U.S. DEPARTMENT OF ENERGY
Office of Science

# B16D-01 - Exploiting Artificial Intelligence for Advancing Earth and Environmental System Science

*Forrest M. Hoffman[1,2], Jitendra Kumar[1], Zachary L. Langford[1], V. Shashank Konduri[3], Auroop R. Ganguly[4], Cheng-En Yang[2], Nathan Collier[1], Min Xu[1], William W. Hargrove[5], Nicki L. Hickmon[6], Scott M. Collis[6], Charuleka Varadharajan[7], and Haruko Wainwright[7]*

*December 12, 2022*

[1]Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2]University of Tennessee, Knoxville, TN, USA
[3]National Ecological Observatory Network, Boulder, CO, USA
[4]Northeastern University, Boston, MA, USA
[5]US Department of Agriculture - Forest Service, Asheville, NC, USA
[6]Argonne National Laboratory, Lemont, IL, USA
[7]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

# Introduction

- Observations of the Earth system are increasing in spatial resolution and temporal frequency, and will grow exponentially over the next 5–10 years
- With Exascale computing, simulation output is growing even faster, outpacing our ability to analyze, interpret and evaluate model results
- Explosive data growth and the promise of discovery through data-driven modeling necessitate new methods for feature extraction, change/anomaly detection, data assimilation, simulation, and analysis


*Frontier at Oak Ridge National Laboratory is the #1 fastest supercomputer on the TOP500 List and the first supercomputer to break the exaflop barrier (May 30, 2022).*

*This article is the second in a two-part series. The first part, "How to Build a Hypercomputer," by Thomas Sterling, appeared in the July 2001 issue.*

## The Do-It-Yourself

# Supercomputer

By William W. Hargrove, Forrest M. Hoffman and Thomas Sterling

Photographs by Kay Chernush

**Scientists have found a cheaper way to solve tremendously difficult computational problems: connect ordinary PCs so that they can work together**

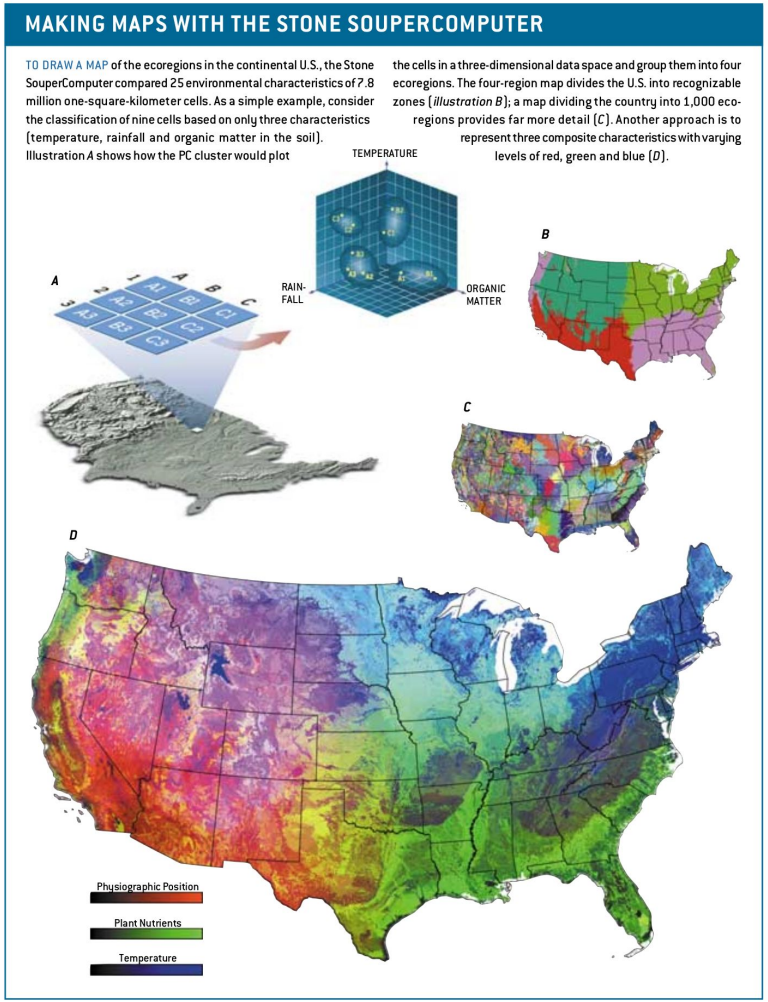
CLUSTER OF PCs at the Oak Ridge National Laboratory in Tennessee has been dubbed the Stone SouperComputer.



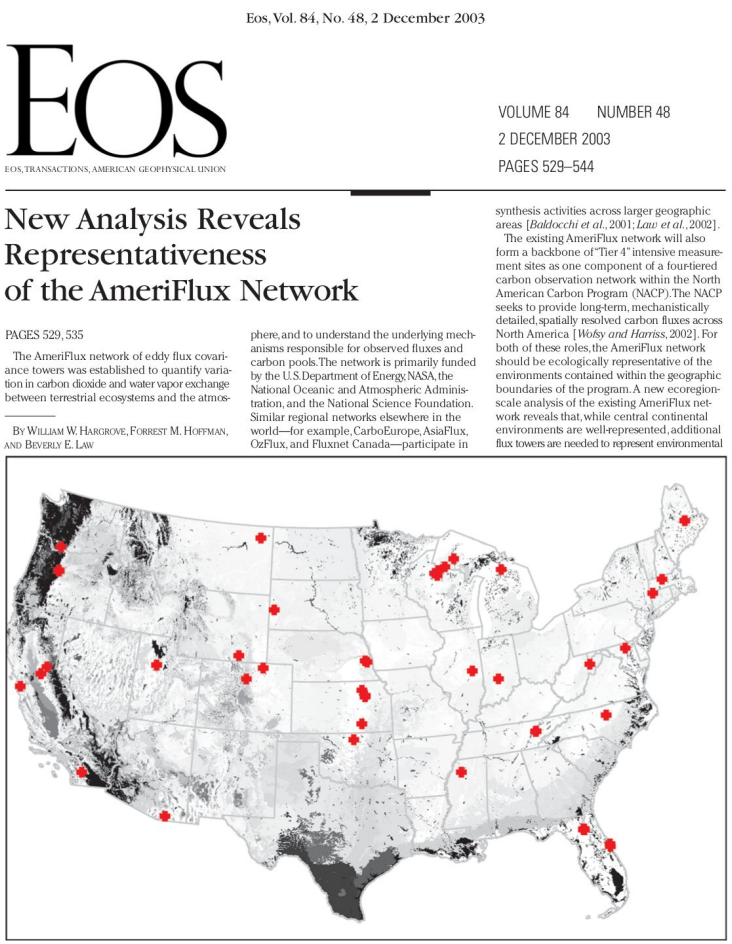
Copyright 2001 Scientific American, Inc.

# Multivariate Geographic Clustering

- Ecoregions have traditionally been created by experts
- Our approach has been to objectively create ecoregions using continuous continental-scale data and clustering
- We developed a highly scalable *k*-means cluster analysis code that uses distributed memory parallelism
- Originally developed on a 486/Pentium cluster, the code now runs on the largest hybrid CPU/GPU architectures on Earth




Hargrove, W. W., F. M. Hoffman, and T. Sterling (2001), The Do-It-Yourself Supercomputer, *Sci. Am.*, 265(2):72–79, https://www.scientificamerican.com/article/the-do-it-yourself-superc/



Eos, Vol. 84, No. 48, 2 December 2003

EOS

EOS, TRANSACTIONS, AMERICAN GEOPHYSICAL UNION

VOLUME 84 NUMBER 48
2 DECEMBER 2003
PAGES 529–544

New Analysis Reveals Representativeness of the AmeriFlux Network

PAGES 529, 535

The AmeriFlux network of eddy flux covariance towers was established to quantify variation in carbon dioxide and water vapor exchange between terrestrial ecosystems and the atmosphere, and to understand the underlying mechanisms responsible for observed fluxes and carbon pools. The network is primarily funded by the U.S. Department of Energy, NASA, the National Oceanic and Atmospheric Administration, and the National Science Foundation. Similar regional networks elsewhere in the world—for example, CarboEurope, AsiaFlux, OzFlux, and Fluxnet Canada—participate in synthesis activities across larger geographic areas [*Baldocchi et al.*, 2001; *Law et al.*, 2002].

The existing AmeriFlux network will also form a backbone of "Tier 4" intensive measurement sites as one component of a four-tiered carbon observation network within the North American Carbon Program (NACP). The NACP seeks to provide long-term, mechanistically detailed, spatially resolved carbon fluxes across North America [*Wofsy and Harriss*, 2002]. For both of these roles, the AmeriFlux network should be ecologically representative of the environments contained within the geographic boundaries of the program. A new ecoregion-scale analysis of the existing AmeriFlux network reveals that, while central continental environments are well-represented, additional flux towers are needed to represent environmental

By William W. Hargrove, Forrest M. Hoffman, and Beverly E. Law

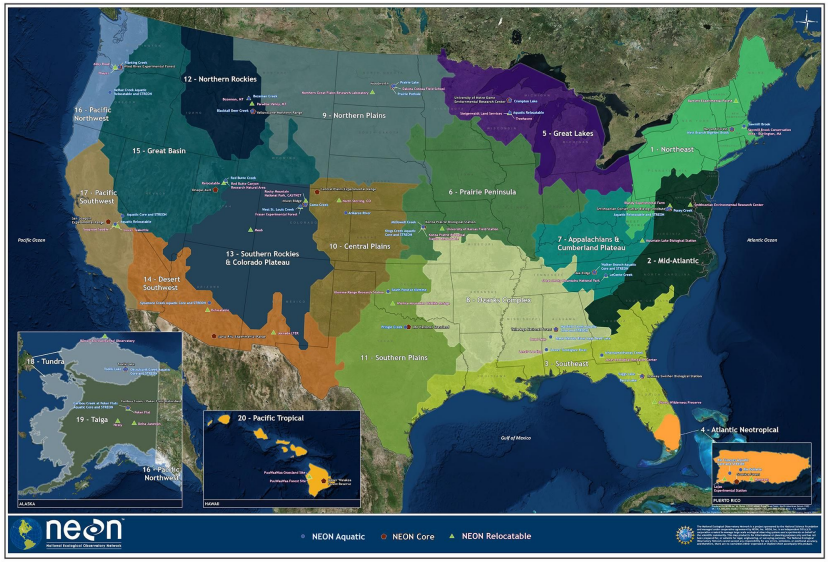
*Fig. 1. The representativeness of an existing spatial array of sample locations or study sites—for example, the AmeriFlux network of carbon dioxide eddy flux covariance towers—can be mapped relative to a set of quantitative ecoregions, suggesting locations for additional samples or sites. Distance in data space to the closest ecoregion containing a site quantifies how well an existing network represents each ecoregion in the map. Environments in darker ecoregions are poorly represented by this network.*

# Network Representativeness

- The *n*-dimensional space formed by the data layers offers a natural framework for estimating representativeness of individual sampling sites
- The Euclidean distance between individual sites in data space is a metric of similarity or dissimilarity
- Representativeness across multiple sampling sites can be combined to produce a map of network representativeness

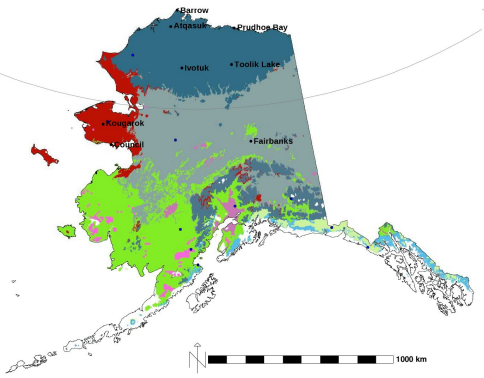
Hargrove, W. W., and F. M. Hoffman (2003), New Analysis Reveals Representativeness of the AmeriFlux Network, *Eos Trans. AGU*, 84(48):529, 535, doi:10.1029/2003EO480001.

# Sampling Network Design

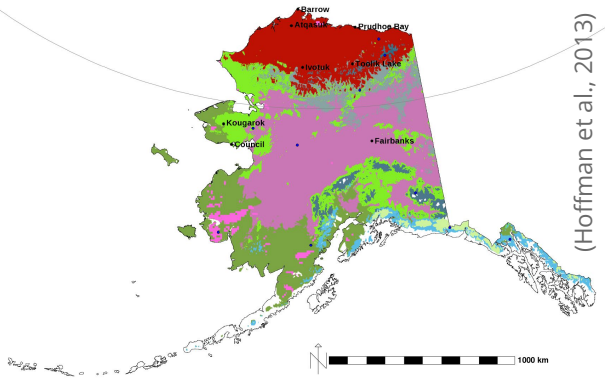



NSF's NEON Sampling Domains

*Gridded data from satellite and airborne remote sensing, models, and synthesis products can be combined to design optimal sampling networks and understand representativeness as it evolves through time*



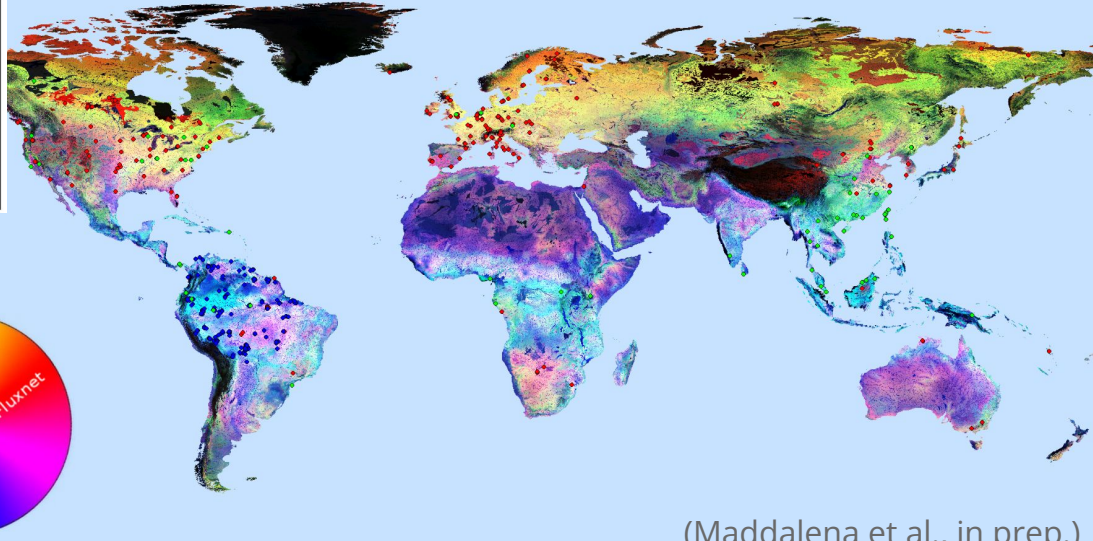

2000–2009




2090–2000

(Hoffman et al., 2013)

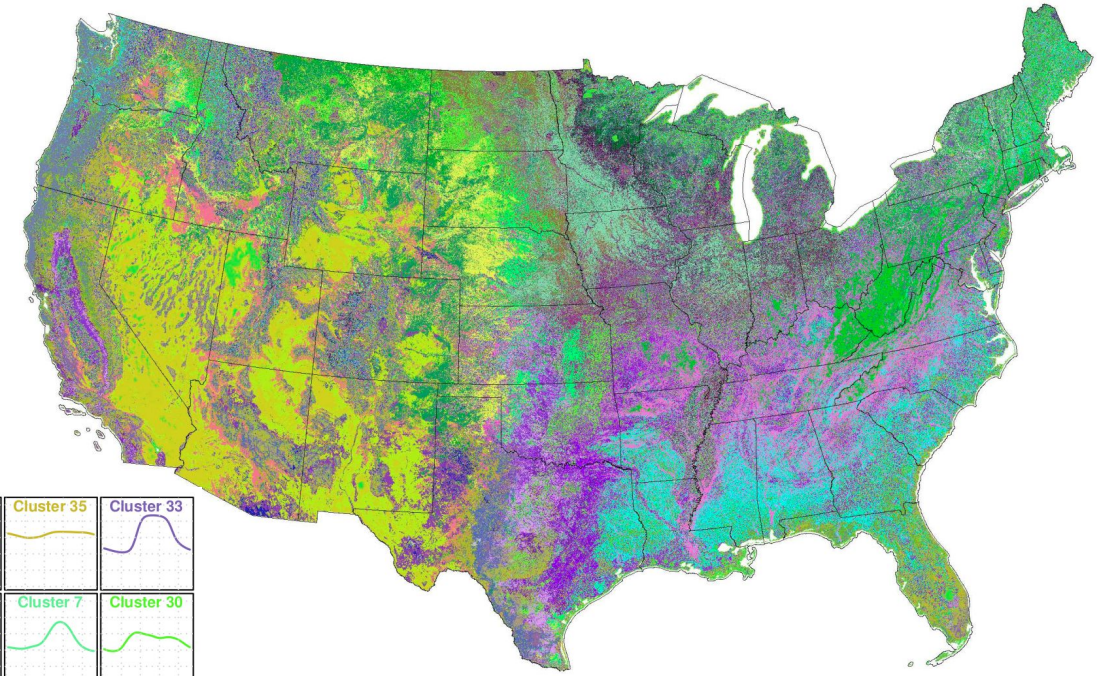
Triple-Network Global Representativeness

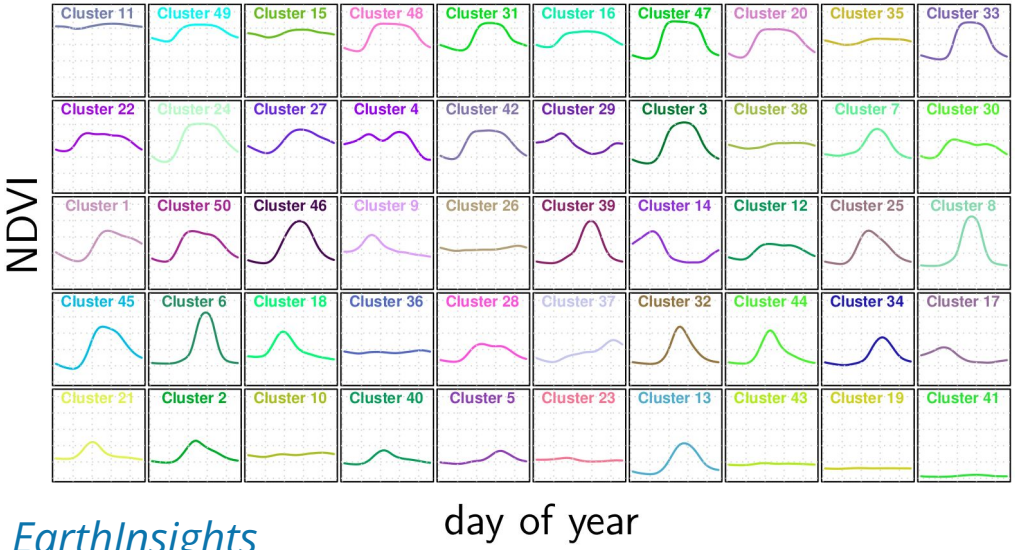



(Maddalena et al., in prep.)

# 50 Phenoregions for year 2012 (Random Colors)

250m MODIS NDVI
Every 8 days (46 images/year)
Clustered from year 2000 to present



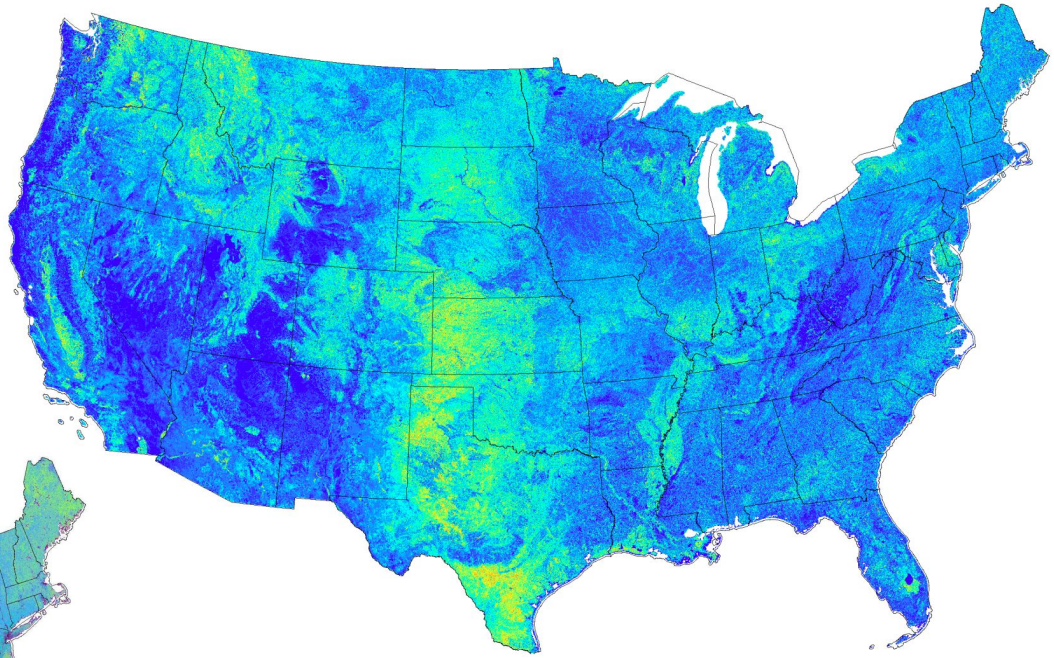


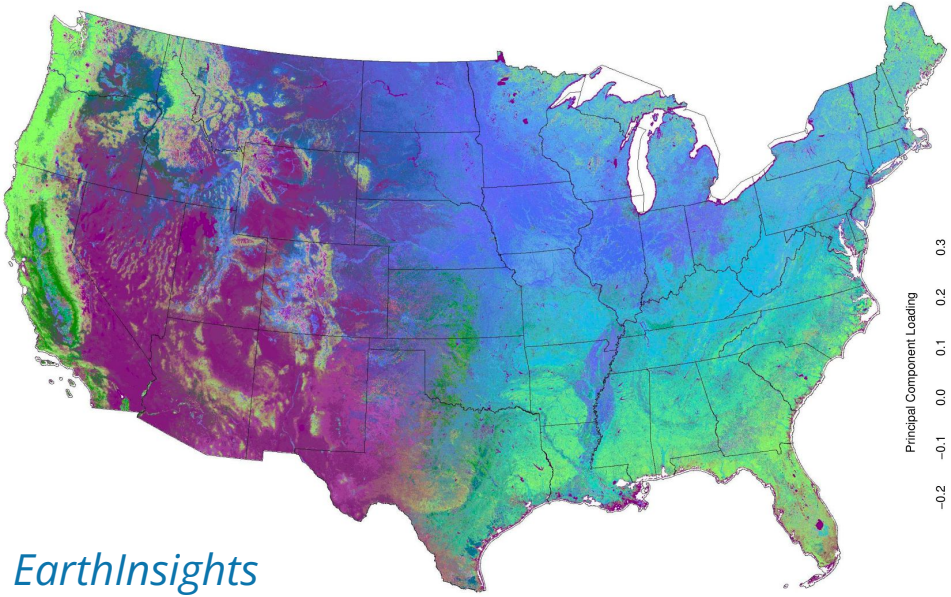

# 50 Phenoregion Prototypes (Random Colors)

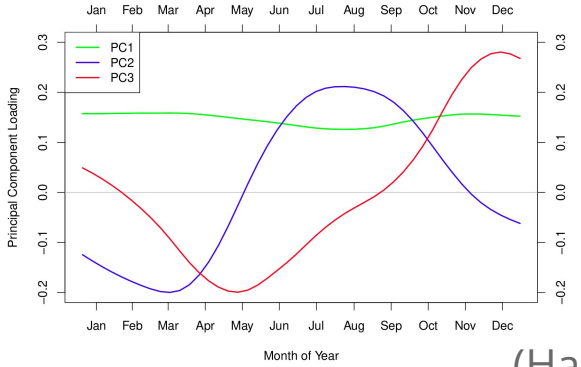
*EarthInsights*

(Hargrove et al., in prep.)

# 50 Phenoregions Persistence and 50 Phenoregions Max Mode (Similarity Colors)








**Principal Components Analysis**

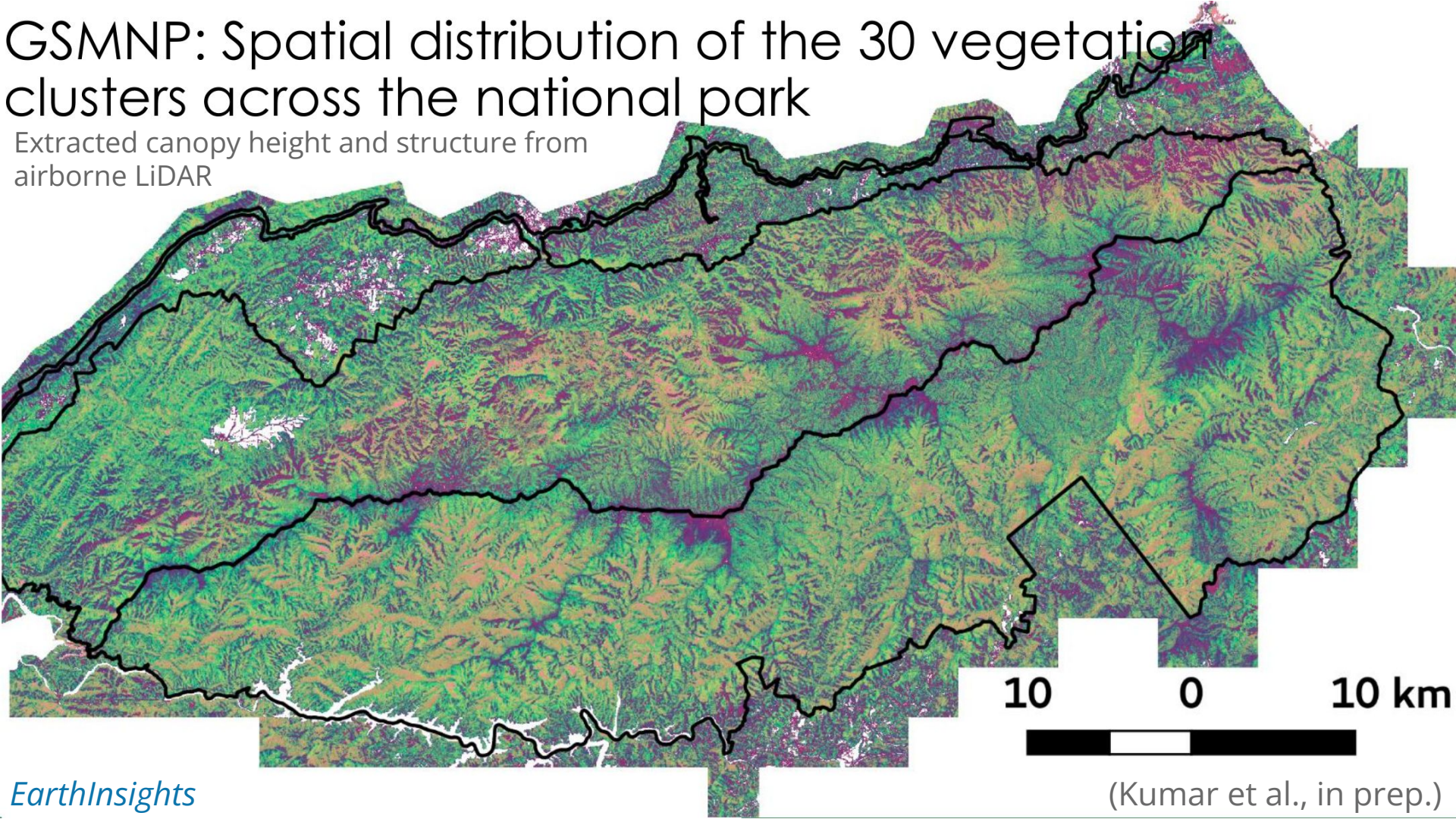
**PC1 ~ Evergreen**
**PC2 ~ Deciduous**
**PC3 ~ Dry Deciduous**

*EarthInsights*

(Hargrove et al., in prep.)

# GSMNP: Spatial distribution of the 30 vegetation clusters across the national park

Extracted canopy height and structure from airborne LiDAR




*EarthInsights*

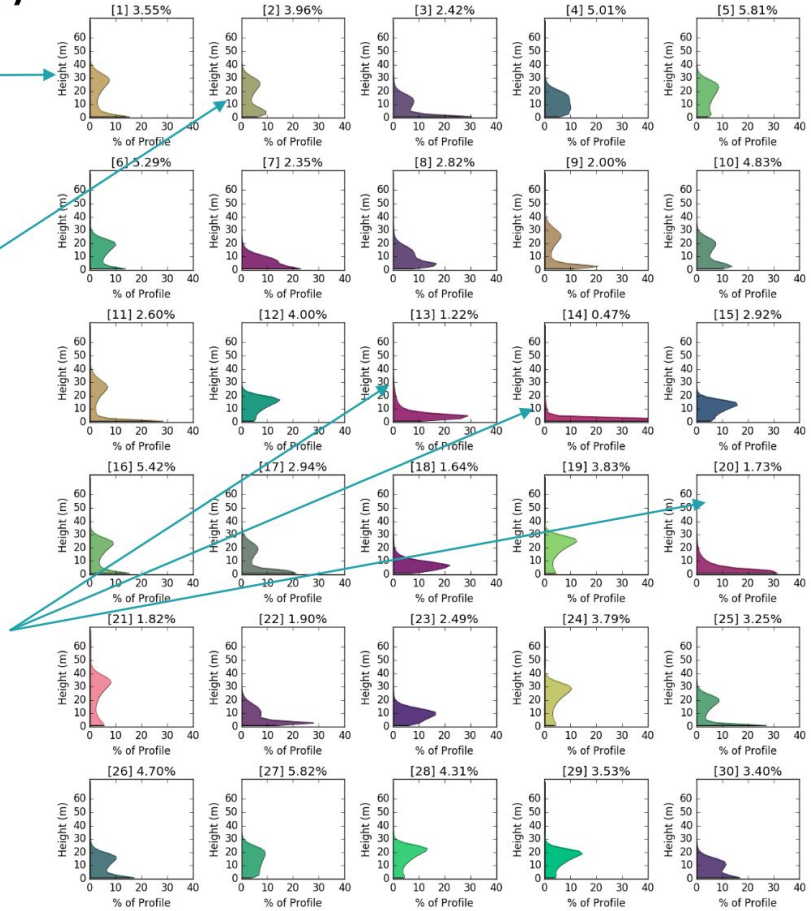
(Kumar et al., in prep.)

# GSMNP: 30 representative vertical structures (cluster centroids) identified

tall forests with low understory vegetation

forests with slightly lower mean height with dense understory vegetation

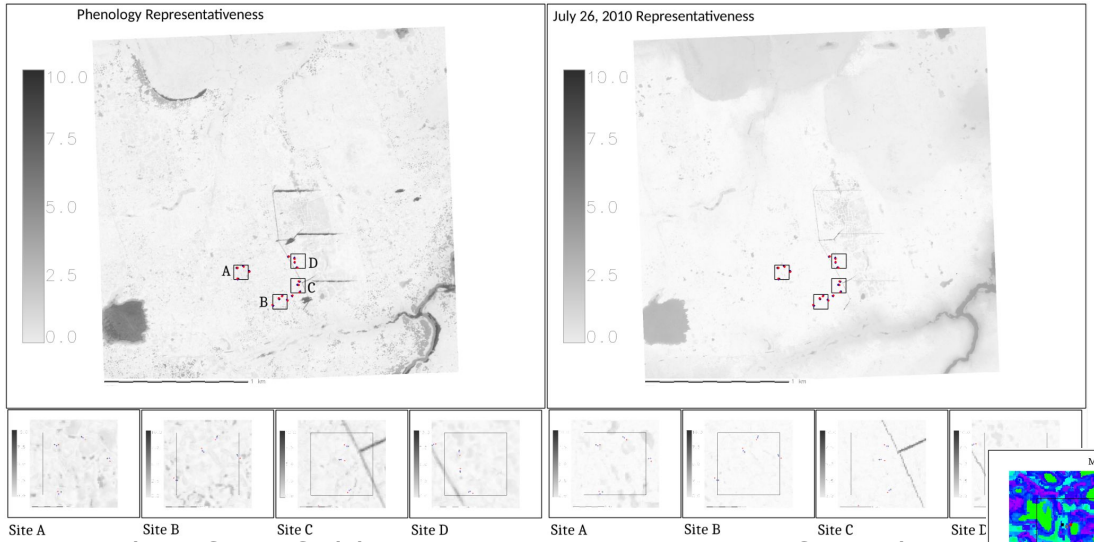
low height grasslands and heath balds that are small in area but distinct landscape type





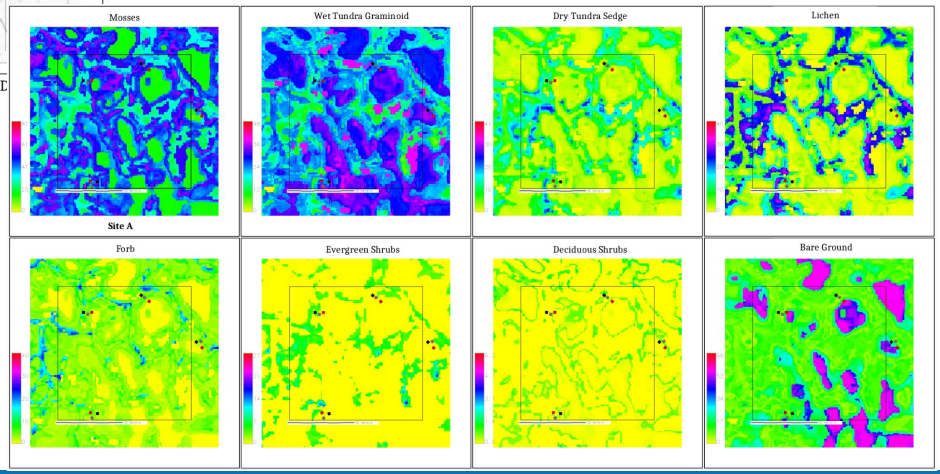

(Kumar et al., in prep.)

# Vegetation Distribution at Barrow Environmental Observatory




Representativeness map for vegetation sampling points in sites A, B, C, and D with phenology (left) and without (right) from WorldView2 multispectral imagery for the year 2010 and LiDAR data

Example plant functional type (PFT) distributions scaled up from vegetation sampling locations

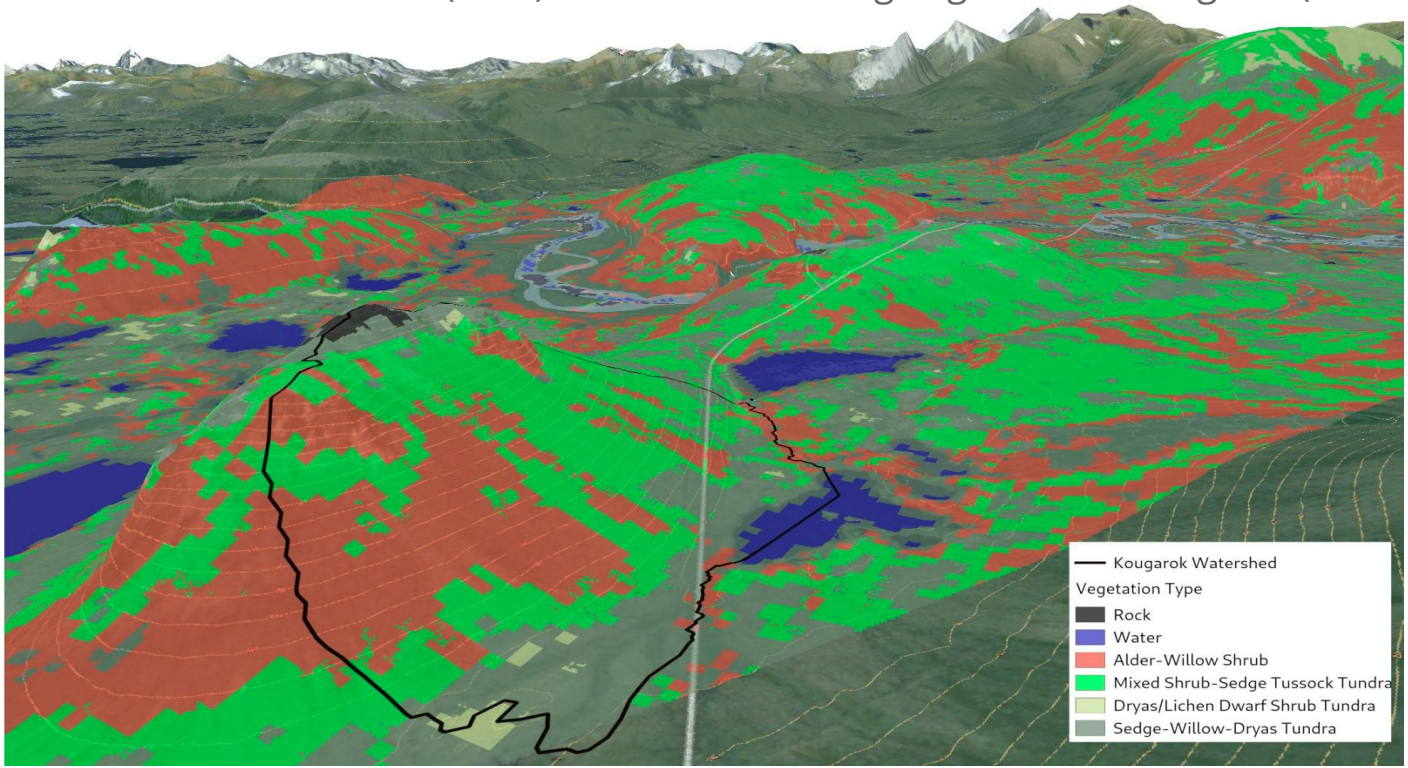



*In situ data from field measurement activities inform the development of wide-scale maps of vegetation distribution through inference using remote sensing data as surrogate variables, and relationships with environmental controls can be extracted*

Langford, Z. L., et al. (2016), Mapping Arctic Plant Functional Type Distributions in the Barrow Environmental Observatory Using WorldView-2 and LiDAR Datasets, *Remote Sens.*, 8(9):733, doi:10.3390/rs8090733.

# Arctic Vegetation Mapping from Multi-Sensor Fusion

Used Hyperion Multispectral and IfSAR-derived Digital Elevation Model, applied cluster analysis, and trained a convolutional neural network (CNN) with Alaska Existing Vegetation Ecoregions (AKEVT)




Langford, Z. L., et al. (2019), Arctic Vegetation Mapping Using Unsupervised Training Datasets and Convolutional Neural Networks, *Remote Sens.*, 11(1):69, doi:10.3390/rs11010069.

# Satellite Data Analytics Enables Within-Season Crop Identification

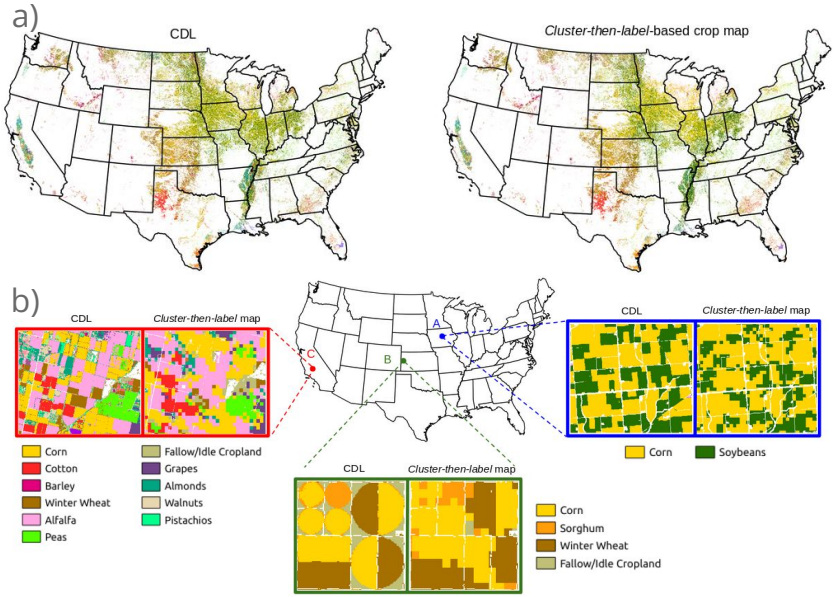
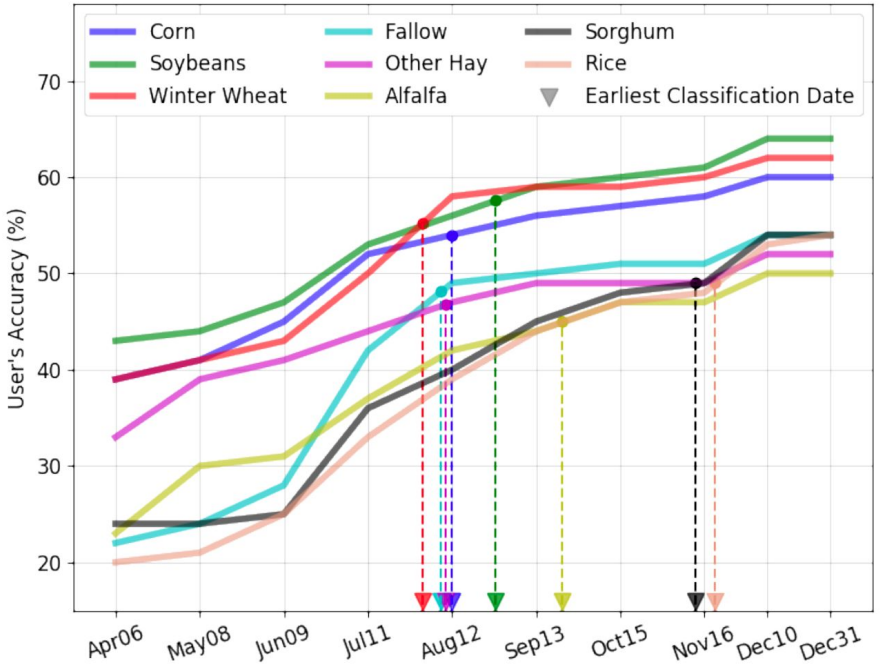


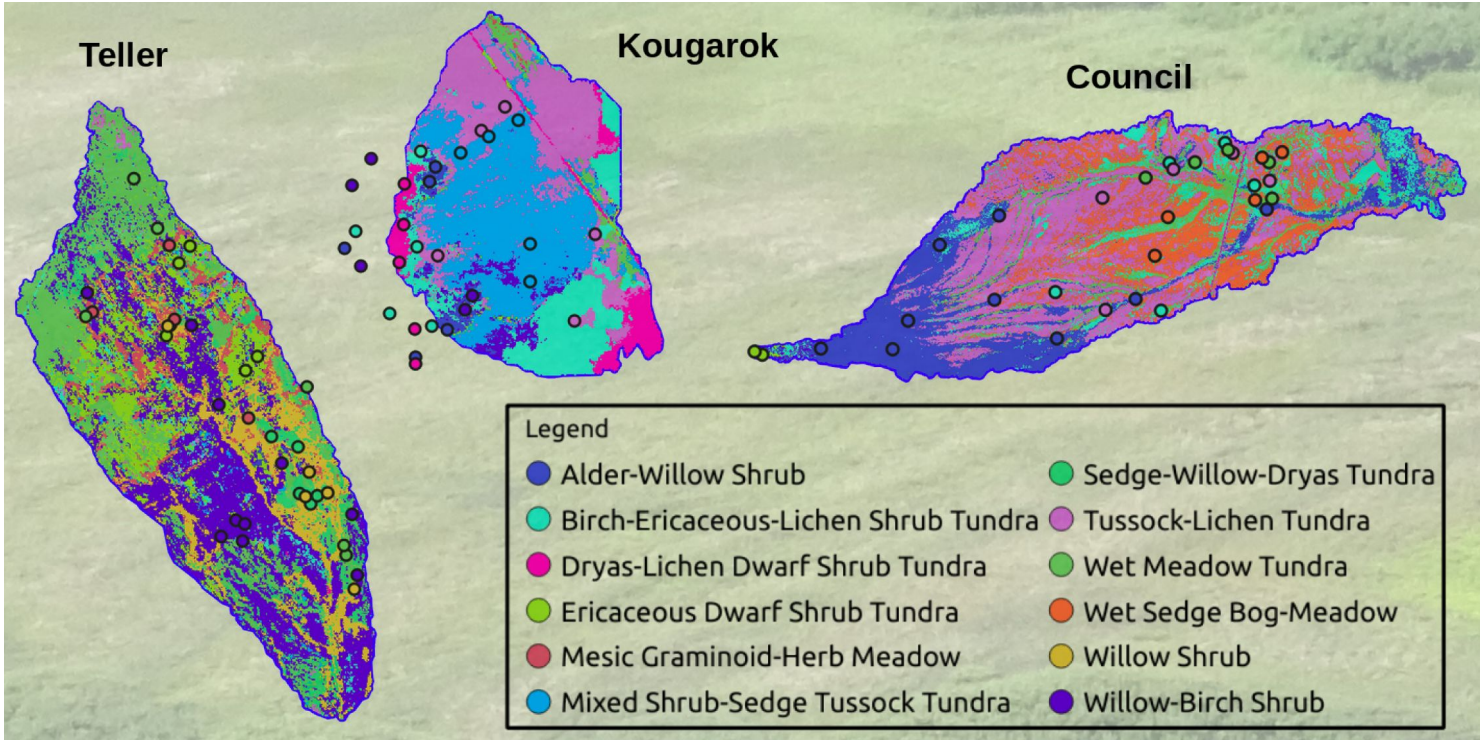

Figure: a) Comparison of cluster-then-label crop map with USDA Crop Data Layer (CDL) shows similar patterns at continental scale. b) Good spatial agreement is found at three selected regions, but cluster-then-label crop maps lack sharpness at field boundaries due to coarser resolution of MODIS data.

## Earliest date for crop type classification




Konduri, V. S., J. Kumar, W. W. Hargrove, F. M. Hoffman, and A. R. Ganguly (2020), Mapping Crops Within the Growing Season Across the United States, *Remote Sens. Environ.*, 251, 112048, doi:10.1016/j.rse.2020.112048.

# Watershed-Scale Plant Communities Determined from DNN and AVIRIS-NG



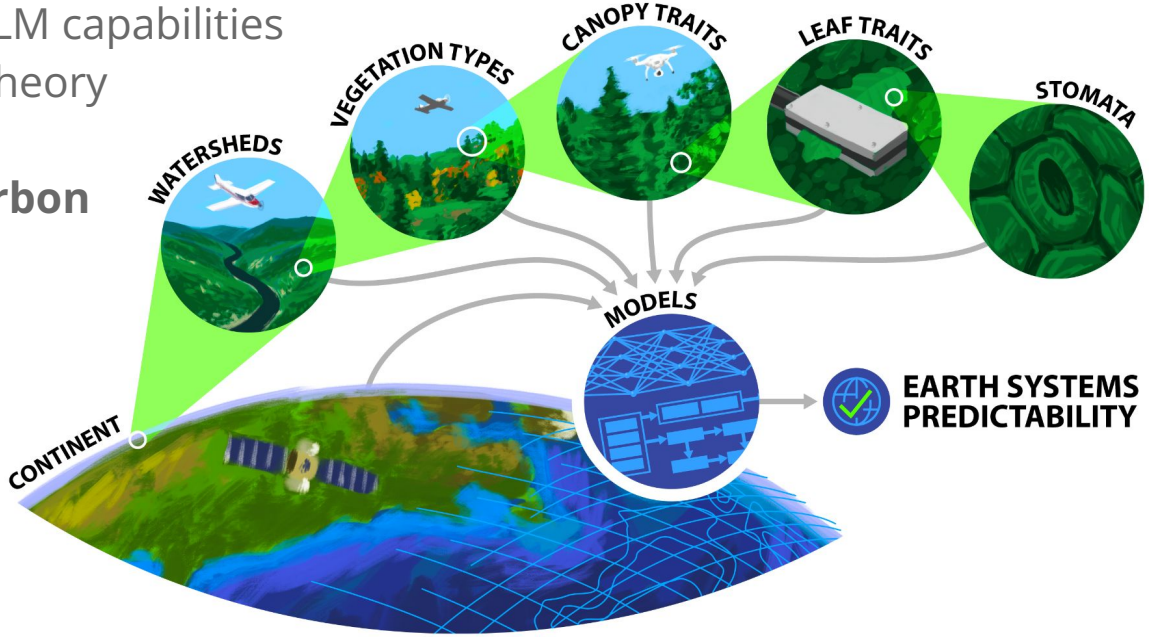

*At the watershed scale, vegetation community distribution follows topographic and water controls. At a fine scale, nutrients limit the distribution of vegetation types.*



(Konduri et al., in prep.)

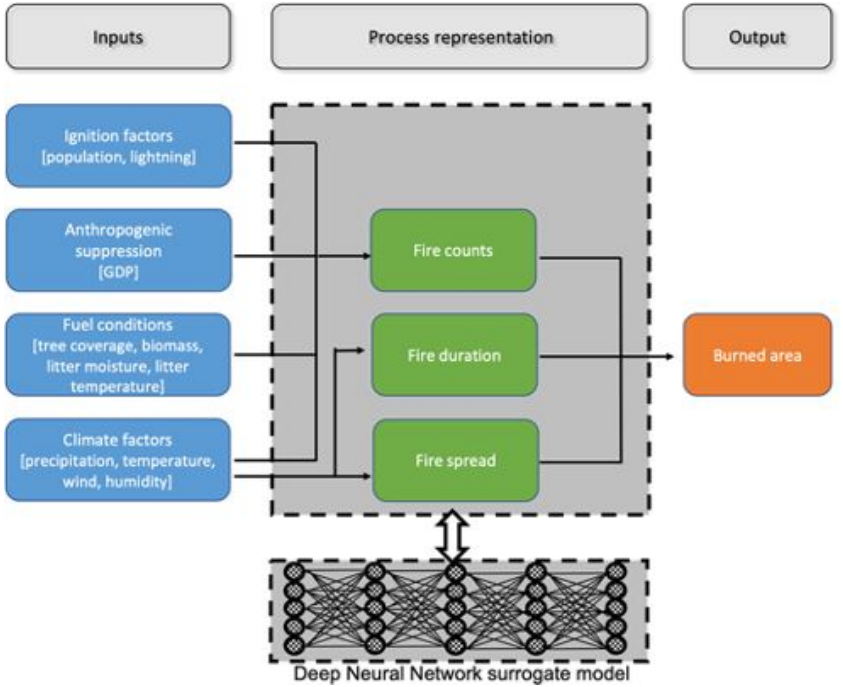
# Machine Learning for Understanding Biospheric Processes

- Widening adoption of deep neural networks and growth of climate data are fueling interest in AI/ML for use in weather and climate and Earth system models
- ML potential is high for improving predictability when (1) *sufficient data are available for process representations* and (2) *process representations are computationally expensive*
- Example methods for improving ELM capabilities by exploring ML and information theory approaches:
  - **Soil organic carbon & radiocarbon**
  - **Wildfire**
  - **Methane emissions**
  - **Ecohydrology**
- All of these applications involve unresolved, subgrid-scale processes that strongly influence results at the largest scales



# Hybrid Modeling of Wildfire Activities

- Improve model simulations of **wildfire processes**, including ignition, fire duration, and spread rate with Deep Neural Network models
- Improve simulated **wildfire emissions** and their impacts on atmospheric properties, including aerosols, greenhouse gases, phosphorus transport, and pollutants
- Improve the projection of near-future and long-term dynamics of wildfire activities
- Accelerate E3SM coupled land–atmosphere modeling activities for wildfire research
- Explore online ML training/validation strategy for E3SM coupled model simulations

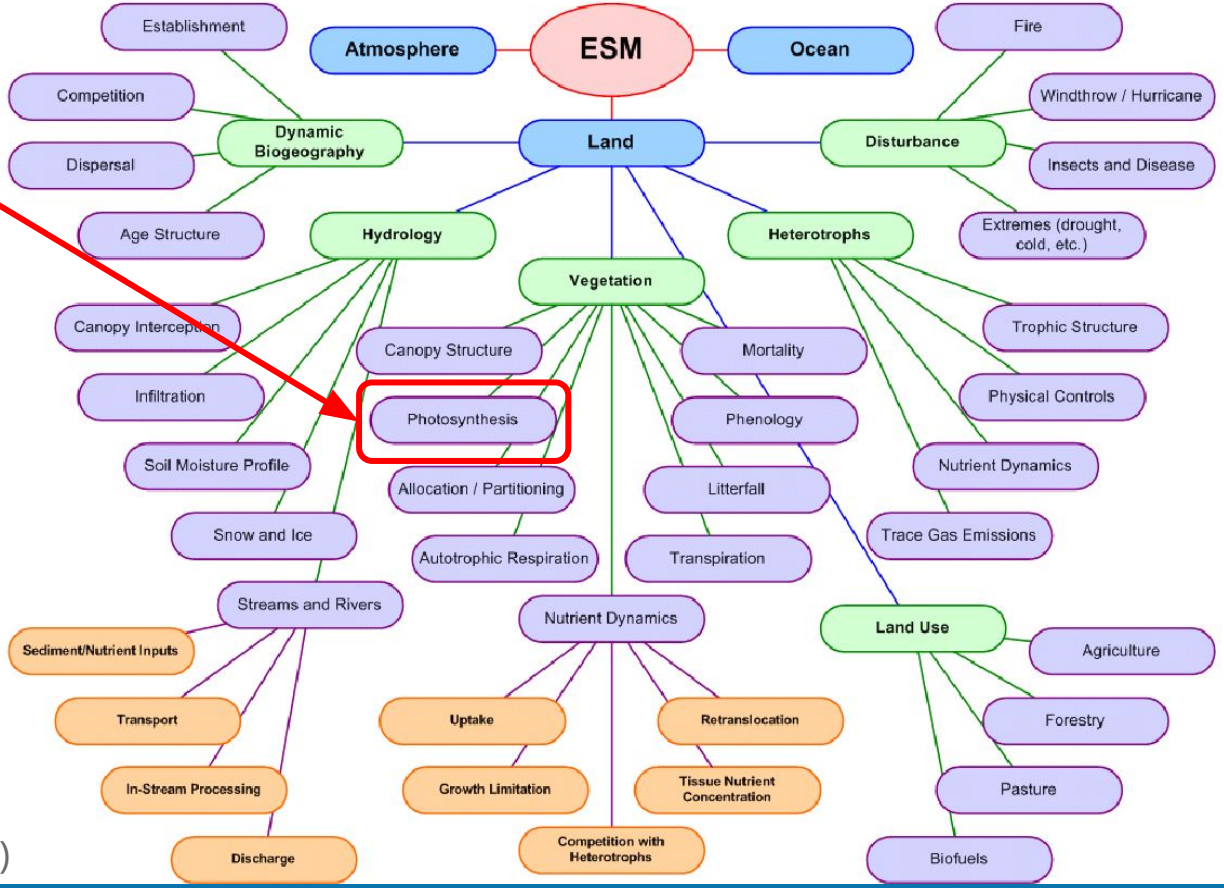



Zhu et al. (2022)

# Hybrid ML/Process-based Modeling for Terrestrial Modeling

In the hierarchy of land model processes, we start with the **photosynthesis** parameterization because

- Multiple hypotheses
- Many leaf-level measurements
- Most computationally intensive part of the land model

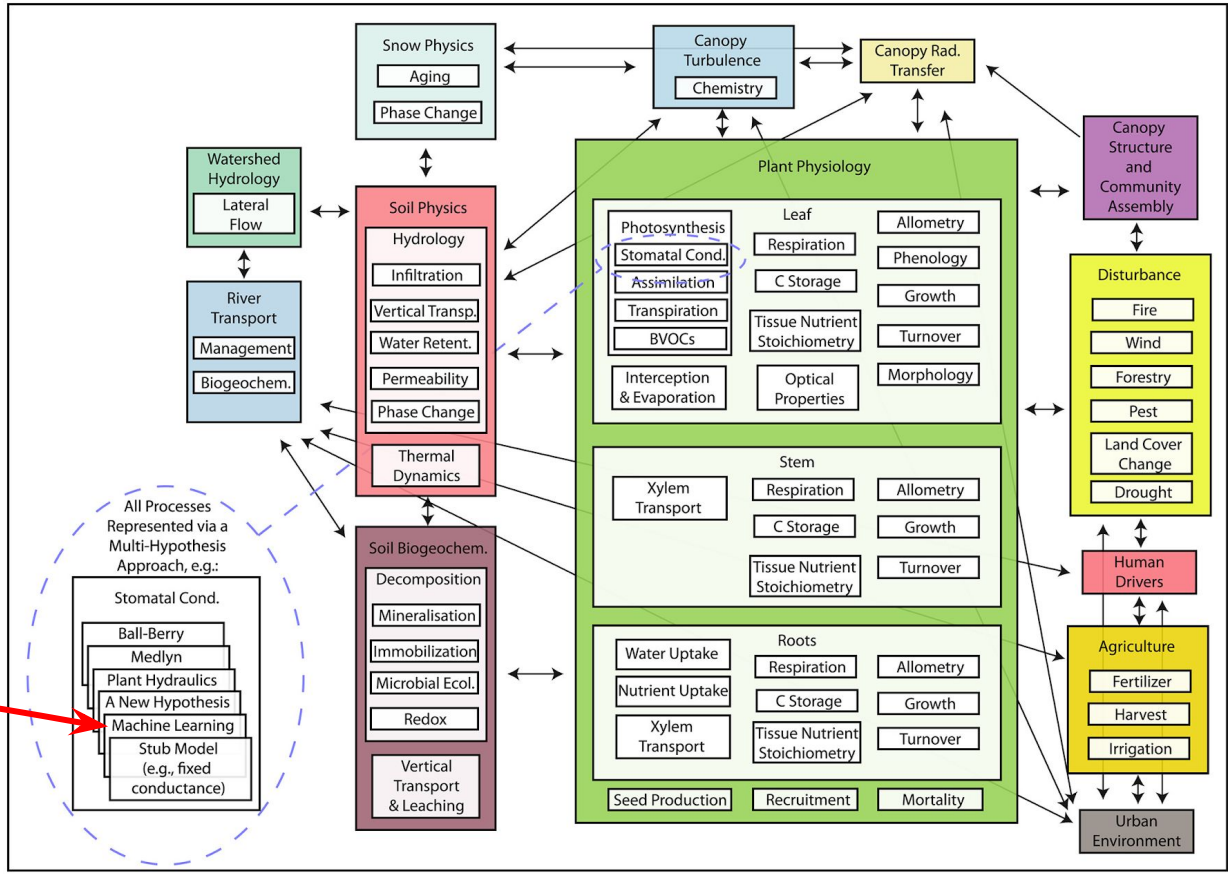



(Figure from P. E. Thornton)

# Hybrid ML/Process-based Modeling for Terrestrial Modeling

Individual processes can be represented in a multi-hypothesis approach, and ML provides an opportunities for (1) a model surrogate module or (2) a data-derived module that can be further explored or used to calibrate other hypotheses, when sufficient data are available.

(Fisher and Koven, 2020)

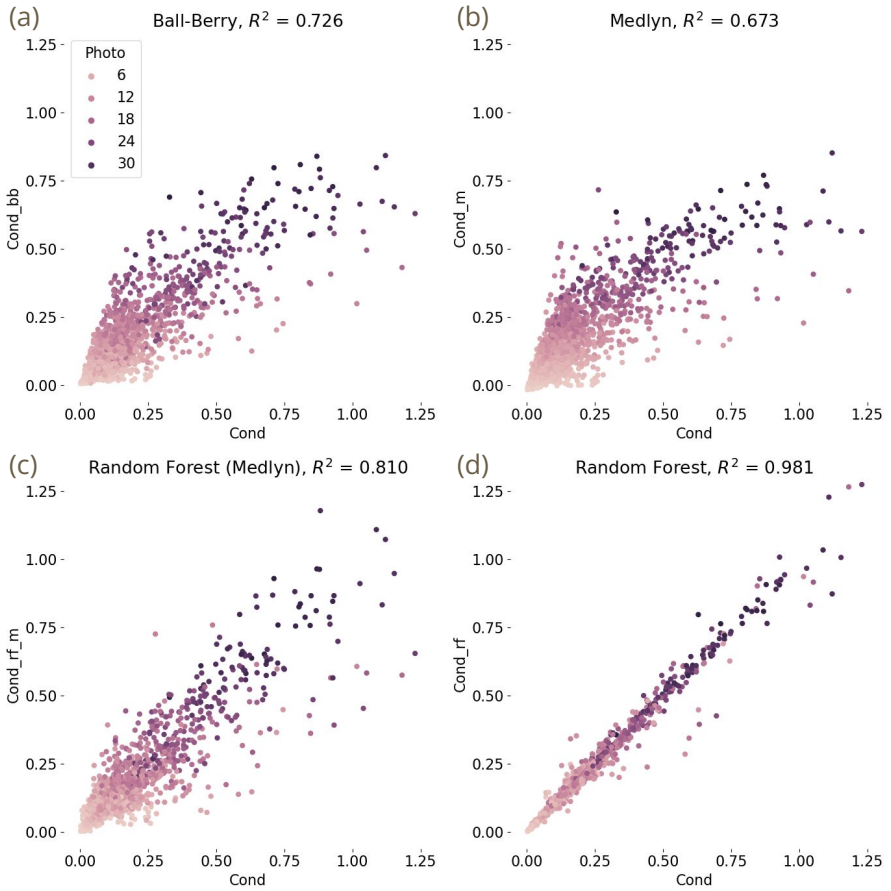



(a) Process Schematic of a Possible Full-Complexity Configuration of a Land Surface Model

# Hybrid Modeling of Photosynthesis and Ecohydrology

- Significant leaf-level data may be used to train ML parameterizations to **improve accuracy** and **computational performance**
- **Estimated stomatal conductance** vs. measured stomatal conductance for (a) Ball-Berry, (b) Medlyn, (c) Random forest (with Medlyn inputs), and (d) Random forest with all inputs from Lin et al. (2015)
- Inputs to the Medlyn parameterization are leaf-level $CO_2$, photosynthesis, and vapor pressure deficit
- Random forest trained on these three inputs (c) performs slightly better than Medlyn
- Random forest trained on more variables (d) achieves an $R^2$ of 0.98

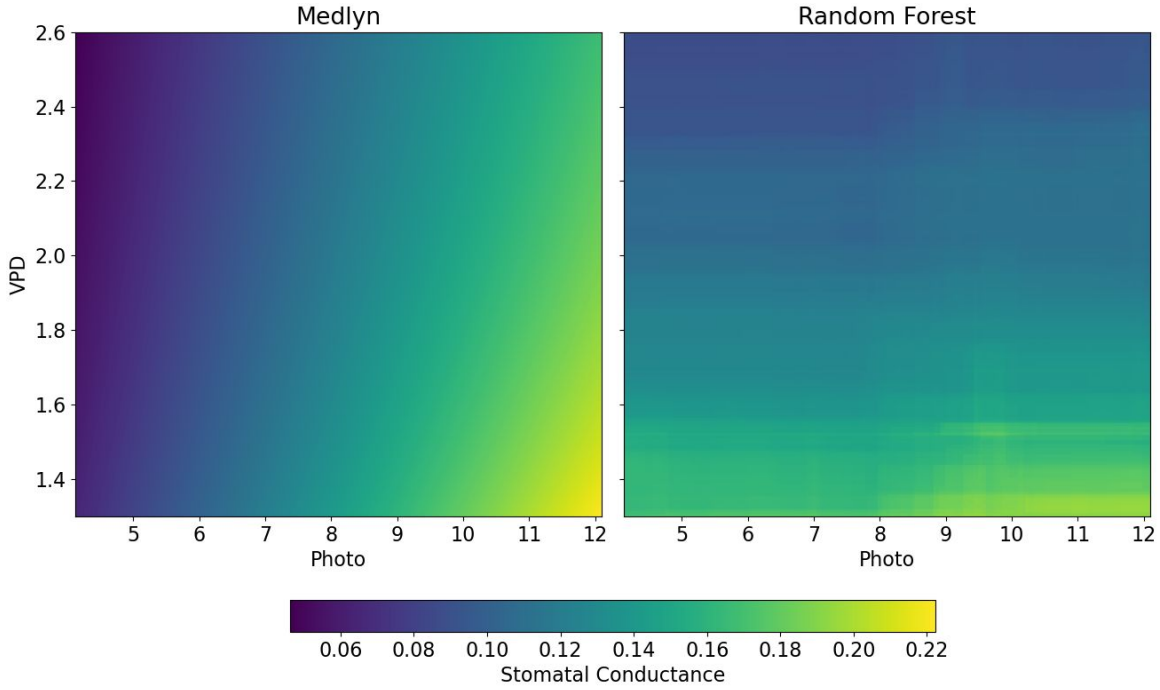
(Massoud, Collier, et al. in prep)



# Hybrid Modeling of Photosynthesis and Ecohydrology

- Most process-based or empirical formulations are continuous
- But ML formulations may exhibit discontinuities in the multi-dimensional space of inputs because of out-of-sample data or artifacts of sampling or precision

- For example, we can see such discontinuities at right for Random Forest in the VPD vs. photosynthesis heat map for stomatal conductance
- These discontinuities are likely to have numerical consequences when attempting to couple a ML parameterization into a hybrid empirical / ML Earth system model




(Massoud, Collier, et al. in prep)

AI4ESP

# ARTIFICIAL INTELLIGENCE FOR EARTH SYSTEM PREDICTABILITY (AI4ESP): CHALLENGES AND OPPORTUNITIES


**FORREST M. HOFFMAN**
Oak Ridge National Laboratory

**CHARULEKA VARADHARAJAN**
**HARUKO WAINWRIGHT**
Lawrence Berkeley National Laboratory

**NICKI L. HICKMON**
**SCOTT M. COLLIS**
Argonne National Laboratory


U.S. DEPARTMENT OF ENERGY

https://ai4esp.org/ https://ai4esp.slack.com/

# AI4ESP | Artificial Intelligence for Earth System Predictability

A multi-lab initiative working with the Earth and Environmental Systems Science Division (EESSD) of the Office of Biological and Environmental Research (BER) to develop a new paradigm for Earth system predictability focused on enabling artificial intelligence across field, lab, modeling, and analysis activities.

White papers were solicited for development and application of AI methods in areas relevant to EESSD research with an emphasis on quantifying and improving Earth system predictability, particularly related to the integrative water cycle and extreme events.

***How can DOE directly leverage artificial intelligence (AI) to engineer a substantial (paradigm-changing) improvement in Earth System Predictability?***

156 white papers were received and read to plan the organization of the **AI4ESP Workshop on Oct 25–Dec 3, 2021**

U.S. DEPARTMENT OF ENERGY

## Earth System Predictability Sessions

- Atmospheric Modeling
- Land Modeling
- Human Systems & Dynamics
- Hydrology
- Watershed Science
- Ecohydrology
- Aerosols & Clouds
- Climate Variability & Extremes
- Coastal Dynamics, Oceans & Ice

## Cross-Cut Sessions

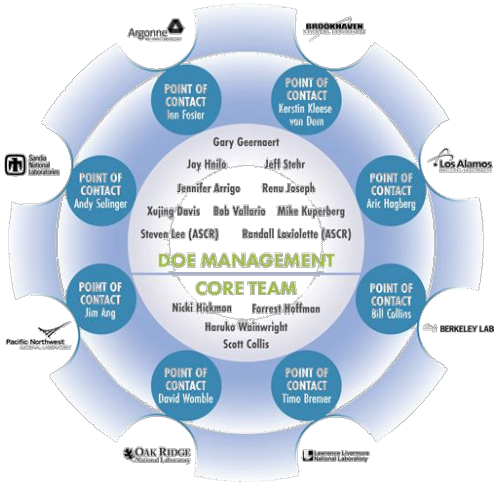
- Data Acquisition
- Neural Networks
- Surrogate models and emulators
- Knowledge-Informed Machine Learning
- Hybrid Modeling
- Explainable/Interpretable/Trustworthy AI
- Knowledge Discovery & Statistical Learning
- AI Architectures and Co-design

## Workshop Report

- Posted on ai4esp.org
- Executive Summary
- Long summary
- Earth science chapters
- Computational science chapters

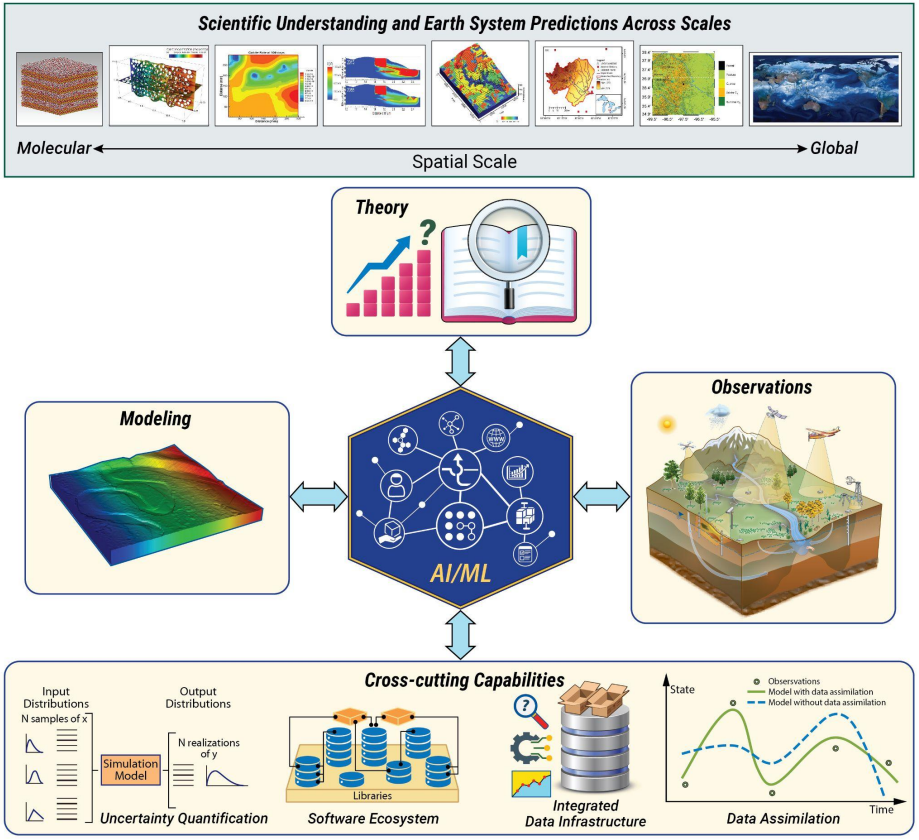
## AMS Special Collection

- Open submissions for new AI for the Earth Systems journal




AI4ESP

# AI4ESP WORKSHOP HIGHLIGHTS



# AI4ESP WORKSHOP HIGHLIGHTS

***Overview of priorities emerging from the AI4ESP workshop across 3 key themes.***

These priorities will help address major challenges for Earth system predictability

## Earth Science Priorities

- New observations
- AI-ready data products
- Data-driven and hybrid models
- Analytical approaches
- Uncertainty quantification, model parametrization & calibration

### *To Tackle Challenges*

- Significant data gaps
- Scaling and heterogeneity
- Extreme events
- Representation of human activities
- Knowledge discovery
- Accurate high-resolution predictions with low bias, uncertainty
- Providing actionable, timely information for decision making

## Computational Science Priorities


- Hybrid models
- Fundamental math and algorithms
- Interpretable, trustworthy AI
- AI-enabled data acquisition
- Data, software, hardware infrastructure

### *To Tackle Challenges*

- Physically consistent predictions for data-driven models
- Computational costs of process models
- Sparse data, extreme values
- Identifying causality
- Interpretable, trustworthy predictions
- Data discovery, access, synthesis
- Model development and comparison

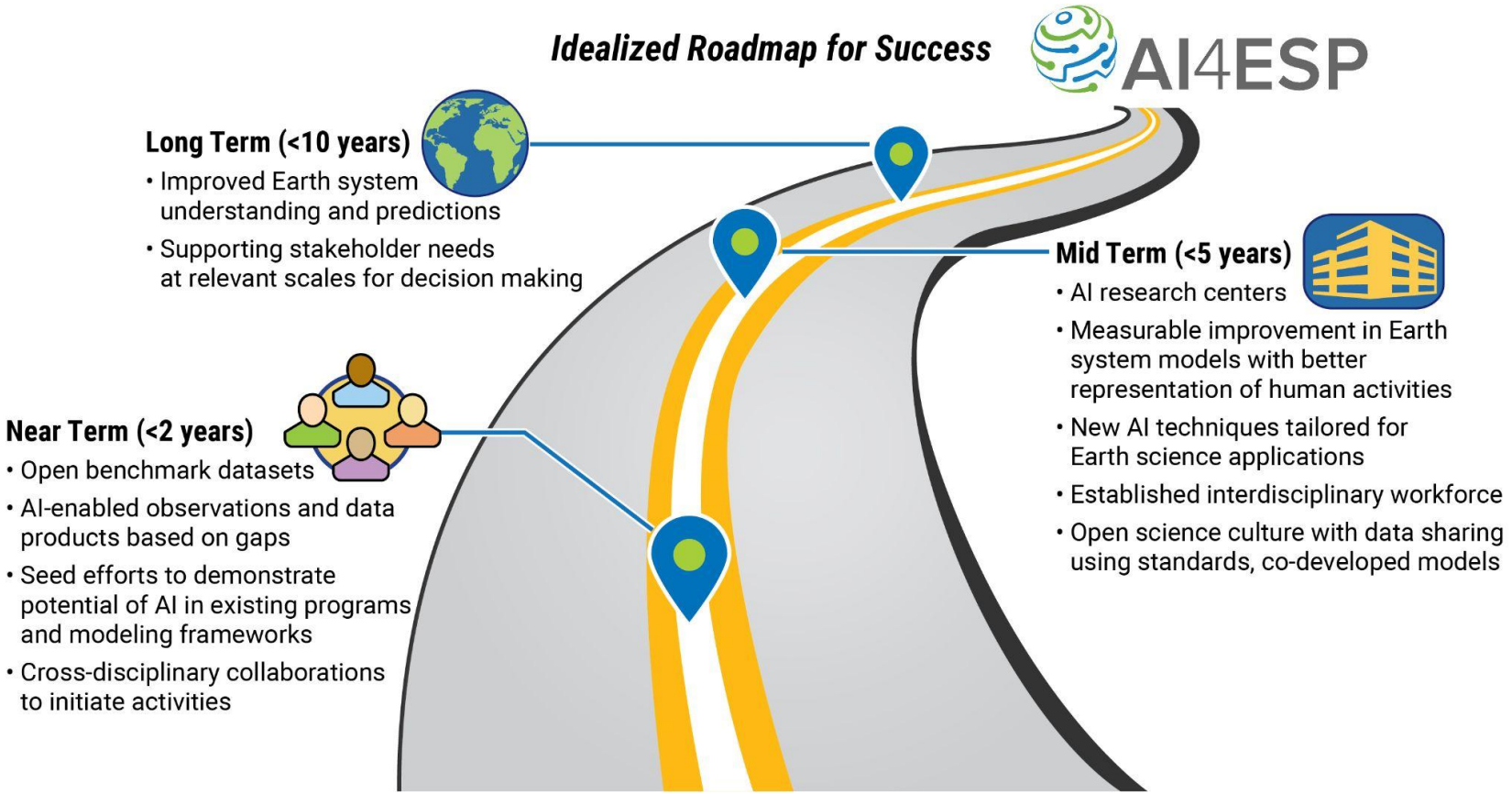
## Programmatic and Cultural Priorities

- AI research centers
- Workforce development
- Codesign infrastructure
- Common standards, benchmarks
- Seed projects, integrate AI into programs
- AI ethics and policies

### *To Tackle Challenges*

- Interdisciplinary scientific research
- Diverse organizational missions
- Personnel lack training in AI/ML
- Using data, communicating across research domains, organizations
- Data bias, model fairness, explainability of predictions

# AI4ESP WORKSHOP HIGHLIGHTS




THANK YOU
U.S. DEPARTMENT OF ENERGY
AI4ESP