U.S. DEPARTMENT OF ENERGY | Office of Science

CESD Cyberinfrastructure Working Groups

Environmental System Science (ESS) PI Meeting

Bolger Center, Potomac, Maryland, USA

April 29, 2019

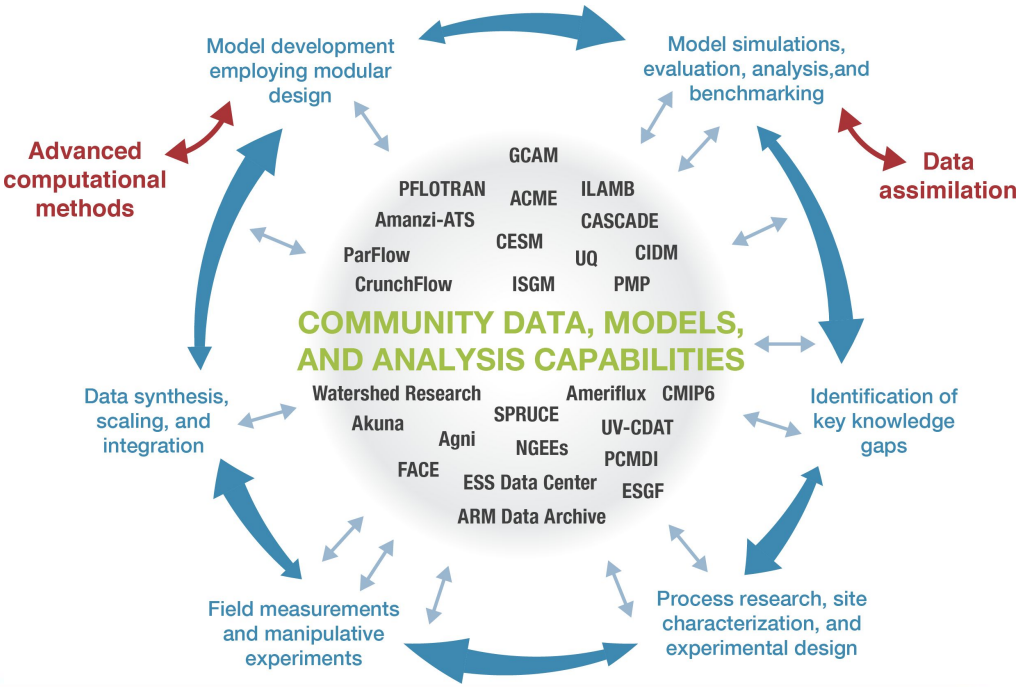
Model–Data Integration

Leads: Forrest M. Hoffman (ORNL) and Xingyuan Chen (PNNL)

Participating Team Members: Bhavna Arora, Bob Bolton, Ben Bond-Lamberty, Eoin Brodie, Laura Condon, Beth Drewniak, Maoyi Huang, Colleen Iversen, Elchin Jafarov, Jitu Kumar, Umakant Mishra, Bill Riley, Joel Rowland, Tim Scheibe, Shawn Serbin, Xiaoying Shi, Peter Thornton, Haruko Wainwright, and Anthony Walker

Model–Data Integration Scope

- Model–data comparison
- Uncertainty quantification (UQ) & data assimilation (DA)
- Management of model results and observational data (with Data Management Working Group)
- Geospatial and remote sensing data analysis
- Data analytics methods and techniques, e.g.,
  - –Data mining
  - –Neural networks
  - –Genetic algorithms
  - –Other machine learning techniques
  - –Visual analytics
- Model–data fusion



Short-Term Goals (2016–2019)

- **Encourage archiving and versioning of publications, data, models, and software tools**
  - Document best practices jointly with other Working Groups
  - Versioning for synthesized & combined data sets (e.g., FLUXNET2015)
  - Digital Object Identifiers (DOIs) for pubs, data, models, and tools
- **Identify available scientific workflows, UQ frameworks, and model–data tools (e.g., ESGF, UV-CDAT, PEcAn, ILAMB)**
  - What workflows are people using and when does one assign a DOI?
  - Develop a user survey to capture initial information
- **Initiate subgroup on geospatial analysis and remote sensing**
  - Google Earth Engine and similar useful tools are rapidly evolving
  - Identify tools and resources for geospatial data analytics
  - Individual community projects have pockets of expertise (e.g., ARM)
- **Advocate for open and standard data formats & conventions**
  - Engage in groups to develop standards and educate users
  - Deploy tools/APIs to transform observational data into model formats
  - Foster API consistency across multi-agency/federated data centers

Short-Term Goals (2016–2019) (continued)

- **Support community activities to make observational data quickly and easily available for model evaluation (e.g., ILAMB)**
  - Sponsor working groups focused on individual data sets and corresponding model metrics
  - Make AmeriFlux, NGEE Arctic, NGEE Tropics, SPRUCE, FACE, and similar data sets rapidly available to modelers by creating benchmarks
- **Organize disparate uncertainty quantification (UQ) activities to foster collaboration and establish best practices**
  - Standardize methods and approaches
  - Create workflows for common modeling frameworks

Progress Since 2016

- **Geospatial analysis and remote sensing**
  - 2017 white paper : **Geospatial Science to Inform Land Surface Models** (Mishra, Serbin, Wainwright, Kumar, Huang, and Chen)
- **Model–data comparison and benchmarking**
  - International Land Model Benchmarking (ILAMB) Workshop and Tools
  - Soil Carbon Dynamics Working Group for data synthesis (2018)
- **Archiving of publications, data, models, & software tools and open data standards & conventions**
  - Data management plan plus software productivity and sustainability requirements for CESD projects
  - Work with new ESS-DIVE
  - Draw on work of ESIP, ISMC, CSDMS, EarthCube
- **Uncertainty quantification (UQ) & data assimilation (DA)**
  - Akuna-CLM, DART-PFLOTRAN, PEcAn
- **Scientific workflows and model & data analysis tools**
  - Jupyter notebooks
- **Community outreach**
  - 2018 AGU Fall Meeting sessions on “Computational Methods and Tools for Model–Data Integration” and “Big Data in the Geosciences”

2018 AGU Fall Meeting

- **Computational Methods and Tools for Model–Data Integration - *F. M. Hoffman (ORNL), X. Chen (PNNL), T. Xu (Utah State U.), and H. Kim (U. Tokyo)***
  - A Bayesian Approach to Soil Biogeochemical Model Comparison - H. W. Xie and S. D. Allison (UC Irvine)
  - (Invited) Efficient Surrogate Modeling Methods to Advance Model-Data Integration - D. Lu (ORNL)
  - Environmental Classification at Scale to Support Global Farming Decisions - P. Salvatore La Rosa et al. (Monsanto Company)
  - Generating Improved Estimates of Streamflow Using Model Averaging of Downscaled Runoff Products Under Uncertainty - M. K. Kallio (Aalto University) et al.
  - Machine Learning Application on Closing Data Gaps in Groundwater Measurements - H. Ren et al. (PNNL)
  - (Invited) Towards improved standardisation of model evaluation using modelevaluation.org - G. Abramowitz (University of New South Wales)
  - Multi-site Critical Zone Process Understanding through Standardized and Automated Data Ingestion and Model-data Coupling - R. Versteeg (Subsurface Insights) et al.
  - Using Sensitivity Analysis as a Tool to Determine the Need for Regeneration of Hydrological and Biogeochemical Predictions - B. Arora et al. (LBNL)

2018 AGU Fall Meeting

- **Big Data in the Geosciences: New Approaches to Storage, Sharing, and Analysis - C. H. David (NASA), F. M. Hoffman (ORNL), H. Alemohammad (Radiant Earth), S. K. Kim (LLNL)**
  - Deep Learning on the Sphere: Convolutional Neural Network on Unstructured Mesh - C. M. Jiang (UC Berkeley and LBNL) et al.
  - (Invited) Lessons Learned in Creating Big Science Data Analysis Solutions for the Cloud - T. Huang (NASA/JPL)
  - NASA Archives in the Cloud with Cumulus - Lauren Frederick (Element 84)
  - (Invited) EarthInsights: Parallel Clustering of Large Earth Science Datasets on the Summit Supercomputer - S. Sreepathi (ORNL) et al.
  - Synthesizing Earth System Model Behavior: From Petabytes to Kilobytes - P. J. Gleckler et al. (LLNL/PCMDI)
  - (Invited) Beyond netCDF: Cloud Native Climate Data with Zarr and XArray - R. P. Abernathey (Columbia University)
  - Faults in the Cloud: Distributed Topographic Template Matching of Fault-related Landforms in Shuttle Radar Topography Mission Data using a Cloud-based Processing Framework - R. Sare and G. E. Hilley (Stanford University)
  - (Invited) Image Super-Resolution and Uncertainty Quantification for Earth Science Data on the NASA Earth Exchange AI Platform - T. J. Vandal (NASA) et al.

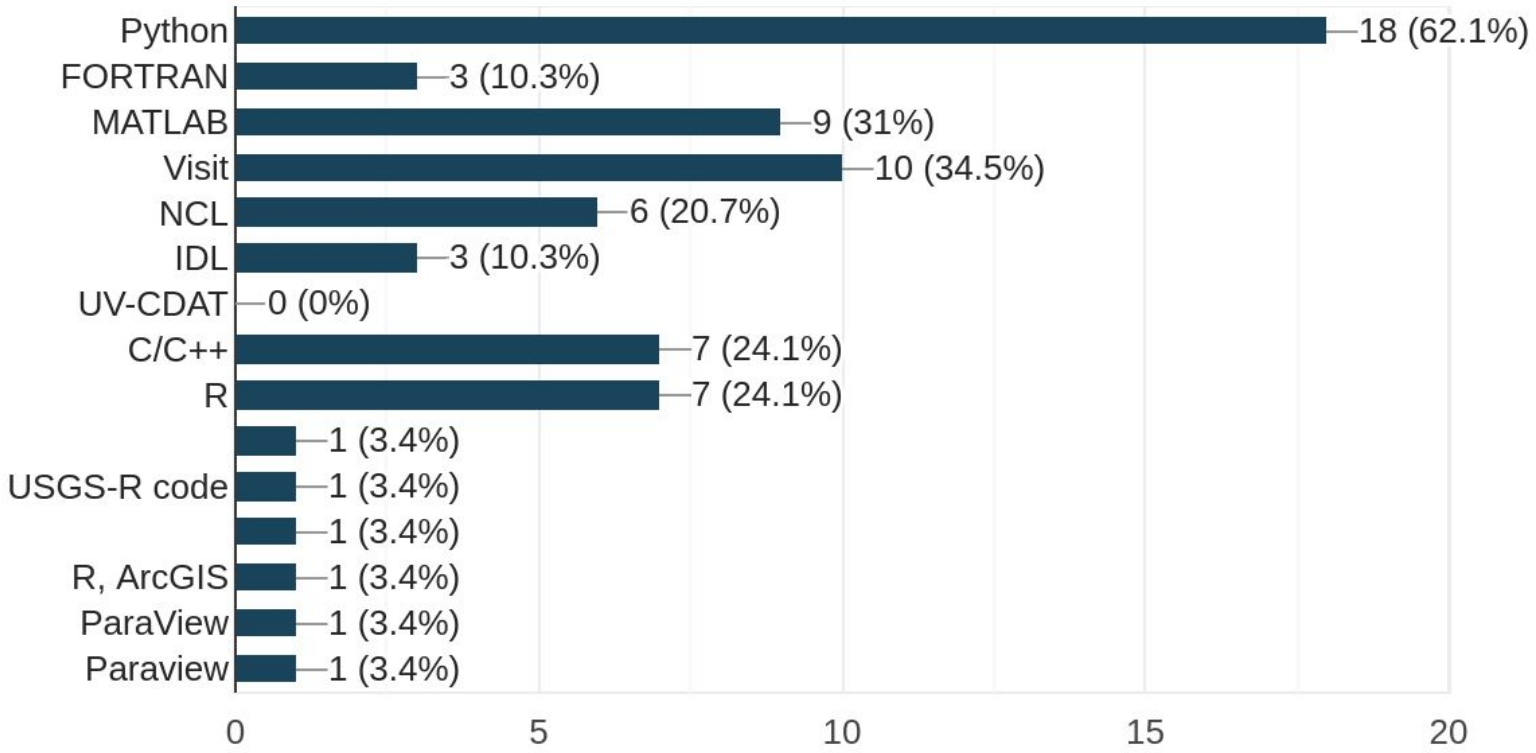
Model–Data Integration Survey Results

- **Community survey on workflows and model–data integration tools was conducted last year**
  - See survey form at https://goo.gl/forms/BdLCDpq1IZckhKPI3
- **Results in following slides come from 30 respondents**
- **Suggestions for topics not covered and future activities will be incorporated for next year**

Model–Data Integration Survey Results

What software tools do you use or prefer for data analysis?

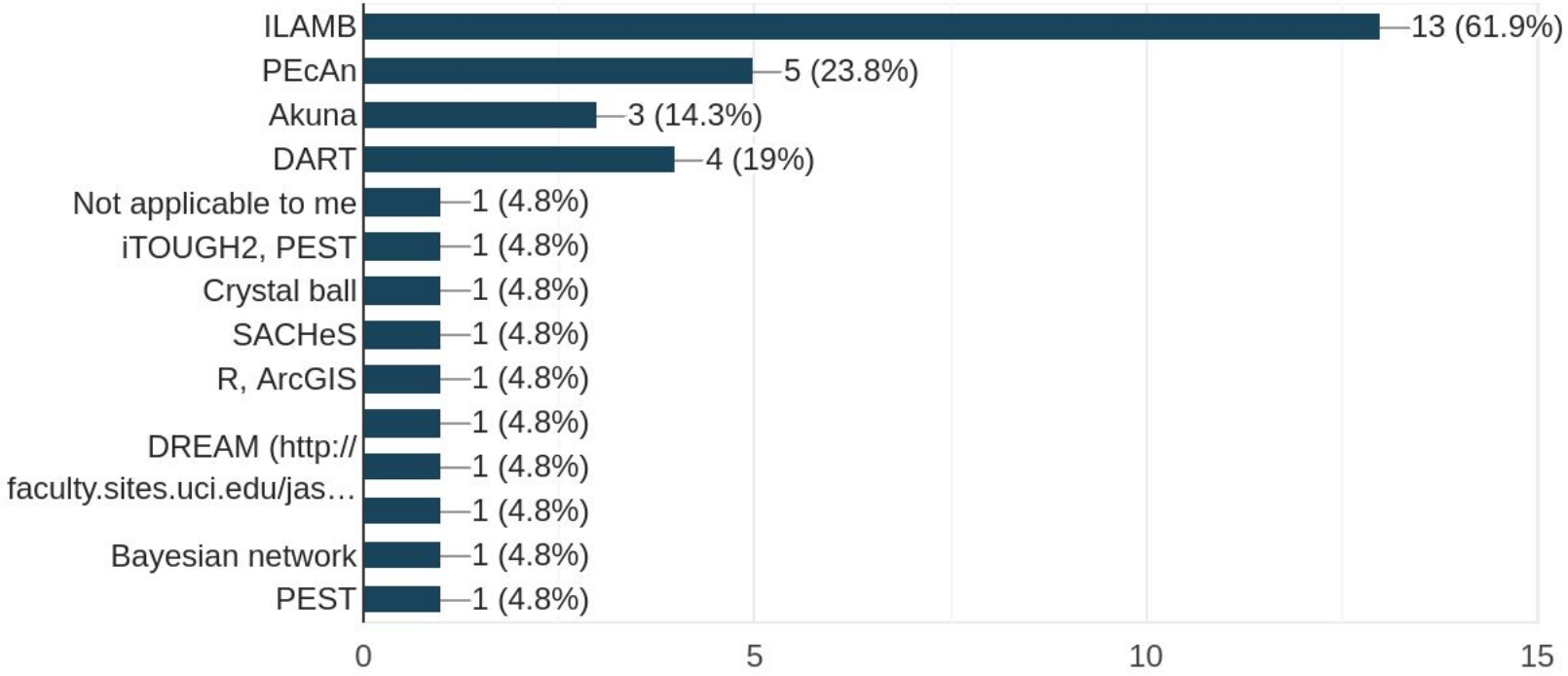
29 responses



Model–Data Integration Survey Results

What software tools do you use or prefer for model evaluation, benchmarking, uncertainty quantification, and data assimilation?

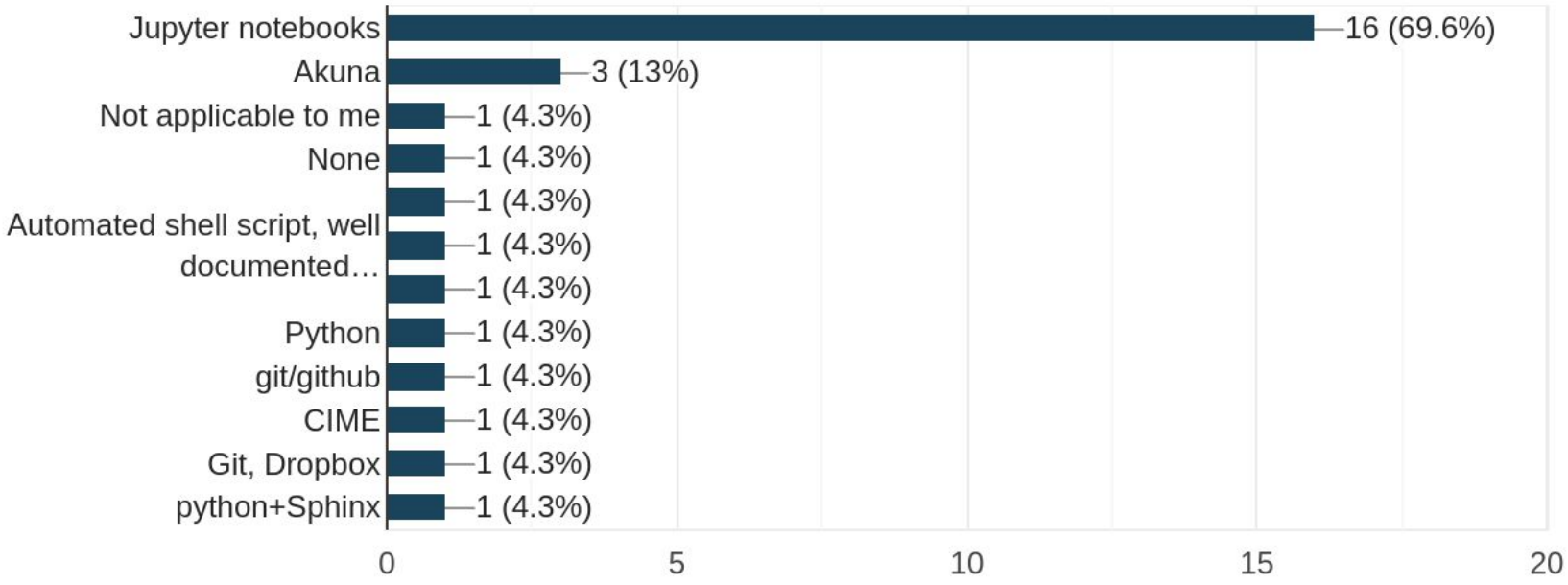
21 responses



Model–Data Integration Survey Results

What software tools do you use or prefer for workflow management, provenance tracking, and archiving?

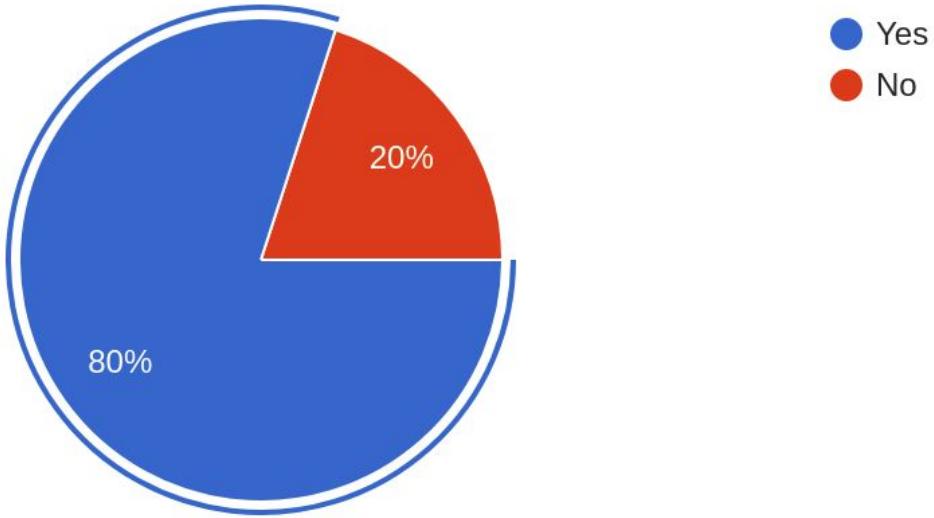
23 responses



Model–Data Integration Survey Results

Do you use high performance computing?

30 responses



Model–Data Integration Survey Results