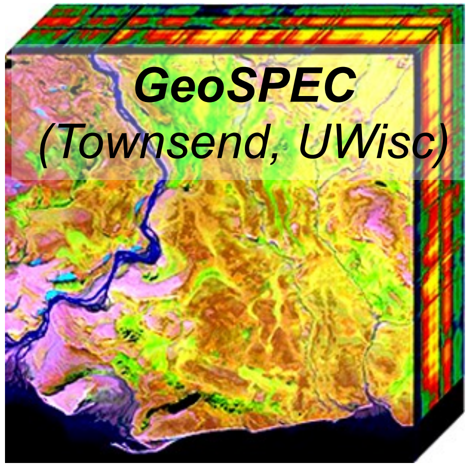
Quakes
(Donnellan, JPL)

Aquaculture
(Uz, GSFC)

Wildfire
(Coen, UCAR)

Air Quality
(Holm, City of LA)

Precipitation
(Beck, UAH))



GeoSPEC
(Townsend, UWisc)

Technologies Currently Being Developed in ACF Projects

ADVANCED ANALYTICS:

- Data Accessibility (Duren, Jetz, Coen)
- Data Fusion (Donnellan, Duren, Jetz, Uz, Coen)
- Big Data Analytics (Hua, Ives, Swenson, Townsend)
- Data Mining (Donnellan)
- On-Demand Product Generation (Hua, Townsend)
- Data Operations Workflows (Zhang)
- Data Incorporation of Metadata, Provenance, Semantics, etc. (Huffer)

AI CAPABILITIES:

- Machine Learning (Beck, Holm, Huffer, Uz)
- Deep Learning (Beck, Holm, Huffer, Uz)
- Data Services Discovery (Zhang)
- Uncertainty Quantification Methods (Ives)

COMPUTATIONAL ENVIRONMENTS:

- Cloud Computing (Beck)

IMPROVED MODELING CAPABILITIES:

- Science Data Model Validation/Automation (Moisan)
- Science Code Development and Reuse (Henze, Moisan)
- Modeling Systems (Martin)
- Model Data Inter-Comparisons (Henze, Swenson)
- Custom Tools (Martin)
- Forecasting/Prediction (Jetz, Swenson, Townsend, Moisan)

ACF Review Schedule – 01/22/2021

January 22nd, 2021	Analytics Collaborative Frameworoks (ACF-Group B) Technical Annual Reviews				
Tech	**Science**	**Name**	**Title**	**Start**	**Stop**
		Le Moigne	Introductions	11:00 AM	11:20 AM
Ceilometers, ML	**PBL**	**Halem**	A Deep Learning LIDAR-based Ceilometer Atmospheric Boundary Layer Height Over CONUS	11:20 PM	12:00 PM
Science Code Development, Model Data Inter-Comparisons	**Atmospheric Composition, Atmos Gas**	**Henze**	Surrogate modeling for atmospheric chemistry and data assimilation	12:00 PM	12:40 PM
Modeling Systems, Custom Tools	**Atmospheric Composition, Atmos Gas**	**Martin**	Development of GCHP to enable improved access to high-res atmospheric modeling	12:40 PM	1:20 PM
Autonomy, ML, Sensor Calibration & Validation	**Atmospheric Composition, Total Ozone and Aerosols**	**Holm**	Predicting What We Breathe: Using ML to Understand Urban Air Quality	1:20 PM	2:00 PM
		Break		***2:00 PM***	***2:20 PM***
Data Fusion, Data Mining	**Earth Surface, Surface deformation**	**Donnellan**	Quantifying Uncertainty and Kinematics of Earthquake Systems (QUAKES-A)	2:20 PM	3:00 PM
Big Data Analytics, On-Demand Products	**Earth Surface, Surface deformation**	**Hua**	Smart On-Demand of SAR ARDs in Multi-Cloud & HPC	3:00 PM	3:40 PM
Data Fusion & Accessibility	**Carbon Cycle, Atmospheric Gas**	**Duren**	Multi-scale Methane Analytic Framework	3:40 PM	4:20 PM
Data Operations Workflows, Data Services Discoverability	**Climate variability, Global / regional climate systems**	**Zhang**	Mining Chained Modules in Analytics Center Framework	4:20 PM	5:00 PM

ACF Review Schedule – 02/05/2021

February 5th, 2021 Analytics Collaborative Framewoks (ACF-Group A) Technical Annual Reviews

Tech	Science	Name	Title	Start	Stop
		Le Moigne	Introductions	11:00 AM	11:20 AM
Data Fusion, Big Data Analytics	Ocean Biology	**Chirayath**	NeMO-Net – The Neural Multi-Modal Observation & Training Network for Global Coral Reef Assessment	11:20 AM	12:00 PM
Autonomy, ML, Data Fusion	Carbon cycle, ocean color	**Schollaert Uz**	Shellfish aquaculture in the Chesapeake bay using AI for water quality	12:00 PM	12:40 PM
Science Data Modeling, Science Code Development	Carbon cycle, ocean color	**Moisan**	NASA Evolutionary Programming Analytic Center (NEPAC)	12:40 PM	1:20 PM
Autonomy, ML, Cloud Computing	Rain Rate, Drop Size, Water & Energy	**Beck**	Cloud-based Analytic Framework for Precipitation Research (CAPRi)	1:20 PM	2:00 PM
Big Data Analytics, Uncertainty Quantification	Carbon cycle, Ecosystems	**Ives**	Statistical tool to analyze large datasets for pattern changes and forecasting	2:00 PM	2:40 PM
		Break		***2:40 PM***	***2:50 PM***
Data Fusion, Data Accessibility	Carbon cycle, Biodiversity	**Jetz**	Biodiversity - Environment Analytic Center Modeling	2:50 PM	3:30 PM
Model Data Intercomparison, Big Data Analytics	Climate variability, bio-diversity	**Swenson**	Canopy condition to continental scale biodiversity forecasts	3:30 PM	4:10 PM
On-Demand Products, Big Data Analytics	Carbon cycle, Biodiversity	**Townsend**	GeoSPEC	4:10 PM	4:50 PM
Autonomy, ML, Metadata	Carbon cycle, Ecosystems	**Huffer**	AMP: An Automated Metadata Pipeline	4:50 PM	5:30 PM

AIST Group Project Review Objectives

- **Regular Annual Reporting Requirements**
  - Individual Programmatic Annual Reviews
  - Technical Annual Reviews Grouped by Topics
- **Establish relationship between awardees**
  - Introduce AIST PIs and their work to one another
  - Enable desired collaborations
  - Potentially share algorithms, codes or cross-cutting ideas
  - GoogleDocs:
    https://docs.google.com/document/d/1CvmgehHflwqDoTKtmrq7bdCm7NMY30bh1u2cHpIv5g8/edit?usp=sharing
- **Present AIST-18 Projects and PIs to broader community**
  - Present AIST-18 projects to NASA ESD Program Managers and partner organizations
  - Support technology infusions and knowledge transfer of AIST projects upon completion.
- **Review Needs in terms of:**
  - ESIP: Project analysis to improve infusion and transition opportunities
  - SMCE (NASA Science Managed Cloud Environment): AWS system access

ESIP
Earth Science Information Partners

Image Credit: NASA

Image Credit: National Geographic

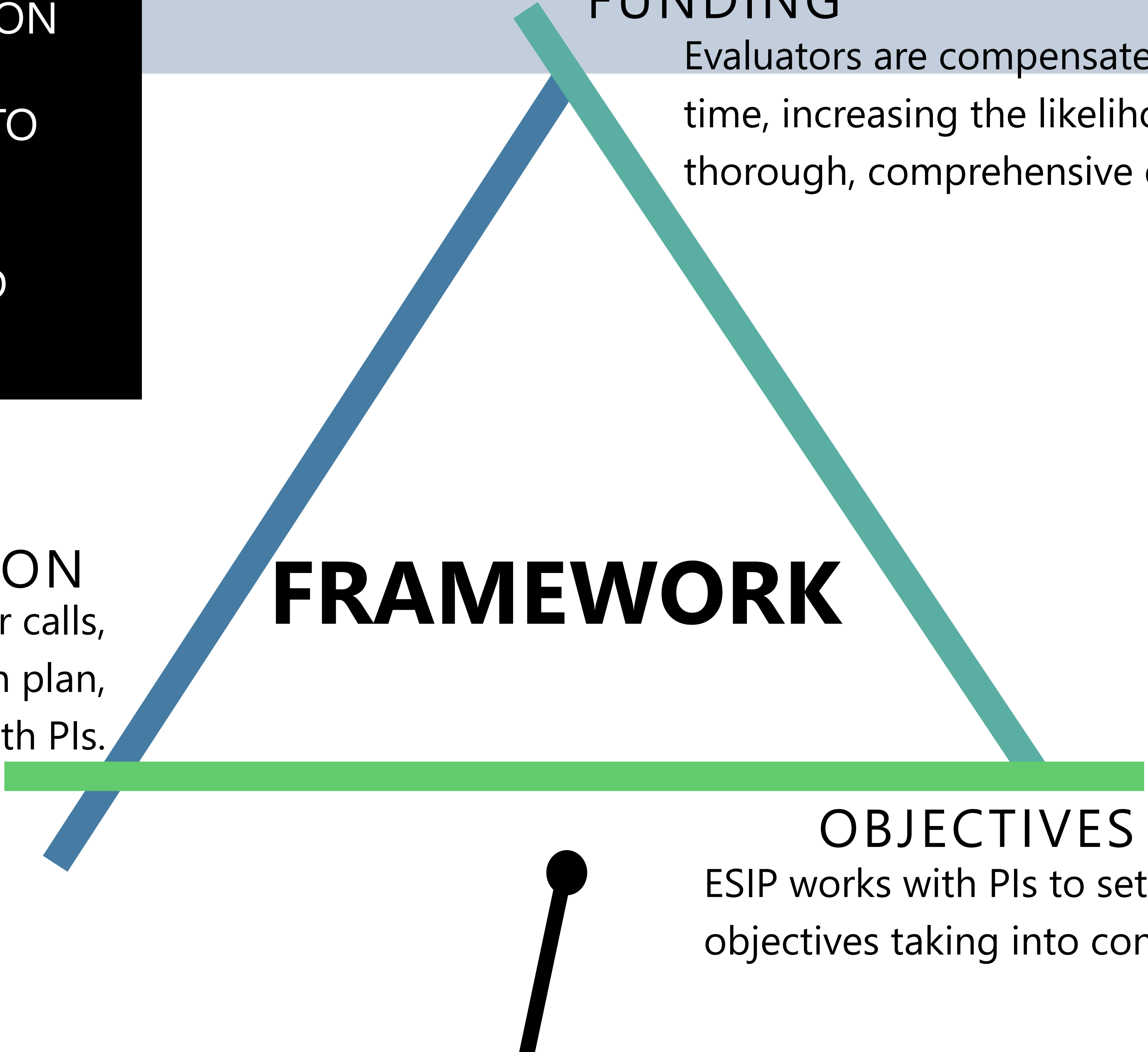
ESIP TECH EVALUATION

ANNIE BURGESS, PHD

AIST | Jan 22, 2021

Supported by:

ESIP PROVIDES AN EVALUATION FRAMEWORK THAT EXPOSES DEVELOPING TECHNOLOGY TO POTENTIAL END-USERS AND ADOPTERS, ULTIMATELY INCREASING ITS UTILITY AND USABILITY.



FRAMEWORK

FUNDING

Evaluators are compensated for their time, increasing the likelihood of a thorough, comprehensive evaluation.

FACILITATION

ESIP facilitates evaluator calls, development of evaluation plan, communication with PIs.

OBJECTIVES

ESIP works with PIs to set specific objectives taking into consideration TRL.

TECHNICAL EXCHANGE MEETING

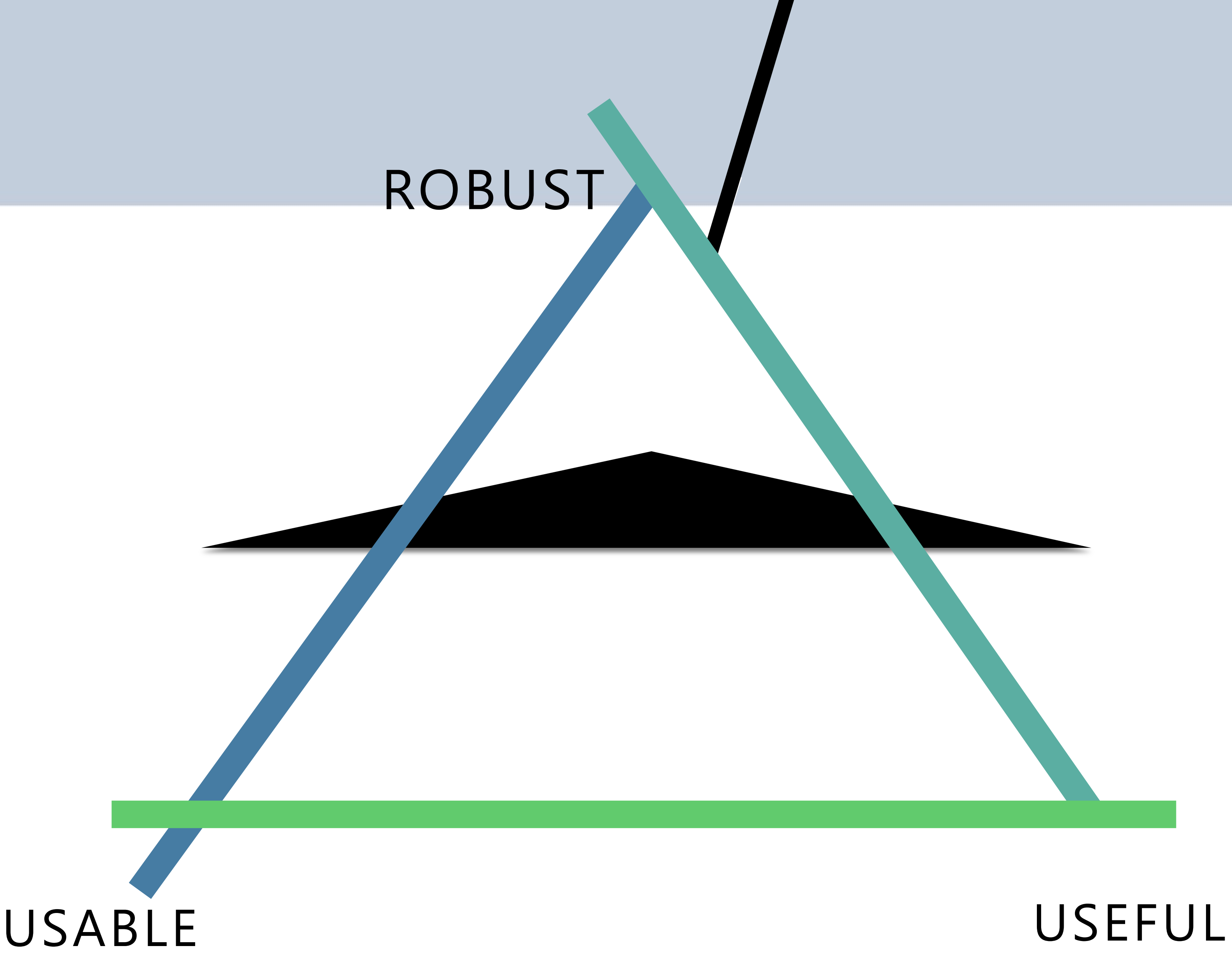
PI team meets evaluators. Big picture to backend... evaluators should have a solid understanding of the purpose and goals of tech.

EVALUATION PERIOD

ESIP coordinates evaluation process. Evaluators meet regularly, requesting information from PIs when necessary.

FINAL REPORT

ESIP works with evaluators to create final report to be shared with PIs & AIST. Reports can be public upon PI request.

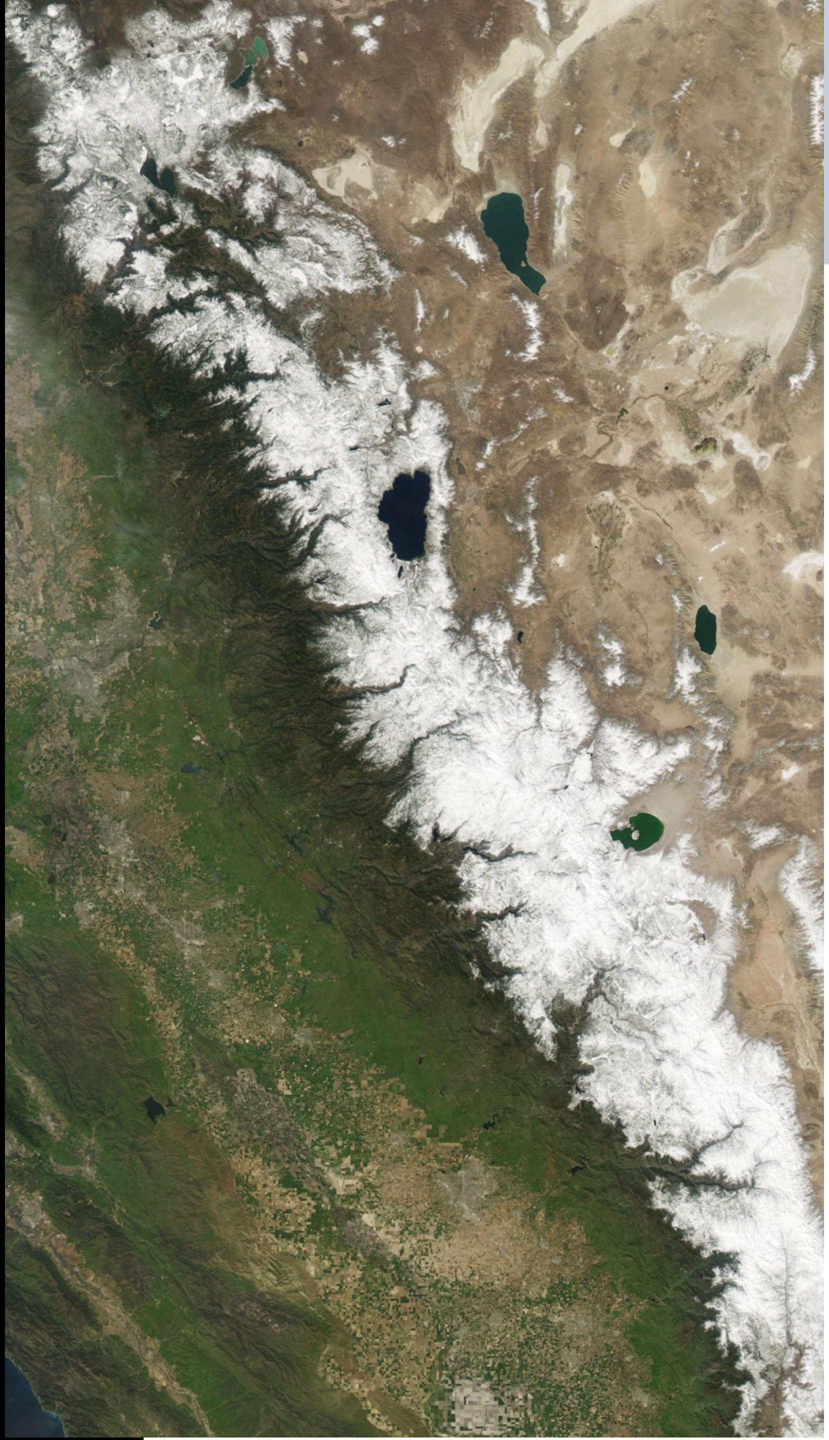

ROBUST
USABLE
USEFUL

ESIP
Earth Science Information Partners
esipfed.org | #ESIPfed

THANK YOU

ANNIE BURGESS, PHD

ANNIEBURGESS@ESIPFED.ORG



ESIP is supported by:

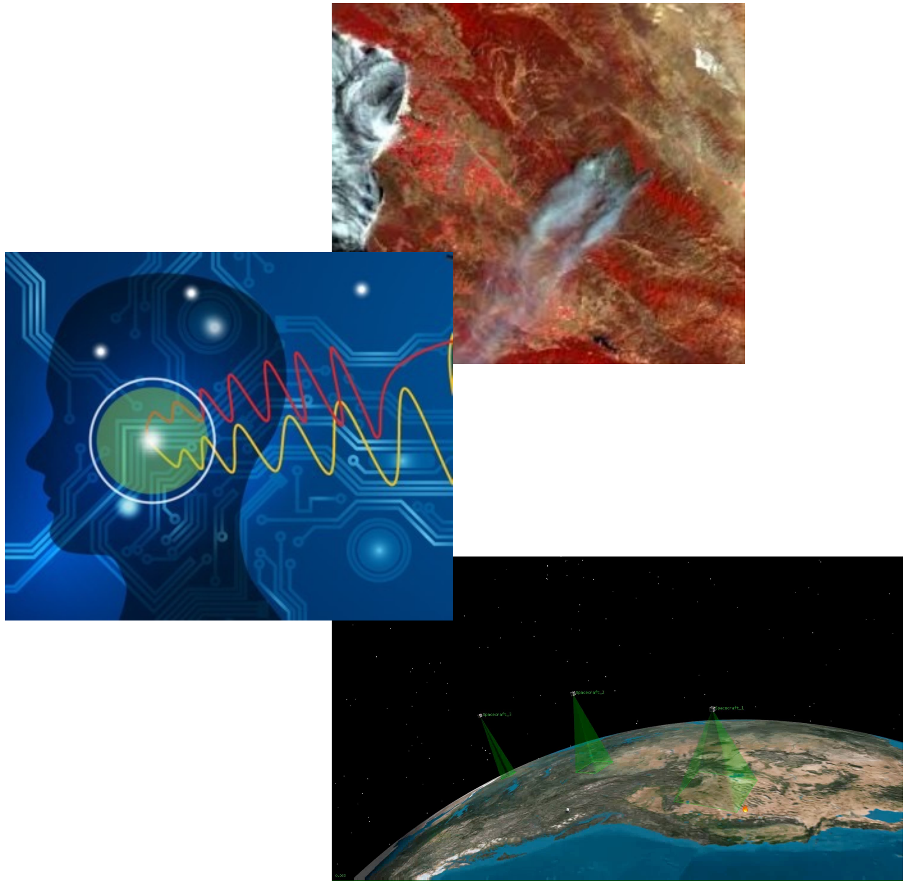
AIST SMCE Options
Marge Cole

- A critical component of the success of AIST projects is access to cost effective, flexible, and scalable compute and storage infrastructure.

- The Science Managed Cloud Environment (SMCE) is a managed Amazon Web Service (AWS) based infrastructure for NASA funded projects that can leverage cloud computing capabilities. This environment is designed to:
  - Provide cloud access to NASA PIs with non-NASA team members.
  - Perform research using new computing capabilities without extensive start-up time.
  - Use new tools and methods from AWS's product catalogue easily and affordably.
  - Scale computing for high-demand, high-bandwidth needs.

- More information at: https://www.nccs.nasa.gov/systems/SMCE

- ***NASA Managed (AWS) Cloud Environment Access***
  - Pay-as-you-go cloud account access with NASA security already built in
  - Enables ease of cloud-based project transition to NASA programs due to NASA level security already requirements already being met.



PI's Introductions

Around the Virtual Room

Towards an R2O Deep Machine Learning Hourly Boundary Layer Height Visualization Product over CONUS from Ceilometer and Satellite based Lidar Aerosol Backscatter

PI M.Halem, CO-PI B. Demoz, CO-Is, P. Nguyen, J. Sleeman,
V. Caicedo, R. Delgado, D. Chapman, Z.Yang,
J. Dorband, P. Gentine

AIST Technical Review (Virtual)

NNH16ZDA001 AIST-16-0091
Supplement P0011(9/22/20-11/30/21)
Jan. 22, 2021

halem@umbc.edu

Presentation Overview

- Prototyping a Ceilometer/Satellite LIDAR backscatter streaming acquisition network.
- AI/ML LIDAR and Model Validated Atmospheric Boundary Layer Height (ABLH)
- Fused Visualization and Aerosol Backscatter Data Archive
- Next Steps and Summary:

  (i) Fully test a Secure, Fault Tolerant, Edge Streaming, Reliable ABLH Network

  (ii) Evaluate NU-WRF-CHEM-GOCART ABLH Data Assimilation

  (iii) Train a NAS[1] AI Emulator for NU-WRF-CHEM parameterizations

  (iv) Embed Deep HED[2] in GOCART/Microphysics and fuse ABLH with PBLH

[1] Neural Architecture Search

[2] Hierarchical Edge Detector

A Deep Learning Ceilometer (LIDAR)-based Atmospheric Boundary Layer Height Product Over CONUS

M. Halem, B. Demoz, UMBC

Objectives:

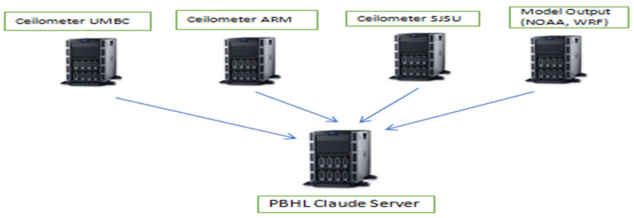
Task 1: Identify, acquire and implement an internet, edge streaming, secure, fault-tolerant ingest L1 system of Ceilometer/Satellite and Model-based LIDAR backscatter observations over the CONUS to generate L2 ABLH products.

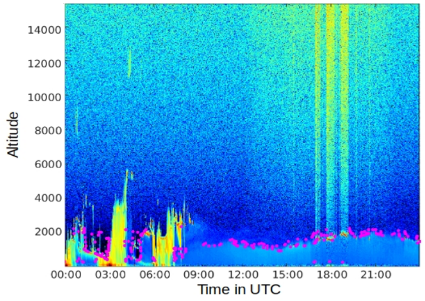
Task 2: Develop and test automated synchronized hybrid L2 ABLH LIDAR processing system for continental wide US profiles combining Machine Learning, Wavelets and Mixture of Experts to generate hourly product with validating error bounds.

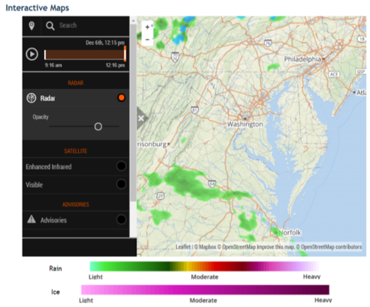
Task 3: Generate point wise, regional and CONUS wide 3-D hourly visualization and longer-term animations. Provide data management, archival and community delivery system of LIDAR Level 1, 2 and 3 products

Task 4: Conduct model output and radiosonde acquisition system for product validation and verifications.

Task 5. Produce quarterly reports and conduct semi-annual reviews and convene external advisory group for system evaluation. .







Approach

- Data Acquisition plan.
  Integrate 4 JCET +3 CSEE ceilometers into automatic data ingest system
- Develop a hybrid machine learning processing system for generating hourly ABLH. Provide Project ATBD or on Giuthub for processing system.
- Validate v1.0 performance and accuracies during op'ns test.
  Identify areas for Improving edge detection method. Continue evaluation of v1.0 methods Add denoising method
  Integrate the LSTM method with the boundary detection method.
- PBLH spatial Visualization.
  Create ABLH spatial maps and dynamic visualizations.
  Fuse UMBC hrly ceilometer ABLH with NOAA PBLH forecast.

Key Milestones

-Acquire 3 NASA Luft ceilometers, install at VA Tech, Bristol PA and NTU ceilometers and conduct 1st end-to-end system test (**10**) of edge streaming ground system Level 1/2/3 operations. 6/21

-Conduct 2nd level 1/2/3 end-end test with ground/satellite and model generated backscatter data in near real time. 9/21

-Produce a robust Ceilometer web-based ABLH hybrid machine learning based system scalble to processing streaming 5-minute data from more than 100 ceilometer stations. 11/21

- Provide a visualization service of PBLH products and generate spatial hourly plots with Zoom capabilities 9/21

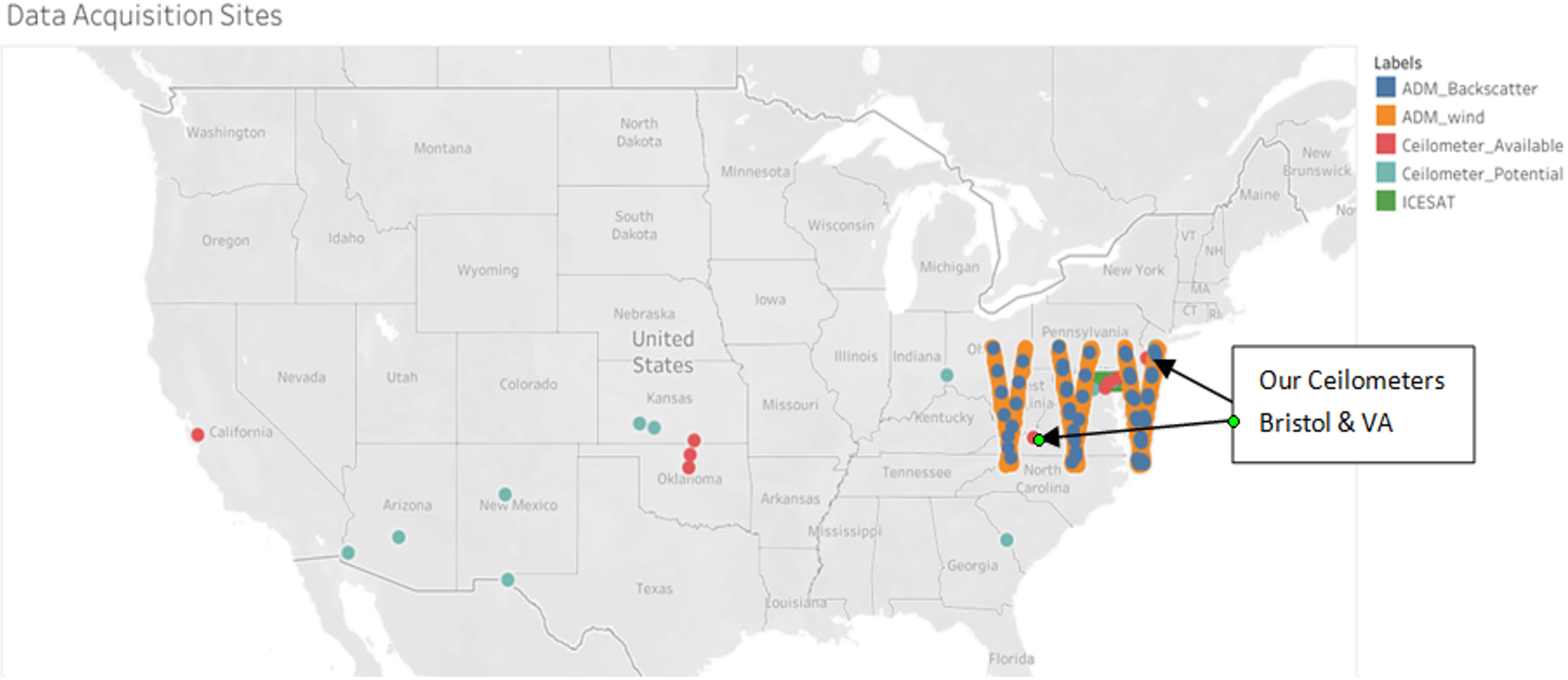
- Demonstrate fault tolerant, secure, edge streaming 2- week end-to end validated test of the unified hybrid ground/space/model AI/ML generation of Regional ABLH web accessible surface 11/21

03/18 AIST-16-00XX

TRL$_{in}$= 5 **TRL$_{fin}$ = 7**

Task 1. Aggregate Acquisition Processing System
P. Nguyen, R. Gite, S. Shivadekar

Task 1: Identify, negotiate, acquire and implement an internet based distributed edge streaming computing system of Level 1 ground-based ceilometer LIDAR PBLH observations over the CONUS.



Our Ceilometers
Bristol & VA

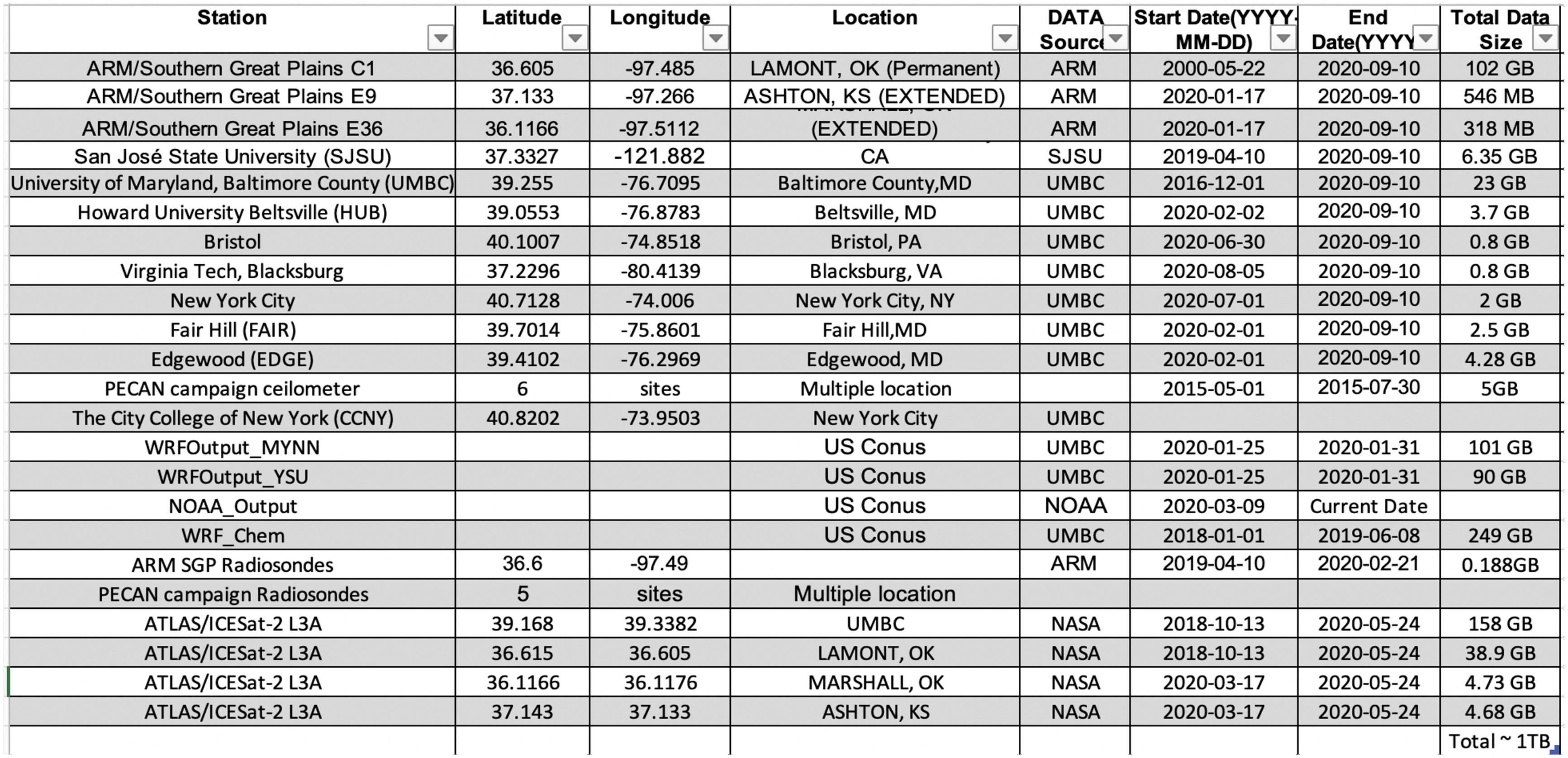
Data source: Ceilometers/Radiosondes from AIST/CSEE grant, UMBC/JCET, DOE/ARM, San Jose University

Model Output: from NOAA (HRRR hourly 13km), our WRF model output.

Satellites IceSat-2 backscatter, ADM wind, backscatter radiation

Ceilometers and Model output

- 2 Ceilometers from grant at Bristol and VA Tech started operation June, August 2020
- **4 Ceilometers from JCET UMBC (1 ceil ~4 years of data from UMBC)**
- 3 Ceilometers from ARM SGP (1 ceil ~20 years from ARM/OK 2011-now, 2 ceil started Jan 2020-now)
- Field campaign PECAN sites 2015 Ceilometers, Radiosondes
- NASA Icesat-2, ESA Aeolus ADM?, Model Output: from NOAA (HRRR hourly at 3km), WRF-CHEM-GOCART



Station	Latitude	Longitude	Location	DATA Sourc	Start Date(YYYY-MM-DD)	End Date(YYYY	Total Data Size
ARM/Southern Great Plains C1	36.605	-97.485	LAMONT, OK (Permanent)	ARM	2000-05-22	2020-09-10	102 GB
ARM/Southern Great Plains E9	37.133	-97.266	ASHTON, KS (EXTENDED)	ARM	2020-01-17	2020-09-10	546 MB
ARM/Southern Great Plains E36	36.1166	-97.5112	MARSHALL, OK (EXTENDED)	ARM	2020-01-17	2020-09-10	318 MB
San José State University (SJSU)	37.3327	-121.882	CA	SJSU	2019-04-10	2020-09-10	6.35 GB
University of Maryland, Baltimore County (UMBC)	39.255	-76.7095	Baltimore County,MD	UMBC	2016-12-01	2020-09-10	23 GB
Howard University Beltsville (HUB)	39.0553	-76.8783	Beltsville, MD	UMBC	2020-02-02	2020-09-10	3.7 GB
Bristol	40.1007	-74.8518	Bristol, PA	UMBC	2020-06-30	2020-09-10	0.8 GB
Virginia Tech, Blacksburg	37.2296	-80.4139	Blacksburg, VA	UMBC	2020-08-05	2020-09-10	0.8 GB
New York City	40.7128	-74.006	New York City, NY	UMBC	2020-07-01	2020-09-10	2 GB
Fair Hill (FAIR)	39.7014	-75.8601	Fair Hill,MD	UMBC	2020-02-01	2020-09-10	2.5 GB
Edgewood (EDGE)	39.4102	-76.2969	Edgewood, MD	UMBC	2020-02-01	2020-09-10	4.28 GB
PECAN campaign ceilometer	6	sites	Multiple location		2015-05-01	2015-07-30	5GB
The City College of New York (CCNY)	40.8202	-73.9503	New York City	UMBC			
WRFOutput_MYNN			US Conus	UMBC	2020-01-25	2020-01-31	101 GB
WRFOutput_YSU			US Conus	UMBC	2020-01-25	2020-01-31	90 GB
NOAA_Output			US Conus	NOAA	2020-03-09	Current Date	
WRF_Chem			US Conus	UMBC	2018-01-01	2019-06-08	249 GB
ARM SGP Radiosondes	36.6	-97.49		ARM	2019-04-10	2020-02-21	0.188GB
PECAN campaign Radiosondes	5	sites	Multiple location				
ATLAS/ICESat-2 L3A	39.168	39.3382	UMBC	NASA	2018-10-13	2020-05-24	158 GB
ATLAS/ICESat-2 L3A	36.615	36.605	LAMONT, OK	NASA	2018-10-13	2020-05-24	38.9 GB
ATLAS/ICESat-2 L3A	36.1166	36.1176	MARSHALL, OK	NASA	2020-03-17	2020-05-24	4.73 GB
ATLAS/ICESat-2 L3A	37.143	37.133	ASHTON, KS	NASA	2020-03-17	2020-05-24	4.68 GB
							Total ~ 1TB

Earth Science Technology Office

Task 1: Ceilometers Acquisition R. Delgardo

- **Procurement of a 3rd Lufft Ceilometer as part of Augmentation (November 2020)**

- **Verification/Validation of Ceilometer Operability at UMBC (Jan. 2021)**
  Instruments evaluation at UMBC before deployment:
  - Signal to noise
  - Overlap factor
  - VPN Data transmission

- **Deployment (**Locations Under Consideration**)**
  South and Southwestern US
  - Dust Storms (Southern Texas/New Mexico/Arizona)
  - Smoke from Agricultural Fires in Central America/Mexico
    *Navajo Technical University
  - Field Campaigns: TRACER (Houston 2021) @ *NASA TOLNET sites

Task 1: NASA Grant Ceilometer Deployment
R. Delgado

- **Locations**

1- Pennsylvania Department of Environmental Protection

Bristol Air Quality Monitoring Station

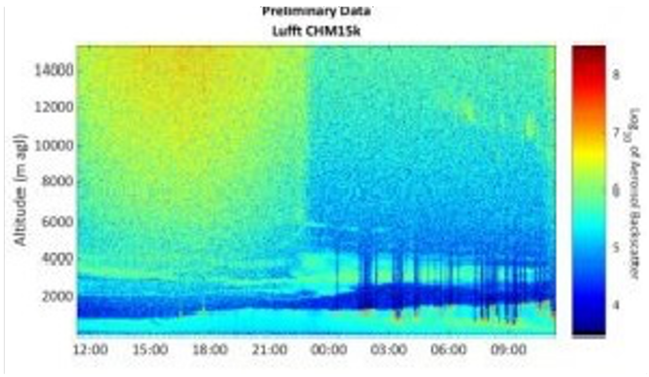
2- Virginia Tech (Elena Lind)

Ceilometer Aerosol Profiling (PBLH) to aid PANDORA profiling retrievals

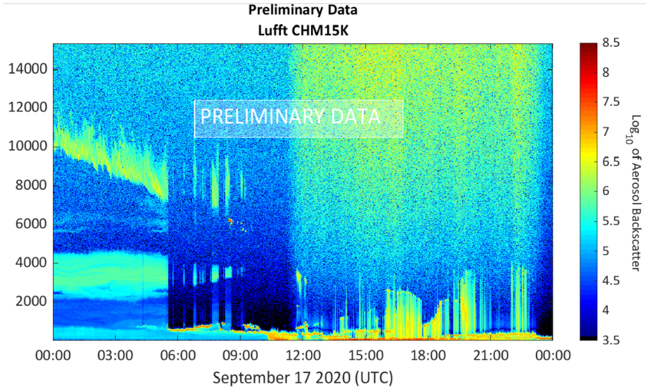
3- Navajo Technical University

Integration of Remote Sensing to Computer/Environmental Science

California Wildfire Smoke
September 17, 2020



Bristol, PA

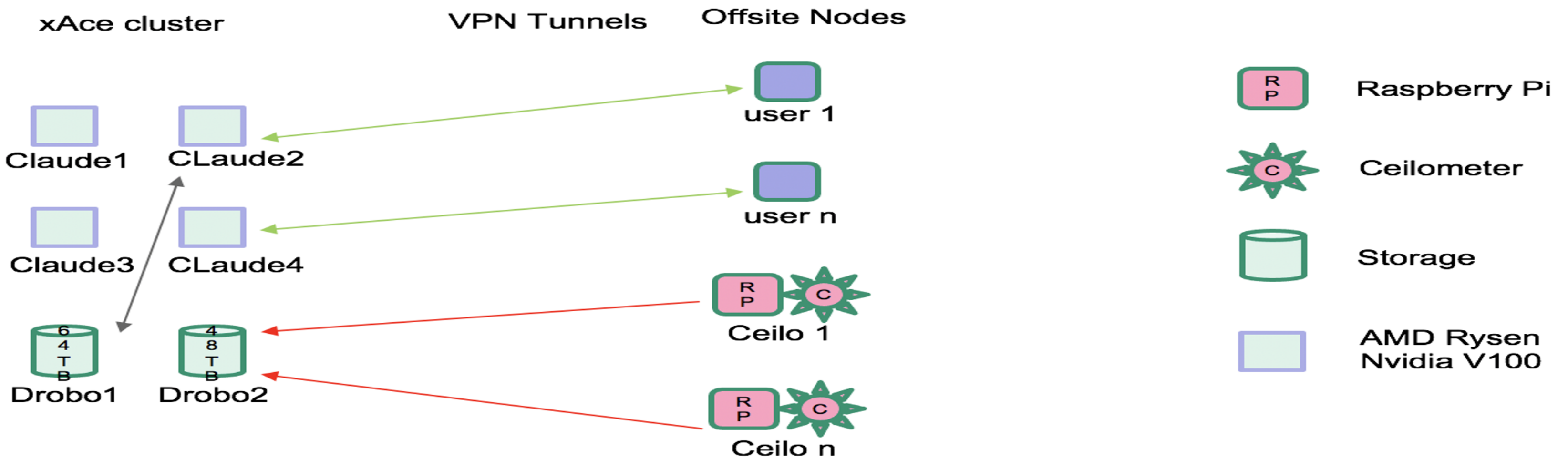


Blacksburg, VA

https://www.timeanddate.com/weather/@5181688/historic?month=9&year=2020

xAce Intra-net
J. Dorband

- Intra-net Security
  - VPN access security (user unique certificate & password)
  - Node security (VPN access & user unique password)
  - Connections in are secure
  - Connections out are open
- Once connected to Intra-net:
  - Access from any machine to any other machine with valid user account
    - User workstation (laptop, desktop)
    - Compute nodes
    - Instrument node
- Instrument node (~ $50 Raspberry Pi)
  - Local data backup from instrument (up to 3 yrs)
  - Periodically passes data on to xAce cluster database
  - Can send data to other offsite nodes/organizations (future)



xAce Hardware Infrastructure

D. Chapmanand J.Dorband

Creation and expansion of Hardware compute infrastructure for Aerosol processing

Claude 1&2 servers ($4K)

- Dual 14.2 Teraflop Nvidia Geforce 2080Ti CUDA capable GPUs (~Nvidia V100)
- 32 Core AMD Ryzen Threadripper 2990wx (~Epyc) 3.0 GHz CPU

Claude 3&4 server (under acquisition) ($6K)

- Dual 36 Teraflop Geforce 3090 CUDA GPUs (~A100)
- 24 Core AMD Ryzen Threadripper 3960X 3.8 Ghz
- 10 Gigabit ethernet NIC and router

Drobo storage configuration ($5K)

- 96 Terabyte Network Attached Storage

xAce Claude and Drobo servers

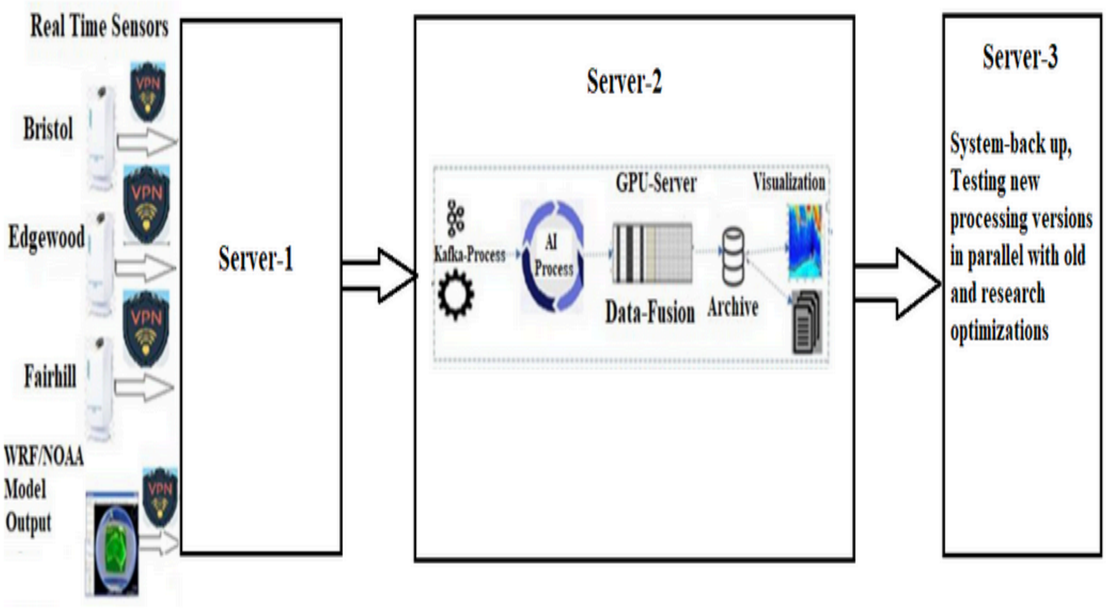
Prototype an Edge Streaming, Secure, Fault Tolerant Automated Data Ingestion and Processing System

- **Develop Data Ingestion** to collect multiple data sources Ceilometers (9 ceilometers) from different organizations, multiple Satellite instruments (ICESat-2, ESA's ADM-Aeolus), Operational model output data products from NOAA.

- **Pulling the data** from NOAA's Model Output (PBLH, HRRR hourly product) and 3 ceilometers data from ARM SGP (automatically)

- **Building Ingestion Server:** use Apache Kafka handles streams of data from multiple ceilometers automatically and backing up pre-processing (raw ceilometer profiles) Level 1B daily data products.

- **Distributed cluster of GPU Servers:** train AI models and process pipelines for improving scalability and throughput and reduce latency.



Task 1 Summary

Current:

- **Deployed/operational 2 Ceilometers and 3rd on order**
- Developed pilot edge streaming, fault tolerant aerosol preprocessing system
- **Ingested, Archived** Ceilometers Level 0 instrument profiles. Produced Level 1B backscatter daily data products from Level 0 backscatter profile from Ceilometers, Satellite Lidars and NOAA's Model Output
- **End to End tested** Data ingesting, Data Preprocessing, ML workflow using Apache Kafka stream automatically.

Plans:

- Continue End to End System Data Acquisition, Preprocessing, ML, Production of L2 ABLH data product, Data Archive and Web Retrieval Services
- **Develop Edge Streaming AI system** using a cluster of GPUs Server for increasing throughput and scalability to ingest and process multiple Ceilometers/Satellite data.

- **Deploy/Acquire** additional Ceilometer data
- **Acquire/Evaluate** the ingest of Satellite Lidar aerosol backscatter
- **Request/Ingest** additional Ceilometer data from other organizations(EPA/ESA)

Task 2. Using Machine Learning to Identify PBL Heights

Dr's Jennifer Sleeman, Vanessa Caicedo and Dorsa Ziaei

Operational method used to estimate and predict heights given what has been learned from past PBL height identification, station location, ceilometer type, and model variables.

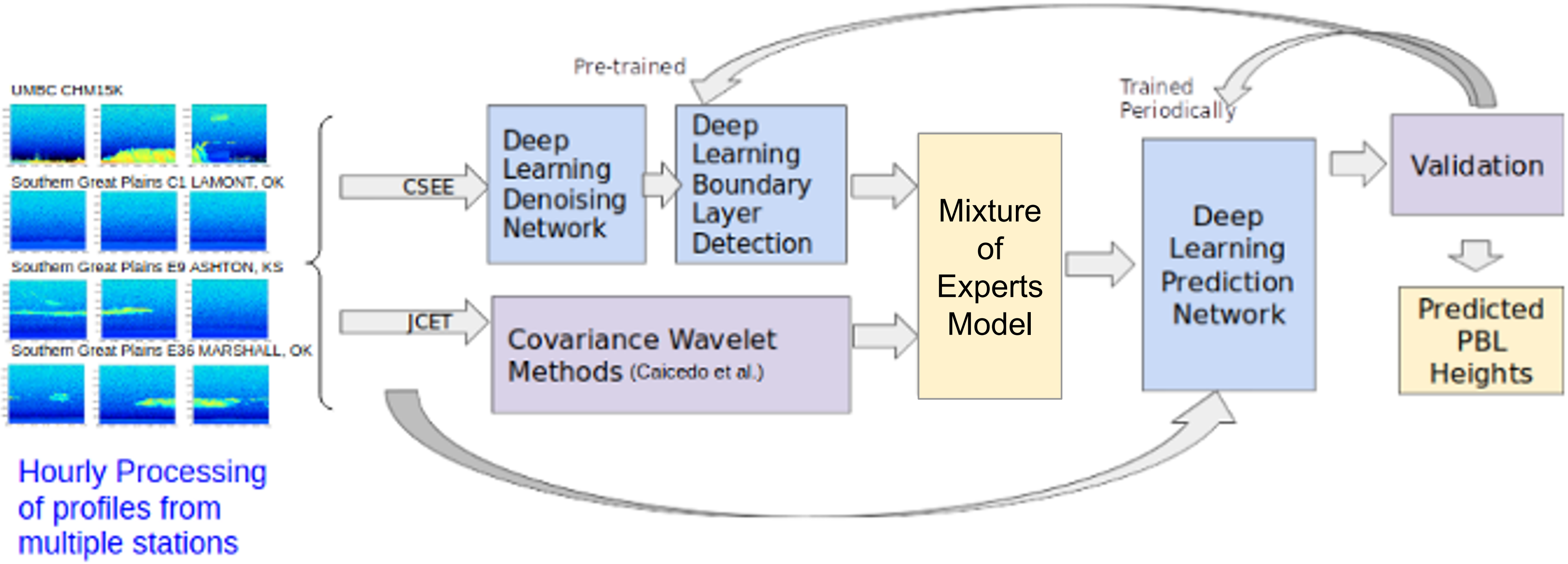
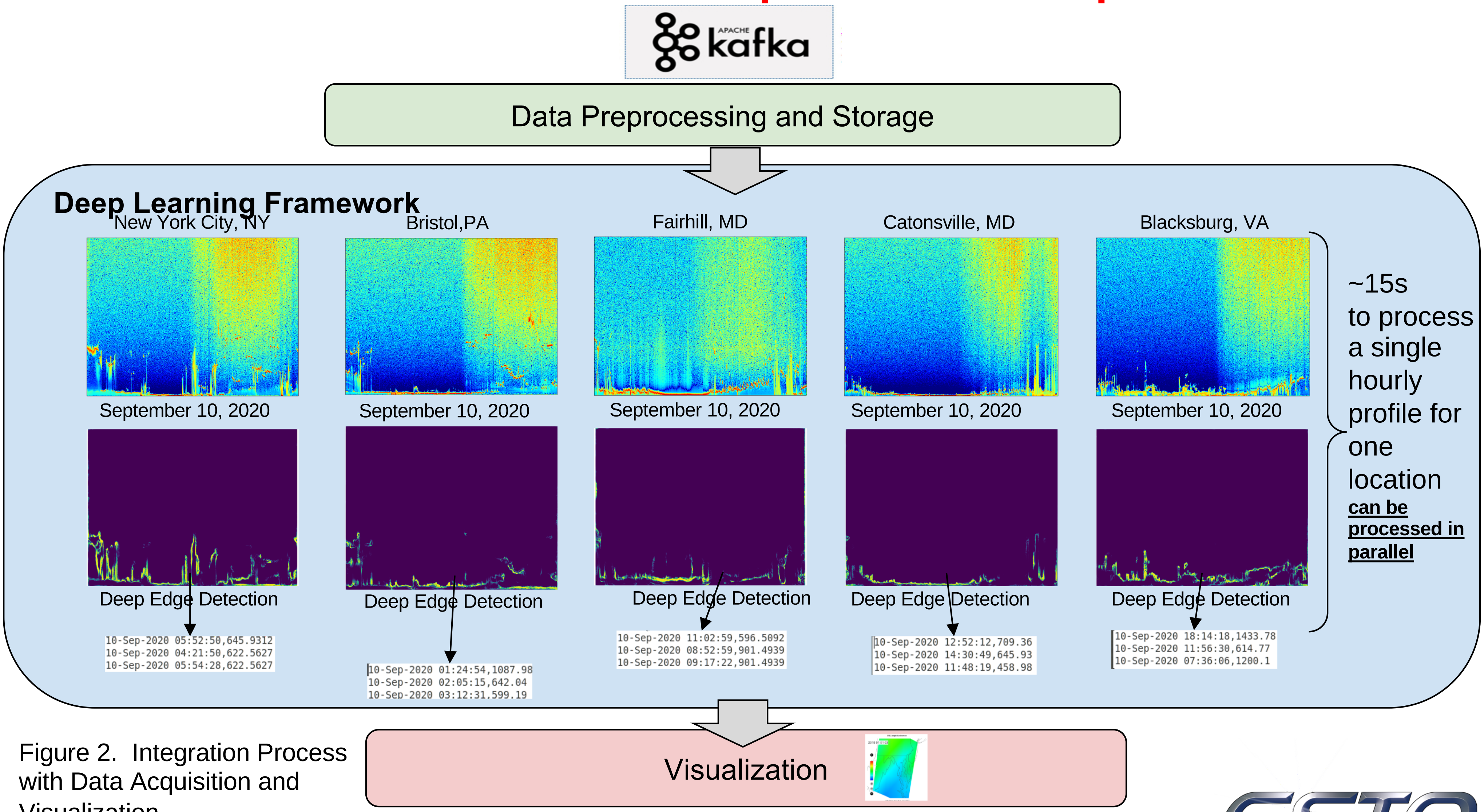

UMBC CHM15K
Southern Great Plains C1 LAMONT, OK
Southern Great Plains E9 ASHTON, KS
Southern Great Plains E36 MARSHALL, OK
Hourly Processing of profiles from multiple stations
CSEE
JCET
Pre-trained
Deep Learning Denoising Network
Deep Learning Boundary Layer Detection
Covariance Wavelet Methods (Caicedo et al.)
Mixture of Experts Model
Trained Periodically
Deep Learning Prediction Network
Validation
Predicted PBL Heights

Figure 1. Integration of Machine Learning Methods for Operational PBLH

Multi-Station Processing and Integration

We have integrated multi-station processing with the data acquisition team and visualization team for a September 10th experiment.



APACHE kafka

Data Preprocessing and Storage

Deep Learning Framework

New York City, NY

September 10, 2020

Deep Edge Detection

```
10-Sep-2020 05:52:50,645.9312
10-Sep-2020 04:21:50,622.5627
10-Sep-2020 05:54:28,622.5627
```

Bristol,PA

September 10, 2020

Deep Edge Detection

```
10-Sep-2020 01:24:54,1087.98
10-Sep-2020 02:05:15,642.04
10-Sep-2020 03:12:31,599.19
```

Fairhill, MD

September 10, 2020

Deep Edge Detection

```
10-Sep-2020 11:02:59,596.5092
10-Sep-2020 08:52:59,901.4939
10-Sep-2020 09:17:22,901.4939
```

Catonsville, MD

September 10, 2020

Deep Edge Detection

```
10-Sep-2020 12:52:12,709.36
10-Sep-2020 14:30:49,645.93
10-Sep-2020 11:48:19,458.98
```

Blacksburg, VA

September 10, 2020

Deep Edge Detection

```
10-Sep-2020 18:14:18,1433.78
10-Sep-2020 11:56:30,614.77
10-Sep-2020 07:36:06,1200.1
```

~15s to process a single hourly profile for one location **can be processed in parallel**

Visualization

Figure 2. Integration Process with Data Acquisition and Visualization

Deep Boundary Detection and Wavelet PBL Height Retrieval Hybrid Method - Mixing Layer Heights Correlation Matrix

The Mixture of Experts method could add 8 additional MLH measurements for the December 2016 Ad-hoc campaign.

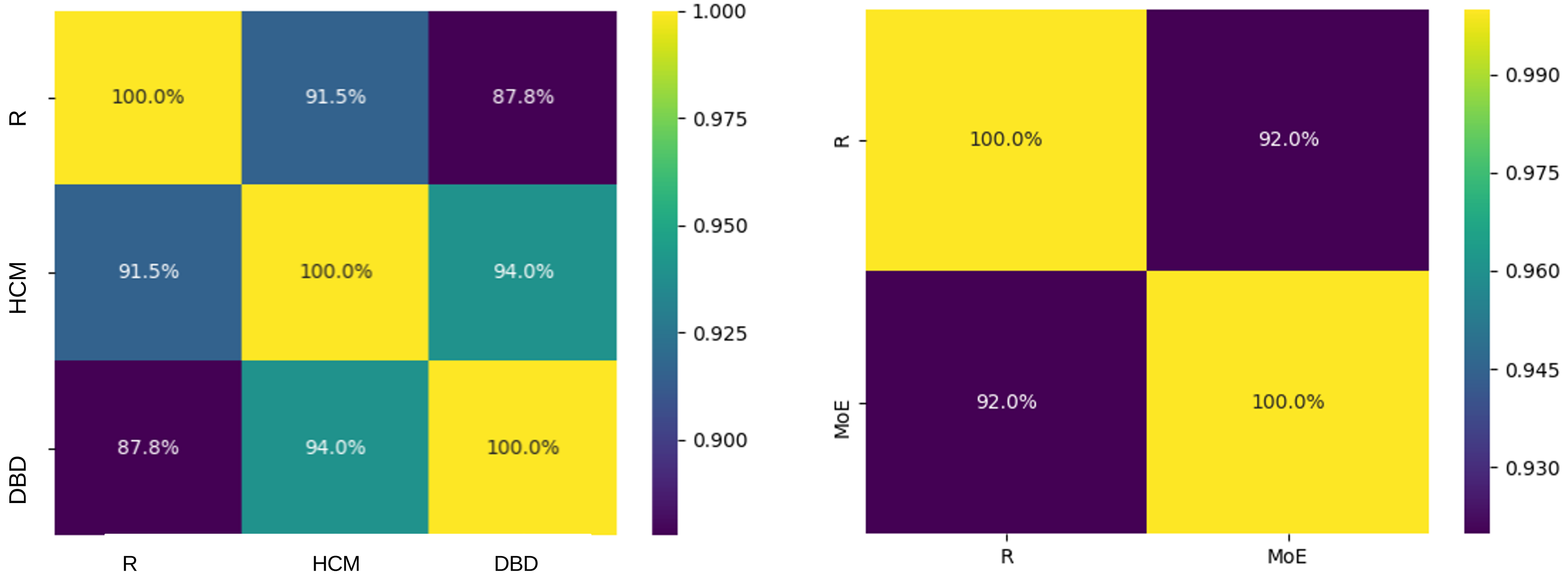
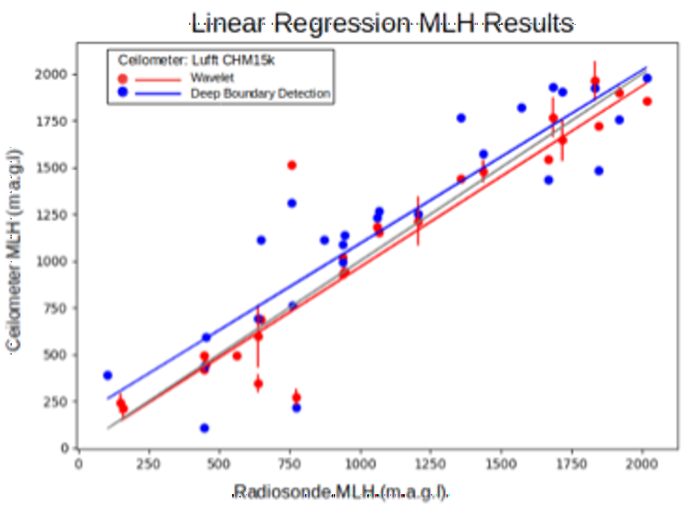


Figure 5. Correlation Matrix of December 2016 Campaign Radiosonde Mixing Layer Heights measurements and the Deep Boundary Detection model (DBD) and the Haar Covariance Method (HCM) compared with The Mixture of Experts Model which combines decisions between the Deep Boundary Detection model (DBD) and the Haar Covariance Method (HCM)

Deep Boundary Detection Method Regression Results

Current Efforts and Updates:

- Continued experimentation of deep boundary detection method including mixing layer heights and cloud based heights without denoising



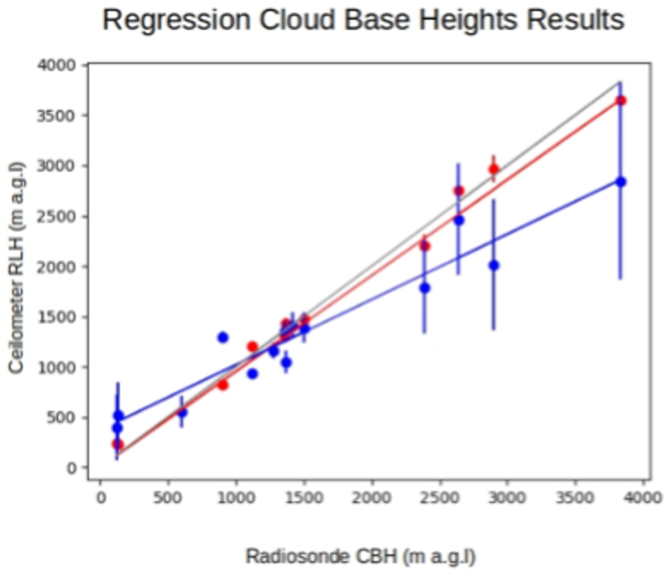
December 2016 Campaign - Total radiosondes: 37

Deep Boundary Detection:

Total Points: 26 – R^2=.80 – Slope=.93

Covariance Method (Caicedo et al. 2020):

Total Points: 26 – R^2=.88 – Slope=.97



December 2016 Campaign - Total radiosondes: 17

Deep Boundary Layer Cloud Base Heights:

Total Points: 14 – R^2=.93 – Slope=.65

Covariance Method (Caicedo et al. 2020) Derived Cloud Base Heights:

Total Points: 11 – R^2=.99 – Slope=.95

Figure 3. Regression Results for MLH and CBH

Mixture of Experts and Multi-Source Convolutional LSTM

- Using multiple sources of data to estimate the PBLH
- Experimenting with a multi-sourced stacked convolutional LSTM
- Learns PBLH over time for given geographical locations using a combination of source data
  - WRF-CHEM model backscatter
  - Ceilometer-based backscatter
  - Satellite-based backscatter

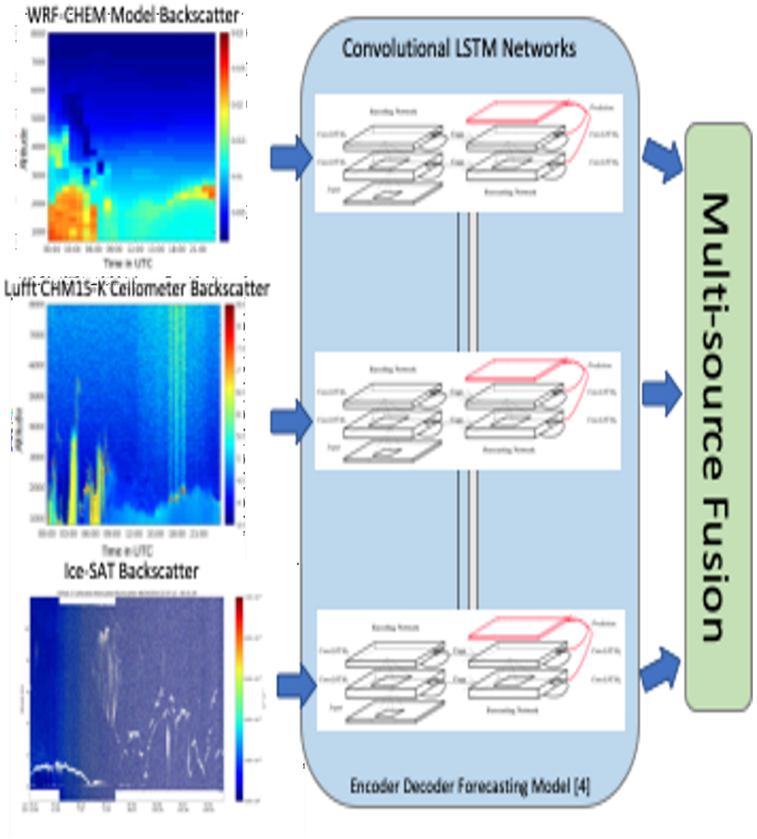

WRF-CHEM Model Backscatter
Time in UTC
Lufft CHM15-K Ceilometer Backscatter
Time in UTC
Ice-SAT Backscatter
Convolutional LSTM Networks
Encoder Decoder Forecasting Model [4]
Multi-source Fusion

Figure 9. Multi-Source Convolutional LSTM.

Extending Deep Boundary Detection to Other Sources: ICESat2

- Current efforts underway to compare a traditional method for estimating PBLH for ICESat2 data and using the Deep Boundary Layer Detection method
- Stacked LSTM to process ICESat2 data (working in combination with the WRF-CHEM model data LSTM and Ceilometer-based LSTM)
- Results will be forthcoming in a future meeting

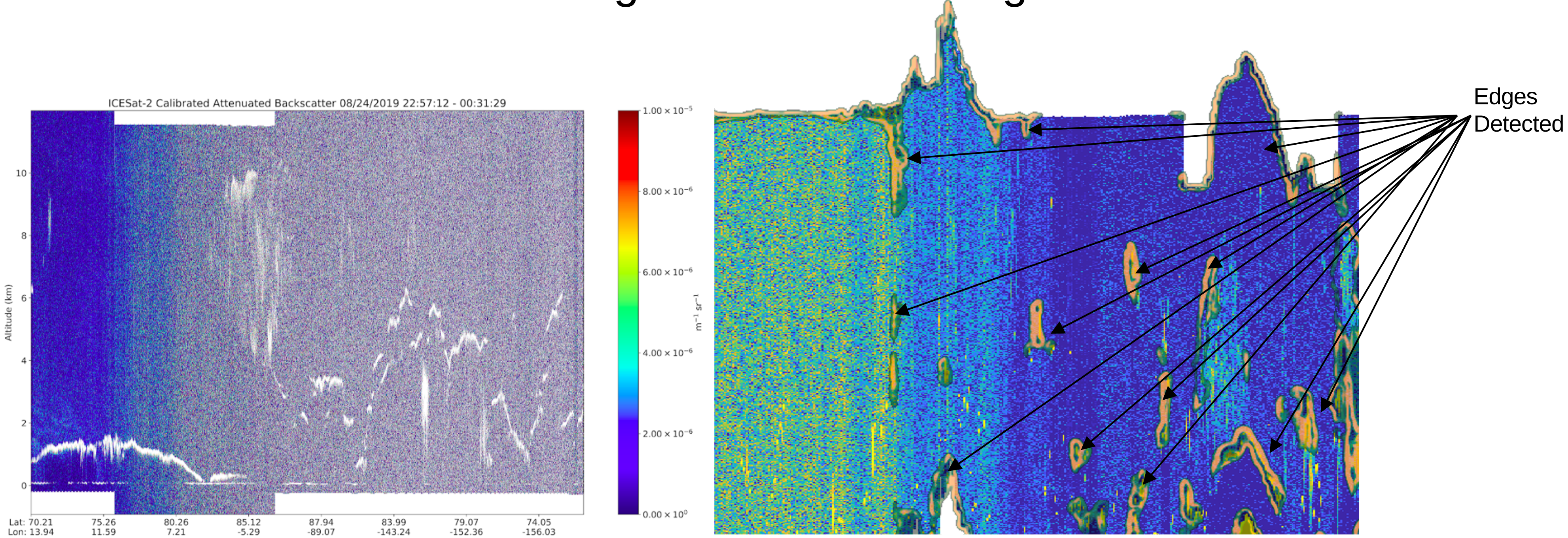
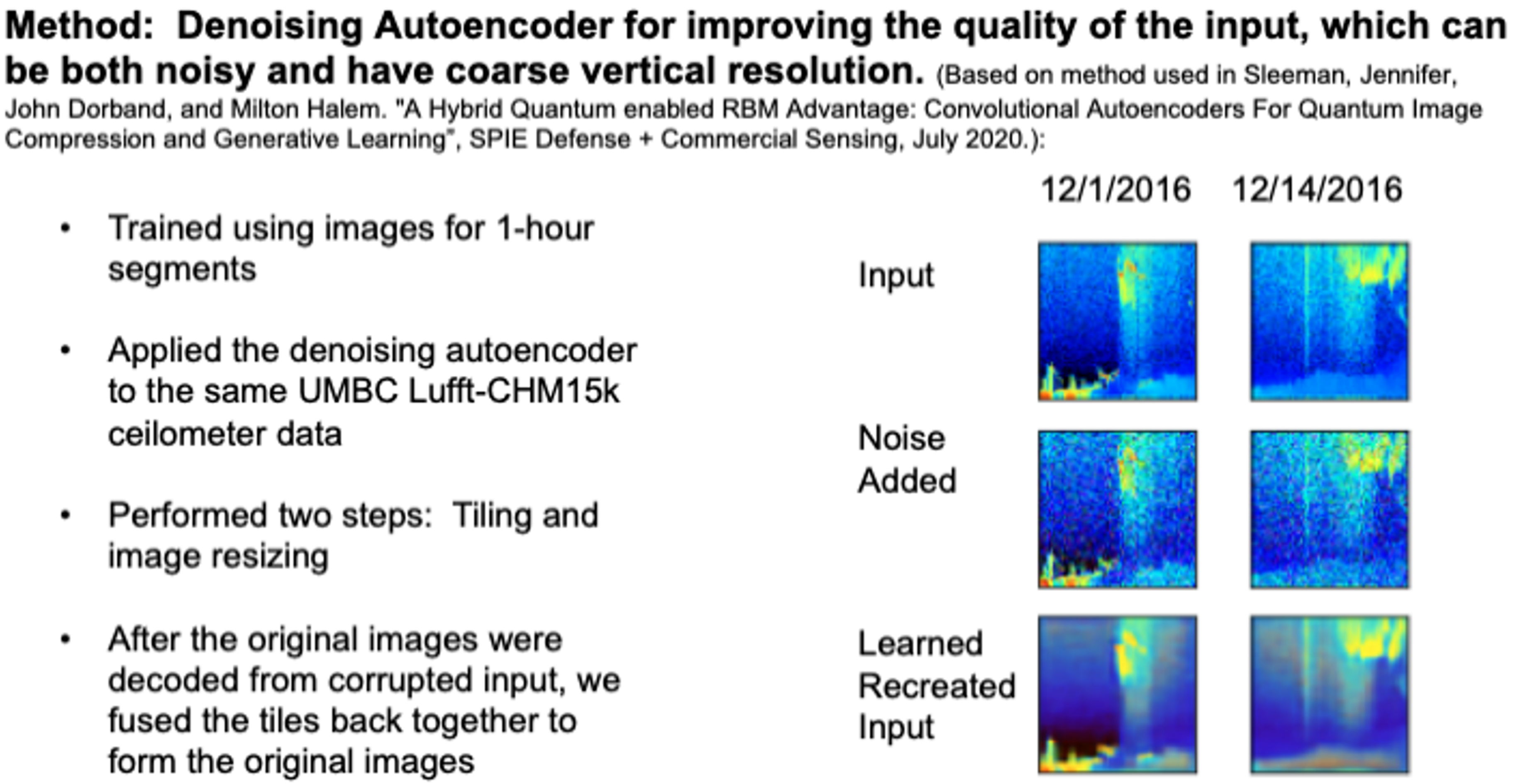


Figure 6a. ICESat2 for Arctic and 6b. ICESat2 -ATL09_20190501105015_05070301_003 With Overlay of Edges Detected for Location around UMBC

Deep Denoising Autoencoder

Early results of the denoising autoencoder. Experiments are underway to evaluate its effect on the deep edge detector (Publication forthcoming).



Method: Denoising Autoencoder for improving the quality of the input, which can be both noisy and have coarse vertical resolution. (Based on method used in Sleeman, Jennifer, John Dorband, and Milton Halem. "A Hybrid Quantum enabled RBM Advantage: Convolutional Autoencoders For Quantum Image Compression and Generative Learning", SPIE Defense + Commercial Sensing, July 2020.):

- Trained using images for 1-hour segments
- Applied the denoising autoencoder to the same UMBC Lufft-CHM15k ceilometer data
- Performed two steps: Tiling and image resizing
- After the original images were decoded from corrupted input, we fused the tiles back together to form the original images

Figure 4. Denoising Autoencoder Early Results.

Comparison of WRF-CHEM Backscatter Exp. With Ceilometers Domain Setup

Experiment Design:

Time period: Sep 9 2020 – Sep 11 2020

Spatial resolution: 9 km × 9 km (mother domain, Northeast);

3 km × 3 km (nest domain, Maryland);

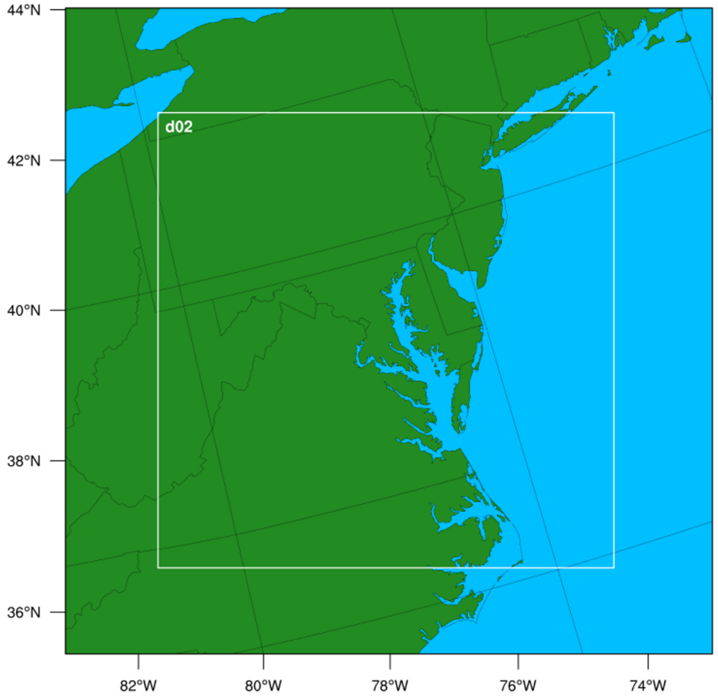
30 Levels

Sensitivity experiments: WRF-Chem (YSU, with chemistry);

WRF-Chem (MYNN, with chemistry).

Atmospheric Processes	WRF-Chem
Shortwave Radiation	RRTMG
Longwave Radiation	RRTMG
Microphysics	WSM5 (Hong et al., 2004)
Cumulus	Grell ensemble
Boundary Layer	YSU or MYNN
Land surface model	Noah LSM
Photolysis	TUV
Gas-phase Mechanism	RADM2
Aerosol process(Dust)	MADE/SORGAM(GOCART)

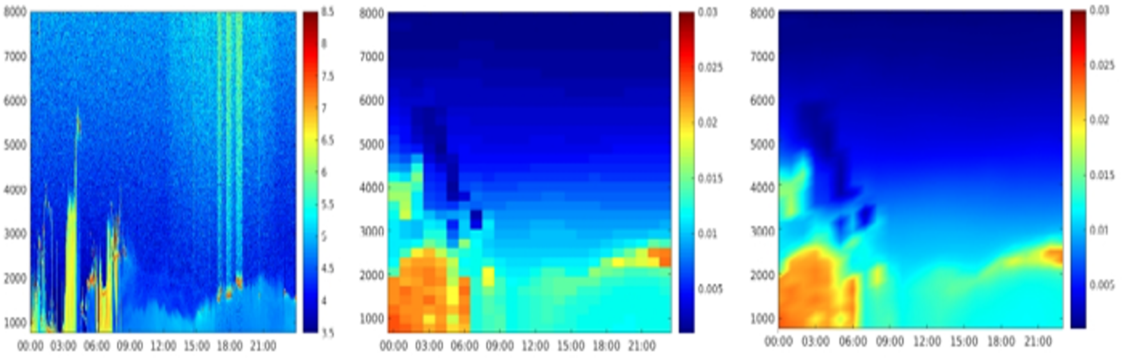
WPS Domain Configuration



Datasets:
Meteorological Data: NARR (North American Regional Reanalysis Data);
Anthropogenic Emission: NEI 2011 (National Emission Inventory)

Extending Deep Edge Detection to WRF-CHEM model

We processed backscatter from WRF-CHEM model and applied HED to output and compared anomalies from Pecan campaign with ceilometer backscatter below



Top. LIDAR Backscatter Image 12/1/2016 UTZ. Passing clouds no rain. 12:00UTZ (7:00AM) Left of image is Night and right is daytime with cloud capped boundary layer. b.) GoCART Model Backscatter Image at 1000 nm, every hour, at 40 levels initialized at Nov. 29, 2016 and c.) an Interpolated GoCART Model Backscatter for December 1, 2016.

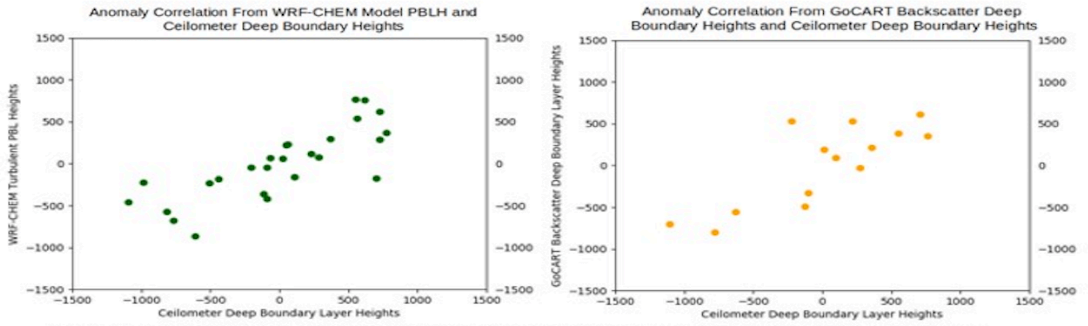


Figure 6. a.) Anomaly Correlation From WRF-CHEM Model PBLH and Ceilometer Deep Boundary Layer Heights (26 points with r2 = .63 and b.) Anomaly Correlation From GoCART Backscatter Deep Boundary Heights and Ceilometer Deep Boundary Heights (14 points with r2 = .66)

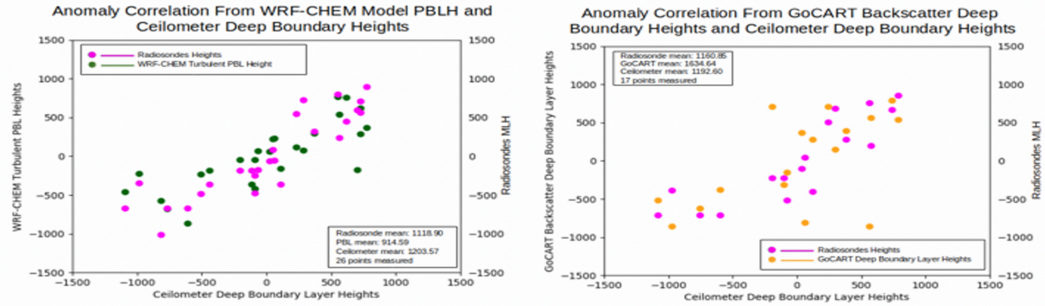


Figure 7. a.) Anomaly Correlation From WRF-CHEM Model PBLH and Ceilometer Deep Boundary Layer Heights and b.) Anomaly Correlation From GoCART Backscatter Deep Boundary Heights and Ceilometer Deep Boundary Heights

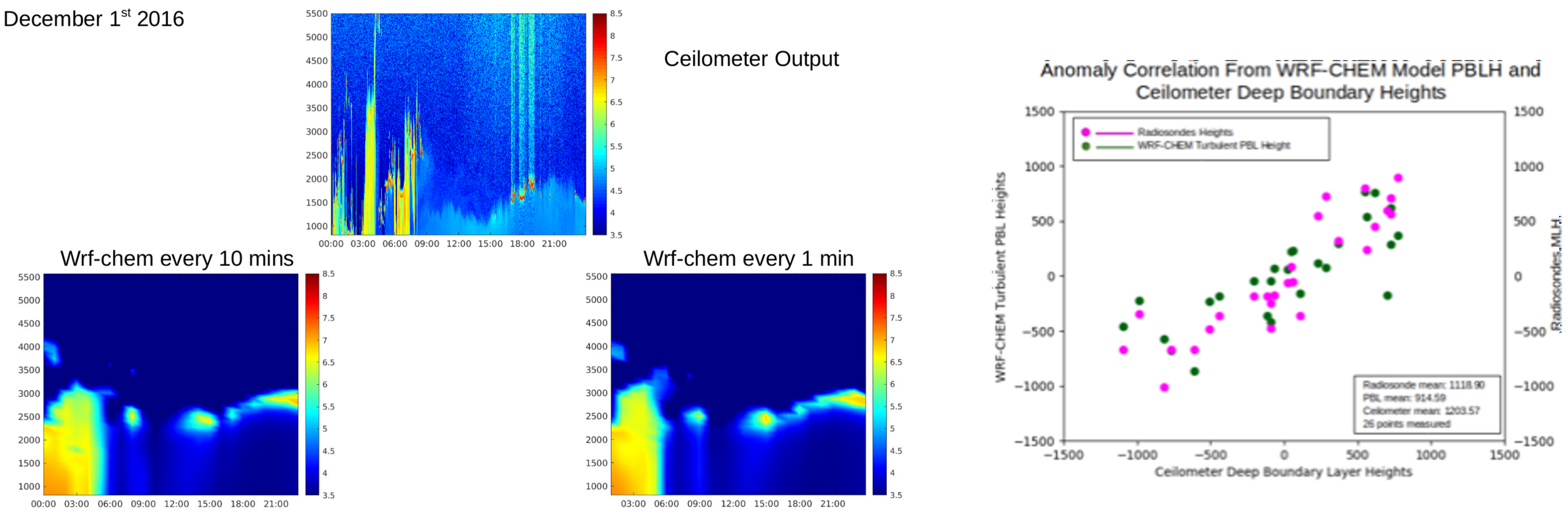
Mid Fig. R2 Corr. of (PBLH,Ceil) = 0.63 for 26 pts. R2 Corr. of (ABLH,Ceil) = 0.66 for 14 pts. Bias of PBLH and backscatter ABLH of opposite sign
Botom (PBLH,Ceil) mean 914, (Rawins,Ceil) mean 1118, Ceilometer mean 1209 for 26pts. Rawinsonde mean 1160, ABLH 1634 Ceilometer 1192 for14pts.

https://www.timeanddate.com/weather/usa/baltimore/historic?month=11&year=2016

Operational System: Using Machine Learning to Identify PBL Heights
Dr's Jennifer Sleeman, Vanessa Caicedo and Dorsa Ziaei

Current Efforts and Updates:

- Continued work on Deep Boundary Layer detection as part of WRF-CHEM
- Evaluating performance for simultaneous processing of 1000’s of geographical locations



December 1st 2016

Ceilometer Output

Wrf-chem every 10 mins

Wrf-chem every 1 min

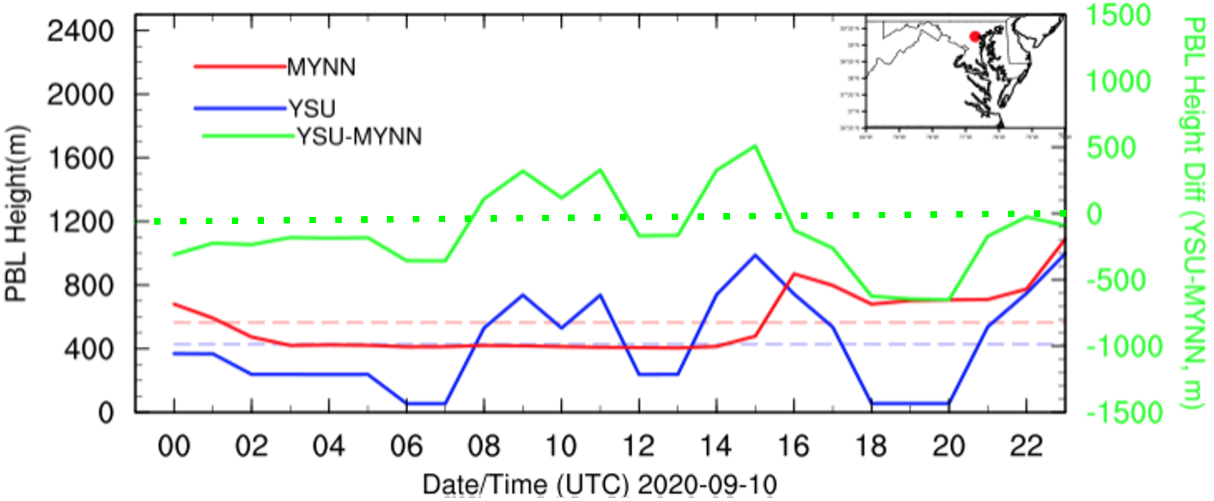
Anomaly Correlation From WRF-CHEM Model PBLH and Ceilometer Deep Boundary Heights

Radiosonde mean: 1118.90
PBL mean: 914.59
Ceilometer mean: 1203.57
26 points measured

Figure 11. Comparing WRF-CHEM models with Ceilometer Output for UMBC Location. Dec 1, 2016.

PBLH Diurnal Variation

Sep 10, 2020



@Catonsville:

- Diurnal cycle;
- YSU: fluctuation; MYNN: smooth
- Afternoon&Nighttime: $PBLH_{YSU} < PBLH_{MYNN}$; Morning: $PBLH_{YSU} > PBLH_{MYNN}$;
- $PBLH_{YSU} = 0$?

Machine Learning Methods to Identify PBL Heights

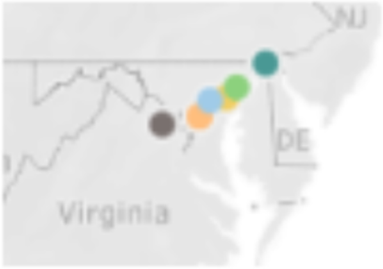
Summary of Progress from the Machine Learning Algorithmic Development Team:

- Simultaneous processing of 5 stations (Proof of concept integration with data acquisition team and visualization team)
- Continued evaluation of the Deep Boundary Layer Performance results including Residual Layer Heights and Cloud Base Heights
- Covariance Method and Deep Boundary Layer Method Mixture of Experts Early Results and Proposed Network
- Breakthrough results with LSTM model using hourly edge detection images for PBLH prediction and Proposed Model for Multi-Source LSTM
- Simultaneous processing of 1000’s of stations for WRF-CHEM integration

Task 3. PBLH Spatial Fusion and Visualization
D. Chapman, P. Bindu

- Objectives
  - Create Level 3 hourly gridded PBLH from ceilometers at 3.2km
  - Hourly fusion with Ceilometer, WRF-Chem Model output PBLH
  - NOAA GFS Model forecasts
  - Web accessible interactive visualization + geobrowser

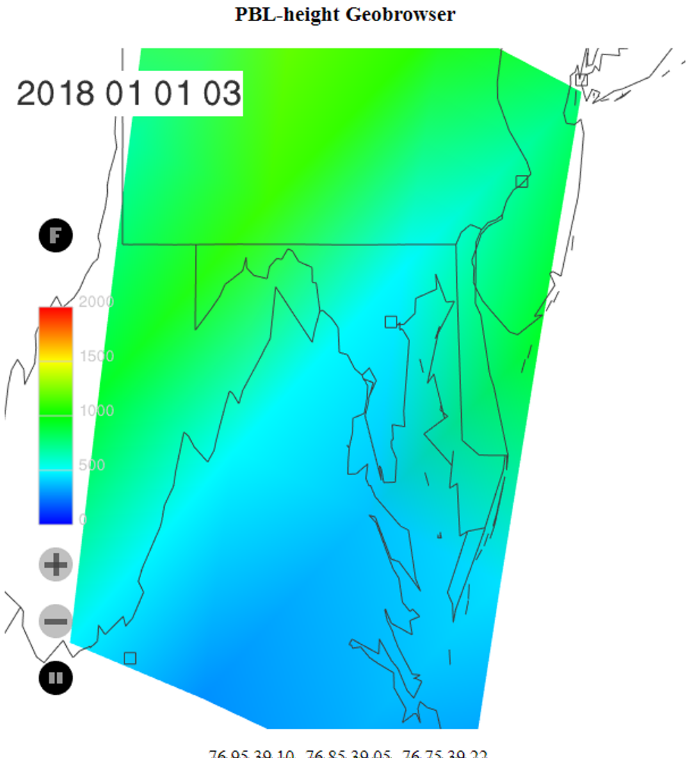
Level 3 Data Fusion



Ceilometer PBLH

WRF-Chem PBLH

PBLH Visualization



PBL-height Geobrowser

PBLH Level 3 Spatial Fusion and Visualization

Key Accomplishments

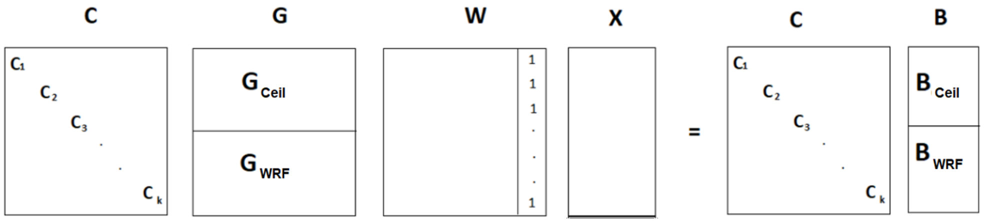
- Integration of Ceilometer profiles from four sites along i95 corridor
  - End to end processing including data streaming, L2 retrievals and L3 gridded PBLH maps
- Fusion of L2 ceilometer profiles with WRF-CHEM model outputs
  - Method of compressive sensing with 2D+time wavelet transform
  - 3.2km resolution and hourly timescales over BW i95 corridor
- Prototype interactive visualization geobrowser
  - Display ceilometer derived L3 gridded PBLH profiles
  - Comparison of L3 product with raw WRF-chem shows large differences in PBLH

Level 3 Gridded PBLH via Compressive Sensing Fusion

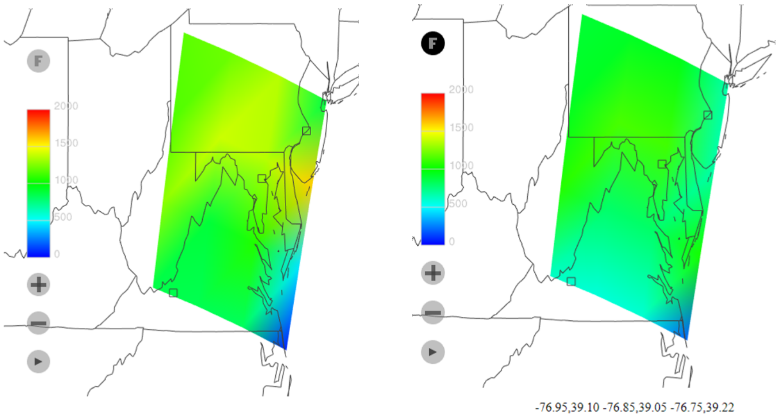
- Compressive Sensing for Level 3 PBLH grid
  - Fusion of Ceilometer Profiles and WRF-chem to infer gridded PBLH at 3.2km resolution.
  - Fusion with WRF-chem model outputs can interpolate between ceilometer point backscatter measurements while maintaining high frequency signal due to surface interaction.
  - PBLH spatial fusion using L1 Compressive Sensing with Wavelet basis space.

$$\min \; ||CGWx - Cb||_2^2 + \lambda ||x||_1$$

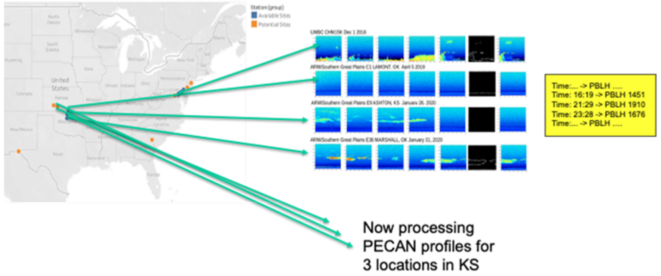
  - PBLH profiles from 5 ceilometers along greater BW metropolitan area.



C: diagonal calibration matrix, G: Sensing matrix, W: wavelet transform
X: Inferred Signal B: Observation vector



Interactive Visualization of L3 Fused PBLH product
Left: wrf-chem PBLH Right: fused PBLH


Station (group)
Available Sites
Potential Sites
United States
UMBC CHM15K Dec 1 2016
ARM/Southern Great Plains C1 LAMONT, OK April 5 2019
ARM/Southern Great Plains E9 ASHTON, KS January 26, 2020
ARM/Southern Great Plains E36 MARSHALL, OK January 31, 2020

Time:... -> PBLH ....
Time: 16:19 -> PBLH 1451
Time: 21:29 -> PBLH 1910
Time: 23:28 -> PBLH 1676
Time:... -> PBLH ....

Now processing PECAN profiles for 3 locations in KS

Ceilometer Stations including PCAN and greater BW metro area

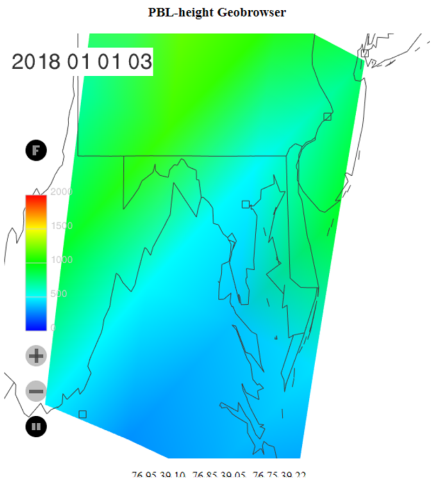
Visualization Features

- Objectives:
  - Create Level 3 hourly gridded PBLH from ceilometers at 3.2km
  - Hourly data fusion of Ceilometer ABLH with WRF-Chem Model output PBLH

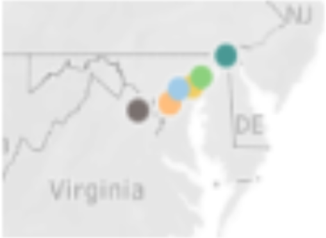
    or NOAA HRRR Model forecasts
  - Web accessible interactive visualization + geobrowser

- Compressive Sensing Fusion of PBLH from Ceilometer and WRF-Chem Model:
- Interactive Data visualization using-
  - U.S. Census Bureau's MAF/TIGER Database
  - HTML5 web servlet technology
  - Integration with Data Archive + Apache

Visualization in Progress



Data Fusion



Ceilometer PBLH

WRF-Chem PBLH

Summary of Accomplishments
B. Demoz, M.Halem

1. **Towards** demonstrating the feasibility of an edge streaming, secure, scalable, fault-tolerant nationwide pilot to ingest, pre-process, infer aerosol boundary layer heights and archive all in near real-time based on ceilometer and remote sensed lidar aerosol backscatter profiles.

2. Developing a validated Hybrid Deep Hierarchical Machine Learning Edge Detection and Covariance Wavelet algorithm for an end-to end hourly Aerosol Boundary Layer Height (ABLH) product from Ceilometer and remote sensed Lidar aerosol backscatter

3. Producing 3-D hourly boundary layer height maps by a compressive sensing fusion methodology from the derived ceilometer ABLH and operational reanalysis PBLH.

Plans Going Forward:

- Complete scaling out the ground-based portion of the edge streaming, fault tolerant, secure prototype NRT ML aerosol backscatter inferring boundary layer height and visualization products by mid summer.

- Test, evaluate and incorporate the ingest of satellite Lidar aerosol backscatter and application of a Deep LSTM derived ABLH product as a compliment to ground based Lidar systems.

- Perform and evaluate data assimilation of ABLH into regional forecast models

- Conduct an OSSE for the proposed NASA Wind Lidar to compliment the NASA Icesat-2 and ESA Aeolus ADM Lidar sensors for a CONUS R2O Regional Subseasonal Forecast.

Publications

2021 Sleeman J., D. Ziaei, Z. Yang, V. Caicedo, C. Calderella, M. Halem, R. Delgado, B. Demoz, " A Deep Multi-Stacked Neural Network Approach for Improved Planetary Boundary Layer Height Estimation" AMS 101st Annual Meeting 8.6 2021

2020 Halem M., J. Sleeman, Z. Yang, M. Chin, D. Watson-Parris, B. Demoz, " Feasibility Studies of Cloud Resolving NU -WRF Subseasonal Forecasts with AI Emulations". AGU IN024 Fall 2020

2020 Gite R., M. Halem, P. Nguyen, " Compressive Sensing and Deep Learning framework for Multiple Satellite Sensor Data Fusion". AGU IN033 Fall 2020

2020 Sleeman J., Vanessa Caicedo, Dorsa Ziaei·, M. Halem, Belay Demoz[1]uben Delgado," Using Machine Learning to Identify Planetary Boundary Layer Heights for Ceilometer-Based LIDAR Backscatter Retrievals". AGU Fall 2020

2020 Yang, Z., M.Halem " Model Evaluation and Assimilation of the Planetary Boundary Layer Height". AGU Fall 2020

2020 Nguyen P., M. Halem,'Satellite Data Fusion of Multiple Observed XCO2 using Compressive Sensing and Deep Learning' IGARSS 9/20

2020 Sleeman J., Z. Yang, V. Caicedo, M. Halem, B. Demoz, "A Deep Machine Learning Approach for LIDAR Based Boundary Layer Heigh Detection" IGARSS 9/20

2020 Ayanzadeh R., M. Halem, T. Finin,"An Ensemble Approach for Compressive Sensing with Quantum Annealers" IGARSS 9/20

2020 Ziaei, D., D. Chapman, Ya. Yesha, M. Halem, “Segmentation of Stem Cell Colonies in Fluorescence Microscopy images with transfer learning” SPIE Conference Medical Imaging, 2020 Houston, TX

2020 Ayanzadeh, R, M. Halem, and T. Finin. "Reinforcement quantum annealing: A hybrid quantum learning automata." Nature: Scientific reports 10, no. 1 (2020): 1 11.

2020 Carroll , B., Belay B. Demoz , David D. Turner , and Ruben Delgado: Lidar observations of a mesoscale moisture transport event impacting convection and comparison to Rapid Refresh model analysis" Published-online: 04 Dec 2020 Collections: Plain Elevated Convection At Night (PECAN)DOI: https://doi.org/10.1175/MWR-D-20-0151.1

2020 Tangborn, A., Demoz, B., Carroll, B. J., Santanello, J., and Anderson, J. L. "Assimilation of lidar planetary boundary layer height observations, Atmos. Meas. Tech. https://doi.org/10.5194/amt-2020-238

2020 Lopez-Coto, Israel; Micheal Hicks; Anna Karion; Ricardo Sakai; Belay Demoz; Kuldeep Prasad; James Whetstone: assessment of Planetary Boundary Layer parameterizations and urban heat island comparison: Impacts and implications for tracer transport J. Appl. Meteor. Climatol. (2020) 59 (10): 1637–1653.https://doi.org/10.1175/JAMC-D-19-0168.1

2019 Nguyen, P, M. Halem,”Deep Learning Models for Predicting C Employing Multivariate Time Series” IEEE Knowledge and Data Discover (KDD) conference Aug. 2019, Anchorage, Alaksa.

Thanks to ESTO/AIST

Surrogate Modeling for Atmospheric Chemistry and Data Assimilation

Daven Henze (PI, CU Boulder, Mechanical Engineering)
Alireza Doostan (co-I/Science PI, CU Boulder, Aerospace Engineering)

AIST-18-0072 Annual Technical Review
1/22/2021

Additional Team Members: Dr. Hee-Sun Choi, Dr. William Tsui, (CU Boulder), Nicolas Bousserez (Collaborator, ECMWF)

Surrogate modeling for atmospheric chemistry and data assimilation

PI: Daven Henze, University of Colorado, Boulder

Objective

- Enhance computational efficiency of air quality (AQ) simulations through development and application of surrogate models for atmospheric gas-phase chemistry
- Demonstrate value through implementation within a widely used global 3-D chemical transport model, the GEOS-Chem 4D-Var chemical data assimilation system
- Provide surrogate-generation toolbox to enable community applications with user-provided chemical mechanisms
- Apply surrogate-based AQ modeling framework for assimilation of geostationary observations of atmospheric composition (TEMPO, pseudo observations of NO2)

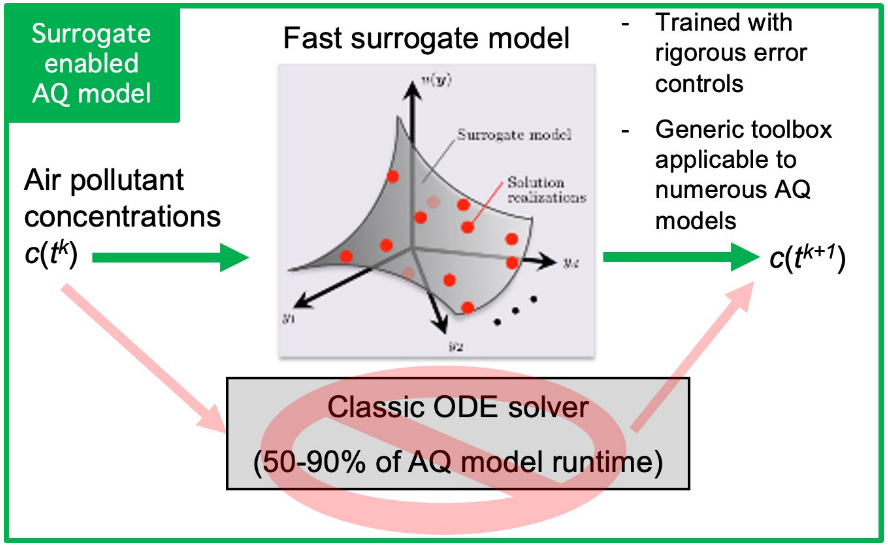

Surrogate enabled AQ model
Fast surrogate model
- Trained with rigorous error controls
- Generic toolbox applicable to numerous AQ models
Air pollutant concentrations
$c(t^k)$
$c(t^{k+1})$
$u(y)$
Surrogate model
Solution realizations
y_1
y_2
y_d
Classic ODE solver
(50-90% of AQ model runtime)

***Fig:** Surrogate improves runtime over classic ODE solver approaches*

Approach

Steps for surrogate model generation:

- Generate a training dataset (107 samples) using global GEOS-Chem High Performance model
- Construct surrogate model with low-rank tensor decomposition using Canonical Polyadic (CP) formalism for compressed sensing (machine learning) and/or DNN
- Implement multi-scale preconditioning to address stiffness of chemical kinetics and regularization and general cross validation for rigorous error control
- Apply and distribute a software development toolbox for surrogate model generation process

Co-PI: Alireza Doostan, University of Colorado, Boulder
Collaborator: Nicolas Bousserez, ECMWF

Key Milestones

• Surrogate of GEOS-Chem chemical solver delivered to GEOS-Chem code repository	01/21
• Surrogate model generating toolbox available to AQ community	07/21
• Surrogate-based GEOS-Chem 4D-Var applied to assimilation of pseudo TEMPO NO2	10/21
• Surrogate model for CAMS delivered to ECMWF for implementation	01/22
• Surrogate-based GEOS-Chem 4D-Var delivered to GEOSChem code repository	01/22

TRL$_{in}$ = 2 **TRL$_{current}$ = 2**

Presentation Contents

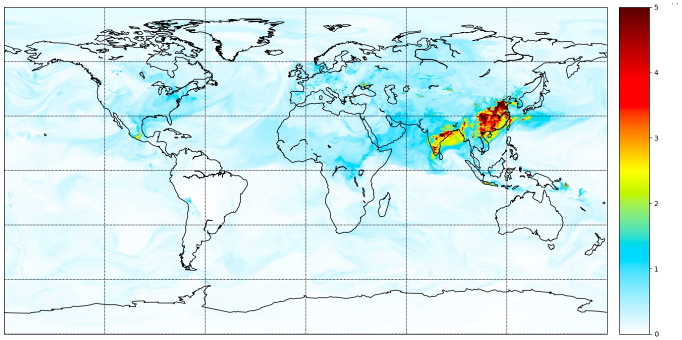
- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Earth Science Technology Office

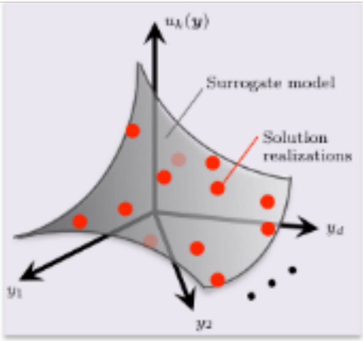
Background / Objectives

Background:

- Computational bottleneck of AQ models is chemistry (50 to 90% of run time)
- Several applications need more efficient models:
  - data assimilation and forecasting (e.g., US NAQFC, ECMWF)
  - higher resolution for health impacts
  - longer simulations for chemistryóclimate
- Previous surrogate modeling attempts inaccurate and/or slow (e.g., Keller and Evans, 2019; Kelp et al., 2020), not focused on data assimilation
- New methods in compressive sensing, tensor decomposition, and machine learning hold promise for parameter space exploration and UQ of large-scale dynamical systems



GEOS-Chem simulation of aerosol sulfate



Construction of high-D surrogates by exploiting sparsity or low-rank structures of the parameter to observable maps.

Relevance:

- R&A and Applications science goals: Atmospheric Composition, Health & Air Quality
- AIST goals: “analytic tools to characterize the natural phenomena or physical processes from data” and “data-driven modeling tools enabling the forecast of future behavior of the phenomena.”
- NASA’s remote sensing of atmospheric composition (e.g., TEMPO)
- Build upon previously funded NASA support for GEOS-Chem 4D-Var (e.g., PI Henze’s NASA grants NNX13AK86G, NNX16AF97G, NNX17AF63G).
- Could contribute to efficiency improvements in other models that use GEOS-Chem’s chemistry routines, such as the NASA GEOS model.

Objectives

Objectives and **Information Technology** (bold) :

- Develop, test, and deliver a **surrogate model for chemistry in GEOS-Chem**
- Generalize **surrogate model generation** procedure within a **software toolbox**
- Demonstrate benefits of **surrogate-based AQ modeling framework for chemical data assimilation** of geostationary observations of atmospheric composition

Science goals:

- Develop new techniques for surrogate modeling of high-dimensional, non-linear, large-scale dynamical chemical systems
- Improve O_3 forecasting through assimilating NO_2 observations from geostationary remote sensing measurements.

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Technical and Science Advancements

- **Problem description**

- Chemical conc. at ! = $!_{!}$ + Δt

- Temperature at ! = $!_{!}$
- Pressure at ! = $!_{!}$
- The rate of constants at ! = $!_{!}$
- Chemical conc. at ! = $!_{!}$

$\mathbf{X}$

Current ambient quality

$$\mathbf{Y}_S = S(\mathbf{X})$$

: Cheap computation
: Fast predictions

$$\wr\wr$$

$$\mathbf{Y} = ODE(\mathbf{X})$$

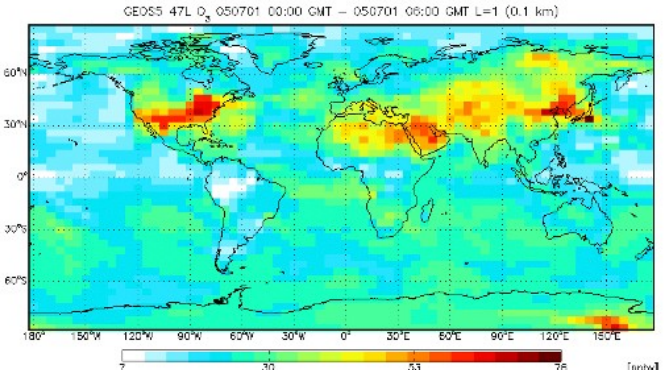
: Expensive computation

Future quality

$\mathbf{Y}_{true}$

$\approx$

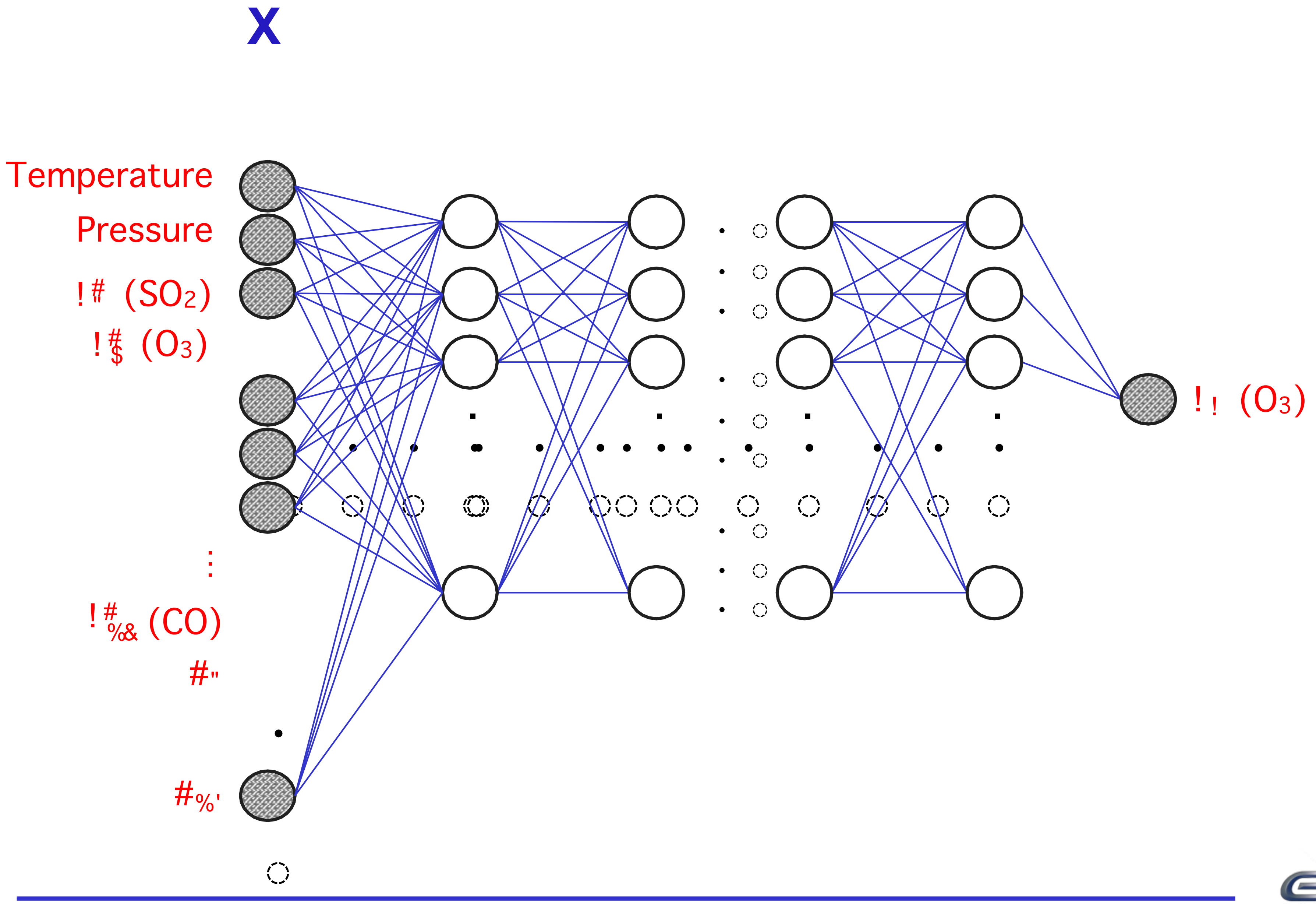
Predictions of future air quality



$\mathbf{Y}$

DNN models for air quality

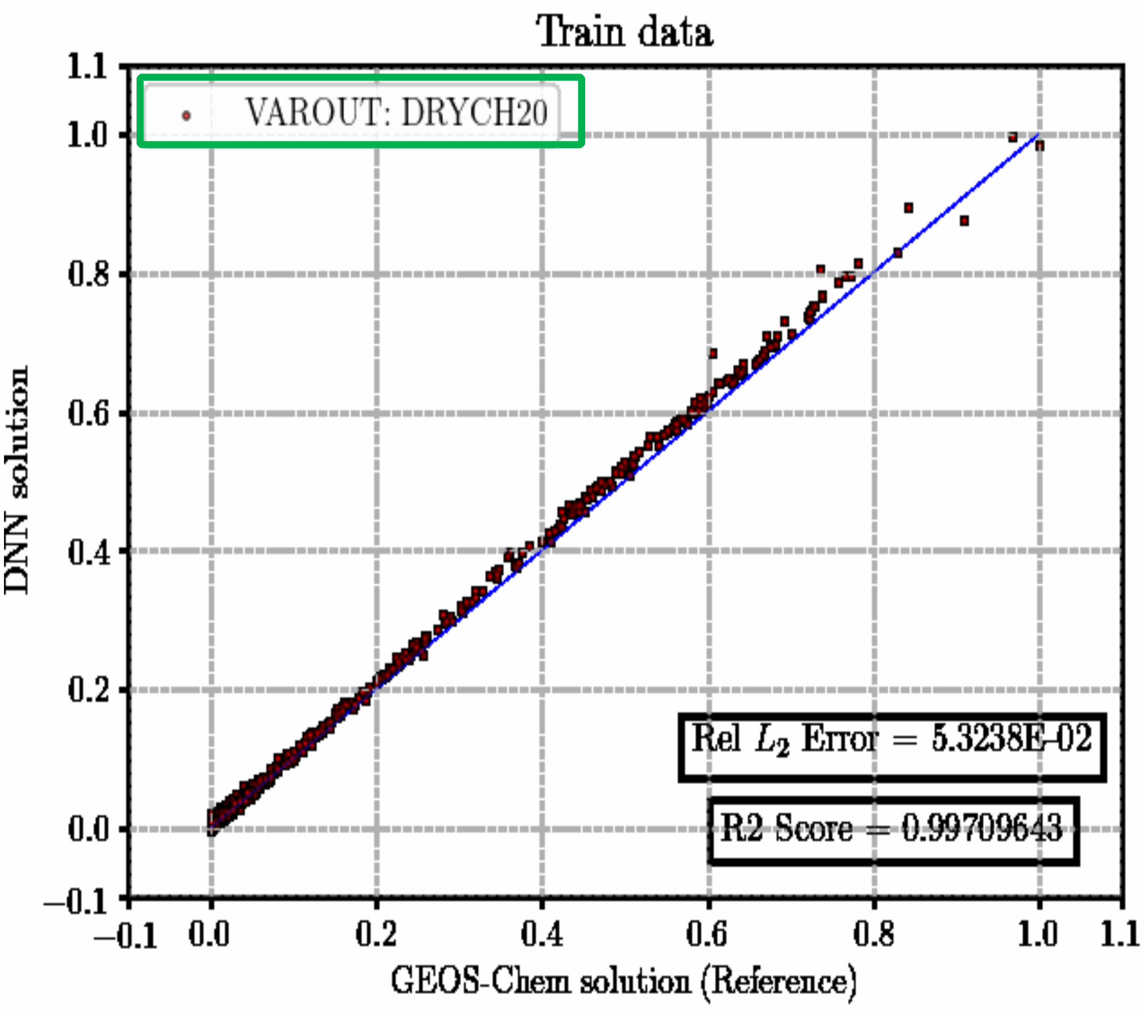
GEOS-Chem 3D model



Performance of DNN models

GEOS-Chem 3D model

- **DNN model (Ver 0)**
  (1) Independent surrogate models for respective chemical species
  (2) $\$_{\$}$ values > 0.98 for all surrogate models



Train
80%

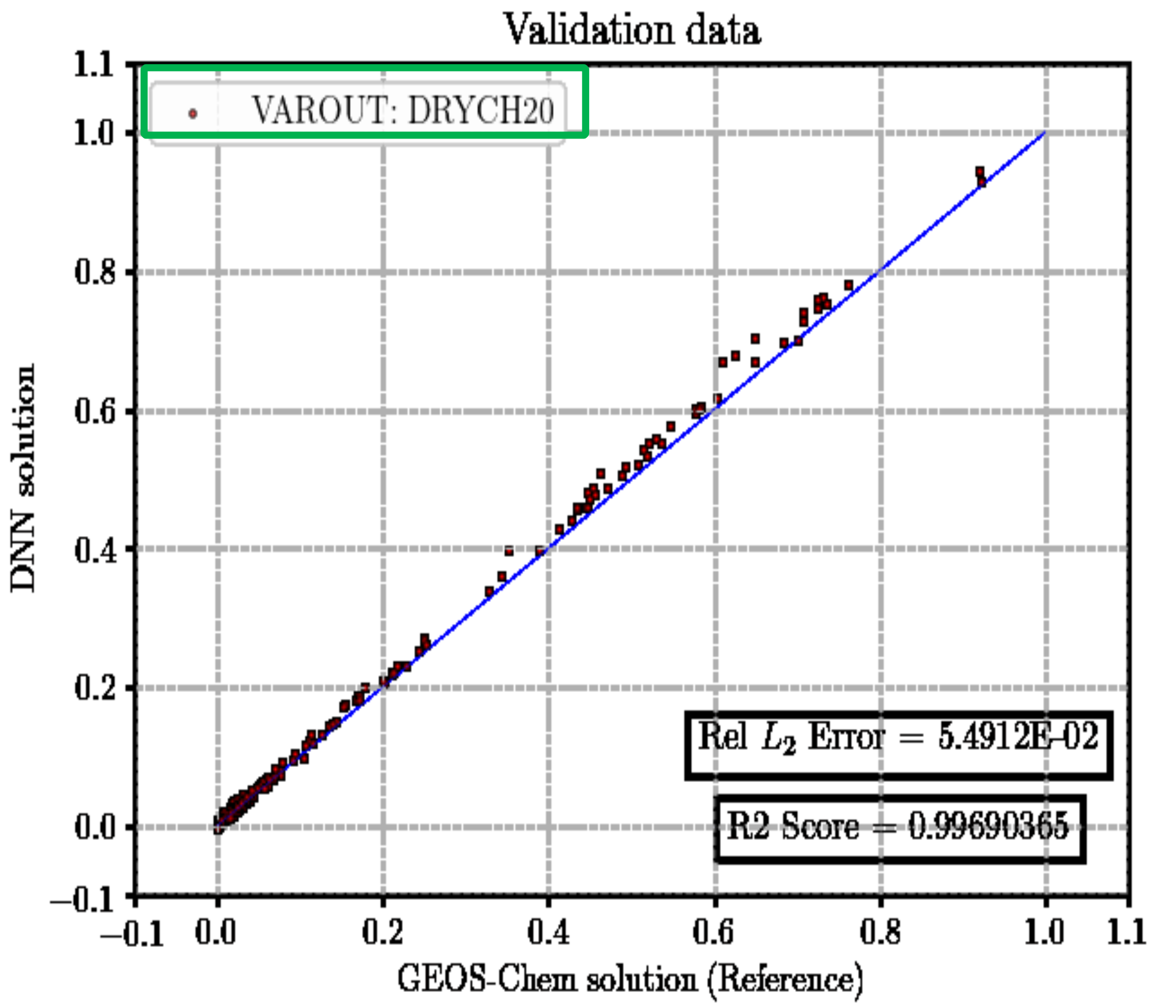
Validation
20%

Samples from GEOS-Chem global 2x2.5

Performance of DNN models

GEOS-Chem 3D model

- **DNN model (Ver 0)**
  (1) Independent surrogate models for respective chemical species
  (2) $\$_{\$}$ values > 0.98 for all surrogate models



Train
80%

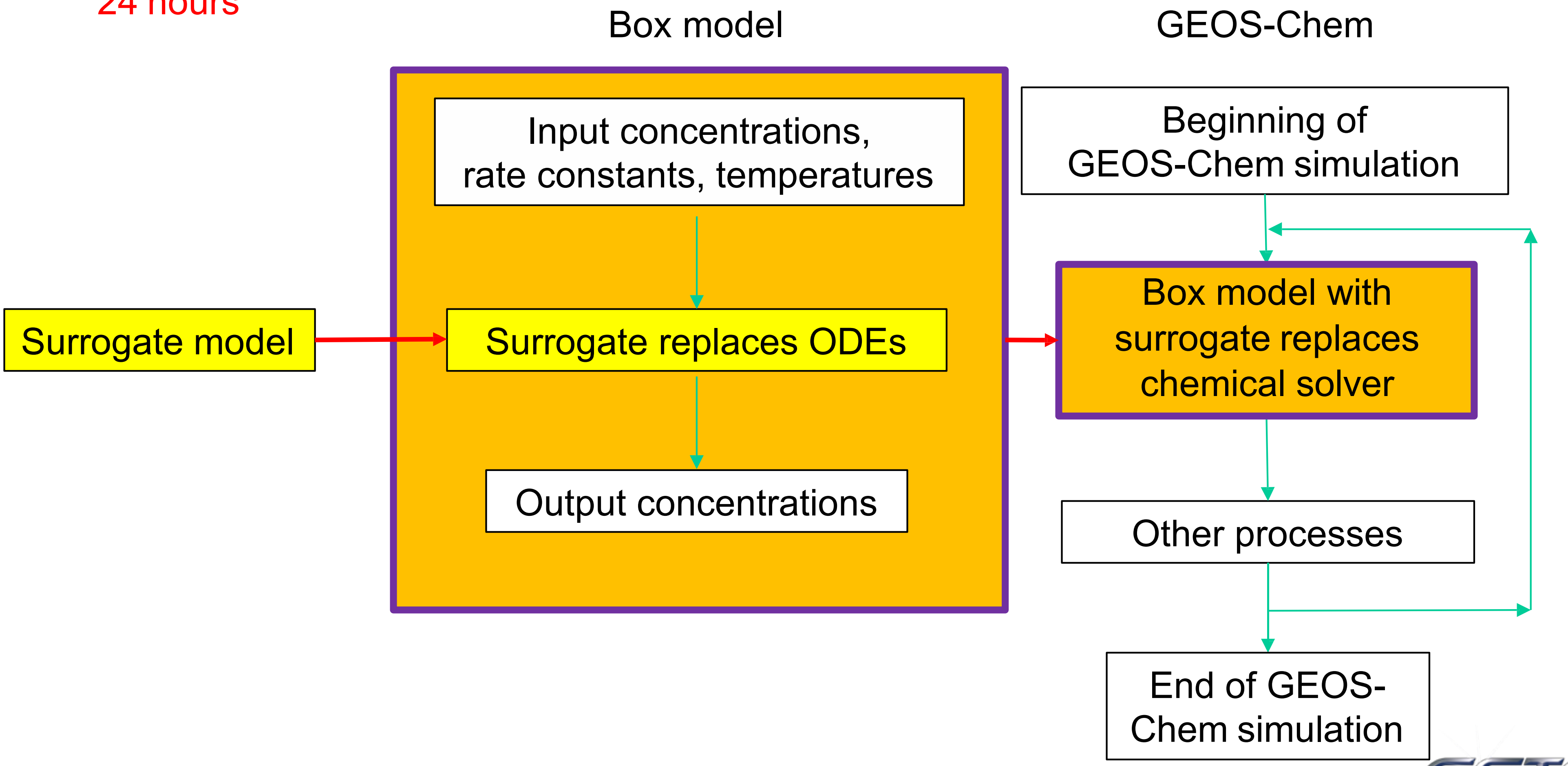
Validation
20%

Samples from GEOS-Chem global 2x2.5

Technical and Science Advancements

Using surrogate models with the box model and GEOS-Chem

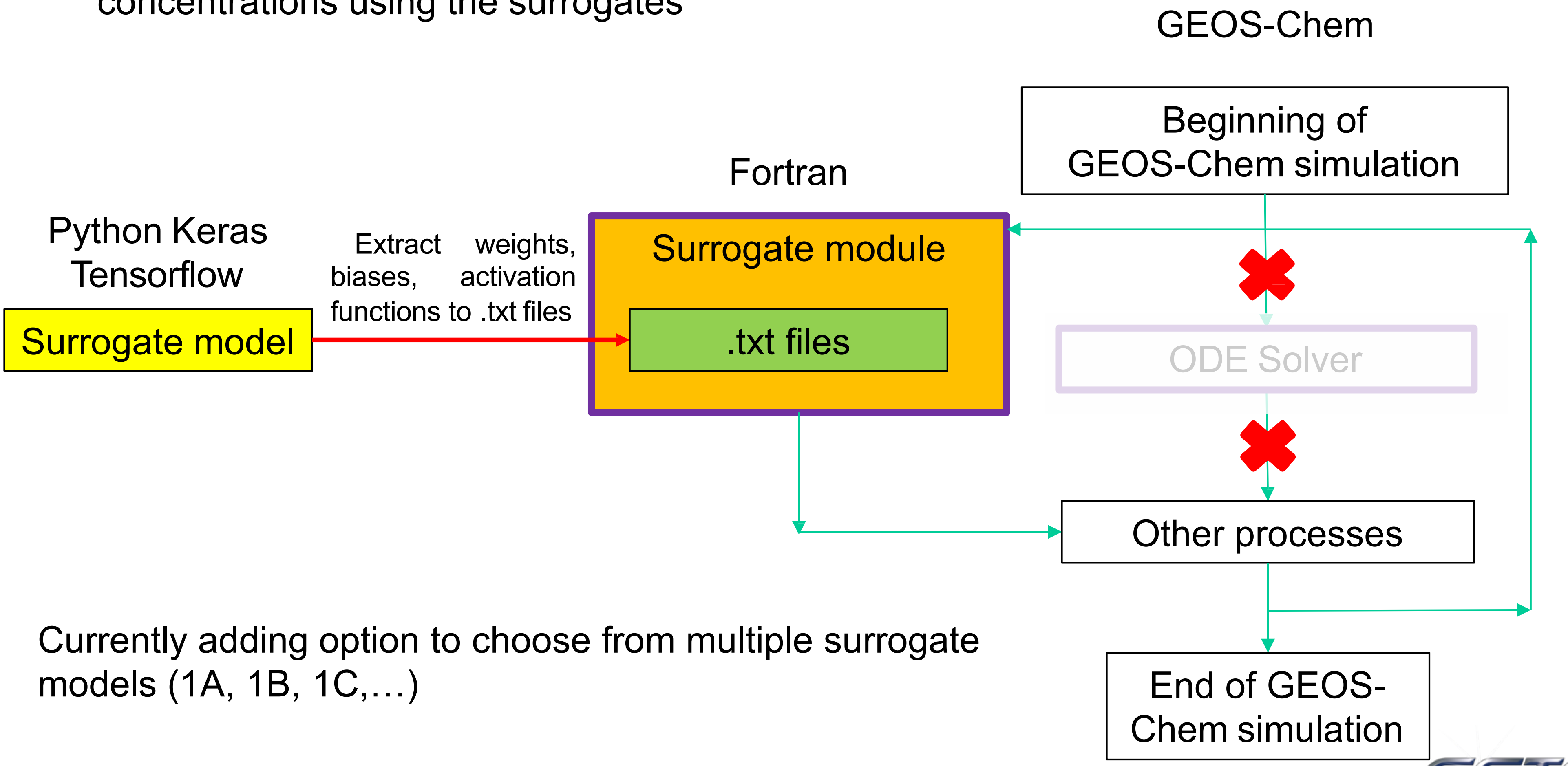
(1) Train surrogate models using one-hour timestep data from GEOS-Chem simulations
(2) Run surrogate models for up to 24 hours and compare to box model
(3) Replace chemical ODE solver in GEOS-Chem with surrogate models and run for up to 24 hours


Box model
GEOS-Chem
Input concentrations, rate constants, temperatures
Surrogate model
Surrogate replaces ODEs
Output concentrations
Beginning of GEOS-Chem simulation
Box model with surrogate replaces chemical solver
Other processes
End of GEOS-Chem simulation

Technical and Science Advancements

Incorporating surrogate model into GEOS-Chem (Fortran)

(1) Surrogate models are trained using Python Keras
(2) Model features are extracted to text files
(3) A separate module was written in fortran which reads the text files to predict chemical concentrations using the surrogates


GEOS-Chem
Beginning of GEOS-Chem simulation
Fortran
Python Keras Tensorflow
Extract weights, biases, activation functions to .txt files
Surrogate module
Surrogate model
.txt files
ODE Solver
Other processes
End of GEOS-Chem simulation

Currently adding option to choose from multiple surrogate models (1A, 1B, 1C,…)

Technical and Science Advancements

Global GEOS-Chem simulations of O_3 for ODEs and surrogate (Ver. 0) over 24 hours

With surrogate

With ODEs

2 hours

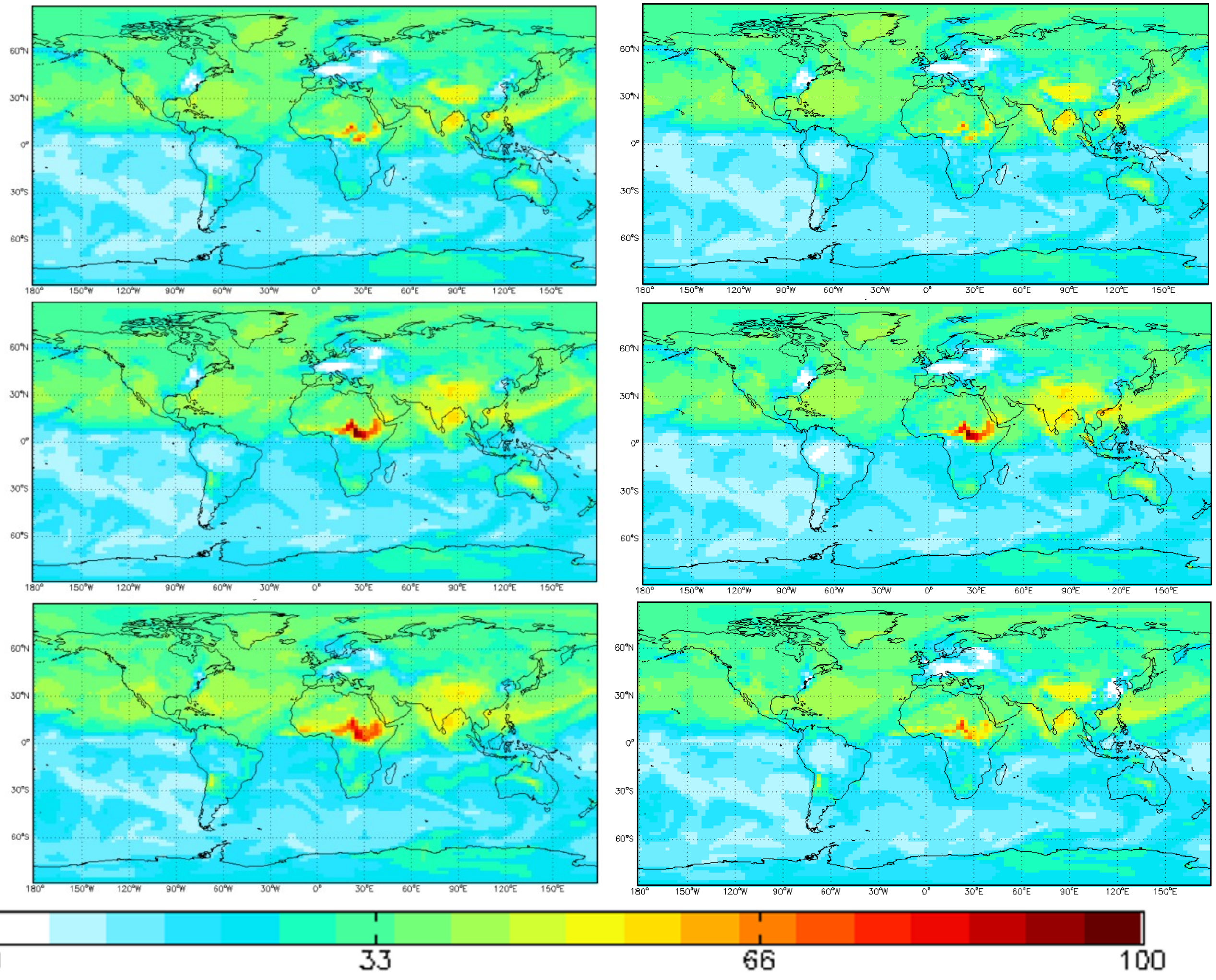
SDA = 1.64

6 hours

SDA = 1.04

24 hours

SDA = 0.64



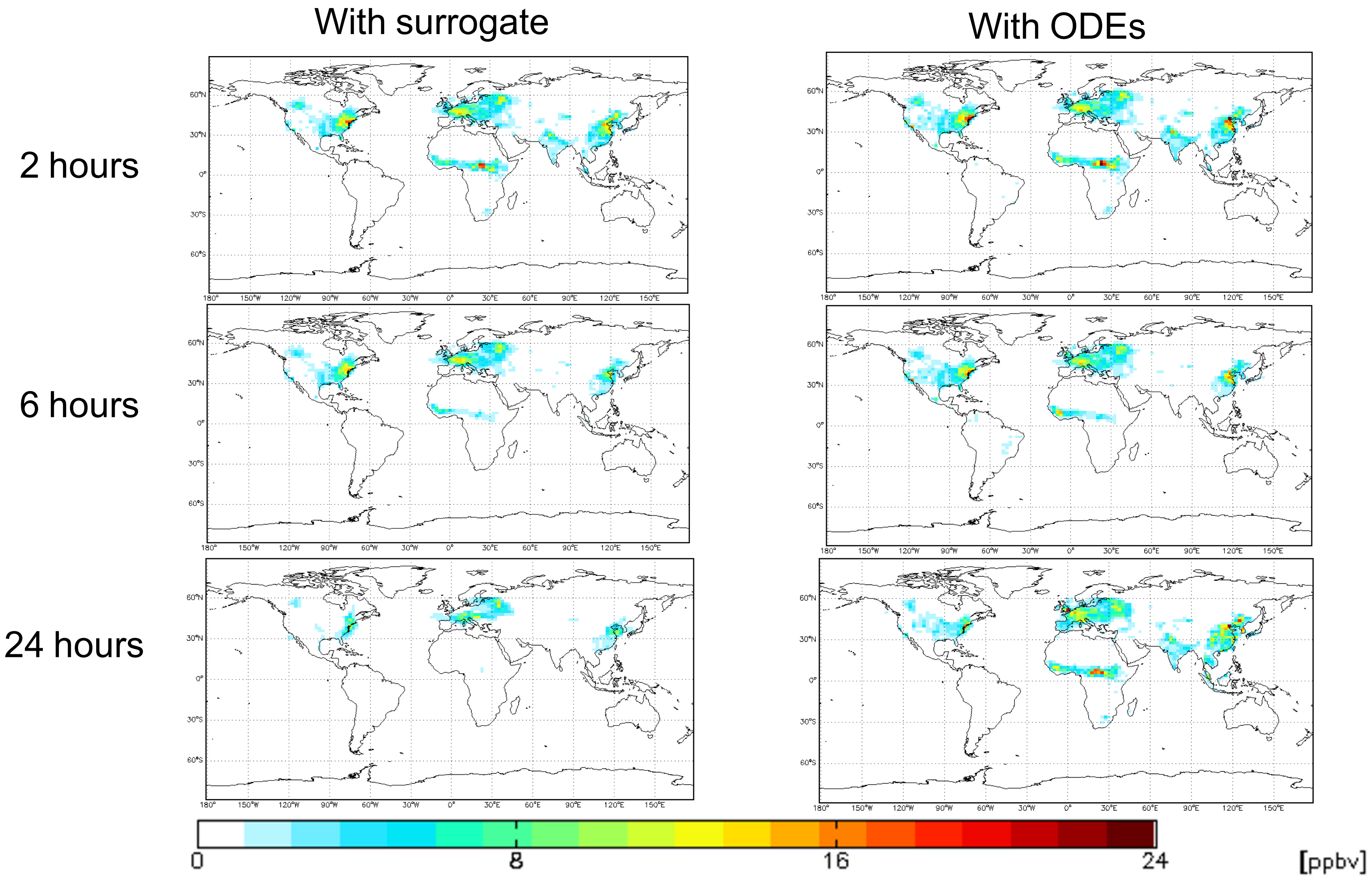
[ppbv]

One real-scaled surrogate per species

Significant digits of accuracy (SDA) - grid cell modified root mean square norm [Sandu et al. (1997), Henze et al. (2007)], typical SDA of ODE solver 1.8-2.0

Technical and Science Advancements

Global GEOS-Chem simulations of NO_2 for ODEs and surrogates over 24 hours

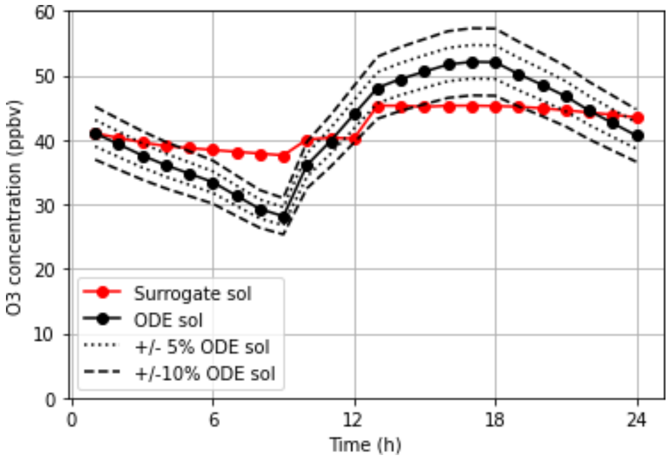


Technical and Science Advancements

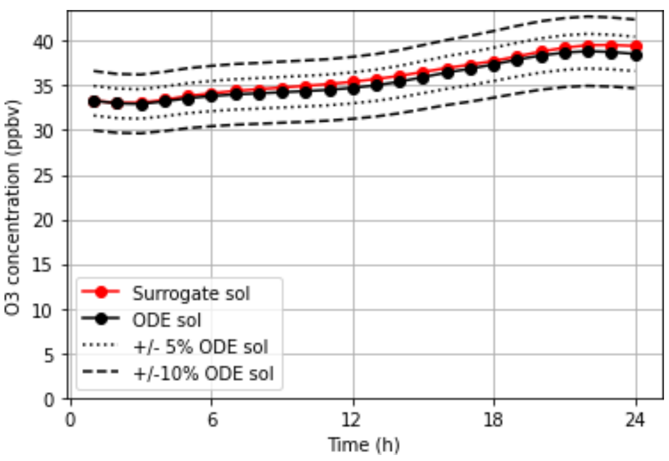
Single grid cell comparison for GEOS-Chem + surrogate (remote regions)

Surrogate predictions of O_3 are largely within 10% difference and not divergent from the ODE solution over 24 hours for remote regions

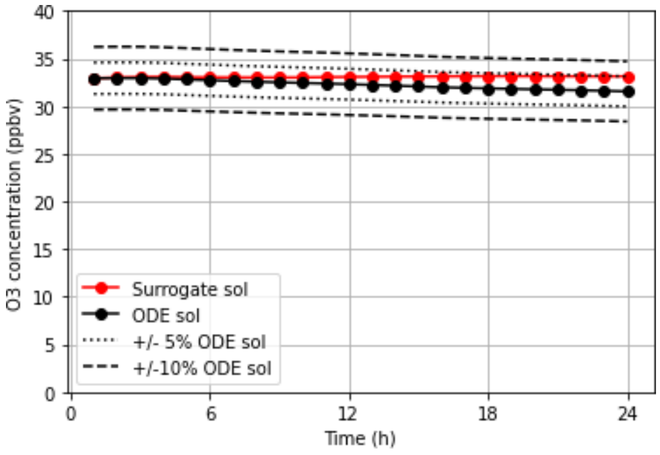
LAT = 14, LON = -15 (remote)



LAT = 42, LON = -50 (remote)



LAT = 60, LON = 90 (remote)

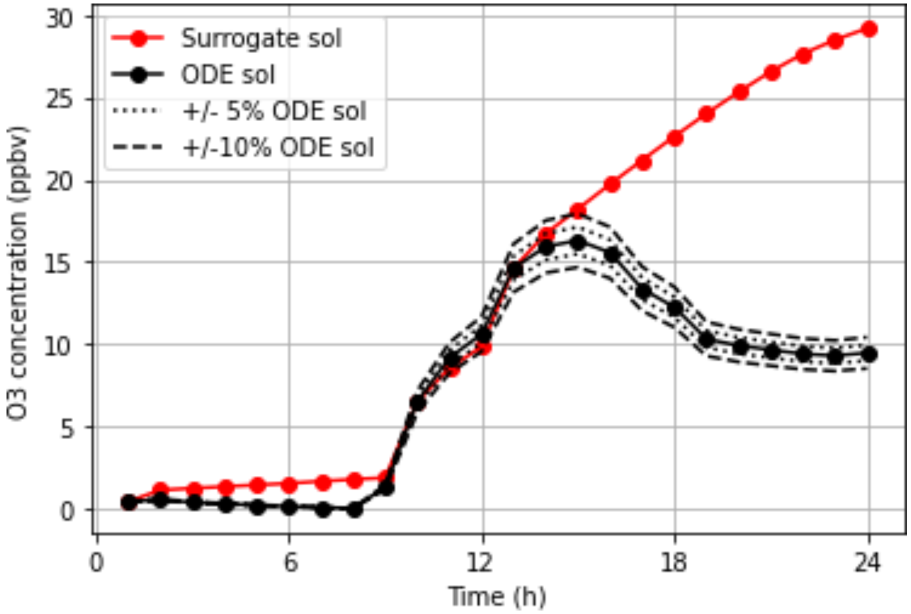


Technical and Science Advancements

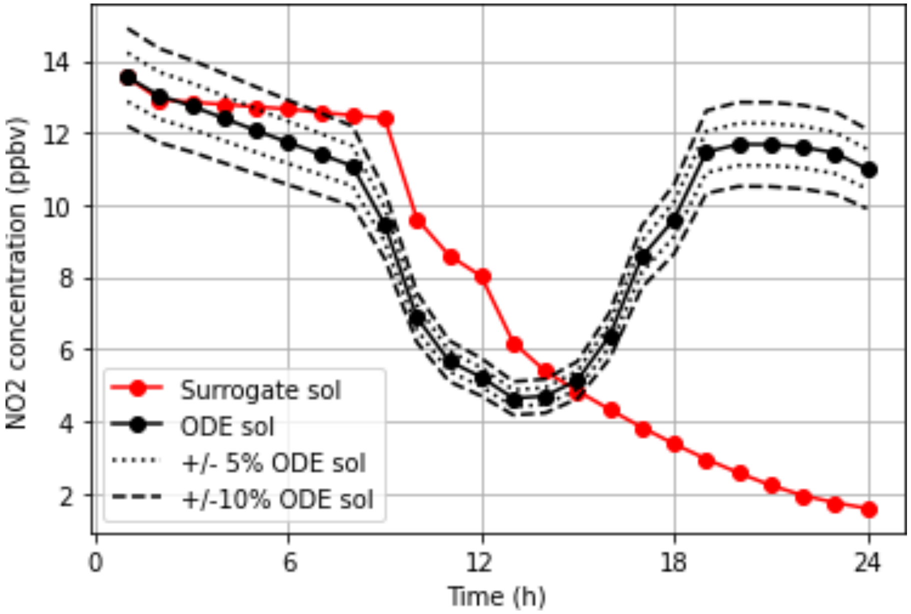
Single grid cell comparison for GEOS-Chem + surrogate (urban regions)

Divergence from ODE solution for the surrogate corresponds to inaccurate NO_2 predictions in many urban regions

O_3 at LAT = 48, LON = 2.5 (urban)

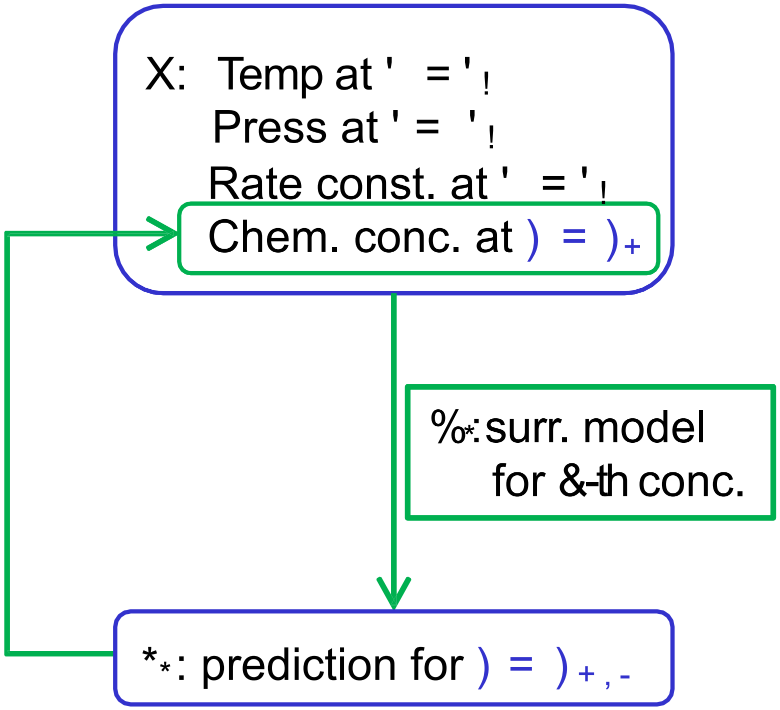


NO_2 at LAT = 48, LON = 2.5 (urban)



Technical and Science Advancements

- **Main Issue:** Inaccuracy of time-transient O3 predictions from surrogate models (Kelp et al, 2020; Sturm and Wexler, 2020)

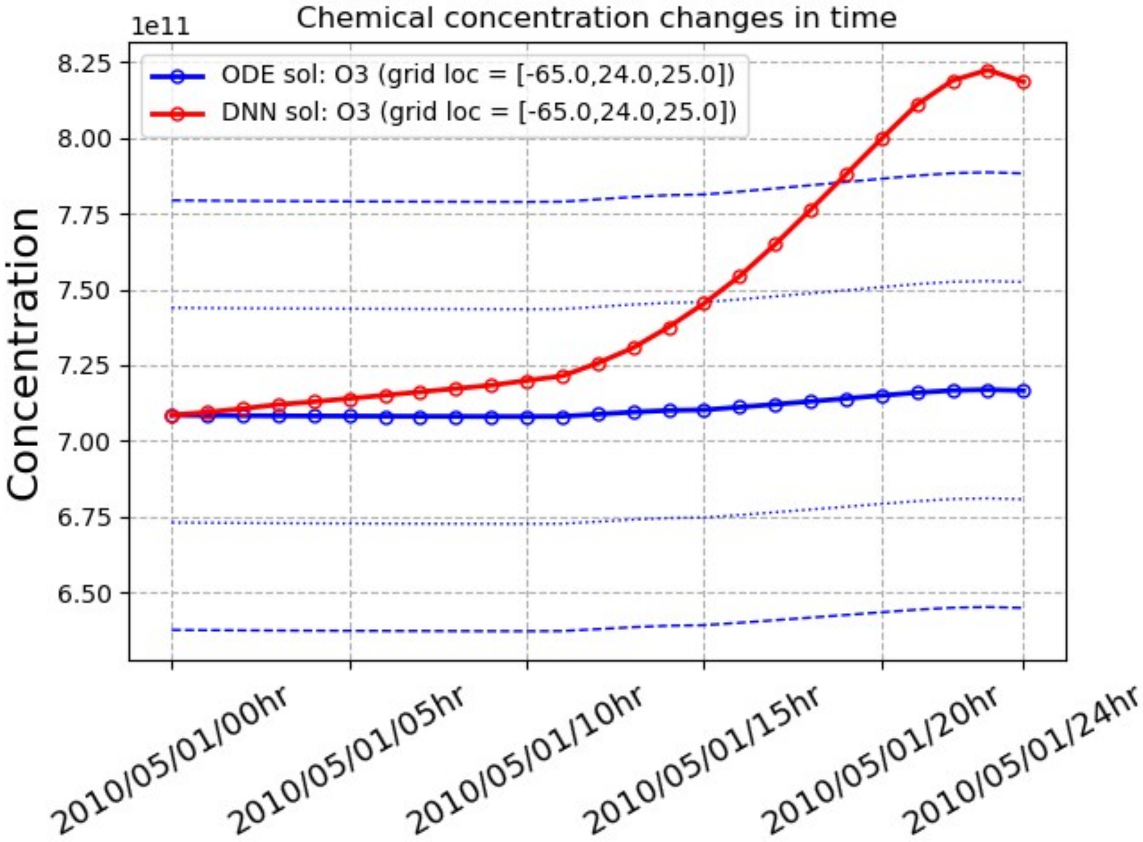


X: Temp at '   = ' !
Press at ' =   ' !
Rate const. at '   = ' !
Chem. conc. at) =) +

%*: surr. model
for &-th conc.

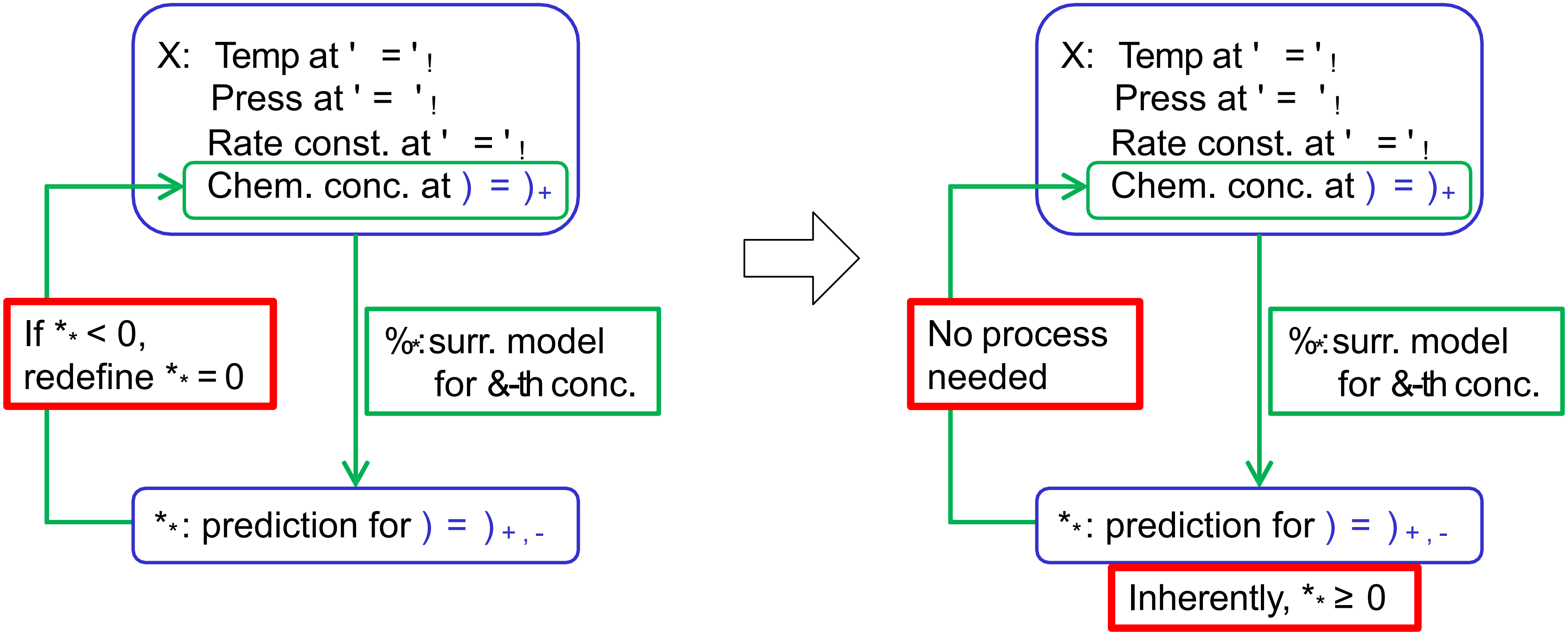
**: prediction for) =) + , -

Surr. Model (Ver.0)



Technical and Science Advancements

- **DNN updates to improve the time-transient O3 prediction**
  (1) Nonnegativity constraint (Ver. 1A)



X: Temp at ' = '!
Press at ' = '!
Rate const. at ' = '!
Chem. conc. at) = $)_+$

If $*_*$ < 0,
redefine $*_*$ = 0

%*: surr. model
for &-th conc.

$*_*$: prediction for) = $)_{+,-}$

X: Temp at ' = '!
Press at ' = '!
Rate const. at ' = '!
Chem. conc. at) = $)_+$

No process
needed

%*: surr. model
for &-th conc.

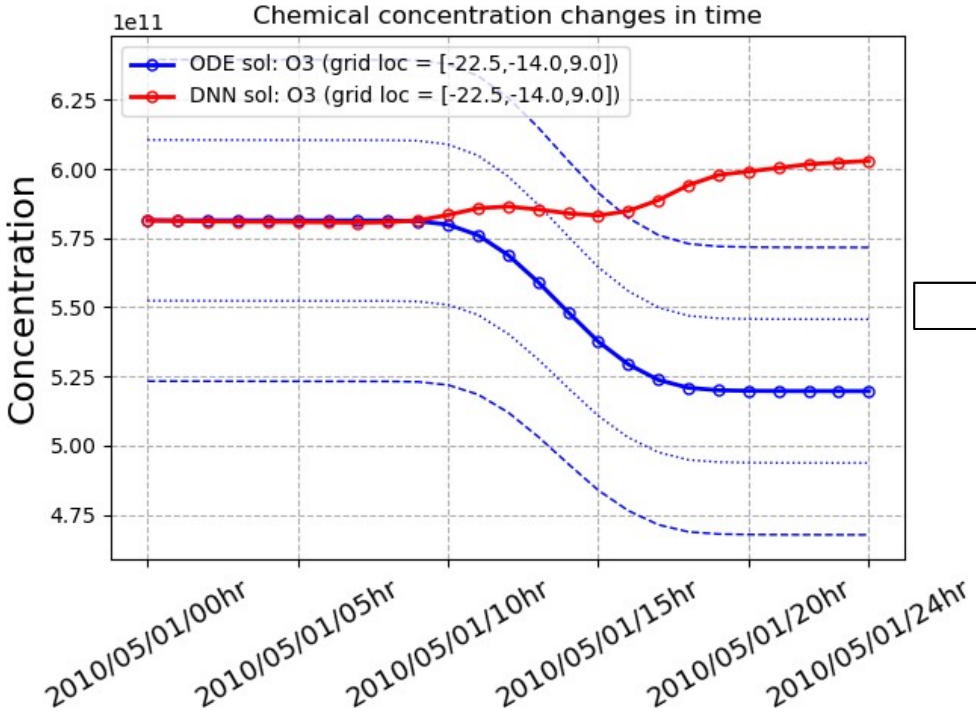
$*_*$: prediction for) = $)_{+,-}$

Inherently, $*_*$ ≥ 0

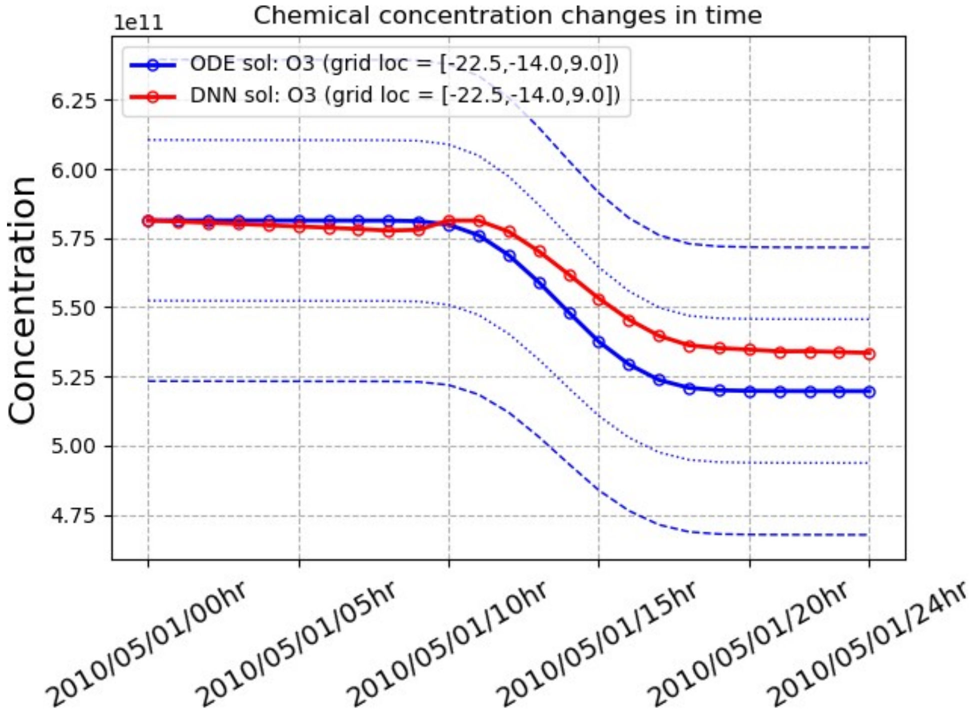
Technical and Science Advancements

- **DNN updates to improve the time-transient O3 prediction**
  (1) Nonnegativity constraint (Ver. 1A)

Before imposing the constraint



After imposing the constraint



Technical and Science Advancements

- **DNN updates to improve the time-transient O3 prediction**

  (1) Nonnegativity constraint

  (2) Chemical / physical regimes

  - By O3 production regime (NOx or VOC limited)
  - By day / night

Technical and Science Advancements

- To improve accuracy of O_3 predictions, we apply regime-specific surrogate models

For high [VOC]/[NO_x],

$$\frac{,\ [-_{(}\,]}{{}_{,}\,'} \propto [/\ -\ _{.}\ ]$$

This is the NO_x-limited regime.

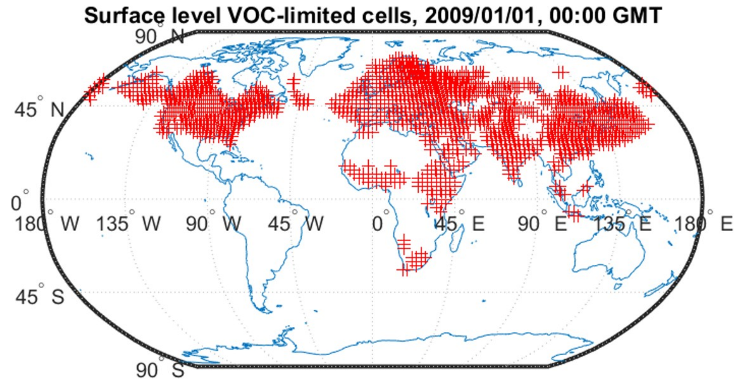
For low [VOC]/[NO_x],

$$\frac{,\ [-_{(}\,]}{{}_{,}\,'} \propto [0 - !\ ], \frac{1}{[/\ -\ _{.}\ ]}$$

This is the VOC-limited regime.

Following the work of Duncan et al. (2010), we determine the regime in each grid cell of GEOS-Chem using input formaldehyde to NO_2 concentration ratios:

- [Form]/[NO_2] < 1 à NO_x-limited
- [Form]/[NO_2] > 2 à VOC-limited
- 1 $\leq$ [Form]/[NO_2] $\leq$ 2 à Neither regime

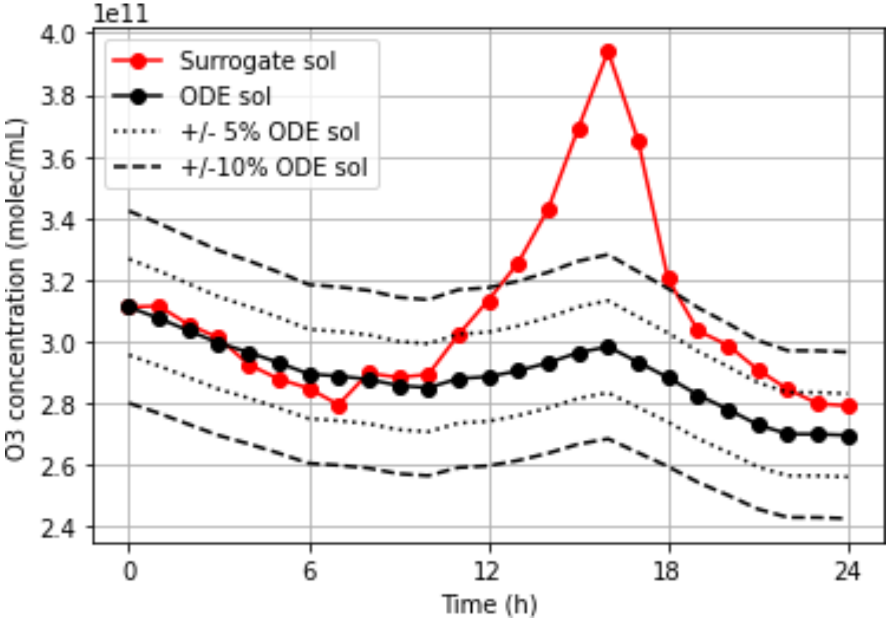


Surface level VOC-limited cells, 2009/01/01, 00:00 GMT

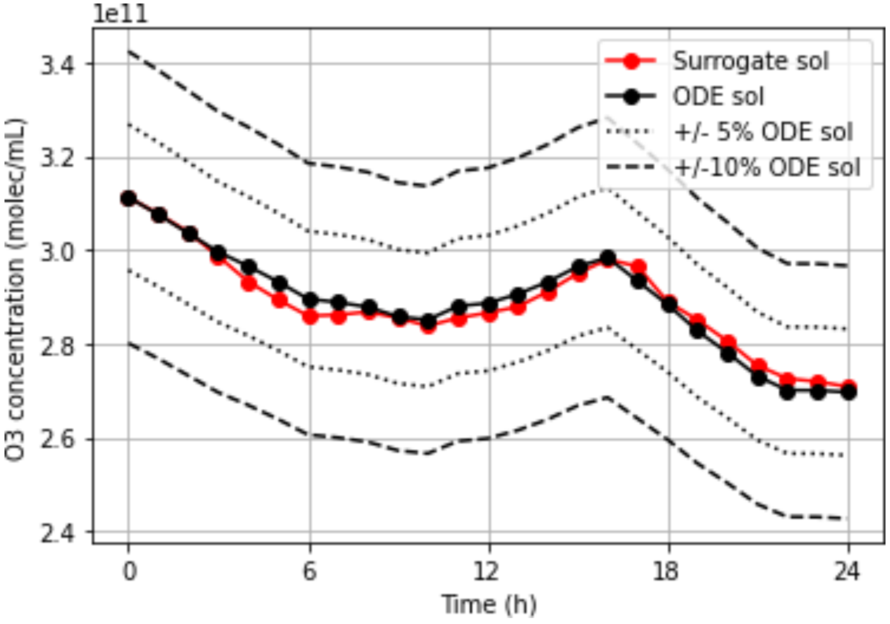
Technical and Science Advancements

- Applying different regimes of O_3 production improves hourly O_3 predictions

Without NO_x/VOC-limited regimes



With NO_x/VOC-limited regimes



- Since regimes are defined by formaldehyde and NO_2 concentrations, the use of the correct surrogate model for O_3 prediction is highly dependent on the accuracy of formaldehyde and NO_2 surrogate models

Technical and Science Advancements

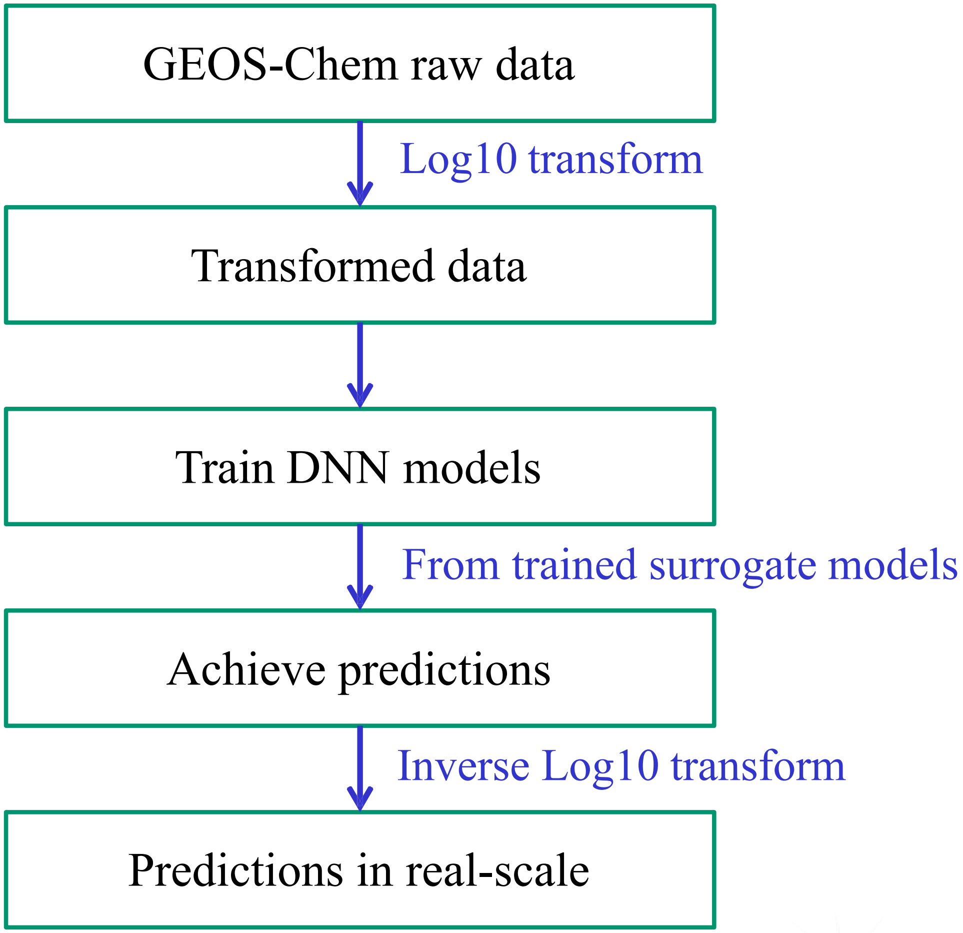
- **DNN updates to improve the time-transient O3 prediction**

  (1) Nonnegativity constraint

  (2) Chemical / physical regimes

  (3) Wide ranges of air condition values

  à **Log-scaled training (Ver. 1C)**



GEOS-Chem raw data

Log10 transform

Transformed data

Train DNN models

From trained surrogate models

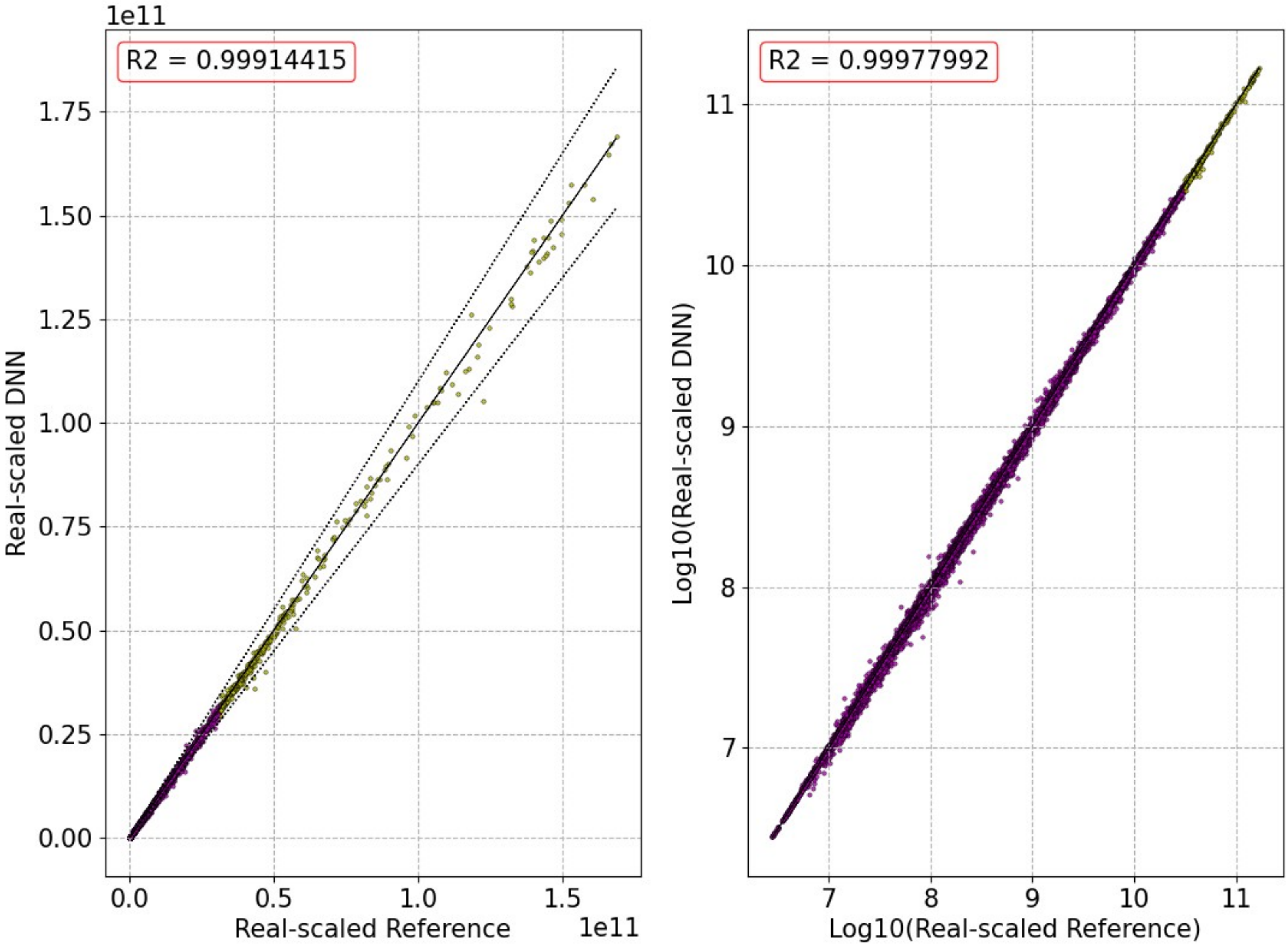
Achieve predictions

Inverse Log10 transform

Predictions in real-scale

Technical and Science Advancements

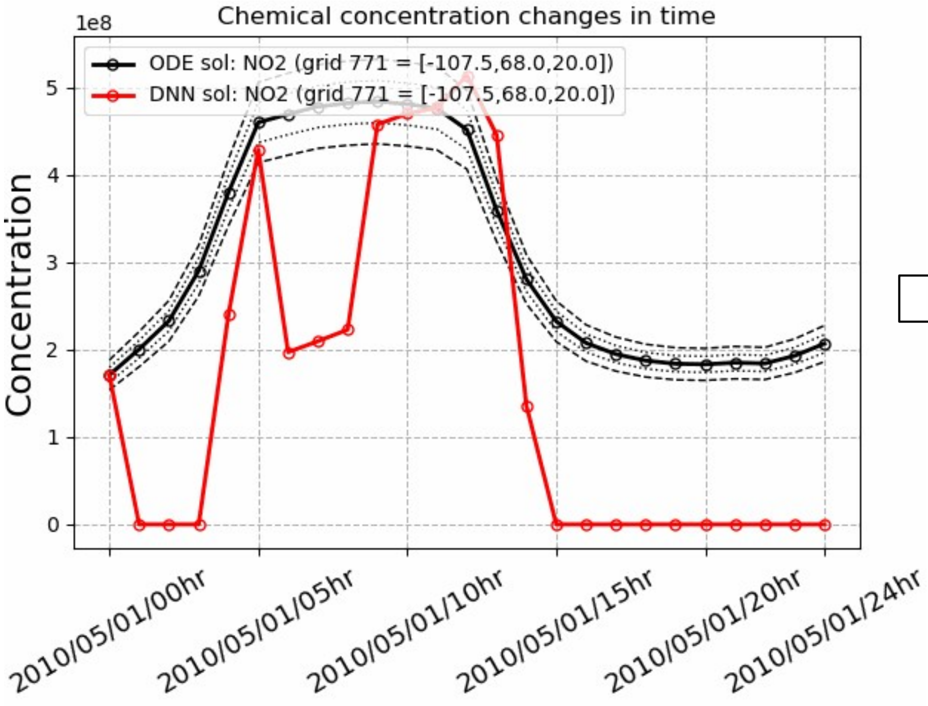
- Log-scaled training results (NO2) – Ver. 1C



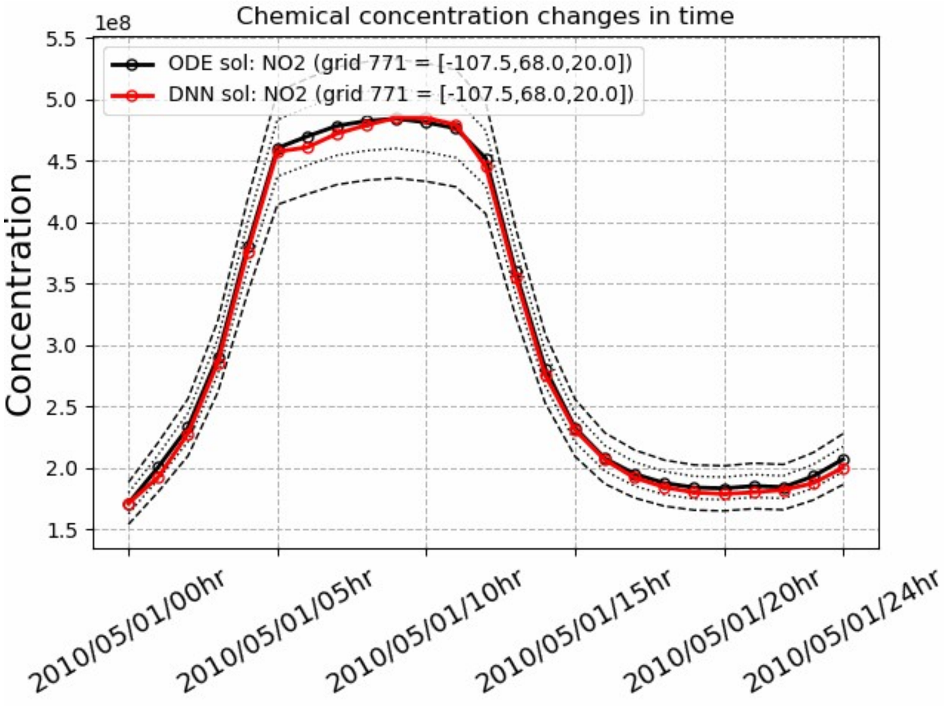
Technical and Science Advancements

- One hour update solutions from NO2 surrogate models

➢ Real-scaled training (Ver. 0)



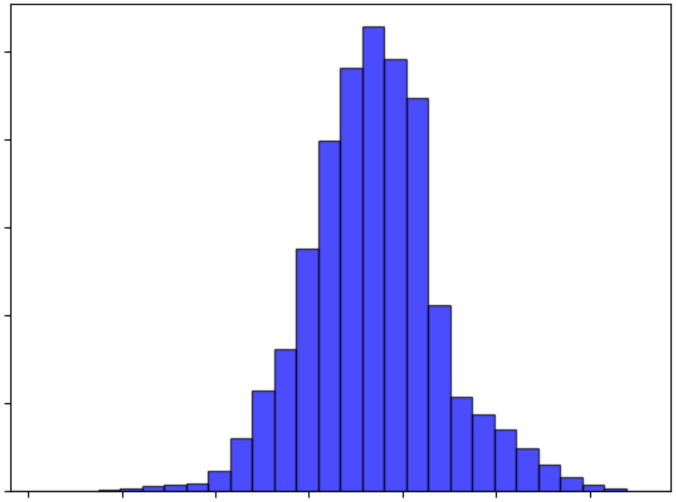
➢ Log-scaled training (Ver. 1C)



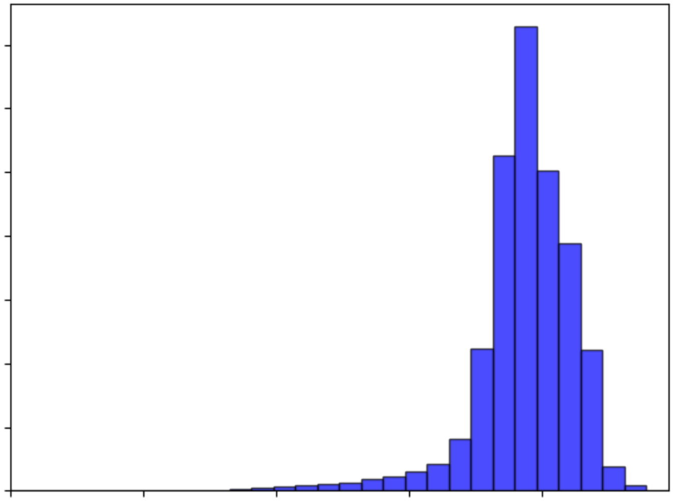
Technical and Science Advancements

- **DNN updates to improve the time-transient O3 prediction**
  (1) Nonnegativity constraint (Ver. 1A)
  (2) Chemical / physical regimes (Ver. 1B)
  (3) Wide ranges of air condition values
  à **Log-scaled training (Ver. 1C)**
  **** Categorization of chemical species depending on the data distributions**

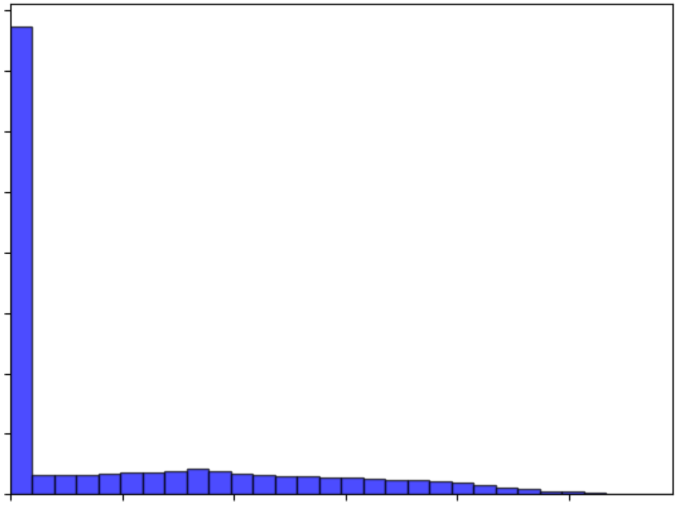
Normal-like



Long shallow front

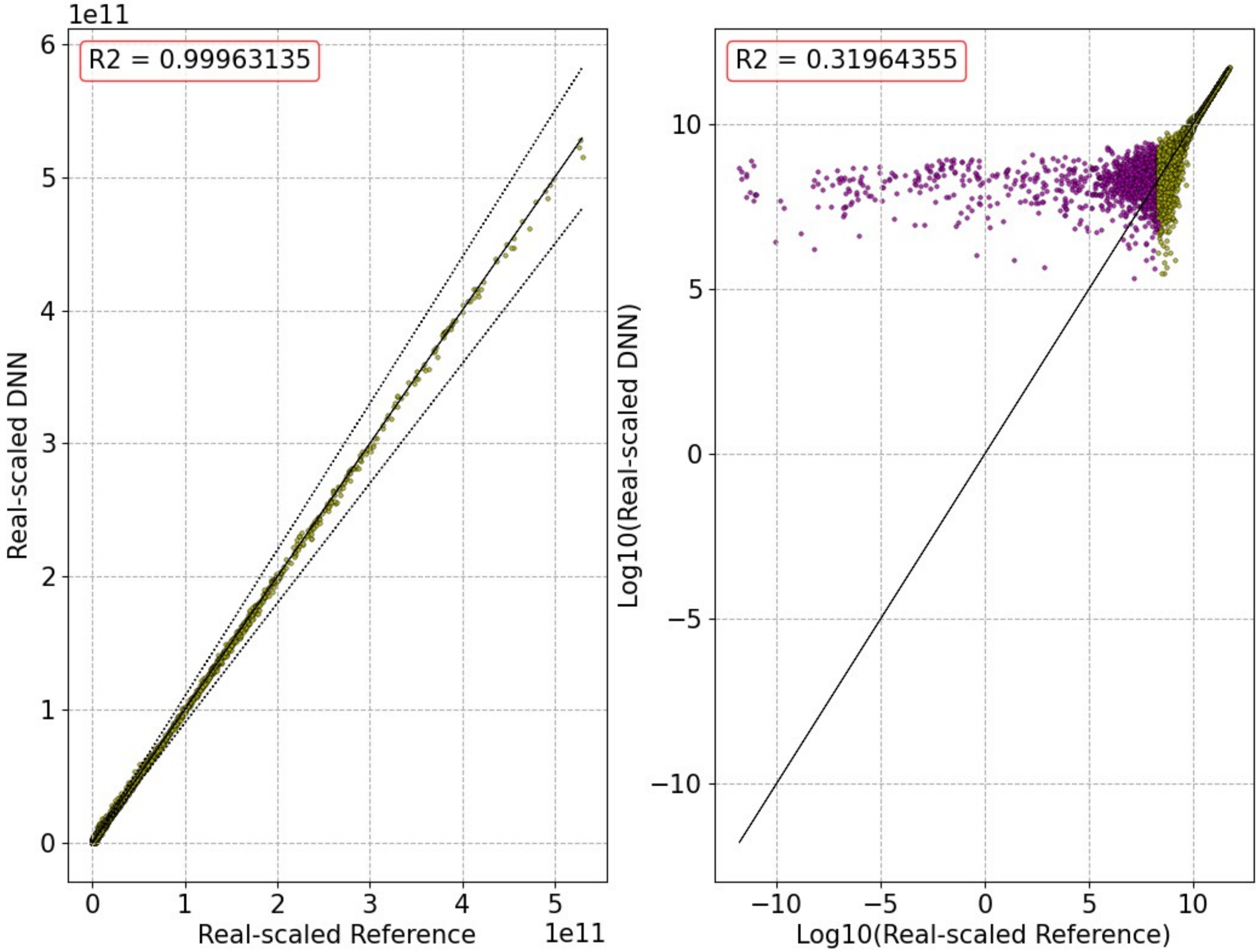


ISOP-like



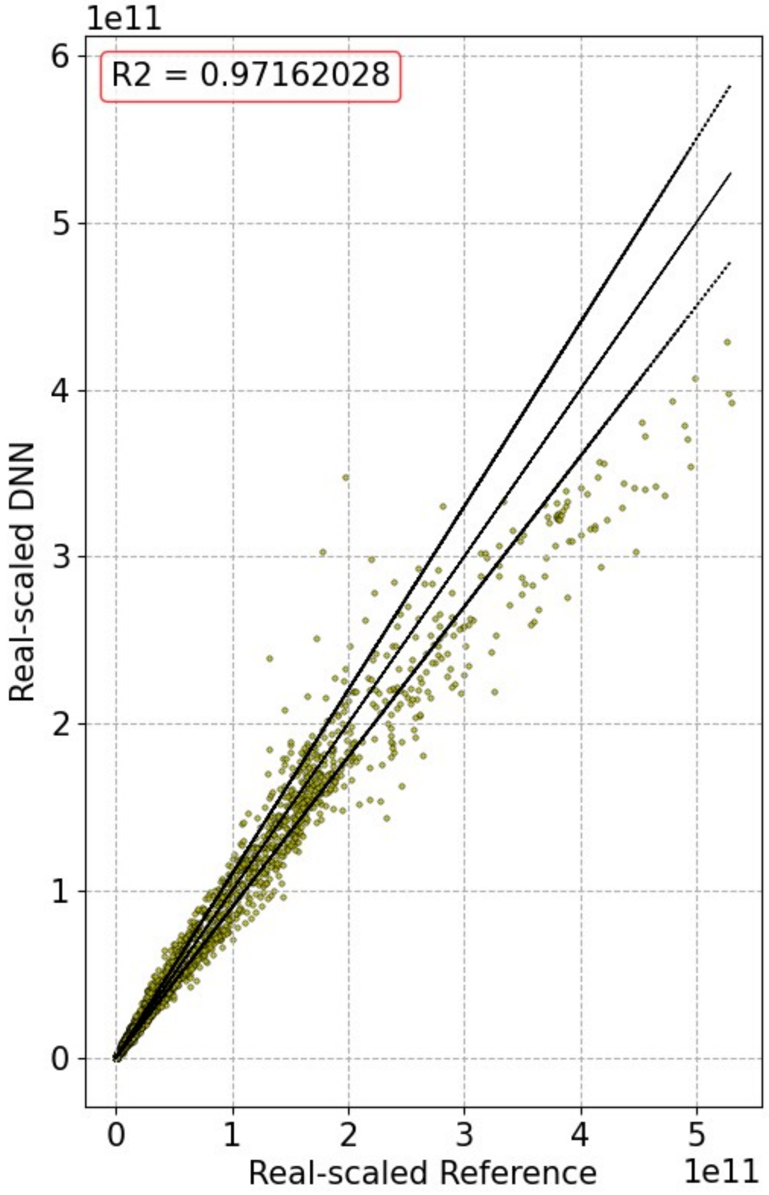
Technical and Science Advancements

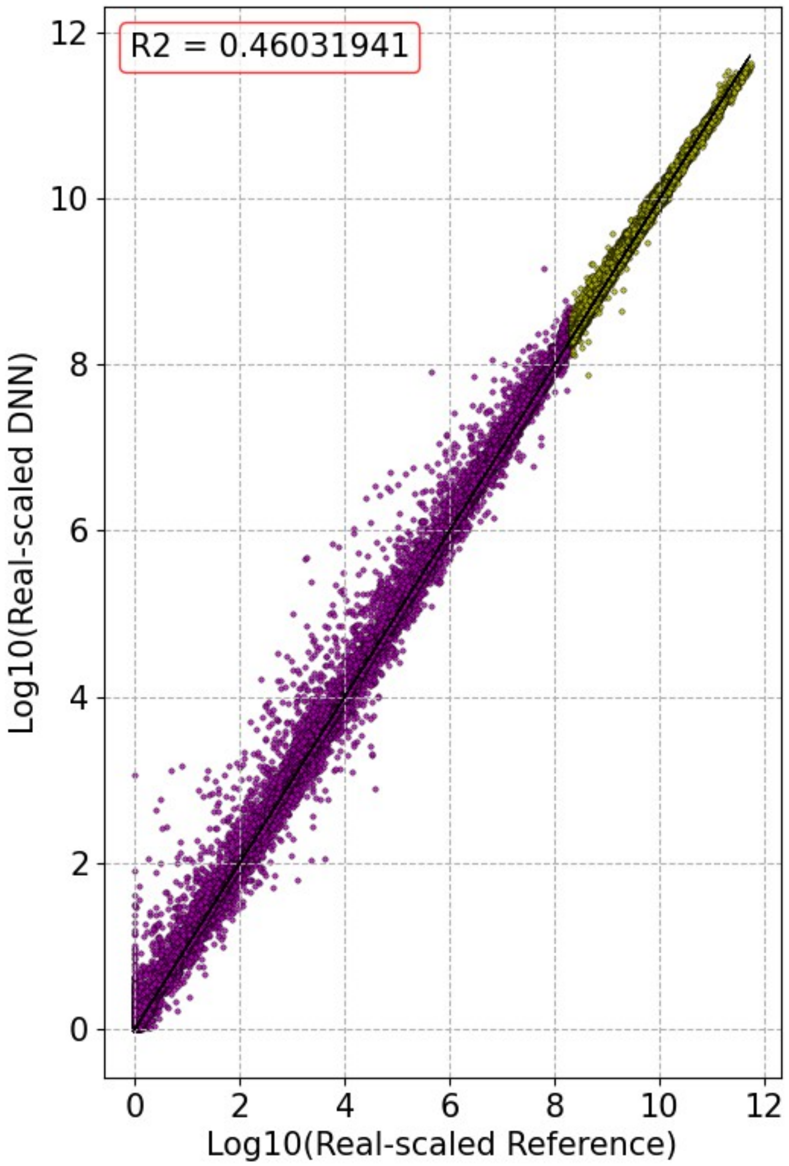
- Log-scaled training results (ISOP) – Ver. 0



Technical and Science Advancements

- Log-scaled training results (ISOP) – Ver. 1C





Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Summary

- Surrogate model development
  - Ensemble of DDN models (one for each individual concentration)
  - Tested / developed in box-model R&D using samples from 3D model

- 3D implementation:
  - Initial implementation (Ver 0) reasonable for a few hrs followed by error growth
  - Now adding flexibility to accommodate multiple surrogate model versions
  - Current computational cost 10% savings w/o any optimization or further parallelization

- Surrogate model updates for improved accuracy:
  - Non-negative constraints
  - Chemical and physical regimes
  - Log-scaling

- Next steps:
  - Evaluate updates (Ver 1A - 1C) in 3D for accuracy and efficiency
  - Explore additional ideas for enhanced stability / accuracy (e.g., time-dependent training)
  - Apply to chemical data assimilation with GEOS-Chem
  - Apply to other models: collect samples from CAMS model (ECMWF collaborator)

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Publications - List of Acronyms

Publications

Conference presentations:

- AMS Atmospheric Chemistry, January, 2021
    - Session on “Machine-learning Applications for Atmospheric Chemistry”

Acronyms

List of Acronyms

3D	3 Dimensional
4D-Var	4Dimensionoal Variational Data Assimilation
AQ	Air Quality
CAMS	Copernicus Atmosphere Monitoring Service (ECMWF’s AQ model)
DNN	Deep Neural Network
ECMWF	European Centre for Medium-Range Weather Forecasts
GCHP	GEOS-Chem High Performance
GEOS	Goddard Earth Observing System
NAQFC	National Air Quality Forecast Center (US national AQ forecasts from NOAA)
ODE	Ordinary Differential Equation
PCE	Polynomial Chaos Expansion
RNN	Recurrent Neural Network
TEMPO	Tropospheric Emissions: Monitoring of Pollution

Development of the High Performance Version of GEOS-Chem (GCHP) to enable broad community access to high-resolution atmospheric chemistry modeling in support of NASA Earth Science

Randall Martin (Washington University)

with contributions (alphabetical) from
Liam Bindle (WashU), Tom Clune (NASA GSFC), Will Downs (Harvard), Sebastian Eastham (MIT), Daniel Jacob (Harvard), Christoph Keller (NASA GSFC), Lizzie Lundgren (Harvard), Jun Meng (WashU/Dalhousie), Steven Pawson (GMAO), Bob Yantosca (Harvard), Jiawei Zhuang (Harvard)

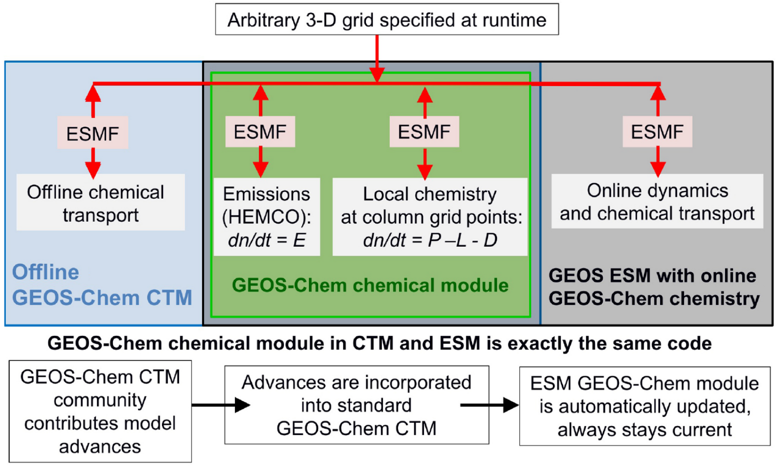
AIST-18-0011 Annual Technical Review

January 22, 2021

Development of the High Performance Version of GEOS-Chem (GCHP) to Enable Broad Community Access to High-resolution Atmospheric Chemistry Modeling in Support of NASA Earth Science

PI: Randall Martin, Washington University

- Develop the High Performance Version of GEOS-Chem (GCHP), a global 3-D chemical transport model, to enable broad community access to high-resolution atmospheric chemistry modeling and chemical data assimilation
- Make GCHP highly accessible by the atmospheric chemistry community to enable the atmospheric chemistry community to better exploit the GEOS system.
- Integrate the following technologies: high performance atmospheric chemistry model; Earth System Modeling Framework; cubed sphere meteorology; stretched grid; multi-node cloud capability; software build system generator; software package manager; software containers.



Arbitrary 3-D grid specified at runtime

ESMF

Offline chemical transport

Offline GEOS-Chem CTM

ESMF

Emissions (HEMCO): *dn/dt = E*

ESMF

Local chemistry at column grid points: *dn/dt = P –L - D*

GEOS-Chem chemical module

ESMF

Online dynamics and chemical transport

GEOS ESM with online GEOS-Chem chemistry

GEOS-Chem chemical module in CTM and ESM is exactly the same code

GEOS-Chem CTM community contributes model advances

Advances are incorporated into standard GEOS-Chem CTM

ESM GEOS-Chem module is automatically updated, always stays current

Schematic of GEOS-Chem chemical module used offline as a chemical transport model or online in an Earth system model with interfaces managed through the Earth system modeling framework.

Make this high-performance version of GEOS-Chem highly accessible by:

- Updating to the current version of the Modeling Analysis and Prediction Layer (MAPL) and enabling seamless updates.
- Improving GCHP performance and portability.
- Generating an operational cubed-sphere archive of GEOS assimilated meteorological data.

- Updated the current MAPL and improved the build system. 05/20
- Developed initial cubed-sphere archive of GEOS assimilated met data. 11/20
- Improved installation through a package manager and software containers. 11/20
- Implement an operational cubed-sphere archive 05/21
- Implement a stretched grid capability in GCHP 09/21

Co-Is/Partners: Daniel Jacob, Harvard; Tom Clune, Christoph Keller, GMAO; Steven Barrett, Sebastian Eastham, MIT

TRL_{in} = 3 **$\text{TRL}_{\text{current}}$ = 5**

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Background and Objectives

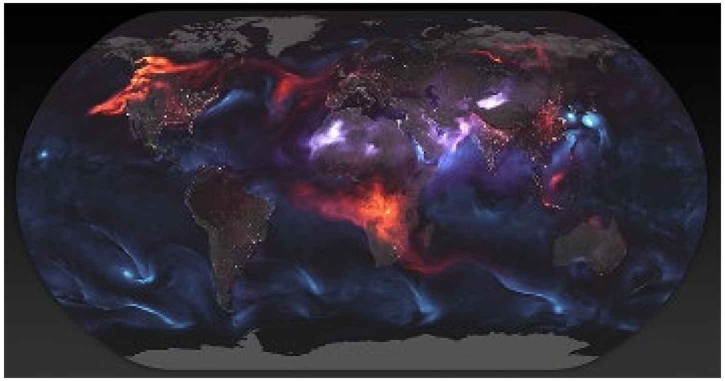
- This project helps meet the R&A and Applications science goals for several cross cutting science areas (highest relevance bolded)
  - **Carbon Cycle**; **Climate Variability**; Water & Energy; **Atmospheric Comp**; Weather; **Eco Forecasting**; Disasters; **Health & Air Quality**; Energy Management; Water & Food; **Fires;** Planetary Boundary Layer; Snow and Ice;

- Overall goal to develop the High Performance Version of GEOS-Chem (GCHP) to enable broad community access to high-resolution atmospheric chemistry modeling and chemical data assimilation

- Performance goals include fully parallelizing the model and enabling the atmospheric chemistry community to better exploit the GEOS system

Objectives

GEOS/GMAO



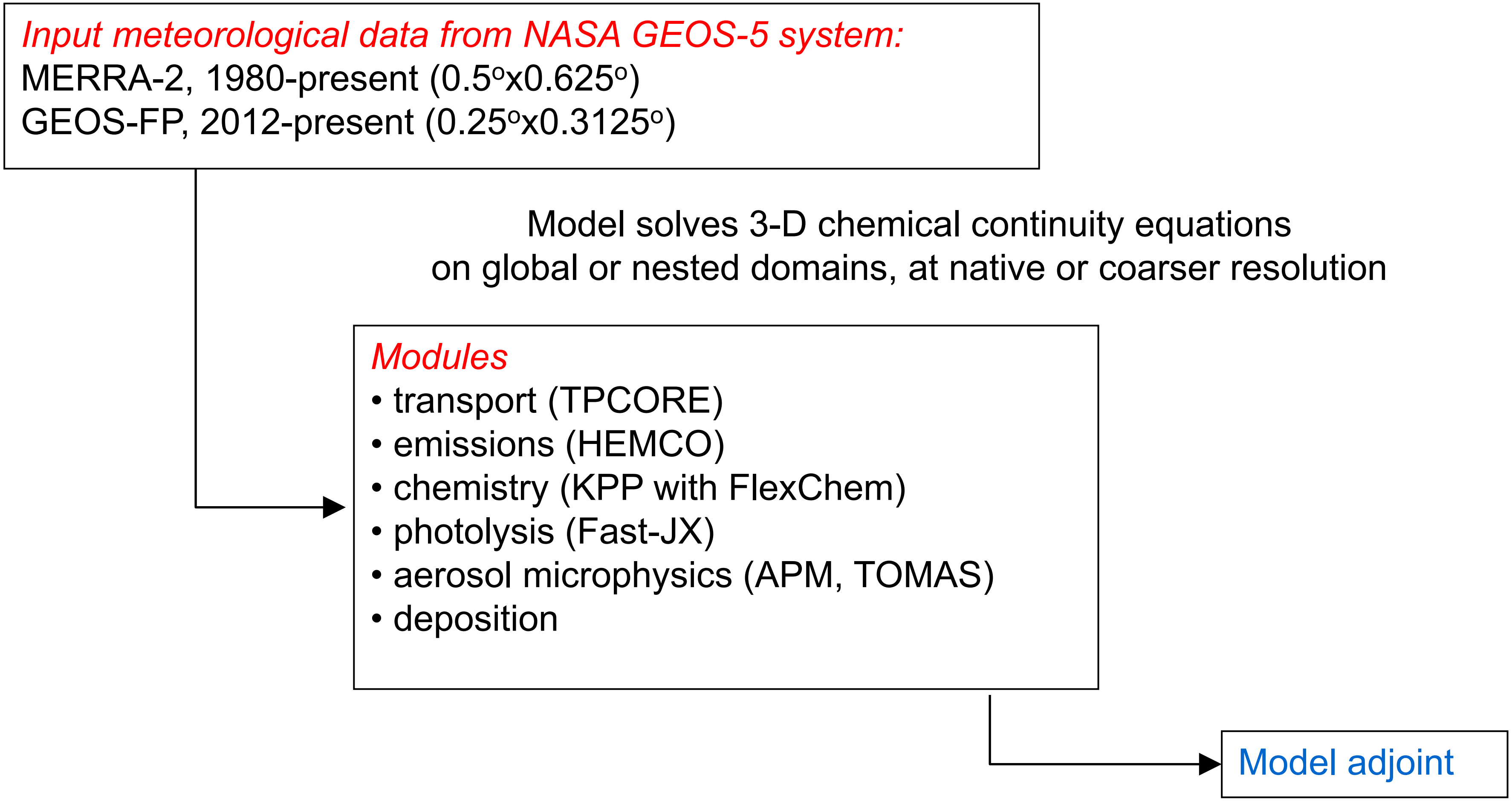
Sophisticated Meteorology & MAPL/ESMF Framework

GEOS Chem

Vibrant Community Seeking Tools to Keep Pace with GEOS

GEOS-Chem Off-line Atmospheric Chemistry Model

Detailed chemical simulation of troposphere and stratosphere



Input meteorological data from NASA GEOS-5 system:
MERRA-2, 1980-present (0.5°x0.625°)
GEOS-FP, 2012-present (0.25°x0.3125°)

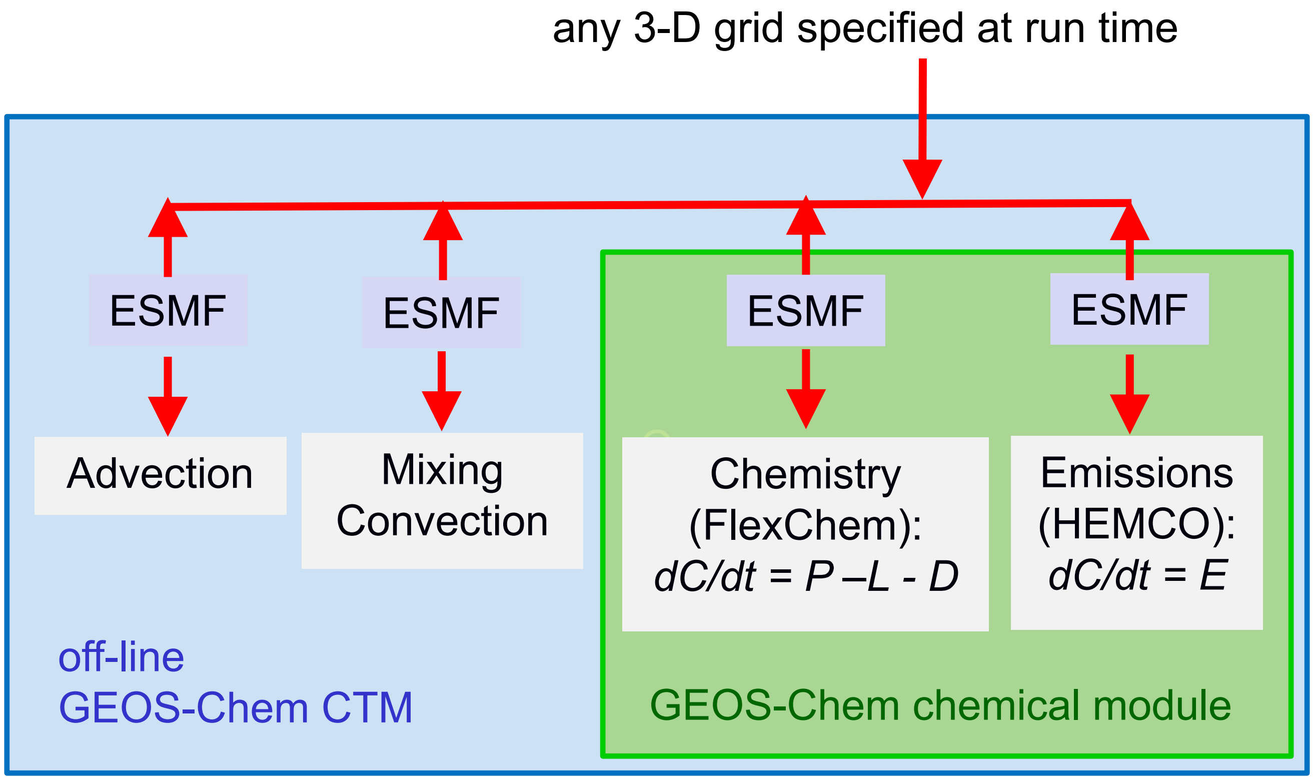
Model solves 3-D chemical continuity equations
on global or nested domains, at native or coarser resolution

Modules

- transport (TPCORE)
- emissions (HEMCO)
- chemistry (KPP with FlexChem)
- photolysis (Fast-JX)
- aerosol microphysics (APM, TOMAS)
- deposition

Model adjoint

GEOS-Chem as Offline and Online Chemical Module



any 3-D grid specified at run time

ESMF

ESMF

ESMF

ESMF

Advection

Mixing Convection

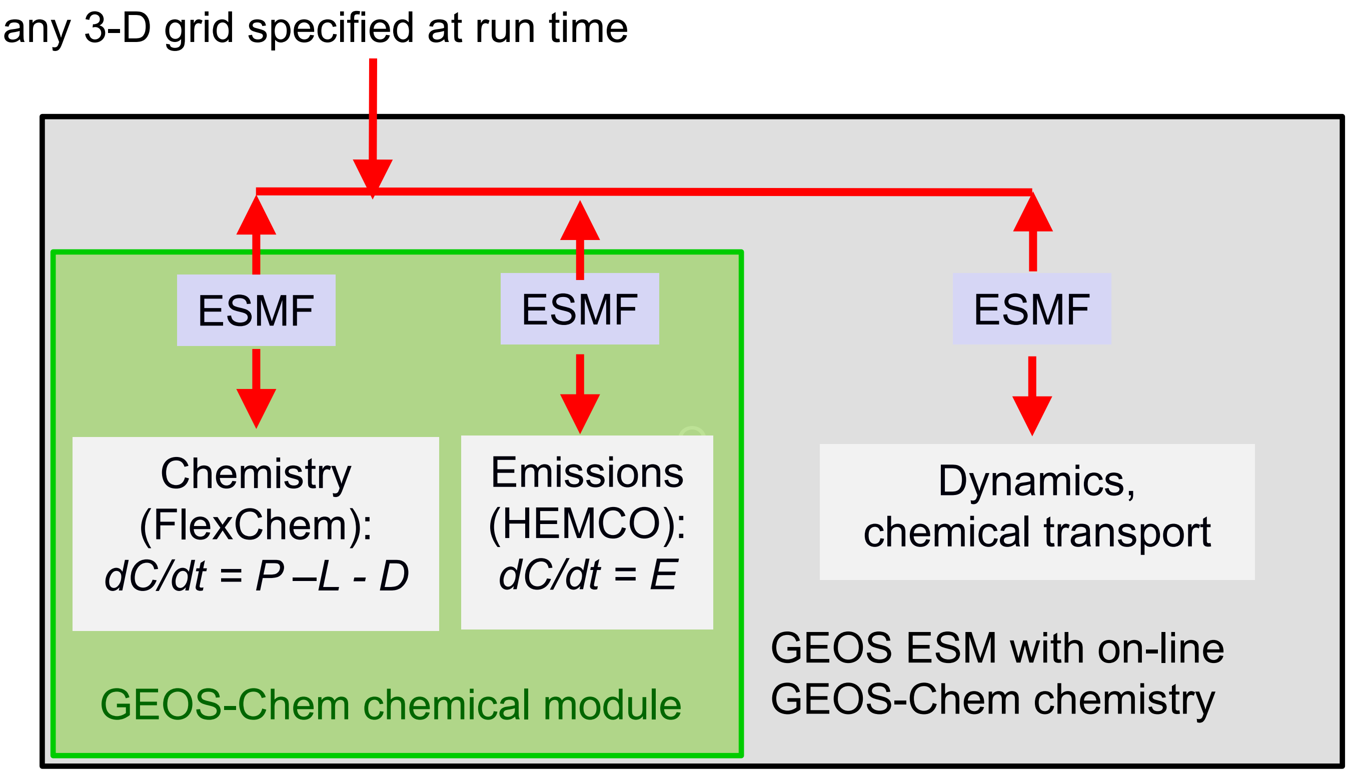
Chemistry (FlexChem): $dC/dt = P - L - D$

Emissions (HEMCO): $dC/dt = E$

off-line GEOS-Chem CTM

GEOS-Chem chemical module

GEOS-Chem as Offline and Online Chemical Module



any 3-D grid specified at run time

ESMF

ESMF

ESMF

Chemistry (FlexChem): *dC/dt = P –L - D*

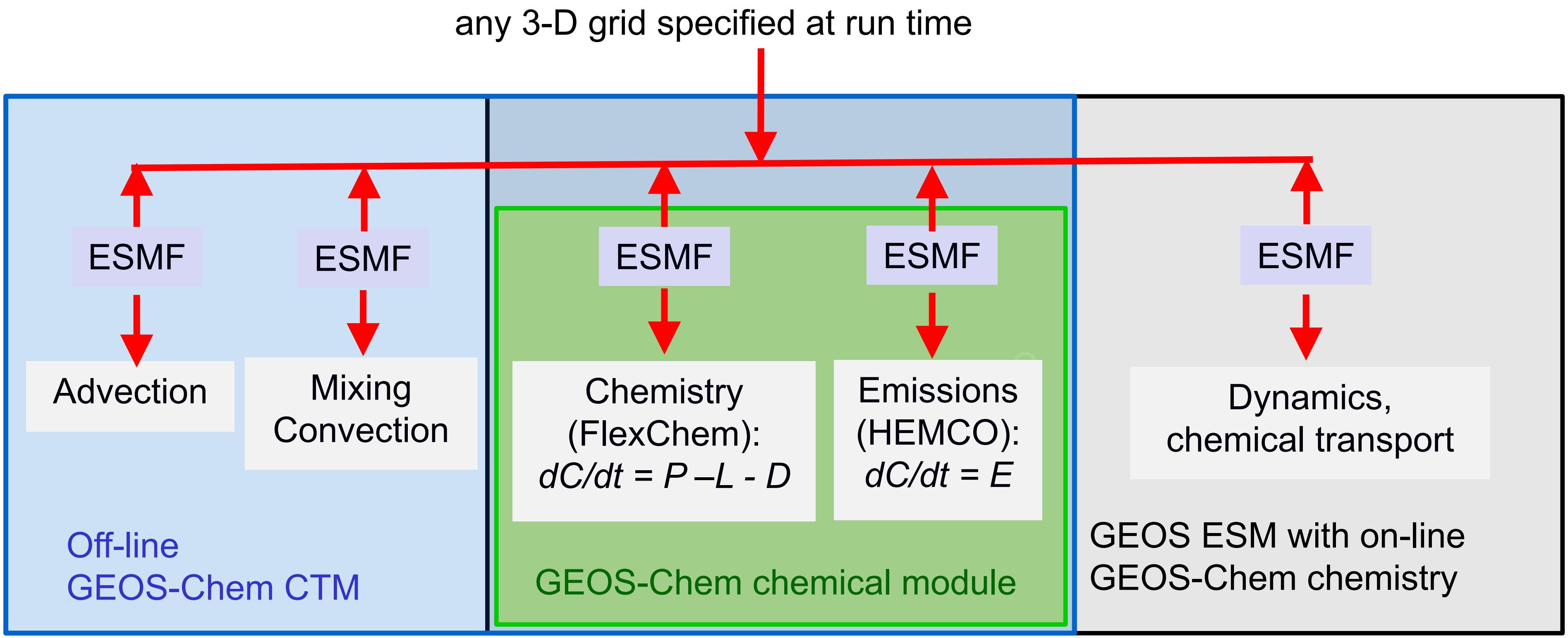
Emissions (HEMCO): *dC/dt = E*

Dynamics, chemical transport

GEOS-Chem chemical module

GEOS ESM with on-line GEOS-Chem chemistry

GEOS-Chem as Offline and Online Chemical Module



any 3-D grid specified at run time

ESMF

Advection

ESMF

Mixing Convection

Off-line GEOS-Chem CTM

ESMF

Chemistry (FlexChem): $dC/dt = P - L - D$

ESMF

Emissions (HEMCO): $dC/dt = E$

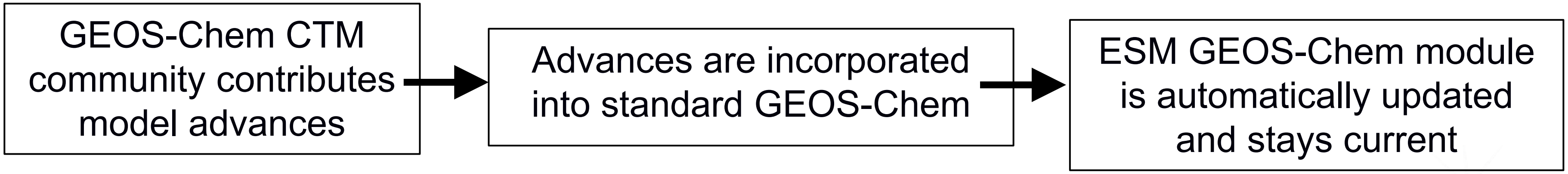
GEOS-Chem chemical module

ESMF

Dynamics, chemical transport

GEOS ESM with on-line GEOS-Chem chemistry

Off-line and on-line GEOS-Chem chemical modules use exactly the same code



GEOS-Chem CTM community contributes model advances

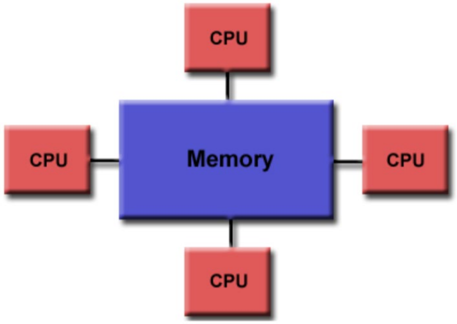
Advances are incorporated into standard GEOS-Chem

ESM GEOS-Chem module is automatically updated and stays current

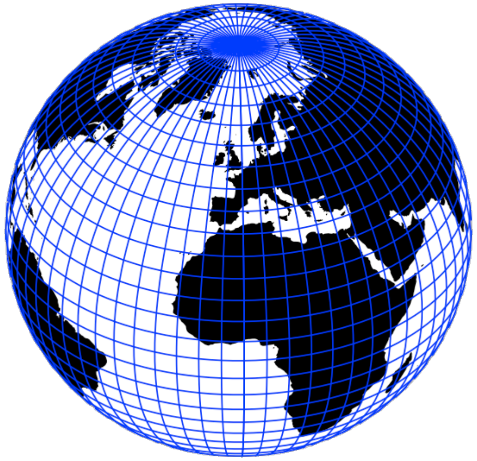
High-Performance GEOS-Chem (GCHP)

GEOS-Chem Classic

Inefficient above 16 Cores
Shared Memory (OpenMP)

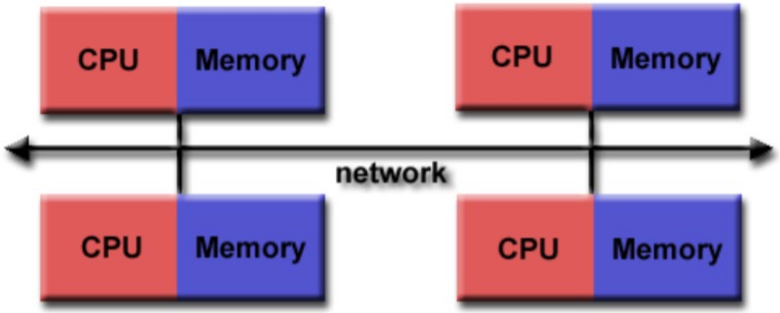


Regular lat-lon

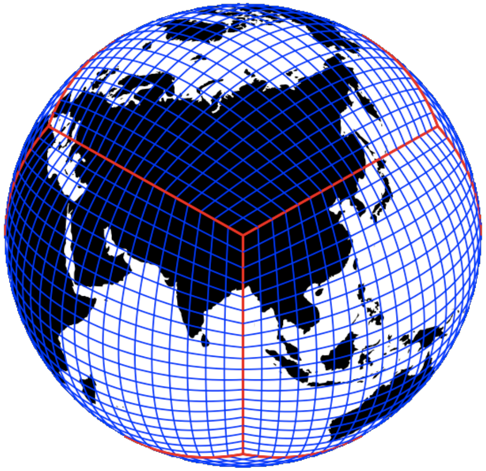


GCHP

Massively Parallel
Distributed Memory (MPI)

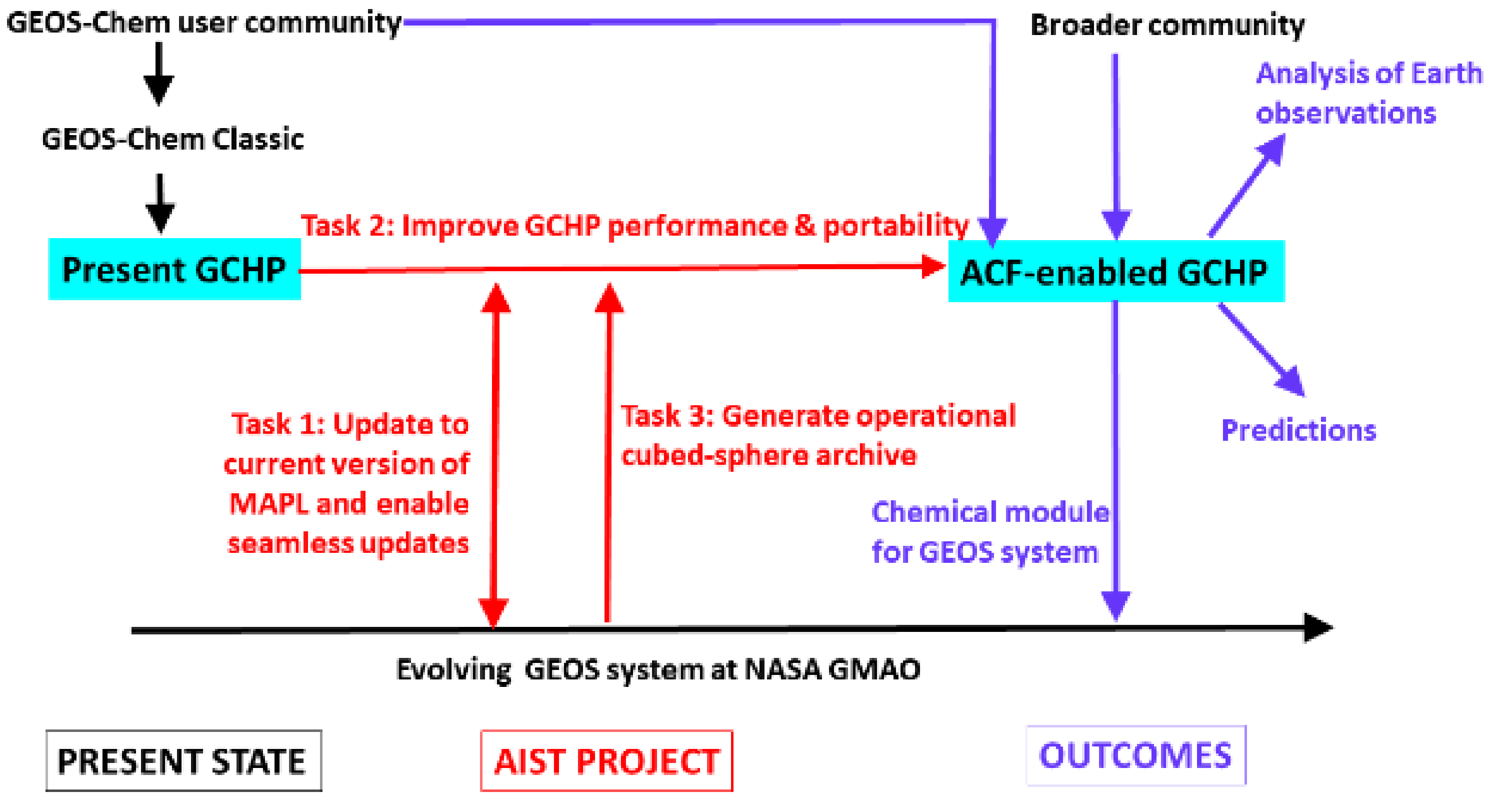


Cubed-sphere



Putman and Lin (2007)

Objectives



GEOS-Chem user community
GEOS-Chem Classic
Present GCHP
Task 2: Improve GCHP performance & portability
ACF-enabled GCHP
Broader community
Analysis of Earth observations
Predictions
Task 1: Update to current version of MAPL and enable seamless updates
Task 3: Generate operational cubed-sphere archive
Chemical module for GEOS system
Evolving GEOS system at NASA GMAO
PRESENT STATE
AIST PROJECT
OUTCOMES

Make the high-performance version of GEOS-Chem (GCHP) highly accessible by the atmospheric chemistry community in sustained partnership with GMAO. Allow the atmospheric chemistry community to better exploit the GEOS system through its applications of GEOS-Chem, and to advance atmospheric chemistry knowledge for the benefit of the GEOS system and NASA's Earth science mission.

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Updating MAPL and Enabling Seamless Updates

GEOS MAPL: ESMF-based software layer which handles communication between atmospheric domains. Initial implementation was manually integrated into GCHP and frozen.

Updating to MAPL 2.2.7 enabled

- Improved parallelization of regridding and I/O
- Improved error diagnostics
- Potential for stretched grid simulations

Using forks of GMAO software repositories as Git submodules enabled

- Seamless pulling of updates
- Promoted collaboration, e.g. grid-box corners, improved error handling

Lizzie Lundgren (Harvard),
Tom Clune (GMAO)

Improved Build System

Problem

- Building (compiling) GCHP was hard for users
- Set up on new cluster required expertise

Why is was hard

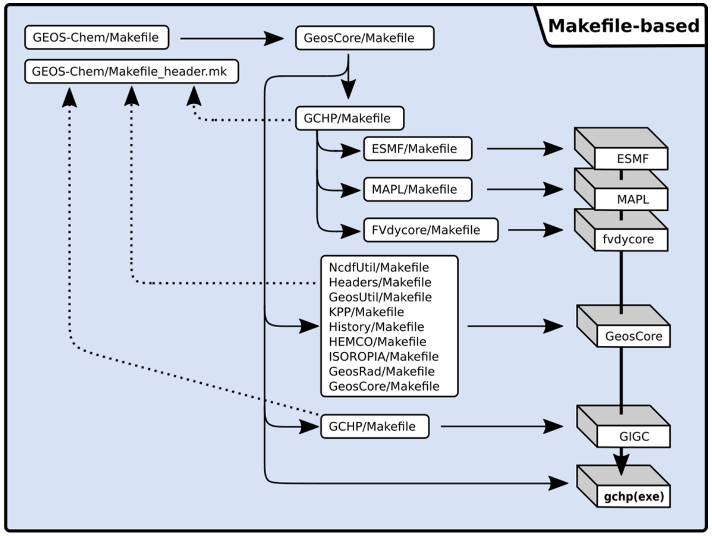
- Written in a low-level language (Make)
- Complex software stack (dependencies)
- Interorganizational code base

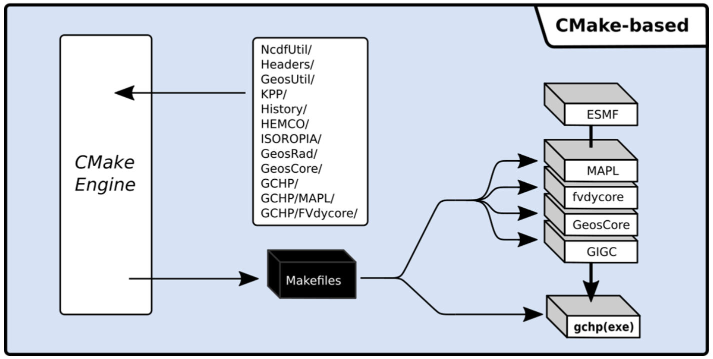
Work completed

- Completely overhauled the build system
  - Written in higher-level language (CMake)
- Higher-level functionality facilitates
  - Interfacing with MAPL's build system
  - A more structured build system
  - Automatically finds software dependencies

Impacts (feedback and experience)

- Much easier to build
- Procedure to build GCHP is simpler/streamlined
- Easier to support/troubleshoot user issues





Liam Bindle (WashU)

Implementation of the Spack Package Manager

- **Challenge**: installation of GCHP was complicated by multiple versions, configurations, platforms, and compilers

- **Spack**: innovative package manager designed to ease installation of scientific software

- **Spack implementation** now provides ‘recipes’ for GCHP dependencies
  - Compilers, MPI, NetCDF libraries, Cmake
  - Significantly streamlines system setup
  - Offers choice of compilers and MPI implementations
  - Includes updated ESMF version

- **Instructions now available** on GCHP Read The Docs

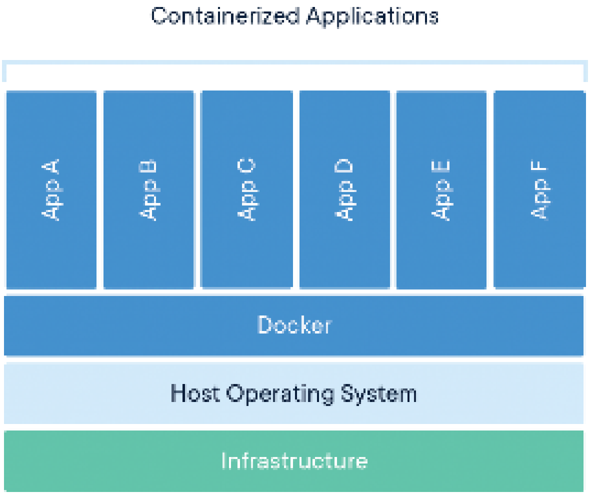
- Will allow creation of GCHP Spack package for single-line setup

Will Downs (Harvard)

Providing GCHP Software Containers

- Containers facilitate fast setup and running of GCHP
  - Include pre-built source code and executable
  - Users only need to install MPI and Singularity

- Now provide GCHP container images on Docker Hub
- Usage instructions available on GCHP Read The Docs
- Ideal for casual users, demonstrations, testing
- Slight performance decrease due to lack of system-specific optimizations



Will Downs (Harvard)

Offline Advection Archive

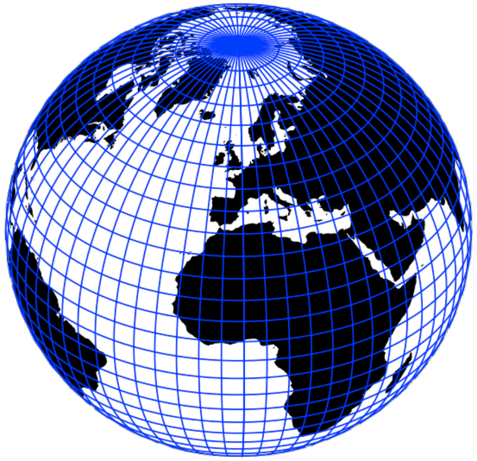
Challenge

- Avoid information loss from unnecessary regridding
- Operational advection fields generated by GMAO on cubed-sphere, regridded to lat-lon for dissemination, and regridded to cubed-sphere for GCHP

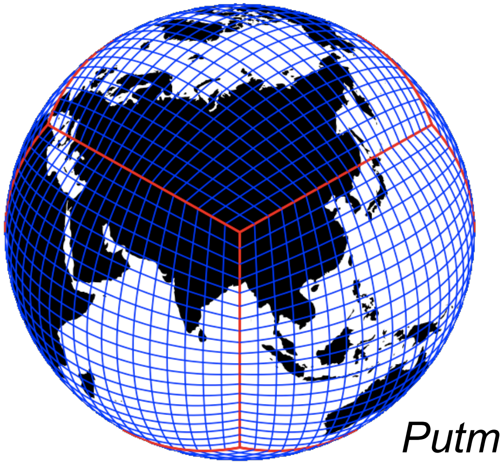
Work completed

- Ability to ingest cubed-sphere data in GCHP
- Generated 2017 MERRA2 archives (hourly C180 resolution)
- Developed mass fluxes transport tracer simulation
- Identified development tasks to eliminate meteorological input preprocessing (for GEOS-Chem)

Lat-lon Winds



Cubed-sphere Mass-Fluxes

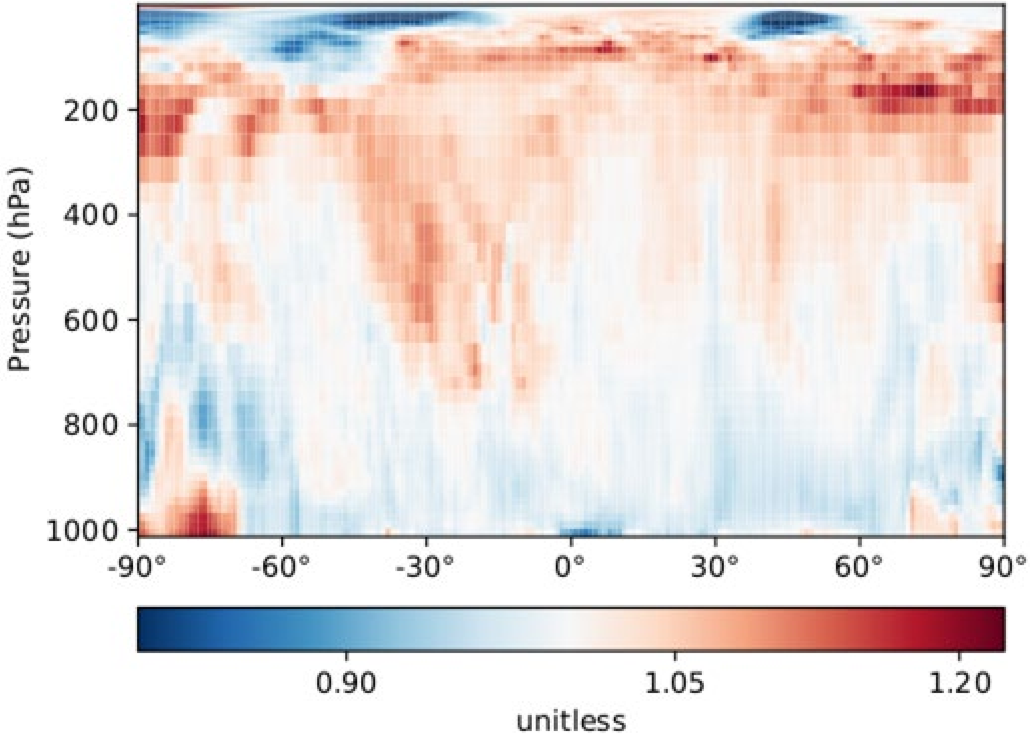


Christoph Keller (GMAO),
Seb Eastham (MIT),
Lizzie Lundgren (Harvard),
Liam Bindle (WashU)

Putman and Lin (2007)

Eliminating Double Regridding Preserves Vertical Motion

Effect of changing from lat-lon winds to cubed-sphere mass fluxes



Plotted: Relative change in ^{222}Rn from switching from LL winds -> CS mass fluxes

Quantities ratioed: Zonal mean ^{222}Rn for July 30, 2017

Simulation: July 2017

Impact:

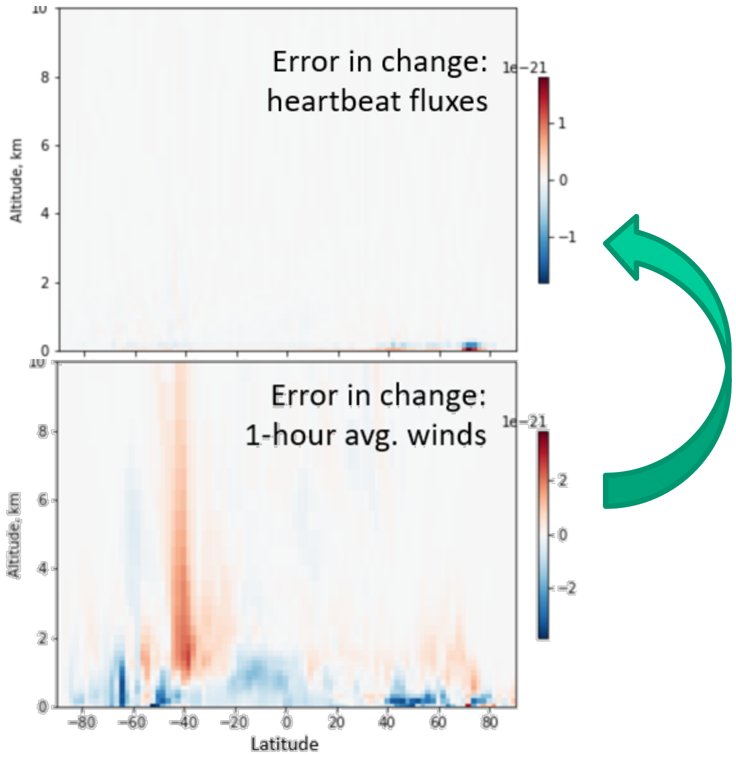
- Better resolves vertical transport

Liam Bindle (WashU)

Mass fluxes for GCHP

Fixing the fixer

- GCHP updated to accept mass fluxes directly from GMAO
- Almost eliminates long-standing CTM error (Jöckel et al, 2001)
- Now extending work to allow flux regridding using ESMF
- Manuscript is in preparation



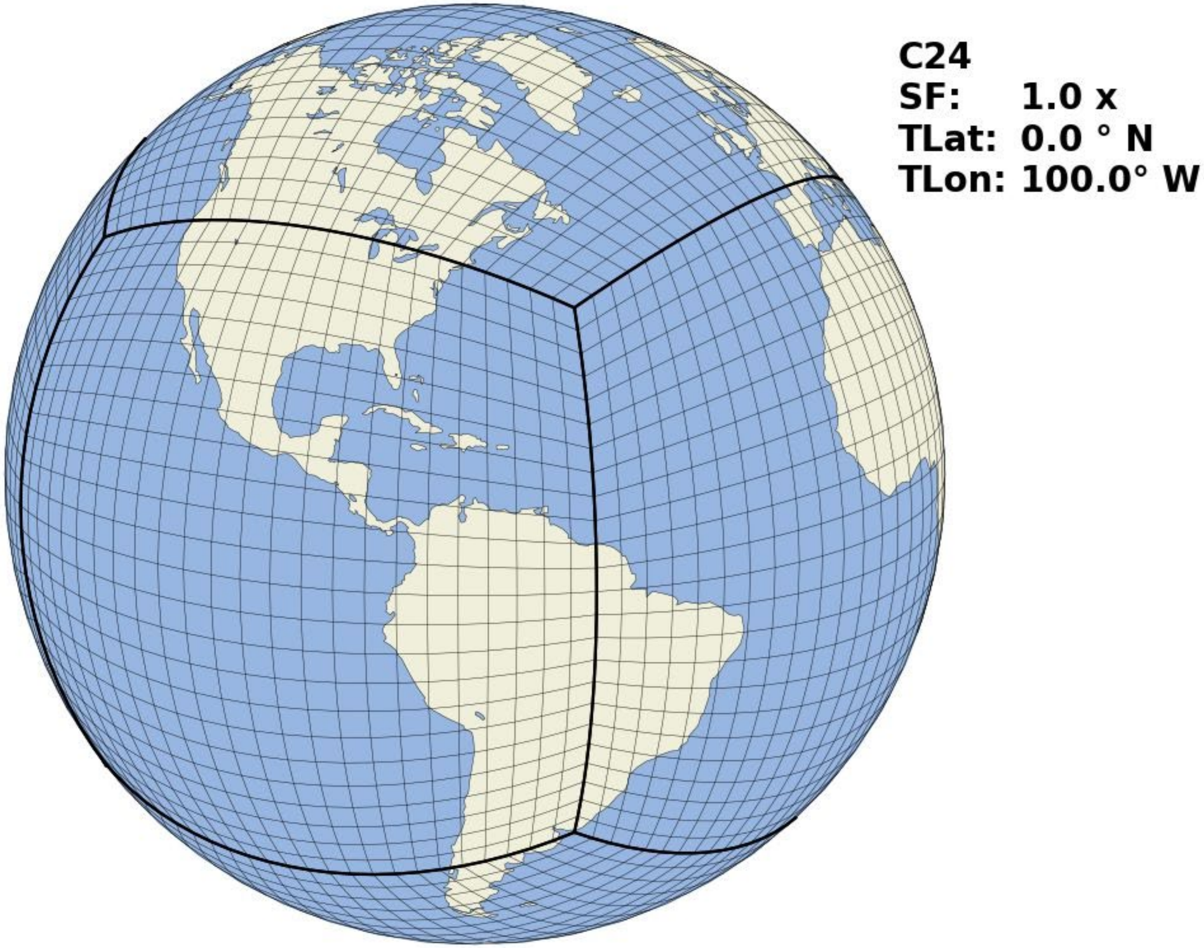
Error in change: heartbeat fluxes

Error in change: 1-hour avg. winds

Using fluxes almost eliminates error in transport of a CH_3I-like tracer

Seb Eastham (MIT)

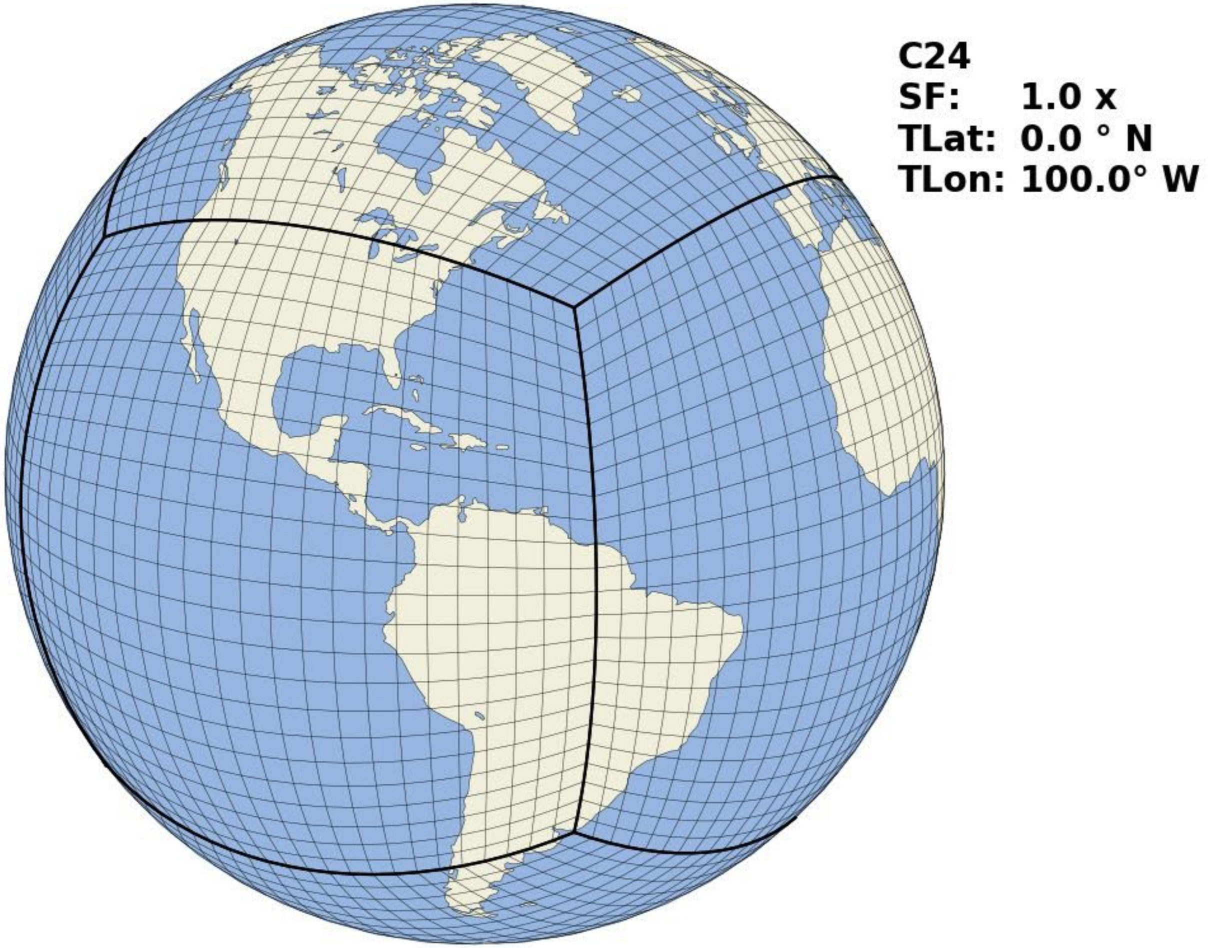
Stretched-grid Capability for Targeted High-resolution Simulations using GCHP



Bindle et al., submitted

Stretched-grid capability for targeted high-resolution simulations using GCHP

- Transformation to the cube-sphere's grid-boxes
- Grid-boxes shrink over target region
- Grid-boxes expand on the opposite face
- No added computational effort



Bindle et al., submitted

Stretched-grid simulation with C720 (12 km) resolution

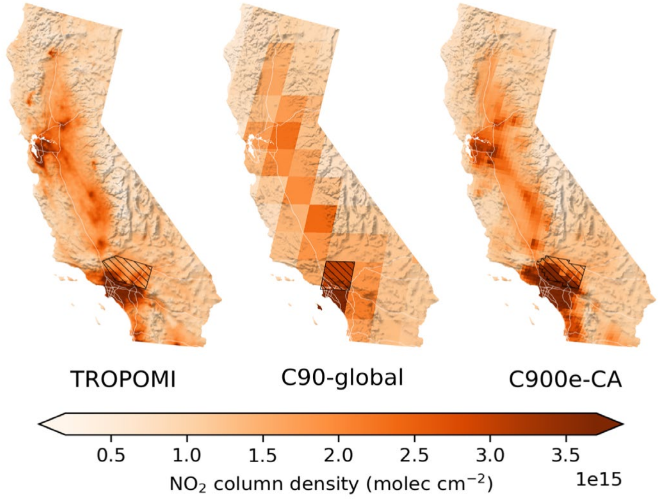
Full chemistry simulation over California (surface ozone)

Complex topography and source structure better represented at fine resolution

Implicit 2-way ‘nesting’

At expense of global C48 (~2° x 2.5°)

Bindle et al., submitted



Grid Independent Emissions Enable Consistent Emissions across Multiple Resolutions

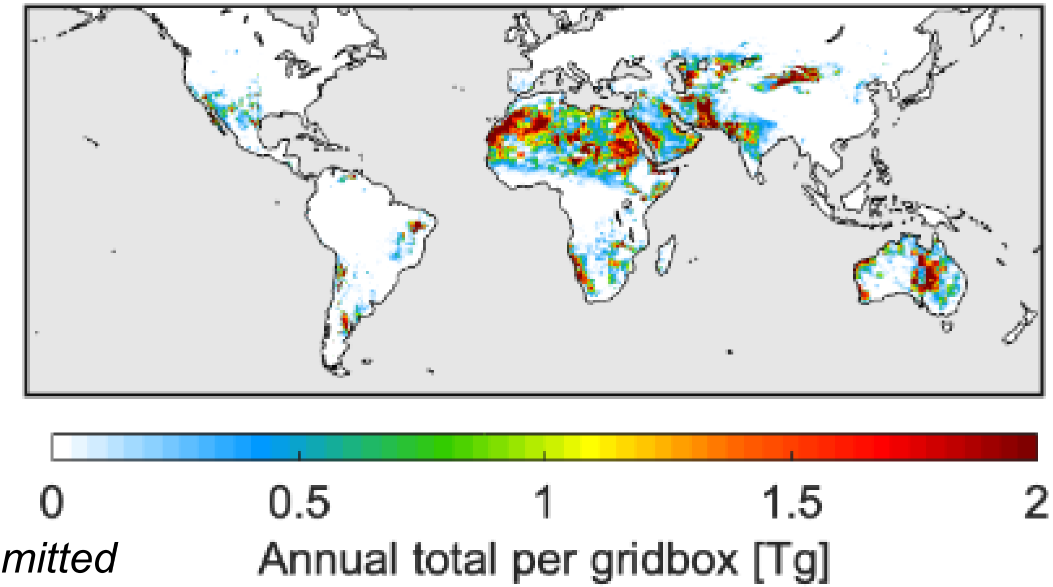
Challenge

- Emissions change with meteorological resolution
- Especially problematic for stretched-grid

Work completed

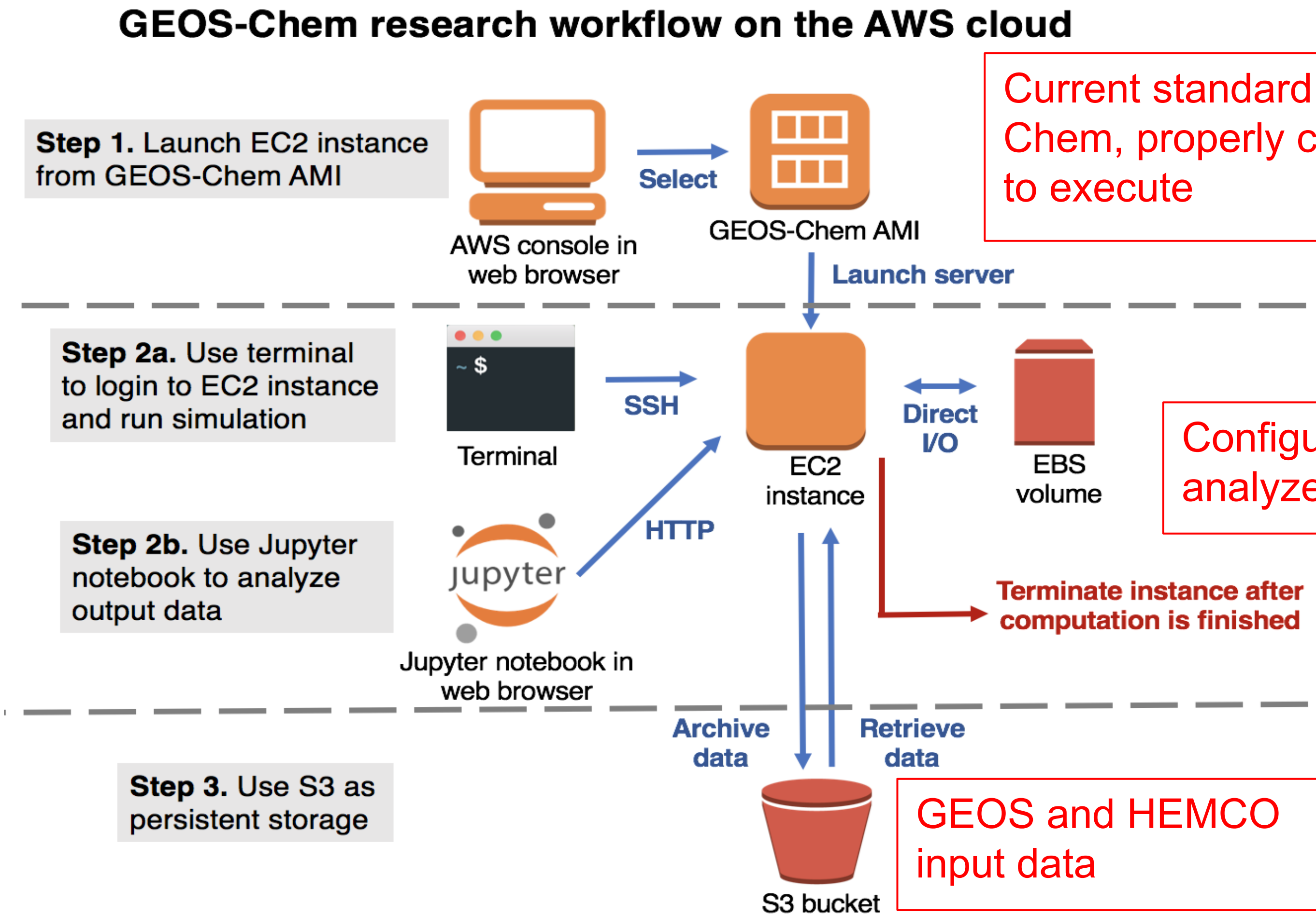
- Contributed to development of grid independent emissions
- Develop archive at native resolution
- Enables representation of emissions at the finest resolution

Mineral Dust Emissions



Meng et al., submitted

GCHP on the AWS cloud



GEOS-Chem research workflow on the AWS cloud

Step 1. Launch EC2 instance from GEOS-Chem AMI

Step 2a. Use terminal to login to EC2 instance and run simulation

Step 2b. Use Jupyter notebook to analyze output data

Step 3. Use S3 as persistent storage

Current standard version of GEOS-Chem, properly configured and ready to execute

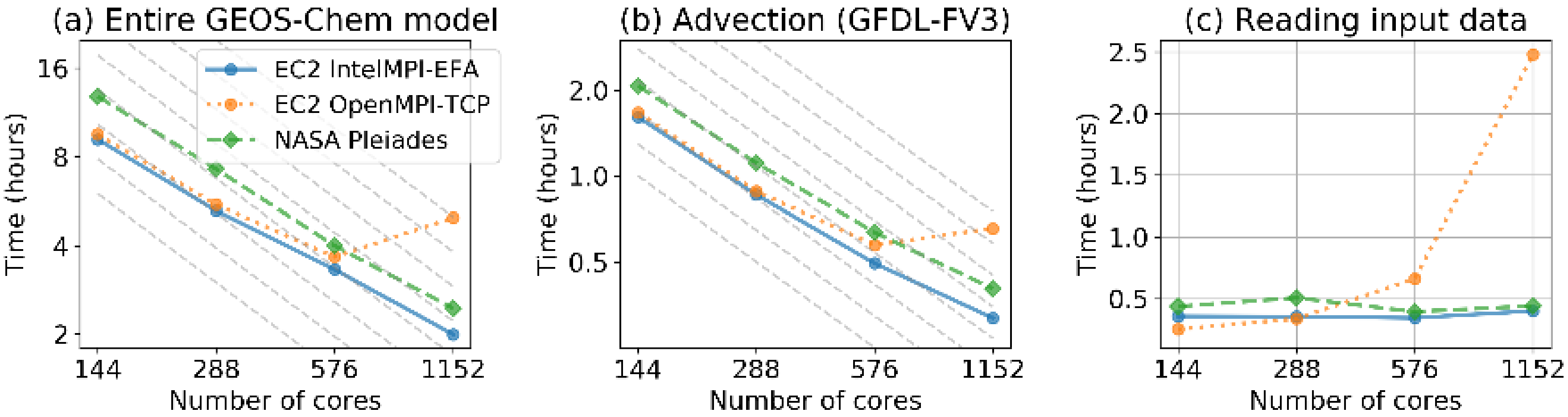
Configure/execute your run, analyze output

GEOS and HEMCO input data

Zhuang et al., JAMES 2020

Progress Toward Multi-node Cloud Capability

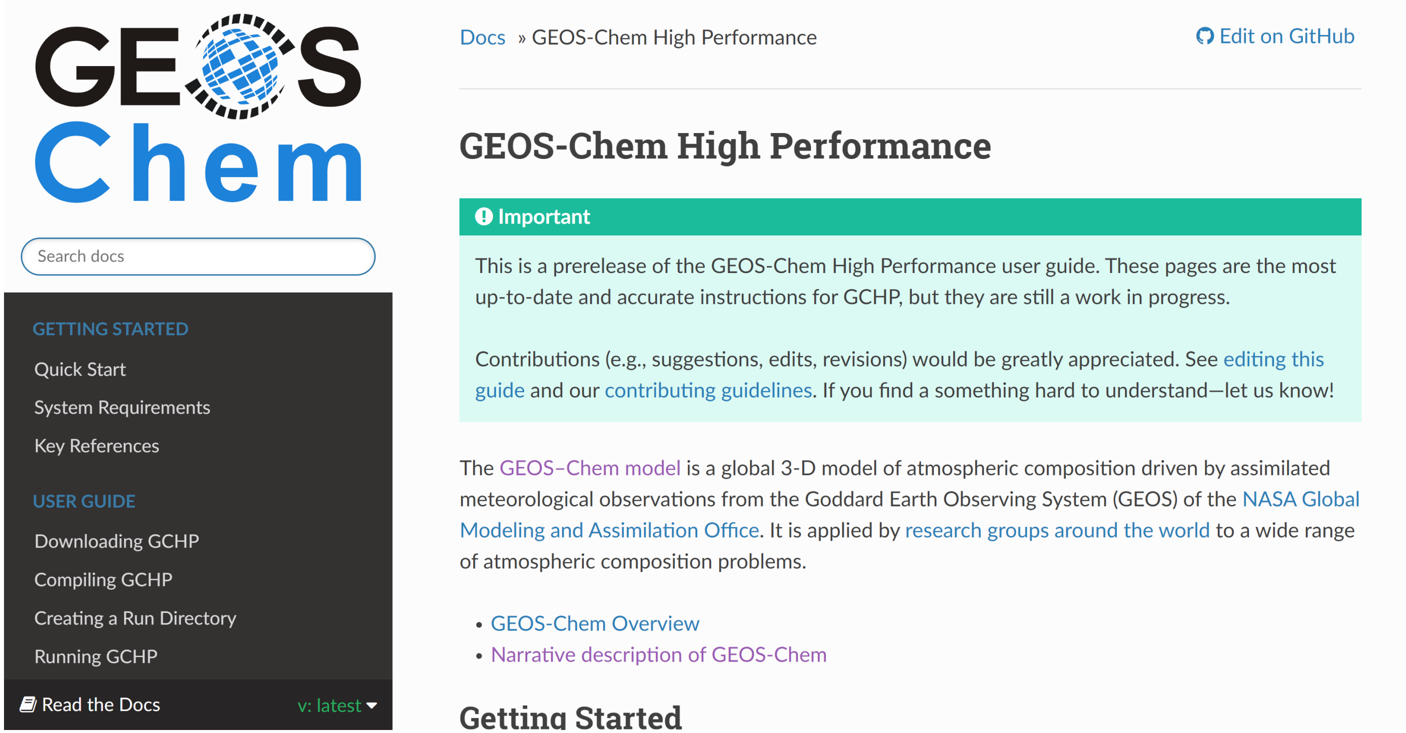
- Demonstration tests with 7-day global 50-km resolution (C180) GEOS-Chem benchmark
- Intel-MPI (with EFA) scales well to 1152 cores; faster than NASA Pleiades by 20%
- OpenMPI (with TCP) cannot scale beyond 576 scores, due to major slow down in I/O and minor slow down in advection.



Zhuang et al., 2020

Supporting Community through Documentation

- **Tutorials on YouTube: https://www.youtube.com/c/geoschem**
- **ReadTheDocs: https://gchp.readthedocs.io/en/latest/**



GEOS-Chem High Performance

Important

This is a prerelease of the GEOS-Chem High Performance user guide. These pages are the most up-to-date and accurate instructions for GCHP, but they are still a work in progress.

Contributions (e.g., suggestions, edits, revisions) would be greatly appreciated. See editing this guide and our contributing guidelines. If you find a something hard to understand—let us know!

The GEOS–Chem model is a global 3-D model of atmospheric composition driven by assimilated meteorological observations from the Goddard Earth Observing System (GEOS) of the NASA Global Modeling and Assimilation Office. It is applied by research groups around the world to a wide range of atmospheric composition problems.

- GEOS-Chem Overview
- Narrative description of GEOS-Chem

Getting Started

GCHP Demonstration

Full chemistry at C360 (~25km) resolution

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Summary of Accomplishments and Future Plans

- **Completed activities as planned**
  - Updated MAPL
  - Enabled seamless updates
  - Improved build system
  - Implemented package manager
  - Implemented containers
  - Generated offline advection archive
  - Enhanced documentation

- **Ongoing work**
  - Complete parallelization assessment and improvement
  - Support multi-node cloud capability
  - Support stretched grid implementation
  - Operationalize cubed-sphere archive

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Publications - List of Acronyms

Publications

Zhuang, J., D.J. Jacob, H. Lin, E.W. Lundgren, R.M. Yantosca, J. Flo Gaya, M.P. Sulprizio, S.D. Eastham, and K. Jorissen, Enabling high-performance cloud computing for Earth science modeling on over a thousand cores: application to the GEOS-Chem atmospheric chemistry model, Journal of Advances in Modeling Earth Systems, doi: 10.1029/2020MS002064, 2020.

Bindle, L., Martin, R. V., Cooper, M. J., Lundgren, E. W., Eastham, S. D., Auer, B. M., Clune, T. L., Weng, H., Lin, J., Murray, L. T., Meng, J., Keller, C. A., Pawson, S., and Jacob, D. J., Grid-Stretching Capability for the GEOS-Chem 13.0.0 Atmospheric Chemistry Model. Geoscientific Model Development, doi: 10.5194/gmd-2020-398, 2020, in review.

Meng, J., Martin, R. V., Ginoux, P., Hammer, M., Sulprizio, M. P., Ridley, D. A., van Donkelaar, A., Grid-independent High Resolution Dust Emissions (v1.0) for Chemical Transport Models: Application to GEOS-Chem (version 12.5.0). Geosci. Model Dev., doi: 10.5194/gmd-2020-380, 2020, in review.

Presentations

Eastham, S. D., Chossière, G., Speth, R. L., & Barrett, S.R.H. The role of aviation and intercontinental transport in local air quality (poster). American Geoscientists Union (AGU) Annual Fall Meeting, 2019.

Eastham, S. D., Monier, E., Rothenberg, D., & Selin, N. Time of emergence for the influence of climate change on surface ozone (presentation). American Meteorological Society (AMS) Annual Meeting, 2020.

Jacob, D.J. and R.V. Martin, GEOS-Chem model overview, Joint keynote presentation, 1st GEOS-Chem Europe Meeting, 1 September 2020.

Martin, R.V., Progressing from Global to Urban Scales for Air Quality Applications, Earth Science Information Partners Virtual Meeting, 15 July 2020.

Martin, R.V., Advancing Understanding of Air Quality from Global to Urban Scales, Frontiers of Atmospheric Science, American Geophysical Union Virtual Conference, December 2020.

List of Acronyms

AMI	Amazon Machine Image
CS	Cubed-Sphere
EC2	Elastic Compute Cloud
EFA	Elastic Fabric Adapter
ESMF	Earth System Modeling Framework
GCHP	GEOS-Chem High Performance
GEOS	Goddard Earth Observation System
GMAO	Global Modeling and Assimilation Office
HEMCO	Harvard-NASA EMission Component
MAPL	Modeling Analysis and Prediction Layer
MPI	Message Passing Interface
S3	Simple Storage Service
TCP	Transmission Control Protocol
TRL	Technology Readiness Level

Predicting What We Breathe

Jeanne Holm (PI, City of Los Angeles)
Dr. Mohammad Pourhomayoun (Co-I, California State University, Los Angeles)
Jeremy Taub (Co-I, OpenAQ)
Dawn Comer (Project Manager, City of Los Angeles)

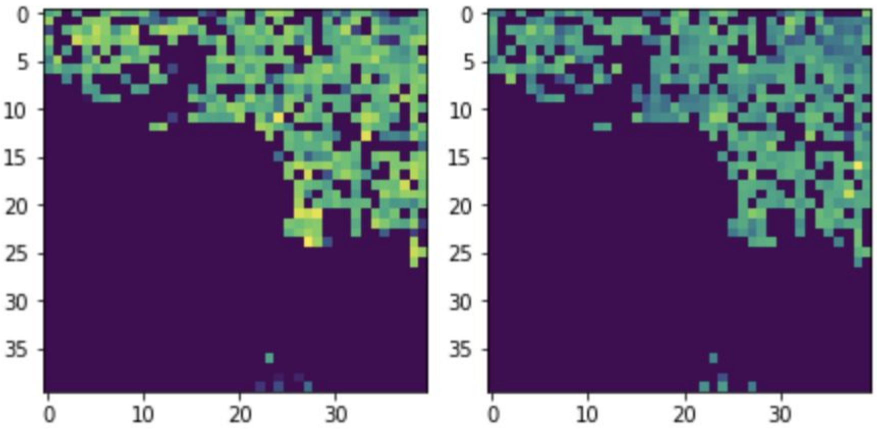
AIST-18-0099 Interim Review
January 22, 2021

Predicting What We Breathe

PI: Jeanne Holm, Deputy Mayor, City of Los Angeles

Objective

- Increase the accessibility and use of space data by using machine learning to help cities predict air quality (AQ) in ways that can be acted upon to improve human health outcomes.
- Provide these tools and algorithms to future Earth science missions (e.g., MAIA) to provide rapid ground truth, combine multiple data sources, and support more rapid use of mission data.



Predicted Air Pollution   Actual Air Pollution

Approach:

- Develop machine learning (ML) algorithms for predictive models for air quality based on measurements of 2.5 micron particulate matter ($PM_{2.5}$) and other air pollutants
- Develop a big data analytics algorithm for integrating ground and space data
- Develop predictive models for health risk using deep learning and machine learning
- Build an open source $PM_{2.5}$ stack for integrating ground and space data
- Create a model for cities with shared attributes to understand predictions and effective interventions

Co-I: Dr. Mohammad Pourhomayoun, Cal State LA

Key Milestones

• Data identification (***Phase 1 complete***)	06/20
• ESTO Science Forum (***Complete***)	06/20
• Identify initial ML models (***Complete***)	07/20
• Develop initial ML algorithm (***Complete***)	12/20
• Identify city interventions and attributes (***Complete***)	11/20
• AGU and CSCI Conferences - 4 papers (***Complete***)	12/20
• Conduct ML training runs (***Phase 1 complete***)	12/20
• Pre- and post-intervention analysis	02/21
• ESTO Science Forum	06/21
• Validate algorithm	10/21
• Publish open source	08/21
• OpenAQ workshops	11/21

$TRL_{in} = 3$

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Background and Objectives

- **AIST Research Focus**
  - Develop machine learning algorithms and models that link ground- and space-based air quality data to
    - ■ Classify patterns
    - ■ Deduce and forecast pollution events
    - ■ Identify AQ similarities amongst megacities

- **Project Objectives**
  - Increase the accessibility and use of space data by using machine learning to help cities predict air quality in ways that will improve human health
  - Provide tools and algorithms to future Earth science missions (such as MAIA) to provide rapid ground truth, conduct data fusion across diverse datasets, and support rapid use of mission data

    1. Create a model for cities to enamine in-situ $PM_{2.5}$, NO_2, PM_{10}, and ozone
    2. Apply machine learning to big datasets from ground and space
    3. Improve decision making on health outcomes in cities

Project Schedule - Overview

Year One (Note that our project start was May 2020)

- 🟢 Identify ground and space-based datasets
- 🟠 Develop a framework to collect and analyze data, look at historical trends and events
- 🟢 Data pre-processing and integration
- 🟠 Select a data architecture and models
- 🟢 Initialize the computational space and migrate data to it
- 🟢 Create, run, and validate initial machine learning algorithms against training data

Year Two

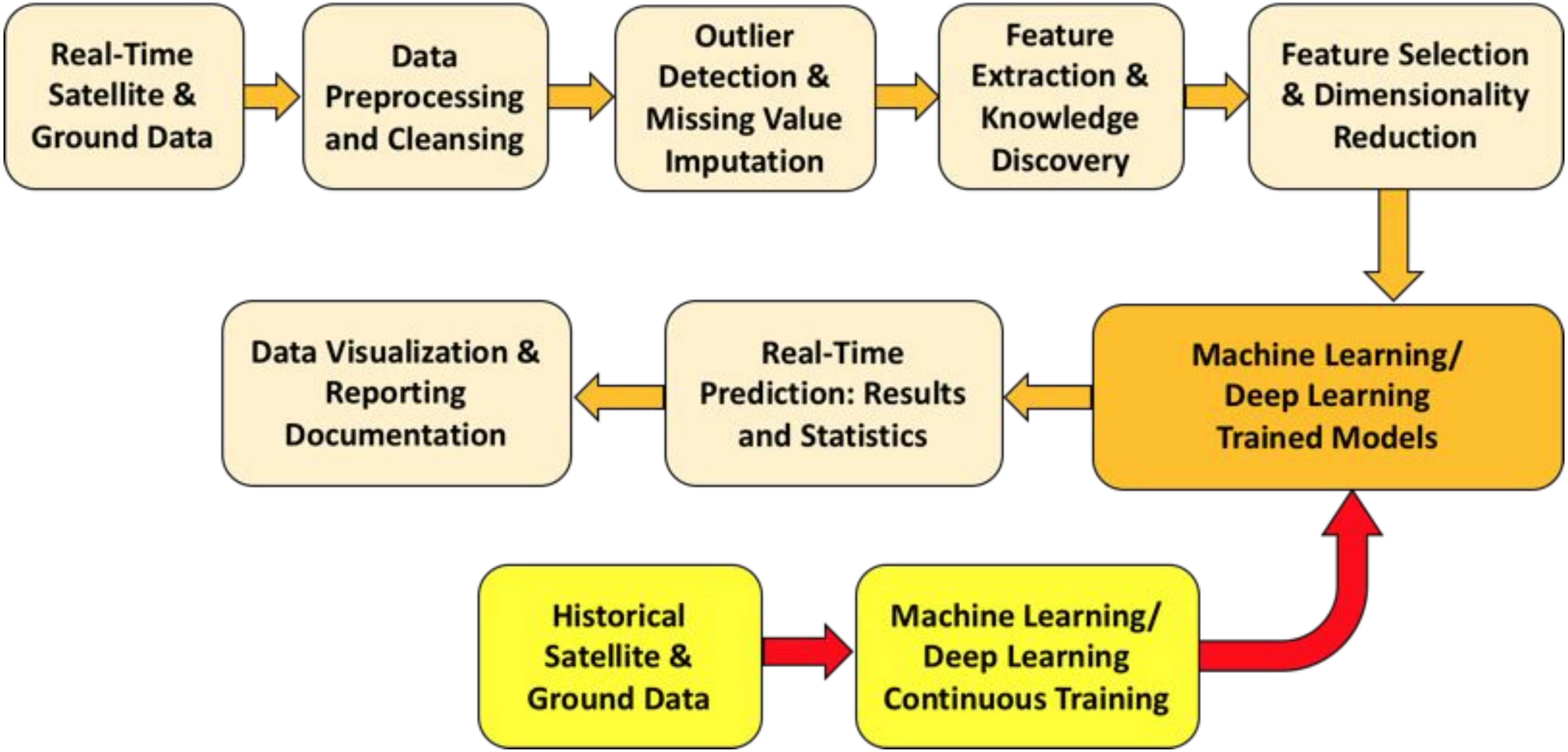
- 🟠 Sister cities will be identified and recruited
- Include possible additional datasets
- Validate the models based on emergent research
- Run and retrain the algorithms against control and expanded data
- Initial open source publication
- 🟠 Regional and international workshops to socialize the models, promote the open source, and gather requirements

Orange = Underway; Green = complete for this phase

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

High-Level Approach to ML Models

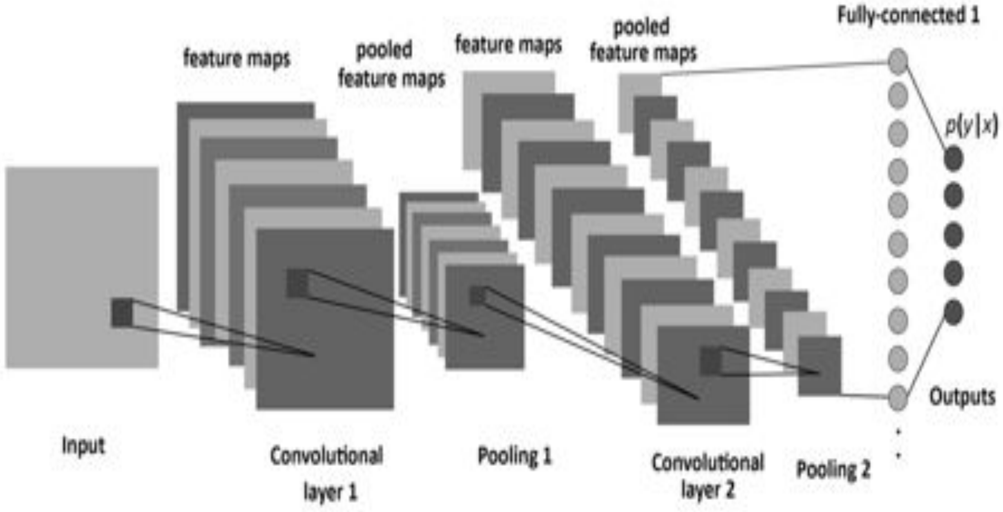

Real-Time Satellite & Ground Data
Data Preprocessing and Cleansing
Outlier Detection & Missing Value Imputation
Feature Extraction & Knowledge Discovery
Feature Selection & Dimensionality Reduction
Machine Learning/ Deep Learning Trained Models
Real-Time Prediction: Results and Statistics
Data Visualization & Reporting Documentation
Historical Satellite & Ground Data
Machine Learning/ Deep Learning Continuous Training

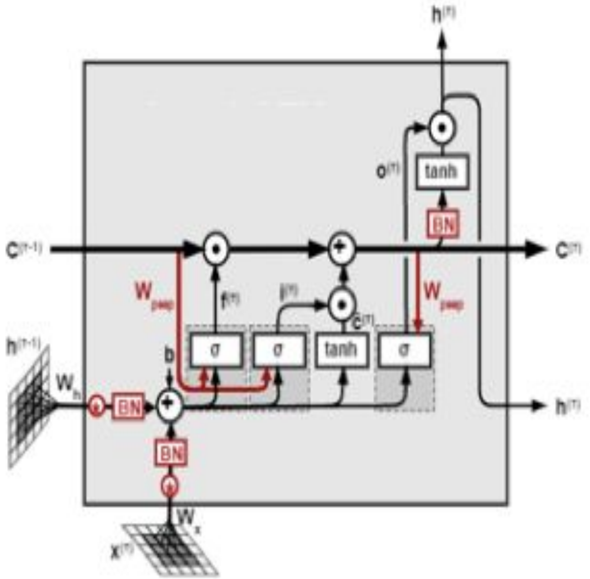
[Ref]: P. Muthukumar, E. Cocom, J. Holm, D. Comer, A. Lyons, I. Burga, Ch. Hasenkopf, and M. Pourhomayoun, “Real-Time Spatiotemporal Air Pollution Prediction with Deep Convolutional LSTM through Satellite Image Analysis,” The 16th Int. Conference on Data Science (ICDATA’20), 2020.

Technical Advances (continued)

Machine Learning Deep Neural Network Models

- Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM): For the **temporal** correlation in the data
- Convolutional Neural Network (CNN): For the **spatial** correlation
- Convolutional RNN/LSTM: For the **spatiotemporal** correlation
- CNN RNN/LSTM: For the **spatiotemporal** correlation

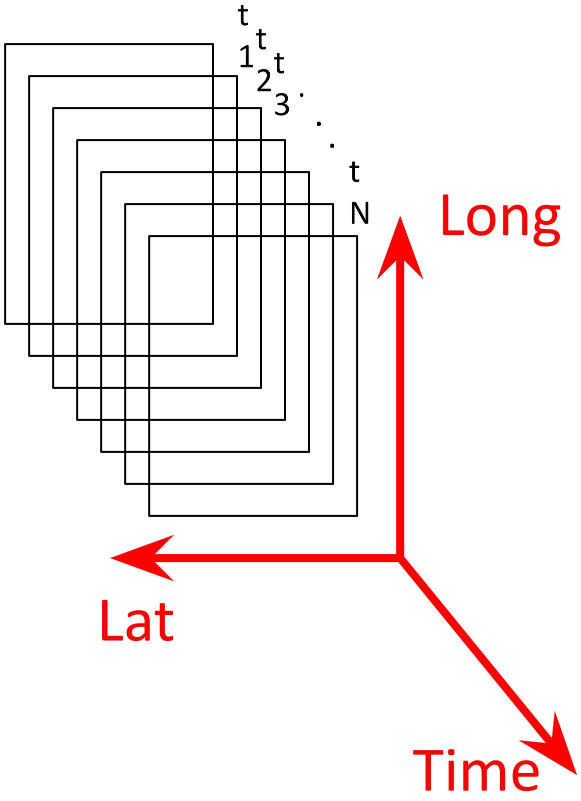




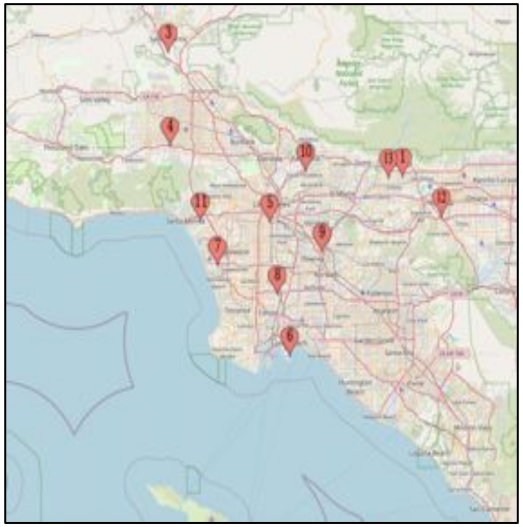
Technical Advances (continued)

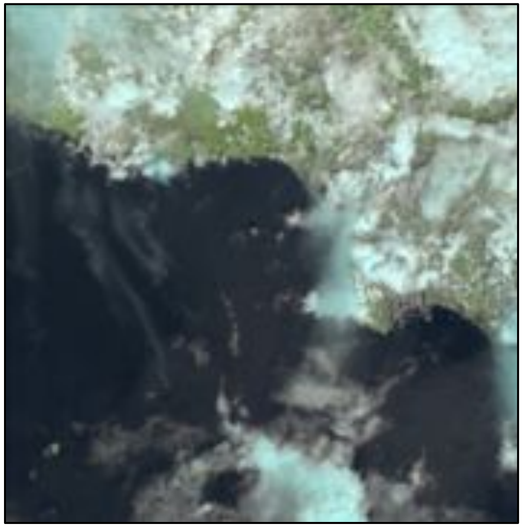
Considering Temporal and Spatial Patterns in the Data

Temporal Correlation



Spatial Correlation

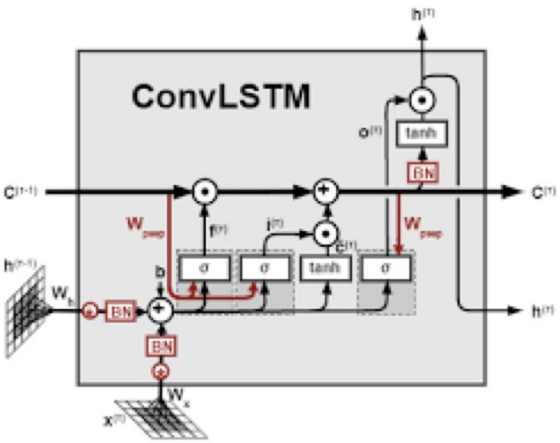


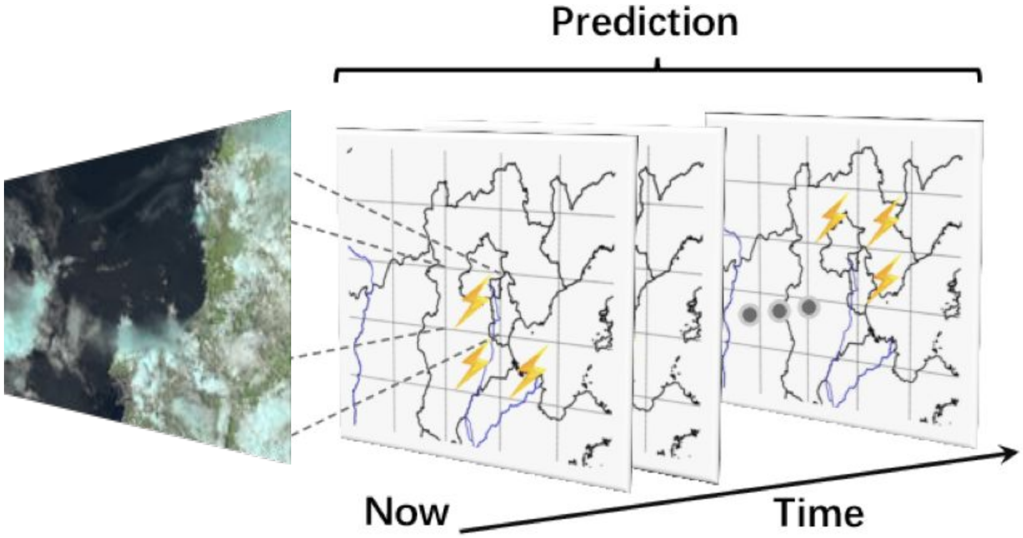


Technical Advances (continued)

Deep Convolutional RNN/LSTM

- Deep Learning model used for learning correlations among spatial + temporal data
- Implements convolution within cells of LSTM
- Input shape: 5D Tensor (Samples, Frames, Rows, Columns, Filters)





Technical Advances (continued)

Predictive Model 1: Predicting **PM2.5** in L.A. County every 46 hours based on satellite observations and ground sensors

Input data

- **Satellite observations NASA MAIAC MODIS**:
  - Spatial Resolution: 1-km/pixel (40x40km)
  - Temporal Resolution: 46-hr frequency
- **Ground-based sensors** (13 in L.A. County), hourly
- **Meteorological data** (L.A. County)

Accuracy	Frame #
91%	Frame 1: 46 hours in future
86%	Frame 2: 4 days in future
84%	Frame 3: 6 days in future
79%	Frame 4: 8 days in future
75%	Frame 5: 10 days in future

[Ref]: P. Muthukumar, E. Cocom, J. Holm, D. Comer, A. Lyons, I. Burga, Ch. Hasenkopf, and M. Pourhomayoun, "Satellite Image Atmospheric Air Pollution Prediction through Meteorological Graph Convolutional Network with Deep Convolutional LSTM," The 2020 International Conference on Computational Science and Computational Intelligence (CSCI'20), 2020.

Technical Advances (continued)

Model Architecture: GraphNN-ConvLSTM

1. GraphNN for Spatiotemporal Meteorological Data

   -Use GNN to create denser, more complex weather data graph (bounded by latitude/longitude as axes) for each timestep (46-hr interval)

2. Unsupervised Learning Graph Representation Learning

   -Intermediate step between GNN and ConvLSTM to convert dense graph to grid-based high-level embeddings in “image” format

3. ConvLSTM Model

   - Inputs: Meteorological Graph Embeddings and processed Satellite Imagery
   - Output: Grid of ground-level air pollutant over LA county every 46 hours

4. 13-Layer Dense Neural Network

   -Flattens ConvLSTM output grid to use as features

   -Output: Predicted air pollutant values in Los Angeles County

[Ref]: P. Muthukumar, E. Cocom, J. Holm, D. Comer, A. Lyons, I. Burga, Ch. Hasenkopf, and M. Pourhomayoun, “Satellite Image Atmospheric Air Pollution Prediction through Meteorological Graph Convolutional Network with Deep Convolutional LSTM,” The 2020 International Conference on Computational Science and Computational Intelligence (CSCI'20), 2020.

Technical Advances (continued)

Predictive Model 2: Predicting **NO_2** in L.A. County every 46 hours based on satellite observations and ground sensors

Input data

- **Satellite observations NASA MAIAC MODIS**:
  - Spatial Resolution: 1-km/pixel (40x40km)
  - Temporal Resolution: 46-hour frequency
- **Ground-based sensors** (13 in L.A. County), hourly
- **Meteorological data** (L.A. County)

Accuracy	Frame #
84%	Frame 1: 46 hours in future
81%	Frame 2: 4 days in future
80%	Frame 3: 6 days in future
73%	Frame 4: 8 days in future
70%	Frame 5: 10 days in future

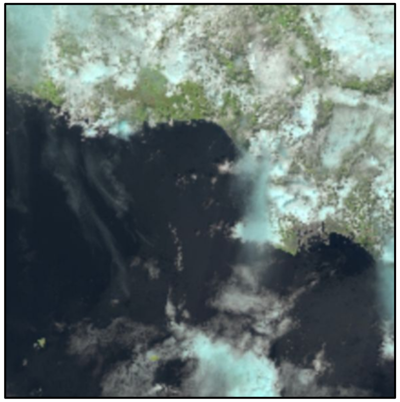
[Ref]: P. Muthukumar, E. Cocom, J. Holm, D. Comer, A. Lyons, I. Burga, Ch. Hasenkopf, and M. Pourhomayoun, "Satellite Image Atmospheric Air Pollution Prediction through Meteorological Graph Convolutional Network with Deep Convolutional LSTM," The 2020 International Conference on Computational Science and Computational Intelligence (CSCI'20), 2020.

Technical Advances (continued)

Predictive Model 3: Predicted NO_2 in L.A. every 46 hours based on satellite observations and ground sensors

Input data

- **Satellite images** (ESA Sentinel-2 Satellite imagery, 945.1 nm spectral band of NO_2)
- **Ground-based sensors**
- **Meteorological data**



Accuracy	Frame #
79%	Frame 1: 46 hours in future
78%	Frame 2: 4 days in future
75%	Frame 3: 6 days in future
70%	Frame 4: 8 days in future
68%	Frame 5: 10 days in future

[Ref]: P. Muthukumar, E. Cocom, J. Holm, D. Comer, A. Lyons, I. Burga, Ch. Hasenkopf, and M. Pourhomayoun, "Real-Time Spatiotemporal Air Pollution Prediction with Deep Convolutional LSTM through Satellite Image Analysis," The 16th Int. Conference on Data Science (ICDATA'20), 2020.

Technical Advances (continued)

Predictive Model 4: Predicting **ozone** in L.A. County every 46 hours based on satellite observations and ground sensors

Input data

- Satellite observations (**NASA MAIAC MODIS**): 1-km/pixel, 46-hr frequency
- **Ground-based sensors** (13 in L.A. County), hourly
- **Meteorological data** (L.A. County)

Accuracy	Frame #
92%	Frame 1: 46 hours in future
89%	Frame 2: 4 days in future
86%	Frame 3: 6 days in future
83%	Frame 4: 8 days in future
76%	Frame 5: 10 days in future

[Ref]: P. Muthukumar, E. Cocom, J. Holm, D. Comer, A. Lyons, I. Burga, Ch. Hasenkopf, and M. Pourhomayoun, "Satellite Image Atmospheric Air Pollution Prediction through Meteorological Graph Convolutional Network with Deep Convolutional LSTM," The 2020 International Conference on Computational Science and Computational Intelligence (CSCI'20), 2020.

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

TRL Assessment

Accomplishments Since Last Review

- Develop ML models from existing satellite and ground level data
- Focus for technology maturation has been on
    - Predictive models for health risk prediction
        - Initial Machine Learning Algorithms are being tested and trained with AQ data sets
    - Predictive models for air quality
        - Using MAIAC and AQMD data to develop 1-10 day predictive models using training data and validation
- Next six months focus will be on
    - Big data analytics algorithms (3)
        - Combining datasets from two satellites and another one of multiple ground sensors for pre-processing (4)
    - Open source $PM_{2.5}$ stack (4)
        - Define components for the phase 1 stack build
        - Operationalize this as all the components in a shared environment with researchers and cities (5)
    - Virtual calibration (3)
        - Show calibration proof of concept on one space- and one ground-based dataset +
    - Predictive models (2)
        - Define useful formats and outputs for health organizations
        - Work with Anthem, AQMD, and Propeller Health for a proof of concept on ingestion and projecting impact (3)

TRL Assessment

Component	Entry TRL	Entry Justification	Exit TRL	Exit Justification
Predictive models for air quality based on several deep RNN models that takes into account both temporal and spatial correlation in ground/space data	3	Models using RNN have not been demonstrated related to air quality data	4	Model will be able to predict 1-2 year later data after undergoing training
Big data analytics algorithms for integrating ground and space data	3	Able to preprocess data based on the type and nature of the data	5	Extract knowledge from the data and prepare it for machine learning
Predictive models for health risk prediction based on deep learning and machine learning algorithms trained on historical data and for air quality predictive model	2	Current health predictions are for long-range forecasts and don't use ML	4	Train the ML algorithms against a historical dataset and predict health risks accurately in the near term
Open source $PM_{2.5}$ stack: Combining open source stack to integrate satellite and ground data for $PM_{2.5}$	4	Tools individually are at TRL 9-10, but unable to easily combine them to provide an integrated view at ground up to 700 km	6	Provide reliable data over time across multiple sources to measure $PM_{2.5}$ for a specific location in Los Angeles
Virtual calibration: Model to provide federation of space data with ground data	3	Under the relationship between $PM_{2.5}$ ground and space data for a given region	4	Use machine learning algorithm to validate calibration of space- or ground-based data

Summary

- Project launched May 18, 2020
  - Spending and obligations are in line with the phasing plan
- Team meets regularly and connects to new partners
  - AQMD
  - Propeller Health
  - OpenAQ
  - SmartAirLA
  - SafeCast
  - Southern California Asthma Association
- Identified initial datasets
- Data processing and integration
- Fine tuning ML model options
- Close coordination with other AIST partners
  - NASA data standards
- Already engaging cities
- Scoping citizen science data collection opportunities with LAPL

Current State

- **Administrative**
  - Project commenced on May 18, 2020 (post COVID-19 delay)
  - Contracts established between the City and OpenAQ and Cal State L.A.
  - Project award formally accepted by City Council
  - Participated in ESIP Winter 2020 meeting
  - Participating in MAIA early adopter meetings
  - Bi-weekly and monthly meetings for core, partners, and community
  - Launched project website - airquality.lacity.org, and project email address - airquality@lacity.org
- **Data Preparation**
  - Identification of ground-based and satellite datasets available from NASA, OpenAQ and existing City department projects
  - Established regular engagement within the AQ data community to collaborate on best practices for accessing and using data (NASA, OpenAQ, L.A. County Health, etc.)
  - Initial use of NASA satellite data for machine learning algorithms
- **Technical Preparation**
  - Data processing and integration
  - Designing machine learning approaches
  - Developing and training machine learning Aagorithms for discovering spatiotemporal patterns in the data and make predictions
-

Current State (continued)

- **Community Engagement**
  - Published and presented 6 peer-reviewed papers and 3 meeting papers (details on slides 33-34)
  - Continued engagement with community advocates (Anthem Blue Cross, Southern California Asthma Association, SmartAirLA, and AQMD)
  - Concept meeting with Agents of Climate augmented reality app for citizen science
  - Initial identification of citizen science project with LA Public Library and SafeCast sensors
  - Identification of AQ sister cities completed
  - Initial identification of AQ interventions to measure

Plan Forward

Next Steps

- Continue evolution of model, algorithms, and validation
- Adding new datasets to the predictive models including more high-resolution satellite observations from NASA and fire/smoke data.
- Continue to identify and integrate local data (health, polluters, traffic, roads, ports) from IOT and in-situ sensors
- Identify gaps in AQ sensor coverage
- Continue to engage citizen scientists (libraries, SafeCast, SmartAirLA, and more), community for environmental justice for awareness and support, and healthcare partners (Propeller Health, Anthem Blue Cross, Southern California Asthma Association) to improve health outcomes
- Share findings via smart city air quality intervention and toolkit (C40 cities, U.N. Sustainable Development Goals Network, Climate Mayors, etc.)
- Develop and conduct training workshops on finding and using air quality data for both LA government and community stakeholder representatives, and for a group of global cities interested in learning more about project models that can be replicated.

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Publications - List of Acronyms

Publications

- Journal / Conference Papers (6 peer-reviewed papers, 3 meeting papers)
    - 2 Papers Published/Presented in 2020 International Conference on Computational Science and Computational Intelligence (CSCI'20: December 16-18, 2020, Las Vegas, USA) https://www.american-cse.org/csci2020/
        - *Satellite Image Atmospheric Air Pollution Prediction through Meteorological Graph Convolutional Network with Deep Convolutional LSTM*
        - *Sensor-Based Air Pollution Prediction Using Deep CNN-LSTM*
    - 2 Abstracts Presented in AGU (American Geophysical Union) Fall Meeting Presentation (December 7-11, 2020) - submissions complete
        - *Particulate Matter Forecasting in Los Angeles County with Ground-Based Sensor Data Analytics*
        - *Real-Time Spatiotemporal NO2 Air Pollution Prediction with Deep Convolutional LSTM through Satellite Image Analytics*
    - Paper presented at ICDATA conference (July 27) : presentation video
        - *Real-Time Spatiotemporal Air Pollution Prediction with Deep ConvLSTM via Satellite Image Analysis*

Publications (continued)

- Journal/Conference Papers
  - Presented project at the Environmental Law Institute (July 29)
    - ELI is supporting the U.S. EPA in an effort to characterize and learn from how states, tribes and local governments are using citizen science in their programs
  - Peer-reviewed paper at International Astronautical Congress (October 11)
  - European Space Agency’s Space for Twin Cities broadcast (November 19)
- Other
  - Project mentioned by Mayor Garcetti @ SCAQMD EJ Conference
  - UN International Day of Clean Air - City of L.A. Social Media (September 7)
  - Clean Air Day - City of L.A. Press Release and Social Media (October 7)
  - Project presented at City of L.A. Chief Sustainability Officer Meeting (November 18)

Partners

- Public
  - City of Los Angeles
  - NASA/JPL
  - Southern California Air Quality Management District
  - SafeCast
- Private
  - OpenAQ
  - SmartAirLA

AQMD

SmartAirLA

- Academic
  - California State University, Los Angeles
  - LA Data Science Federation
- Organizations
  - Mayor Garcetti leads the C40 Cities
  - Climate Mayors

Acronyms

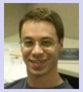
AQMD	South Coast Air Quality Management District
ML	Machine learning
Cal State LA	California State University, Los Angeles
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
CNN	Convolutional Neural Network

QUANTIFYING UNCERTAINTY AND KINEMATICS OF EARTHQUAKE SYSTEMS (QUAKES-A) ANALYTIC CENTER FRAMEWORK

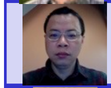
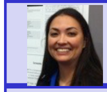
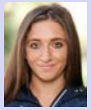
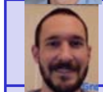
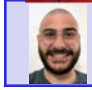
Andrea Donnellan (PI, Jet Propulsion Laboratory, California Institute of Technology)
Jay Parker (Co-I, Jet Propulsion Laboratory, California Institute of Technology),
Robert Granat (Co-I, City College of New York),
Marlon Pierce (Co-I, Indiana University),
John Rundle (Co-I, University of California Davis),
Lisa Grant Ludwig (Co-I, University of California Irvine)

AIST-18-001 Annual Technical Review
January 22, 2021

Investigators

	Name	Org	Position	Role
	Andrea Donnellan	JPL	PI	Oversight, testing and evaluation
	Jay Parker	JPL	Co-I	InSAR edge detection and displacement estimation
	Robert Granat	CCNY	Co-I	Data fusion and uncertainty quantification
	Marlon Pierce	Indiana U	Co-I	Science gateway analytic center framework
	John Rundle	UC Davis	Co-I	Geodetic/seismicity forecasting
	Lisa Grant Ludwig	UC Irvine	Co-I	Target communities interface

Current Team Members and Students

	Name	Org	Level	Role
	Brian Hawkins	JPL	Staff	UAVSAR GNSS adjustment
	Jun Wang	Indiana U	Staff	GIS, web services
	Michael Heflin	JPL	Staff	GNSS time series/velocity field
	Maggi Glasscoe	JPL	Staff	Response and Hazard
	Nathan Pulver	JPL/CPP	B.S.	UAVSAR GNSS adjustment
	Megan Mirkhanian	UC Irvine	Ph.D.	User guide and community engagement
	Nick Mowery	Indiana U	B.S.	Science gateway
	Cameron Saylor	UC Davis	Ph.D.	Radar analysis
	Gregory Lyzenga	JPL	Staff	Data and modeling
	Juan Carlos Beltran	UC Riverside	B.S.	UAVSAR time series analysis
	Joe Yazbeck	UC Davis	Ph.D.	Radar damage assessment

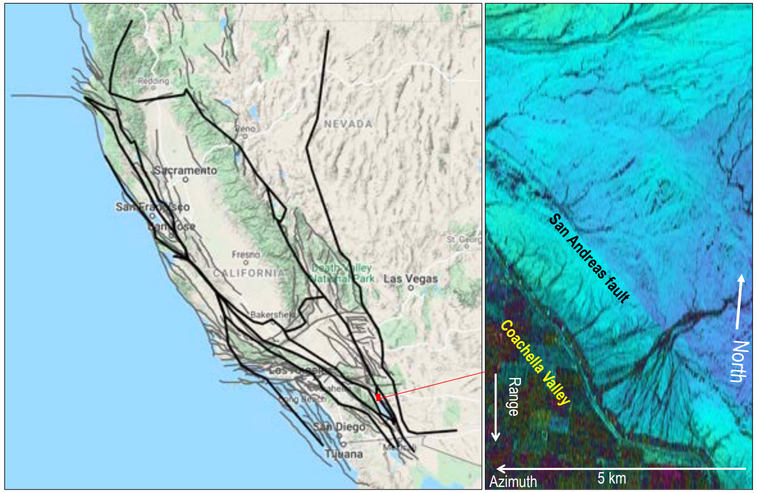
Quantifying Uncertainty and Kinematics of Earthquake Systems (QUAKES-A)

PI: Andrea Donnellan, JPL/Caltech

Objective

Create a uniform crustal deformation reference model for the active plate margin of California

- Fused InSAR, topographic, and GNSS geodetic imaging data
- Quantify uncertainties for the reference model
- Improve earthquake forecast models
- Improve understanding of the physical processes leading to and following earthquakes



Left: Cluster boundaries (black) for k=9 and faults (gray)
Right: Baseline adjusted interferogram showing fault slip

Approach

- Infuse GNSS network solutions into UAVSAR baseline estimation and extract features from data
- Develop cluster analysis to identify and rank active fault systems spatially and temporally
- Fuse/interpolate all available geodetic imaging data to provide a uniformly sampled deformation field based in part on results from the clustering analysis
- Assimilate and correlate the crustal deformation products into seismicity-based earthquake forecasts and back test to understand possible improvements.

Co-Is/Partners: R. Granat, J. Parker (JPL), M. Pierce (IU), J. Rundle (UCD), L. Grant Ludwig (UCI) / Partners: SCEC, FEMA, US and CA Geological Surveys

Key Milestones

- InSAR Adjustment/Machine Learning	Nov/20
- Reference Model (Data Fusion)	Apr/21
- Uncertainty Quantification	Aug/21
- Geodetic/Seismicity Earthquake Forecasts	Nov/21

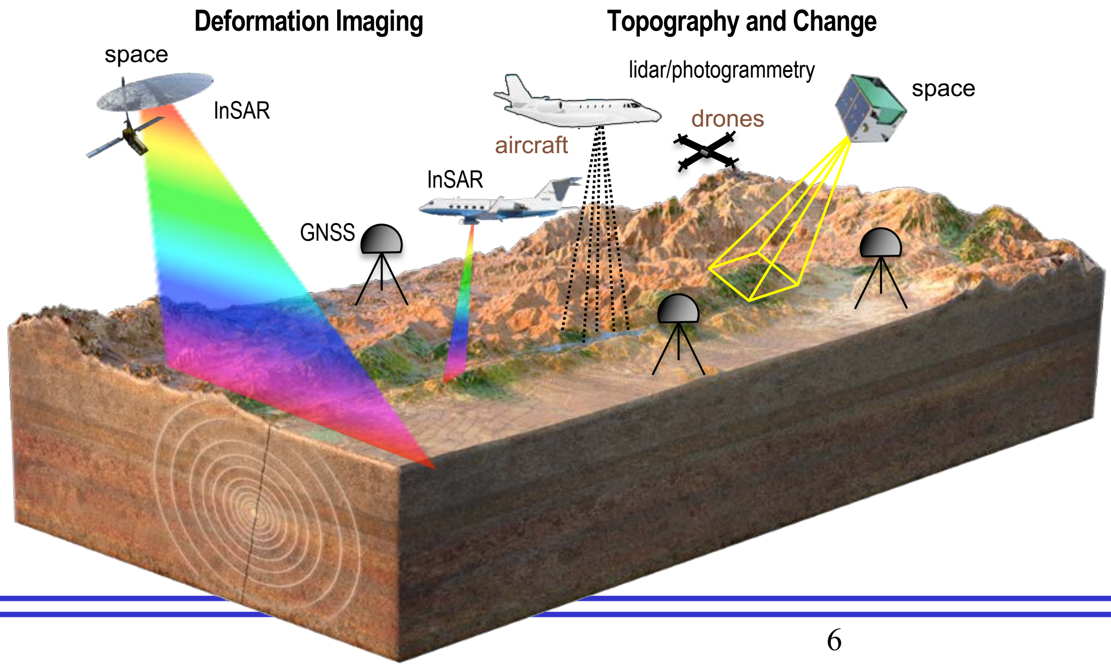
TRL_{in} = 3 $TRL_{current}$ = 4

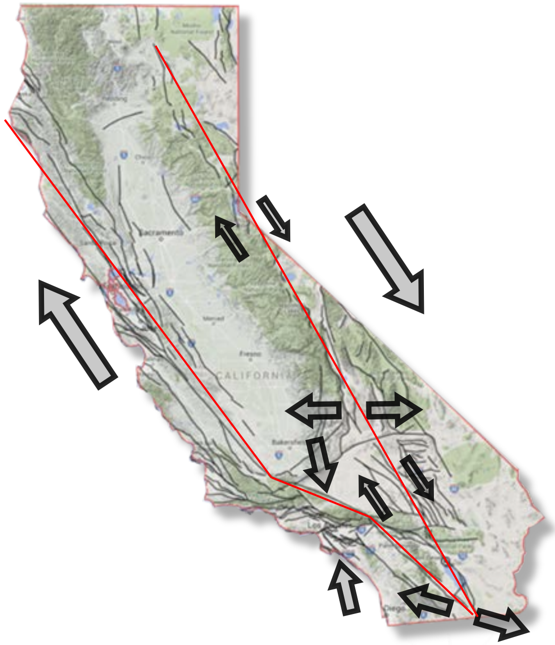
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Background and Objectives

- **Background:** Crustal deformation measurements provide inside into earthquake processes
  - Data come from various instruments of differing characteristics
  - Facilitates understanding tectonic, crustal deformation, and earthquake processes a goal of NASA’s Earth Surface and Interior program.
- **Objective:** Create a uniform crustal deformation reference model for the active plate margin of California
  - Harmonize data products in a time-dependent adaptive gridded product
  - Quantify uncertainties
  - Deploy in a science gateway (GeoGateway)

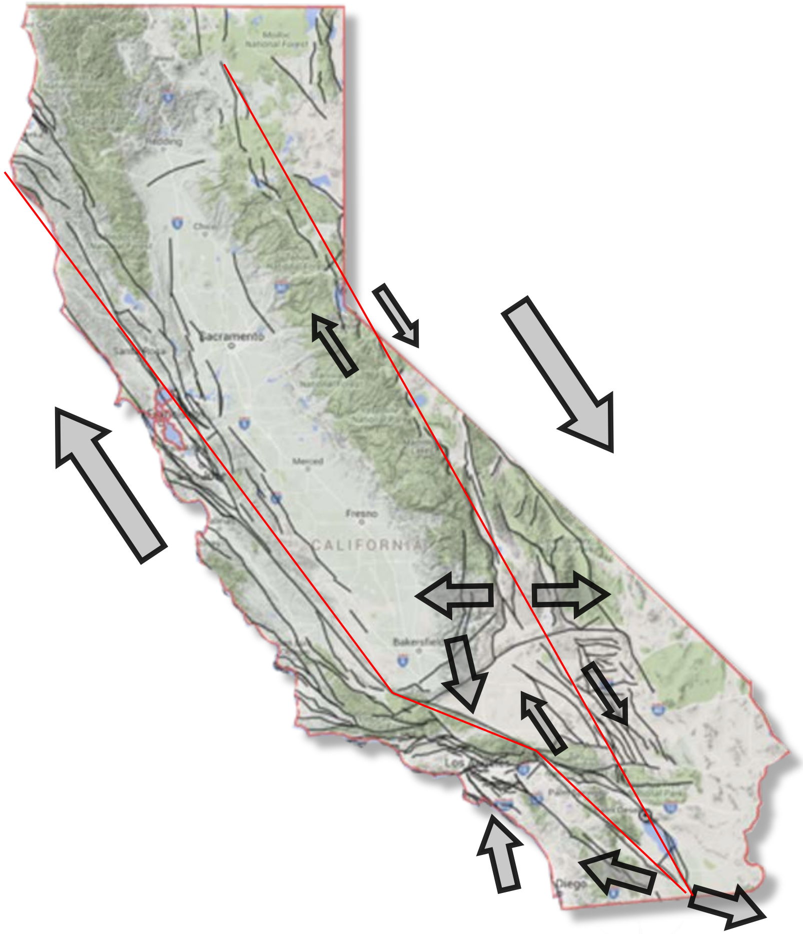




Objective

Create a uniform crustal deformation reference model for the active plate margin of California

- Fused InSAR, topographic, and GNSS geodetic imaging data
- Quantify uncertainties for the reference model
- Improve earthquake forecast models
- Improve understanding of the physical processes leading to and following earthquakes



Presentation Contents

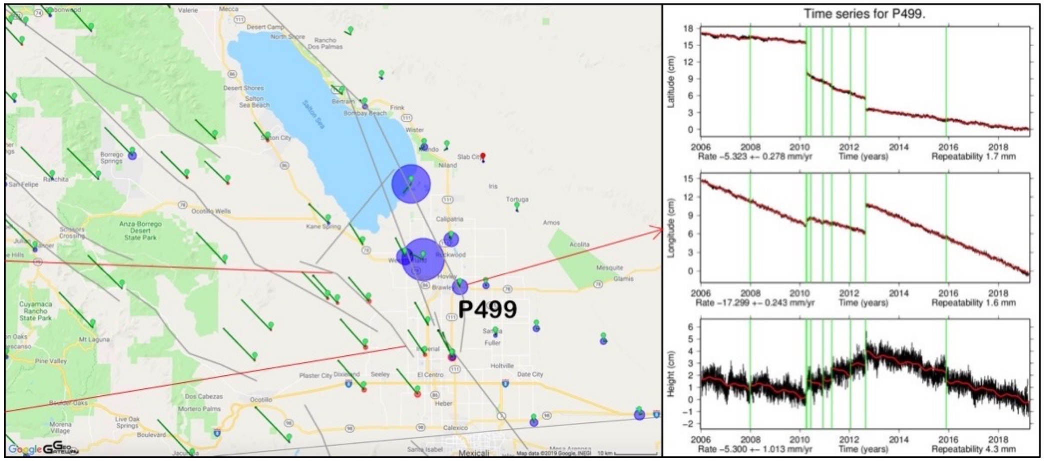
- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

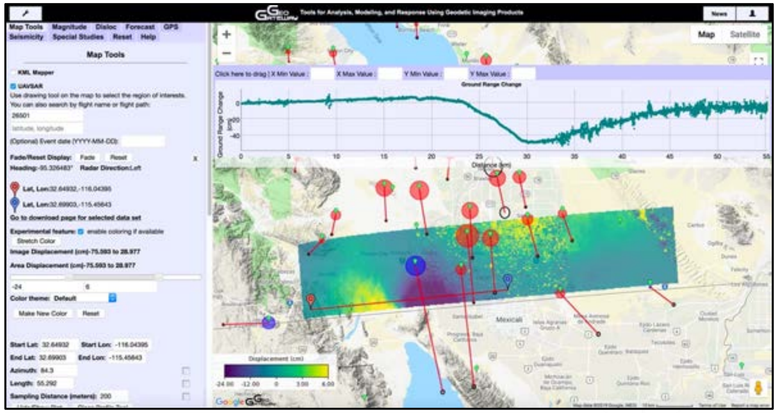
GNSS

UAVSAR

Technical and Science Advancements

- Fuse InSAR, topographic, and GNSS geodetic imaging data
  - Use GNSS data to adjust UAVSAR baseline estimate (position difference between first and second pass)
  - Extract features in InSAR images
  - Develop clustering algorithms to identify deformation boundaries in GNSS data
- Quantify uncertainties for the reference model
- Improve earthquake forecast models
- Improve understanding of the physical processes leading to and following earthquakes
- Uniform crustal deformation model serves as reference for modeling and analysis





Map Tools

☐ KML Mapper

☑ UAVSAR

Use drawing tool on the map to select the region of interests.
You can also search by flight name or flight path:

26501

latitude, longitude

(Optional) Event date (YYYY-MM-DD):

Fade/Reset Display: Fade Reset

Heading:-95.326483° Radar Direction:Left

Lat, Lon:32.64932,-116.04395

Lat, Lon:32.69903,-115.45643

Go to download page for selected data set

Experimental feature: ☑ enable coloring if available

Stretch Color

Image Displacement (cm)-75.593 to 28.977

Area Displacement (cm)-75.593 to 28.977

-24 6

Color theme: Default

Make New Color Reset

Start Lat: 32.64932 Start Lon: -116.04395

End Lat: 32.69903 End Lon: -115.45643

Azimuth: 84.3

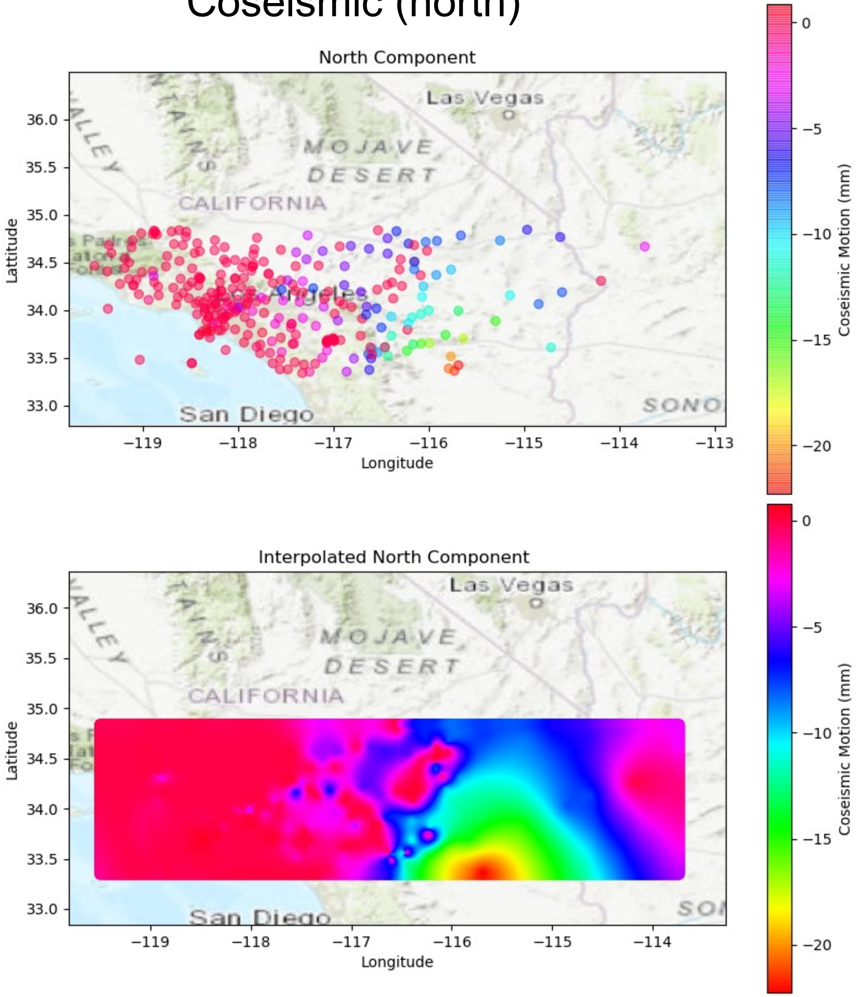
Length: 55.292

Sampling Distance (meters): 200

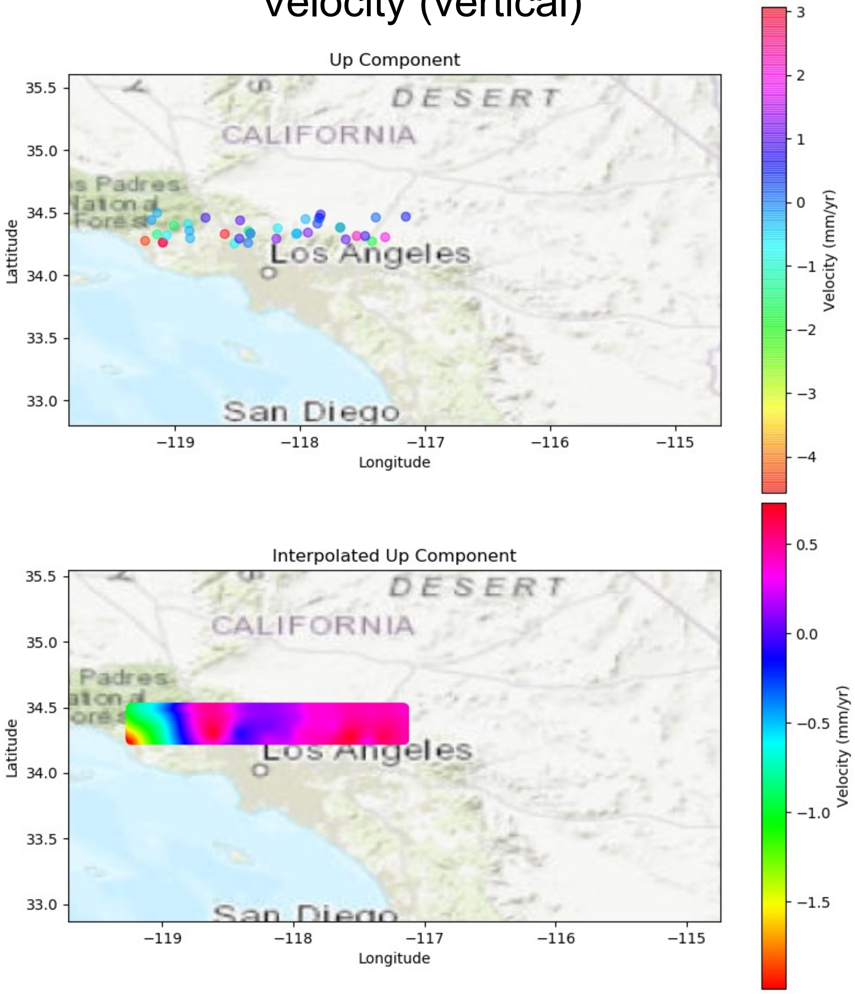
Click here to drag | X Min Value : X Max Value : Y Min Value : Y Max Value :

Interpolation

Coseismic (north)

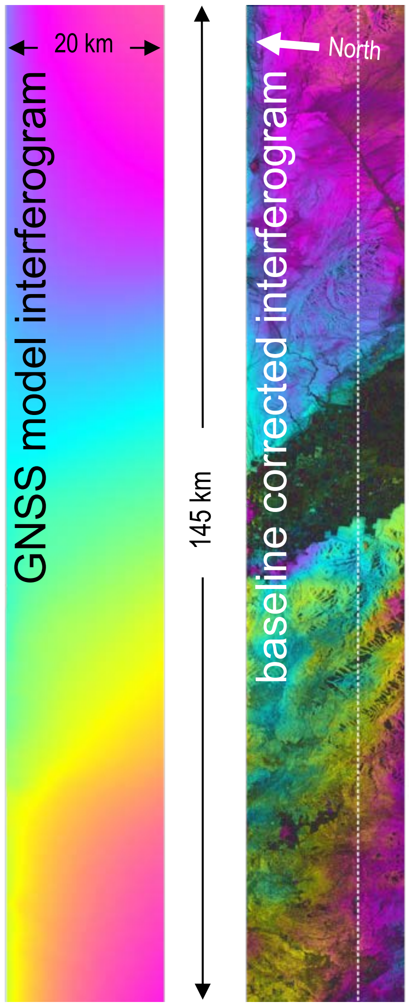


Velocity (vertical)

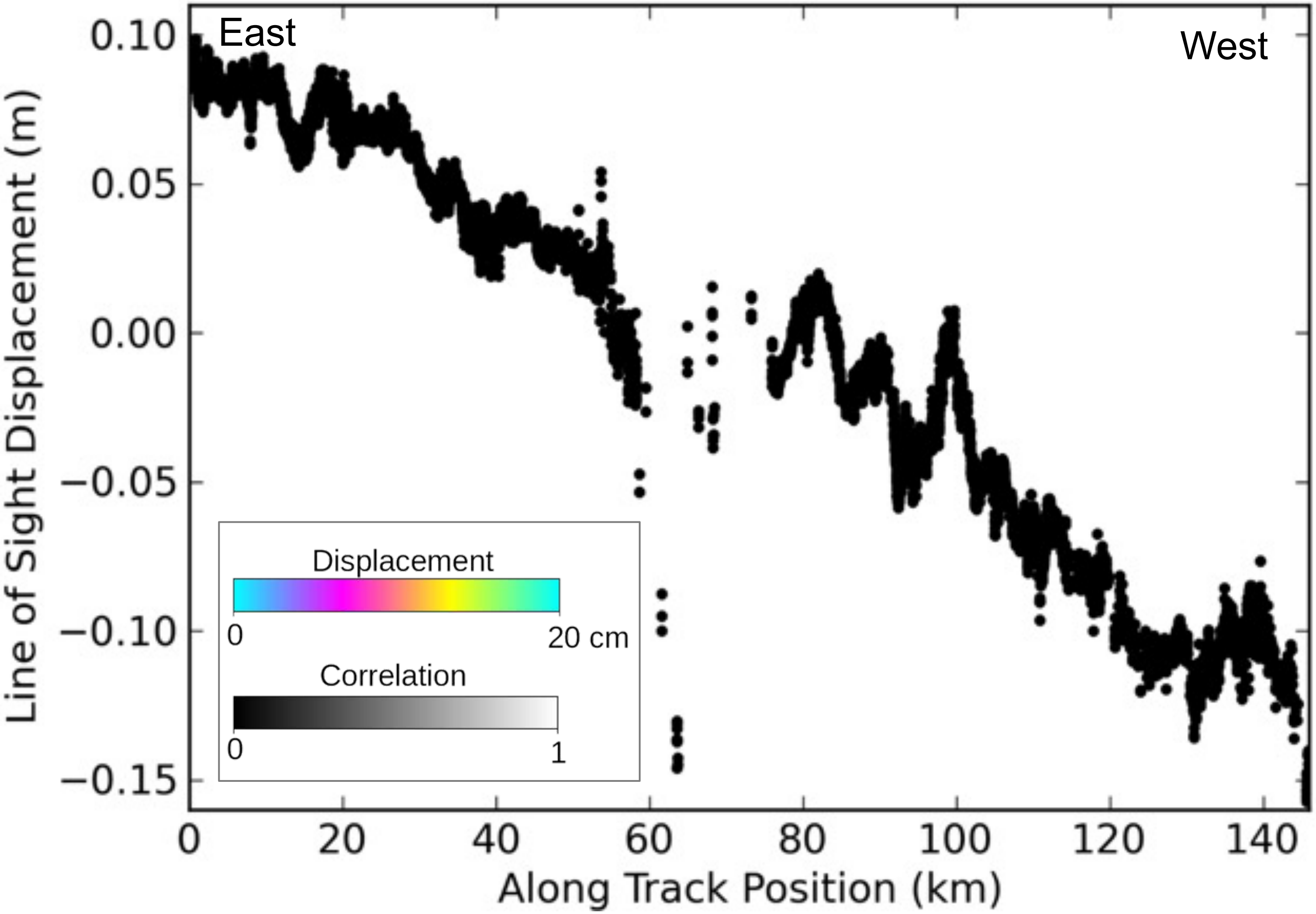


1. Creates synthetic interferogram for UAVSAR baseline adjustment
2. Creates initial uniform posting gridded deformation field

UAVSAR Baseline Adjustment



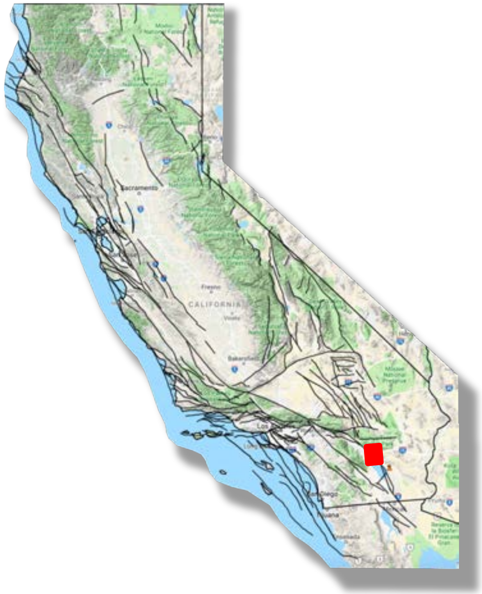
Line-of-sight motion across image for velocity product

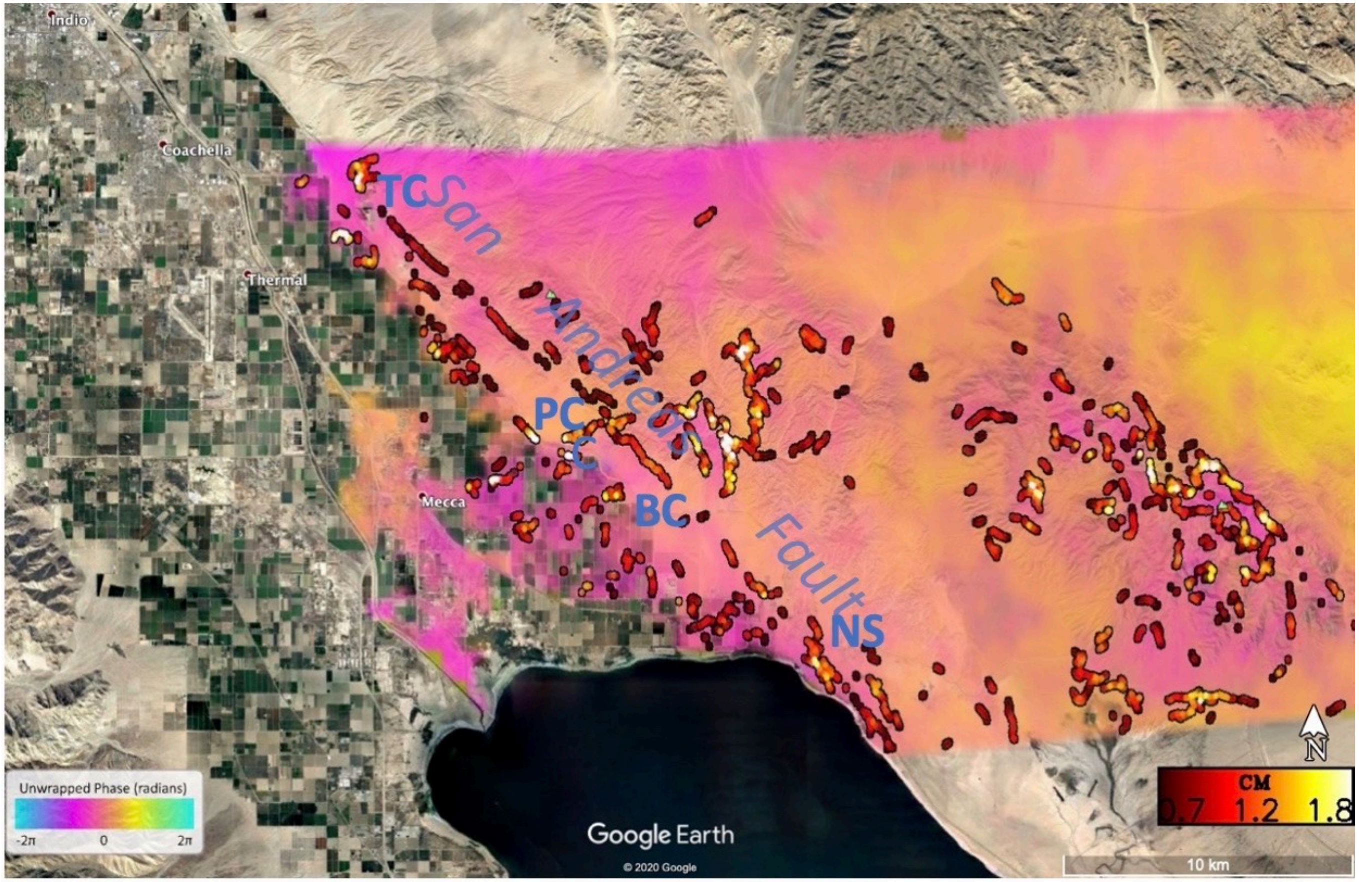


Enables extraction of plate tectonic motion and variations

Insar Edge Detection

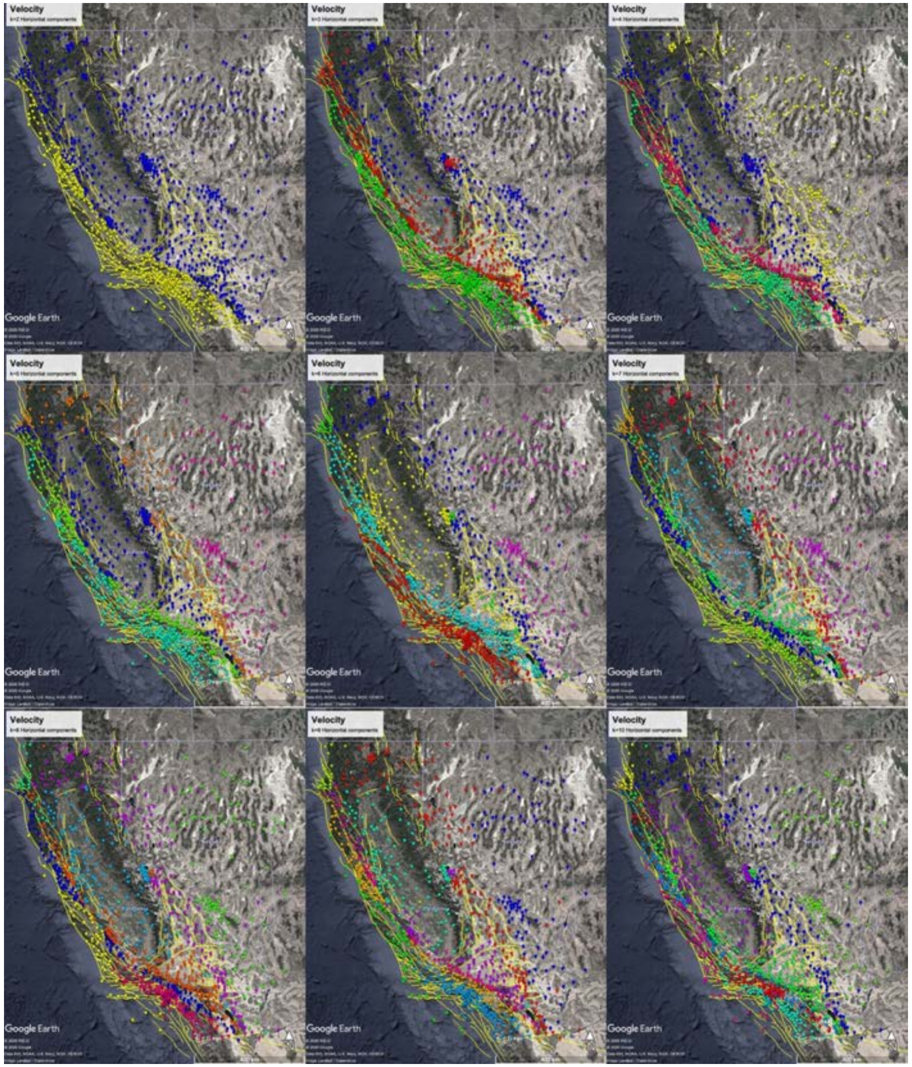
Detected edges and amplitude of slip

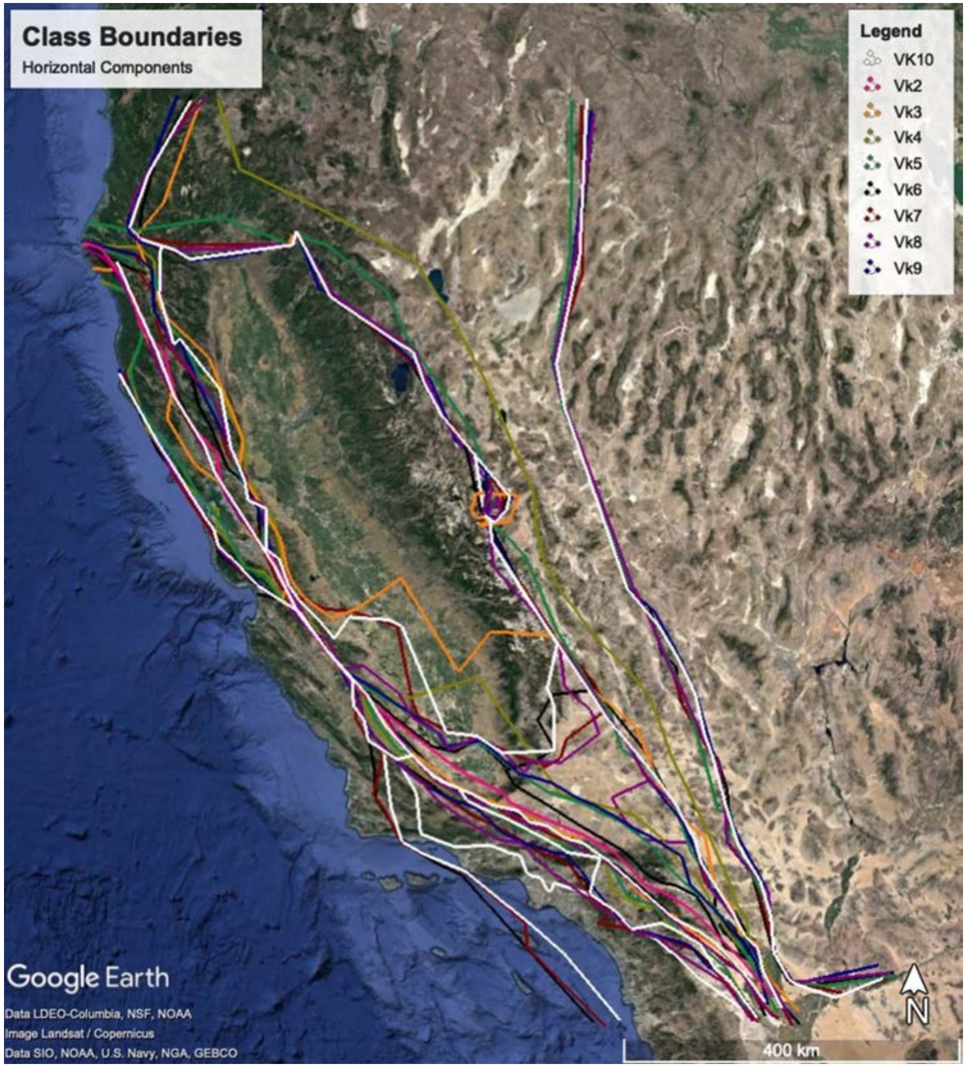




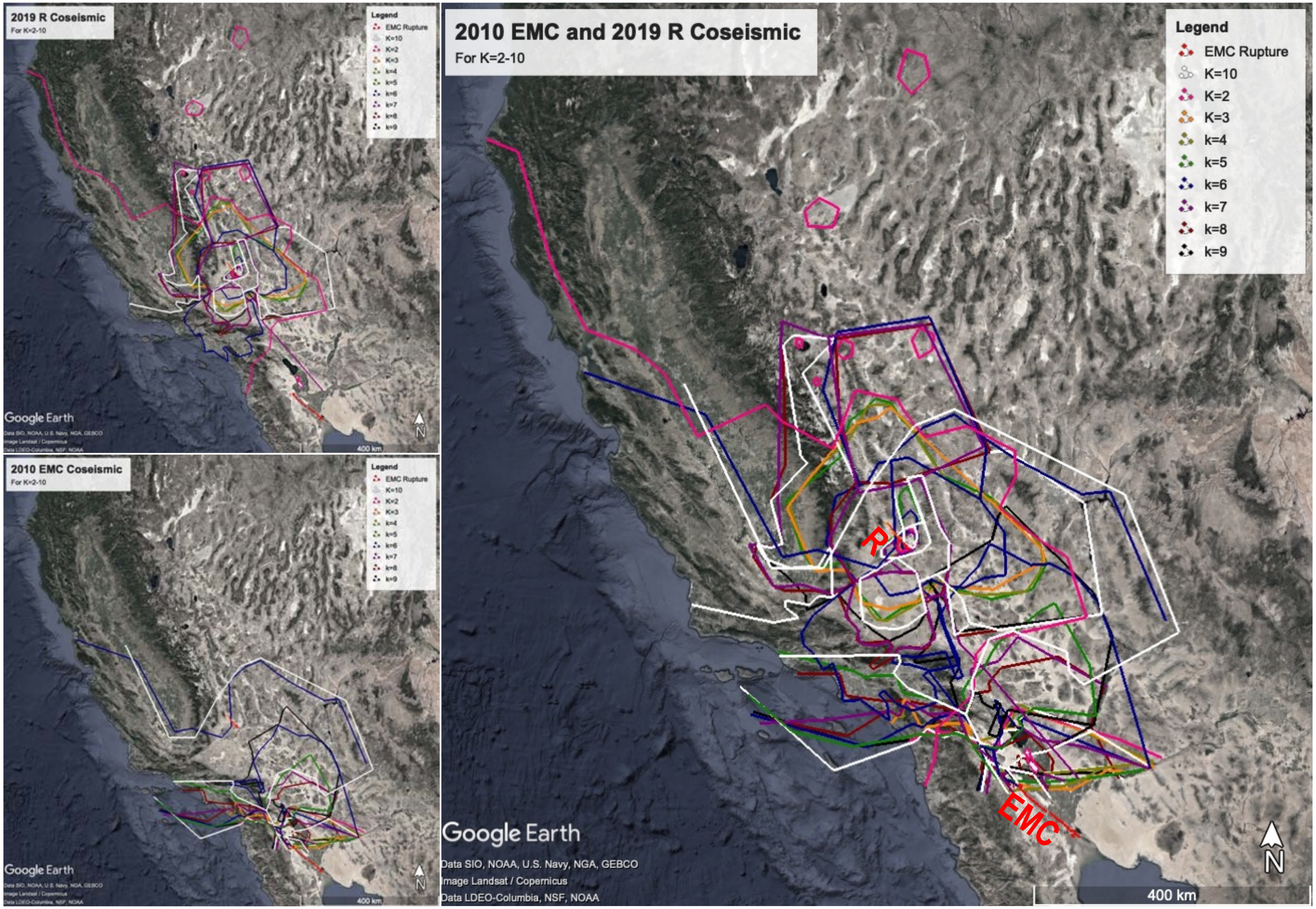
GNSS Clustering

Velocities k=2-10

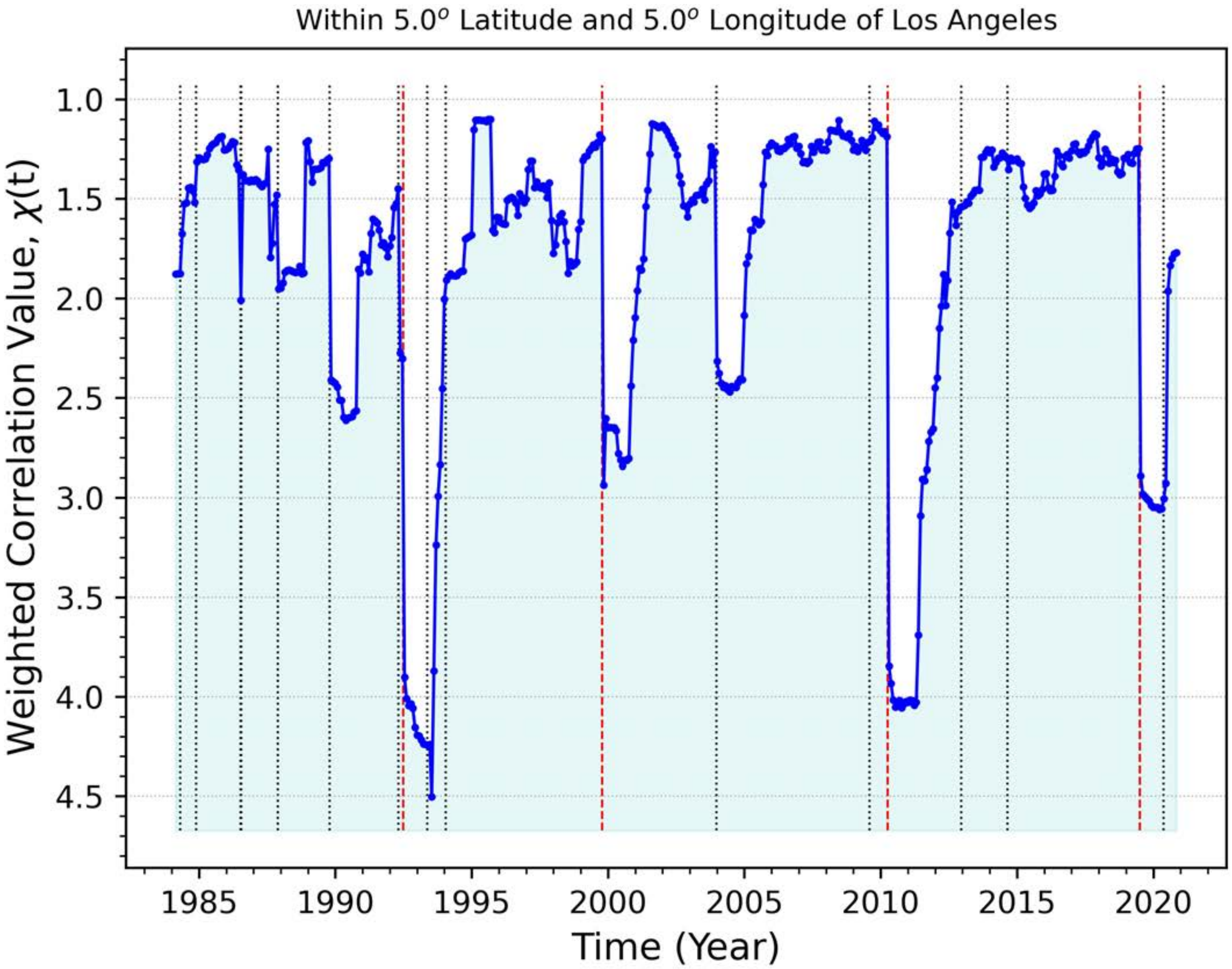

Velocity
k=2 Horizontal components
Velocity
k=3 Horizontal components
Velocity
k=4 Horizontal components
Velocity
k=5 Horizontal components
Velocity
k=6 Horizontal components
Velocity
k=7 Horizontal components
Velocity
k=8 Horizontal components
Velocity
k=9 Horizontal components
Velocity
k=10 Horizontal components
Google Earth


Class Boundaries
Horizontal Components
Legend
VK10
Vk2
Vk3
Vk4
Vk5
Vk6
Vk7
Vk8
Vk9
Google Earth
Data LDEO-Columbia, NSF, NOAA
Image Landsat / Copernicus
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
400 km
N

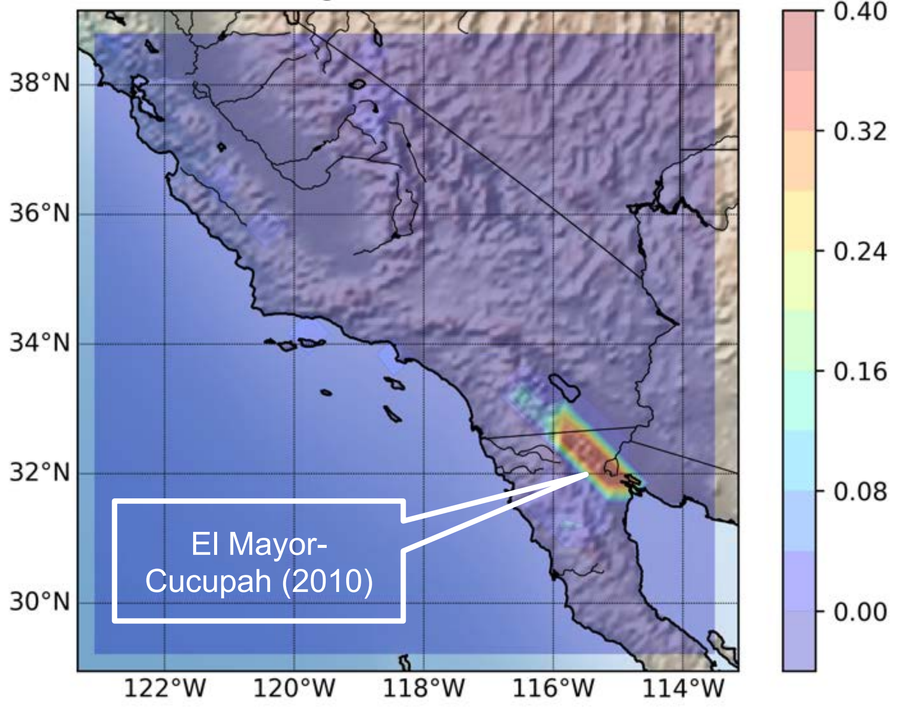
Coseismic Clustering Results



Regional Seismicity Correlation Timeseries

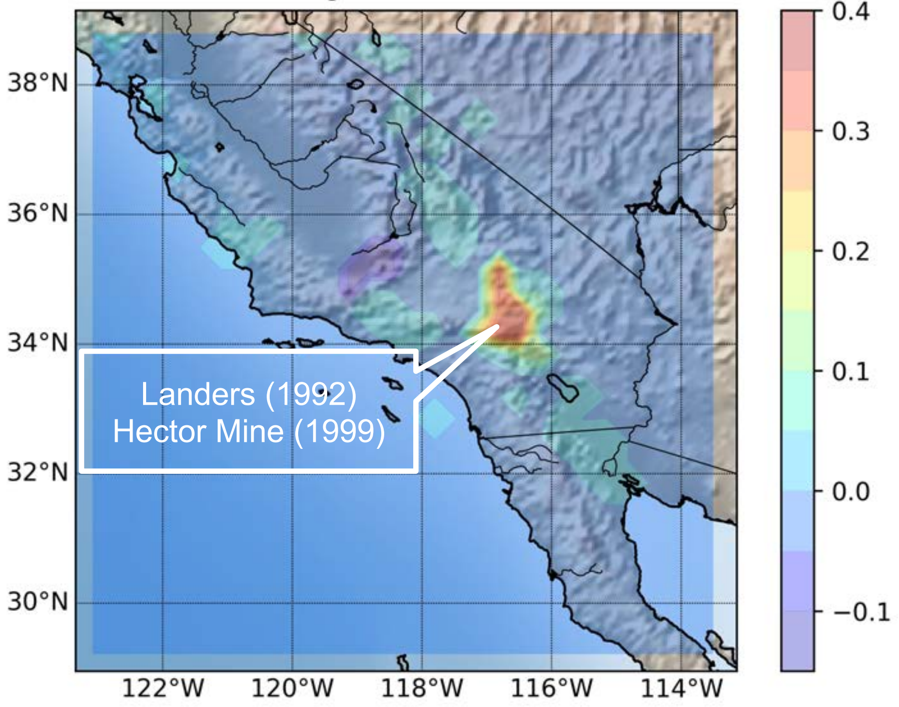


Eigen Patterns from Seismicity


EigenPattern 1 From 1950/01/01 To 2021/01/18

Percentage of Correlation: 6.22%

El Mayor-Cucupah (2010)


EigenPattern 2 From 1950/01/01 To 2021/01/18

Percentage of Correlation: 5.5%

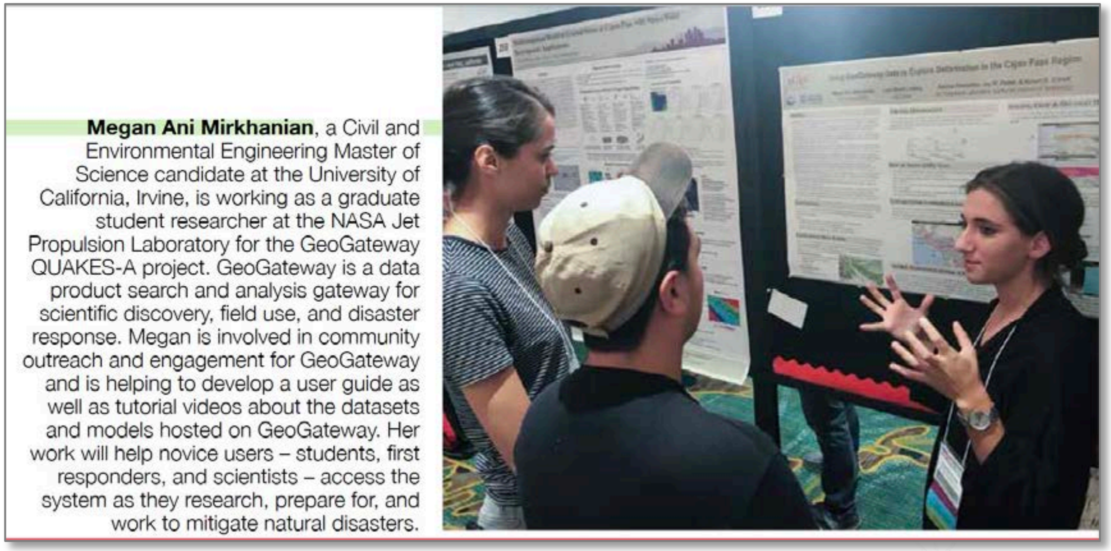
Landers (1992)
Hector Mine (1999)

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Summary of Accomplishments and Future Plans

- Developed and demonstrated method and workflow for carrying out UAVSAR baseline adjustment
- InSAR feature extraction methodology was completed and demonstrated (Parker et al, in preparation)
- Clustering algorithms were developed to identify deformation boundaries in GNSS data (Granat et al, in preparation)
- Uncertainty quantification methods are under consideration and evaluation
- GNSS clustering methodology is being used to guide the development of geodetic/seismicity earthquake forecasts
- GeoGateway has been rewritten using new standards and was released in December
- Userguide was developed and the team taught a workshop on the use of GeoGateway at the Annual Geological Society of America Meeting in October
- Student Megan Mirkhanian was featured in the annual ESTO report



Megan Ani Mirkhanian, a Civil and Environmental Engineering Master of Science candidate at the University of California, Irvine, is working as a graduate student researcher at the NASA Jet Propulsion Laboratory for the GeoGateway QUAKES-A project. GeoGateway is a data product search and analysis gateway for scientific discovery, field use, and disaster response. Megan is involved in community outreach and engagement for GeoGateway and is helping to develop a user guide as well as tutorial videos about the datasets and models hosted on GeoGateway. Her work will help novice users – students, first responders, and scientists – access the system as they research, prepare for, and work to mitigate natural disasters.

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Publications - List of Acronyms

Publications

- Donnellan, A., J. Parker, M. Heflin, M. Glasscoe, G. Lyzenga, M. Pierce, J. Wang, J. Rundle. L. Grant Ludwig, R. Granat, M. Mirkhanian, 2021, Improving Access to Geodetic Imaging Crustal Deformation Data Using GeoGateway, Earth Science Informatics, DOI: 10.1007/s12145-020-00561-7.
- Granat, R., A. Donnellan, M. Heflin, G. Lyzenga. M. Glasscoe, J. Parker, M. Pierce, J. Wang, J. Rundle, L. Grant Ludwig, in preparation, Clustering Analysis Methods for GNSS Observations: A Data-Driven Approach to Identifying California's Major Faults, Earth and Space Science
- Parker. J, A. Donnellan, R. Bilham, L Grant Ludwig, J. Wang, M. Pierce, N. Mowery, in preparation, Highly Resolved 2010 Triggered Creep on the Coachella Segment, San Andreas Fault, Earth and Space Science.
- Rundle, John B, and Andrea Donnellan, Nowcasting Earthquakes in Southern California With Machine Learning: Bursts, Swarms, and Aftershocks May Be Related to Levels of Regional Tectonic Stress, Earth and Space Science 7.9 (2020): e2020EA001097.
- Rundle, John B, Andrea Donnellan, James Crutchfield and Geoffrey Fox, Nowcasting earthquakes: Imaging the earthquake cycle in California with Machine Learning, to be submitted to Earth and Space Science.
- Rundle, John B., Seth Stein, Andrea Donnellan, Donald L Turcotte, William Klein and Cameron Saylor, The Complex Dynamics of Earthquake Fault Systems: New Approaches to Forecasting and Nowcasting of Earthquakes, revised, Reports on Progress in Physics (invited)
- Saylor, Cameron, John B Rundle, Andrea Donnellan , in review, Estimating Fault Configurations From InSAR Data Using A Genetic Algorithm, Earth and Space Science
- Parker, J. A. Donnellan, M. Glasscoe, submitted, Survey of Transverse Range Fire Scars in Ten Years of UAVSAR Polarimetry, Earth and Space Science.

List of Acronyms

QUAKES	Quantifying Uncertainty and Kinematics of Earthquake Systems
GNSS	Global Navigation Satellite System
InSAR	Interferometric Synthetic Aperture Radar
UAVSAR	Uninhabited Aerial Vehicle Synthetic Aperture Radar

Smart On-Demand Analysis of Multi-Temporal and Full Resolution SAR ARDs in Multi-Cloud & HPC

Hook Hua (PI, JPL)

Science Data System: Gerald Manipon (Co-I, JPL), Mohammed Karim (JPL), Marjorie Lucas (JPL), Zhangfan Xing (JPL), Joseph Jacob (JPL), Alex Dunn (JPL), Dustin Lo (JPL), Susan Neely (JPL)

Systems Engineer: Rishi Verma (JPL)

Flood and Damage Assessment: Sang-Hu Yun (Co-I, JPL), Jungkyo Jung (Co-I , JPL)

Solid Earth Science: Susan Owen (Co-I, JPL), David Bekaert (Co-I, JPL), Eric Fielding (Co-I, JPL)

Intern: David Tran

AIST-18-0085 Annual Technical Review

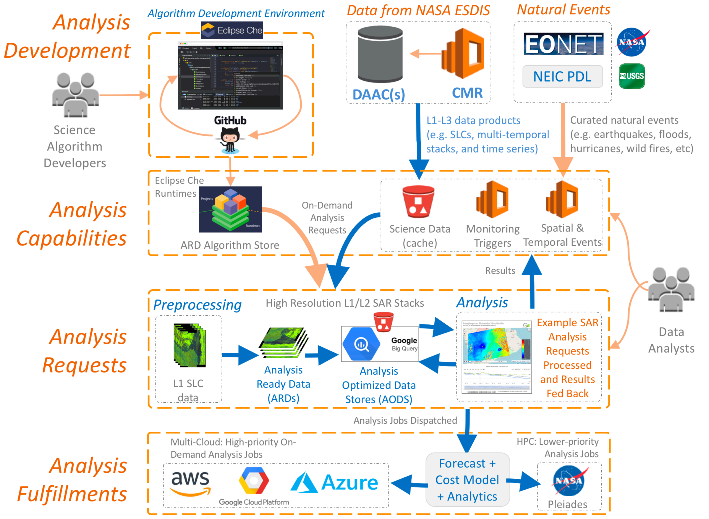
Friday, January 22, 2021

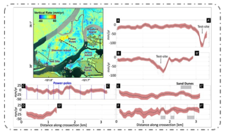
Smart On-Demand Analysis of Multi-Temporal and Full Resolution SAR ARDs in Multi-Cloud & HPC

PI: Hook Hua / JPL

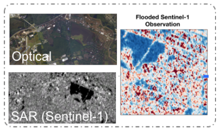
Objective:

- Address pain-points in the complexities of large-scale **algorithm development and on-demand analysis** of handling voluminous SAR measurements at full resolution from L1 SLCs to L3 time series.
- Increase **multi-temporal and full resolution SAR** data use as well as facilitate algorithm development and analysis for higher fidelity surface deformation and urgent response use cases.
- Enable **algorithm development** and deployment **at scale** in multi-cloud & HPC environment
- **Mitigate costs** of large-scale SAR data analysis

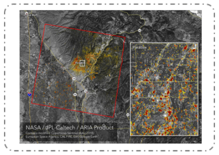

Analysis Development
Algorithm Development Environment
Eclipse Che
GitHub
Science Algorithm Developers
Data from NASA ESDIS
DAAC(s)
CMR
Natural Events
EONET
NEIC PDL
NASA
USGS
L1-L3 data products (e.g. SLCs, multi-temporal stacks, and time series)
Curated natural events (e.g. earthquakes, floods, hurricanes, wild fires, etc)
Analysis Capabilities
Eclipse Che Runtimes
Projects
Runtimes
ARD Algorithm Store
On-Demand Analysis Requests
Science Data (cache)
Monitoring Triggers
Spatial & Temporal Events
Results
Analysis Requests
Preprocessing
High Resolution L1/L2 SAR Stacks
L1 SLC data
Analysis Ready Data (ARDs)
Google Big Query
Analysis Optimized Data Stores (AODS)
Analysis
Example SAR Analysis Requests Processed and Results Fed Back
Data Analysts
Analysis Jobs Dispatched
Analysis Fulfillments
Multi-Cloud: High-priority On-Demand Analysis Jobs
aws
Google Cloud Platform
Azure
Forecast + Cost Model + Analytics
HPC: Lower-priority Analysis Jobs
NASA
Pleiades



Full Resolution Time Series Analysis



Multi-temporal-based Flood Assessment Maps



Multi-temporal-based Damage Proxy Maps

Approach:

- Generation of SAR Analysis Ready Data (**ARD**) using science notebook-based **algorithm development environment** where algorithms are deployed as *runtimes*
- On-demand analysis *runtimes* are run across multi-cloud (**AWS, Google Cloud Platform, and Microsoft Azure**) and **NASA HPC (Pleiades)** environments.
- Enabling "**smart on-demand**" where analysis are **ML-forecast and cost-model-informed** to help address the cost of large-scale analysis jobs across multi-cloud. E.g. optimizing for fast processing vs lower costs requests.
- Demonstration use cases for multi-temporal and full resolution SAR ARDs for **solid earth and urgent response**.

Co-Is: Gerald Manipon, Sang-Ho Yun, Eric Fielding, Jungkyo Jung, David Bekaert, Susan Owen, JPL

Key Milestones

Initial cloud-native SDS with EONET events	8/20
Multi-Temporal and Full Resolution SAR prototype ARD using Sentinel-1A/B	10/20
Integrate algorithm development environment with on-demand cloud science data processing	1/21
Analysis Processing on Multi-Cloud	10/21
Smart On-Demand Analysis with ML Forecasting and Estimation	1/22
Tech demo of time series and Change Detection (DPM and/or FPM) analysis from ADE	1/22

TRL$_{in}$ = 4

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Background

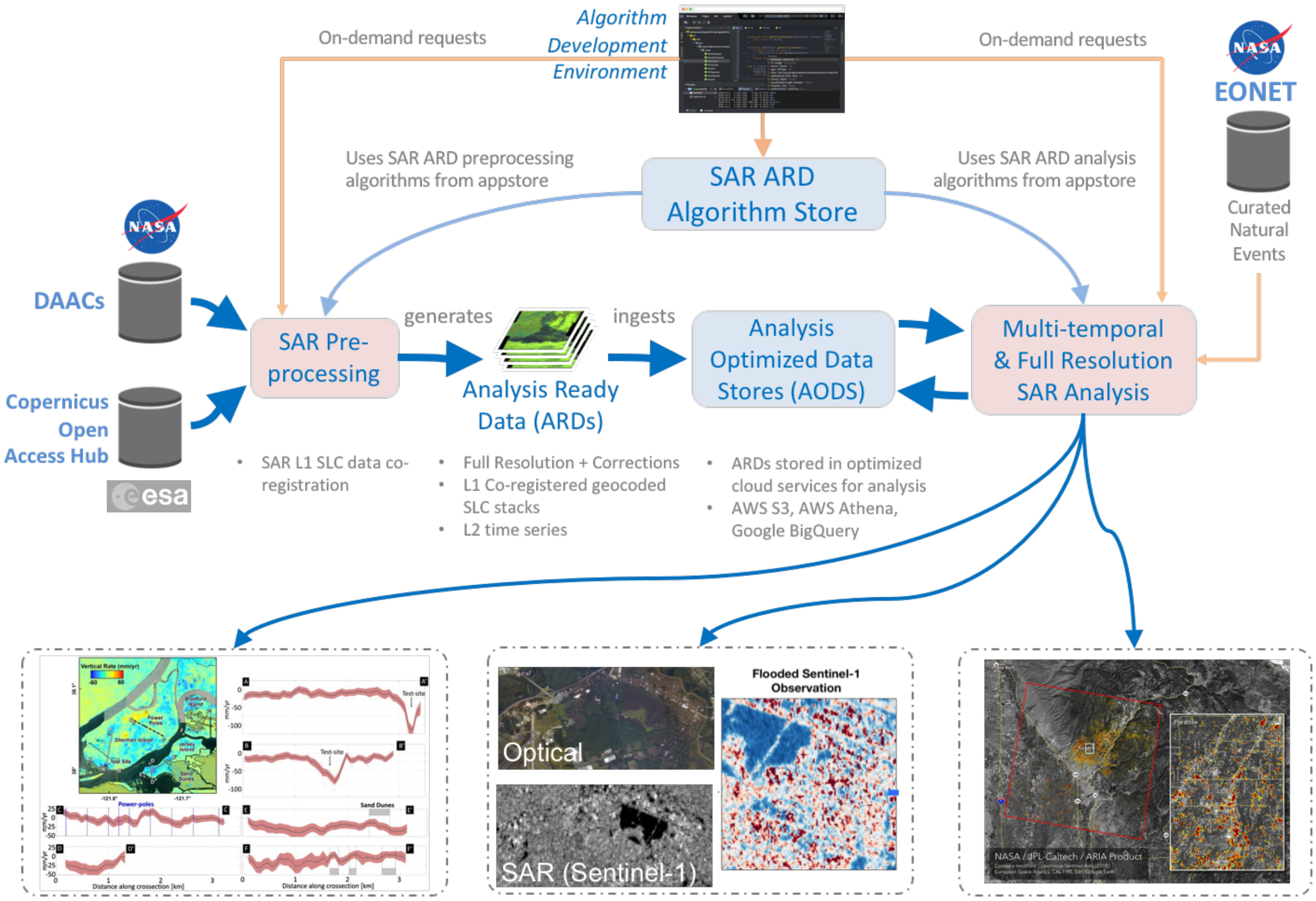
- Motivation
  - *Increasing **gap** between SDS in cloud capability vs algorithm development needs*
  - SAR data can aid in **decision making** for floods, earthquakes, and other monitoring and response scenarios where rapid information for **situational awareness** is required.
  - Increasing international SAR observations
  - SAR intrinsically **high** data volume, compute, and variety of algorithm analysis methods.

- Analytic Collaborative Framework (ACF)
  - *Address disconnect between **algorithm development and large-scale Science Data Systems (SDSes)** in the cloud*
  - *Enables more rapid time to market from algorithm development to data product generation, production, validation*
  - *Facilitating algorithm development of **multi-temporal and full resolution SAR analysis***
  - *Prototype an Analysis Ready Data (**ARD**) for SAR*

Objectives

- Address need for **rapid & scalable algorithm development** environment
- Provides pathways for algorithms to **run at large-scale** science data systems and corresponding **efficient handling of voluminous datasets**.
- Increase **accessibility** of **multi-sensor SAR analysis** to users
- **Cost-efficient** computational capacity for these larger L2 and L3 analysis is already becoming a bottleneck for effective algorithm development and analysis.
- Assess Analysis Ready Data (**ARD**) approach to SAR to consolidate algorithm development
- Demonstrate **multi-cloud** (AWS, Google Cloud Platform, Azure) and **NASA HEC** approach to on-demand processing
- Leverage **Machine Learning**-based cost optimization across multi-cloud

Objectives / Tech Advance



Algorithm Development Environment

On-demand requests

On-demand requests

SAR ARD Algorithm Store

Uses SAR ARD preprocessing algorithms from appstore

Uses SAR ARD analysis algorithms from appstore

EONET

Curated Natural Events

DAACs

Copernicus Open Access Hub

SAR Pre-processing

generates

Analysis Ready Data (ARDs)

ingests

Analysis Optimized Data Stores (AODS)

Multi-temporal & Full Resolution SAR Analysis

- SAR L1 SLC data co-registration

- Full Resolution + Corrections
- L1 Co-registered geocoded SLC stacks
- L2 time series

- ARDs stored in optimized cloud services for analysis
- AWS S3, AWS Athena, Google BigQuery

Full Resolution Time Series Analysis

(critical infrastructure monitoring, landslides, etc)

Multi-temporal-based Flood Assessment Maps

Multi-temporal-based Damage Proxy Maps

Presentation Contents

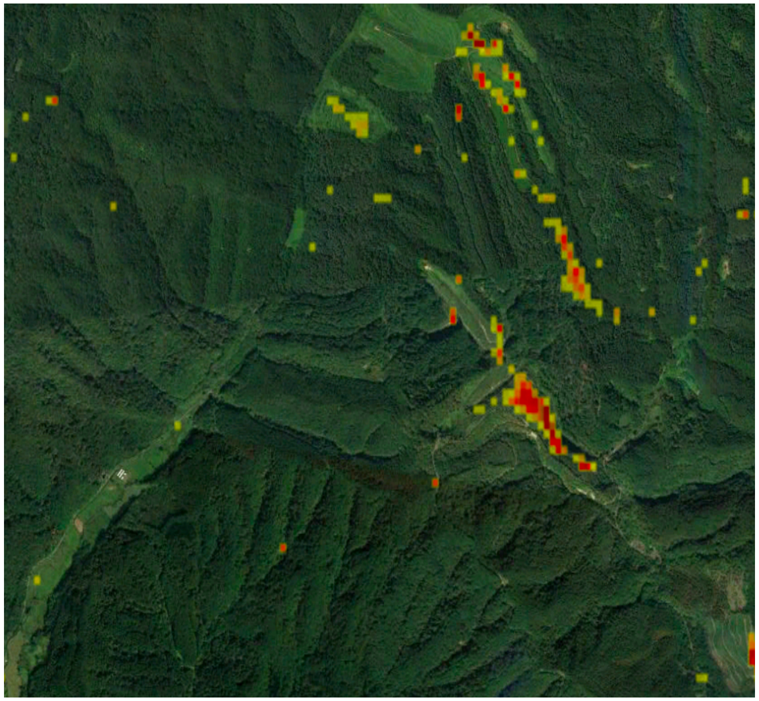
- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

ESTO
Earth Science Technology Office

Need for Algorithm Development--at Scale

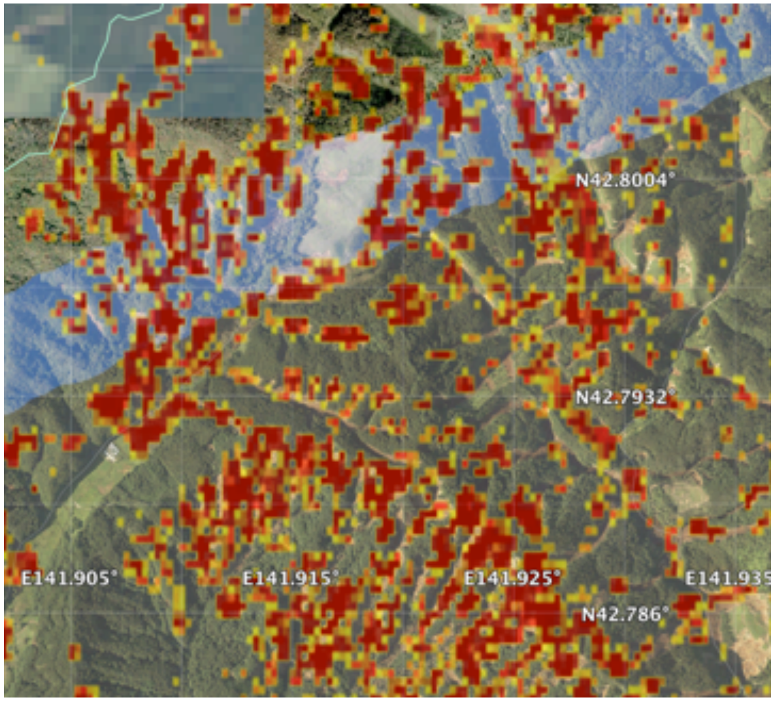
Source: Sang-Ho Yun, Jungkyo Jung

DPM1



Before/After Scenes
Processing: 1 hour
“Downloading”: 1.5 hours

DPM2/3



Time Series of Scenes
Processing: 26 days
“Downloading”: 40 hours

Landslides Triggered by the M6.6 Hokkaido Earthquake (Sept 2018)

NISAR and SWOT On-demand Needs

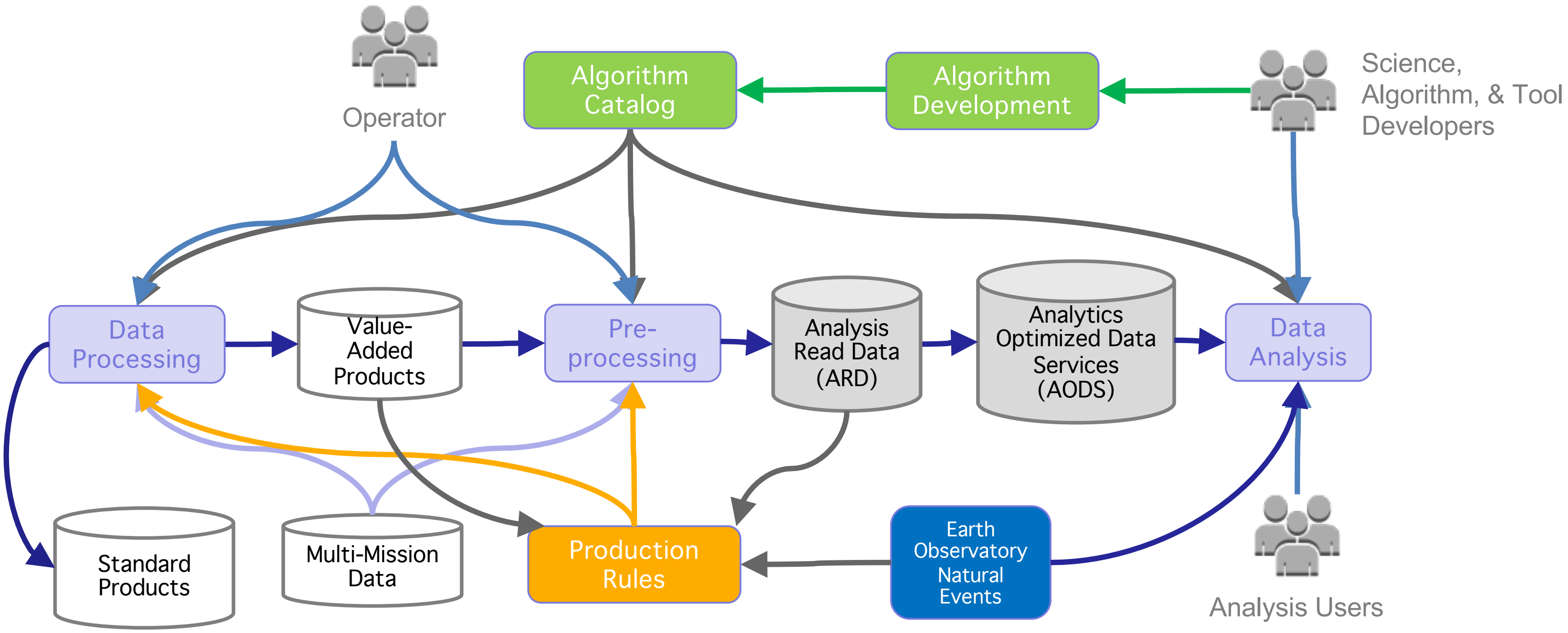
This AIST's technology demonstration is in alignment with NISAR and SWOT's on-demand needs :

1. **Type A: "Tunable" On-Demand Processing**
   - *"Bring your own parameters" scenario*
   - Trigger SDS to run standard product PGEs with custom tunable parameters.
     - Example: Re-run L2 GUNW generation but with nearest 3 neighbor pairing strategy (small-scale and large-scale processing in AWS).

2. **Type B: Science Notebook Development Environment (for L1-L3 Cal/Val and ADT)**
   - *"Bring your own code" scenario*
   - A Juypter notebook algorithm development environment that is collocated with SDS
     - Example: Running ISCE3 in a Juypter notebook next to L1 SLC data generated by SDS
   - Running notebooks at-scale in SDS
     - Example: Running global biomass estimate using custom L2 biomass model

3. **Type C: Automatic Generation of Custom Products in Keep-Up Mode**
   "Subscription" scenario
   - Triggering your own code or custom parameters based on new data stream
   - Allows custom code for urgent response and forward stream processing.
     - Example: Set up a variant of coherence change detection algorithm to run automatically for any new L1 SLC acquisitions.

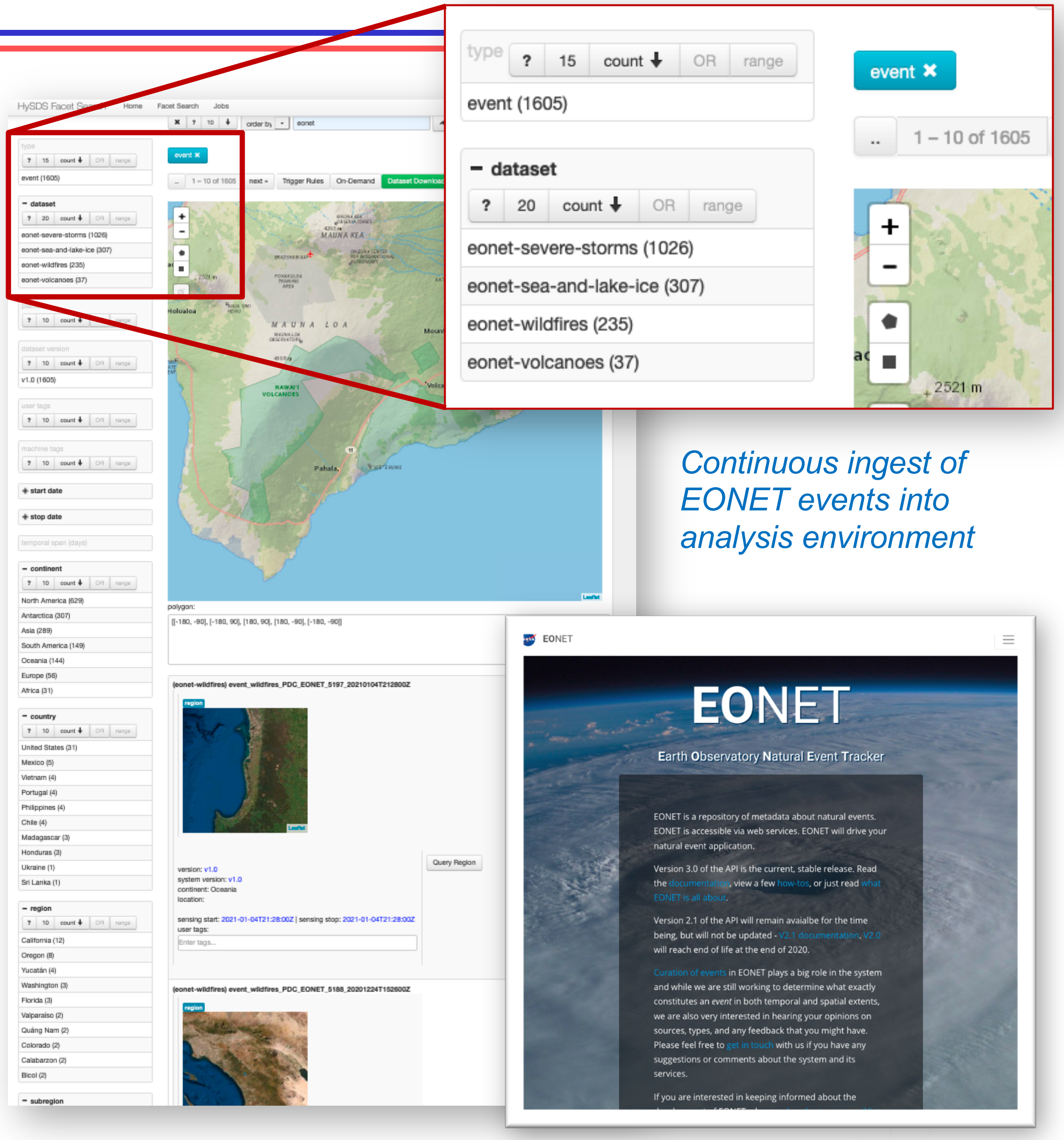
Key Concepts



- Algorithm development environment (Jupyter notebooks)
- Collocated in cloud with science data processing
- Algorithm test bed –at scale
- SAR ARDs for easier analysis
- Events catalog to natural events
- Production Rules Triggers to link events to automated analysis via user’s notebooks

Integration of NASA EONET Events

- Goal: to provide natural events as "triggers" for automating data processing with "notebook algorithms"
- NASA Earth Observatory Natural Event Tracker (EONET)
  - Providing a curated source of continuously updated natural event metadata.
- Curated Events
  - **Severe Storms:** Tropical Cyclones
    - National Hurricane Center
    - Joint Typhoon Warning Center
  - **Volcanoes**
    - Smithsonian/USGS Weekly Volcanic Activity Report
  - **Wildfires**
    - Alberta Wildfire
    - British Columbia Wildfire Service
    - California Department of Forestry and Fire Protection
    - InciWeb
    - Manitoba Wildfire Program
    - Pacific Disaster Center
  - **Sea and Lake Ice:** icebergs
    - National Ice Center



type
? 15 count ⬇ OR range
event (1605)

− dataset
? 20 count ⬇ OR range
eonet-severe-storms (1026)
eonet-sea-and-lake-ice (307)
eonet-wildfires (235)
eonet-volcanoes (37)

dataset version
? 10 count ⬇ OR range
v1.0 (1605)

user tags
? 10 count ⬇ OR range

machine tags
? 10 count ⬇ OR range

+ start date

+ stop date

temporal span (days)

− continent
? 10 count ⬇ OR range
North America (629)
Antarctica (307)
Asia (289)
South America (149)
Oceania (144)
Europe (56)
Africa (31)

− country
? 10 count ⬇ OR range
United States (31)
Mexico (5)
Vietnam (4)
Portugal (4)
Philippines (4)
Chile (4)
Madagascar (3)
Honduras (3)
Ukraine (1)
Sri Lanka (1)

− region
? 10 count ⬇ OR range
California (12)
Oregon (8)
Yucatán (4)
Washington (3)
Florida (3)
Valparaíso (2)
Quảng Nam (2)
Colorado (2)
Calabarzon (2)
Bicol (2)

− subregion

polygon:
[[-180, -90], [-180, 90], [180, 90], [180, -90], [-180, -90]]

(eonet-wildfires) event_wildfires_PDC_EONET_5197_20210104T212800Z

version: v1.0
system version: v1.0
continent: Oceania
location:

sensing start: 2021-01-04T21:28:00Z | sensing stop: 2021-01-04T21:28:00Z
user tags:
Enter tags...

(eonet-wildfires) event_wildfires_PDC_EONET_5188_20201224T152600Z

EONET is a repository of metadata about natural events. EONET is accessible via web services. EONET will drive your natural event application.

Version 3.0 of the API is the current, stable release. Read the documentation, view a few how-tos, or just read what EONET is all about.

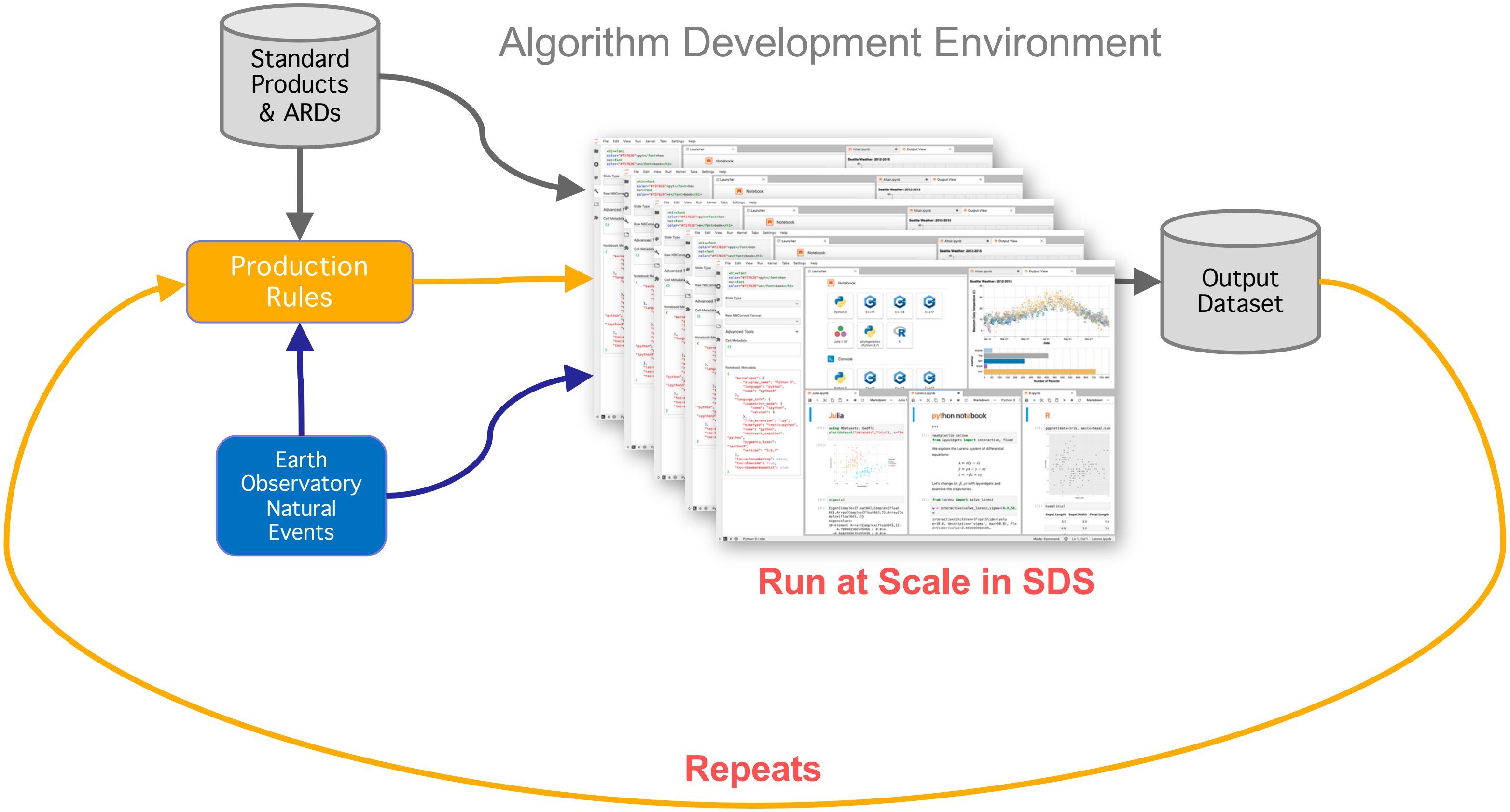
Version 2.1 of the API will remain avaialbe for the time being, but will not be updated - V2.1 documentation. V2.0 will reach end of life at the end of 2020.

Curation of events in EONET plays a big role in the system and while we are still working to determine what exactly constitutes an *event* in both temporal and spatial extents, we are also very interested in hearing your opinions on sources, types, and any feedback that you might have. Please feel free to get in touch with us if you have any suggestions or comments about the system and its services.

If you are interested in keeping informed about the

Continuous ingest of EONET events into analysis environment

Orchestration of Jupyter Notebooks—at scale



Algorithm Development Environment

Standard Products & ARDs

Production Rules

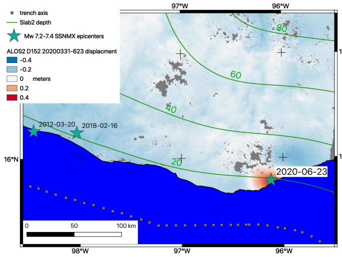
Earth Observatory Natural Events

Output Dataset

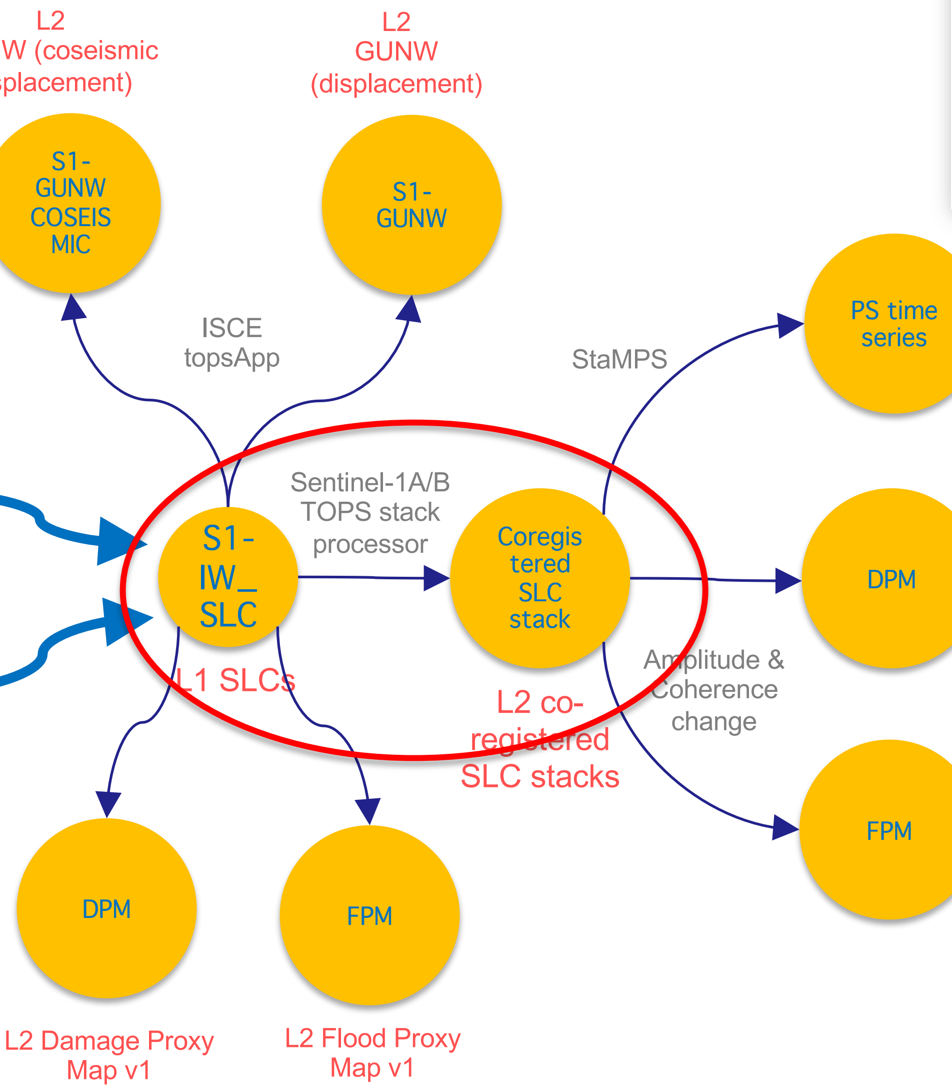
Run at Scale in SDS

Repeats

On-demand SAR Analysis and Products with Sentinel-1A/B



Source: Eric Fielding (JPL)



L2 GUNW (coseismic displacement)

S1-GUNW COSEIS MIC

L2 GUNW (displacement)

S1-GUNW

ISCE topsApp

StaMPS

ASF DAAC

Copernicus Open Access Hub

S1-IW_SLC

L1 SLCs

Sentinel-1A/B TOPS stack processor

Coregistered SLC stack

L2 co-registered SLC stacks

Amplitude & Coherence change

PS time series

L3 high-resolution time series

DPM

L3 Damage Proxy Map v2+

FPM

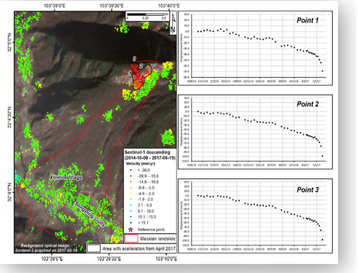
L3

DPM

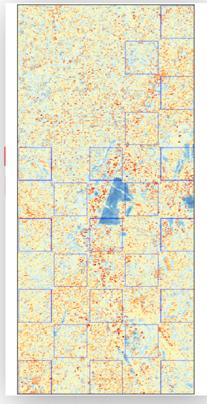
FPM

L2 Damage Proxy Map v1

L2 Flood Proxy Map v1



General Relationship of Strain and time of a series of creep deformation (Adapted from Saito, 1965)

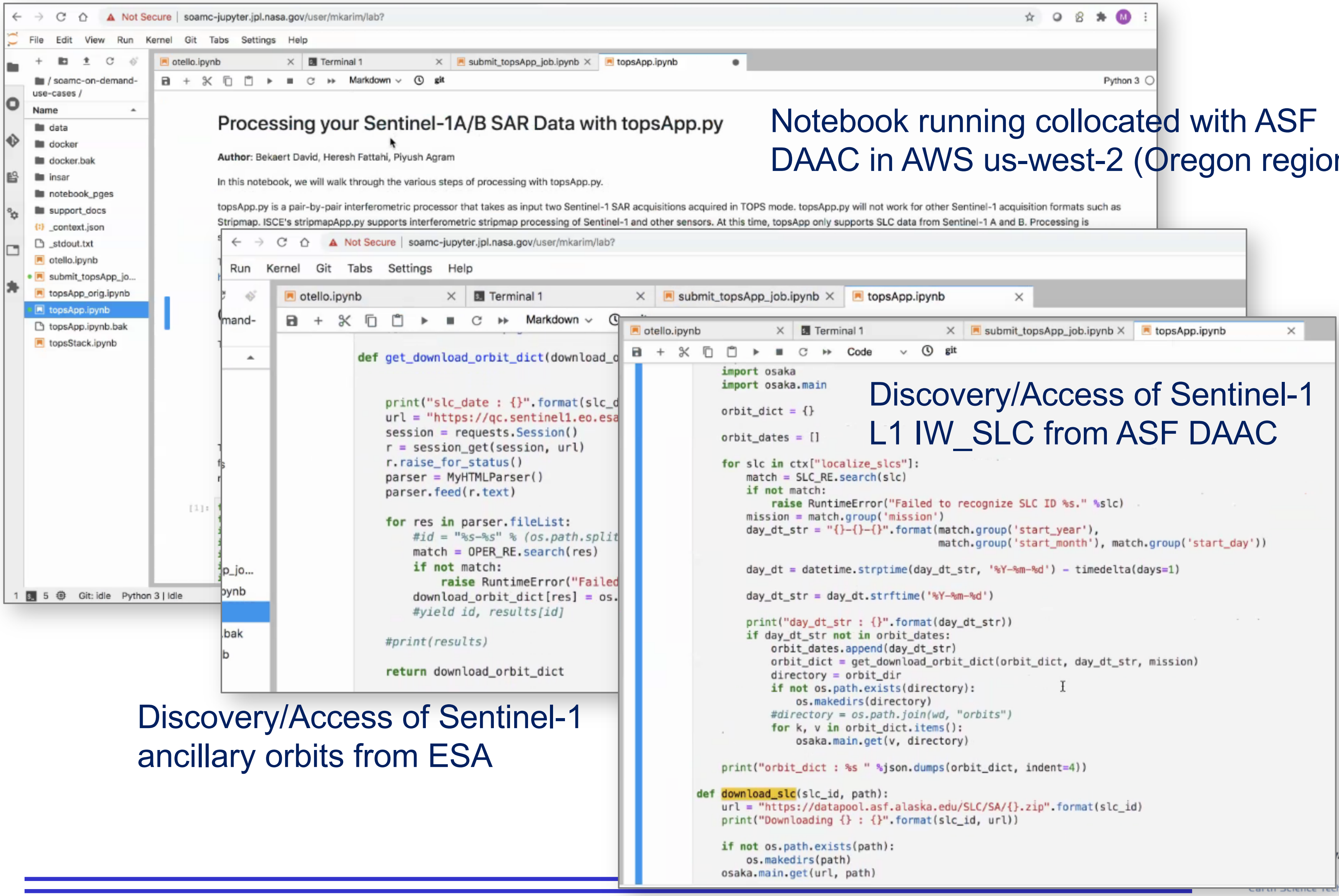


Example on-demand tunable parameters:

- Range looks
- Azimuth looks
- Filter strength
- Different DEMs
- Different phase unwrappers
- InSAR network pairings

ESTO Earth Science Technology Office

SAR Algorithms in Jupyter Notebooks Collocated with DAAC in AWS



Processing your Sentinel-1A/B SAR Data with topsApp.py

Author: Bekaert David, Heresh Fattahi, Piyush Agram

In this notebook, we will walk through the various steps of processing with topsApp.py.

topsApp.py is a pair-by-pair interferometric processor that takes as input two Sentinel-1 SAR acquisitions acquired in TOPS mode. topsApp.py will not work for other Sentinel-1 acquisition formats such as Stripmap. ISCE's stripmapApp.py supports interferometric stripmap processing of Sentinel-1 and other sensors. At this time, topsApp only supports SLC data from Sentinel-1 A and B. Processing is

```
def get_download_orbit_dict(download_o


    print("slc_date : {}".format(slc_d
    url = "https://qc.sentinel1.eo.esa
    session = requests.Session()
    r = session_get(session, url)
    r.raise_for_status()
    parser = MyHTMLParser()
    parser.feed(r.text)

    for res in parser.fileList:
        #id = "%s-%s" % (os.path.split
        match = OPER_RE.search(res)
        if not match:
            raise RuntimeError("Failed
        download_orbit_dict[res] = os.
        #yield id, results[id]

    #print(results)

    return download_orbit_dict
```

```
    import osaka
    import osaka.main

    orbit_dict = {}

    orbit_dates = []

    for slc in ctx["localize_slcs"]:
        match = SLC_RE.search(slc)
        if not match:
            raise RuntimeError("Failed to recognize SLC ID %s." %slc)
        mission = match.group('mission')
        day_dt_str = "{}-{}-{}".format(match.group('start_year'),
                                       match.group('start_month'), match.group('start_day'))

        day_dt = datetime.strptime(day_dt_str, '%Y-%m-%d') - timedelta(days=1)

        day_dt_str = day_dt.strftime('%Y-%m-%d')

        print("day_dt_str : {}".format(day_dt_str))
        if day_dt_str not in orbit_dates:
            orbit_dates.append(day_dt_str)
            orbit_dict = get_download_orbit_dict(orbit_dict, day_dt_str, mission)
            directory = orbit_dir
            if not os.path.exists(directory):
                os.makedirs(directory)
            #directory = os.path.join(wd, "orbits")
            for k, v in orbit_dict.items():
                osaka.main.get(v, directory)

    print("orbit_dict : %s " %json.dumps(orbit_dict, indent=4))

def download_slc(slc_id, path):
    url = "https://datapool.asf.alaska.edu/SLC/SA/{}.zip".format(slc_id)
    print("Downloading {} : {}".format(slc_id, url))

    if not os.path.exists(path):
        os.makedirs(path)
    osaka.main.get(url, path)
```

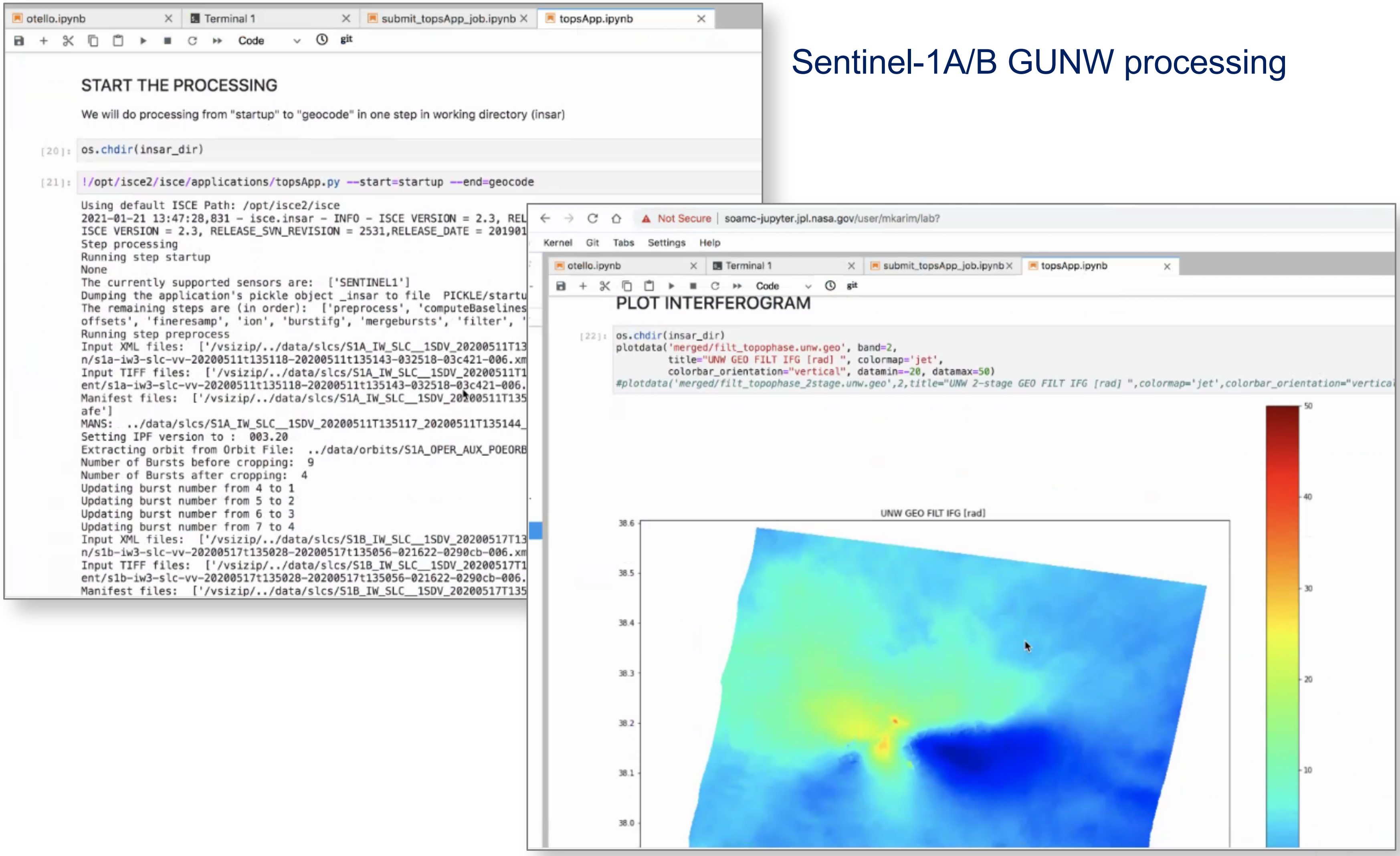
Notebook running collocated with ASF DAAC in AWS us-west-2 (Oregon region)

Discovery/Access of Sentinel-1 L1 IW_SLC from ASF DAAC

Discovery/Access of Sentinel-1 ancillary orbits from ESA

SAR Algorithms in Jupyter Notebooks Collocated with DAAC in AWS

Sentinel-1A/B GUNW processing



START THE PROCESSING

We will do processing from "startup" to "geocode" in one step in working directory (insar)

```
[20]: os.chdir(insar_dir)

[21]: !/opt/isce2/isce/applications/topsApp.py --start=startup --end=geocode

Using default ISCE Path: /opt/isce2/isce
2021-01-21 13:47:28,831 - isce.insar - INFO - ISCE VERSION = 2.3, REL
ISCE VERSION = 2.3, RELEASE_SVN_REVISION = 2531,RELEASE_DATE = 201901
Step processing
Running step startup
None
The currently supported sensors are:  ['SENTINEL1']
Dumping the application's pickle object _insar to file  PICKLE/startu
The remaining steps are (in order):  ['preprocess', 'computeBaselines
offsets', 'fineresamp', 'ion', 'burstifg', 'mergebursts', 'filter', '
Running step preprocess
Input XML files:  ['/vsizip/../data/slcs/S1A_IW_SLC__1SDV_20200511T13
n/s1a-iw3-slc-vv-20200511t135118-20200511t135143-032518-03c421-006.xm
Input TIFF files:  ['/vsizip/../data/slcs/S1A_IW_SLC__1SDV_20200511T1
ent/s1a-iw3-slc-vv-20200511t135118-20200511t135143-032518-03c421-006.
Manifest files:  ['/vsizip/../data/slcs/S1A_IW_SLC__1SDV_20200511T135
afe']
MANS:  ../data/slcs/S1A_IW_SLC__1SDV_20200511T135117_20200511T135144_
Setting IPF version to :  003.20
Extracting orbit from Orbit File:  ../data/orbits/S1A_OPER_AUX_POEORB
Number of Bursts before cropping:  9
Number of Bursts after cropping:  4
Updating burst number from 4 to 1
Updating burst number from 5 to 2
Updating burst number from 6 to 3
Updating burst number from 7 to 4
Input XML files:  ['/vsizip/../data/slcs/S1B_IW_SLC__1SDV_20200517T13
n/s1b-iw3-slc-vv-20200517t135028-20200517t135056-021622-0290cb-006.xm
Input TIFF files:  ['/vsizip/../data/slcs/S1B_IW_SLC__1SDV_20200517T1
ent/s1b-iw3-slc-vv-20200517t135028-20200517t135056-021622-0290cb-006.
Manifest files:  ['/vsizip/../data/slcs/S1B_IW_SLC__1SDV_20200517T135
```

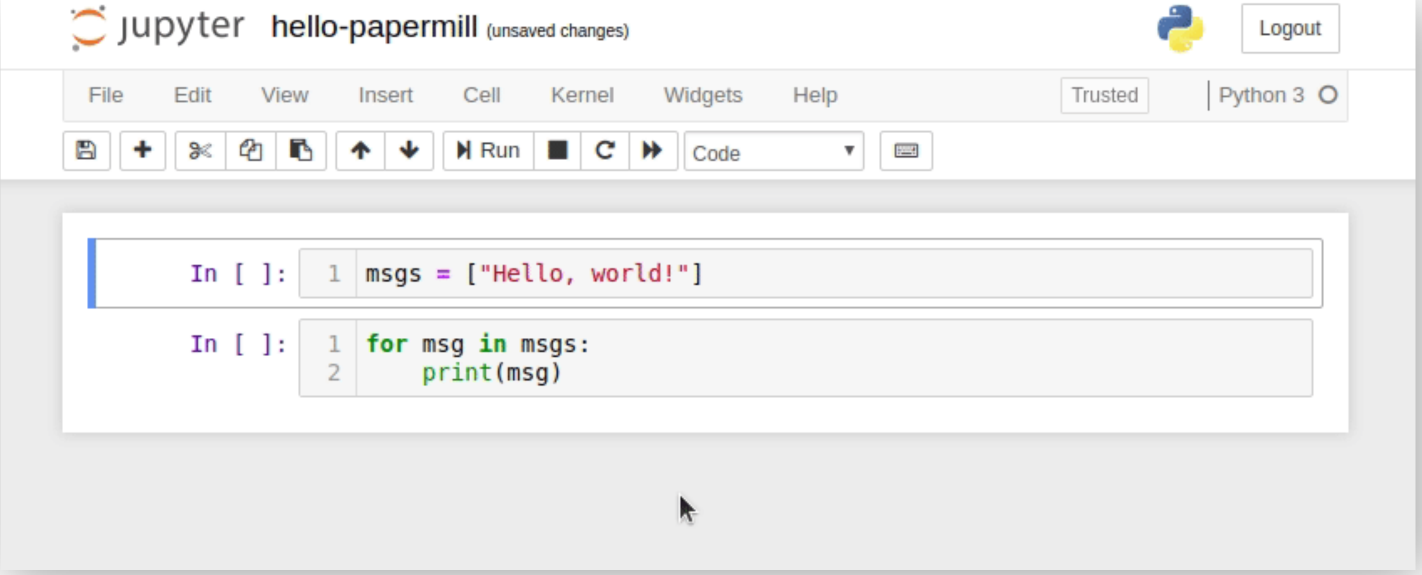
PLOT INTERFEROGRAM

```
[22]: os.chdir(insar_dir)
      plotdata('merged/filt_topophase.unw.geo', band=2,
               title="UNW GEO FILT IFG [rad] ", colormap='jet',
               colorbar_orientation="vertical", datamin=-20, datamax=50)
      #plotdata('merged/filt_topophase_2stage.unw.geo',2,title="UNW 2-stage GEO FILT IFG [rad] ",colormap='jet',colorbar_orientation="vertica
```

Automating Science Notebooks into Executable Containers

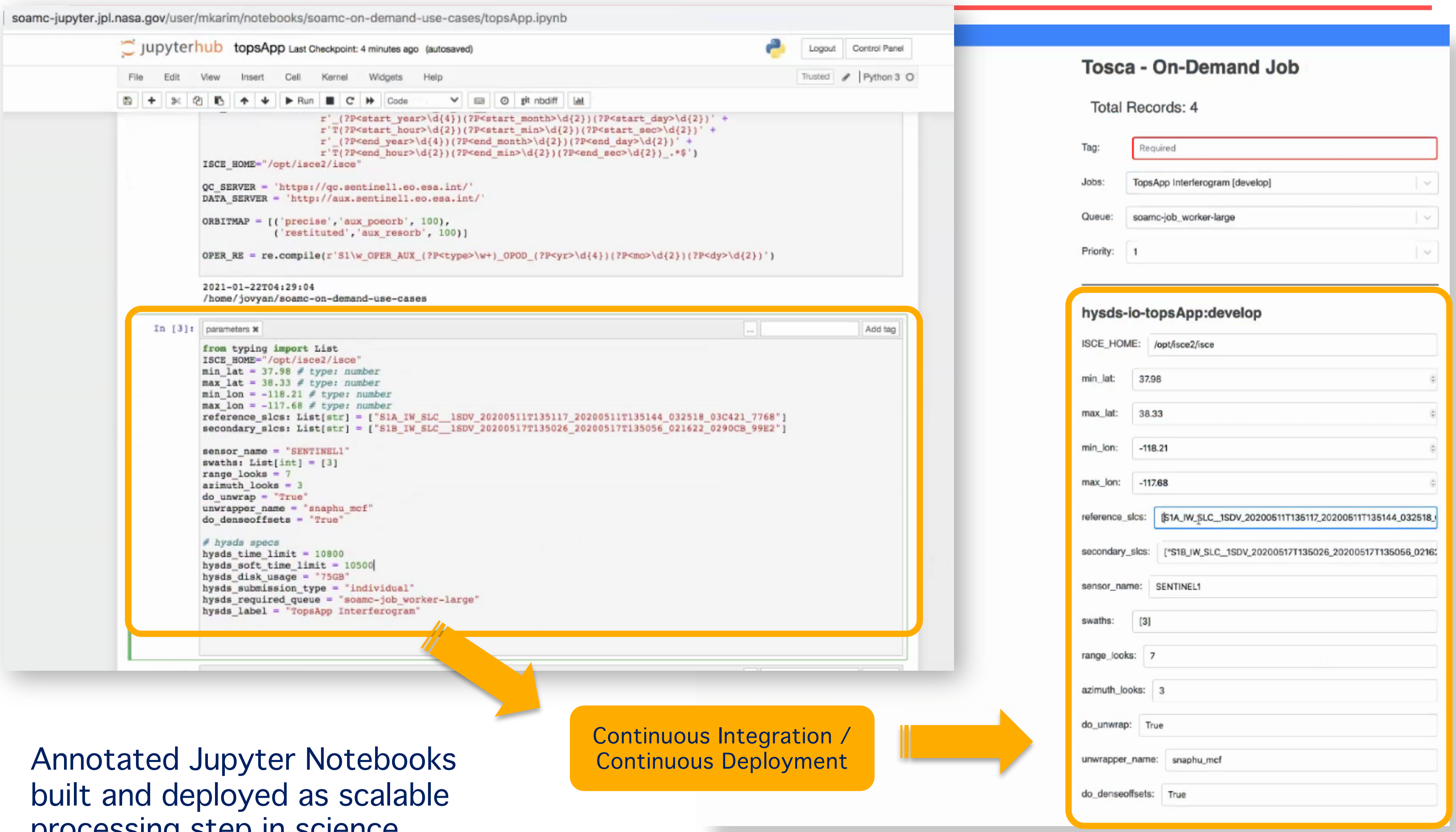
- Enable running same Jupyter notebooks at scale in SDS
  - Enables running large analysis with notebooks across collection of data

- Automated generation of Jupyter notebooks as executable containers
  - Building annotated science notebooks to execute with open source tool `papermill`, then Containerize, and deploy to SDS--to run at scale



```
In [ ]: msgs = ["Hello, world!"]

In [ ]: for msg in msgs:
            print(msg)
```

Registering Science Notebooks to Run at Scale in Science Data Systems (SDS)



soamc-jupyter.jpl.nasa.gov/user/mkarim/notebooks/soamc-on-demand-use-cases/topsApp.ipynb

```
                     r'_(?P<start_year>\d{4})(?P<start_month>\d{2})(?P<start_day>\d{2})' +
                     r'T(?P<start_hour>\d{2})(?P<start_min>\d{2})(?P<start_sec>\d{2})' +
                     r'_(?P<end_year>\d{4})(?P<end_month>\d{2})(?P<end_day>\d{2})' +
                     r'T(?P<end_hour>\d{2})(?P<end_min>\d{2})(?P<end_sec>\d{2})_.*$')
ISCE_HOME="/opt/isce2/isce"

QC_SERVER = 'https://qc.sentinel1.eo.esa.int/'
DATA_SERVER = 'http://aux.sentinel1.eo.esa.int/'

ORBITMAP = [('precise','aux_poeorb', 100),
            ('restituted','aux_resorb', 100)]

OPER_RE = re.compile(r'S1\w_OPER_AUX_(?P<type>\w+)_OPOD_(?P<yr>\d{4})(?P<mo>\d{2})(?P<dy>\d{2})')
```

```
2021-01-22T04:29:04
/home/jovyan/soamc-on-demand-use-cases
```

```
from typing import List
ISCE_HOME="/opt/isce2/isce"
min_lat = 37.98 # type: number
max_lat = 38.33 # type: number
min_lon = -118.21 # type: number
max_lon = -117.68 # type: number
reference_slcs: List[str] = ["S1A_IW_SLC__1SDV_20200511T135117_20200511T135144_032518_03C421_7768"]
secondary_slcs: List[str] = ["S1B_IW_SLC__1SDV_20200517T135026_20200517T135056_021622_0290CB_99E2"]

sensor_name = "SENTINEL1"
swaths: List[int] = [3]
range_looks = 7
azimuth_looks = 3
do_unwrap = "True"
unwrapper_name = "snaphu_mcf"
do_denseoffsets = "True"

# hysds specs
hysds_time_limit = 10800
hysds_soft_time_limit = 10500
hysds_disk_usage = "75GB"
hysds_submission_type = "individual"
hysds_required_queue = "soamc-job_worker-large"
hysds_label = "TopsApp Interferogram"
```

Continuous Integration / Continuous Deployment

Tosca - On-Demand Job

Total Records: 4

Tag: Required

Jobs: TopsApp Interferogram [develop]

Queue: soamc-job_worker-large

Priority: 1

hysds-io-topsApp:develop

ISCE_HOME: /opt/isce2/isce

min_lat: 37.98

max_lat: 38.33

min_lon: -118.21

max_lon: -117.68

reference_slcs: [S1A_IW_SLC__1SDV_20200511T135117_20200511T135144_032518_

secondary_slcs: ["S1B_IW_SLC__1SDV_20200517T135026_20200517T135056_0216:

sensor_name: SENTINEL1

swaths: [3]

range_looks: 7

azimuth_looks: 3

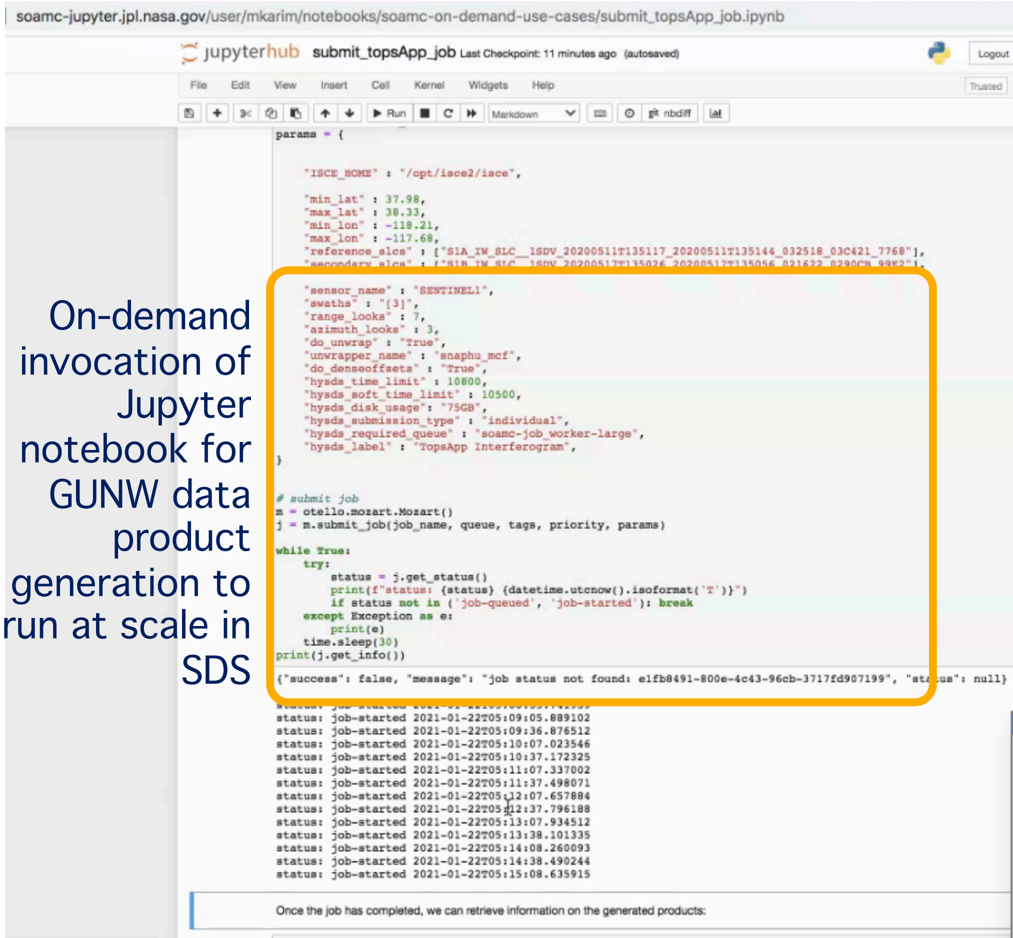
do_unwrap: True

unwrapper_name: snaphu_mcf

do_denseoffsets: True

Annotated Jupyter Notebooks built and deployed as scalable processing step in science data system

On-demand SAR Analysis in SDS at Scale, from Jupyter Notebooks


soamc-jupyter.jpl.nasa.gov/user/mkarim/notebooks/soamc-on-demand-use-cases/submit_topsApp_job.ipynb
jupyterhub submit_topsApp_job Last Checkpoint: 11 minutes ago (autosaved)
Logout
Control Panel
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Markdown
nbdiff

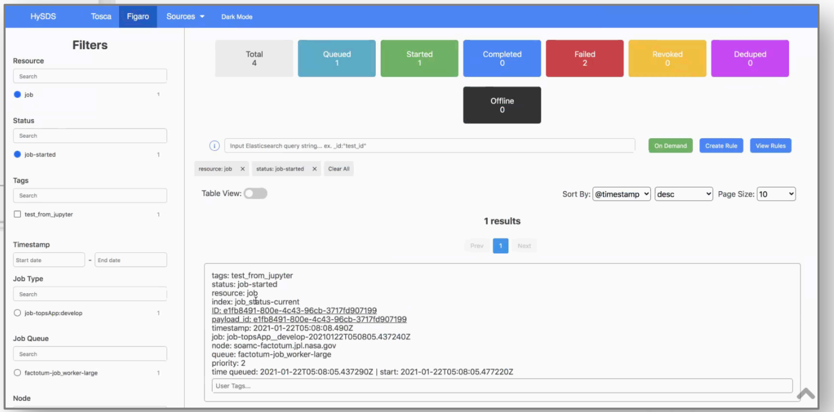
```
params = {

    "ISCE_HOME" : "/opt/isce2/isce",

    "min_lat" : 37.98,
    "max_lat" : 38.33,
    "min_lon" : -118.21,
    "max_lon" : -117.68,
    "reference_slcs" : ["S1A_IW_SLC__1SDV_20200511T135117_20200511T135144_032518_03C421_7768"],
    "secondary_slcs" : ["S1B_IW_SLC__1SDV_20200517T135026_20200517T135056_021622_0290CB_99E2"],

    "sensor_name" : "SENTINEL1",
    "swaths" : "[3]",
    "range_looks" : 7,
    "azimuth_looks" : 3,
    "do_unwrap" : "True",
    "unwrapper_name" : "snaphu_mcf",
    "do_denseoffsets" : "True",
    "hysds_time_limit" : 10800,
    "hysds_soft_time_limit" : 10500,
    "hysds_disk_usage": "75GB",
    "hysds_submission_type" : "individual",
    "hysds_required_queue" : "soamc-job_worker-large",
    "hysds_label" : "TopsApp Interferogram",
}


# submit job
m = otello.mozart.Mozart()
j = m.submit_job(job_name, queue, tags, priority, params)

while True:
    try:
        status = j.get_status()
        print(f"status: {status} {datetime.utcnow().isoformat('T')}")
        if status not in ('job-queued', 'job-started'): break
    except Exception as e:
        print(e)
    time.sleep(30)
print(j.get_info())
```

```
{"success": false, "message": "job status not found: e1fb8491-800e-4c43-96cb-3717fd907199", "status": null}
status: job-started 2021-01-22T05:09:05.889102
status: job-started 2021-01-22T05:09:36.876512
status: job-started 2021-01-22T05:10:07.023546
status: job-started 2021-01-22T05:10:37.172325
status: job-started 2021-01-22T05:11:07.337002
status: job-started 2021-01-22T05:11:37.498071
status: job-started 2021-01-22T05:12:07.657884
status: job-started 2021-01-22T05:12:37.796188
status: job-started 2021-01-22T05:13:07.934512
status: job-started 2021-01-22T05:13:38.101335
status: job-started 2021-01-22T05:14:08.260093
status: job-started 2021-01-22T05:14:38.490244
status: job-started 2021-01-22T05:15:08.635915
```

Once the job has completed, we can retrieve information on the generated products:

On-demand invocation of Jupyter notebook for GUNW data product generation to run at scale in SDS

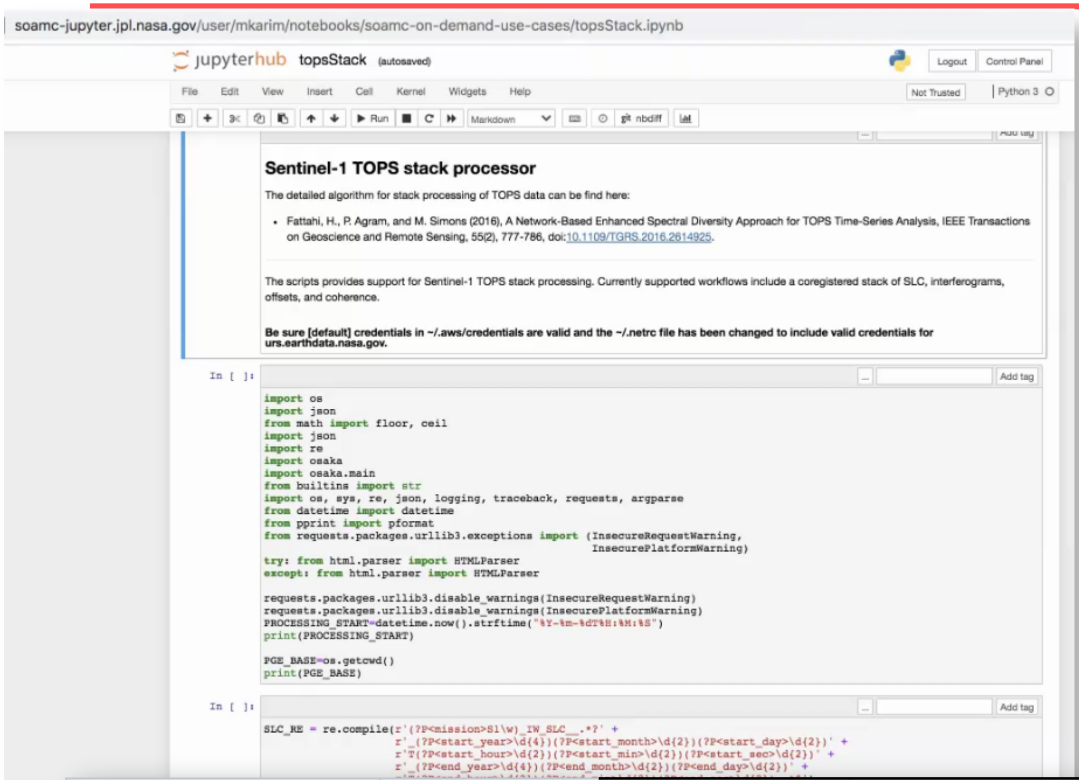
- Outer top-level driver notebook can be used to do map-reduce of mapping n-stacks to n-distributed jobs
- Original SAR analysis notebook deployed as Containerized processing step
  - Distributed data access
  - process a single product (e.g. SLC stack, GUNW)



tags: test_from_jupyter
status: job-started
resource: job
index: job_status-current
ID: e1fb8491-800e-4c43-96cb-3717fd907199
payload_id: e1fb8491-800e-4c43-96cb-3717fd907199
timestamp: 2021-01-22T05:08:08.490Z
job: job-topsApp__develop-20210122T050805.437240Z
node: soamc-factotum.jpl.nasa.gov
queue: factotum-job_worker-large
priority: 2
time queued: 2021-01-22T05:08:05.437290Z | start: 2021-01-22T05:08:05.477220Z

(right) On-demand notebook dispatched and running in SDS in auto-scaling fleet **using lower-cost AWS spot market**

ARD-like Coregistered SLC Stack Generation Example


soamc-jupyter.jpl.nasa.gov/user/mkarim/notebooks/soamc-on-demand-use-cases/topsStack.ipynb
jupyterhub topsStack (autosaved)
Logout
Control Panel
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3
Run
Markdown
nbdiff
Add tag

Sentinel-1 TOPS stack processor

The detailed algorithm for stack processing of TOPS data can be find here:

- Fattahi, H., P. Agram, and M. Simons (2016), A Network-Based Enhanced Spectral Diversity Approach for TOPS Time-Series Analysis, IEEE Transactions on Geoscience and Remote Sensing, 55(2), 777-786, doi:10.1109/TGRS.2016.2614925.

The scripts provides support for Sentinel-1 TOPS stack processing. Currently supported workflows include a coregistered stack of SLC, interferograms, offsets, and coherence.

Be sure [default] credentials in ~/.aws/credentials are valid and the ~/.netrc file has been changed to include valid credentials for urs.earthdata.nasa.gov.

In [ ]:

```
import os
import json
from math import floor, ceil
import json
import re
import osaka
import osaka.main
from builtins import str
import os, sys, re, json, logging, traceback, requests, argparse
from datetime import datetime
from pprint import pformat
from requests.packages.urllib3.exceptions import (InsecureRequestWarning,
                                                  InsecurePlatformWarning)
try: from html.parser import HTMLParser
except: from html.parser import HTMLParser

requests.packages.urllib3.disable_warnings(InsecureRequestWarning)
requests.packages.urllib3.disable_warnings(InsecurePlatformWarning)
PROCESSING_START=datetime.now().strftime("%Y-%m-%dT%H:%M:%S")
print(PROCESSING_START)

PGE_BASE=os.getcwd()
print(PGE_BASE)
```

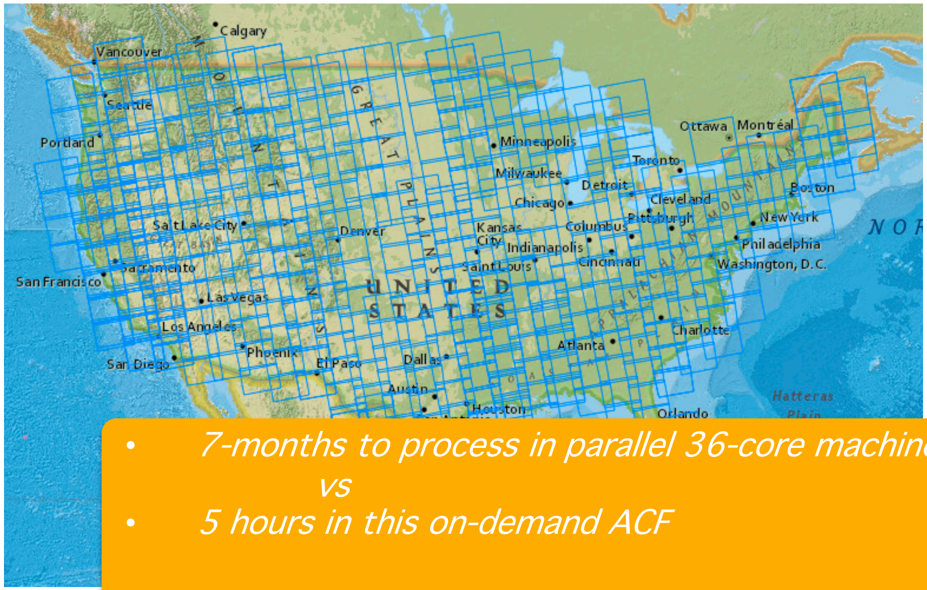
In [ ]:

```
SLC_RE = re.compile(r'(?P<mission>S1\w)_IW_SLC__.*?' +
                    r'_(?P<start_year>\d{4})(?P<start_month>\d{2})(?P<start_day>\d{2})' +
                    r'T(?P<start_hour>\d{2})(?P<start_min>\d{2})(?P<start_sec>\d{2})' +
                    r'_(?P<end_year>\d{4})(?P<end_month>\d{2})(?P<end_day>\d{2})' +
```

	c5d.9xlarge (36 vCPU, 72 GiB)	c5.24xlarge (96 vCPU, 192 GiB)	x1e.2xlarge (8 vCPU, 244 GiB)
1 year (~30 SLCS, 4 bursts)	7 hrs, 24 mins, 46 secs	4 hrs, 38 mins, 33 secs	
2 years (~60 SLCS, 4 bursts)	13 hrs, 37 mins, 39 secs	8 hrs, 16 min, 46 secs	
1.7 years (54 SLCS) Beirut	2.63 hrs (50% HT)		
2.7 years (84 SLCS) Beirut	4.09 hrs (50% HT)		
0.7 years (26 SLCs) Beirut			2.76 hrs (50% HT)

- Coregistration of of SLCs into geocoded stacks
- ARD-like stack as basis of other SAR analysis
  - Damage proxy maps
  - Flood proxy map
  - High resolution displacement time series
- Ported to run in Jupyter notebook and deployable into SDS
- Updates to align with latest ISCE2 open source development
- Benchmarked and optimized performance runs with multi-core parellelization

Example Potential of SAR Analysis at Scale with Notebooks



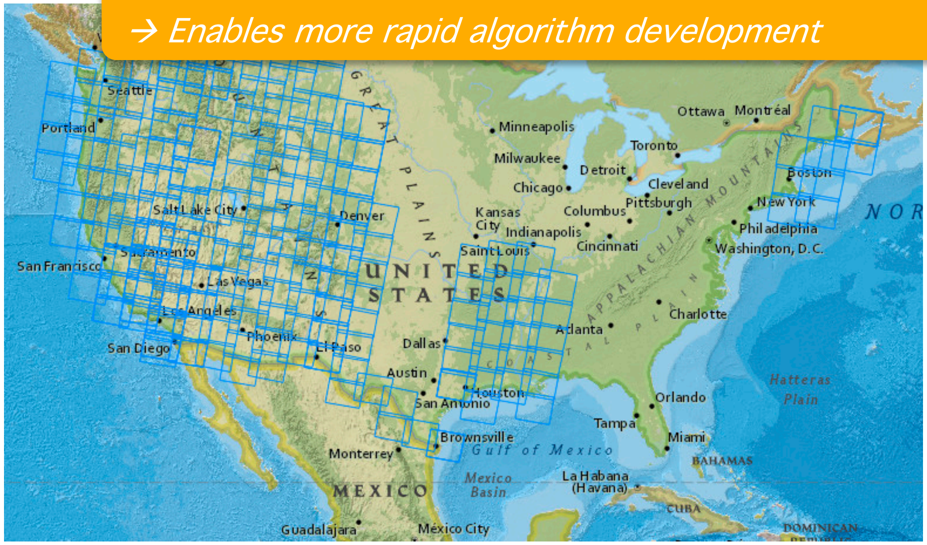
(left) Sentinel-1A/B ascending track over U.S. : ~650 parallel stack processor jobs running at scale

- 7-months to process in parallel 36-core machine
  vs
- 5 hours in this on-demand ACF

→ Enables more rapid algorithm development

- Approach for ARD-like Sentinel-1 SLC stack generation—at scale
  - Decompose each SLC footprint temporal stack generation to be handled by its own Jupyter notebook instance.
  - **Coarse grain parallelization:** scale up parallel SLC stack notebooks to run in parallel in SDS in AWS
  - **Fine graine parallelization:** each notebook leverages multi-core processing
  - Leverage **lower costs AWS spot market instances** for deploying Jupyter notebooks at scale
- Each SLC footprint stack processing is deployed to run at scale in SDS via Containerized Jupyter notebooks

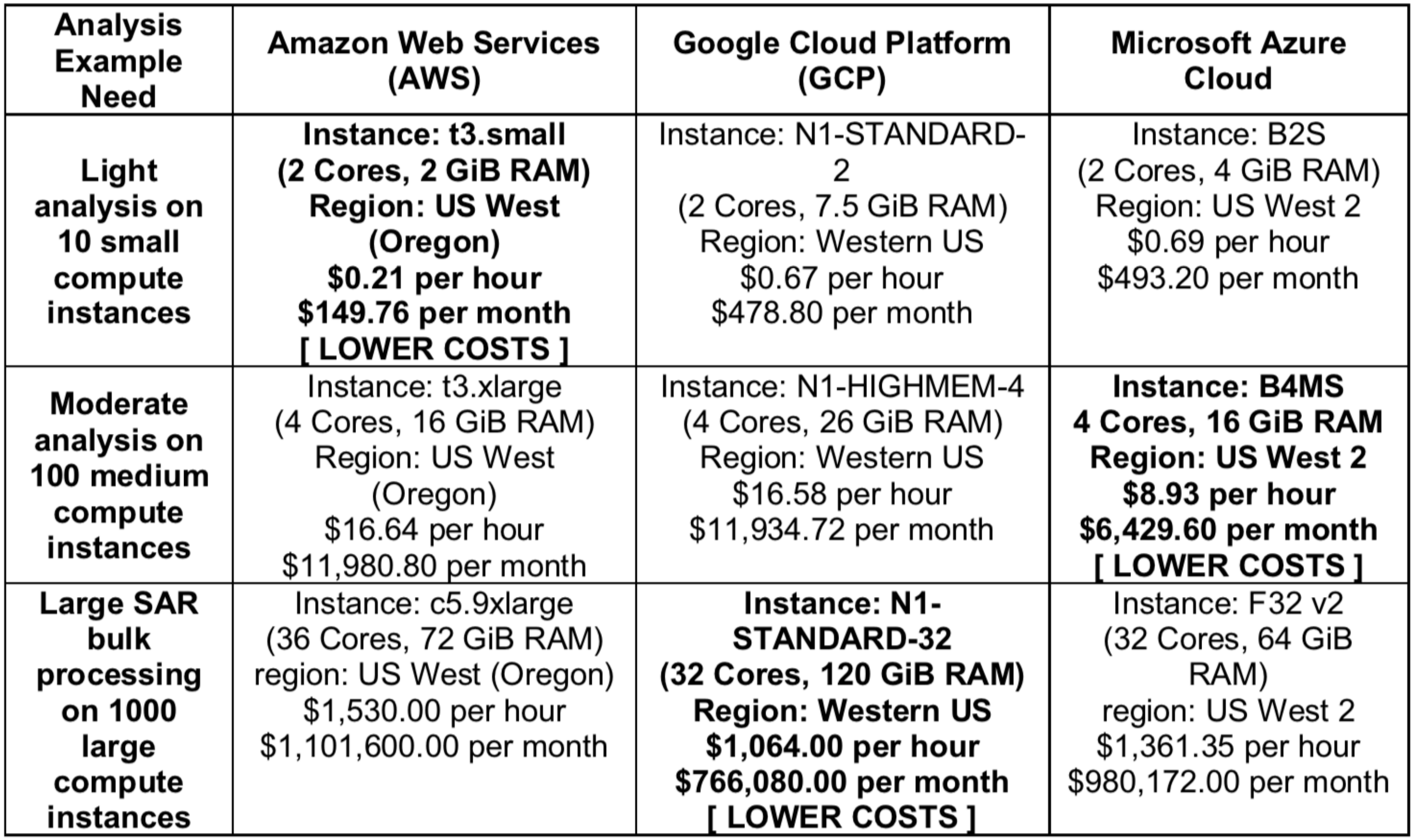
- ** Operational costs of these kinds of large processing jobs are outside the scope of AIST technology demonstration*



Sentinel-1A/B descending track over U.S. : ~426 parallel stack processor jobs running at scale

Addressing SAR Analysis Cloud Costs

- Large compute needs and costs of SDSes in both NISAR and SWOT
- Address vendor lock-in issues
- Early cost analysis shows potential for savings across multi-cloud



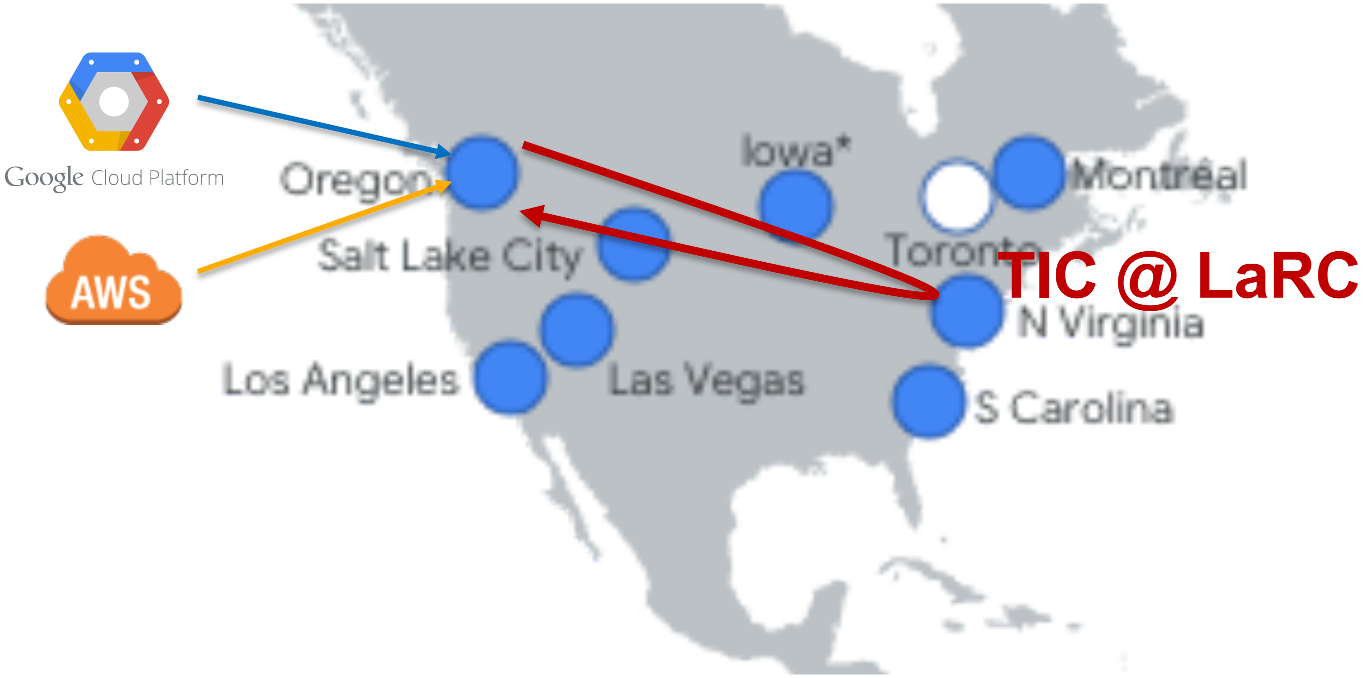
Analysis Example Need	Amazon Web Services (AWS)	Google Cloud Platform (GCP)	Microsoft Azure Cloud
Light analysis on 10 small compute instances	**Instance: t3.small** **(2 Cores, 2 GiB RAM)** **Region: US West (Oregon)** **$0.21 per hour** **$149.76 per month** **[LOWER COSTS]**	Instance: N1-STANDARD-2 (2 Cores, 7.5 GiB RAM) Region: Western US $0.67 per hour $478.80 per month	Instance: B2S (2 Cores, 4 GiB RAM) Region: US West 2 $0.69 per hour $493.20 per month
Moderate analysis on 100 medium compute instances	Instance: t3.xlarge (4 Cores, 16 GiB RAM) Region: US West (Oregon) $16.64 per hour $11,980.80 per month	Instance: N1-HIGHMEM-4 (4 Cores, 26 GiB RAM) Region: Western US $16.58 per hour $11,934.72 per month	**Instance: B4MS** **4 Cores, 16 GiB RAM** **Region: US West 2** **$8.93 per hour** **$6,429.60 per month** **[LOWER COSTS]**
Large SAR bulk processing on 1000 large compute instances	Instance: c5.9xlarge (36 Cores, 72 GiB RAM) region: US West (Oregon) $1,530.00 per hour $1,101,600.00 per month	**Instance: N1-STANDARD-32** **(32 Cores, 120 GiB RAM)** **Region: Western US** **$1,064.00 per hour** **$766,080.00 per month** **[LOWER COSTS]**	Instance: F32 v2 (32 Cores, 64 GiB RAM) region: US West 2 $1,361.35 per hour $980,172.00 per month

Multi-Cloud Onboarding via NASA Managed Cloud Environments (MCE)

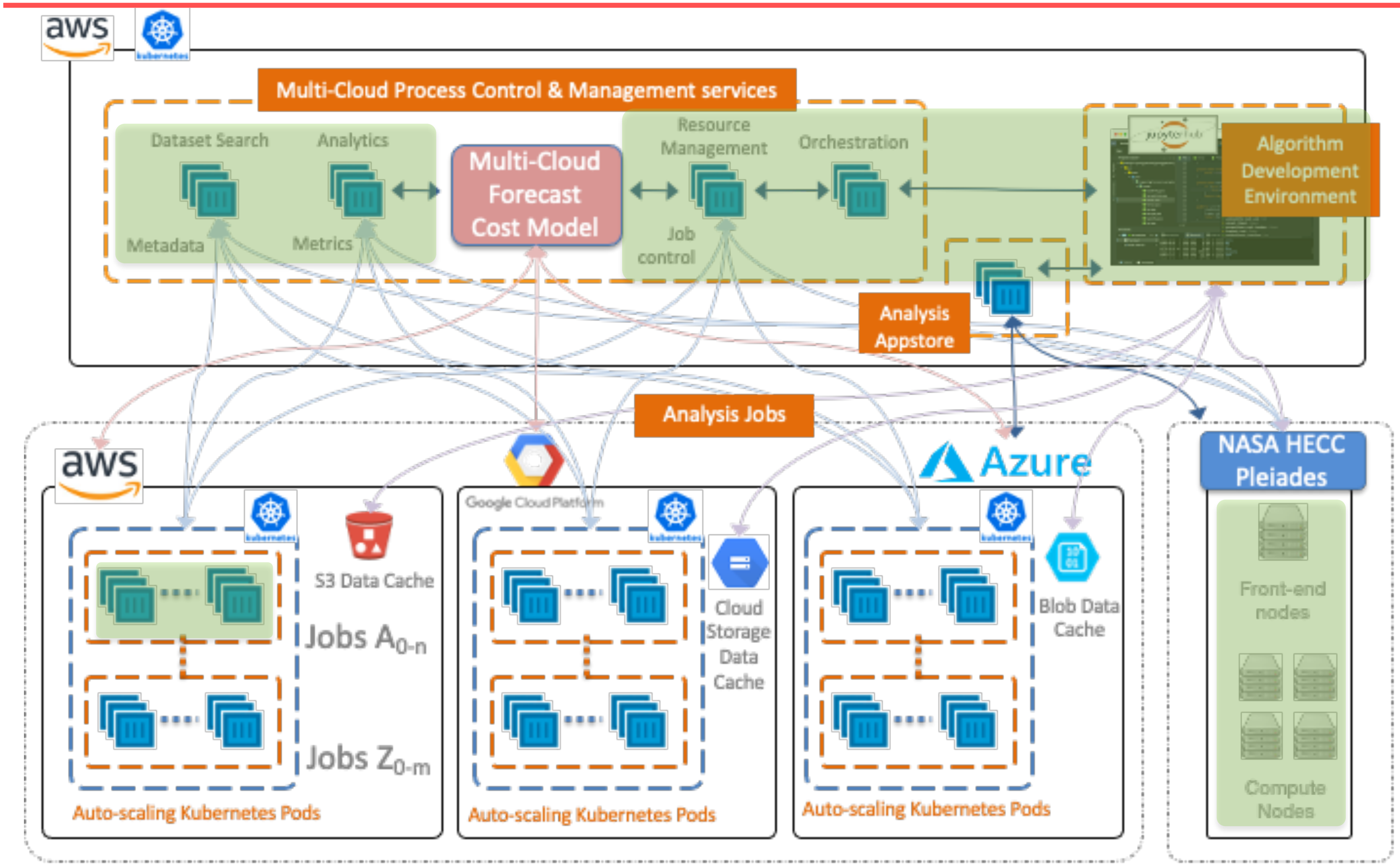
- Access to Google Cloud Platform (GCP), Azure, and AWS for NASA requires going through an MCE
  - FedRAMP
  - Cybersecurity compliance
  - Consolidated accounting, billing
  - EAR and ITAR compliance
- AWS onboarded via JPL’s “MCE”
- AIST SMCE supports AWS.
- MSFC has MCE that has early onboarding of GCP and Azure
  - *MCE in cloud vendor is behind LaRC firewall*
- Identified extraneous data egress and costs via Trusted Internet Connection (TIC)
  - TIC is an OMB/DHS IT security mandate (OMB MEMO M-08-05)
- Currently assessing alternate MCEs to onboard into GCP and Azure

Impact of Trusted Internet Connection (TIC) to Distributed Multi-Cloud Analysis

- TIC Mandate requires all data leaving a federal agency (federal IP address space) to first pass through a TIC for data monitoring and traffic analysis
- NASA deployed TIC architecture:
  - Goddard Space Flight Center (GSFC)
  - Johnson Space Center (JSC)
  - Ames Research Center (ARC)
  - Marshal Space Flight Center (MSFC)
- Traffic from MCE GCP in Oregon region back to AWS Oregon will “trombone” from Oregon to LaRC
- Same cloud region to cloud region data transfer may also incur “tromboning” of data and therefore add full egress costs
- Assessing if can setup compute nodes outside of TIC boundary



End-to-End On-demand Analysis with Multi-Cloud & HEC



Multi-Cloud:
High-priority On-Demand Analysis Jobs

HECC:
Lower-priority Analysis Jobs

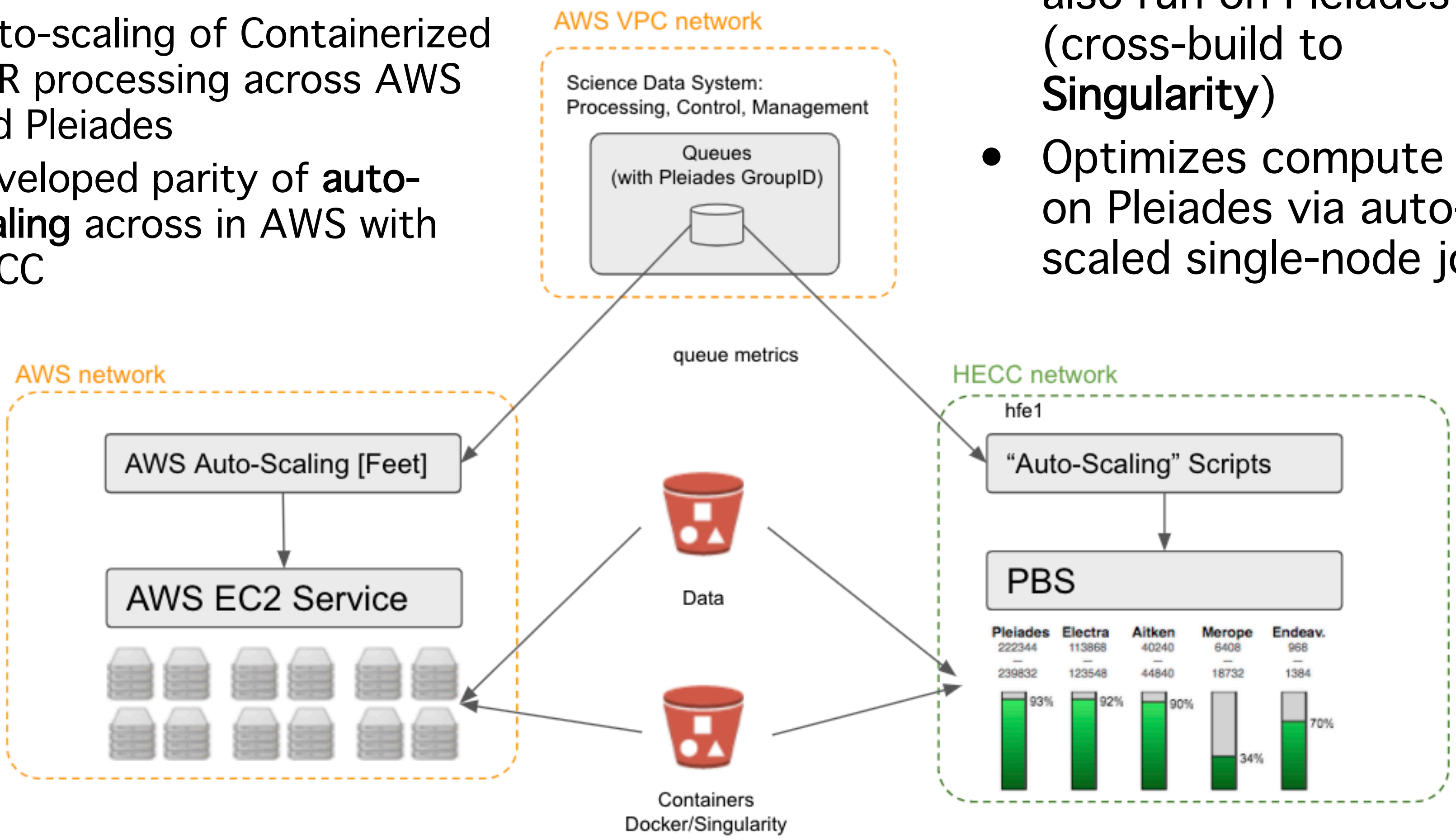
Normalization of Algorithm Deployment Across Multi-cloud + HPC

- Seeking container deployment solution that is:
  - Supported across multiple cloud vendors and NASA's HPC environment (Pleiades)
  - Compatible with Kubernetes by selecting an appropriate container runtime.
- Considered Docker, Podman, Singularity
- Docker is not supported on Pleiades due to security concerns
  - Large user base; native in AWS, Google, Azure. Support in HySDS (used by NISAR, SWOT, SMAP in cloud)
  - **Requires root access** to build and run containers. This violates Pleiades security protocols.
- Podman is a remarkably complete drop in replacement for Docker, with some shortcomings
  - Identical syntax to Docker for common operations
  - Can run in rootless mode required for HPC!
  - But in rootless mode, has a number of shortcomings. Most notably, **lacks rootless support for NFS and parallel filesystems.** This is a major limitation on Pleaides where NFS mounts and Lustre are extensively used.
- Singularity is a good compromise:
  - Portable containers that can be built and run by nonprivileged users.
  - Wide support on HPC systems

Prototyped Auto-scaling Compute Across AWS and NASA Pleiades

- Augmented effort started under ESI funding for ARIA in HEC
- Auto-scaling of Containerized SAR processing across AWS and Pleiades
- Developed parity of **auto-scaling** across in AWS with HECC

- Algorithms deployed to run at scale in AWS can also run on Pleiades (cross-build to **Singularity**)
- Optimizes compute use on Pleiades via auto-scaled single-node jobs

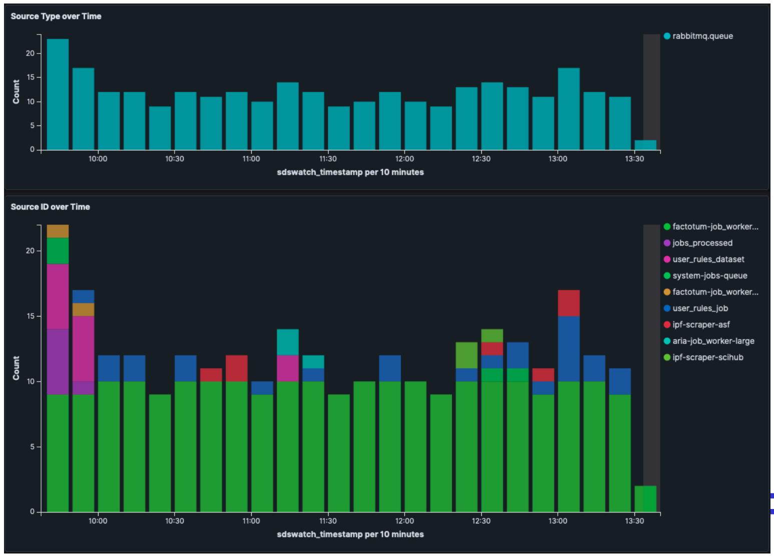


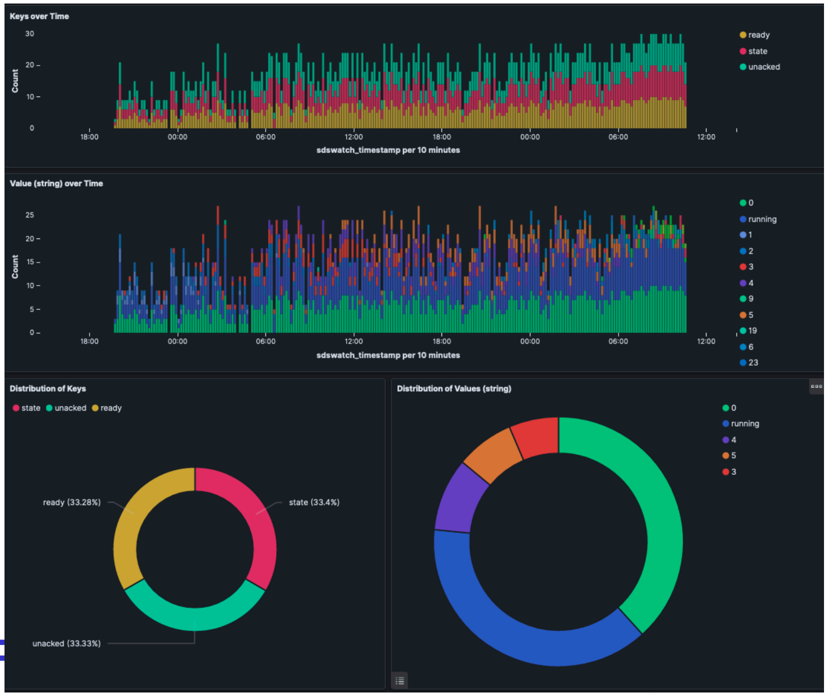
On-demand Analysis Metrics - Towards Machine Learning-based Cost Optimization

- To enable machine learning-based multi-cloud process migration, need to train machine learning model of temporal forecasting of analysis workloads in SDS
- Need for collecting detailed analytics of
  - Notebook processing steps in SDS
  - SDS performance metrics
- Intern (David Tran) worked on SDSWatch tool
  - Collects on-demand processing system metrics
  - Analytics of data system
  - *Metrics as input to ML forecasting for cost estimation*

Example showing distribution over time of jobs-processed and its states

Example showing distribution over time of running jobs over time 9am-1pm PT sampling





Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Summary of Accomplishments

- Setup **SDS** in AWS and on-premise (JPL) using open source HySDS (same system used by **NISAR, SWOT, OCO-2 in cloud, SMAP in cloud**)
- Setup **Jupyter Hub** as the algorithm development environment (ADE) with demonstration notebooks
- Demonstration SAR algorithms in **Jupyter notebooks** for
  - **Sentinel-1 coregistered SLC stacks**
  - **Sentinel-1 GUNW**
- Integrated ADE and SDS for running Jupyter notebooks **on-demand and at scale in SDS**
- Initial Design for on-demand **multi-cloud (AWS, GCP, Azure) and HEC Pleiades**
- **Metrics collection** prototype of on-demand for later use in ML forecasting for cost optimization

Infusion Plans with NISAR and SWOT

- Coordination with SDSes from NISAR and SWOT on this AIST contributing to the on-demand algorithm development and test bed environment
  - For algorithm improvement and data product improvement
  - SWOT
    - Interests in hydrology algorithm development environment
  - NISAR
    - Science teams already started exploring science notebooks for algorithms
    - Cal/Val and ADT
- Similar to this AIST, algorithm development environment (ADE) and processing control and management (PCM) system deployed with NISAR SDS
  - Access to S1-GUNW (Sentinel-1A/B variant of NISAR L2 GUNW standard product)
  - L1 geocoded SLC stacks from Sentinel-1A/B
  - Demonstration of “executable notebooks” running at scale via SDS
- **NISAR’s similar on-demand system (AIST contribution) will be demoed at the next NISAR Science Team meeting in February 2021.**
- Interest in ML-Forecasting-based cloud optimization for lowering costs of on-demand analysis

Future Plans

- Invite beta users to use the on-demand Jupyter environment for testing algorithm development and running in SDS
- Demonstration of ADE+PCM for additional SAR algorithm development at scale
  - Coregistered SLC
  - Mintpy time series
  - ML classification of SAR coregistered SLCs and time series for anomaly event detection
- Continue coordination with NISAR and SWOT
- Coordination with OCIO for on-boarding multi-cloud vendors (Google Cloud Platform and Azure)
- Updates to Containerized Jupyter deployment onto HEC Pleiades for compatibility with multi-cloud
- ML-based forecasting from metrics for multi-cloud cost model

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Publications - List of Acronyms

Publications

Journal / Conference Papers

- IGARSS 2020: abstract accepted for, *'Anomaly Detection and On-Demand Algorithm-Based Analysis Center Framework For Multi-Temporal SAR ARDs'*

Dissertations

- n/a

Other

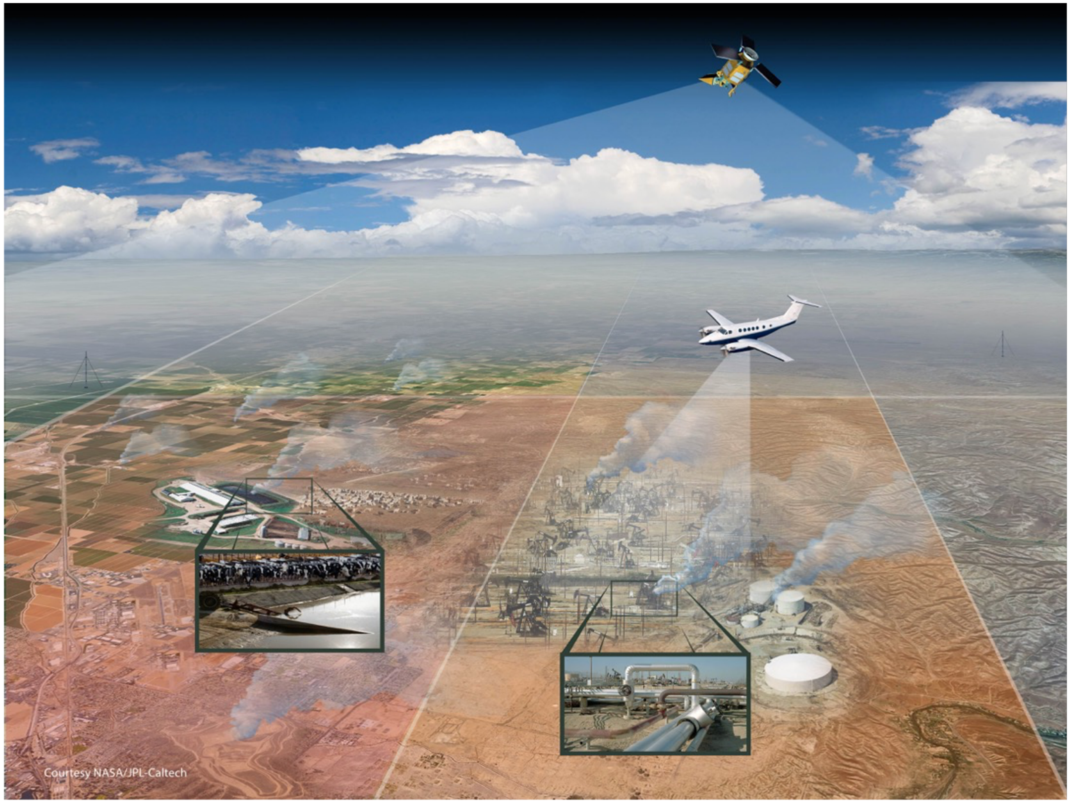
- n/a

Acronyms

List of Acronyms

ADE	Algorithm Development Environment
ADT	Algorithm Development Team
ARD	Analysis Ready Data
AODS	Analysis Optimized Data Services
AWS	Amazon Web Services
DPM	Damage Proxy Map
EONET	Earth Observatory Network Event Tracker
FPM	Flood Proxy Map
GCP	Google Cloud Services
HEC	High End Computing
HPC	High Performance Computing
HySDS	Hybrid Cloud Science Data System
InSAR	Interferometric Synthetic Aperture Radar
PGE	Product Generation Executive
PS time series	Persistent Scatter time series
SAR	Synthetic Aperture Radar
SDS	Science Data System
SLC	Single Looks Complex

Multi-scale Methane Analytic Framework (M2AF)


Courtesy NASA/JPL-Caltech

Riley Duren (PI, University of Arizona/JPL, Caltech)
Natasha Stavros (Exiting PDM, JPL, Caltech)/
Judy Lai (Entering PDM, JPL, Caltech)

AIST-18-0044 2020 Review

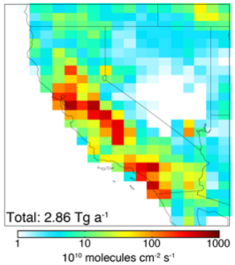
22 Jan 2021

Multi-scale Methane Analytic Framework (M2AF)

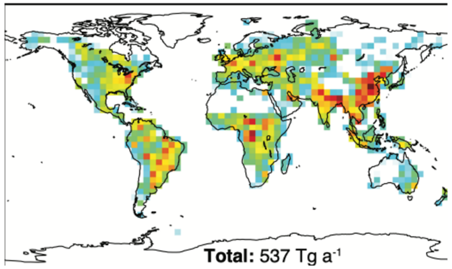
PI: Riley Duren, University of Arizona and Jet Propulsion Laboratory

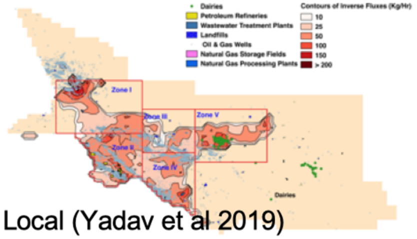
- Develop and mature technologies to support the data discovery, efficient processing, analysis and use of methane data from multiple satellite and airborne observations, surface measurements and modeling systems from global to facility (point source) scales.
- Test and demonstrate system using existing diverse methane data sets for California with stakeholder participation.

Regional (Wecht et al 2014)

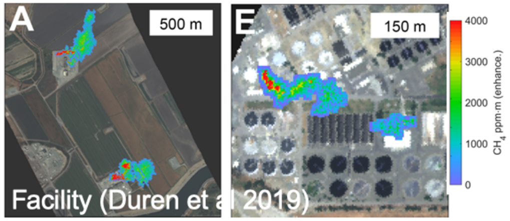


Global (Turner et al 2015)




Dairies
Petroleum Refineries
Wastewater Treatment Plants
Landfills
Oil & Gas Wells
Natural Gas Storage Fields
Natural Gas Processing Plants
Contours of Inverse Fluxes (Kg/Hr)
10
25
50
100
150
> 200
Zone I
Zone II
Zone III
Zone IV
Zone V

Local (Yadav et al 2019)



Facility (Duren et al 2019)

Leverage and extend nascent component capabilities by:

- Optimizing workflow for GEOS-chem flux inversions (global 2 deg/N. America 50 km), enabling annual updates and improved attribution to key emission sectors
- Extending prototype multi-observation local scale flux inversion system (e.g., HRRR 3 km scale) to a more generalized capability for priority regions
- Optimizing workflow for facility scale point-source analysis to reducing latencies from >6 months to 2 weeks
- Integrating the above into a common, searchable system for discovery, fusion and assessment

Co-Is: J. Worden, J. Jacob, D. Cusworth, V. Yadav, A. Thorp, N. Stavros, JPL; D. Jacob, Harvard

Requirements, architecture, design complete	6/2020
System Test 1: local and regional (CA) analytics	12/2020
Deploy workflow for California state-scale analytics	6/2021
System Test 2: N America emissions analytics	12/2021

TRL_{in} = 3 $TRL_{current}$ = 3

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Background / Objectives

- Improving understanding of methane as a major climate forcing agent, is key to the *tracking and characterizing the mechanisms of environmental change* objective in NASA’s Strategic Plan

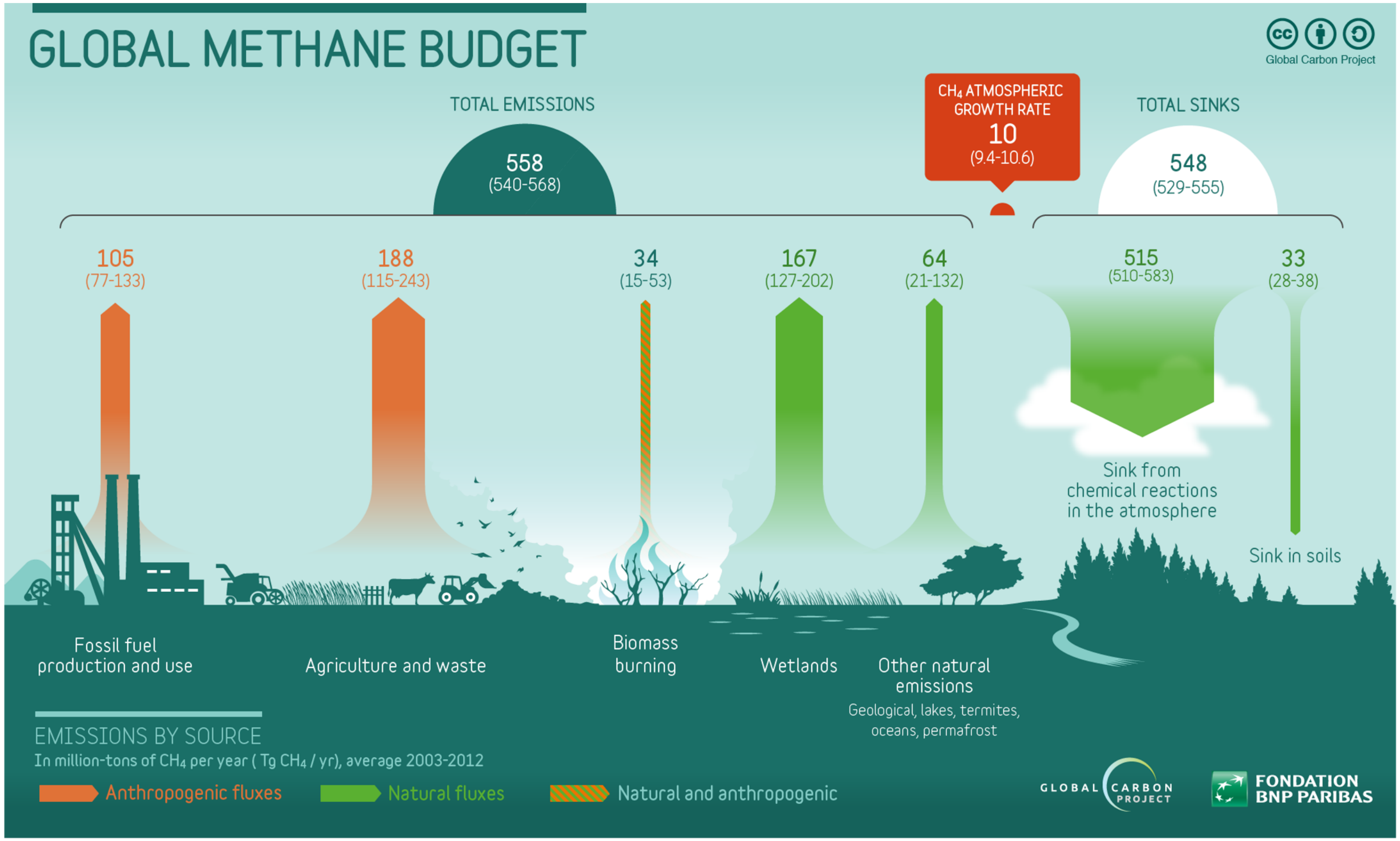
- M^{2}AF **contributes to the US Carbon Cycle Science Plan** objectives:
  - *(Goal-1) provide clear and timely explanation of past and current variations observed in atmospheric CO2 and CH4–and the uncertainties surrounding them*
  - *(Goal-6) address decision maker needs for current and future carbon cycle information and provide data and projections that are relevant, credible, and legitimate*

- M^{2}AF is **responsive to NASA’s Carbon Cycle and Ecosystems** focus by reducing uncertainty in:
  - (*Goal-1) how the global carbon cycle, terrestrial and aquatic ecosystems are changing*
  - (*Goal-3) future changes in global methane cycling as inputs for improved climate change projections*

- M^{2}AF aims to reduce risk, cost, and time for delivering products from current and future Earth Science missions as **highlighted by the 2017 Earth Decadal Survey**:
  - priority for measurements of *methane fluxes and trends at global and regional scales with quantification of point sources and identification of source types* (Earth System Explorer, Greenhouse Gas thrust)

- M^{2}AF is **responsive to NASA Applied Sciences Program** as it is endorsed by public and private sector stakeholders indicating interest and strong potential for infusing the technologies

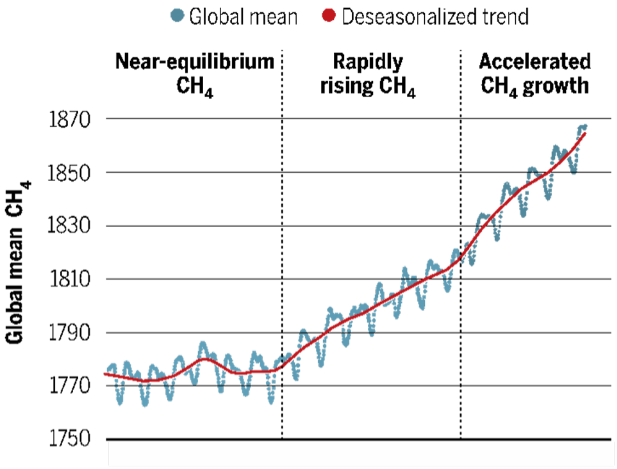
Why methane?



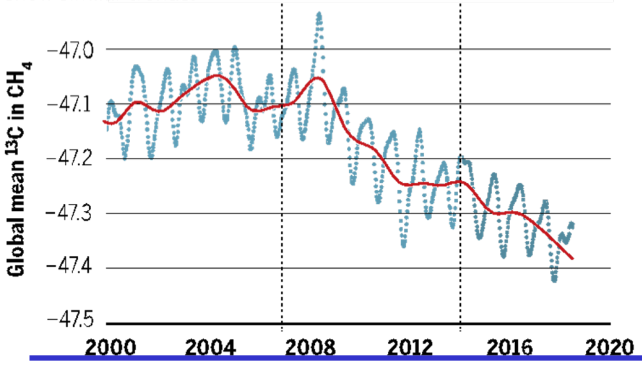
- Methane: #2 anthropogenic climate forcing agent and ozone precursor
- Large uncertainty (50% to unknown) across many scales
- ~ 34X and 86X global warming potential of CO_2 on 100 and 20 yr horizons

Methane growth rate: causes are poorly understood….

Data from U.S. National Oceanic and Atmospheric Administration observing stations show that global mean atmospheric CH_4 started to rise in 2007, with a sharper increase beginning in 2014 (*2*).



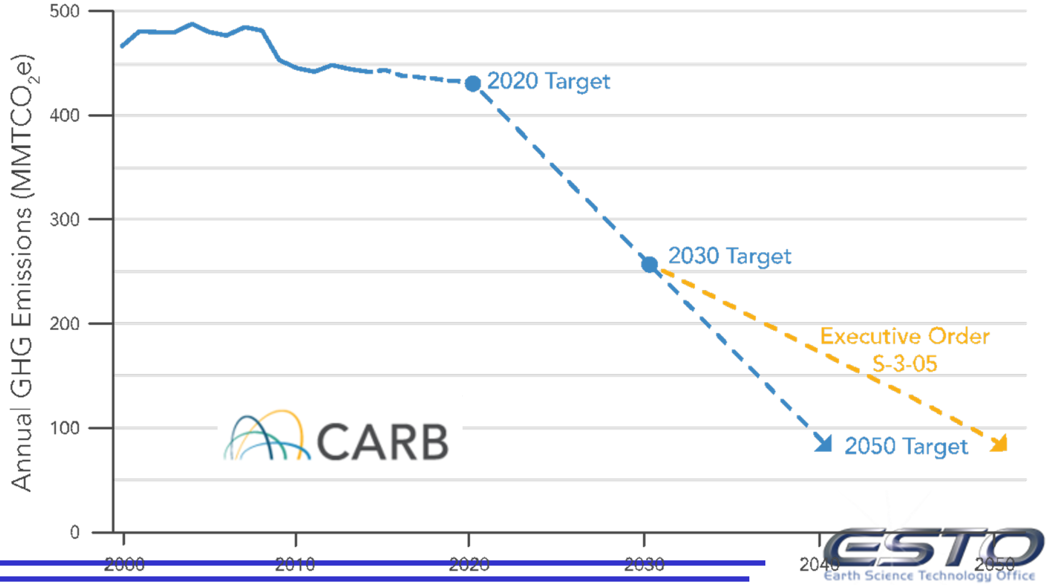
At the same time, the proportion of ^{13}C in CH_4 has been falling, providing insight into possible sources for the additional CH_4. Measurements from other observing station networks show similar trends.



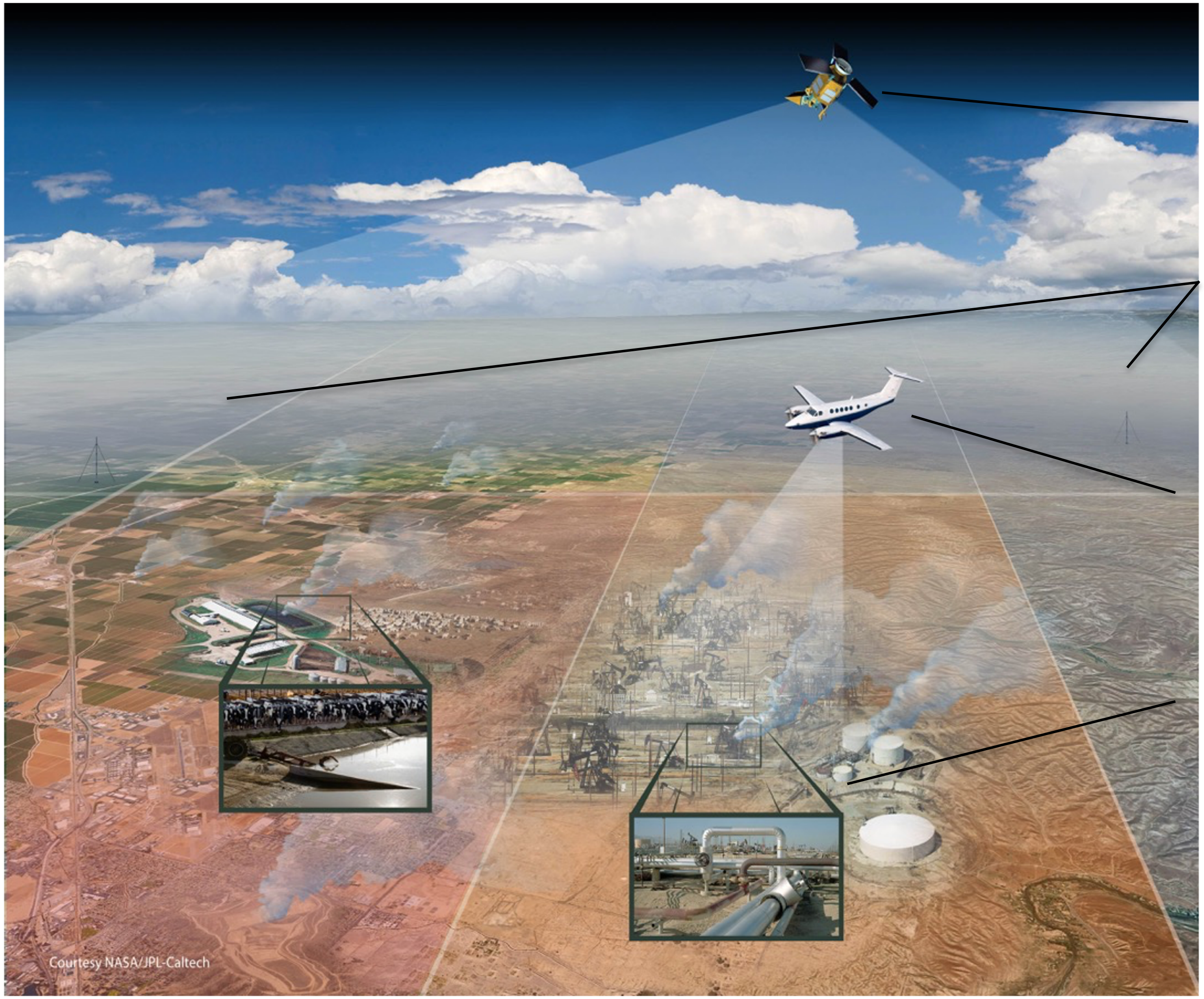
Fletcher and Schaefer, *Science*, 2019

…and currently incompatible with greenhouse gas mitigation goals

California Greenhouse Gas mitigation targets



Tiered Observing & Analysis Strategy



Courtesy NASA/JPL-Caltech

(1) Satellites: Global mappers and point source mappers

(2) Regional & local surface in-situ networks (towers)

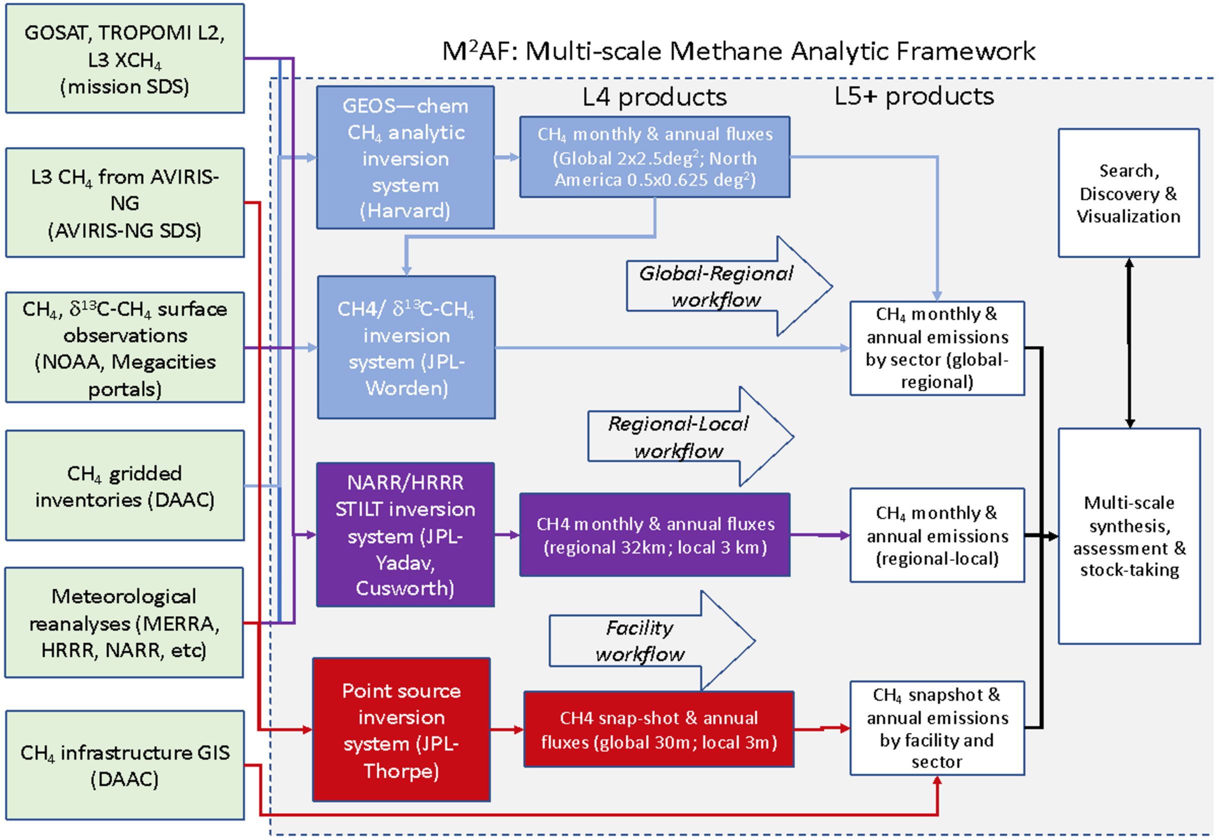
(3) Airborne surveys: Local-regional net fluxes & point-source mappers

(4) On-site and on-road surveys

Specific use-cases drive measurement strategies, spatio-temporal sampling, detection limits, and instrument precision requirements

Objectives / Technical Advance

- Improve component workflows to reduce methane data product (Levels 4 and 5) latency and integrate common core functions
- Create new tools for on-demand analytics including fusion across multiple products and spatial scales
- Improved data search, discovery and visualization capabilities of Methane data



M^2AF: Multi-scale Methane Analytic Framework

GOSAT, TROPOMI L2, L3 XCH_4 (mission SDS)

L3 CH_4 from AVIRIS-NG (AVIRIS-NG SDS)

CH_4, $\delta^{13}C\text{-}CH_4$ surface observations (NOAA, Megacities portals)

CH_4 gridded inventories (DAAC)

Meteorological reanalyses (MERRA, HRRR, NARR, etc)

CH_4 infrastructure GIS (DAAC)

L4 products

L5+ products

GEOS—chem CH_4 analytic inversion system (Harvard)

CH_4 monthly & annual fluxes (Global 2x2.5deg^2; North America 0.5x0.625 deg^2)

CH4/ $\delta^{13}C\text{-}CH_4$ inversion system (JPL-Worden)

Global-Regional workflow

CH_4 monthly & annual emissions by sector (global-regional)

Regional-Local workflow

NARR/HRRR STILT inversion system (JPL-Yadav, Cusworth)

CH4 monthly & annual fluxes (regional 32km; local 3 km)

CH_4 monthly & annual emissions (regional-local)

Facility workflow

Point source inversion system (JPL-Thorpe)

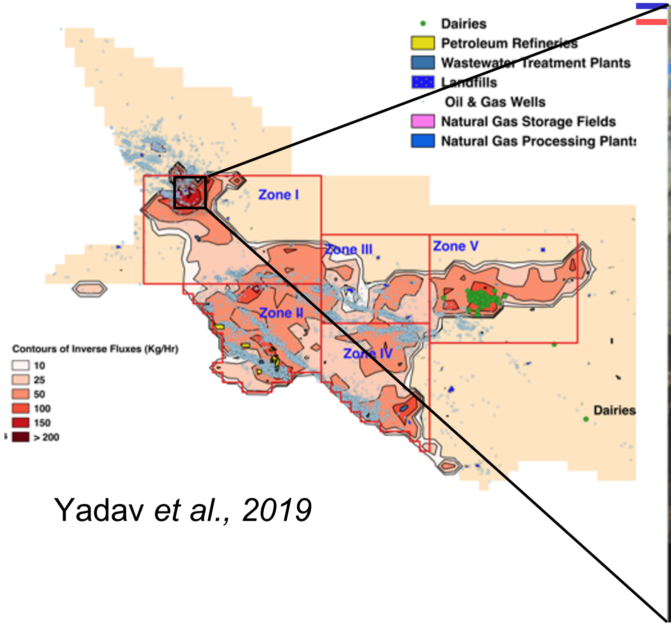
CH4 snap-shot & annual fluxes (global 30m; local 3m)

CH_4 snapshot & annual emissions by facility and sector

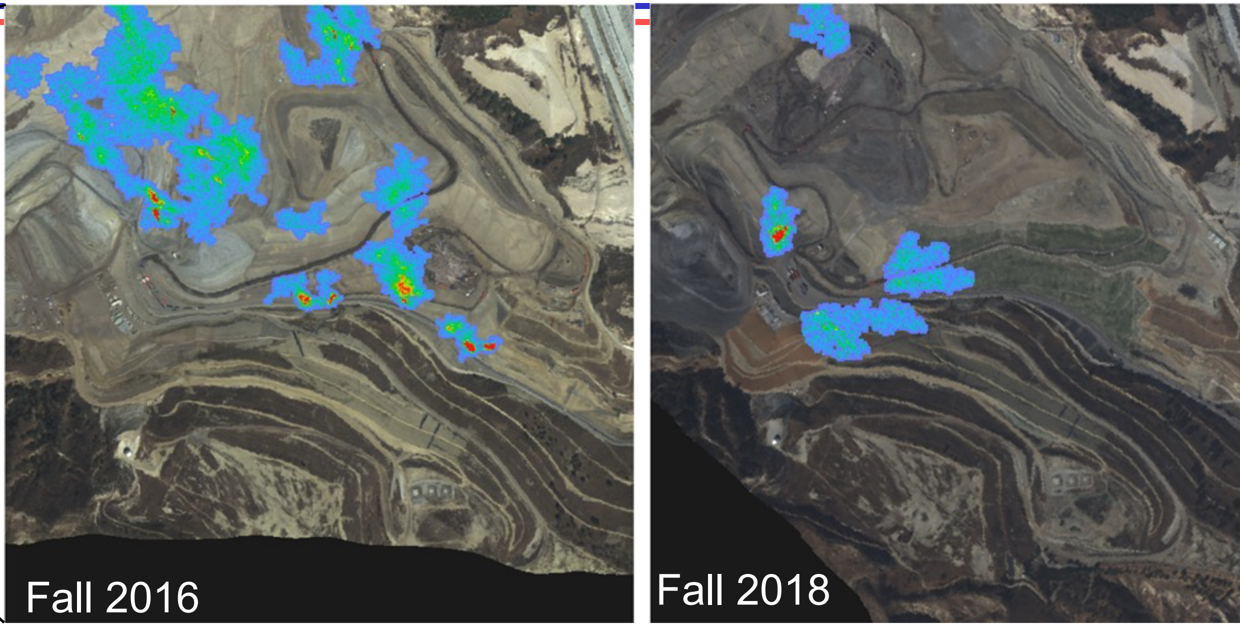
Search, Discovery & Visualization

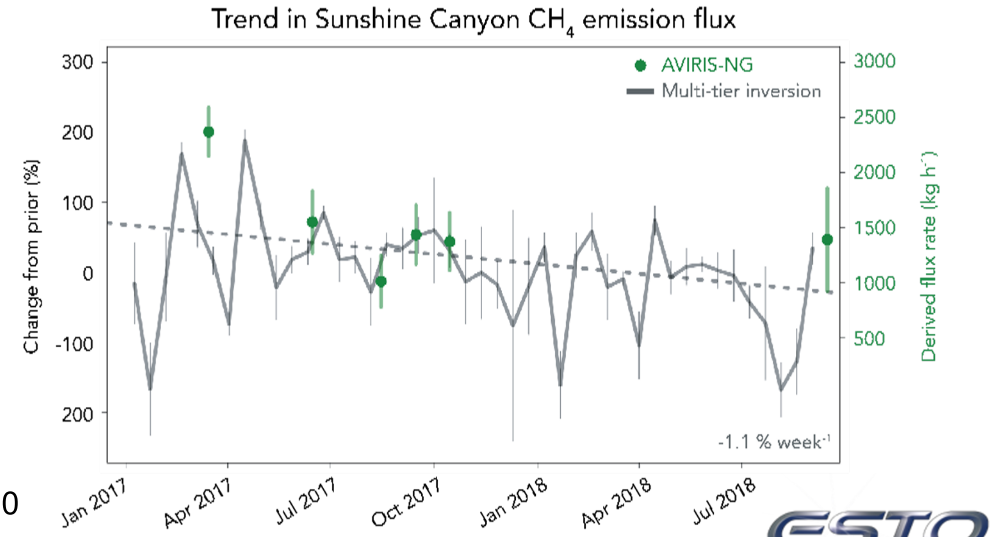
Multi-scale synthesis, assessment & stock-taking

Tiered observing system in action: Landfill emissions mitigation


Dairies
Petroleum Refineries
Wastewater Treatment Plants
Landfills
Oil & Gas Wells
Natural Gas Storage Fields
Natural Gas Processing Plants
Zone I
Zone II
Zone III
Zone IV
Zone V
Dairies
Contours of Inverse Fluxes (Kg/Hr)
10
25
50
100
150
> 200

Yadav *et al., 2019*





Cusworth *et al.*, 2020

Leverage existing Methane Source Finder data portal for on-demand Analytics

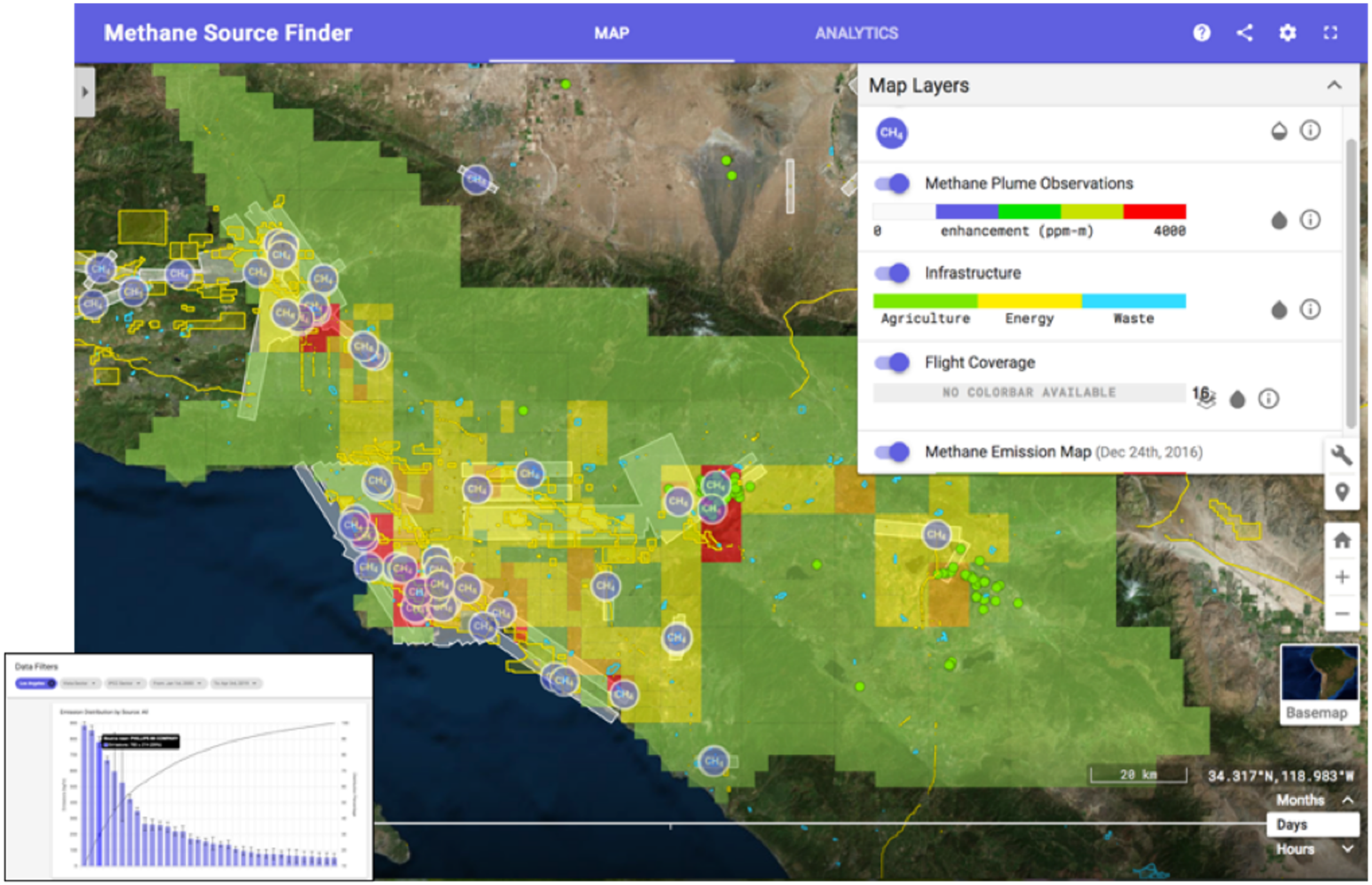


Figure 7: Data portal and user interface for M^{2}AF data search, discovery and visualization that will leverage our previous ACCESS Methane Source Finder project. This example illustrates multi-scale methane data for the Los Angeles basin: a local scale methane emissions map at 3km resolution (red-yellow-green overlay), individual methane point sources with emissions estimates (blue circles), and infrastructure GIS layer as well as metadata and plotting functions. The portal includes hooks for adding regional and global scale methane data products.

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Q3 Update Overview

- Requirements, architecture and interface definitions
- Workflow refinements
  - Global
  - Regional
  - Local
- Cross-scale workflow integration
- Complete Test 1 – TRL advancement

Current State: Global Workflow Refinements

- Streamlining bottom-up inventory workflows and updating data of:
  - 2017 global fuel emissions (expected end of summer), other years ongoing
  - 2017 and 2018 EPA (expected end of summer), other years ongoing
  - Wetlands using a combination of process-based information from recent studies along with more empirical approach involving comparison of our WETCHARTS ensemble models to satellite data

- COMPLETED Top-down fluxes using GOSAT in 2017-18
- Next using TROPOMI in 2019+ for top-down fluxes

Current State: Regional Workflow Refinements

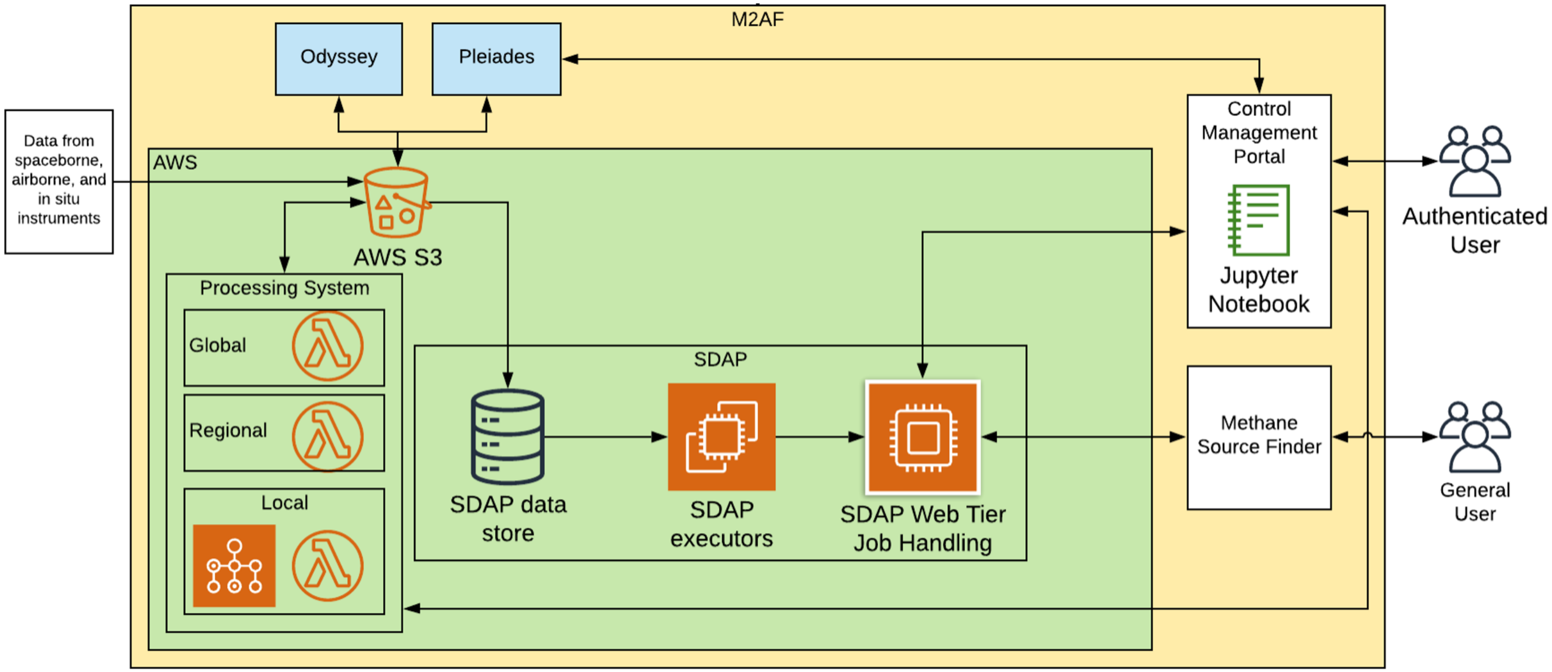
- Regional Workflow Refinements
  - Implemented regional STILT inversion frameworks and deployed in Permian Basin and Los Angeles, in-progress in Central California
  - Working version sector attribution over CONUS; currently scaling up to include entire global domain
  - Looking at pre- and post-COVID inversions Permian Basin, Los Angeles, Central California, among others

Current State: Local Workflow Refinements

- Local Workflow Refinements
  - Demonstrated operational, automated methane data pipeline that can accommodate multiple instruments (AVIRIS-NG/GAO)
    - Latency reduced from months to days
  - Additional workflow testing using recent airborne campaigns over California and Permian to compare post-COVID to pre-COVID previously acquired data
  - Completed verification and validation of data pipeline for point source identification
  - Developing interfaces between current ad-hoc multi-sensor on premise (AVIRIS SDS) and a cloud (AWS) software deployment for seamless multi-sensor integration

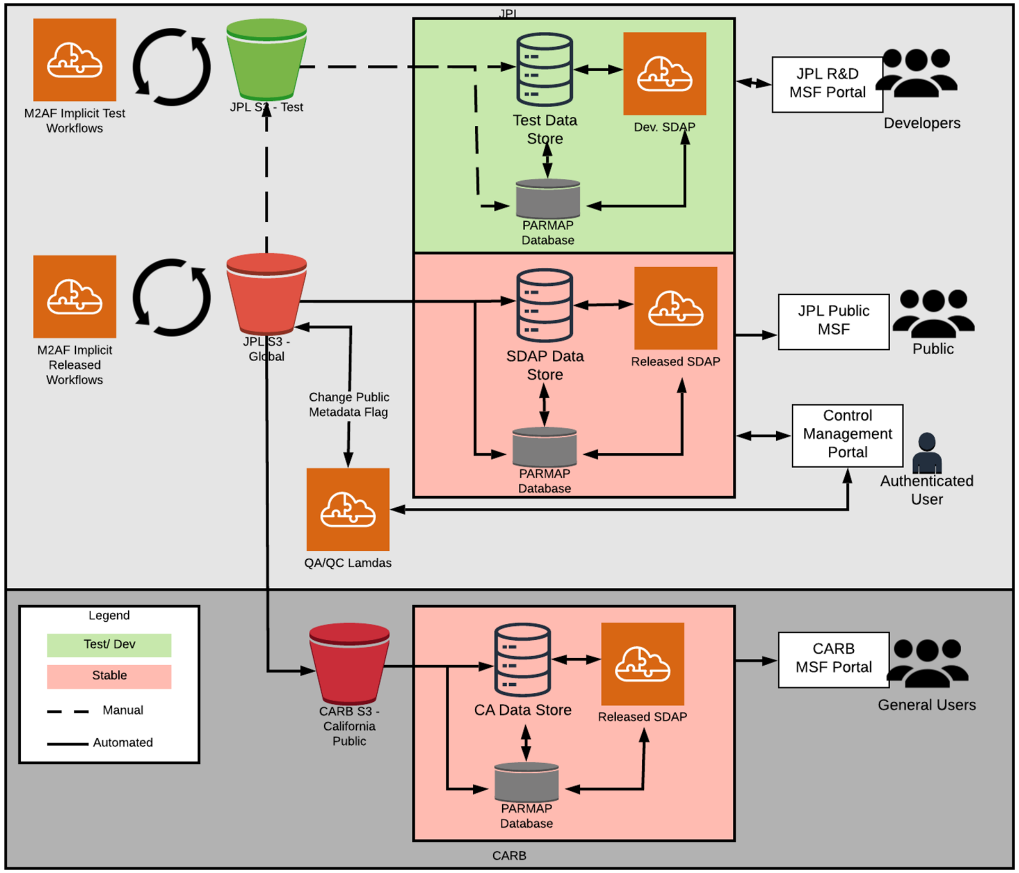
Current State: Cross-Scale Integration

- System Design complete
  - Implicit workflow management system on AWS using lambda and batch
  - Interfaces with two supercomputers
  - SDAP for on-demand analytics
- Development of local workflows in the AWS
- Development of interfaces for streamlined regional workflow deployment on Pleiades
- Two user portals:
  - Public users- Methane Source Finder
  - Authenticated users - Control Management Portal



Current State: Cross-Scale Integration

- Development of interfaces for streamlined regional workflow deployment on Pleiades:
    1. Developer MSF – testing added functionalities without disrupting operations
    2. Control Management Portal (CMP) – for “blessed” collaborators/science team use and looks like MSF with additional tabs at top: 1) submitting a Pleiades job and 2) QA/QC.
    3. CARB MSF – general public MSF updated with on-the-fly analytics



Completed Test 1 and 2

- Test 1: Plumbing Automated Workflows
  - Regional
    - Link Control Management Portal to submit job to Pleiades
    - Run regional forward model STILT on Pleiades
    - AWS inversion run
  - Local - Mostly automated plume list generation in AWS
- Test 2: Local Workflows Tested and SDAP Deployed
  - Wind ingest to SDAP data store; includes readers for file and projection conversion
  - Run end-to-end local plume workflow including batch CNN through extended plume list and source aggregator
  - Reader for extended plume list from new domain to display in MSF
  - Tested developments

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Summary

- After COVID-related impacts and staffing changes, we are now fully staffed
- Good progress on Global, Regional, and Local workflows as well as cross-scale integration
- Test 1 and 2 demonstrated migration of workflows to serverless AWS and on-demand super computer job submission
- Analysis underway for summer airborne campaigns and COVID impact assessment
  - Framework is supporting other R&A program funded tasks

Plan Forward

June - Test 3: Regional Analytics through SDAP displayed in MSF and QA/QC Portal

Task	Task Name	START DATE	END DATE
AIST.T1.01	Sector Emissions Attribution	4/15/2020	2/28/2021
AIST.T6.11	Ingest regional datasets to SDAP for testing	1/15/2021	5/30/2021
AIST.T6.13	CMP job status update implementation	1/15/2021	5/30/2021
AIST.T6.14	JPL CMP integration with JPL Public MSF via authentication	1/15/2021	5/30/2021
AIST.T6.12	QA/QC Portal integration to CPM	1/15/2021	12/31/2021
AIST.T5.04	Add support (imaging, query & analysis) for regional datasets	1/15/2021	5/30/2021
AIST.T1.02	Streamline annual bottom-up inventory generation	4/15/2020	5/30/2021
AIST.T6.11	Deploy to AWS and test (version 3): Deploy workflow for state-scale analytics	6/1/2021	6/30/2021

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Publications - List of Acronyms

Publications/Conferences/Meetings

Date	Category	What	Publisher	Who
Feb 2020	Paper	Fast and accurate retrieval of methane concentrations from imaging spectrometer data using sparsity prior	IEEE Transactions on Geoscience and Remote Sensing	Foote, Dennison, Thorpe, et al.
Mar 2020	Paper	Synthesis of methane observations across scales: Strategies for deploying a multi-tiered observing network	GRL	Cusworth, Duren, Thorpe, Yadav
Mar 2020	Paper	Using remote sensing to detect, validate, and quantify methane emissions from California solid waste operations	ERL	Cusworth, Duren, Thorpe
Oct 2020	Paper	Attribution of the accelerating increase in atmospheric methane during 2010–2018 by inverse analysis of GOSAT observations	ACP	Daniel Jacob

Date	Category	What	Presentation	Location	Who
June 3, 2020	Conference	16th international workshop on greenhouse gas measurements from space		Darmstadt, Germany	Cusworth/Thorpe
June 23, 2020	Meeting	NASA's 17th annual Earth Science Technology Forum	Multi-Scale Methane Analytic Framework	Virtual	Stavros
May 4, 2020	Meeting	KISS COVID-19 Virtual Study		Virtual	Cusworth
Dec 2020	Conference	AGU Fall Meeting	15 posters/presentations	Virtual	All

Acronyms

List of Acronyms

AVIRIS-ng	Airborne Visible Infrared Imaging Spectrometer Next Generation
CH4	Methane
DAAC	Data Active Archive Center
GEOS	Geostationary Operational Environment Satellite
GOSAT	Greenhouse Gases Observing Satellite
HEC	High End Computing
HRRR-STILT	High-Resolution Rapid Refresh - Stochastic Time-Inverted Lagrangian Transport
IDS	NASA Inter-Disciplinary Science Program
M2AF	Multi-scale Methane Analytic Framework
MERRA	Modern-Era Retrospective analysis for Research and Applications
MSF	Methane Source Finder
NARR	North American Regional Reanalysis
SDS	Science Data System
TROPOMI	TROPOspheric Monitoring Instrument

Mining Chained Modules in Analytic Center Framework

Jia Zhang (PI, Southern Methodist University)
Seungwon Lee (Co-I, JPL)
Ramakrishna Nemani (Co-I, Ames)
Alex Goodman (Co-I, JPL)
Benyang Tang (Co-I, JPL)

AIST-18-0059 Annual Review
01/22/2021

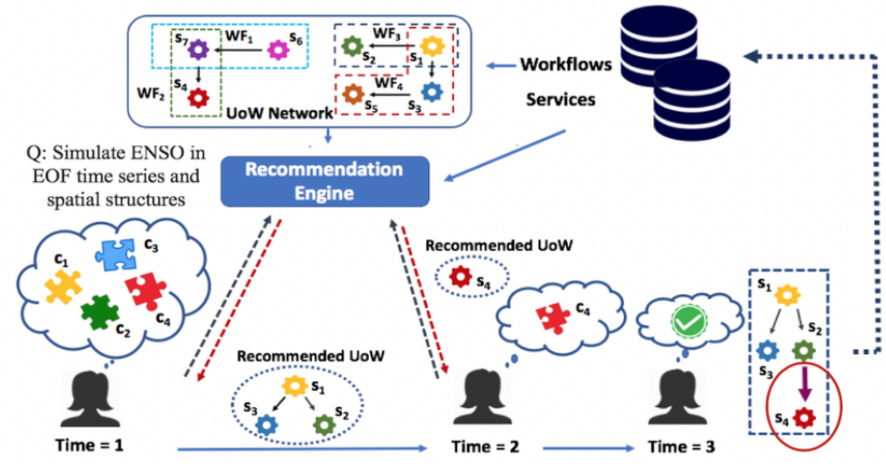
Mining Chained Modules in Analytic Center Framework

PI: Jia Zhang, Southern Methodist University

Objective

Build a workflow tool as a building block for ACF, capable of recommending chained software modules throughout an Earth science data analytics workflow design.

- Develop algorithms to mine software usage history and construct a knowledge network;
- Develop algorithms to explore reusable software module chains from knowledge network;
- Develop an intelligent service that provides personalized recommend-as-you-go support to help users design workflow.



Approach

- Develop Climate Model Diagnostic Analyzer (CMDA) workflows using Jupyter Notebook;
- Develop techniques to parse Jupyter notebooks to extract service usage dependencies ;
- Develop techniques to construct a knowledge network to store and retrieve mined knowledge;
- Develop techniques to identify and extract reusable service chain snippets cross workflows.
- Develop reference templates to guide Earth scientists to design workflows suing Jupyter Notebook;

Co-Is: Seungwon Lee, JPL: Ramakrishna Nemani, Ames; Alex Goodman, Benyang Tang, JPL

Key Milestones

Phase 1: CMDA Jupyter notebook examples	04/20
Phase 1: Algorithms to analyze CMDA notebooks	06/20
Phase 1: Network analysis algorithms	08/20
Phase 1: Workflow recommendation system	09/20
Phase 1: User test; CMDA notebooks	10/20
Phase 2: Notebook templates; refined notebook analysis	01/21
Phase 2: Notebook templates; Enhanced workflow tool	07/21
Phase 2: JPL Summer School	09/21
Phase 2: User testing and documentation	10/21

$TRL_{in} = 3$ $TRL_{current} = 3$

Team Members

Jia Zhang (PI, Professor, Southern Methodist University)

Seungwo Lee (Co-I, JPL)

Alex Goodman (Co-I, JPL)

Ramakrishna Nemani (Co-I, Ames)

Benyang Tang (Co-I, JPL)

Kyle Pearson (Co-I, JPL)

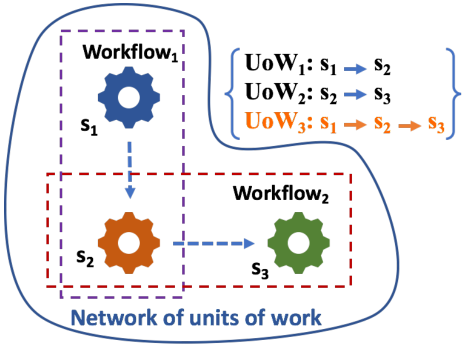
Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Background and Objectives

- NASA is building Analytic Center Framework (ACF) as a collaboration platform for community users to harmonize existing tools, data and computing environments.
  - In the next 5-10 years, it can be anticipated many data analytics tools and models will be published onto NASA ACF as reusable modules.
- A large number of software modules will make it difficult for Earth scientists to choose from.
  - How to help Earth scientists find suitable software modules at ACF from a sea of available candidates and use them productively?
- This AIST project targets for the next 5-10-year timeframe, aiming to develop a unique and important building block for ACF:

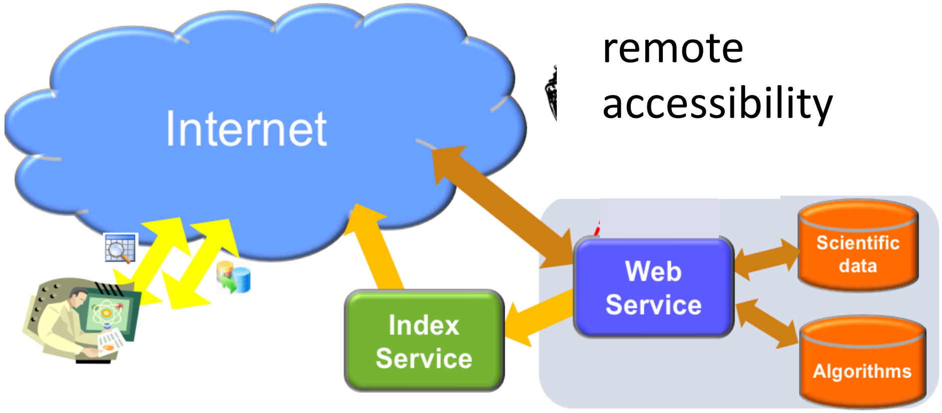
a workflow tool capable of recommending chained software snippets when a geoscientist designs a data analytics workflow



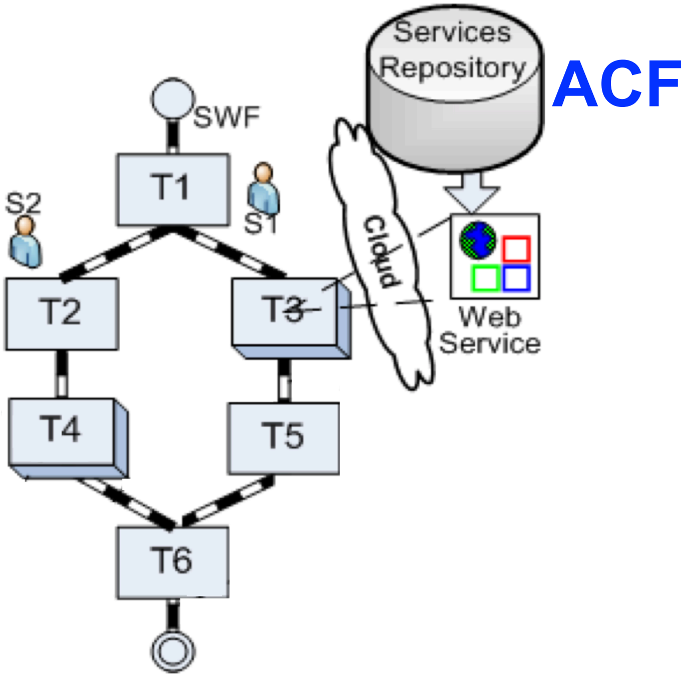
Workflow_1
s_1
UoW_1: $s_1 \rightarrow s_2$
UoW_2: $s_2 \rightarrow s_3$
UoW_3: $s_1 \rightarrow s_2 \rightarrow s_3$
Workflow_2
s_2
s_3
Network of units of work

Service Oriented Science

- Scientists expose data and computational algorithms as remotely accessible web services



Application Programming Interface (API)



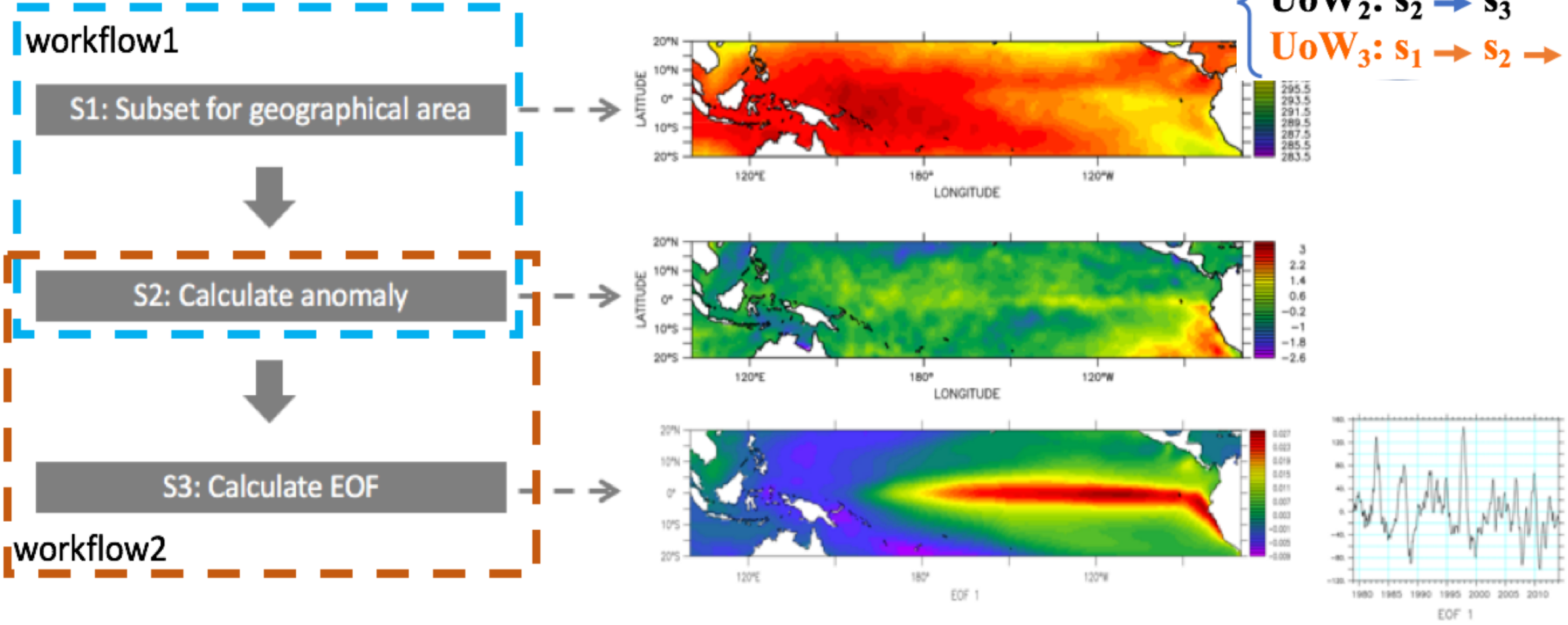
Service reuse can help scientists focus on science in data analytics procedure (workflow)

Intelligent Service Oriented Workflow Recommendation

[Foster-Science'05]

Unit of Work

$$\left\{\begin{array}{l} UoW_1: s_1 \rightarrow s_2 \\ UoW_2: s_2 \rightarrow s_3 \\ UoW_3: s_1 \rightarrow s_2 \rightarrow s_3 \end{array}\right\}$$



Mine service usage history (workflow provenance) and identify reusable, and maybe unprecedented, service chain snippets (**UoW**) to facilitate automatic data analytics workflow development.

Presentation Contents

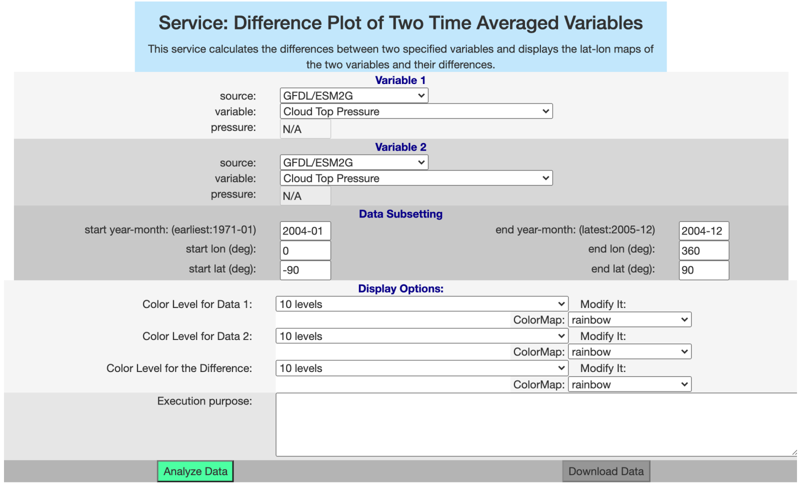
- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Task 1: Develop CMDA workflows using Jupyter Notebook

- We created a Jupyter Notebook using CMDA webservices and Python function calls.
- The Jupyter Notebook calls a CMDA service as a HTTP GET request.
- The Jupyter Notebook provides an interactive input configuration for the CMDA service call.
- The Jupyter Notebook provides an interactive output plotting for the CMDA output data.
- The Jupyter Notebook prepares the CMDA service output data as Xarray Dataset object.
- Further analysis steps are implemented in Python function calls.
- Each Jupyter Notebook provides a scientific workflow representing one or more CMDA service calls and other Python function calls.

Task 1: Develop CMDA workflows using Jupyter Notebook

CMDA Service Web Interface



Service: Difference Plot of Two Time Averaged Variables

This service calculates the differences between two specified variables and displays the lat-lon maps of the two variables and their differences.

Variable 1

source: GFDL/ESM2G
variable: Cloud Top Pressure
pressure: N/A

Variable 2

source: GFDL/ESM2G
variable: Cloud Top Pressure
pressure: N/A

Data Subsetting

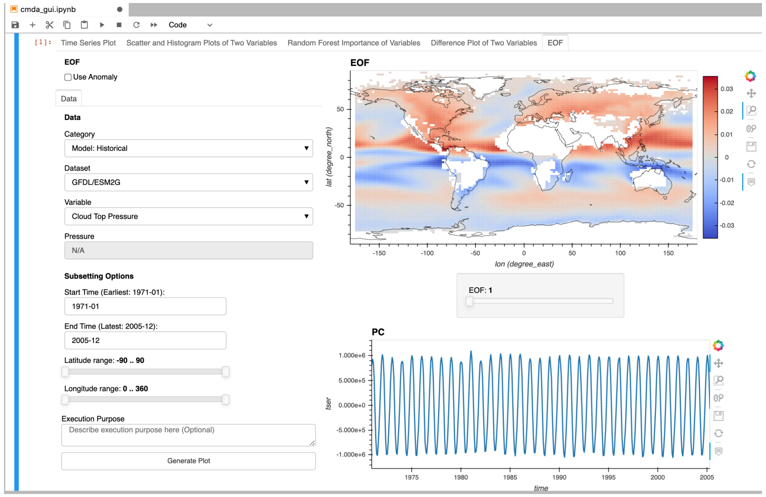
start year-month: (earliest:1971-01) 2004-01
end year-month: (latest:2005-12) 2004-12
start lon (deg): 0
end lon (deg): 360
start lat (deg): -90
end lat (deg): 90

Display Options:

Color Level for Data 1: 10 levels Modify It: ColorMap: rainbow
Color Level for Data 2: 10 levels Modify It: ColorMap: rainbow
Color Level for the Difference: 10 levels Modify It: ColorMap: rainbow

Execution purpose:

CMDA Service Jupyter Notebook Interface



cmda_gui.ipynb

[1]:

EOF

□ Use Anomaly

Data

Category: Model: Historical
Dataset: GFDL/ESM2G
Variable: Cloud Top Pressure
Pressure: N/A

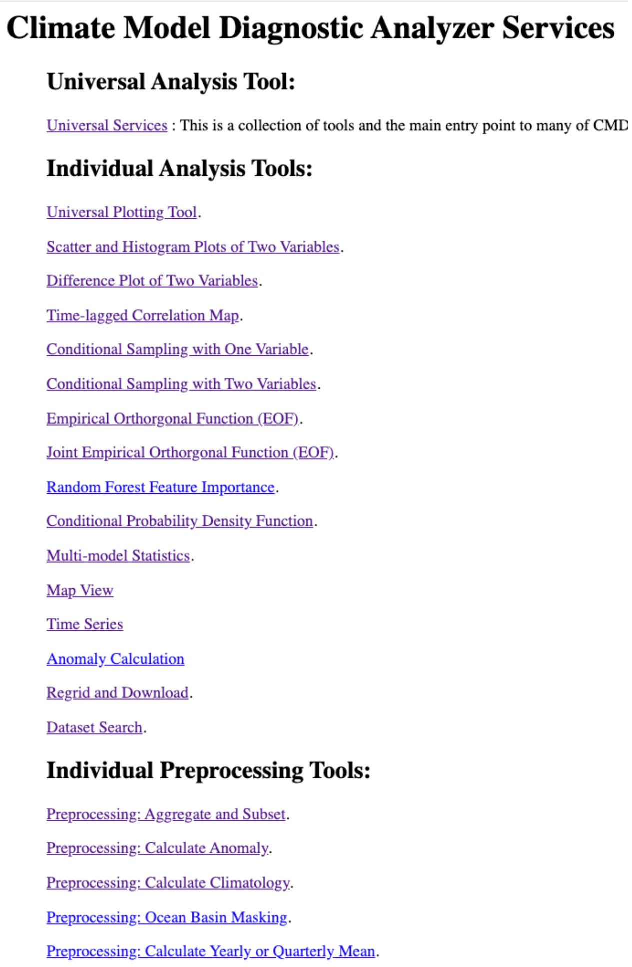
Subsetting Options

Start Time (Earliest: 1971-01): 1971-01
End Time (Latest: 2005-12): 2005-12
Latitude range: **-90 .. 90**
Longitude range: **0 .. 360**

Execution Purpose: Describe execution purpose here (Optional)

EOF: **1**

Task 1: Develop CMDA workflows using Jupyter Notebook



Climate Model Diagnostic Analyzer Services

Universal Analysis Tool:

Universal Services : This is a collection of tools and the main entry point to many of CMD

Individual Analysis Tools:

Universal Plotting Tool.

Scatter and Histogram Plots of Two Variables.

Difference Plot of Two Variables.

Time-lagged Correlation Map.

Conditional Sampling with One Variable.

Conditional Sampling with Two Variables.

Empirical Orthorgonal Function (EOF).

Joint Empirical Orthorgonal Function (EOF).

Random Forest Feature Importance.

Conditional Probability Density Function.

Multi-model Statistics.

Map View

Time Series

Anomaly Calculation

Regrid and Download.

Dataset Search.

Individual Preprocessing Tools:

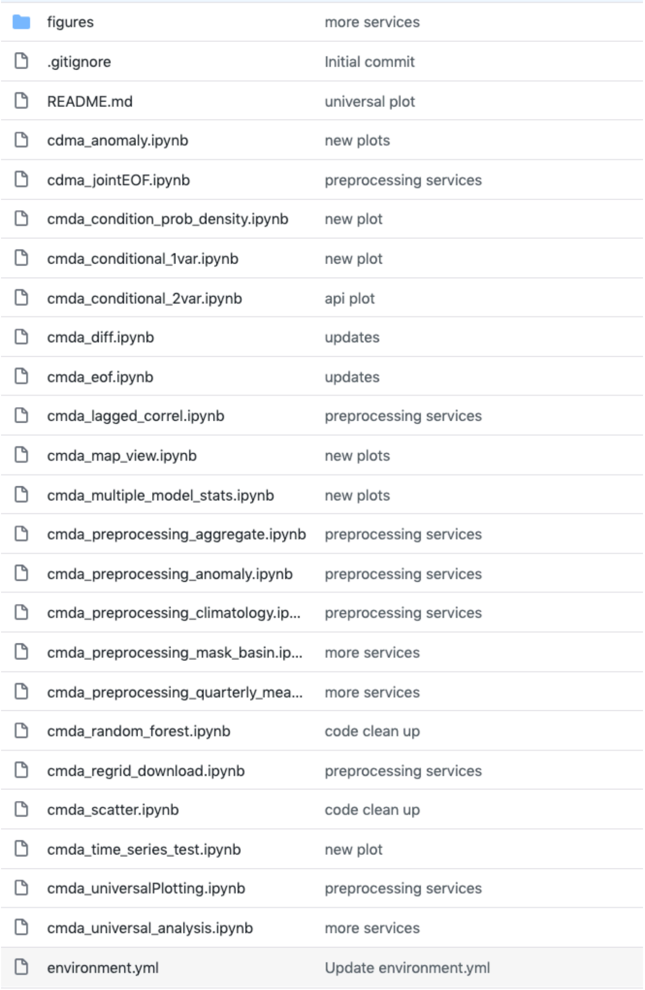
Preprocessing: Aggregate and Subset.

Preprocessing: Calculate Anomaly.

Preprocessing: Calculate Climatology.

Preprocessing: Ocean Basin Masking.

Preprocessing: Calculate Yearly or Quarterly Mean.



📁	figures	more services
🗋	.gitignore	Initial commit
🗋	README.md	universal plot
🗋	cdma_anomaly.ipynb	new plots
🗋	cdma_jointEOF.ipynb	preprocessing services
🗋	cmda_condition_prob_density.ipynb	new plot
🗋	cmda_conditional_1var.ipynb	new plot
🗋	cmda_conditional_2var.ipynb	api plot
🗋	cmda_diff.ipynb	updates
🗋	cmda_eof.ipynb	updates
🗋	cmda_lagged_correl.ipynb	preprocessing services
🗋	cmda_map_view.ipynb	new plots
🗋	cmda_multiple_model_stats.ipynb	new plots
🗋	cmda_preprocessing_aggregate.ipynb	preprocessing services
🗋	cmda_preprocessing_anomaly.ipynb	preprocessing services
🗋	cmda_preprocessing_climatology.ip...	preprocessing services
🗋	cmda_preprocessing_mask_basin.ip...	more services
🗋	cmda_preprocessing_quarterly_mea...	more services
🗋	cmda_random_forest.ipynb	code clean up
🗋	cmda_regrid_download.ipynb	preprocessing services
🗋	cmda_scatter.ipynb	code clean up
🗋	cmda_time_series_test.ipynb	new plot
🗋	cmda_universalPlotting.ipynb	preprocessing services
🗋	cmda_universal_analysis.ipynb	more services
🗋	environment.yml	Update environment.yml

Task 1: Develop CMDA workflows using Jupyter Notebook

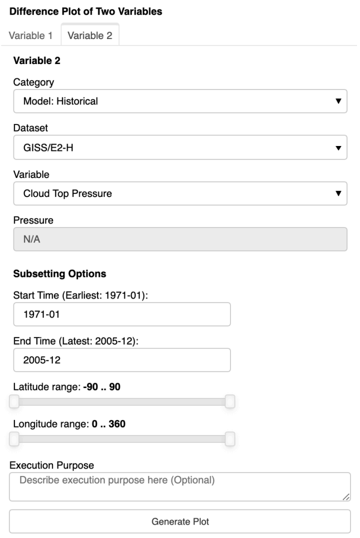
Example of One-Step CMDA Workflow in Jupyter Notebook

1. Interactive input configuration

2. REST API call to CMDA service

3. Output data download and processing

4. Interactive output visualization



Difference Plot of Two Variables

Variable 2

Category: Model: Historical ▼

Dataset: GISS/E2-H ▼

Variable: Cloud Top Pressure ▼

Pressure: N/A

Subsetting Options

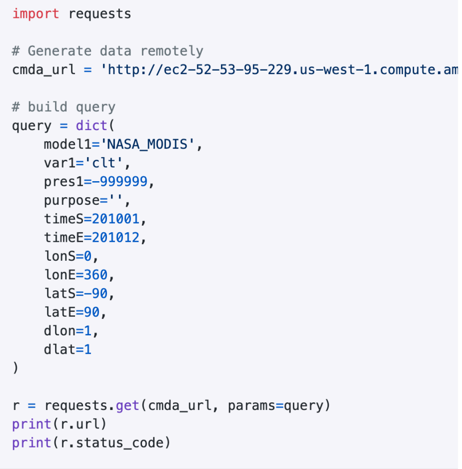
Start Time (Earliest: 1971-01): 1971-01

End Time (Latest: 2005-12): 2005-12

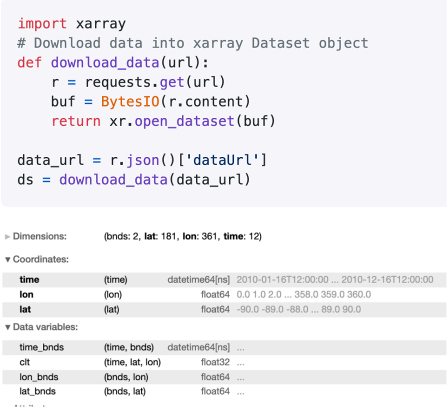
Latitude range: **-90 .. 90**

Longitude range: **0 .. 360**

Execution Purpose: Describe execution purpose here (Optional)

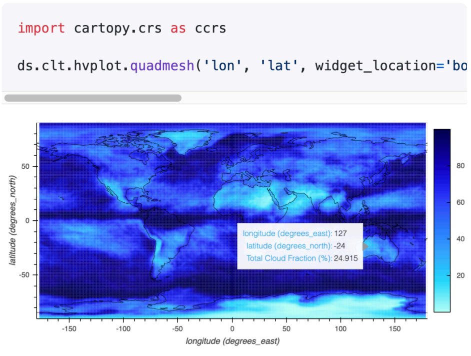


```python
import requests

# Generate data remotely
cmda_url = 'http://ec2-52-53-95-229.us-west-1.compute.am

# build query
query = dict(
    model1='NASA_MODIS',
    var1='clt',
    pres1=-999999,
    purpose='',
    timeS=201001,
    timeE=201012,
    lonS=0,
    lonE=360,
    latS=-90,
    latE=90,
    dlon=1,
    dlat=1
)

r = requests.get(cmda_url, params=query)
print(r.url)
print(r.status_code)
```



```python
import xarray
# Download data into xarray Dataset object
def download_data(url):
    r = requests.get(url)
    buf = BytesIO(r.content)
    return xr.open_dataset(buf)

data_url = r.json()['dataUrl']
ds = download_data(data_url)
```

```
▸ Dimensions:      (bnds: 2, lat: 181, lon: 361, time: 12)
▾ Coordinates:
    time           (time)             datetime64[ns]  2010-01-16T12:00:00 ... 2010-12-16T12:00:00
    lon            (lon)                     float64  0.0 1.0 2.0 ... 358.0 359.0 360.0
    lat            (lat)                     float64  -90.0 -89.0 -88.0 ... 89.0 90.0
▾ Data variables:
    time_bnds      (time, bnds)       datetime64[ns]  ...
    clt            (time, lat, lon)          float32  ...
    lon_bnds       (bnds, lon)               float64  ...
    lat_bnds       (bnds, lat)               float64  ...
```



```python
import cartopy.crs as ccrs

ds.clt.hvplot.quadmesh('lon', 'lat', widget_location='bo
```

longitude (degrees_east): 127
latitude (degrees_north): -24
Total Cloud Fraction (%): 24.915

The Jupyter notebooks contain interactive plots with bokeh

Task 1: Develop CMDA workflows using Jupyter Notebook

Example 1 of Multi-Step Scientific Workflow

Question: Calculate the global net radiative flux imbalance at Top of Atmosphere (TOA).

1. Calculate the time-averaged radiation fluxes.

```
from cmda import ServiceViewer
import numpy as np
import panel as pn
pn.extension()
app = ServiceViewer()
rsdt = app.open_url('http://api.jpl-cmda.org/svc/mapView? model1=NASA_CERES&var1=rsdt& ...)
rsut = app.open_url('http://api.jpl-cmda.org/svc/mapView? model1=NASA_CERES&var1=rsdt& ...)
rlut = app.open_url('http://api.jpl-cmda.org/svc/mapView? model1=NASA_CERES&var1=rlut& ...)
```

2. Calculate the net radiative flux.

```
rad_net = rsdt.rsdt - rsut.rsut - rlut.rlut
```

```
xarray.DataArray
latitude: 180
longitude: 360
array([[ -39.3181 , -39.3181 , -39.3181 , ..., -39.3181 , -39.3181 , -39.3181 ], ….
```

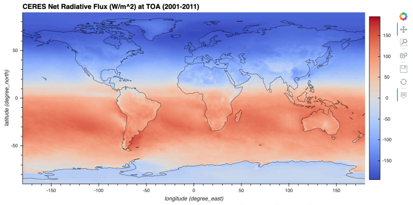
3. Calculate the space-averaged net radiative flux.

```
rad_net_space_averaged = rad_net.weighted(np.cos(np.deg2rad(rad_net.latitude))).mean(('longitude', 'latitude'))
```

```
xarray.DataArray
array(8.01082829)
Coordinates: (0)
Attributes: (0)
```

4. Interactive output visualization

```
import cartopy.crs as ccrs
rad_net.hvplot.quadmesh('longitude', 'latitude',
                        title='CERES Net Radiative Flux (W/m^2) at TOA (2001-2011)', geo=True,
                        projection=ccrs.PlateCarree(),
                        crs=ccrs.PlateCarree(), coastline=True,
                        width=800, rasterize=True)
```



Task 1: Develop CMDA workflows using Jupyter Notebook

Example 2 of Multi-Step Scientific Workflow

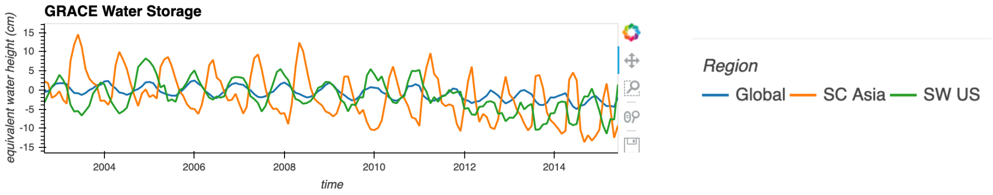
Question: Investigate the seasonality of GRACE land water storage in comparison with AIR surface air temperature and TRIMM precipitation.

1. Calculate the time series of GRACE water storage in 3 regions.

```
from cmda import ServiceViewer
import numpy as np
import panel as pn
pn.extension()
app = ServiceViewer()
grace_global = app.open_url('http://api.jpl-cmda.org/svc/timeSeries? model1=NASA_GRACEvlatS1=-90&vlatE1=90&vlonS1=0&vlonE1=360)
grace_sc_asia = app.open_url('http://api.jpl-cmda.org/svc/timeSeries? model1=NASA_GRACE&vlatS1=23&vlatE1=35&vlonS1=66&vlonE1=96)
grace_sw_us = app.open_url('http://api.jpl-cmda.org/svc/timeSeries? model1=NASA_GRACE&vlatS1=31&vlatE1=42&vlonS1=236&vlonE1=258)
grace = xr.concat([grace_global, grace_sc_asia, grace_sw_us], dim='Region').assign_coords(Region=['Global', 'SC Asia', 'SW US']).squeeze()
```

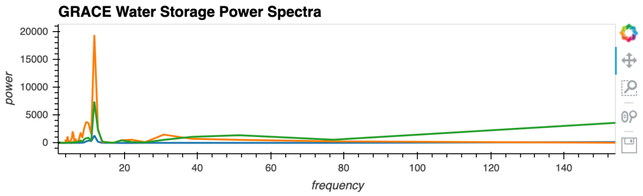
2. Plot the GRACE time series to see general patterns.

```
grace.hvplot(x='time', y='variable', by='Region', title='GRACE Water Storage', legend='bottom')
```



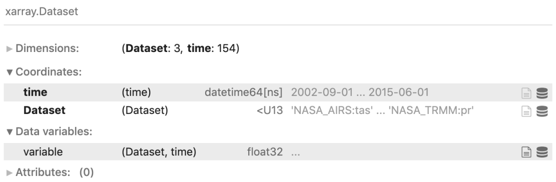
3. Calculate the power spectrum of the GRACE time series.

```
f, p = signal.periodogram(grace.variable, 1/12, detrend='linear’)
f[f == 0] = np.nan
grace['power'] = ('Region', 'frequency’),
p grace = grace.assign_coords(frequency=(1/(12*f)))
grace.power.hvplot(by='Region', title='GRACE Water Storage Power
Spectra', legend='bottom')
```



4. Calculate the time series of AIRS temperature and TRIMM precipitation in South Central Asia.

```
ds_sca = app.open_url('http://api.jpl-
cmda.org/svc/timeSeries?purpose=&timeS=200209&timeE=201506&
model1=NASA_AIRS&var1=tas&pres1=-
999999&vlatS1=23&vlatE1=35&vlonS1=66&vlonE1=96&model2=NA
SA_GRACE&var2=zl&pres2=-
999999&vlatS2=23&vlatE2=35&vlonS2=66&vlonE2=96&model3=NA
SA_TRMM&var3=pr&pres3=-
999999&vlatS3=23&vlatE3=35&vlonS3=66&vlonE3=96&nVar=3')
```



```
xarray.Dataset

▶ Dimensions:      (Dataset: 3, time: 154)
▼ Coordinates:
    time           (time)            datetime64[ns]  2002-09-01 ... 2015-06-01
    Dataset        (Dataset)                   <U13  'NASA_AIRS:tas' ... 'NASA_TRMM:pr'
▼ Data variables:
    variable       (Dataset, time)          float32  ...
▶ Attributes:   (0)
```

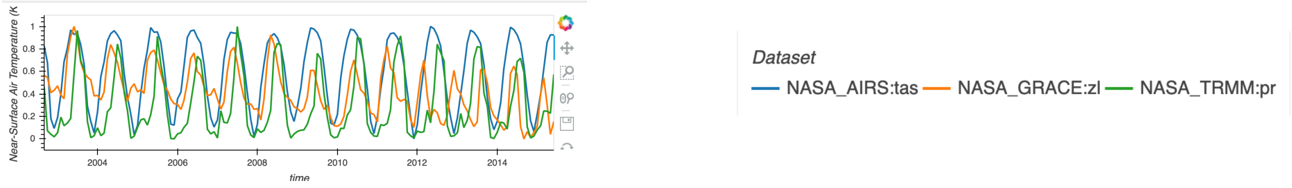
Task 1: Develop CMDA workflows using Jupyter Notebook

Example 2 of Multi-Step Scientific Workflow

Question: Investigate the seasonality of GRACE land water storage in comparison with AIR surface air temperature and TRIMM precipitation.

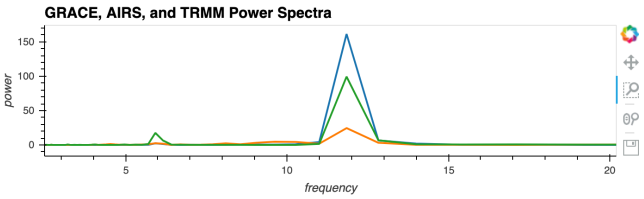
5. Plot the GRACE, AIRS, TRIMM time series to see general patterns.

```
ds_sca.hvplot(x='time', y='variable', by='Dataset', legend='bottom')
```



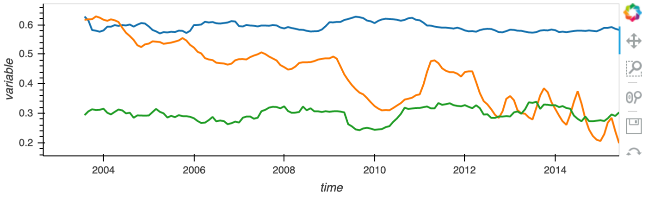
6. Calculate the power spectrum of the time series.

```
f, p = signal.periodogram(ds_sca.variable, 1/12, detrend='linear')
f[f == 0] = np.nan
ds_sca['power'] = ('Dataset', 'frequency'), p
ds_sca = ds_sca.assign_coords(frequency=(1/(12*f)))
ds_sca.power.hvplot(by='Dataset', title='GRACE, AIRS, and TRMM
Power Spectra', legend='bottom')
```



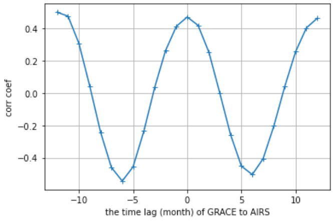
7. Calculate the interannual variability of the three parameters.

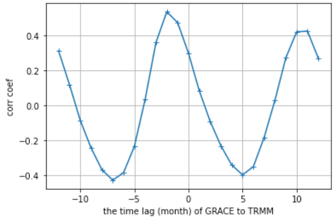
```
ds_sca.rolling(time=12).mean().hvplot(x='time', y='variable',
by='Dataset', legend='bottom')
```



8. Calculate the time-lagged correlation between GRACE and TRMM/AIRS.

```
lags = np.arange(-12,13)
corr9 = np.zeros((len(lags),))
count = -1
for lag in lags:
  count += 1
  if lag>0: grace1 = GRACE_ts[lag:]; airs1 = AIRS_ts[:-lag]
  elif lag==0: grace1 = GRACE_ts;  airs1 = AIRS_ts
  else: lag2 = -lag; airs1 = AIRS_ts[lag2:]; grace1 = GRACE_ts[:-lag2]
  corr9[count] = np.corrcoef(grace1,airs1)[0,1]
```





Task 1: Develop CMDA workflows using Jupyter Notebook

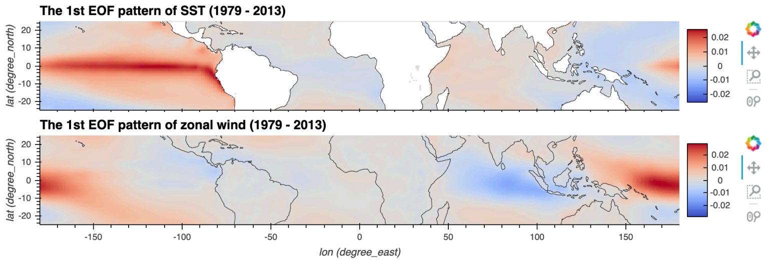
Example 3 of Multi-Step Scientific Workflow

Question: EOF and time-correlation analysis of the tropical zonal wind and sea surface temperature

1. Calculate the EOF of ECMWF zonal wind and sea surface temperature anomalies.

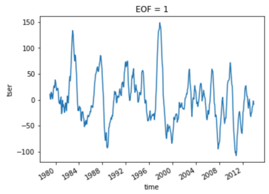
```
from cmda import ServiceViewer
import numpy as np
import panel as pn
pn.extension()
app = ServiceViewer()
east_wind = app.open_url('http://api.jpl-cmda.org/svc/EOF?model1=ECMWF_interim&var1=uas&pres1=-999999&purpose=&lonS=-
180&lonE=180&latS=-25&latE=25&timeS=197901&timeE=201312&anomaly=1') sst = app.open_url('http://api.jpl-
cmda.org/svc/EOF?model1=ECMWF_interim&var1=tos&pres1=-999999&purpose=&lonS=-180&lonE=180&latS=-
25&latE=25&timeS=197901&timeE=201312&anomaly=1')
```

2. Plot the first EOF spatial patterns.

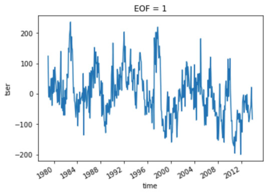

The 1st EOF pattern of SST (1979 - 2013)
The 1st EOF pattern of zonal wind (1979 - 2013)
lat (degree_north)
20
10
0
-10
-20
0.02
0.01
0
-0.01
-0.02
-150
-100
-50
0
50
100
150
lon (degree_east)

3. Plot the first EOF time series.

EOF time series of SST

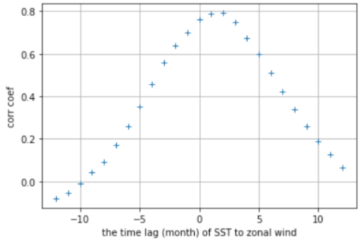


EOF time series of Zonal Wind



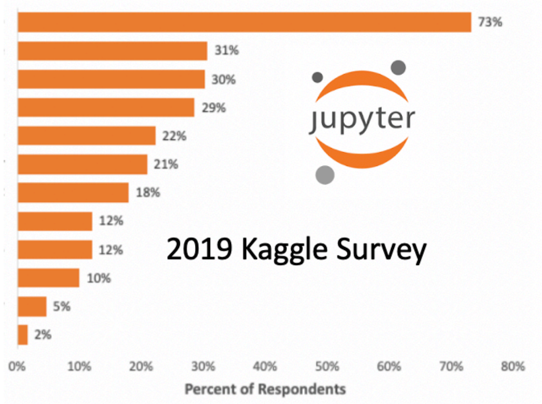
4. Calculate the time-lagged correlation of SST to zonal wind.

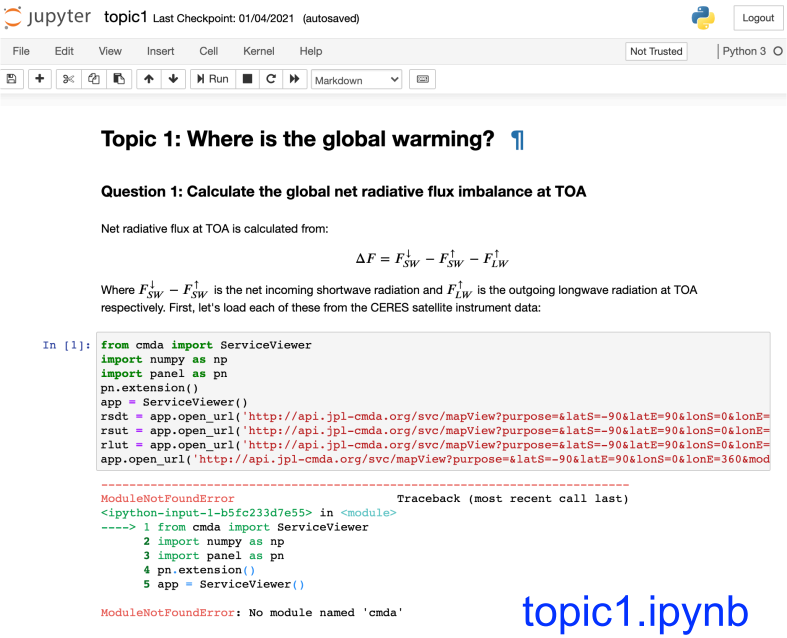
```
lags = np.arange(-12,13)
corr9 = np.zeros((len(lags),))
count = -1
ww = east_wind.tser[0].values
ss = sst.tser[0].values
for lag in lags:
  count += 1
  if lag>0: sst1 = ss[lag:] wind1 = ww[:-lag]
  elif lag==0: sst1 = ss wind1 = ww
  else: lag2 = -lag wind1 = ww[lag2:] sst1 = ss[:-lag2]
corr9[count] = np.corrcoef(sst1,wind1)[0,1]
```



Task 2: Algorithms to Analyze CMDA Notebooks

- *De facto* choice for data science
- Commonly comprises rich descriptions and explanations, which are helpful as context for machines to learn toward explainability
  - Used for enrich service usage scenarios





Topic 1: Where is the global warming? ¶

Question 1: Calculate the global net radiative flux imbalance at TOA

Net radiative flux at TOA is calculated from:

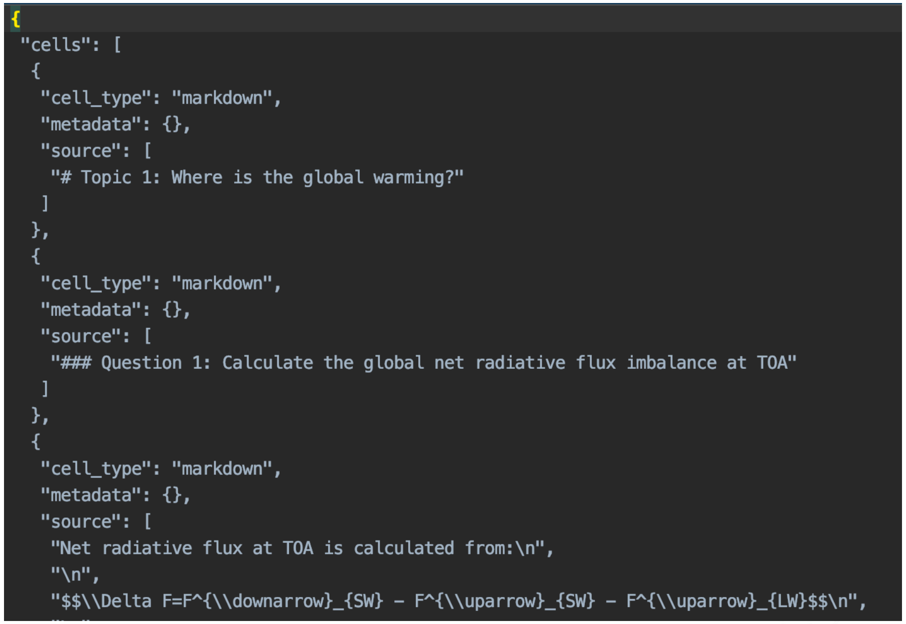
$$\Delta F = F^{\downarrow}_{SW} - F^{\uparrow}_{SW} - F^{\uparrow}_{LW}$$

Where $F^{\downarrow}_{SW} - F^{\uparrow}_{SW}$ is the net incoming shortwave radiation and $F^{\uparrow}_{LW}$ is the outgoing longwave radiation at TOA respectively. First, let's load each of these from the CERES satellite instrument data:

```
In [1]: from cmda import ServiceViewer
        import numpy as np
        import panel as pn
        pn.extension()
        app = ServiceViewer()
        rsdt = app.open_url('http://api.jpl-cmda.org/svc/mapView?purpose=&latS=-90&latE=90&lonS=0&lonE=
        rsut = app.open_url('http://api.jpl-cmda.org/svc/mapView?purpose=&latS=-90&latE=90&lonS=0&lonE=
        rlut = app.open_url('http://api.jpl-cmda.org/svc/mapView?purpose=&latS=-90&latE=90&lonS=0&lonE=
        app.open_url('http://api.jpl-cmda.org/svc/mapView?purpose=&latS=-90&latE=90&lonS=0&lonE=360&mod
```

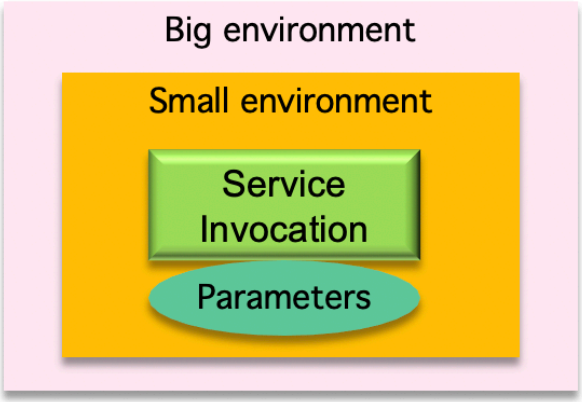
```
---------------------------------------------------------------------------
ModuleNotFoundError                       Traceback (most recent call last)
<ipython-input-1-b5fc233d7e55> in <module>
----> 1 from cmda import ServiceViewer
      2 import numpy as np
      3 import panel as pn
      4 pn.extension()
      5 app = ServiceViewer()

ModuleNotFoundError: No module named 'cmda'
```

topic1.ipynb



```
{
 "cells": [
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "# Topic 1: Where is the global warming?"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "### Question 1: Calculate the global net radiative flux imbalance at TOA"
   ]
  },
  {
   "cell_type": "markdown",
   "metadata": {},
   "source": [
    "Net radiative flux at TOA is calculated from:\n",
    "\n",
    "$$\\Delta F=F^{\\downarrow}_{SW} - F^{\\uparrow}_{SW} - F^{\\uparrow}_{LW}$$\n",
```

Model to Analyze CMDA Notebooks



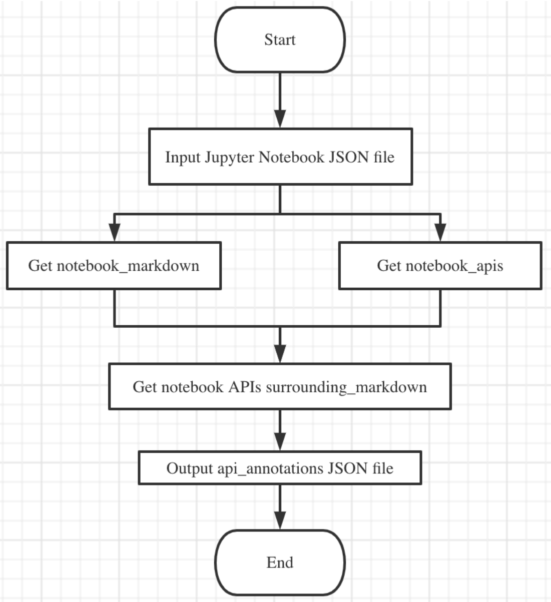
Big environment

Small environment

Service Invocation

Parameters

Cells: cell_type, metadata, source, execution_count
Describe each cell's information
Metadata: kernelspec, language_info
Describe source file information



Start

Input Jupyter Notebook JSON file

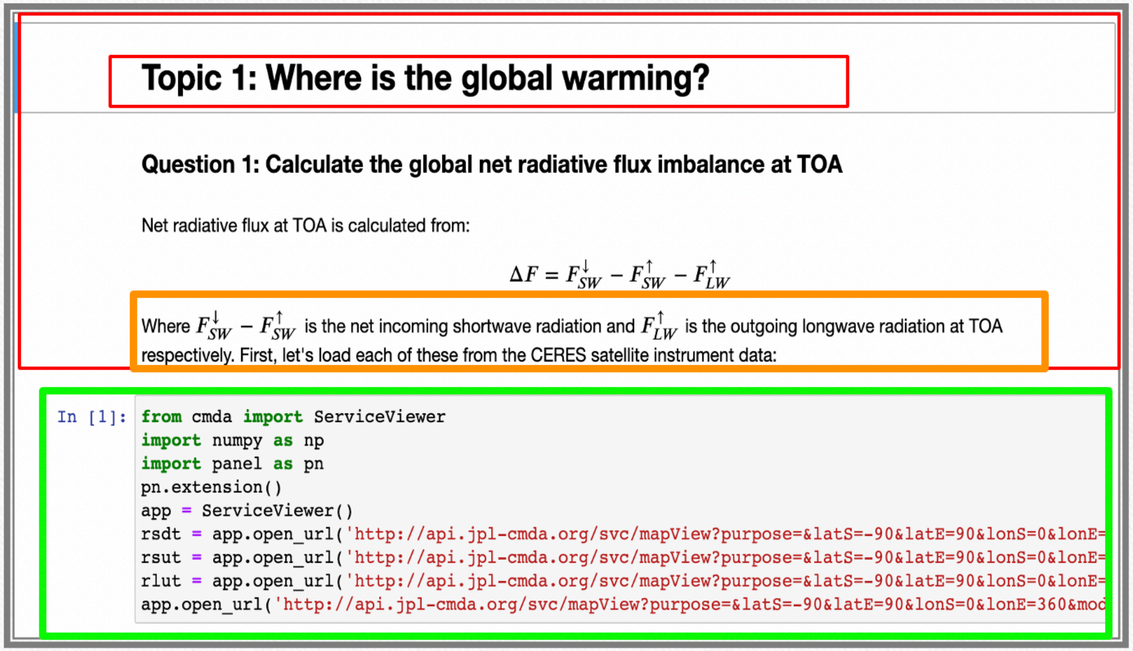
Get notebook_markdown

Get notebook_apis

Get notebook APIs surrounding_markdown

Output api_annotations JSON file

End



Topic 1: Where is the global warming?

Question 1: Calculate the global net radiative flux imbalance at TOA

Net radiative flux at TOA is calculated from:

$$\Delta F = F_{SW}^{\downarrow} - F_{SW}^{\uparrow} - F_{LW}^{\uparrow}$$

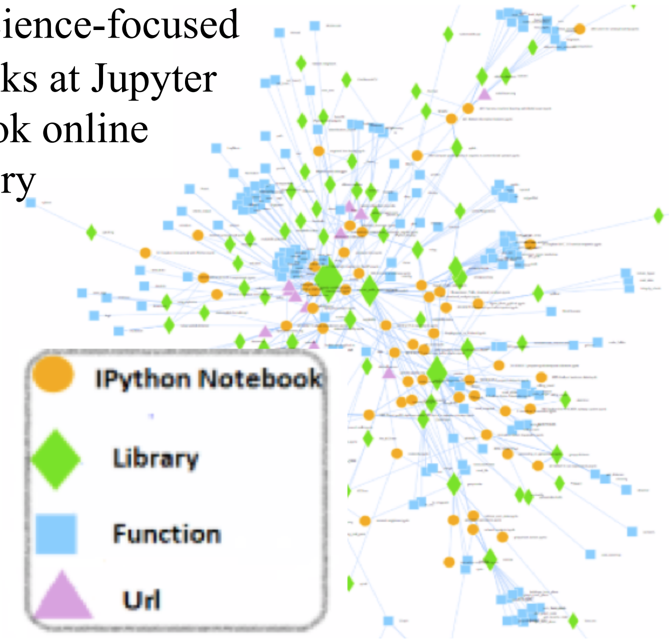
Where $F_{SW}^{\downarrow} - F_{SW}^{\uparrow}$ is the net incoming shortwave radiation and $F_{LW}^{\uparrow}$ is the outgoing longwave radiation at TOA respectively. First, let's load each of these from the CERES satellite instrument data:

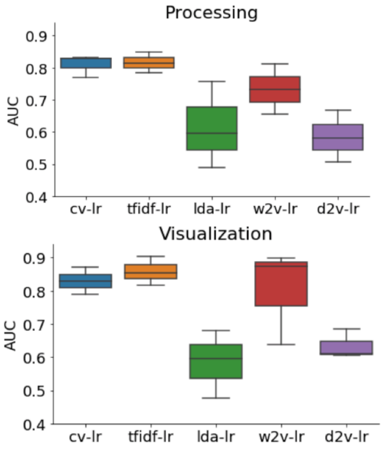
```
In [1]: from cmda import ServiceViewer
        import numpy as np
        import panel as pn
        pn.extension()
        app = ServiceViewer()
        rsdt = app.open_url('http://api.jpl-cmda.org/svc/mapView?purpose=&latS=-90&latE=90&lonS=0&lonE=
        rsut = app.open_url('http://api.jpl-cmda.org/svc/mapView?purpose=&latS=-90&latE=90&lonS=0&lonE=
        rlut = app.open_url('http://api.jpl-cmda.org/svc/mapView?purpose=&latS=-90&latE=90&lonS=0&lonE=
        app.open_url('http://api.jpl-cmda.org/svc/mapView?purpose=&latS=-90&latE=90&lonS=0&lonE=360&mod
```

rlut = app.open_url('http://api.jpl-cmda.org/svc/mapView?purpose=&latS=-90&latE=90&lonS=0&lonE=360&model1=NASA_CERES&var1=rlut&pres1=-999999&vtimeS1=200101&vtimeE1=201112&vmonths1=1&vmonths1=2&vmonths1=3&vmonths1=4&vmonths1=5&vmonths1=6&vmonths1=7&vmonths1=8&vmonths1=9&vmonths1=10&vmonths1=11&vmonths1=12&scale=0&nVar=0')

Unit of Work & Intent Analysis

Earth science-focused notebooks at Jupyter Notebook online repository





code
markdown
parameters

Definition (Unit of Work - UoW) A unit of work is a connected directed graph $uow = < S', E' >$ extracted from a directed graph of a workflow $w = < S, E >$: $uow \subseteq w$, iff:

1) $S' \subseteq S$
2) $E' \subseteq E$
3) uow is connected.

Definition (Network of Units of Work) A network of units of work over M workflows is defined as $N_{uow} = < S'', E'' >$ where $S'' = \bigcup_{j=1}^{M} S_j$ is the set of all services included in all workflows, and $E'' = \bigcup_{j=1}^{M} E_j$ is the set of all the edges in the network each labeled with a workflow identifier.

Definition (Service Intent) The intent of a service $s \in S''$ is defined as ϕ_s which shows a distribution of topics over the Intent space of the network N_{uow}, where its $|T|$-dimensional vector of probabilities sums up to 1: $\sum_{i=1}^{|T|} p_{i,s} = 1$.

Definition (Intent of Unit of Work) The Intent of a unit of work u is defined as $\phi_u = < \phi_{1,u}, \phi_{2,u}, ..., \phi_{|T|,u} >$, where the intent value can be calculated using a Softmax function σ such that:

$$\phi_u = \sigma(\sum_{s\in u} \phi_{i,s}) = \frac{e^{\sum_{s\in u} \phi_{i,s}}}{\sum_{j=1}^{|T|} e^{\sum_{s\in u} \phi_{j,s}}} \qquad (1)$$

Task 3. Network Analysis Algorithm

Definition (Service Cluster) A service cluster SC_j associated to a conceptual service $c_{j,q} \in C_q$ is a set of services $\{s_i\}$ for which $sim(\phi_{s_i}, \phi_{c_{j,q}}) \geq \lambda$.

Definition (User's Aggressiveness) The aggressiveness of user's search query q at time t, is the willingness of the user to accept risks and bigger workflows as the result of the recommendation system at time t. The aggressiveness can be calculated as follows:

$$A_{q,t} = \delta * |C_q| + \eta * (|C_q| - |S_{W_t}|) \tag{2}$$

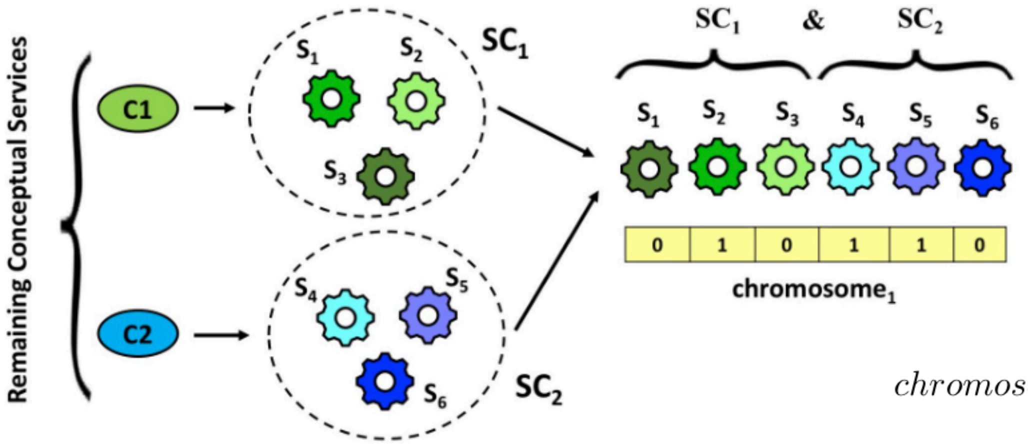
Definition (Search Query) A search query at time t is a triple q_t $=< G_q, W_t, A_t >$, where G_q $=< \phi_q, C_q >$ is the final goal of the user which contains ϕ_q as the user's intent and C_q as the list of user's desired conceptual services identified, respectively. W_t represents the current partial workflow, and A_t is the user's current aggressiveness which

Definition (UoW Recommendation Problem) Given a search query q at time t, the Unit of Work Recommendation Problem aims to find a connected subgraph from the UoW network to maximize weighted coverage of the conceptual services C_q, while keeping the noise among the services less than or equal to the user's aggressiveness $A_{q,t}$. Hence, this problem can be formulated as an optimization problem as follows:

$$\begin{aligned}
\text{maximize} \quad & \sum_{c_j \in C'_q} sim(\phi_{c_j}, \phi_q) \cdot cov(c_j) \\
\text{subject to} \quad & \sum_{s_i \in S} (1 - sim(\phi_{s_i}, \phi_q)) \cdot sel(s_i) \leq A_t \\
& connected(subgraph\{s_i | sel(s_i) = 1\}) \\
& \textstyle\sum_{\forall s_i \in SC_j} sel(s_i) \geq cov(c_j), \quad \forall c_j \in C'_q \\
& cov(c_j) \in \{0, 1\}, \quad j = 1, ..., N_{C'_q} \\
& sel(s_i) \in \{0, 1\}, \quad i = 1, ..., N
\end{aligned} \tag{5}$$

Theorem 1. *UoW Recommendation Problem is NP-hard.*

Task 3.1 Genetic Algorithm-based Solution



$$chromosome = [C_1 \& C_2 \& \ldots \& C_N] \tag{6}$$

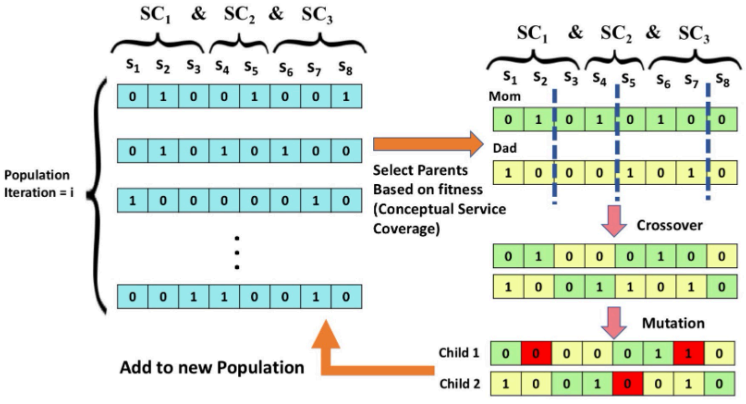
$$chromosome = [SC_1 \& SC_2 \& \ldots \& SC_N] \tag{7}$$

$$chromosome = [s_{11} s_{12} \ldots s_{1i} \& s_{21} s_{22} \ldots s_{2j} \& \ldots \& s_{n1} s_{n2} \ldots s_{nk}] \tag{8}$$

$$fitness(ch_i) = \begin{cases} 0 : & \text{if } invalid(ch_i) \\ \sum_{\{j \mid \forall_k gene_k = 1, s_k \in SC_j\}} \varpi_j : & \text{otherwise} \end{cases} \tag{9}$$

A chromosome is considered *invalid* in the following three cases:

1) $\forall SC_j \in SC_q : |\{s_k | \forall_k gene_k = 1, s_k \in SC_j\}| > 1$
2) $subgraph\{s_i | \forall_i gene_i = 1\}!connected$
3) $\sum_{s_i \in S} (1 - sim(\phi_{s_i}, \phi_q)).gene_i > A_t$


SC_1 & SC_2 & SC_3
S_1 S_2 S_3 S_4 S_5 S_6 S_7 S_8
Population Iteration = i
Select Parents Based on fitness (Conceptual Service Coverage)
Mom
Dad
Crossover
Mutation
Child 1
Child 2
Add to new Population

Population Iteration = i:

0 1 0 0 1 0 0 1

0 1 0 1 0 1 0 0

1 0 0 0 0 0 1 0

⋮

0 0 1 1 0 0 1 0

Mom: 0 1 0 1 0 1 0 0

Dad: 1 0 0 0 1 0 1 0

Crossover:

0 1 0 0 0 1 0 0

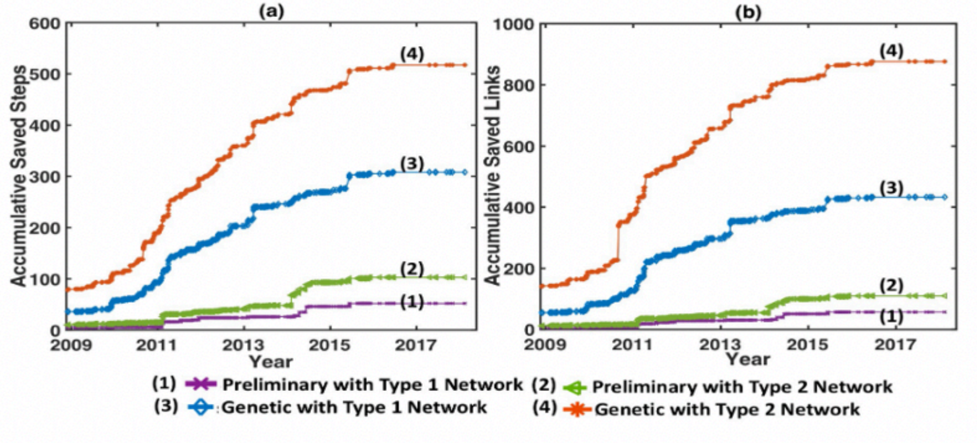
1 0 0 1 1 0 1 0

Mutation:

Child 1: 0 0 0 0 0 1 1 0

Child 2: 1 0 0 1 0 0 1 0

Experimental Results



Development efforts saved comparison
(a) Steps saved
(b) (b) Links saved

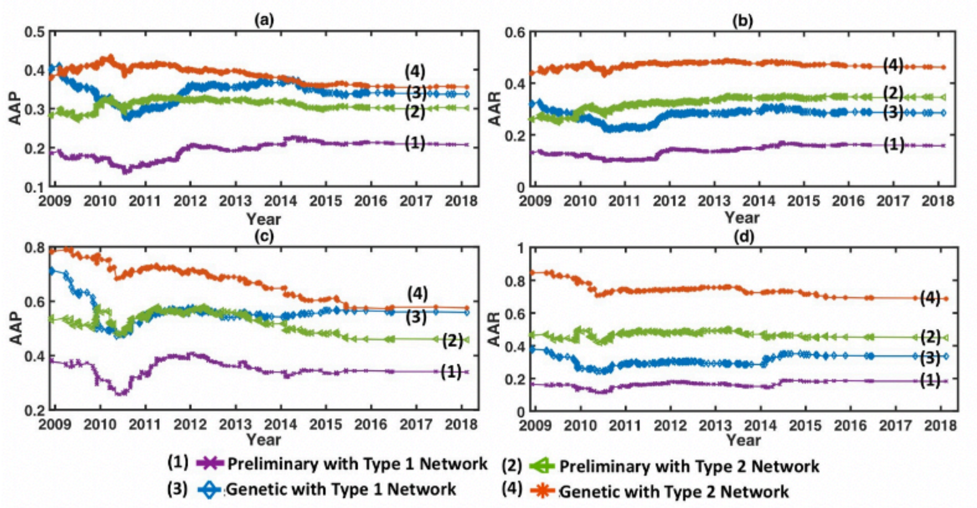


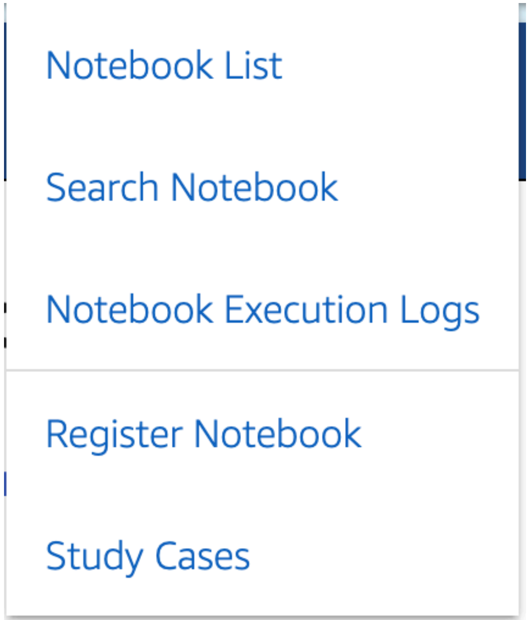
TABLE 1: Summary of Testbed

# Unique workflows	2,030
# Workflow versions	3,277
# Unique services	513
# Unique service operations	1,248
# Workflows with at least one service	1,719
# Workflows with at least two services	511

- Testbed design
  - myExperiment.org
  - 2007-2018
- Testing scenarios
  - Every workflow was treated as a user search query.
    - For each workflow, UoW network contains all in prior workflows
    - For each workflow, UoW network remains almost the same
- Baseline methods
  - Semantics
  - Pattern

Task 4: Workflow Recommendation System

- On top of
  - Open NASA Earth Exchange (OpenNEX) platform
  - CMAC App Store project
- Recommender system
  - Browse all notebooks
  - Search notebook
  - Manage notebook execution logs
  - Register new notebook
  - Publish interesting notebook usages
- Used at 2020 JPL Summer School



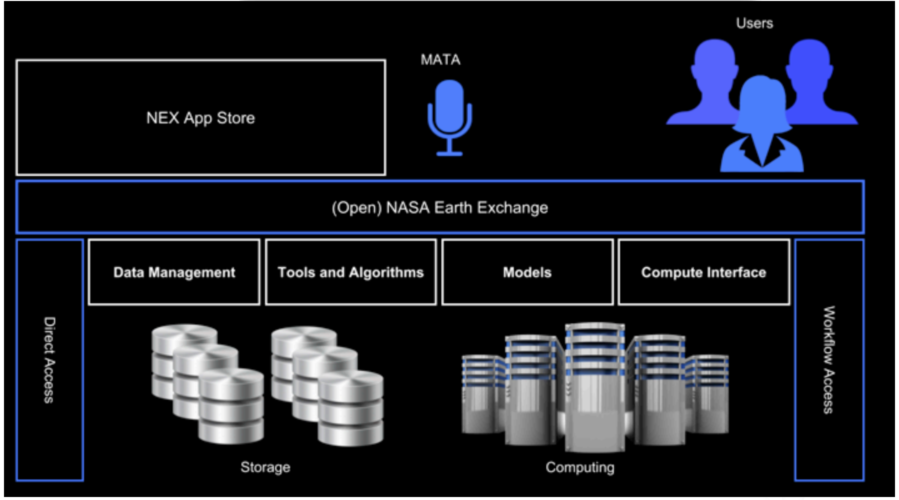
Notebook List

Search Notebook

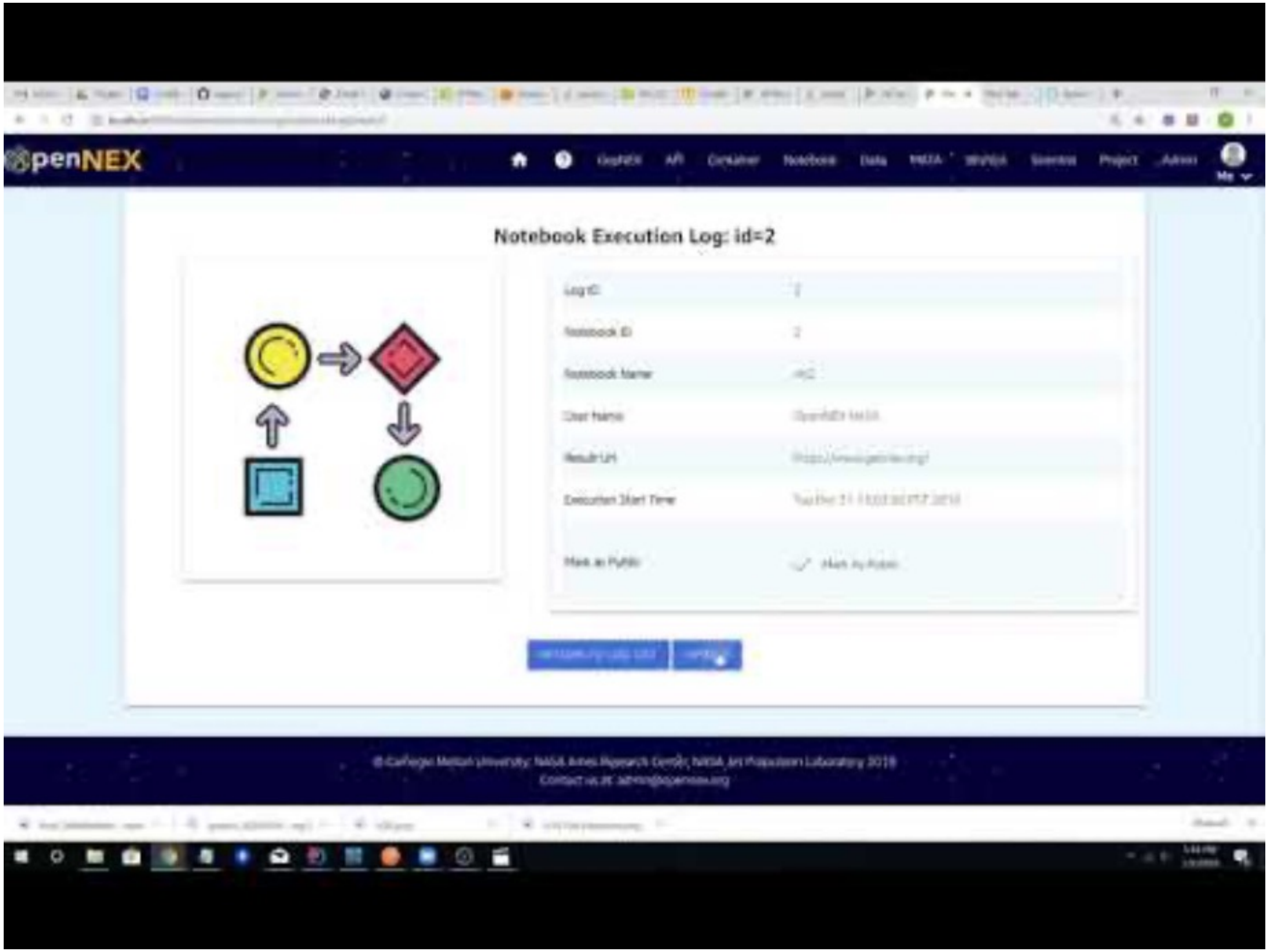
Notebook Execution Logs

Register Notebook

Study Cases



Demo on Workflow Tool



Notebook Execution Log: id=2

Log ID	2
Notebook ID	2
Notebook Name	[illegible]
User Name	OpenNEX NASA
Result Url	[illegible]
Execution Start Time	[illegible]
Mark as Public	✓ Mark As Public

© Carnegie Mellon University; NASA Ames Research Center; NASA Jet Propulsion Laboratory 2018
Contact us at: admin@opennex.org

Task 5: User Test (2020 Virtual NASA Summer School)

- CMDA and OpenNEX App Store were used to support the virtual NASA Summer School on Satellite Observations and Climate Models in 2020.
- The NASA Summer School brings together the next generation of climate scientists to engage with premier climate scientists.
- The summer school students perform a group research project using CMDA analysis tools and OpenNEX collaboration supporting tools.
- We provided both the CMDA service web interface (original) and the CMDA service Jupyter Notebook interface (new).
- The survey after the virtual summer shows that about 50% of the students used the web interface and the other 50% of students used the Jupyter Notebook interface.

https://opennex.org

2020 NASA Summer School on Satellite Observations and Climate Models

Web Interface: http://api.jpl-cmda.org
Jupyter Notebook Interface: http://hub.jpl-cmda.org

Group Research Topics

1. Where is global warming?
2. Tropical variability and analysis of the El Nino-Southern Oscillation (ENSO) forcing
3. Variability of clouds and precipitation
4. Land water storage variability
5. Sensitivity of equilibrium climate to physical parameterizations
6. Added values of high-resolution downscaling

https://opennex.org

2020 NASA Summer School on Satellite Observations and Climate Models

Students will be engaged in group projects for hands-on exercise of using satellite observation data and climate model outputs for climate science research. In the group projects, the students will explore satellite observation and climate model data to study one of the following six topics:

1. Where is global warming?
2. Tropical variability and analysis of the El Nino-Southern Oscillation (ENSO) forcing
3. Variability of clouds and precipitation
4. Land water storage variability
5. Sensitivity of equilibrium climate to physical parameterizations
6. Added values of high-resolution downscaling

All relevant datasets are provided by the following analysis tools:

- Web-based Tool: http://api.jpl-cmda.org
- Jupyter-Notebook-based Tool: http://jpl-cmda.org

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Future Plans
- Publications - List of Acronyms

Summary of Accomplishments and Future Plans

• Task 1: CMDA Jupyter notebook examples	04/20
• Task 2: Algorithms to analyze CMDA notebooks	06/20
• Task 3: Network analysis algorithms	08/20
• Task 4: Workflow recommendation system	09/20
• Task 5: User test; CMDA notebooks	10/20

• Task 6: Notebook templates; refined notebook analysis	01/21
• Task 7: Notebook templates; Enhanced workflow tool	07/21
• Task 8: JPL Summer School	09/21
• Task 9: User testing and documentation	10/21

Presentation Contents

- Background and Objectives
- Technical and Science Advancements
- Summary of Accomplishments and Plans Forward
- Publications - List of Acronyms

Publications

"Unit of Work Supporting Generative Scientific Workflow Recommendation"
(J. Zhang, M. Pourreza, S. Lee, R. Nemani, and T.J. Lee)
International Conference on Service Oriented Computing (ICSOC)

"Mining Units of Work in Scientific Workflow Provenance"
under revision at IEEE Transactions on Services Computing

Acronyms

ACF	Analytic Center Framework
API	REST web service, remotely accessible software component
Workflow	Multi-step data analytics procedure, also known as mashup
REST	REpresentational State Transfer
UoW	Unit of Work
CMDA	Climate Model Diagnostic Analyzer
NEX	NASA Earth eXchange
OpenNEX	Open NASA Earth eXchange
Notebook	Juypter notebook
Provenance	Data analytics history, data analytics procedure execution logs
SOC	Service Oriented Computing
ENSO	El Nino-Southern Oscillation
SST	Sea-Surface Temperatures
EOF	Empirical Orthogonal Function
NLP	Natural Language Processing
NRC	National Research Council
IPCC	Intergovernmental Panel on Climate Change
AR6	Assessment Report