Applying Data Science in Earth Science to Understand Climate

Deb Agarwal
daagarwal@lbl.gov
Scientific Data Division (https://crd.lbl.gov/divisions/scidata/), Interim Director
Lawrence Berkeley National Laboratory
Berkeley Institute for Data Science - Affiliate

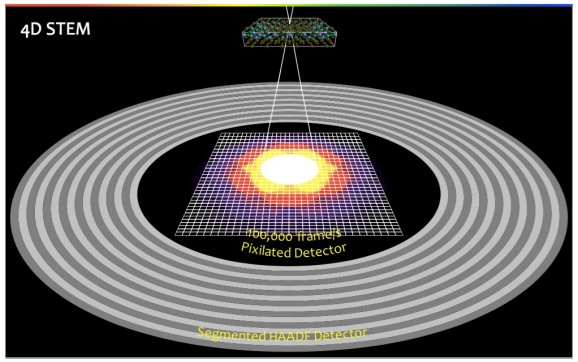
Acknowledgements

● Thank you to the many researchers and collaborators at Berkeley Lab and across the DOE national labs and academia whose work is described in this talk. Without their contributions, none of this work would be possible. A particular thank you for slides and images:

- Lavanya Ramakrishnan
- Charuleka Varadharajan
- Suren Byna
- John Wu
- Talita Perciano
- Juli Mueller
- Michael Wehner
- Gilberto Pastorello
- Dan Gunter
- You-Wei Cheah
- Forrest Hoffman
- AI4ESP Workshop participants
- John Shalf
- Michael Wehner
- Prabhat
- Bin Dong
- Shreyas Cholia
- Dennis Baldocchi
- Margaret Torn
- Sebastien Biraud

Climate Data Comes From a Wide Array of Instruments

Environmental sensors



Electron microscopes

Drones

Sequencers

Accelerators and Light sources

Satellites

Hard to measure and rapidly changing natural systems

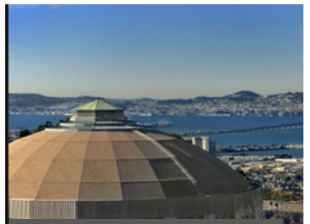
Climate Data Science Challenges are Large-scale and Complex

Volume

Variety/ Complexity

Velocity

Molecular Foundry



ALS

JGI

JOINT GENOME INSTITUTE

Data Rates:
ALS (2027) - 25 PB/year
JGI (~2023) - 30 PB/day
Molecular Foundry/NCEM (2022)- 30 PB/year (generated), 1 to 5 PB/year (saved)

ESS-DIVE
Deep Insight for Earth Science Data

AmeriFlux Network

nmdc
National Microbiome Data Collaborative

Watershed Function
Scientific Focus Area

The Materials Project
materialsproject.org

KBase

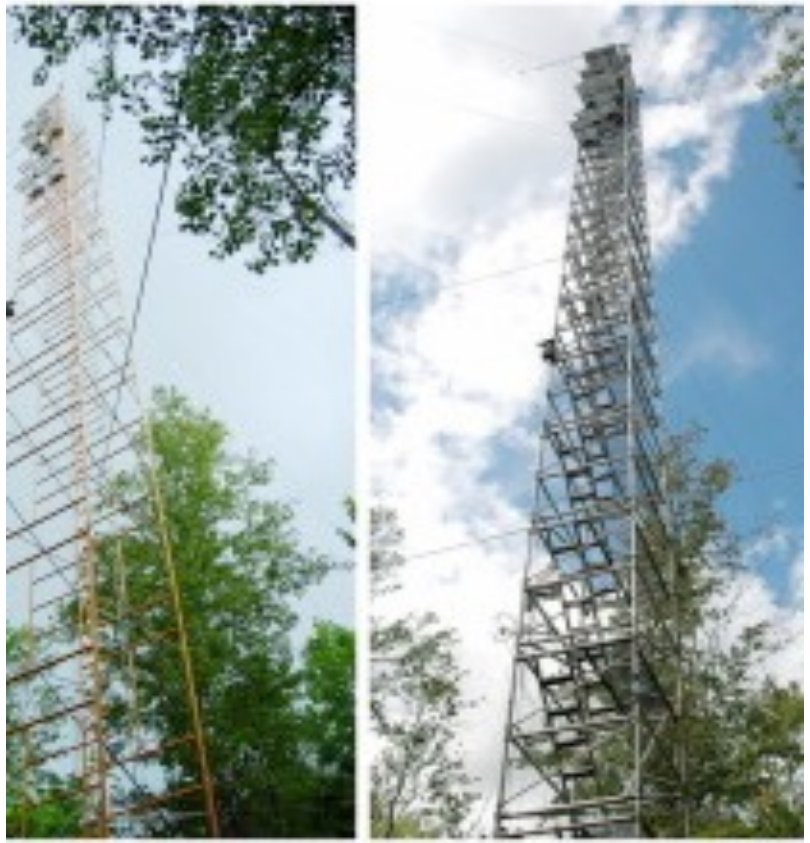
- Chemistry labs (e.g. EMN projects) Bldg 70, 33/34
- Biology "wet" labs (JBEI)
- Seismic monitoring
- Energy measurements
- Cryo-EM
- [Planned] earth-sciences "pods"

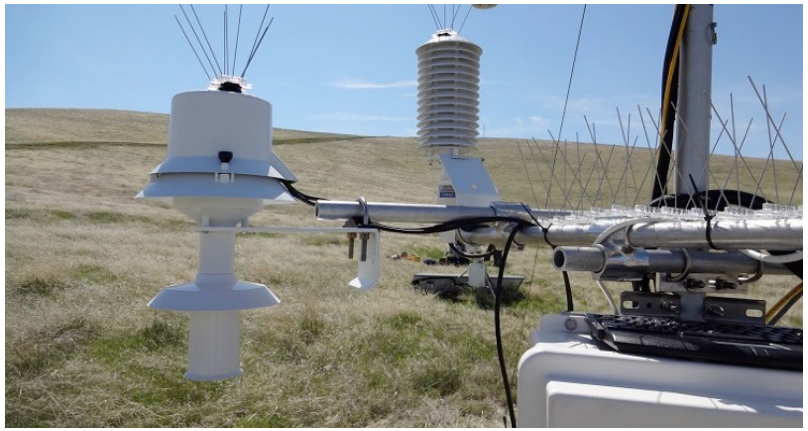
BioEPIC

NERSC

ESnet

Some Example Climate Data Projects

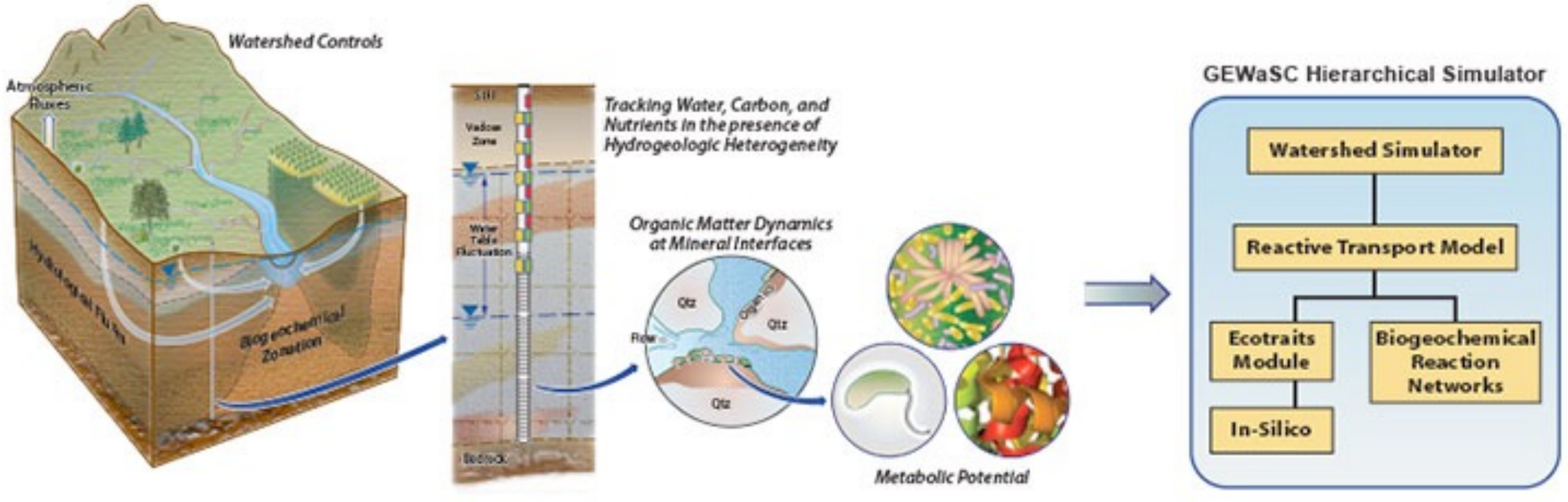




AmeriFlux Understanding Carbon Flux in the Americas

https://ameriflux.lbl.gov/

Watershed SFA - Understanding Watershed Function



https://ess.science.energy.gov/lbnl-watershed-function-sfa/

Our Science Teams are Large and Diverse

Margaret Torn
AMP Lead PI, AmeriFlux Management Project
Lawrence Berkeley National Lab
510-495-2223
mstorn@lbl.gov
bio website

Dennis Baldocchi
AMP Co-PI, Core Site PI, SSC
UC Berkeley / Tonzi/Vaira/Delta sites
510-642-2874
baldocchi@berkeley.edu
bio website

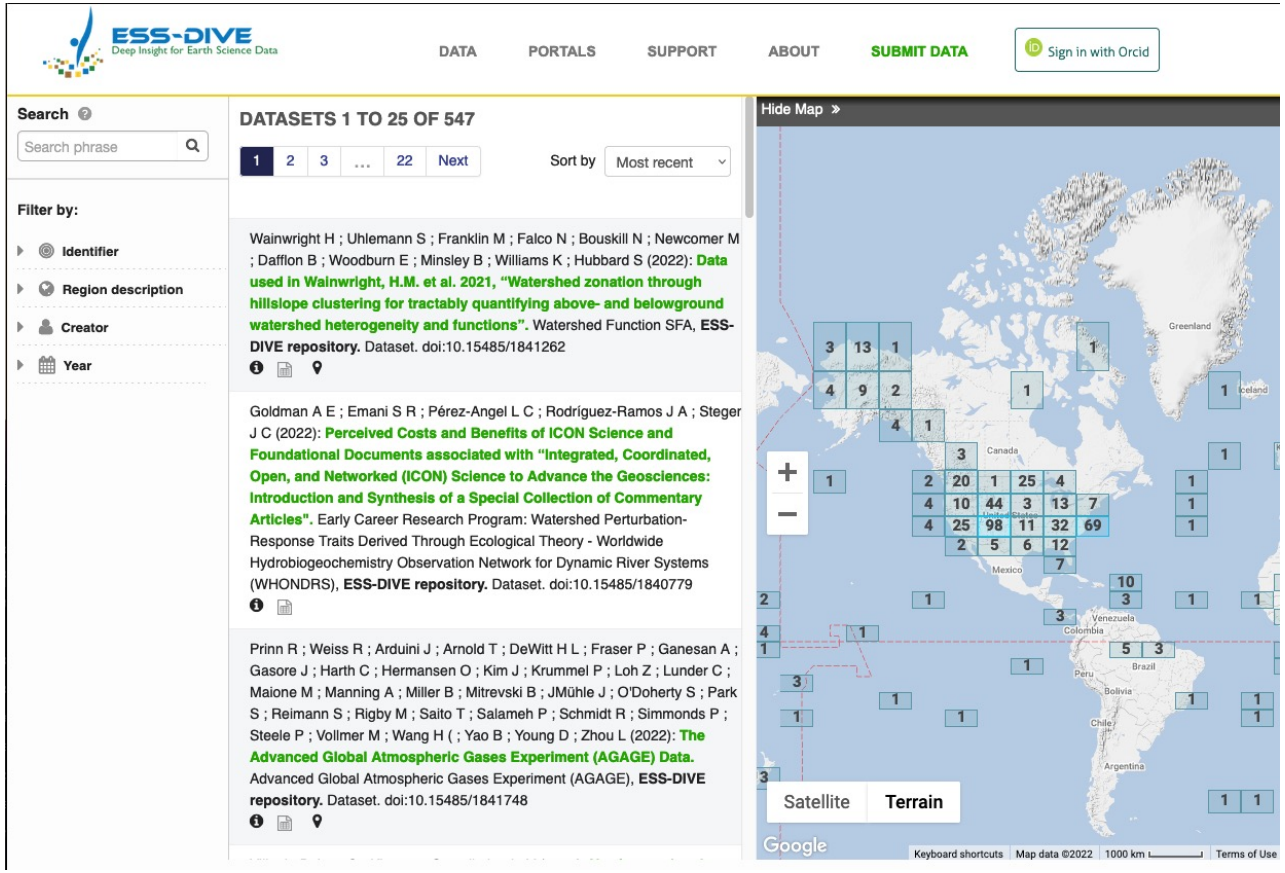
Sébastien Biraud
Deputy project lead, AMP Tech Team Lead
Lawrence Berkeley National Lab
510-486-6084
scbiraud@lbl.gov
bio website

Deb Agarwal
AMP Data Team Lead
Lawrence Berkeley National Lab
510-486-7078
daagarwal@lbl.gov
bio website

Trevor Keenan
AMP Outreach and Network Coordination Lead
Lawrence Berkeley National Laboratory/ UC Berkeley
trevorkeenan@lbl.gov
bio website

AmeriFlux Project Leadership

Environmental Systems Science Data Repository – ESS-DIVE



- Long-term archiving of data from the DOE Earth Systems Science projects

FAIR

- Data **Findable** through search and DOI
- **Accessible** through APIs and web
- **Interoperable** standardized data formats and metadata accessible through APIs
- **Reusable** based on citation and data versioning

https://ess-dive.lbl.gov/

Understanding the Less Obvious Climate Data Challenges

Climate Data are Diverse, Multi-scale, Multi-modal, and Difficult to Collect

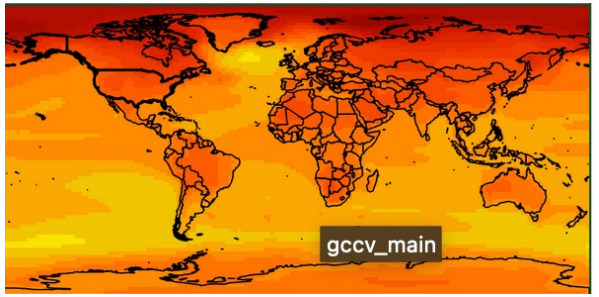
Ground-based Measurements

Limitations
- Sparse geographically
- Uneven distribution
- Limited duration
- Difficult measurements
- Only historical

Remote Measurements

Limitations
- Frequency low typically
- Limited variables
- Limited duration
- Only historical

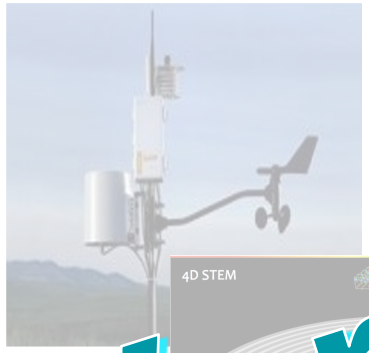
Simulations

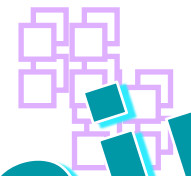


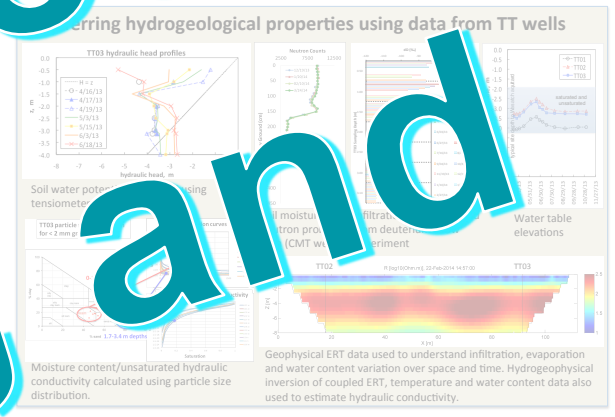
Limitations
- Limited representation of processes

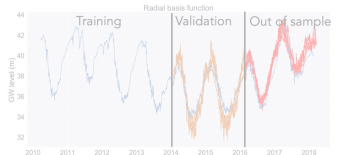
Climate Science Needs Help to Address the Whole Data Lifecycle

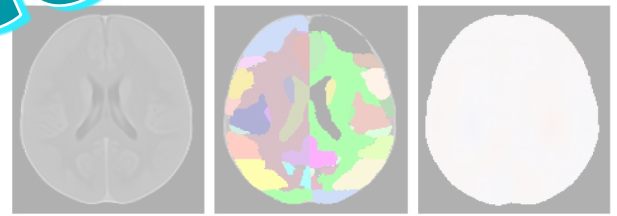
Data Generation Data Management Data Analytics





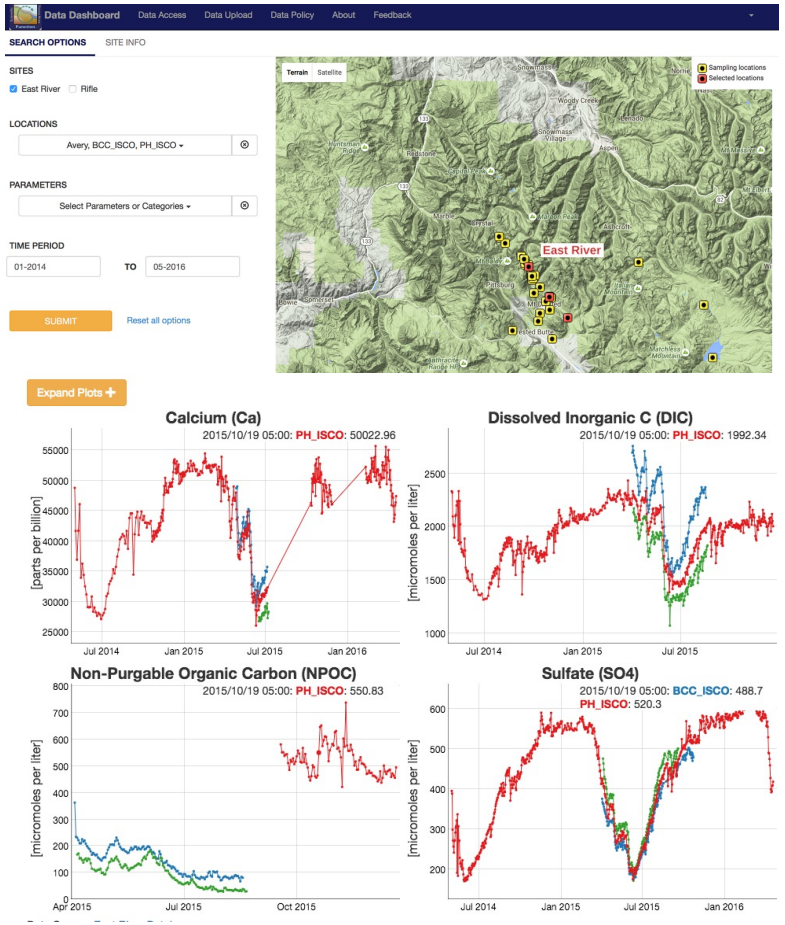






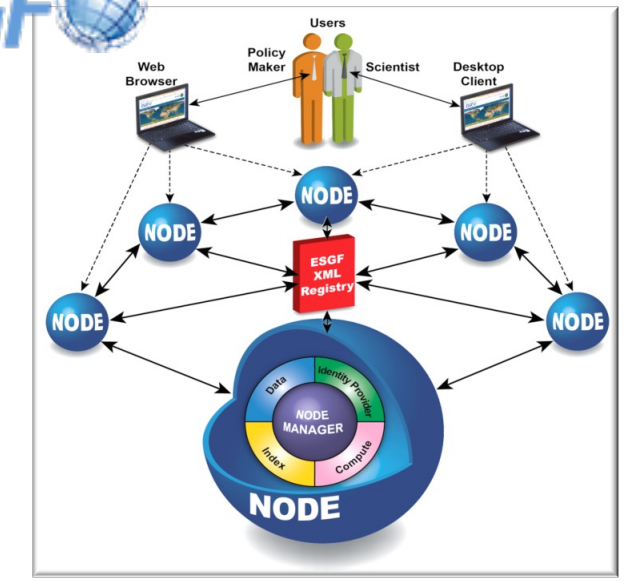
Versioning, Reuse, Annotation, and Metadata

Addressing Usability of the Data



Making Climate Data Available

ESGF



KBase
PREDICTIVE BIOLOGY

EMSL

FLUXNET
A Global Network

AMERIFLUX

IGSN

ESS-DIVE
Deep Insight for Earth Science Data

NATIONAL SCIENCE FOUNDATION
LTER NETWORK
LONG TERM ECOLOGICAL RESEARCH

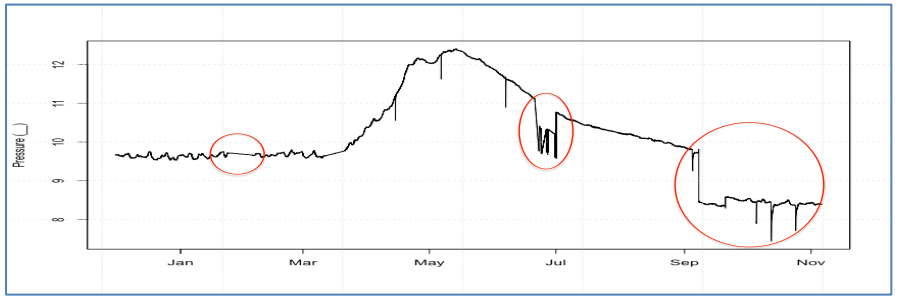
NASA EARTHDATA
OPEN ACCESS FOR OPEN SCIENCE

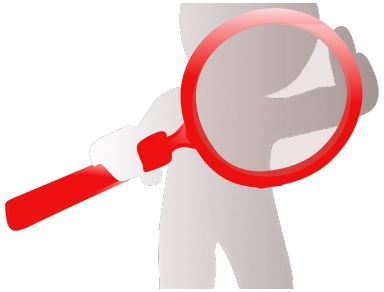
Making Climate Data Usable

- Licensing clear and findable
- Format consistent and interpretable
- Units / methodology described and compatible
- Metadata about the data available
- Citation/credit requirements defined
- QA/QC applied gaps filled?
- Assumptions/circumstances of data collection
- Uncertainty assessment
- Papers/analyses performed

Most Data Requires Significant Processing Before it is Ready to Use

QA/QC

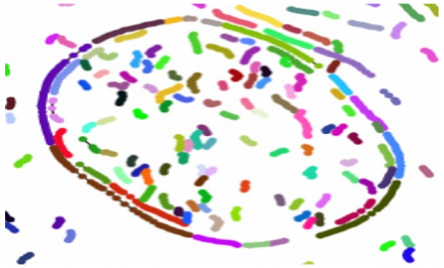




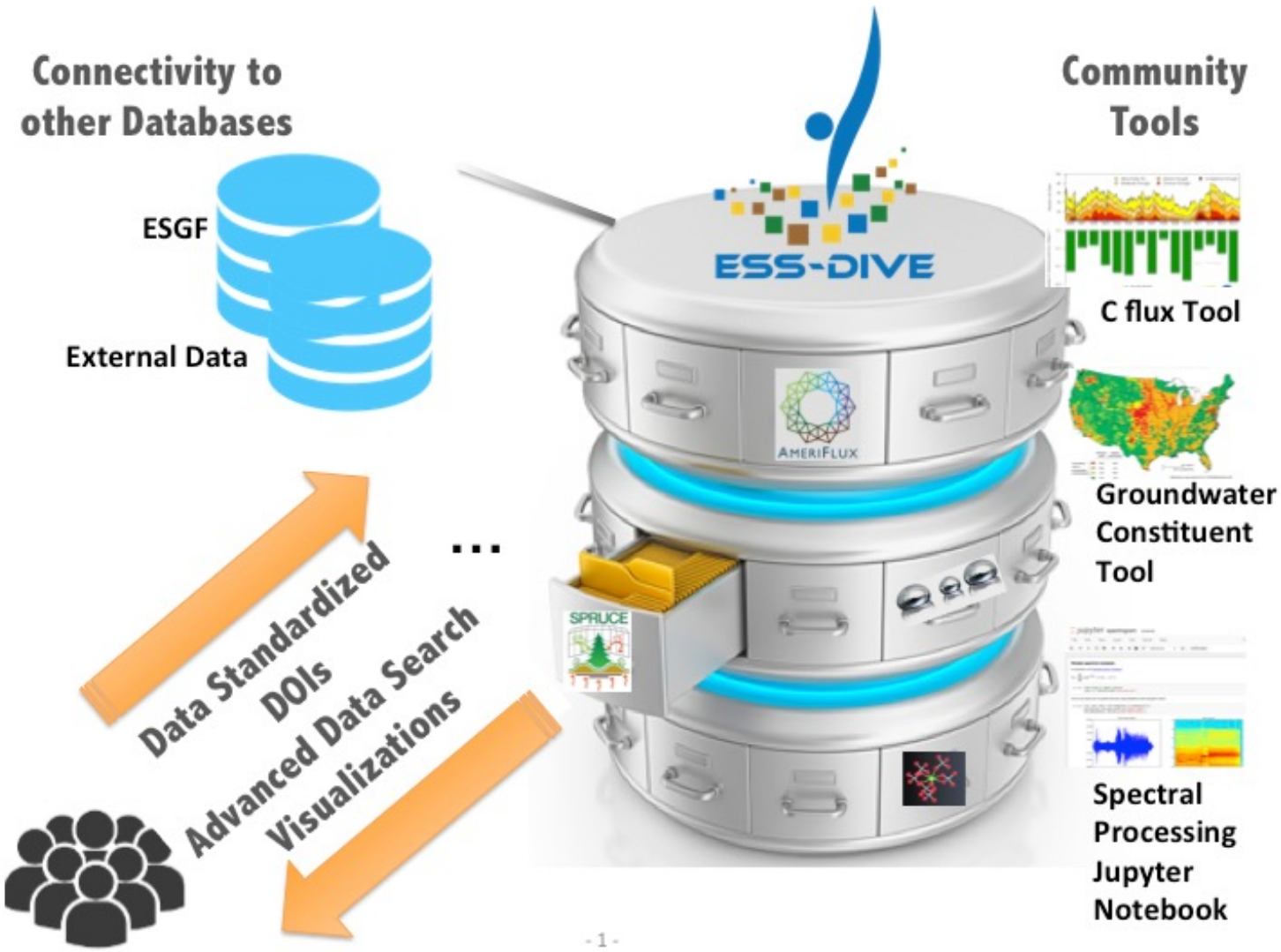
Identification and removal or gap-filling of missing/bad data

Enhancement

Transformed Representation (Graph)



Data Repositories Are Core to the Usability and Reusability of Data



HDF5 Has Been Emerging as a Standard Data Format in Science

HETEROGENEOUS DATA –

HDF® supports n-dimensional datasets and each element in the dataset may itself be a complex object.

EASY SHARING +

CROSS PLATFORM +

FAST I/O +

BIG DATA +

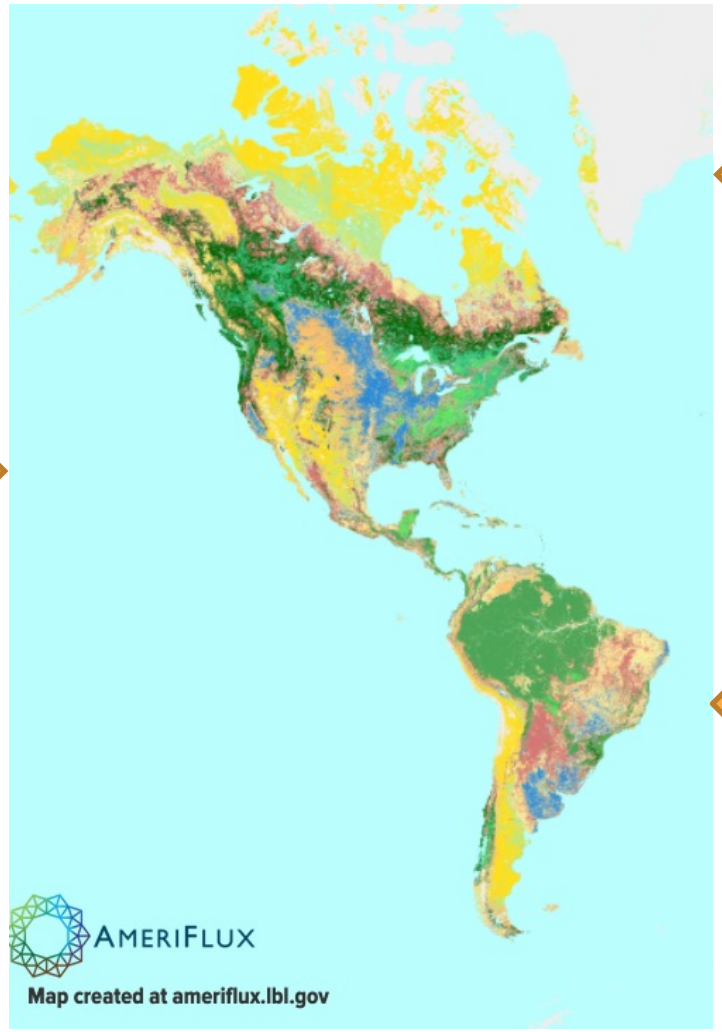
KEEP METADATA WITH DATA +

https://www.exascaleproject.org/research-project/exaio/
https://www.hdfgroup.org/solutions/hdf5/

Making Climate Data Usable (2)

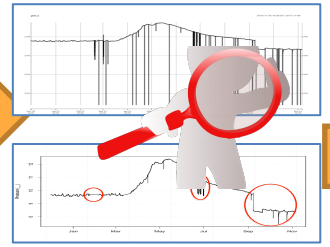
- Appropriate scale
- Geolocation matches need
- Frequency and duration matches need
- Needed variables are measured
- Metadata are available
- Compatible with other datasets

Datasets Often Are Integrated Across Scales to Build the Needed Dataset



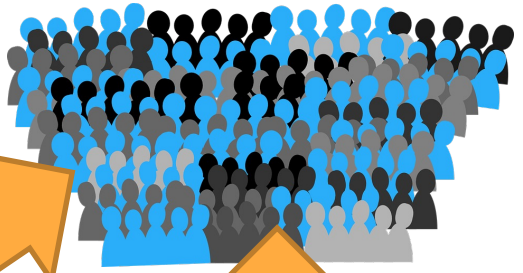
AmeriFlux Data Lifecycle

QA/QC



Base Data Product

Data Repository



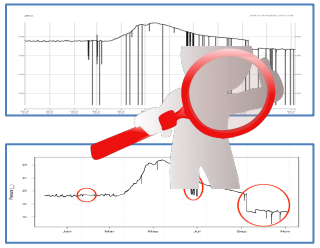
Analytics

Fluxnet Data Product

QA/QC

Raw Data Storage

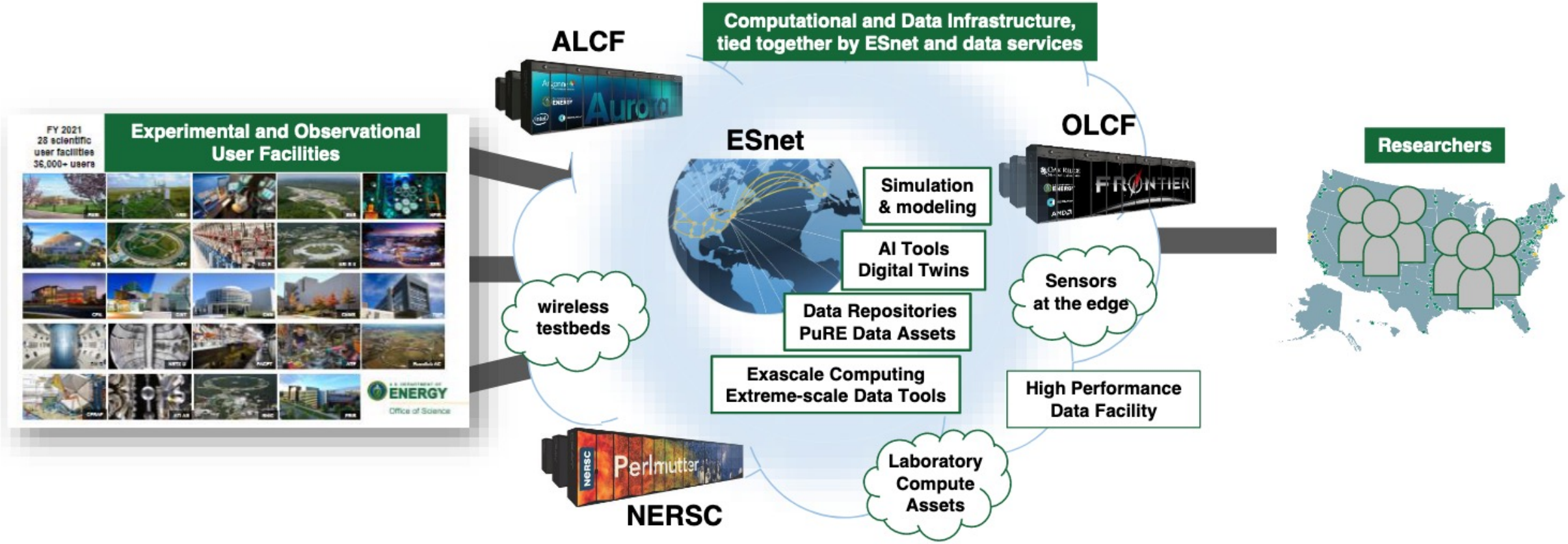
QA/QC



Processing Flux Assessment, etc

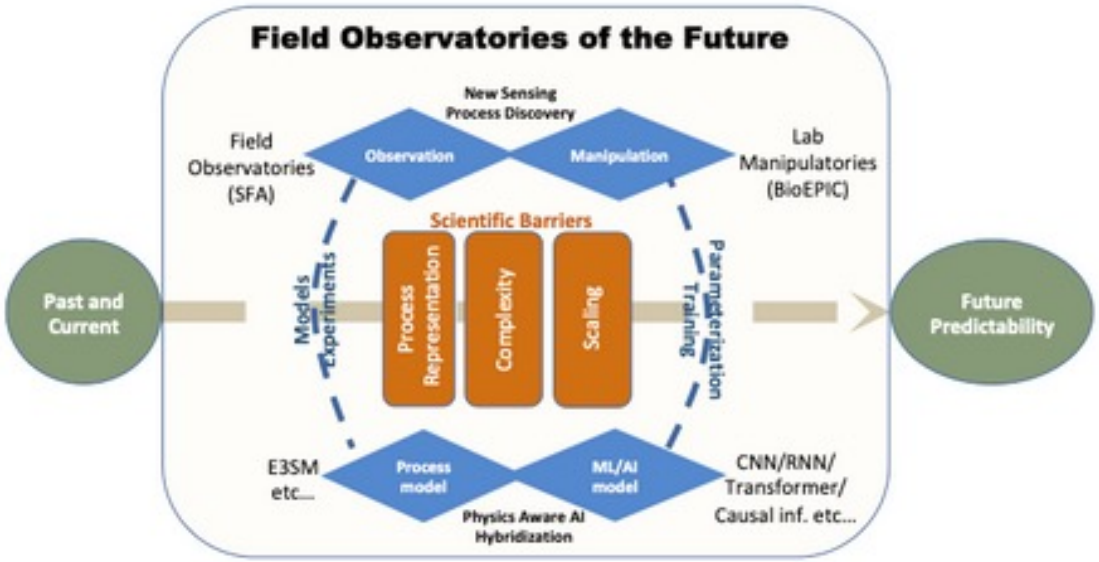
Weather Simulation

Incipient ecosystem: Office of Science User Facilities



Source: Ben Brown ASCAC Talk – Sept 29, 2021

Conceptual example to improve water budget estimates in mountainous regions

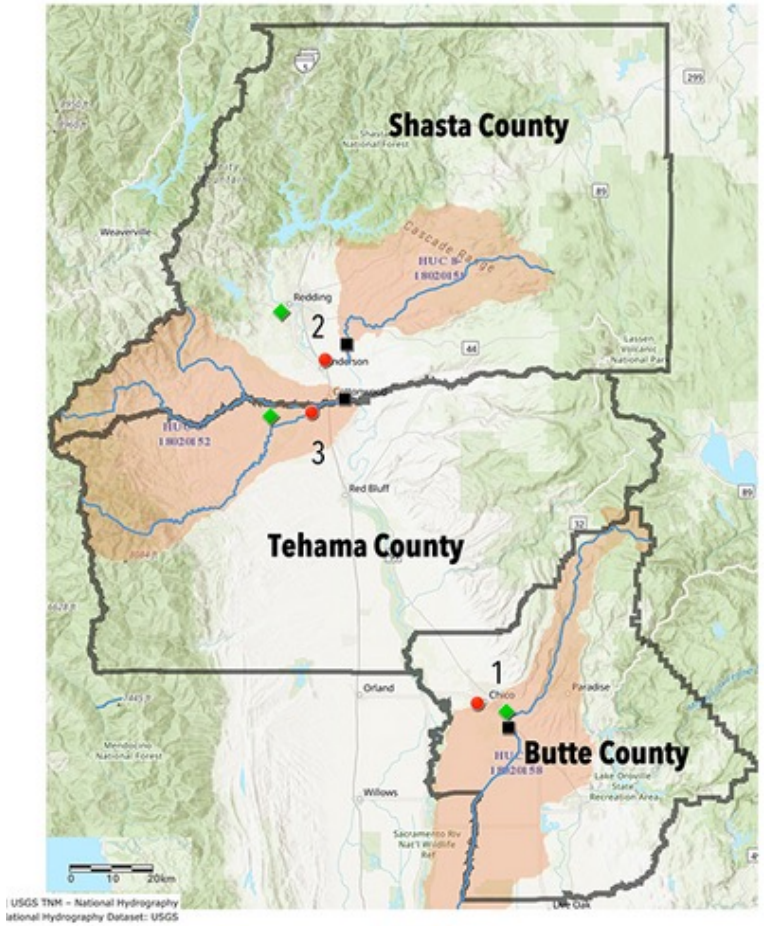


Near real-time decisions on 'what', 'when' and 'where' to sample to enhance scientific understanding and Earth system predictability

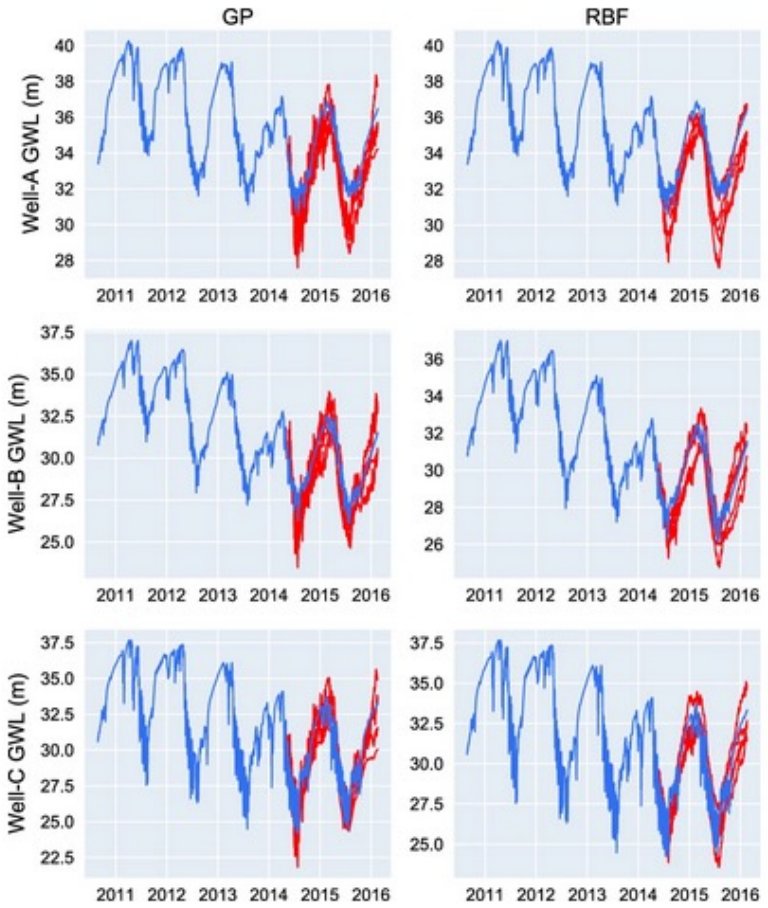
Slide from a recent ML4Sci talk by Charu varadharajan

Groundwater Prediction Using Deep Learning

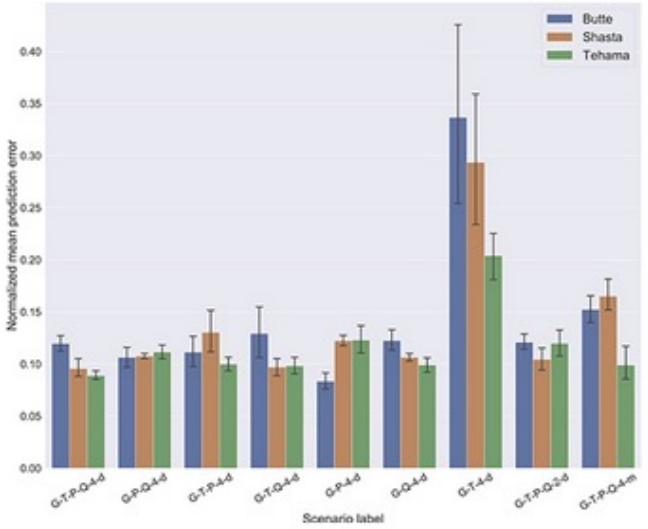
Daily groundwater predictions in 3 counties 2016--2018 with different SGMA prioritization



Multiple deep learning models (MLP, CNN, LSTM) with novel hyperparameter optimization

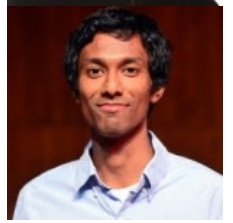
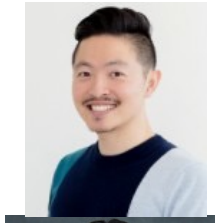


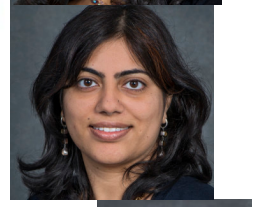
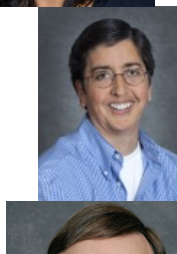
Robust sensitivity analysis to different inputs



Once trained, models can be run very quickly and can predict reliably for upto ~1 year

Mueller et al. J. of Global Optimization, 2020
Sahu et al., Fron. in Water, 2020

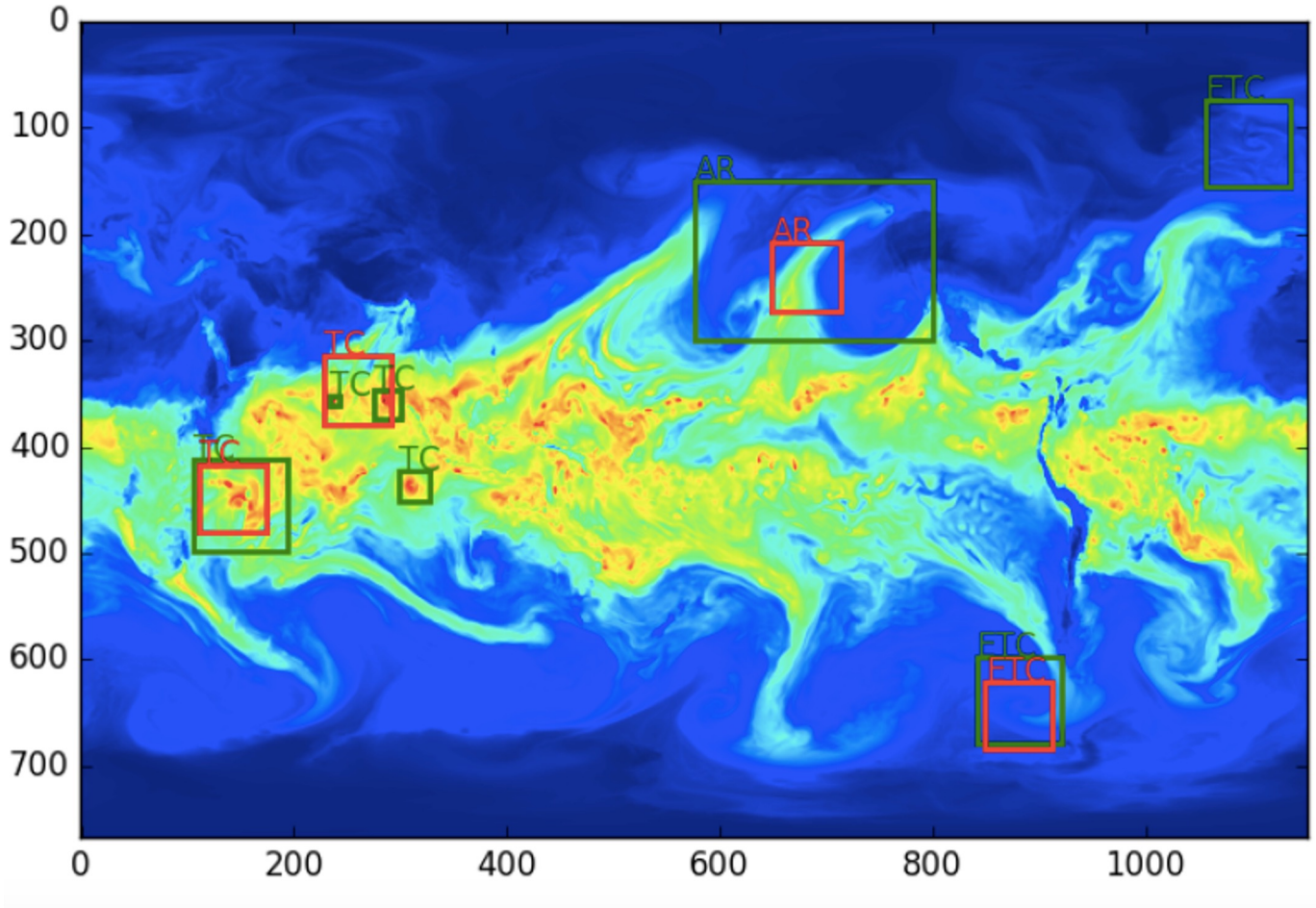



Climate Science is Increasingly Using Machine Learning

- FourCastNet – trying to build a climate model using ML
  - Climate model outputs used as the training data
  - Looking at expanding into use of observation data
- Storm prediction using ML to find and count predicted storm number and size
  - Climate model outputs used as inputs
- Digital twins
- Forcasting
- …

Deep Learning: Classification + Regression Results

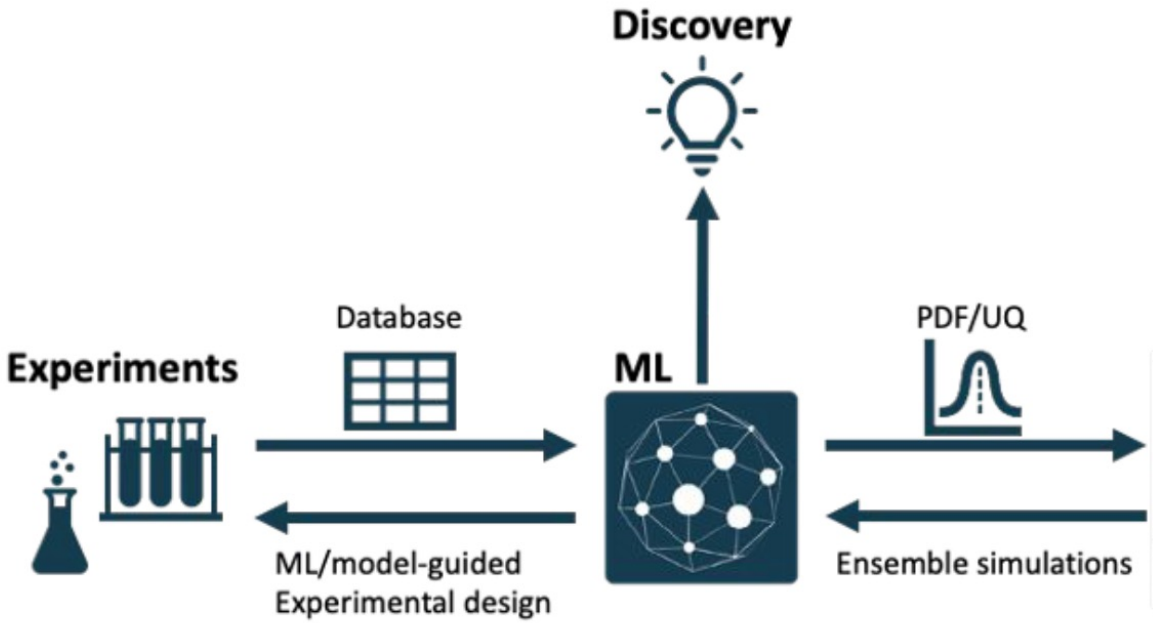
Identify extreme storms in climate simulations



Ground Truth
Prediction

Contributors: Thorsten Kurth, Jian Yang, Ioannis Mitliagkas, Chris Pal, Nadathur Satish, Narayanan Sundaram, Amir Khosrowshahi, Michael Wehner, Bill Collins, Prabhat

New Framework to Unify Community-Wide Experimental Geochemistry Data for Model Intercomparison and Machine Learning (Mavrik Zavarin, LLNL)



- Laboratory experiments are critical for BGC modeling
  - But often experiment data are too small to apply ML/AI
- Framework to create large community-wide dataset for AI/ML for inter-model comparison and comprehensive UQ
- ML/AI developments: surrogate models, data assimilation, hybrid geochemical model (physics model for aqueous; ML for surface)

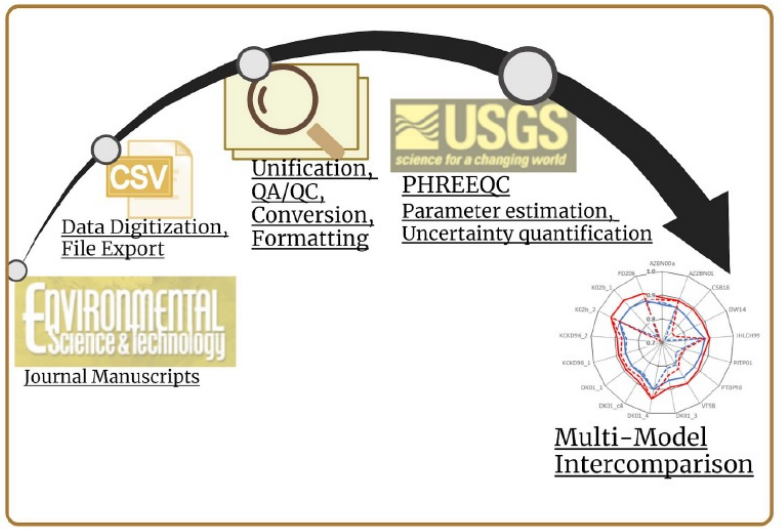
First Paper: This study demonstrates a comprehensive workflow

(1) Mine experimental data from the literature

(2) Perform systematic parameter estimation and uncertainty quantification

(3) Perform model intercomparison.

- Global multi-institutional data, and 17 surface complexation models.



Mavrik Zavarin, Elliot Chang, Haruko Wainwright, Nicholas Parham, Rahul Kaukuntla, Jadallah Zouabe, Amanda Deinhart, Victoria Genetti, Sam Shipman, Frank Bok, and Vinzenz Brendler, **Community Data Mining Approach for Surface Complexation Database Development,** *Environmental Science & Technology* **Article ASAP** , DOI: 10.1021/acs.est.1c07109

AI-Constrained Ecohydrology for Improving Earth System Predictions

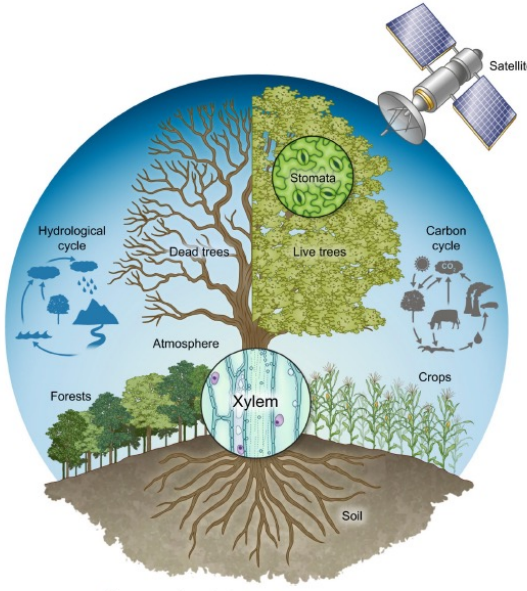
Collaboration among ORNL, LANL, Penn State, et al.

Contact: Forrest Hoffman

Project to prototype machine learning-based parameterizations for stomatal conductance and photosynthesis

- Photosynthesis is a computationally expensive part of land models and leaf-level flux and phenology data is available
- Use combinations of leaf-level and plant hydrodynamics data to build ML models of C_3, C_4, and CAM vegetation
- Investigate ML approaches for scaling to canopies and watersheds
- Prototype hybrid ML-/process-based components within the E3SM Land Model (ELM)
- Future efforts:
  - Conduct regional and global simulations to benchmark different combinations of process-based and ML modules
  - Explore approaches for building hybrid modeling interfaces within ELM

Nature



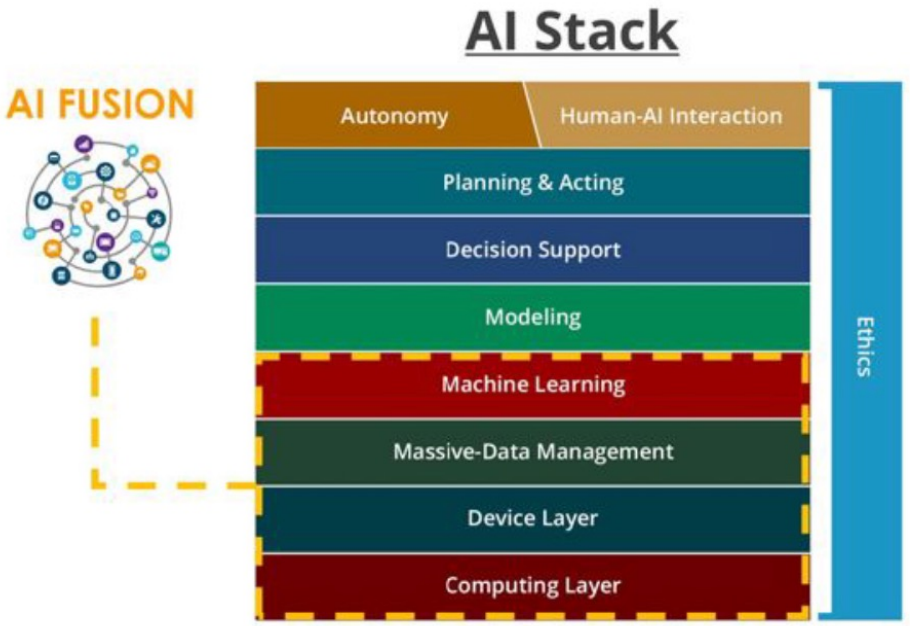
McDowell et al. (2019)

AI4ESP WORKSHOP HIGHLIGHTS

Codesign Is Critical

Codesign advanced computing, software, hybrid ML/physical models, observations and future Earth system modeling capabilities

- Common/consistent language & format
- Merged products (standardization, interoperability)
- Adaptive data & parameter selection
- Computation using large datasets without moving
- Specialized AI/ML code & architecture
- Training and benchmarking datasets and hybrid model design



College of Engineering, Carnegie Mellon University

OPEN CLASSIFICATION OF REGIMES IN THE SOUTHEAST USA (OPENCRUMS-USA).

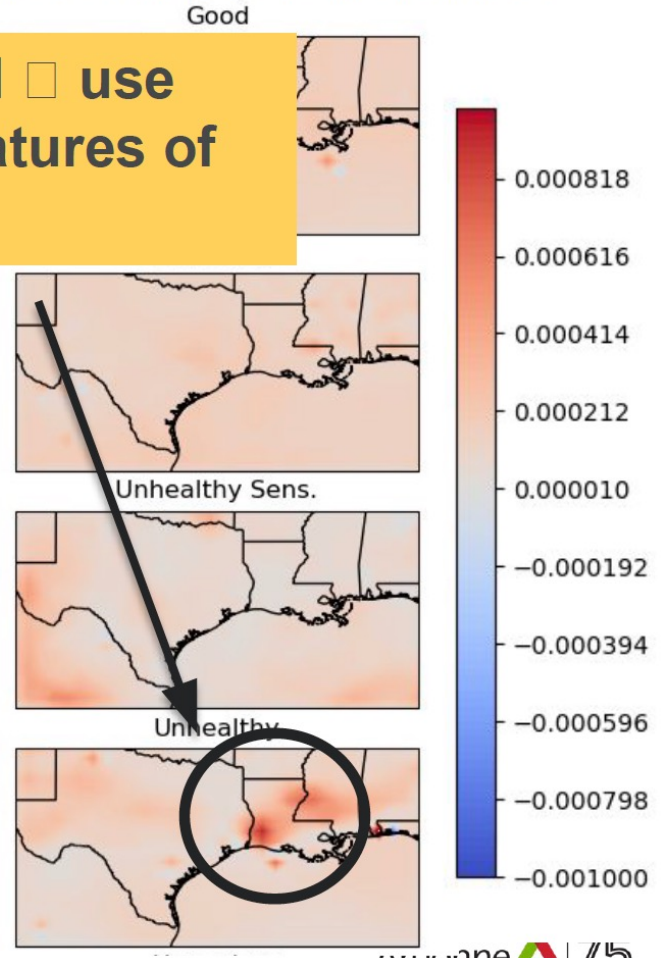
Exploring ML techniques for classifying large climate datasets

- AI4ESP seed project to explore best ML techniques for reducing reanalysis datasets + classifying aerosol/meteorological regimes over TRACER (Houston) + ARM SE USA
- Have tested CNNs and PCA for dimensionality reduction on 10 years of MERRA2 aerosol data
- 17 parallel CNNs show ~97% accuracy of classifying MERRA2 data by EPA AQI over Houston
- Future: explore classification on ERA5 for meteorology + more techniques
- Release all materials as open cookbooks for workforce development

https://github.com/rcjackson/opencrums

rBC mass relevance

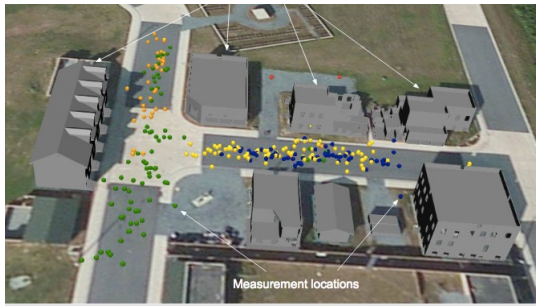
Explainable AI use RLP to see features of interest



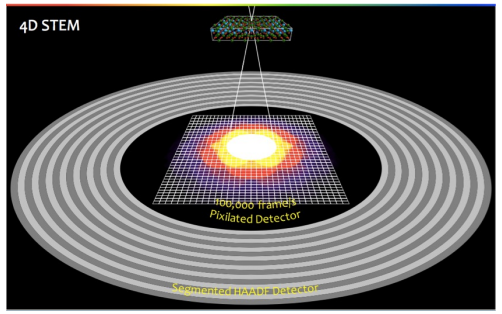
ScienceSearch: Automated Metadata using Machine Learning

Goal: Investigate ML techniques to generate automated metadata that will enable search on data

Use Cases:



LIDAR Sensors



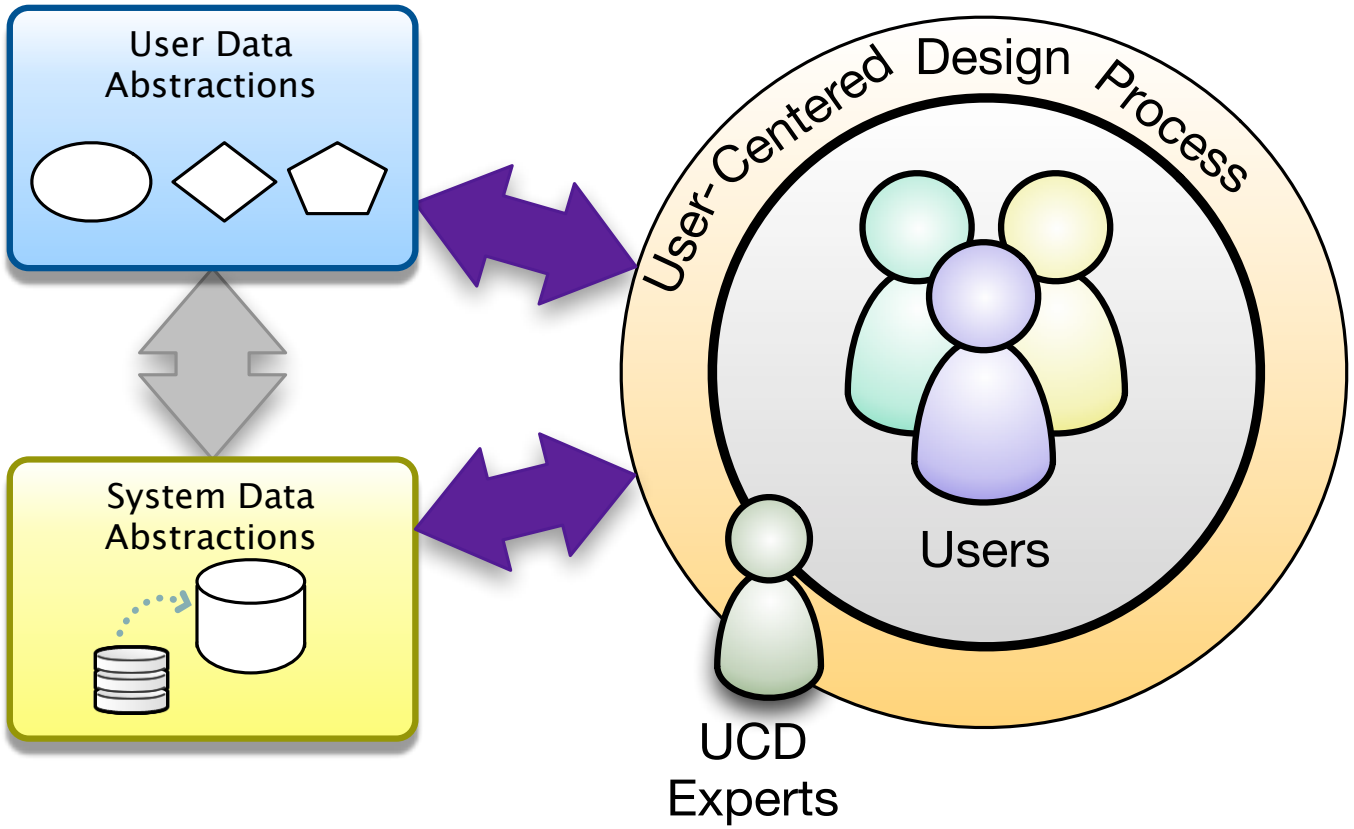
High speed electron detectors

Supercomputing: files and job logs

Urgency: Without robust searching capabilities, ability to fuse multimodal data from instruments, compare results across domains, and reproduce scientific results, remain labor intensive, one-off efforts

Team: Katie Antypas, Lavanya Ramakrishnan, Gunther Weber, Joe M. Hellerstein

Usable Data Abstractions

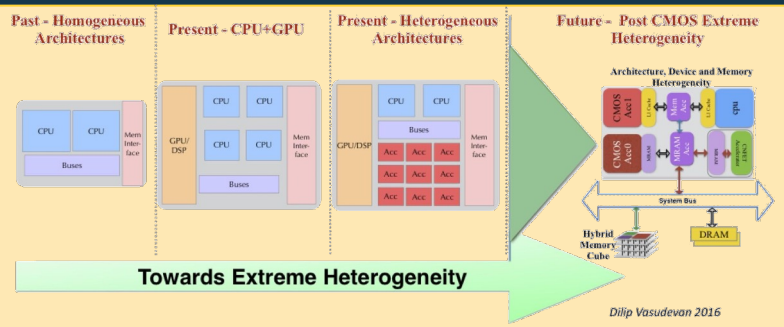


PI: Lavanya Ramakrishnan

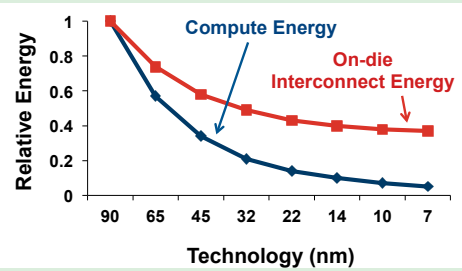
Adapting to HPC Machines of the Future

HPC Architecture and Storage Are Continuing to Morph

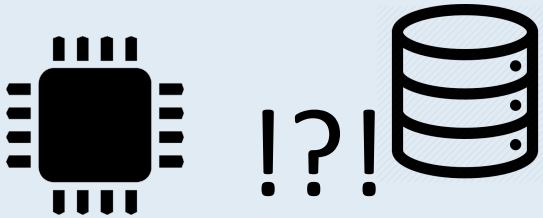
- **Extreme Heterogeneity:**
  - Architecture specialization will lead to diverse modes of acceleration
  - Challenges code generation and software integration



- **Memory Locality**
  - Dominates now and will dominate more if accelerators are effective
  - Heterogeneous memory spaces (each accelerator has its own memory)



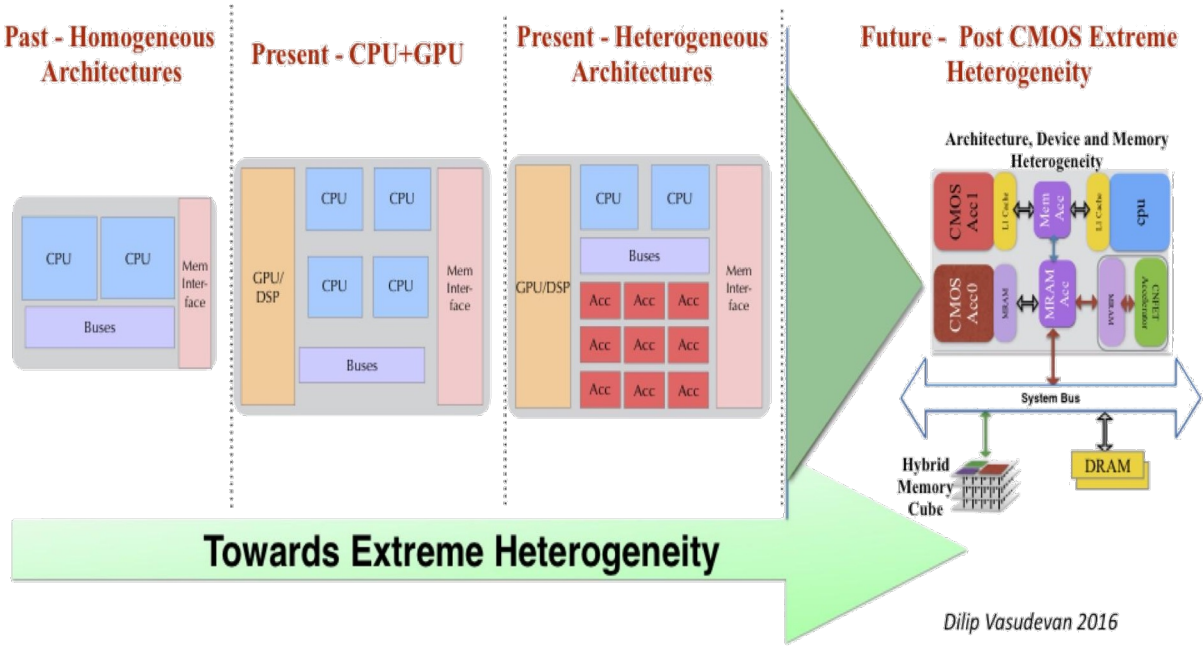
- **Memory/Storage Transformation**
  - Breaks old storage paradigms (based on TapeIO)
  - Storage indistinguishable from memory (and vice-versa)



Extreme Heterogeneity Challenges

Challenges

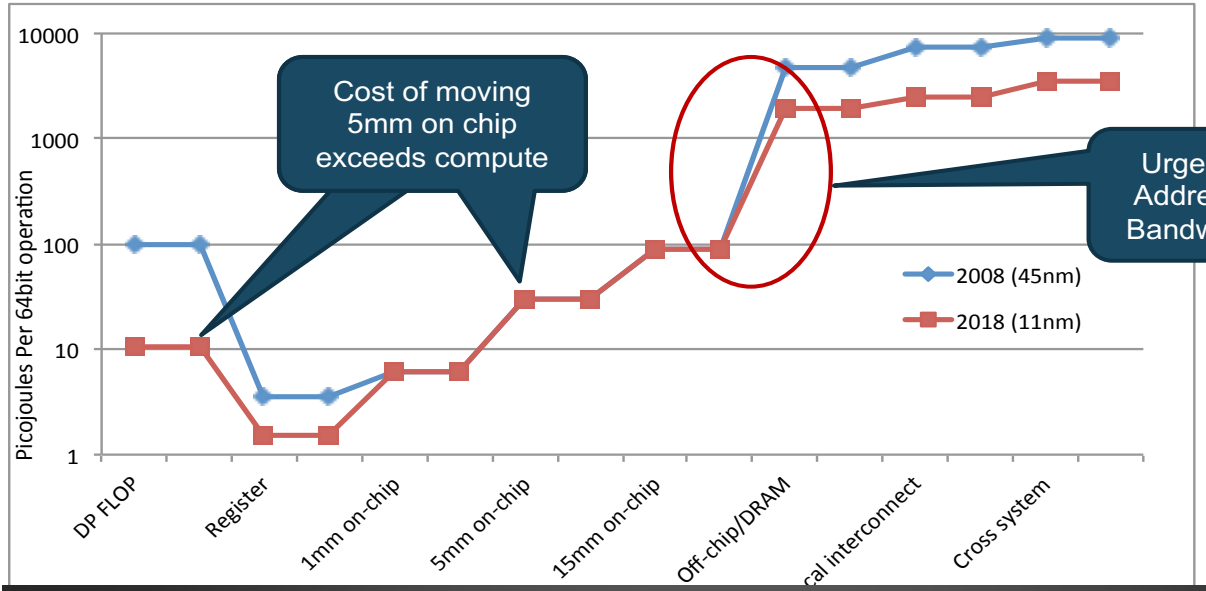
- **Specialization is most immediate path to performance growth post exascale**
- **Trend toward diverse accelerators**

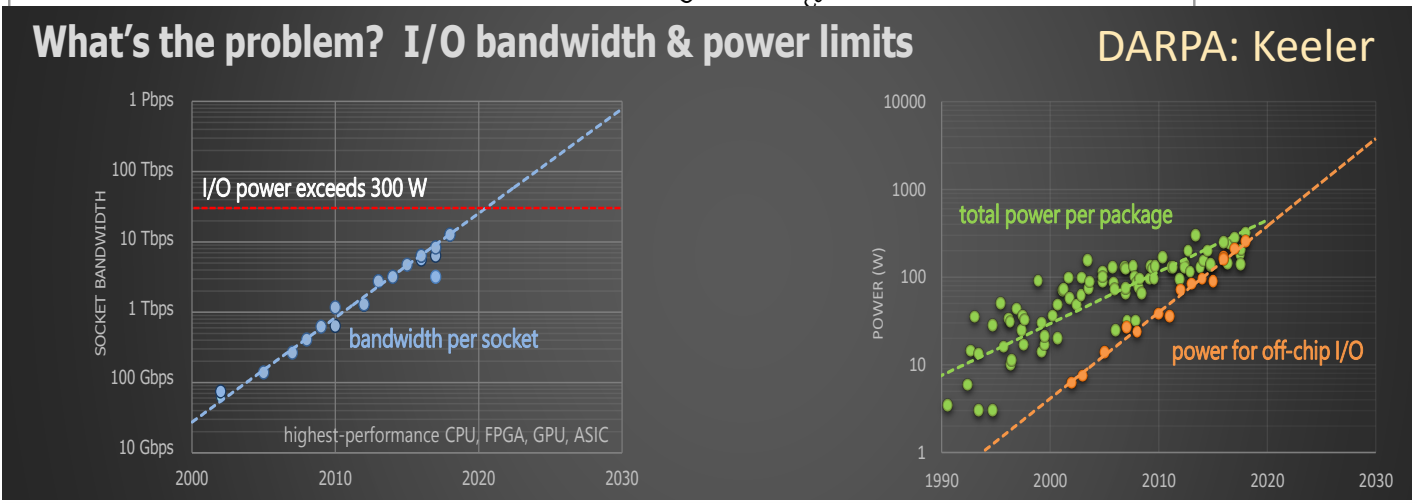


Limits of Current Practice

- **ECP: $10-15M per application to refactor for GPUs & Manycore**
- **We are not prepared to handle** *dozens of heterogeneous integrated accelerators*

Memory Locality Challenge





Challenge: Data Movement Costs More than Compute

- Even worse if accelerators are successful

Limits of Current Practice

- Adding more and more OpenMP/OpenACC directives to manage data movement
- Accept consequences of being BW limited

New Approaches:

1. **Data-centric programming models**
2. **Software to Exploit Rack Disaggregation**
3. **Data Accelerators:** Not just compute accelerators alone (*MSR-Catapult*)

Memory Centric Programming Model: *inverts compute-centric paradigm*

Enables Compiler and RTS to automate common data movement optimizations

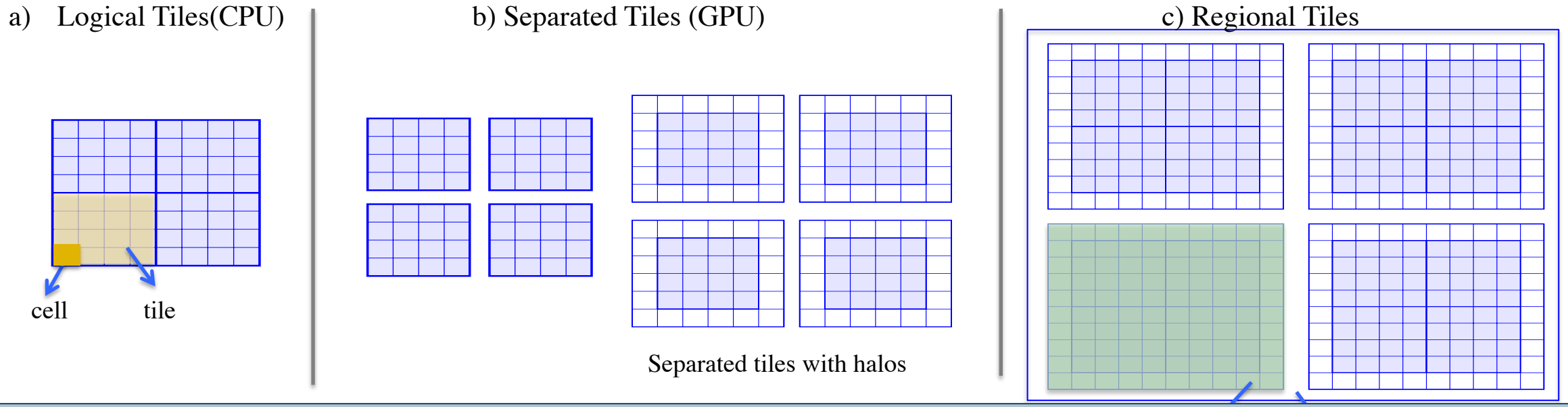
Data Arrays and Loops operate on logical data layout (not physical data layout)

- Arrays are opaque (logical) objects
- Loops operate on logical arrays
- RTS determines optimal physical layout
- Loops/Iterators match layout

Data-localized Parallel Loop *(bind to data layout)*

```
Allocate3DArray(tiledA);  // runtime system chooses tiling pattern
do tileno=1, ntiles (tiledA)
     do j=lo(2), hi(2)
       do i=lo(1), hi(1)
          tiledA(i,j)= ...
       end do
   end do
end do
```

And if layout changes, all loops in code don't need rewrite!



Currently implemented as TiDA Design Pattern for AMReX, but can elevate to programming model for broader impact!

Adapted from IRA presentation by John Shalf

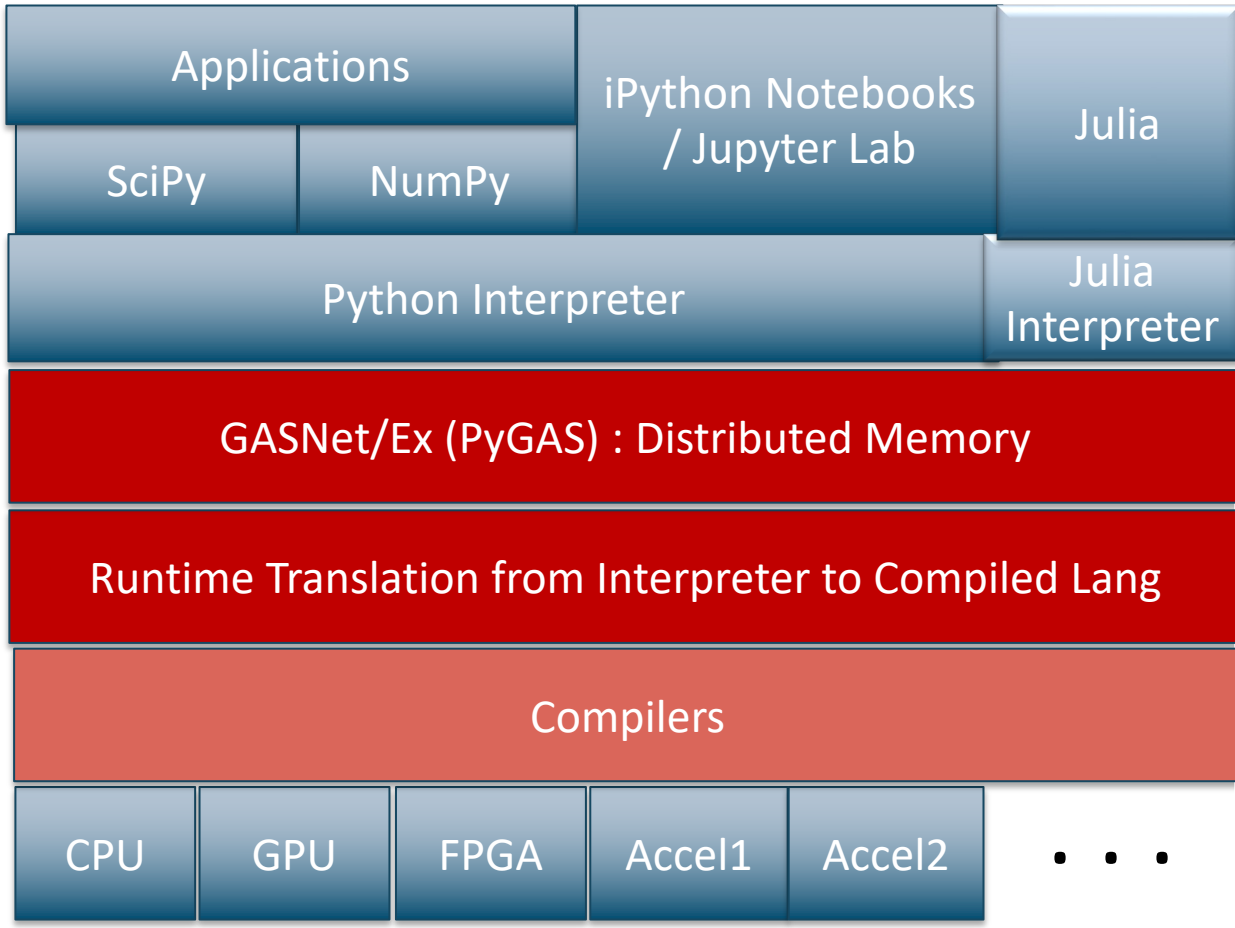
Top-Down: Embracing and Accelerating the Productivity Stack for Data
(user facing programming interfaces)

- **Challenge:** **Productivity languages (Python, Julia) are rapidly gaining traction for numerical analysis in the sciences (e.g. SciPy and NumPy).**

- **Limits of Current Practice**: Calls C++ Libraries for speed
  - Need to develop robust communication & parallelism
  - Python interpreter ~1000x slower than compiled language

- **Research Agenda:** **Python extensions for more efficient distributed memory and Acceleration**
  - Extend PGAS to Python and develop distributed array abstractions for HPC Scalability (PyGAS)
  - Develop code-generation capabilities for diverse accelerators (EH)
  - Enable Python code to be compiled dynamically to target (SEJITS)

- **Leverages** **existing and past investments in GASNet/Ex, PyGAS, SEJITS, iPython, and Jupyter**.



Enabling Science: A new data programing model on arrays

Inspiring by: a Tensor can be defined as a multidimensional array and proper transformation rules

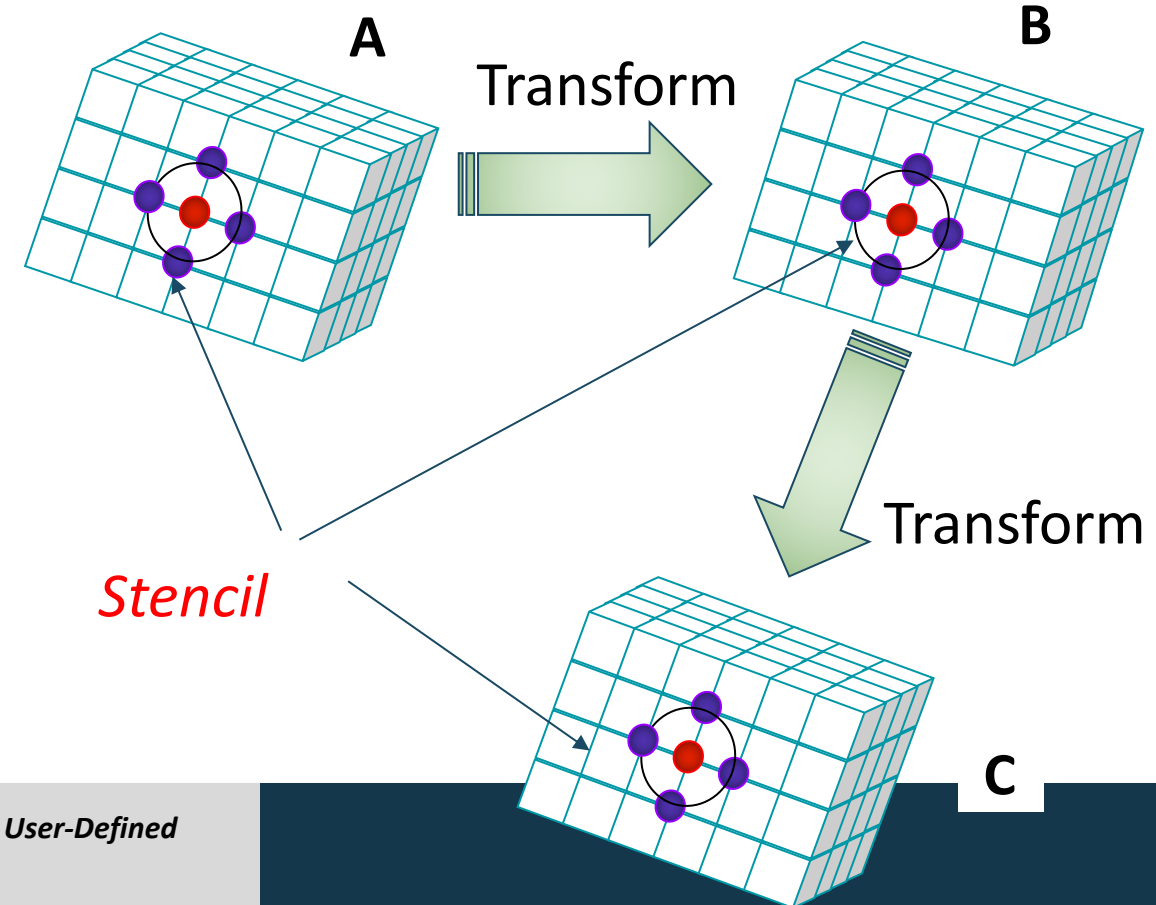
1, An *Array*-native Data Programming Model namelv SLOPE

2, A *Stencil*-based Abstract Data Type

3, A single *Transform* operation

4, HPC friendly

- MPI based Single Program Multiple Data (SPMD) Pattern
- Directly on scientific data formats, e.g., HDF5, ADIOS, PNetCDF
- Manual/auto-chunking & ghost zone management
- Distributed array cache
- Async ghost zone exchange
- Check-point
- In-place modification semantic
- Multi-arrays supports



Bin Dong, Kesheng Wu, Surendra Byna, Jialin Liu, Weijie Zhao, Florin Rusu, ***"ArrayUDF: User-Defined Scientific Data Analysis on Arrays",*** *HPDC 2017*

Conclusions

- The observation data for climate is challenging to manage and to integrate
- Combining ML and models is becoming more common
- Digital twins for the climate models are beginning to glimmer on the horizon
- Self-driving observation systems are coming
- We are reaching the practical scaling limit of the PDE/ODE approach to modeling the the climate
- There are many areas of data management where new research could significantly improve the easy of doing science
- Don’t forget the user – no matter how cool it is - - - if the user can’t use it, it is not useful