U.S. DEPARTMENT OF ENERGY

Office of Science

Exploiting Artificial Intelligence for Advancing Earth and Environmental System Science

Forrest M. Hoffman[1,2], Jitendra Kumar[3], Zachary L. Langford[4], V. Shashank Konduri[5], Nathan Collier[1], Min Xu[1], and William W. Hargrove[6]

June 2, 2022

[1]Computational Earth Sciences Group, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2]Department of Civil & Environmental Engineering, University of Tennessee, Knoxville, TN, USA
[3]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[4]Cyber Resilience & Intelligence Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[5]NASA Goddard Space Flight Center, Greenbelt, MD, USA
[6]US Department of Agriculture, Forest Service, Eastern Forest Environmental Threat Assessment Center, Asheville, NC, USA

Future Strategy Webinar Series 2022

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 evaluate and benchmark model results
- Explosive data growth and the promise of discovery through data-driven modeling necessitate new methods for feature extraction, change detection, data assimilation, simulation, and analysis

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

FOCUS | NEXT-GENERATION SUPERCOMPUTERS

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.

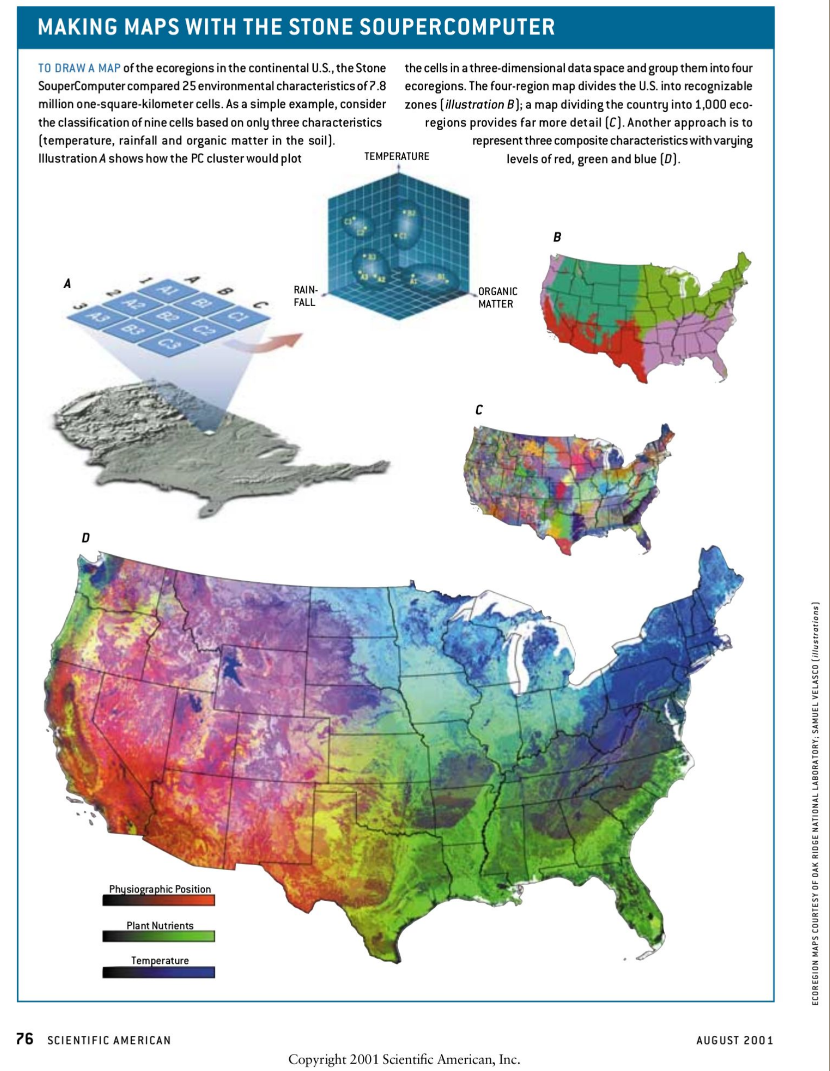
Copyright 2001 Scientific American, Inc.

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/

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

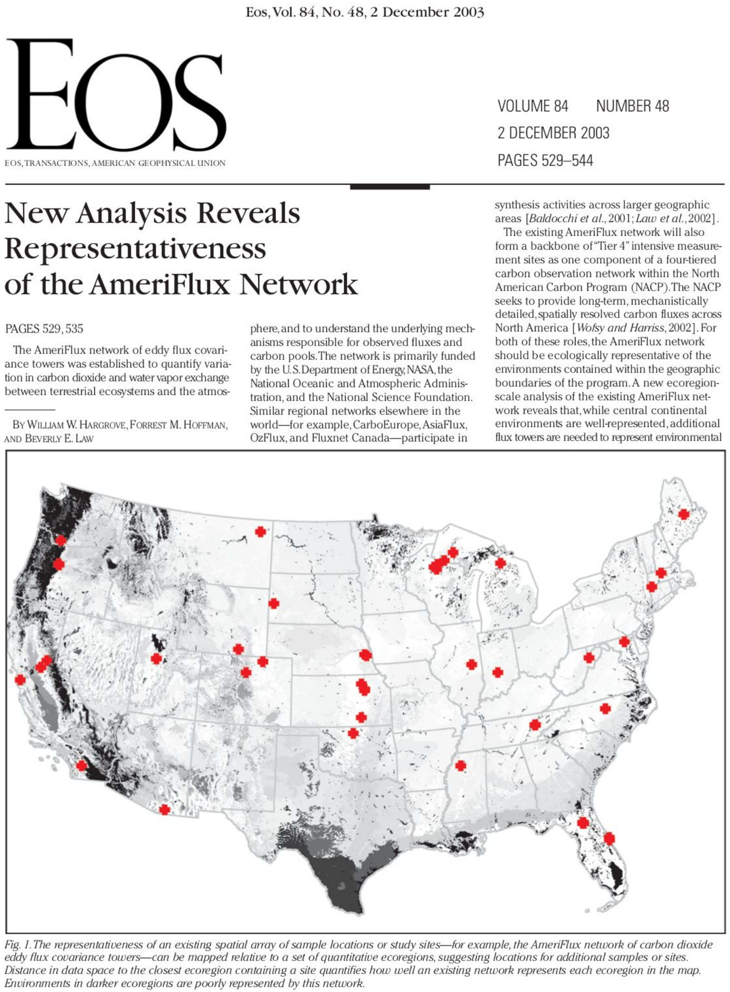
Hargrove, W. W., F. M. Hoffman, and T. Sterling (2001), The Do-It-Yourself Supercomputer, *Sci. Am.*, 265(2):72–79, https://www.scientificamerican.com/article/the-do-it-yourself-superc/



MAKING MAPS WITH THE STONE SOUPERCOMPUTER

TO DRAW A MAP of the ecoregions in the continental U.S., the Stone SouperComputer compared 25 environmental characteristics of 7.8 million one-square-kilometer cells. As a simple example, consider the classification of nine cells based on only three characteristics (temperature, rainfall and organic matter in the soil). Illustration *A* shows how the PC cluster would plot the cells in a three-dimensional data space and group them into four ecoregions. The four-region map divides the U.S. into recognizable zones (*illustration B*); a map dividing the country into 1,000 ecoregions provides far more detail (*C*). Another approach is to represent three composite characteristics with varying levels of red, green and blue (*D*).

Network Representativeness


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.

- 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.

Optimizing Sampling Networks

- Our group produced this network representativeness map for the authors from global climate, edaphic, and elevation and topography data
- Dark areas, including most of the Indian subcontinent, were poorly represented by the constellation of eddy covariance flux towers participating in FLUXNET in the year 2007

Sundareshwar, P. V., et al. (2007), Environmental Monitoring Network for India, *Science*, 316(5822):204–205, doi:[10.1126/science.1137417](https://doi.org/10.1126/science.1137417).

POLICY FORUM

CORRECTED 8 JUNE 2007; SEE LAST PAGE

ENVIRONMENT

Environmental Monitoring Network for India

An integrated monitoring system is proposed for India that will monitor terrestrial, coastal, and oceanic environments.

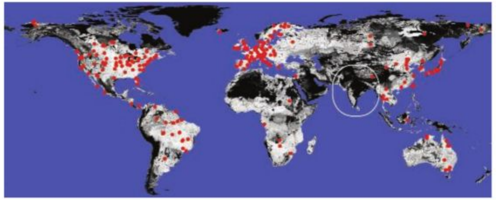
P. V. Sundareshwar,* R. Murtugudde, G. Srinivasan, S. Singh, K. J. Ramesh, R. Ramesh, S. B. Verma, D. Agarwal, D. Baldocchi, C. K. Baru, K. K. Baruah, G. R. Chowdhury, V. K. Dadhwal, C. B. S. Dutt, J. Fuentes, Prabhat K. Gupta, W. W. Hargrove, M. Howard, C. S. Jha, S. Lal, W. K. Michener, A. P. Mitra, J. T. Morris, R. R. Myneni, M. Naja, R. Nemani, R. Purvaja, S. Raha, S. K. Santhana Vanan, M. Sharma, A. Subramaniam, R. Sukumar, R. R. Twilley, P. R. Zimmerman

Understanding the consequences of global environmental change and its mitigation will require an integrated global effort of comprehensive long-term data collection, synthesis, and action (*1*). The last decade has seen a dramatic global increase in the number of networked monitoring sites. For example, FLUXNET is a global collection of >300 micrometeorological terrestrial-flux research sites (see figure, right) that monitor fluxes of CO_2, water vapor, and energy (*2–4*). A similar, albeit sparser, network of ocean observation sites is quantifying the fluxes of greenhouse gases (GHGs) from oceans and their role in the global carbon cycle (*5*, *6*). These networks are operated on an ad hoc basis by the scientific community. Although FLUXNET and other observation networks cover diverse vegetation types within a 70°S to 30°N latitude band (*3*) and different oceans (*5*, *6*), there are not comprehensive and reliable data from African and Asian regions. Lack of robust scientific data from these regions of the world is a serious impediment to efforts to understand and mitigate impacts of climate and environmental change (*5*, *7*).

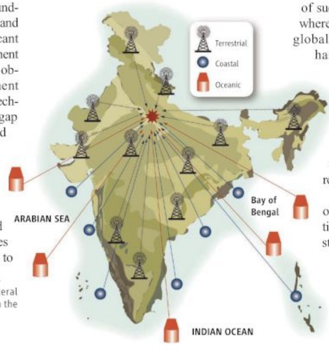
The Indian subcontinent and the surrounding seas, with more than 1.3 billion people and unique natural resources, have a significant impact on the regional and global environment but lack a comprehensive environmental observation network. Within the government of India, the Department of Science and Technology (DST) has proposed filling this gap by establishing INDOFLUX, a coordinated multidisciplinary environmental monitoring network that integrates terrestrial, coastal, and oceanic environments (see figure, right).

In a workshop held in July 2006 (*8*), a team of scientists from India and the United States developed the overarching objectives for the proposed INDOFLUX. These are to provide a scientific understanding (i) of the coupling of atmospheric, oceanic, and terrestrial environments in India; (ii) of the nature and pace of environmental change in India; and (iii) of subsequent impacts on provision of ecosystem services. Also, in order to evaluate what will enable India to sustain its natural resources, these goals include an assessment of the vulnerability and consequent risks to its social and natural systems.

Climate change will alter the regional biosphere-climate feedbacks and land-ocean coupling. Although global models reliably predict the trend in the impact of climate change on India's forest resources, the magnitude of such change is uncertain (*9*). Similarly, whereas all oceans show the influence of global warming (*10*), the Indian Ocean has shown higher-than-average surface warming, especially during the last five decades (*11*, *12*). This warming may have global impacts (*13*, *14*), even though the impact on the Indian summer monsoons is not well understood (*15*, *16*). These uncertainties highlight the need for regional models driven by regional data. As the hypoxia observed in the Gulf of Mexico is related to agricultural practices in the watershed (*17*), Indian Ocean studies also indicate couplings between mainland activities and offshore and



Current monitoring sites in FLUXNET. Sites are shown in red, and global representativeness is estimated by Global Multivariate Clustering Analysis (*24–26*). Darker areas are poorly represented by the existing FLUXNET towers. Environmental similarity was calculated from a set of variables (precipitation, temperature, solar flux, total soil carbon and nitrogen, bulk density, elevation, and compound topographic index) at a resolution of 4 km.



A schematic of the INDOFLUX proposal. Placement of stations reflects different climactic, vegetation, and land-use areas. Final locations will be determined as part of the formal science plan.

CREDITS (TOP TO BOTTOM): W. W. HARGROVE/P. V. SUNDARESHWAR (SOURCE)

The authors were members of an indo–U.S. bilateral workshop on INDOFLUX. Affiliations are provided in the supporting online material.

*Author for correspondence. E-mail: pvs@sdsmt.edu

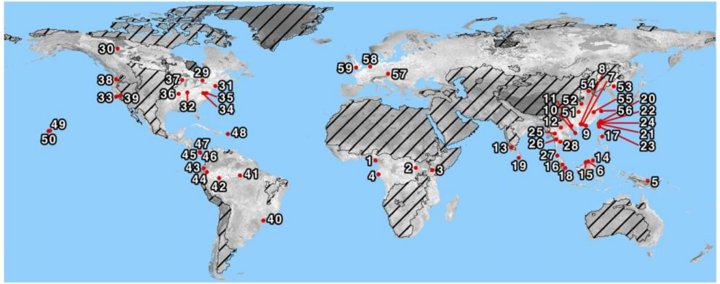


Fig. 1 Map of the CTFS-ForestGEO network illustrating its representation of bioclimatic, edaphic, and topographic conditions globally. Site numbers correspond to ID# in Table 2. Shading indicates how well the network of sites represents the suite of environmental factors included in the analysis; light-colored areas are well-represented by the network, while dark colored areas are poorly represented. Stippling covers nonforest areas. The analysis is described in Appendix S1.

Table 1 Attributes of a CTFS-ForestGEO census

Attribute	Utility
Very large plot size	Resolve community and population dynamics of highly diverse forests with many rare species with sufficient sample sizes (Losos & Leigh, 2004; Condit *et al.*, 2006); quantify spatial patterns at multiple scales (Condit *et al.*, 2000; Wiegand *et al.*, 2007a,b; Detto & Muller-Landau, 2013; Lutz *et al.*, 2013); characterize gap dynamics (Feeley *et al.*, 2007b); calibrate and validate remote sensing and models, particularly those with large spatial grain (Mascaro *et al.*, 2011; Réjou-Méchain *et al.*, 2014)
Includes every freestanding woody stem ≥1 cm DBH	Characterize the abundance and diversity of understory as well as canopy trees; quantify the demography of juveniles (Condit, 2000; Muller-Landau *et al.*, 2006a,b).
All individuals identified to species	Characterize patterns of diversity, species-area, and abundance distributions (Hubbell, 1979, 2001; He & Legendre, 2002; Condit *et al.*, 2005; John *et al.*, 2007; Shen *et al.*, 2009; He & Hubbell, 2011; Wang *et al.*, 2011; Cheng *et al.*, 2012); test theories of competition and coexistence (Brown *et al.*, 2013); describe poorly known plant species (Gereau & Kenfack, 2000; Davies, 2001; Davies *et al.*, 2001; Sonké *et al.*, 2002; Kenfack *et al.*, 2004, 2006)
Diameter measured on all stems	Characterize size-abundance distributions (Muller-Landau *et al.*, 2006b; Lai *et al.*, 2013; Lutz *et al.*, 2013); combine with allometries to estimate whole-ecosystem properties such as biomass (Chave *et al.*, 2008; Valencia *et al.*, 2009; Lin *et al.*, 2012; Ngo *et al.*, 2013; Muller-Landau *et al.*, 2014)
Mapping of all stems and fine-scale topography	Characterize the spatial pattern of populations (Condit, 2000); conduct spatially explicit analyses of neighborhood influences (Condit *et al.*, 1992; Hubbell *et al.*, 2001; Uriarte *et al.*, 2004, 2005; Rüger *et al.*, 2011, 2012; Lutz *et al.*, 2014); characterize microhabitat specificity and controls on demography, biomass, etc. (Harms *et al.*, 2001; Valencia *et al.*, 2004; Chuyong *et al.*, 2011); align on the ground and remote sensing measurements (Asner *et al.*, 2011; Mascaro *et al.*, 2011).
Census typically repeated every 5 years	Characterize demographic rates and changes therein (Russo *et al.*, 2005; Muller-Landau *et al.*, 2006a,b; Feeley *et al.*, 2007a; Lai *et al.*, 2013; Stephenson *et al.*, 2014); characterize changes in community composition (Losos & Leigh, 2004; Chave *et al.*, 2008; Feeley *et al.*, 2011; Swenson *et al.*, 2012; Chisholm *et al.*, 2014); characterize changes in biomass or productivity (Chave *et al.*, 2008; Banin *et al.*, 2014; Muller-Landau *et al.*, 2014)

Optimizing Sampling Networks

- The CTFS-ForestGEO global forest monitoring network is aimed at characterizing forest responses to global change
- The figure at left shows the global representativeness of the CTFS-ForestGEO sites in 2014
- Non-forested areas are masked with hatching, and as expected, they are consistently darker than the forested regions, which are represented to varying degrees by the monitoring sites

Anderson-Teixeira, K. J., et al. (2015), CTFS-ForestGEO: A Worldwide Network Monitoring Forests in an Era of Global Change, *Glob. Change Biol.*, 21(2):528–549, doi:10.1111/gcb.12712.

Representativeness for Alaska

Data Layers

Table: 37 characteristics averaged for the present (2000–2009) and the future (2090–2099).

Description	Number/Name	Units	Source
Monthly mean air temperature	12	°C	GCM
Monthly mean precipitation	12	mm	GCM
Day of freeze	mean	day of year	GCM
	standard deviation	days	
Day of thaw	mean	day of year	GCM
	standard deviation	days	
Length of growing season	mean	days	GCM
	standard deviation	days	
Maximum active layer thickness	1	m	GIPL
Warming effect of snow	1	°C	GIPL
Mean annual ground temperature at bottom of active layer	1	°C	GIPL
Mean annual ground surface temperature	1	°C	GIPL
Thermal offset	1	°C	GIPL
Limnicity	1	%	NHD
Elevation	1	m	SRTM

Hoffman, F. M., J. Kumar, R. T. Mills, and W. W. Hargrove (2013), Representativeness-Based Sampling Network Design for the State of Alaska, *Landscape Ecol.*, 28(8):1567–1586, doi:10.1007/s10980-013-9902-0.

Landscape Ecol (2013) 28:1567–1586
DOI 10.1007/s10980-013-9902-0

RESEARCH ARTICLE

Representativeness-based sampling network design for the State of Alaska

Forrest M. Hoffman · Jitendra Kumar · Richard T. Mills · William W. Hargrove

Received: 13 February 2013 / Accepted: 31 May 2013 / Published online: 20 June 2013

Abstract Resource and logistical constraints limit the frequency and extent of environmental observations, particularly in the Arctic, necessitating the development of a systematic sampling strategy to maximize coverage and objectively represent environmental variability at desired scales. A quantitative methodology for stratifying sampling domains, informing site selection, and determining the representativeness of measurement sites and networks is described here. Multivariate spatiotemporal clustering was applied to down-scaled general circulation model results and data for the State of Alaska at 4 km^2 resolution to define multiple sets of ecoregions across two decadal time periods. Maps of ecoregions for the present (2000–2009) and future (2090–2099) were produced, showing how combinations of 37 characteristics are distributed and how they may shift in the future. Representative sampling locations are identified on present and future ecoregion maps. A representativeness metric was developed, and representativeness maps for eight candidate sampling locations were produced. This metric was used to characterize the environmental similarity of each site. This analysis provides model-inspired insights into optimal sampling strategies, offers a framework for up-scaling measurements, and provides a down-scaling approach for integration of models and measurements. These techniques can be applied at different spatial and temporal scales to meet the needs of individual measurement campaigns.

Keywords Ecoregions · Representativeness · Network design · Cluster analysis · Alaska · Permafrost

F. M. Hoffman (✉)
Computer Science & Mathematics Division, Climate Change Science Institute (CCSI), Oak Ridge National Laboratory, Oak Ridge, TN, USA
e-mail: forrest@climatemodeling.org

F. M. Hoffman · J. Kumar · R. T. Mills
Environmental Sciences Division, Climate Change Science Institute (CCSI), Oak Ridge National Laboratory, Oak Ridge, TN, USA
e-mail: jkumar@climatemodeling.org

R. T. Mills
e-mail: rmills@ornl.gov

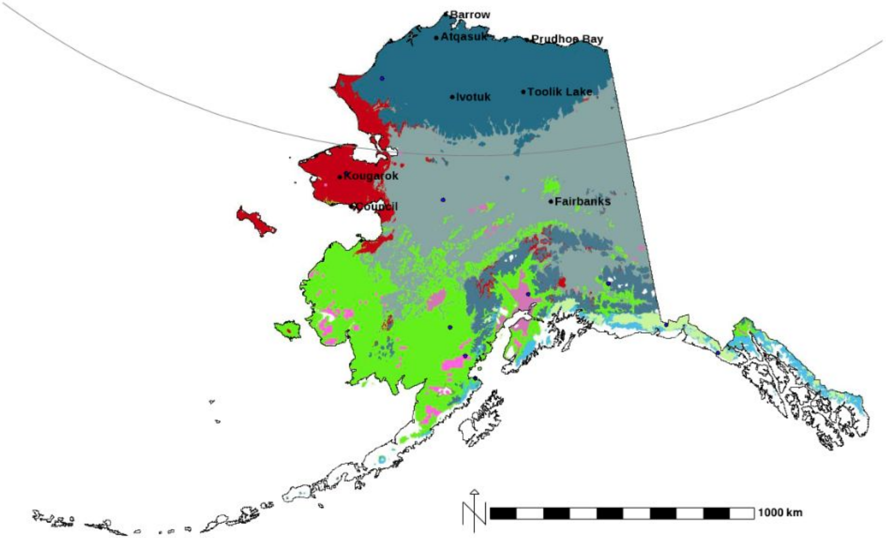
W. W. Hargrove
Eastern Forest Environmental Threat Assessment Center, USDA Forest Service, Southern Research Station, Asheville, NC, USA
e-mail: hnw@geobabble.org

Introduction

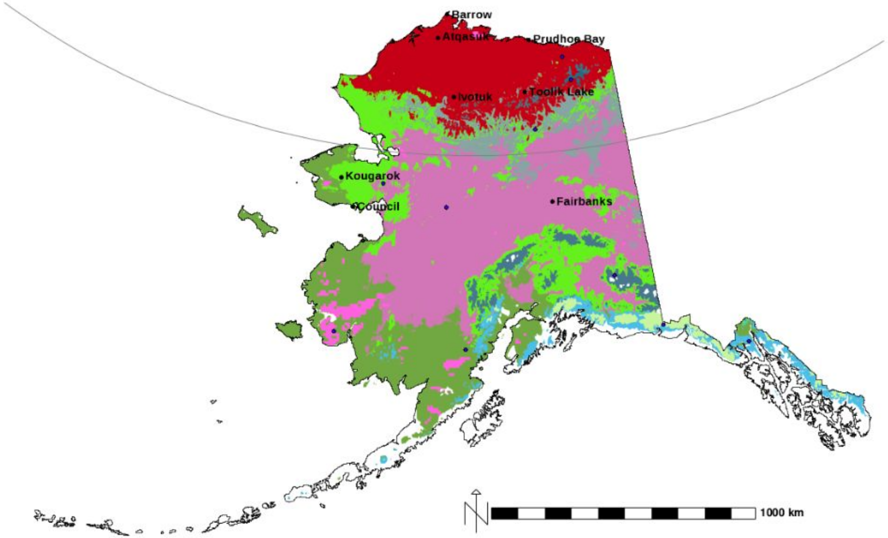
The Arctic contains vast amounts of frozen water in the form of sea ice, snow, glaciers, and permafrost. Extended areas of permafrost in the Arctic contain soil organic carbon that is equivalent to twice the size of the atmospheric carbon pool, and this large stabilized

10 Alaska Ecoregions, Present and Future

(Hoffman et al., 2013)



2000–2009

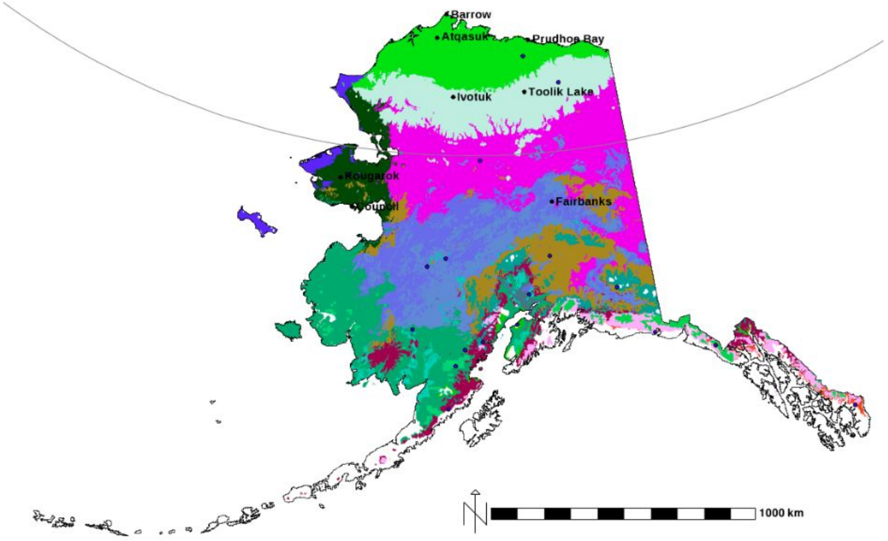


2090–2099

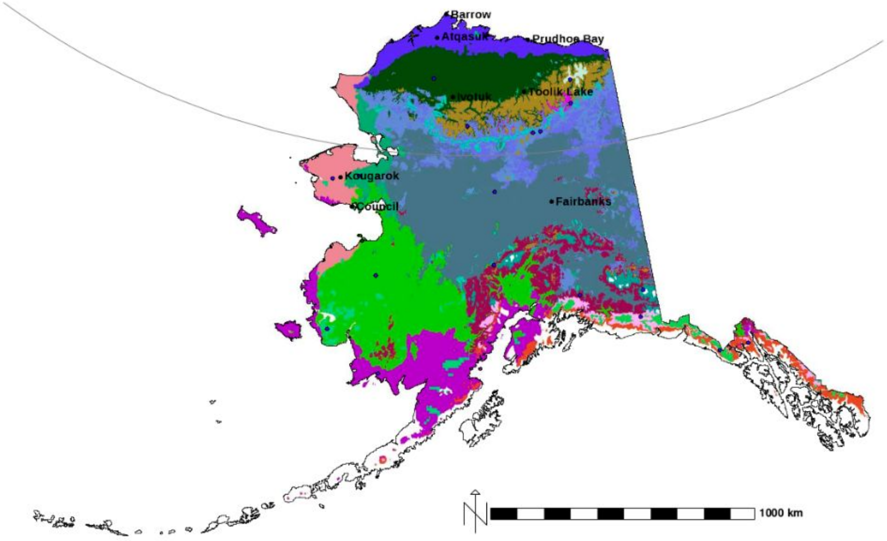
- Since the random colors are the same in both maps, a change in color represents an environmental change between the present and the future.
- At this level of division, the conditions in the large boreal forest become compressed onto the Brooks Range and the conditions on the Seward Peninsula “migrate” to the North Slope.

20 Alaska Ecoregions, Present and Future

(Hoffman et al., 2013)



2000–2009



2090–2099

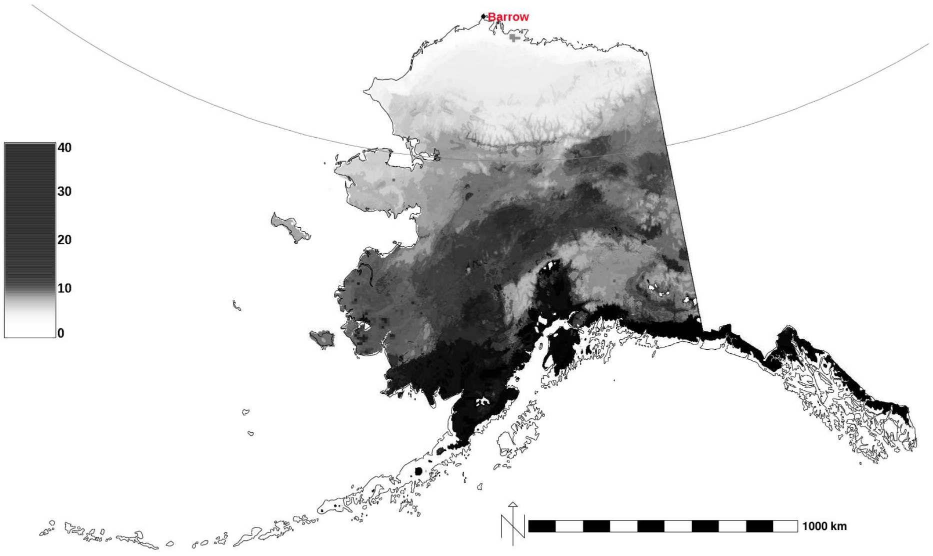
- Since the random colors are the same in both maps, a change in color represents an environmental change between the present and the future.
- At this level of division, the two primary regions of the Seward Peninsula and that of the northern boreal forest replace the two regions on the North Slope almost entirely.

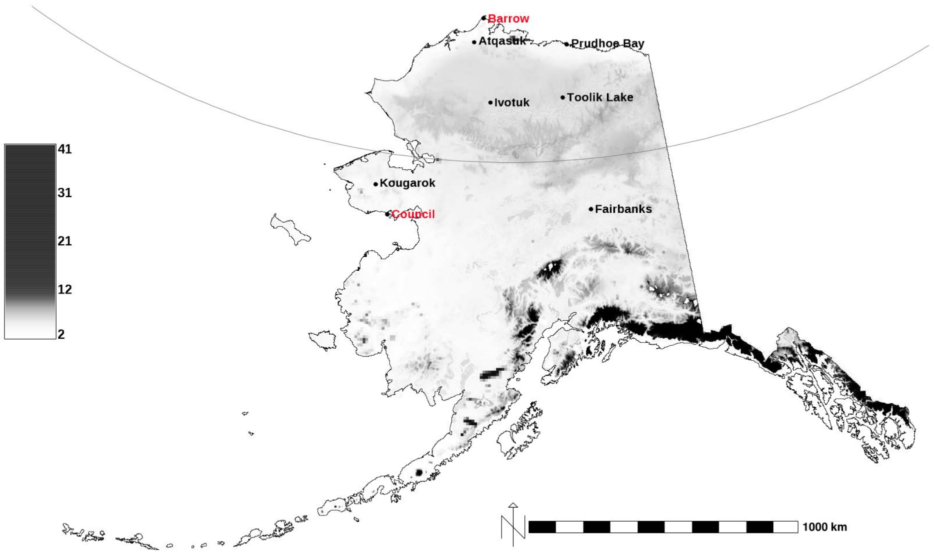
Sampling Site Representativeness

- This representativeness analysis uses the standardized *n*-dimensional data space formed from all input data layers
- In this data space, the Euclidean distance between a sampling location (like Barrow) and every other point is calculated
- These data space distances are then used to generate grayscale maps showing the similarity, or lack thereof, of every location to the sampling location
- In the subsequent maps, white areas are well represented by the sampling location or network, while dark and black areas as poorly represented by the sampling location or network
- This analysis assumes that the climate surrogates maintain their predictive power and that no significant biological adaptation occurs in the future

Network Representativeness: Barrow vs. Barrow + Council

(Hoffman et al., 2013)





Light-colored regions are well represented and dark-colored regions are poorly represented by the sampling location listed in **red**.

State Space Dissimilarities: 8 Sites, Present (2000–2009)

Table: Site state space dissimilarities for the present (2000–2009).

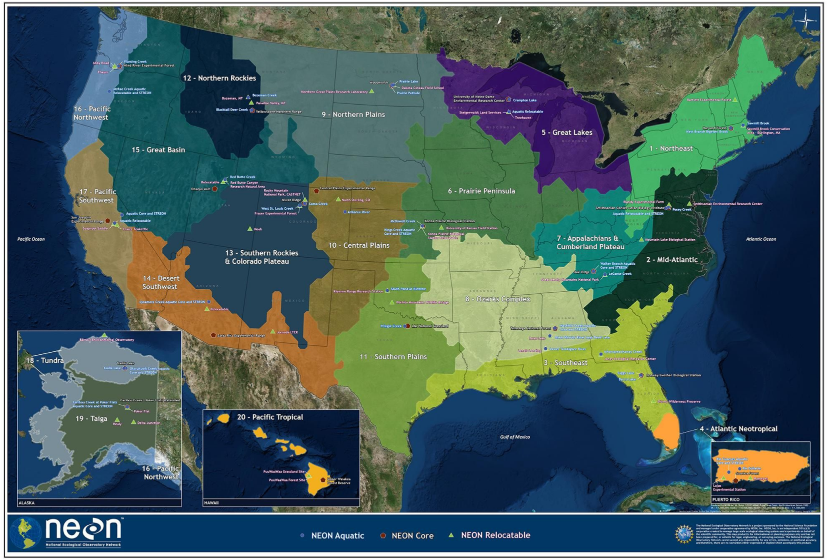
Sites	Council	Atqasuk	Ivotuk	Toolik Lake	Kougarok	Prudhoe Bay	Fairbanks
Barrow	9.13	4.53	5.90	5.87	7.98	3.57	12.16
Council		8.69	6.37	7.00	2.28	8.15	5.05
Atqasuk			5.18	5.23	7.79	1.74	10.66
Ivotuk				1.81	5.83	4.48	7.90
Toolik Lake					6.47	4.65	8.70
Kougarok						7.25	5.57
Prudhoe Bay							10.38

State Space Dissimilarities: 8 Sites, Present and Future

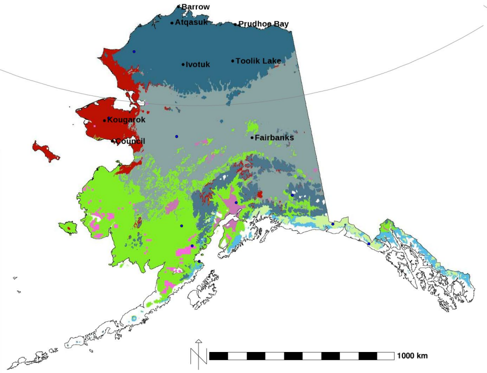
Table: Site state space dissimilarities between the present (2000–2009) and the future (2090–2099).

		Future (2090–2099)							
	Sites	Barrow	Council	Atqasuk	Ivotuk	Toolik Lake	Kougarok	Prudhoe Bay	Fairbanks
Present (2000–2009)	Barrow	3.31	9.67	4.63	6.05	5.75	9.02	3.69	11.67
	Council	8.38	1.65	8.10	5.91	6.87	3.10	7.45	5.38
	Atqasuk	6.01	9.33	2.42	5.46	5.26	8.97	2.63	10.13
	Ivotuk	7.06	7.17	5.83	1.53	2.05	7.25	4.87	7.40
	Toolik Lake	7.19	7.67	6.07	2.48	1.25	7.70	5.23	8.16
	Kougarok	7.29	3.05	6.92	5.57	6.31	2.51	6.54	5.75
	Prudhoe Bay	5.29	8.80	3.07	4.75	4.69	8.48	1.94	9.81
	Fairbanks	12.02	5.49	10.36	7.83	8.74	6.24	10.10	1.96

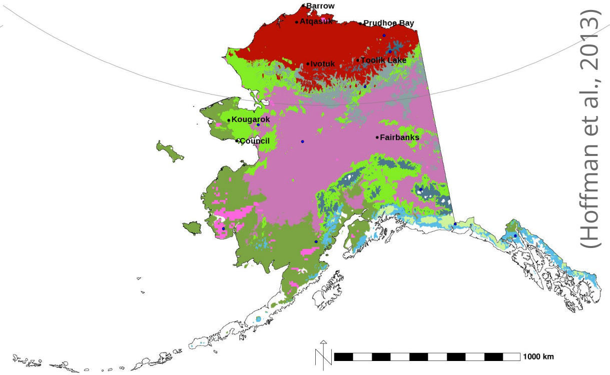
Sampling Network Design



NSF's NEON Sampling Domains



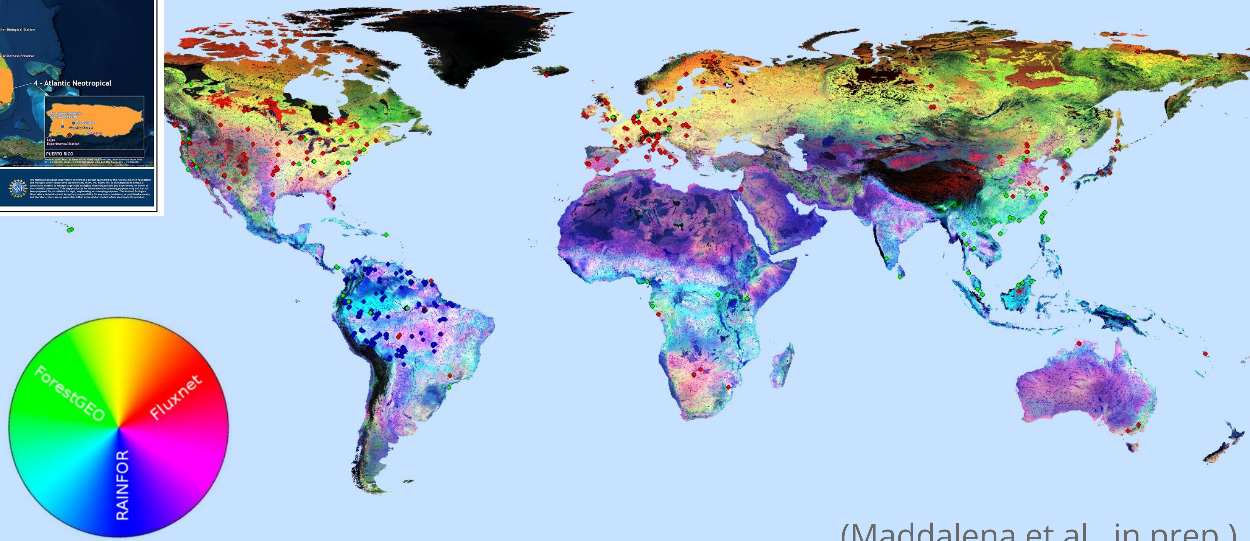
2000–2009



2090–2000

(Hoffman et al., 2013)

Triple-Network Global Representativeness

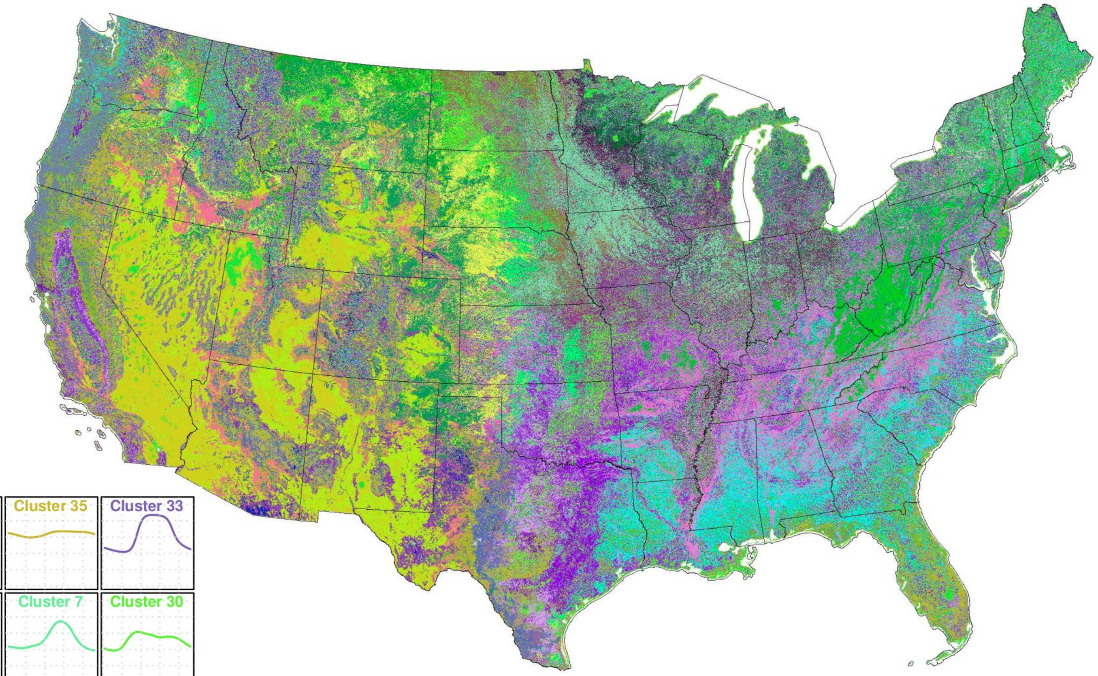


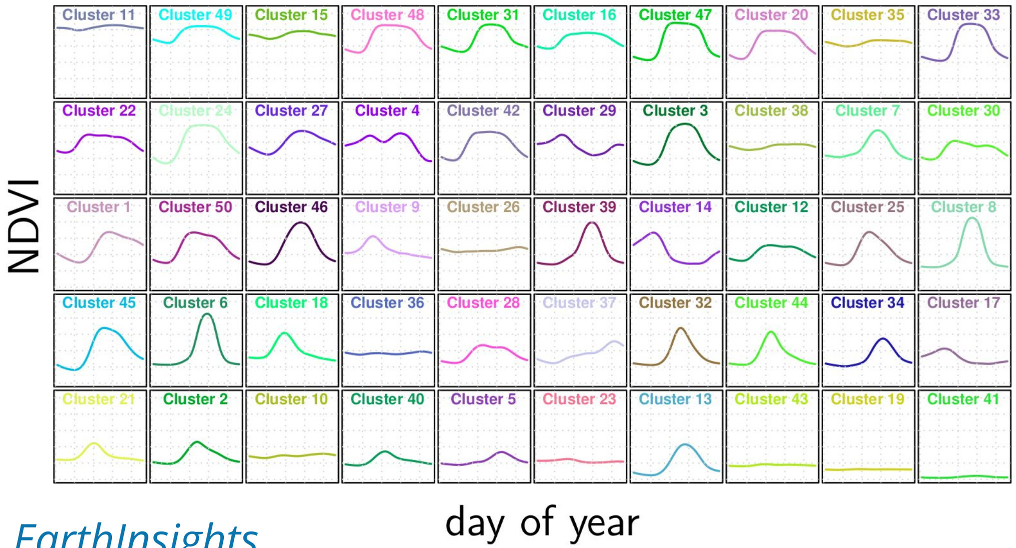
(Maddalena et al., in prep.)

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

50 Phenoregions for year 2012 (Random Colors)

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



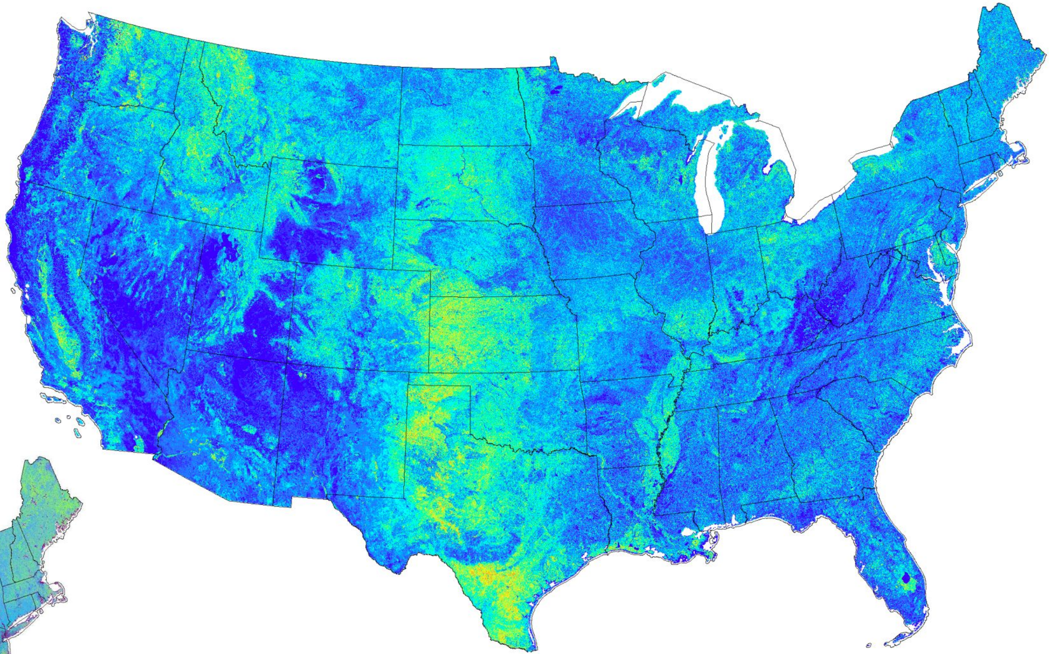


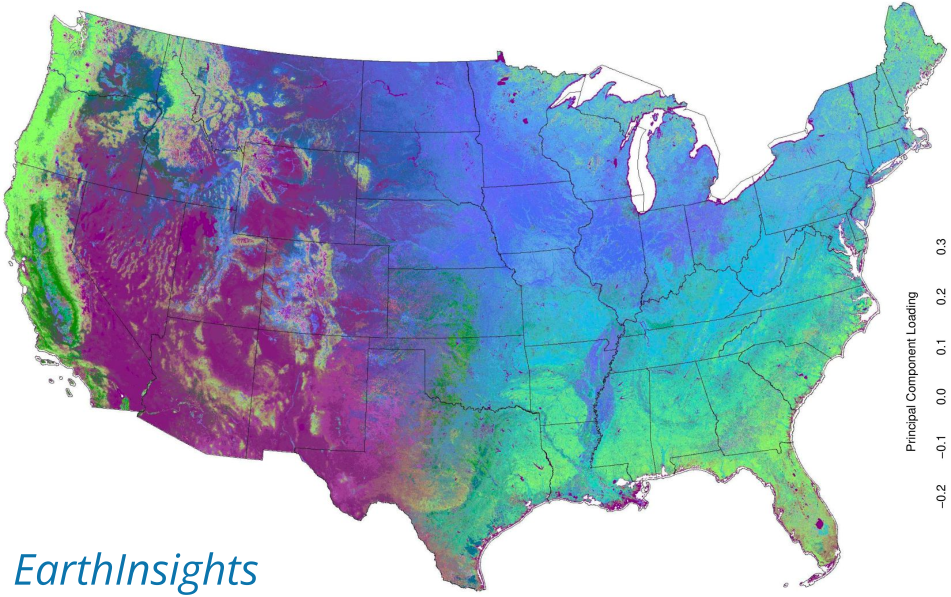
50 Phenoregion Prototypes (Random Colors)

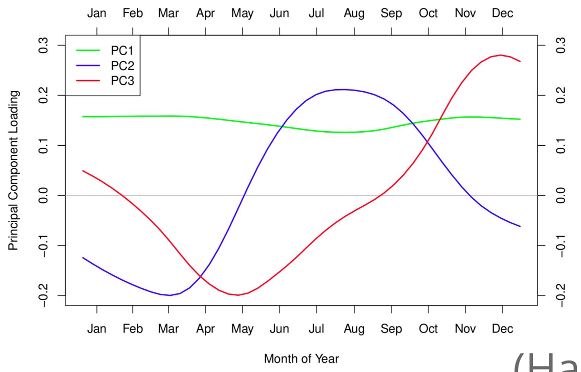
EarthInsights

(Hargrove et al., in prep.)

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







Principal Components Analysis

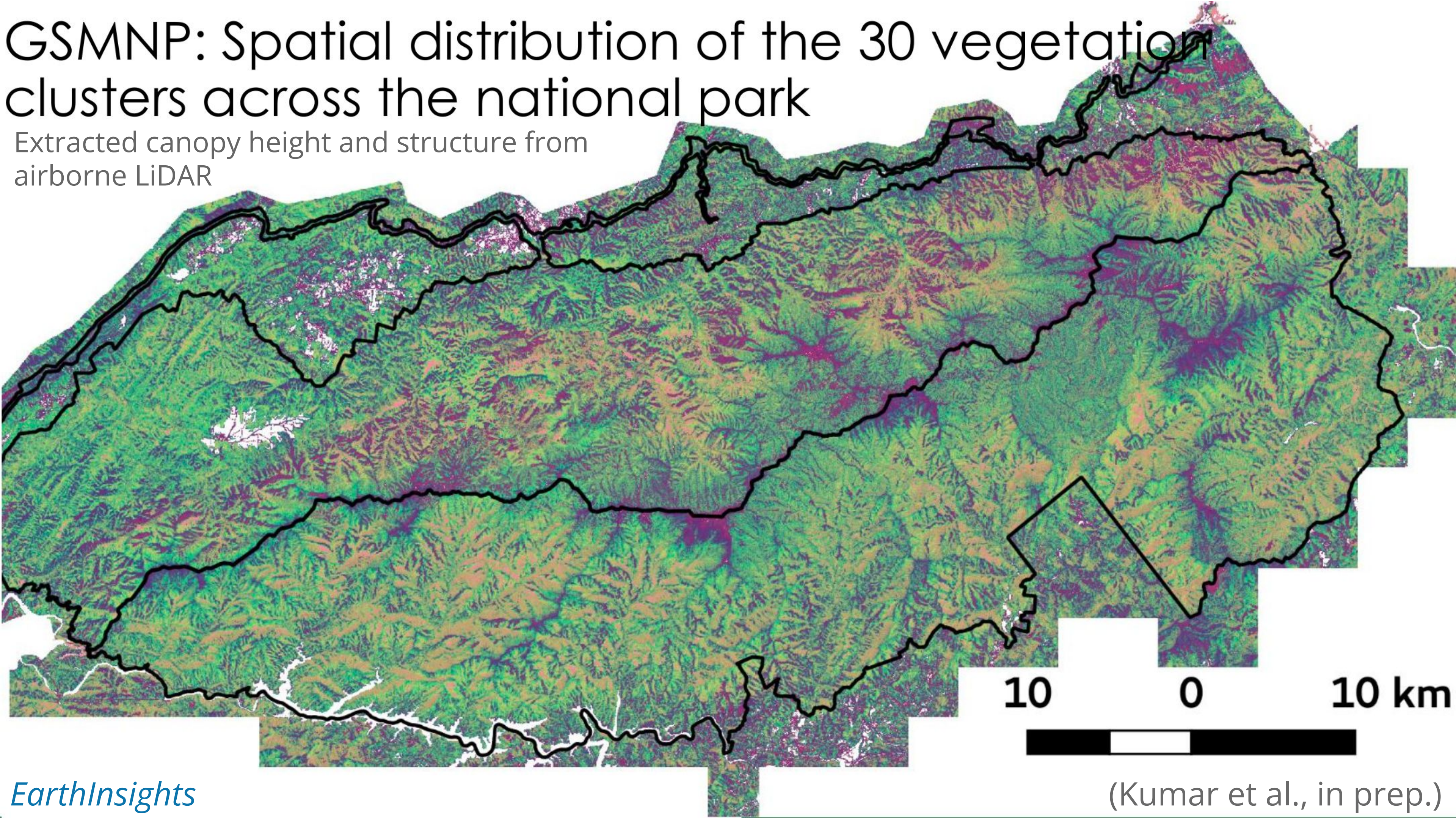
PC1 ~ Evergreen
PC2 ~ Deciduous
PC3 ~ Dry Deciduous

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

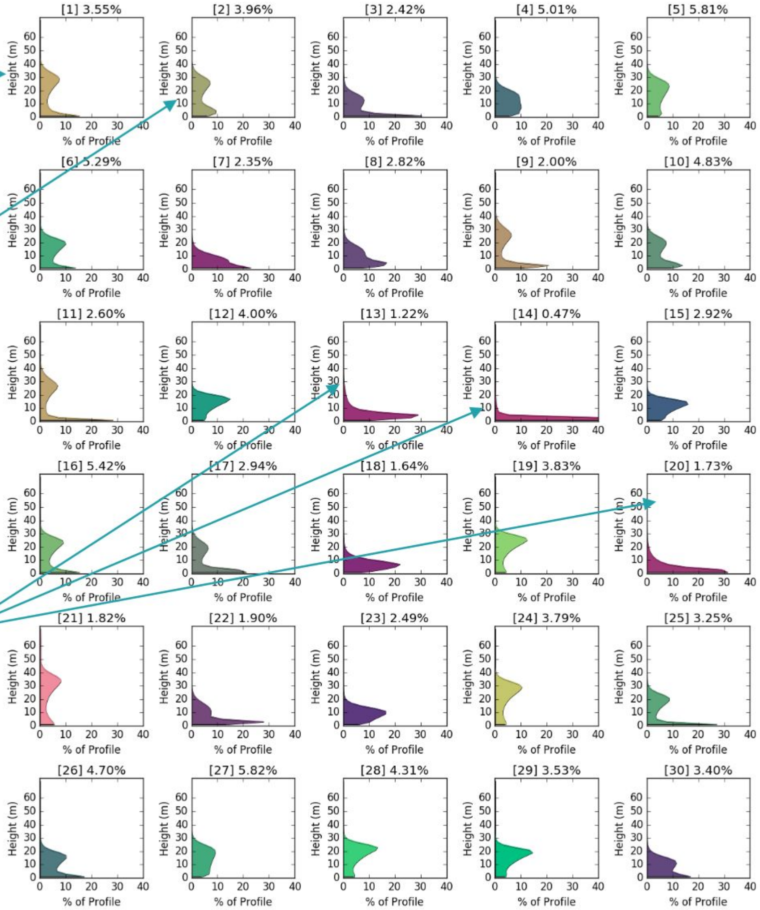
(Kumar et al., in prep.)

GSMNP: 30 representative vertical structures (cluster centroids) identified

tall forests with low understory vegetation

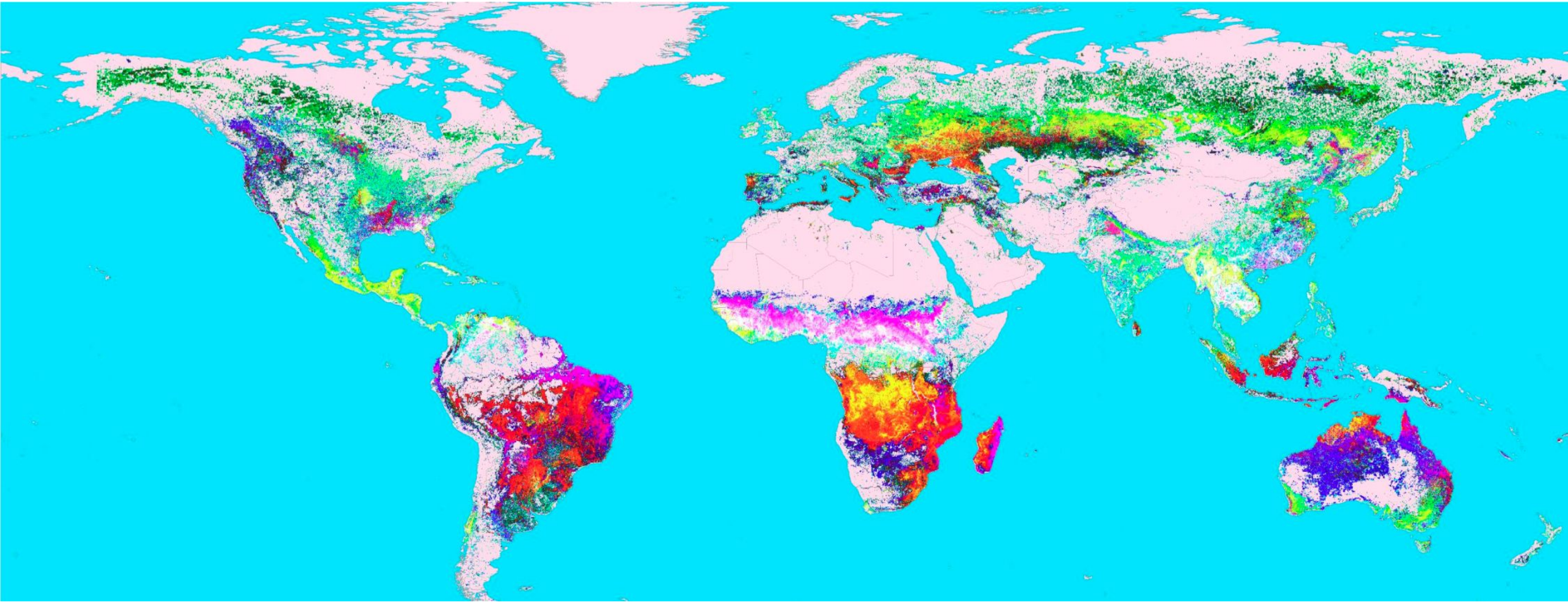
forests with slightly lower mean height with dense understory vegetation

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



(Kumar et al., in prep.)

Global Fire Regimes

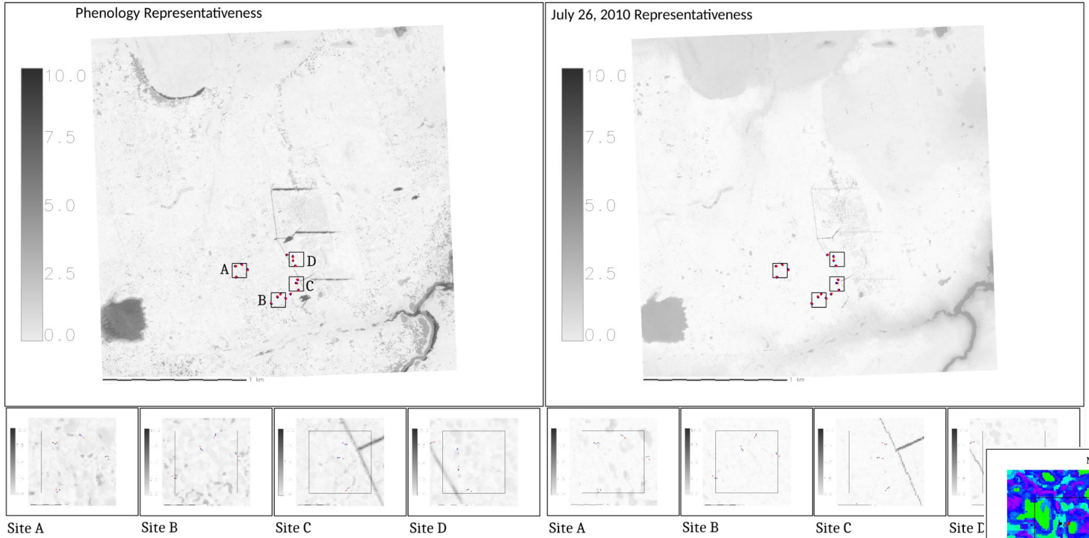


Regions that exhibit similar fire seasonality globally

From MODIS "Hotspots" at 1 km resolution from 2002–2018

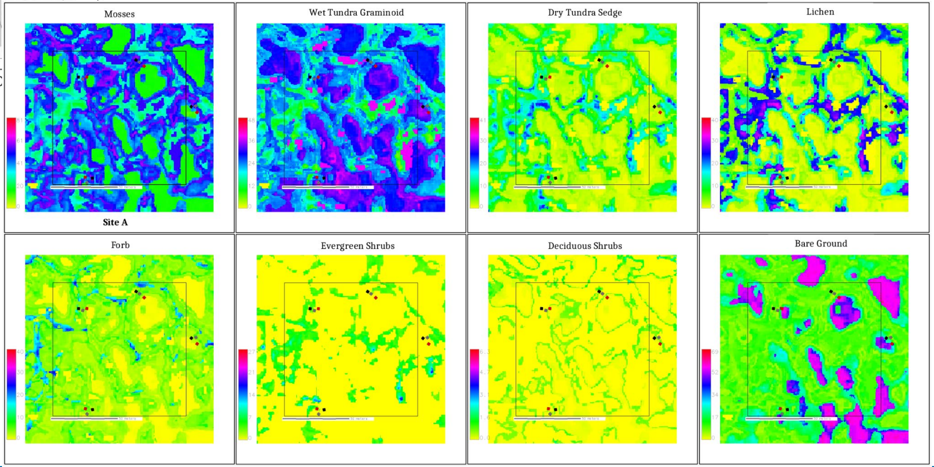
(Norman et al., submitted)

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

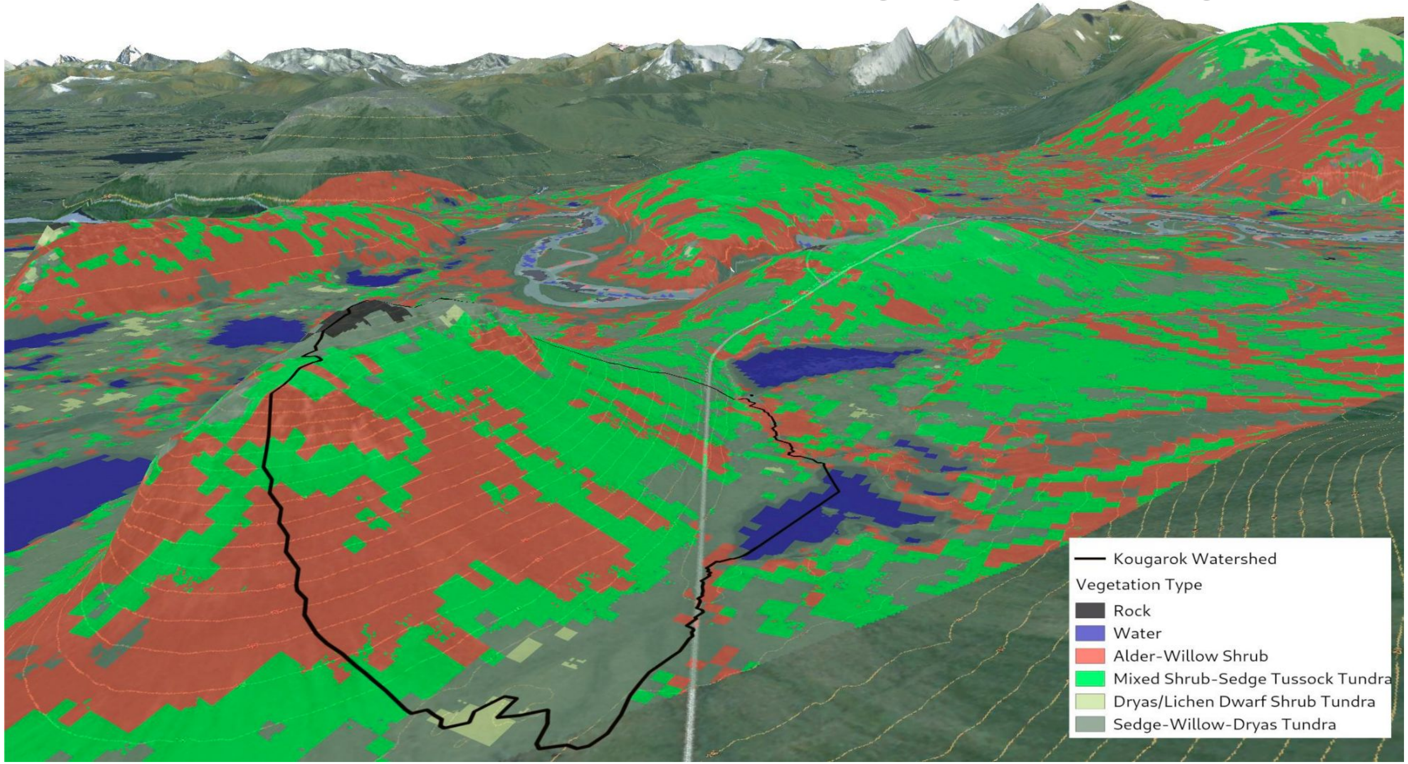


In situ data from field measurement activities inform the development of wide-scale maps of vegetation distribution through inference using remote sensing data as surrogate variables, and relationships with environmental controls can be extracted

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](https://doi.org/10.3390/rs11010069).

Satellite Data Analytics Enables Within-Season Crop Identification

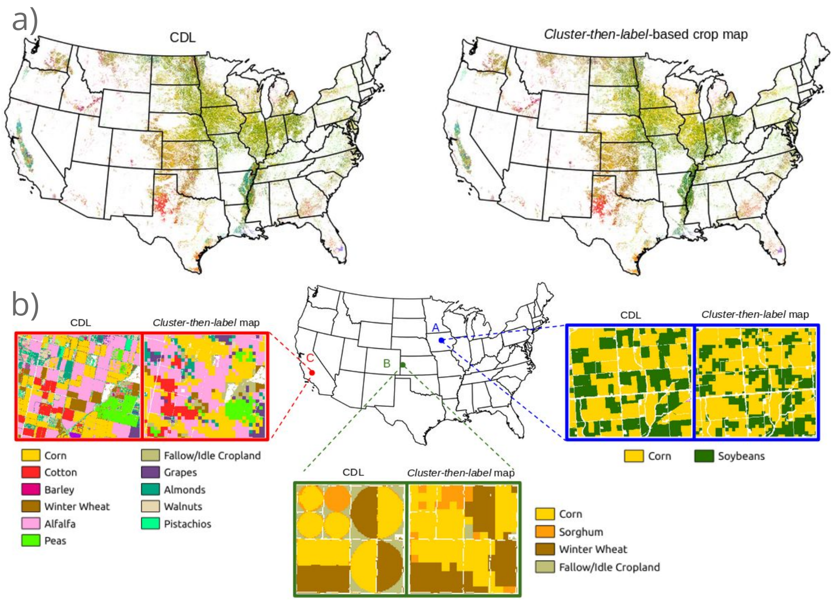
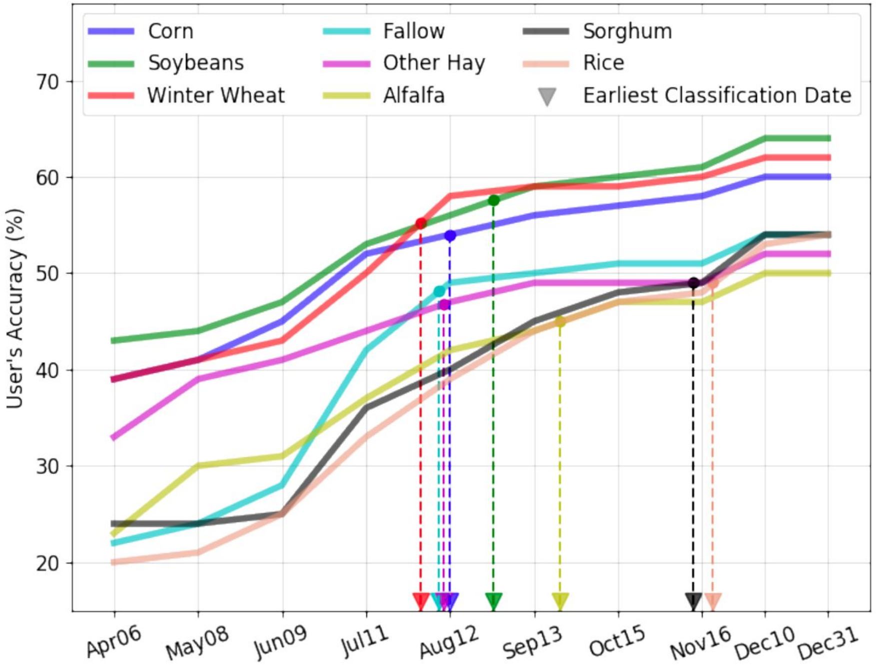


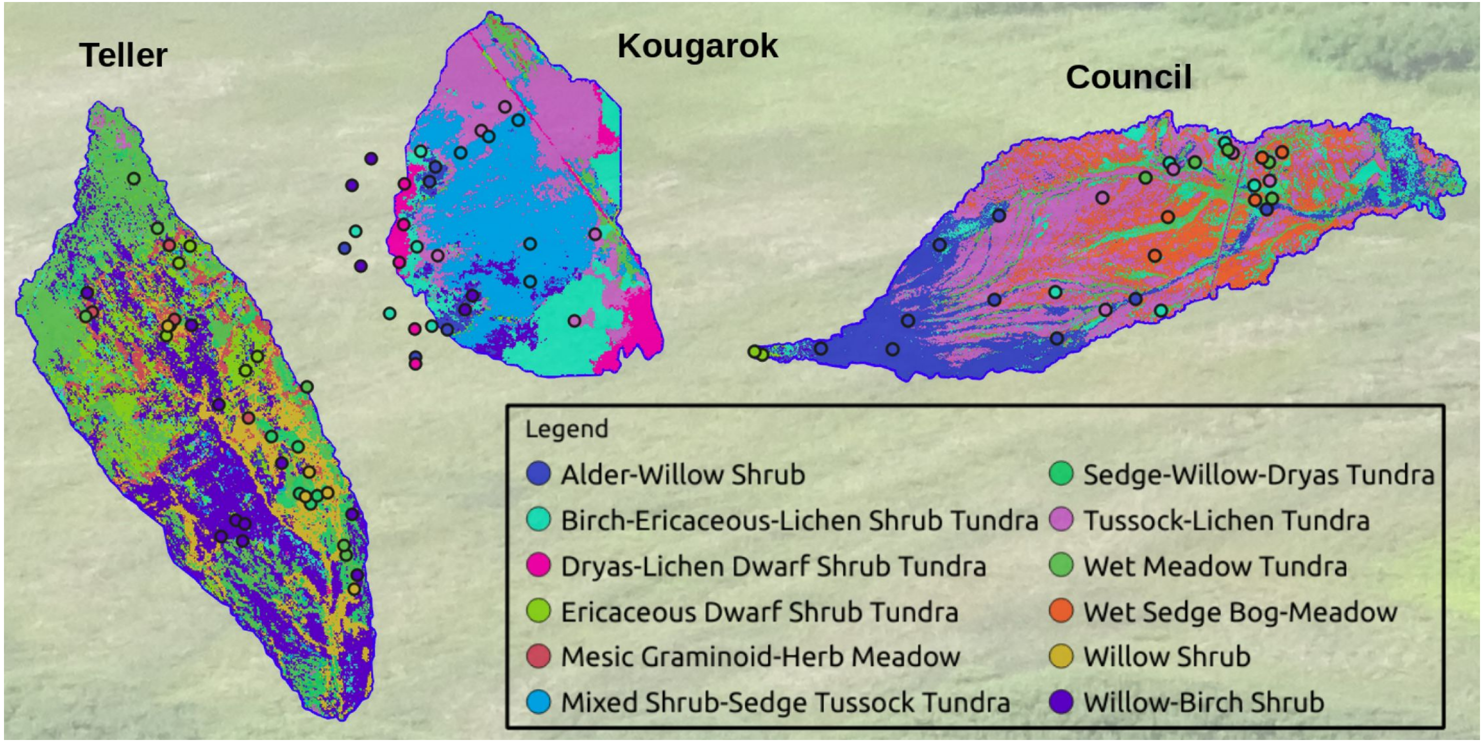
Figure: a) Comparison of cluster-then-label crop map with USDA Crop Data Layer (CDL) shows similar patterns at continental scale. b) Good spatial agreement is found at three selected regions, but cluster-then-label crop maps lack sharpness at field boundaries due to coarser resolution of MODIS data.

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



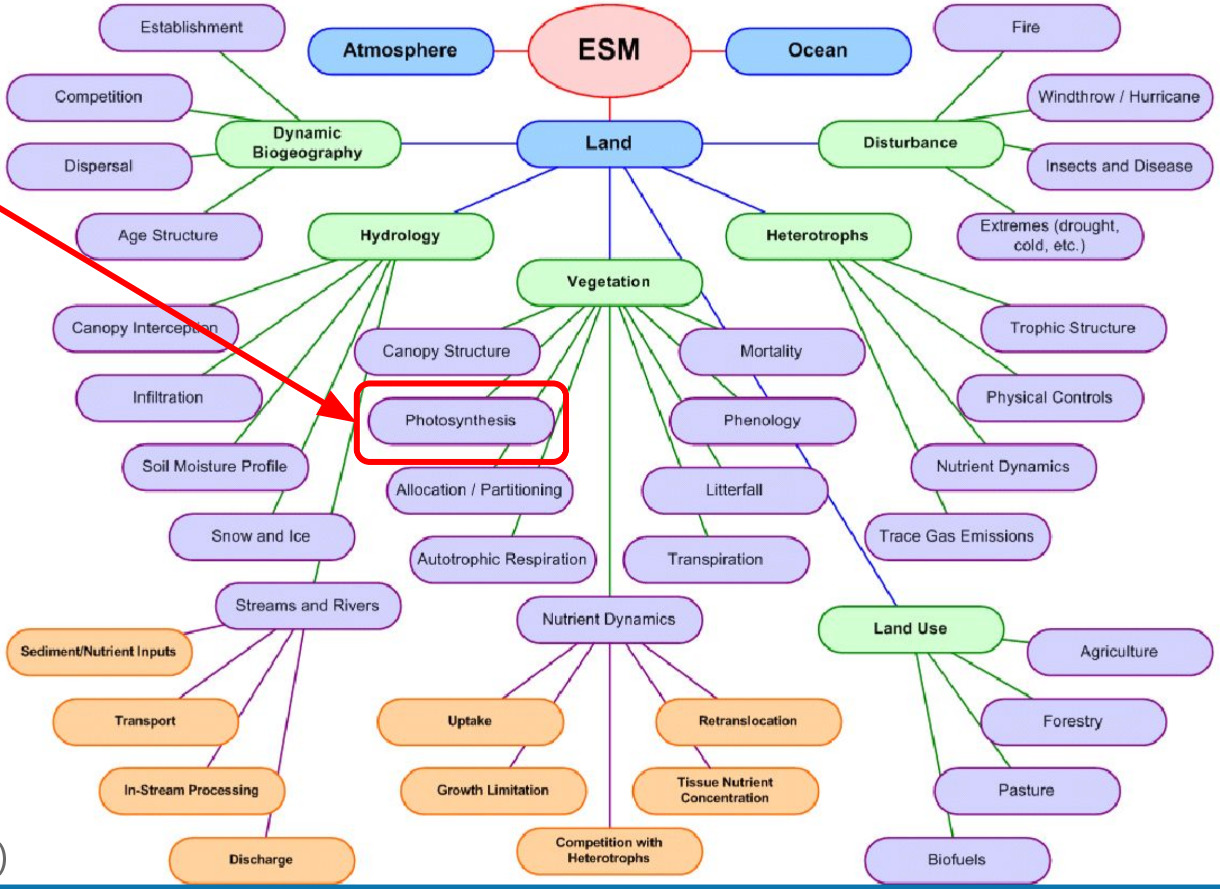
At the watershed scale, vegetation community distribution follows topographic and water controls. At a fine scale, nutrients limit the distribution of vegetation types.

(Konduri et al., in prep.)

Hybrid ML/Process-based Modeling for Terrestrial Modeling

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

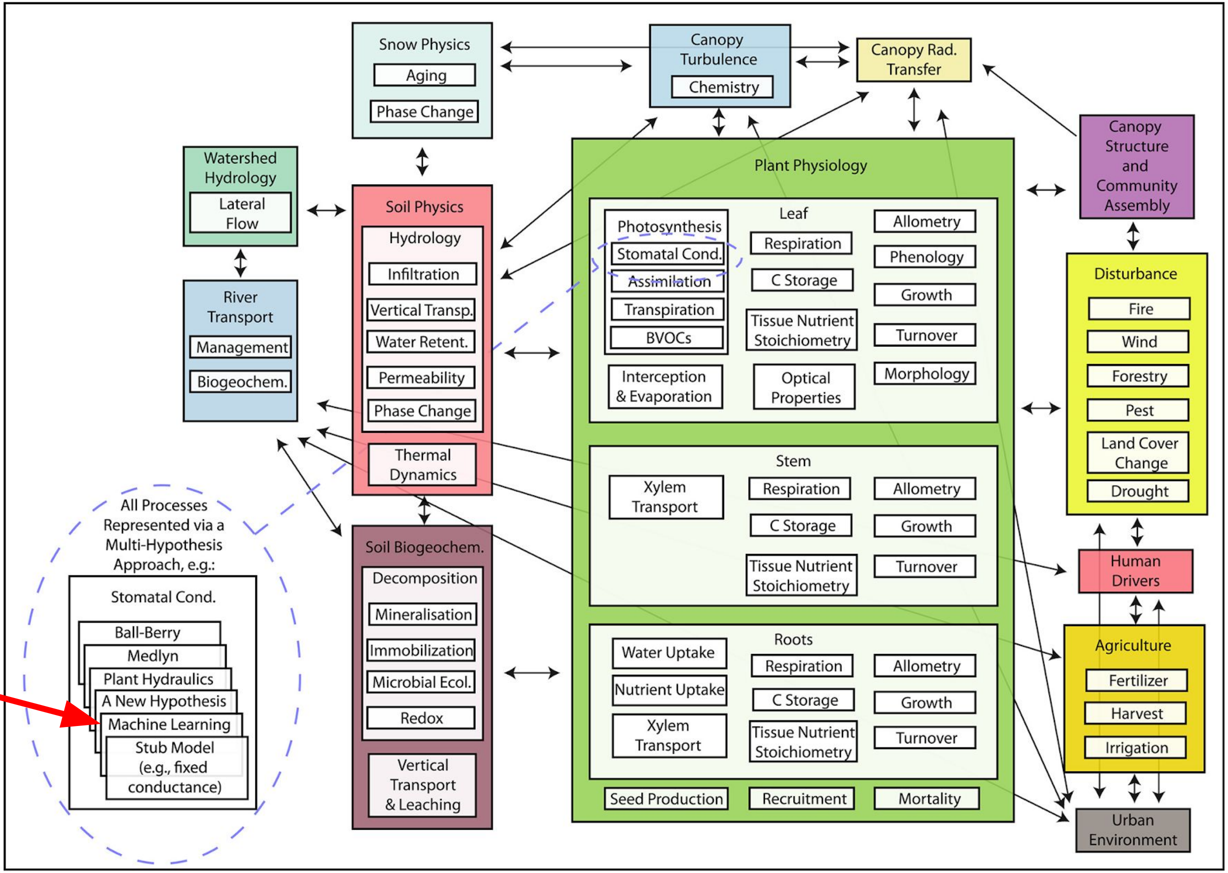
- Multiple hypotheses
- Many leaf-level measurements
- Most computationally intensive part of the land model



(Figure from P. E. Thornton)

Hybrid ML/Process-based Modeling for Terrestrial Modeling

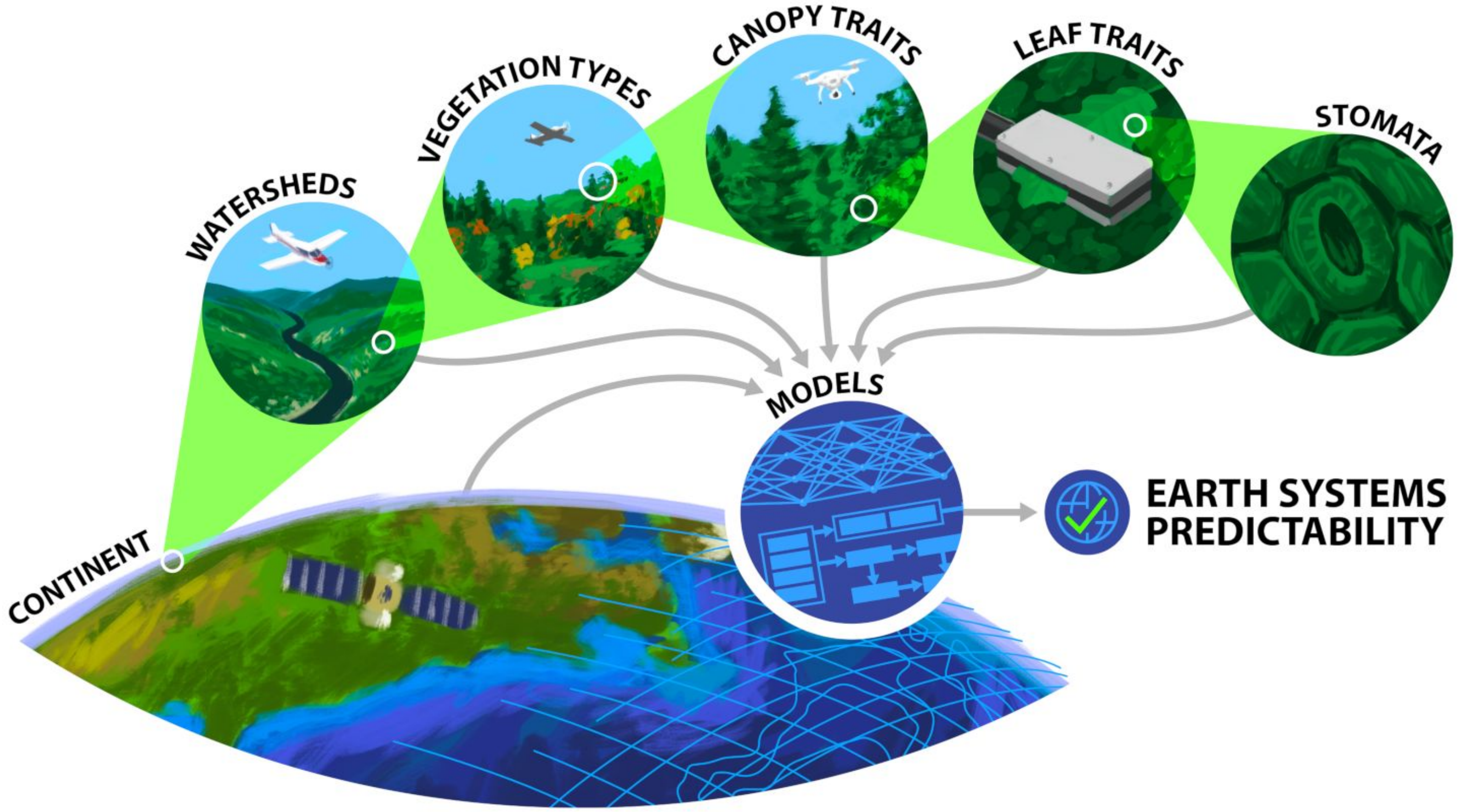
Individual processes can be represented by a multi-hypothesis approach, and ML provides an opportunity for a data-derived hypothesis that can be further explored or used to calibrate other hypotheses, when sufficient data are available.



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

(Fisher and Koven, 2020)

Spanning Spatial & Temporal Scales for Ecosystem Modeling



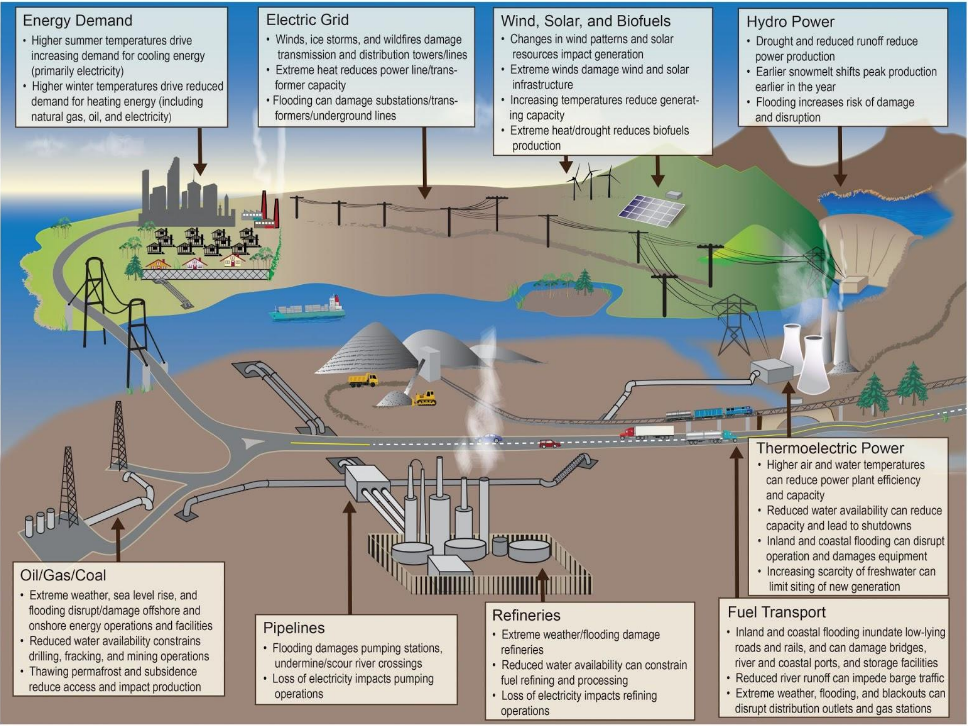
AI FOR SCIENCE TOWN HALL

Earth and Environmental Sciences

Forrest M. Hoffman (ORNL),
Rao Kotamarthi (ANL),
Haruko Wainwright (LBNL),
and the EES Writing Team

U.S. DEPARTMENT OF ENERGY | Office of Science

Grand Challenge #1



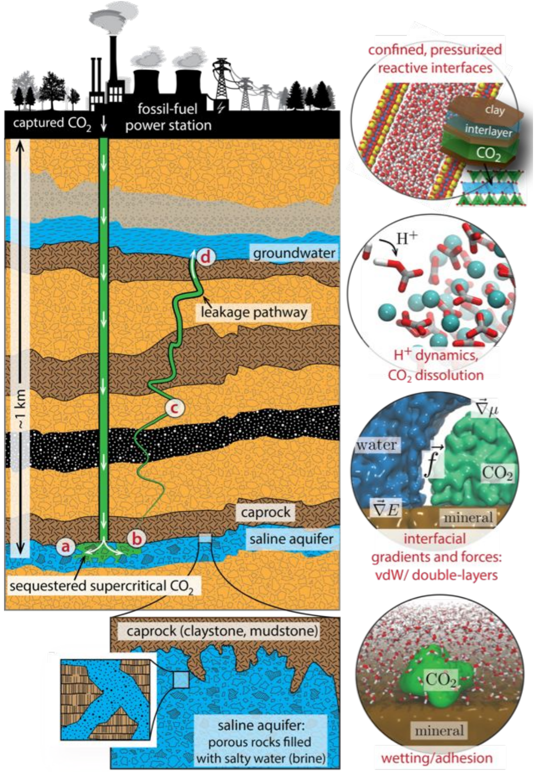
Project environmental risk and develop resiliency in a changing environment

- Increasing frequency of weather extremes and changing environment pose risks to energy infrastructure and the built environment
- Sparse observations and inadequate model fidelity limit the ability to identify vulnerability, mitigate risks, and respond to disasters

Grand Challenge #1

- New tools are needed to accelerate projection of weather extremes and impacts on energy infrastructure
- Building resiliency to address evolving risks will benefit from integration of smart sensing systems, built-for-purpose models, ensemble forecasts to quantify uncertainty, and dynamic decision support systems for critical infrastructure

Grand Challenge #2

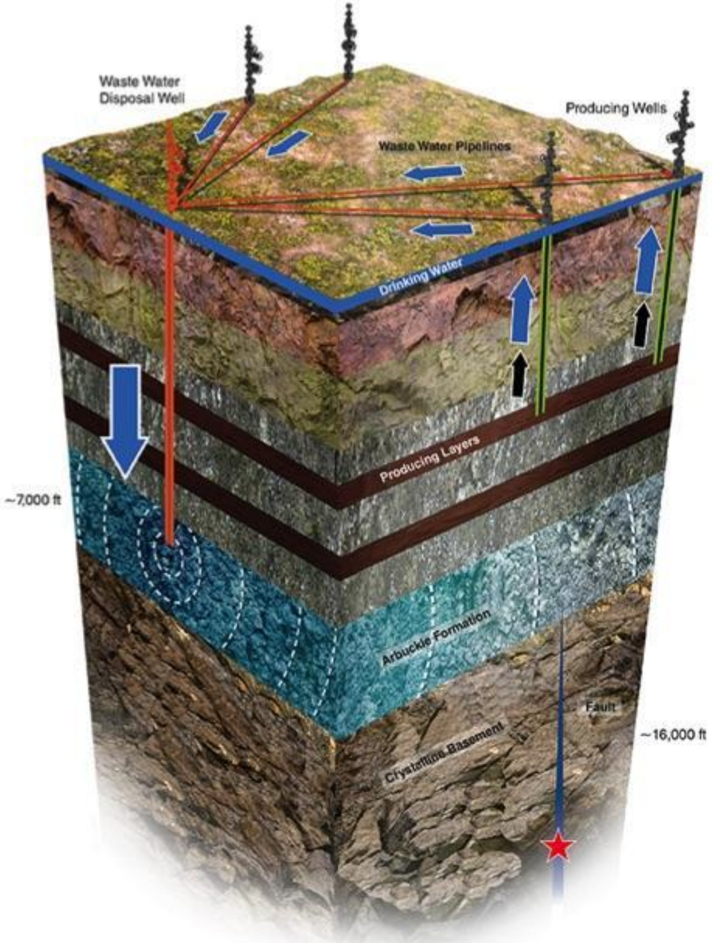


Characterize and modify subsurface conditions for responsible energy production, CO_2 storage, and contaminant remediation

- National energy security and transition to renewable energy resources relies on utilization of subsurface reservoirs for energy production, carbon storage, and spent nuclear fuel storage
- Subsurface data are uncertain, disparate, diverse, sparse, and affected by scaling issues
- Subsurface process models are incomplete, uncertain, and frequently unreliable for prediction

Grand Challenge #2

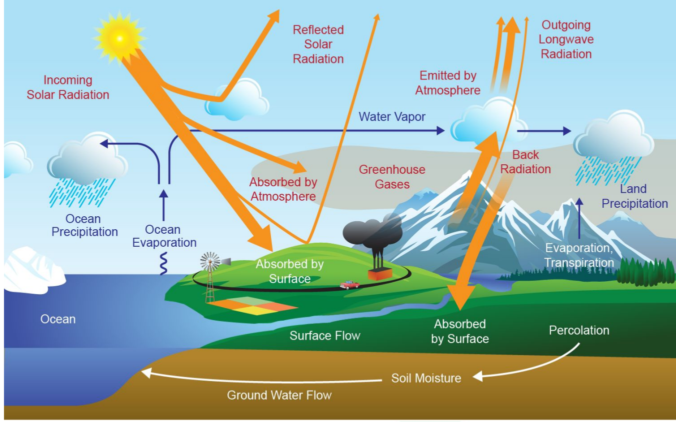
- We need to substantially increase hydrocarbon extraction efficiency, discover and exploit hidden geothermal resources, reduce induced seismicity and other impacts, improve geologic CO_2 storage, and predict long-term fate and transport of contaminants
- Mitigating risks requires improved subsurface characterization and assimilation of real-time data streams into predictive models of geological and ecological processes



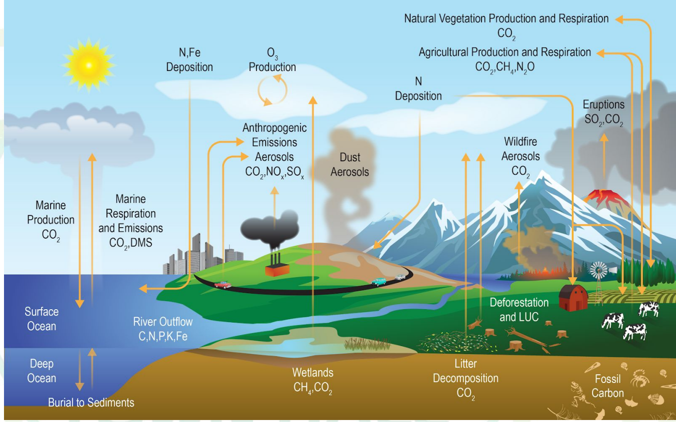
Grand Challenge #3

Develop a predictive understanding of the Earth system under a changing environment

- To advance the nation's energy and infrastructure security, a foundational scientific understanding of complex and dynamic hydrological, biological, and geochemical processes and their interactions is required (across atmosphere, ocean, land, ice)
- Knowledge must be incorporated into Earth system models to project future climate conditions for various scenarios of population, socioeconomics, and energy production and use



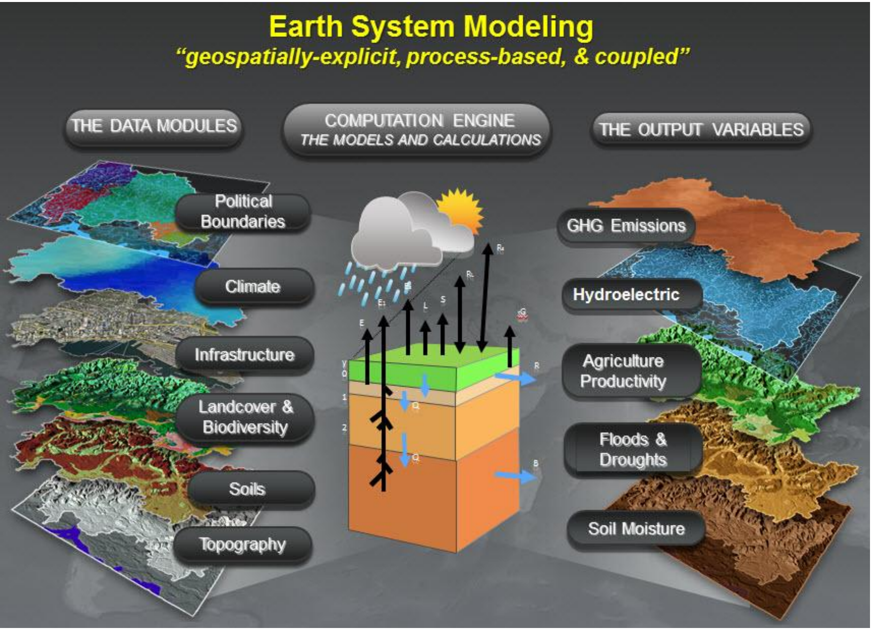
Energy & Water Cycles



Carbon & Biogeochemical Cycles

Washington DC Town Hall
October 22-23

Grand Challenge #3

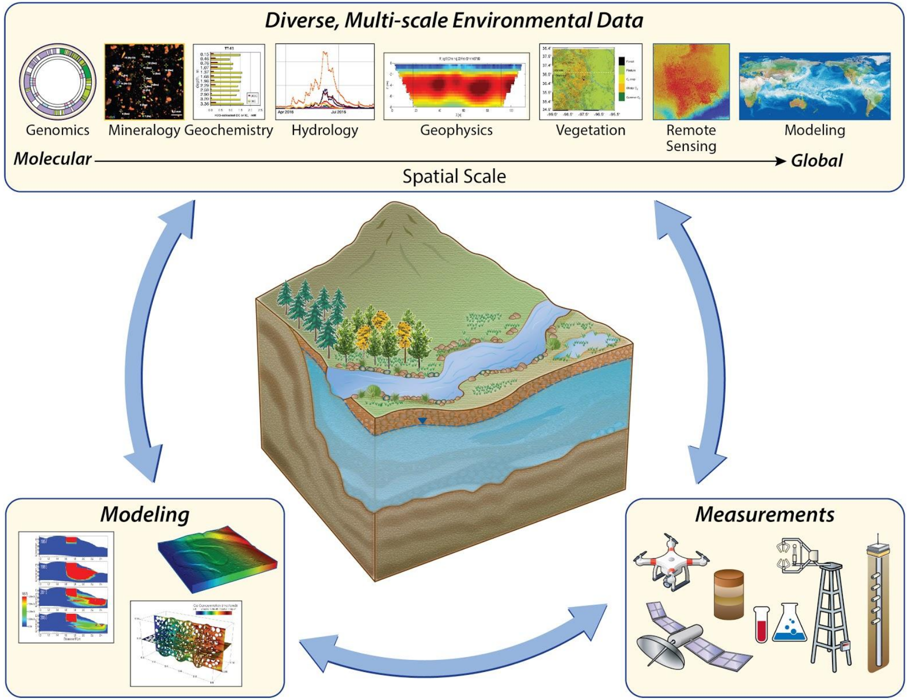


- Accurate predictions are needed to quantify changes in atmospheric and ocean circulation and weather extremes, to close the carbon cycle, and to understand responses and feedbacks of human, terrestrial, and marine ecosystems to environmental change
- Advances in genomics and bioscience data need to be leveraged to provide detailed understanding of plant–microbial interactions and their adaptations and feedbacks to the changing environment

Grand Challenge #4

Ensure global water security under a changing environment

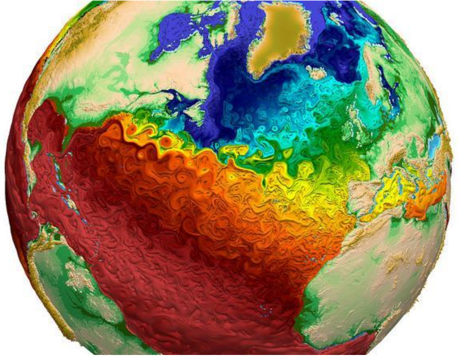
- Water resources are critical for energy production, human health, food security, and economic prosperity
- Water availability and water quality are impacted by environmental change, weather extremes, and disturbances such as wildfire and land use change



Grand Challenge #4

- Methods are needed to integrate disparate and diverse multi-scale data with models of watersheds, rivers, and water utility infrastructure
- Predictions of water quality and quantity require data-driven models and smart sensing systems
- Water resource management must account for changes in weather extremes, population, and economic growth

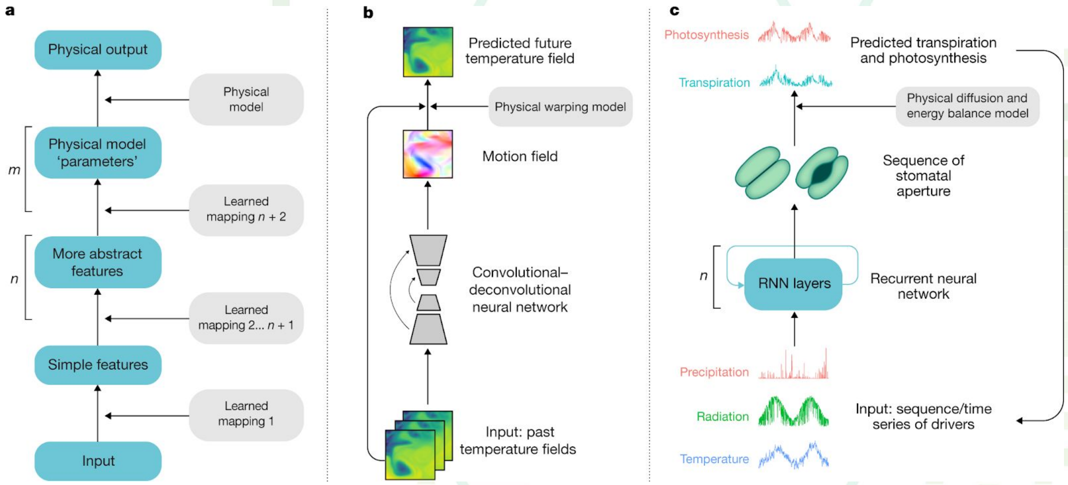
Accelerating Development



The near-term (5–10 years) priorities are to:

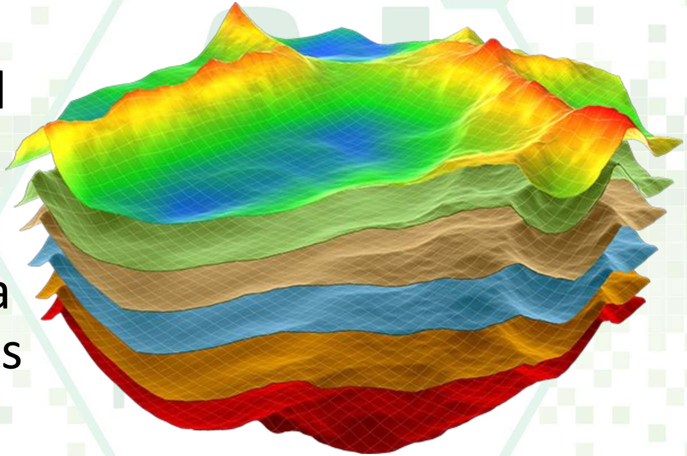
- Develop hybrid process-based/AI modeling frameworks for Exascale systems
- Develop strategies for mapping hybrid components on GPU/CPU based on computational density and communications patterns
- Develop physics / chemistry / biology-constrained ML
- Develop explainable AI and ML methods for hypothesis generation and testing

Hybrid Approaches to Earth Science Simulation (Reichstein et al., 2019)



Expected Outcomes

- Model testbeds and surrogate models are expected to yield insights into process understanding across all Grand Challenges
- Data-driven and physics-constrained hybrid models are expected to stimulate new discovery and bridge space and time scales
- Integrated models of Earth system processes and energy/built infrastructure will enhance national energy and water security through simulation
- AI methods will enable effective use of large data streams for energy production, predictive process understanding, and environmental resiliency



AI4ESP

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

DOE Environmental System Science (ESS) PI Meeting

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

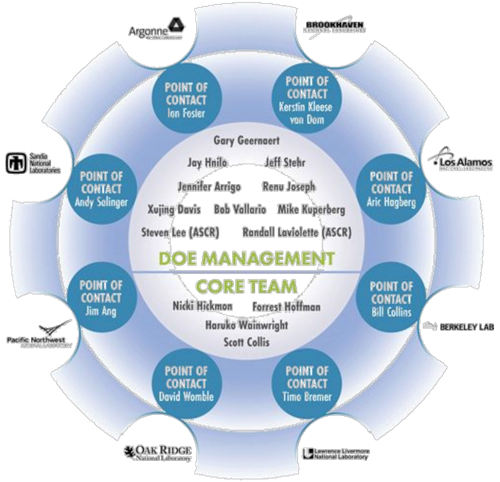
- Data Acquisition
- Neural Networks
- Surrogate models and emulators
- Knowledge-Informed Machine Learning
- Hybrid Modeling
- Explainable/Interpretable/Trustworthy AI
- Knowledge Discovery & Statistical Learning
- AI Architectures and Co-design

Workshop Report

- Chapters for each session have been written and reviewed
- Summary chapters are being written now
- Final review and approval expected soon after July 1, 2022

AMS Special Collection

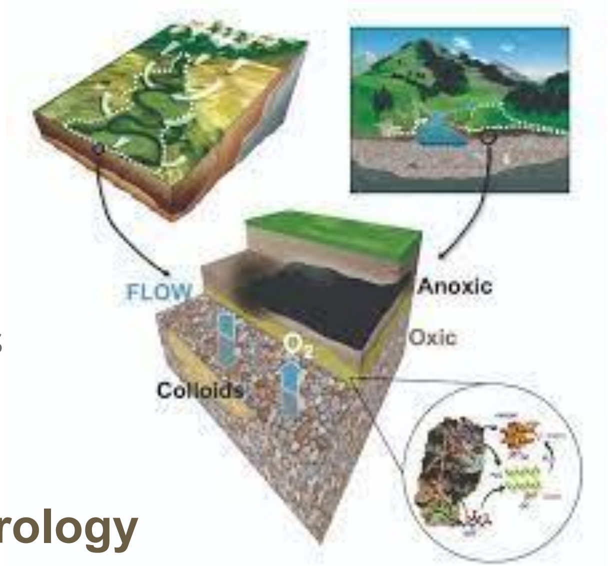
- Proposal recently submitted for all AMS journals



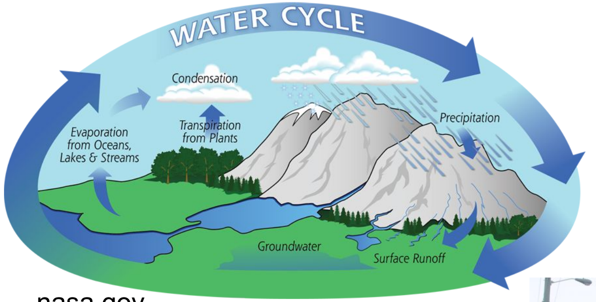
AI4ESP

AI4ESP White Papers: Earth System Predictability Topics

- **Watershed science**
  - Hydro-Biogeochemistry, Soil biogeochemistry
  - Water quality
  - Lab-to-field, field-to-regional scale analysis
  - Experimental data, sensor networks (rapid responses), and experimental/network designs



ess.science.energy.gov



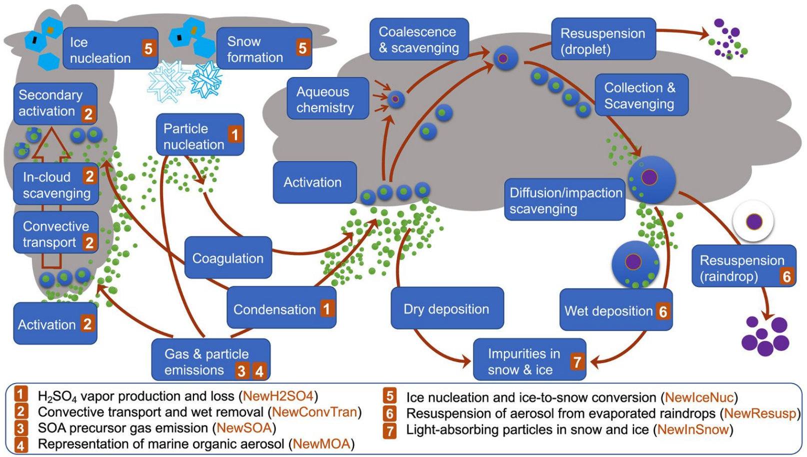
nasa.gov

- **Hydrology**
  - Water resources
  - Precipitation-induced hazards (floods etc)
  - Weather/hydrological monitoring
  - Groundwater to surface water models
  - Mountain hydrology
  - Regional to continental scale

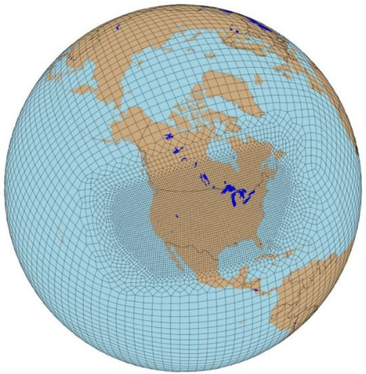
climate.gov

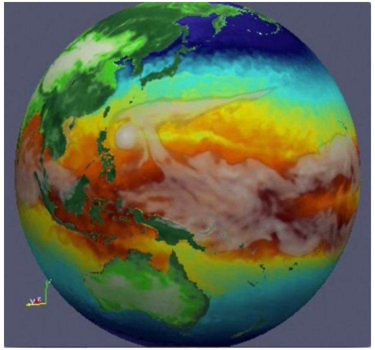
AI4ESP White Papers: Earth System Predictability Topics

- **Atmospheric Modeling**
  - Convection and turbulence
  - Surface Fluxes
  - Radiation
  - Model Tuning
  - General concepts that can generalized to other ESMs components



e3sm.org





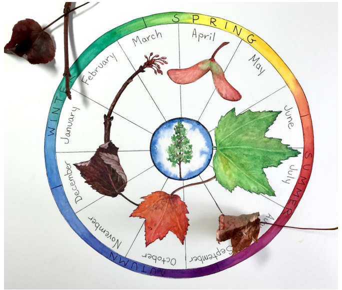
e3sm.org

- **Aerosols and Clouds**
  - Cloud Classification
  - Aerosol cloud interactions

AI4ESP White Papers: Earth System Predictability Topics

- **Land Modeling**
  - Agriculture / Crops
  - Leaf Phenology
  - Streamflow / Water Availability
  - Wildfire
  - Satellite Data Assimilation
- **Ecohydrology**
  - Stomatal Conductance / Photosynthesis
  - Plant Hydraulics and Growth
  - Evapotranspiration
  - Soil Moisture
  - Soil Hydrology

Getty Images



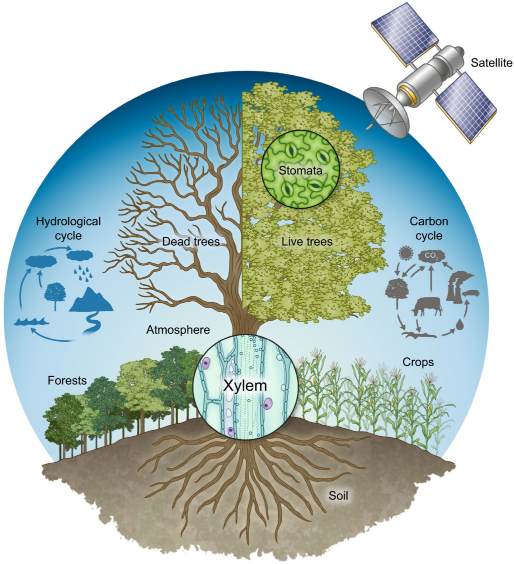
Adkins Arboretum

wallpaperbetter.com

ABC7 News

drought.gov

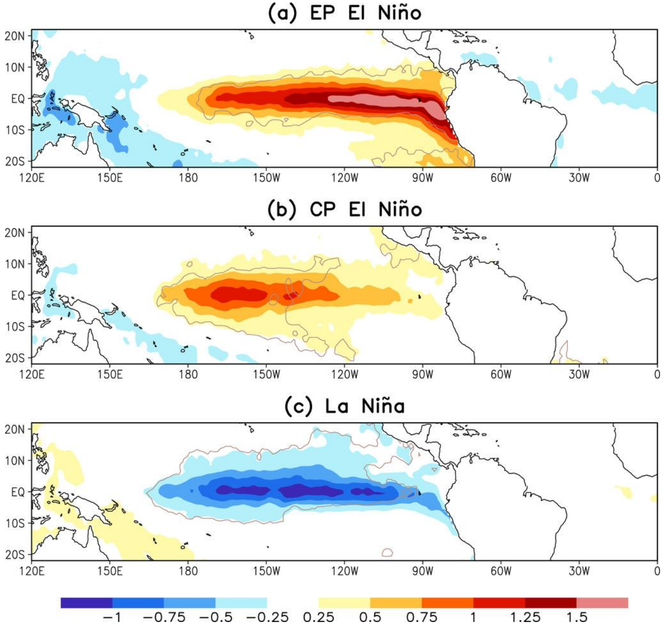
Nature



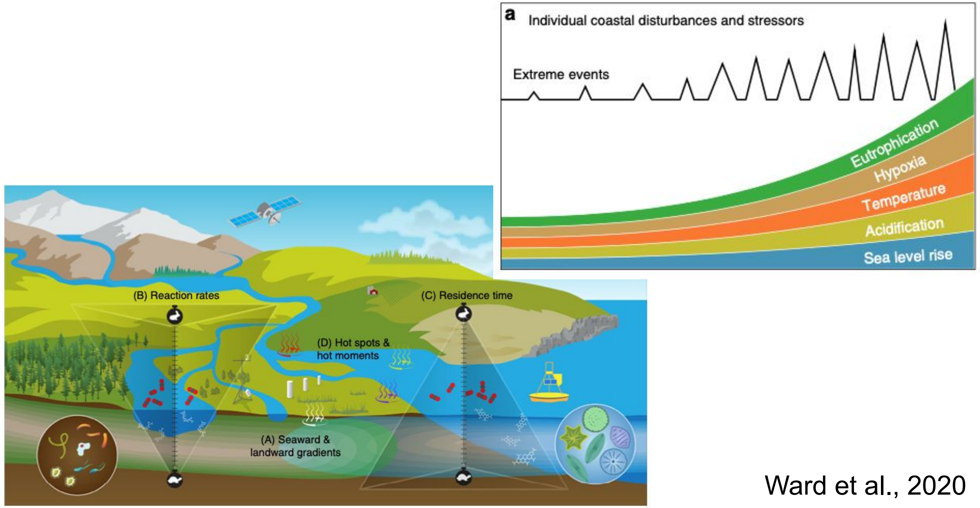
McDowell et al. (2019)

AI4ESP White Papers: Earth System Predictability Topics

- **Climate variability and Extremes**
  - TCs, ARs, Compound/Cascading events
  - Predictability
  - Circulation/climate variability (ENSO, NAO etc)
  - Telecommunication



Wang et al., 2014

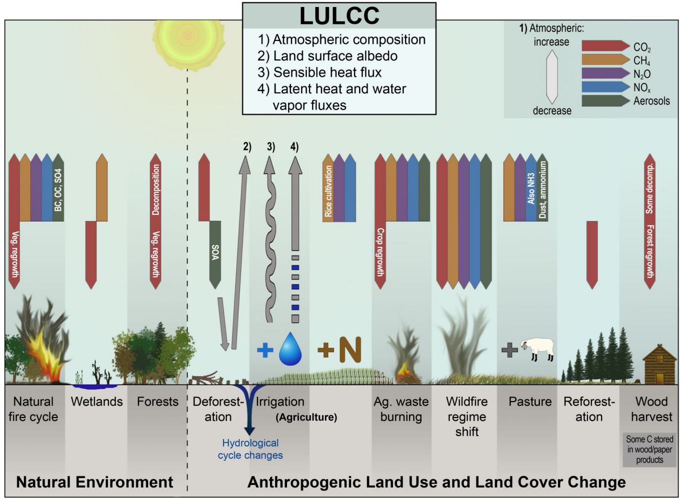


Ward et al., 2020

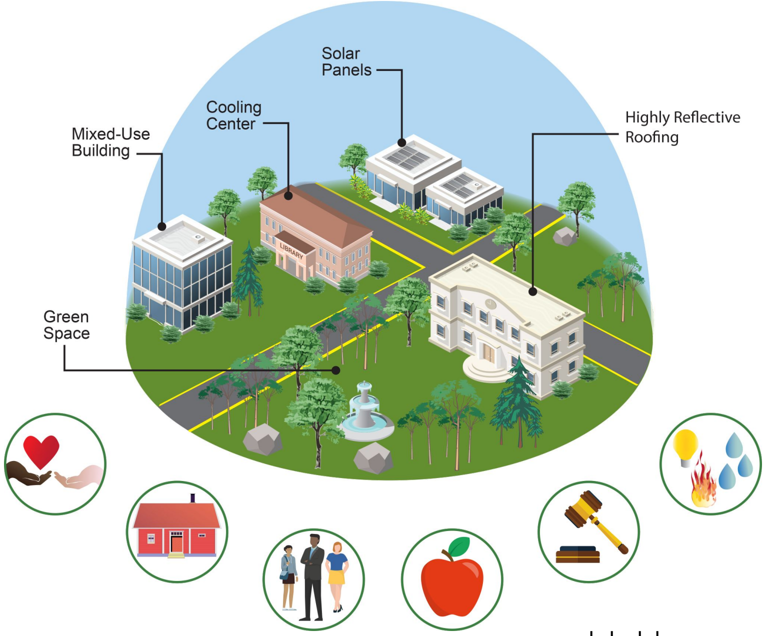
- **Coastal dynamics, Ocean/Ice**
  - Ocean/land/ice interface
  - Sea-level rise, storm surge
  - Coastal ecosystem/carbon cycling

AI4ESP White Papers: Earth System Predictability Topics

- **Human Systems and Dynamics**
  - Human activities/population
  - Energy-water-land nexus
  - Agriculture
  - Urban environment
  - Land use/cover changes



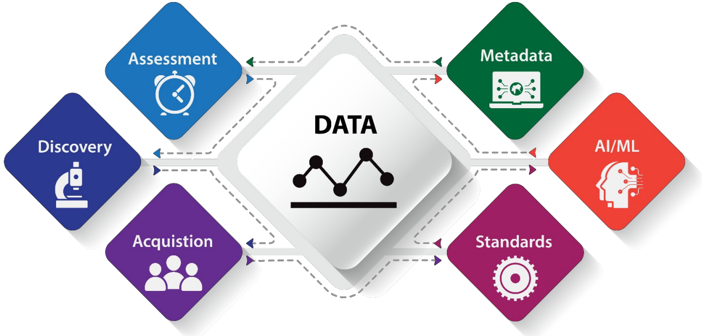
globalchange.gov



globalchange.gov

AI4ESP: Cross-cutting Topics

- Data Acquisition to Distribution
- Neural Networks
- Surrogate Models and Emulators
- Knowledge-Informed Machine Learning
- Hybrid Modeling
- Explainable and Trustworthy AI
- Knowledge Discovery & Statistical Learning
- AI Architectures and Co-design



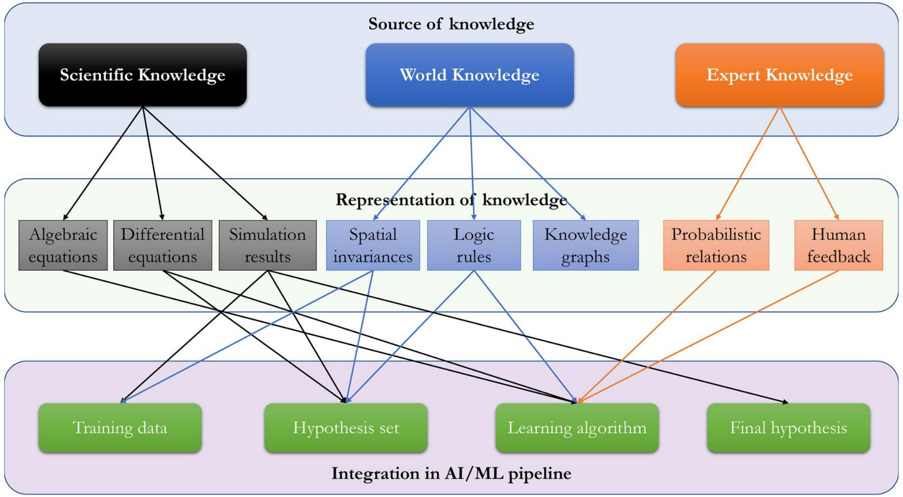


Figure adapted from Von Reuden et al. (2021)

Highlights Across All Sessions

Science

- AI/ML can accelerate next-generation integrated models to support decision-making that incorporate complex natural and human processes at sufficient resolutions
- Broad consensus on need for deep integration of process-based and ML models (hybrid models)
- Challenges: scaling, sub-grid representation, model calibration/UQ, extreme events, human systems
- Data gaps are vast – more observations informed by model needs, AI-ready products
- Results must be robust, explainable, & trustworthy

Data, Software, Infrastructure

- Need benchmark data and model intercomparison approaches
- Computational infrastructure for integration of process & ML models, data assimilation and synthesis
- Use ML to accelerate data-model and model-observation pipelines

Culture

- Workforce development across domain and computational scientists
- Interdisciplinary research centers focused on AI4ESP

AI-Constrained Ecohydrology for Improving Earth System Predictions

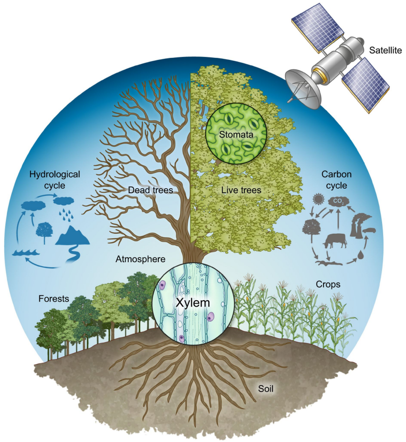
Collaboration among ORNL, LANL, Penn State, et al.

Contact: Forrest M. 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 are 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)

The Future is Bright for AI/ML in Earth System Science

A Convergence of New Technology, Explosive Data Growth, and Free Tools

- High performance computing (exascale in big centers and commercial cloud)
- Large data storage resources (commercial and on-premise cloud)
- High speed networks (e.g., ESnet) and data movement technologies (Globus)
- Satellites (shoebox CubeSats) and airborne (drones) platforms
- Cheap (free!) and easy-to-use ML tools (PyTorch, Keras, Scikit-Learn)

Future Applications Could Revolutionize Our Understanding and Ability to Predict

- Poorly understood processes and mechanisms can be mimicked with adequate amounts of data and advanced ML techniques
- Explainable AI and systematic approaches to modeling could lead to new scientific discoveries and improved understanding of the Earth system
- Predictions of complex, nonlinear, large-scale phenomena and natural hazards could be predicted with increasing accuracy

THANK YOU

U.S. DEPARTMENT OF ENERGY

AI4ESP