EESSD – ESS Cyberinfrastructure Working Groups

Model–Data Integration

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

Working Group Members: Bhavna Arora, Ben Bond-Lamberty, Eoin Brodie, Beth Drewniak, Nancy Hess, Colleen Iversen, Elchin Jafarov, Peishi Jiang, Tim Johnson, Jitu Kumar, Douglas Mans, Umakant Mishra, Daniel Ricciuto, Bill Riley, Joel Rowland, Tim Scheibe, Shawn Serbin, Xiaoying Shi, Benjamin Sulman, Peter Thornton, Haruko Wainright, Anthony Walker, Erica Woodburn, Chonggang Xu

Model–Data Integration Working Group Scope

- Model–data comparison, model evaluation and benchmarking
- Uncertainty quantification (UQ) and 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
  - Machine learning, neural network models
  - Genetic algorithms
  - Visual analytics
- Hybrid ML/process-based models & data-driven models

U.S. DEPARTMENT OF ENERGY | Office of Science
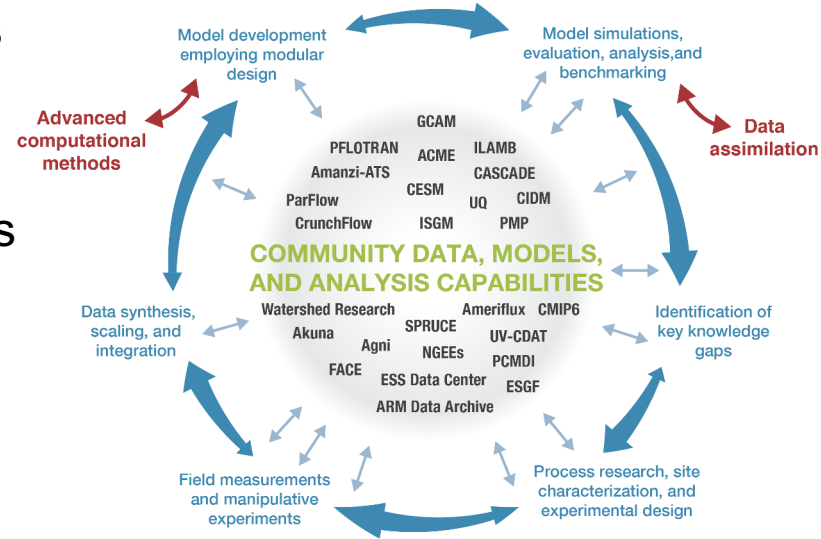
Progress Toward Shared Cyberinfrastructure

- Geospatial analysis and remote sensing
  - 2017 white paper : *Geospatial Science to Inform Land Surface Models* (Mishra, Serbin, Wainwright, Kumar, Huang, and Chen)
  - Shiklomanov, A. N., B. A. Bradley, K. M. Dahlin, A. M. Fox, C. M. Gough, F. M. Hoffman, E. M. Middleton, S. P. Serbin, L. Smallman, W. K. Smith (2019), Enhancing Global Change Experiments through Integration of Remote-sensing Techniques, *Front. Ecol. Environ.*, 17(4):215–224, doi:10.1002/fee.2031.
  - Remote sensing coordination for NGEE Arctic (Kumar, Serbin, et al.)

- Model–data comparison and benchmarking
  - International Land Model Benchmarking (ILAMB) Workshop (2016) and Tool Releases
  - Soil Carbon Dynamics Working Group for data synthesis (2018)
 - RUBISCO-AmeriFlux Working Group for model–data integration (2019)

Progress Toward Shared Cyberinfrastructure

- Archiving of publications, data, models, & software tools and open data standards & conventions
    - Data Management Plan plus software productivity and sustainability requirements for EESSD projects
    - Work with ESS-DIVE and the Earth System Grid Federation (ESGF)
    - Draw on work of community projects and consortia:
        - Federation of Earth Science Information Partners (ESIP)
        - International Soil Modeling Consortium (ISMC)
        - Community Surface Dynamics Modeling System (CSDMS)
        - Pangeo – A community platform for Big Data geoscience

- Uncertainty quantification (UQ) & data assimilation (DA)
    - Akuna-CLM, DART-PFLOTRAN, PEcAn-ILAMB integration

# Progress Toward Shared Cyberinfrastructure

- Scientific workflows and model & data analysis tools
  - Jupyter Notebooks, Google Earth Engine
  - TensorFlow, PyTorch, Scikit-Learn

- Model–Data Integration WG Survey on workflows and tools (2018–2019)

- Community outreach
  - 2018 AGU Fall Meeting sessions on "Computational Methods and Tools for Model–Data Integration" and "Big Data in the Geosciences"
  - 2019 AGU Fall Meeting sessions on "Innovation and Exploration of Observations and Earth System Models Using Machine Learning and Big Data Analysis"
  - DOE CMIP6 Hackathon for RGMA Projects (2019)
  - Workshops on Data Mining in Earth System Science (DMESS) at the IEEE International Conference on Data Mining (ICDM)

DART-PFLOTRAN (Collaboration with NCAR)
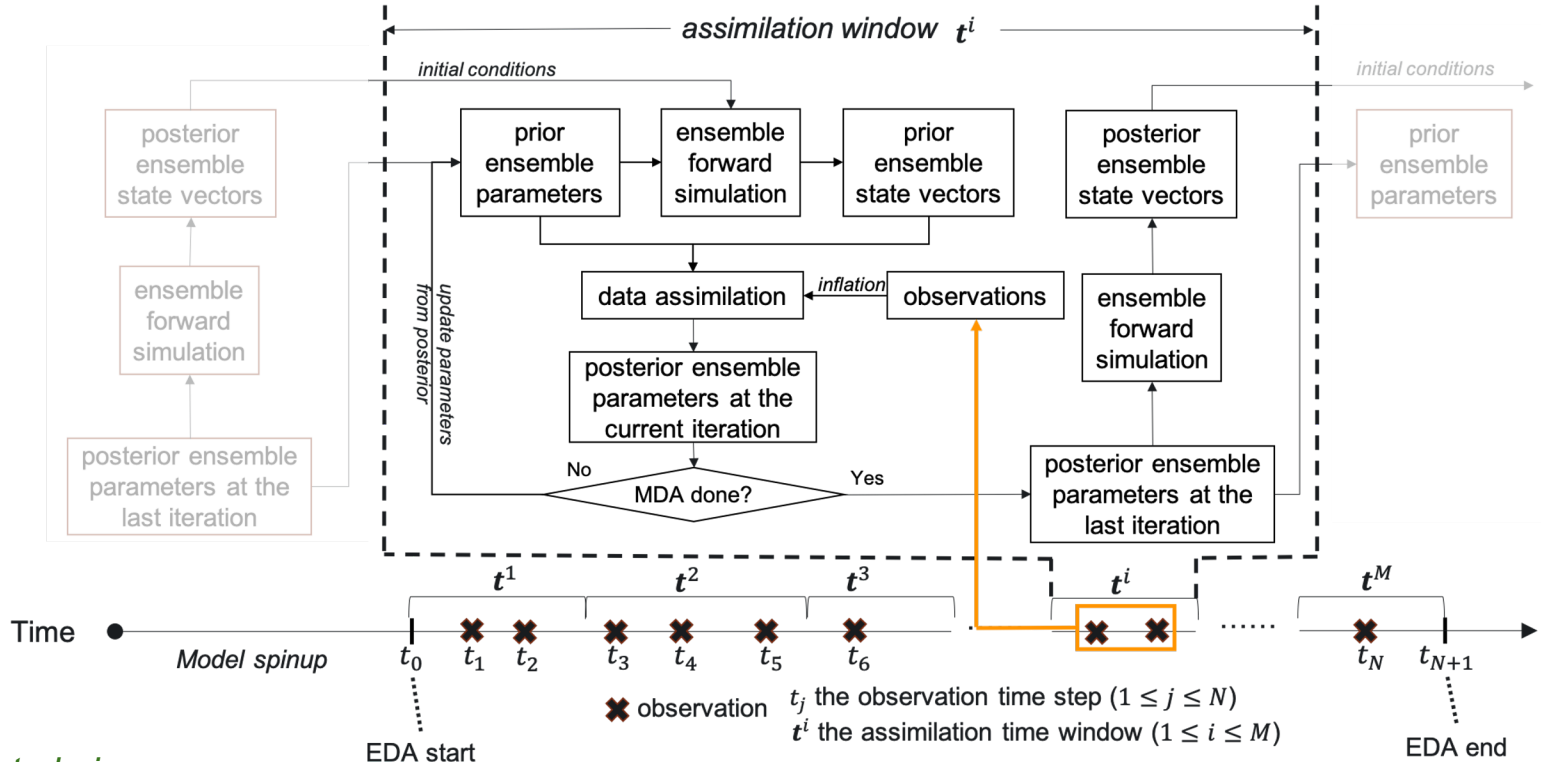
Jiang et al., in prep

KBase Pipeline to Inform Reactive Transport Modeling
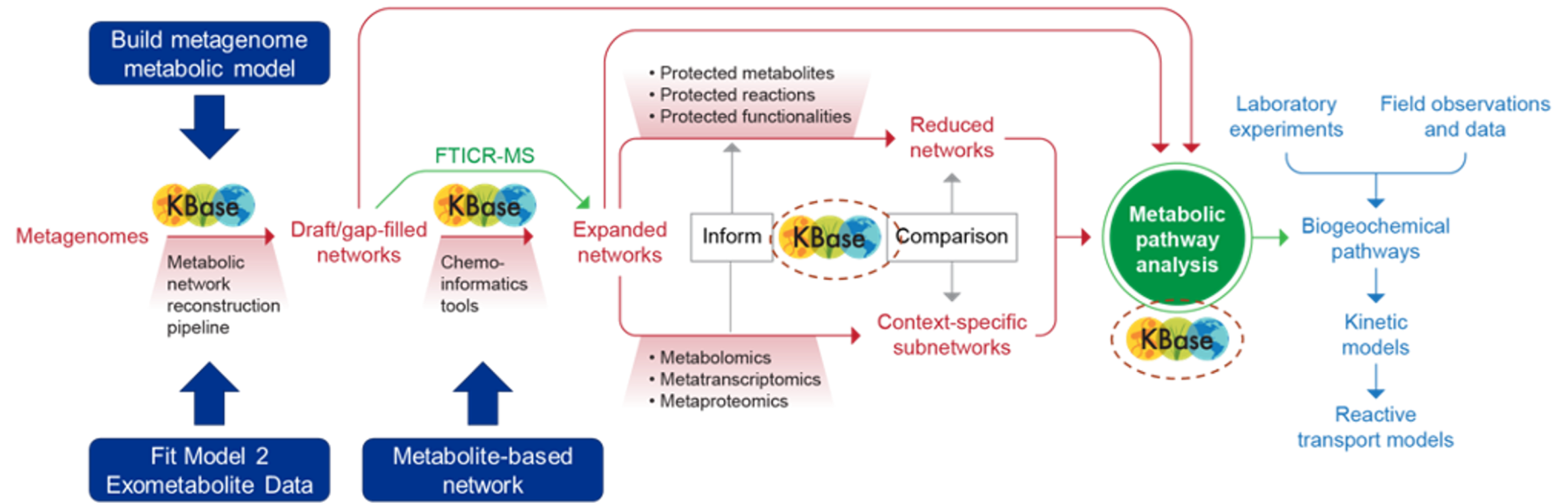
IDEAS-Watersheds PNNL SBR SFA Partnership
Song et al., submitted

U.S. DEPARTMENT OF **ENERGY** | Office of Science

Exasheds: Advancing Computational Science with **Machine Learning**

Combine modern **data-driven** approaches with advanced integrated surface-subsurface hydrological and **biogeochemical models** leveraging leadership-class computing facilities

5-year vision: Hyper-resolution, process-explicit hydrobiogeochemical simulations at river basin scales taking full advantage of diverse and spatially extensive data and providing feedback to design of distributed networks
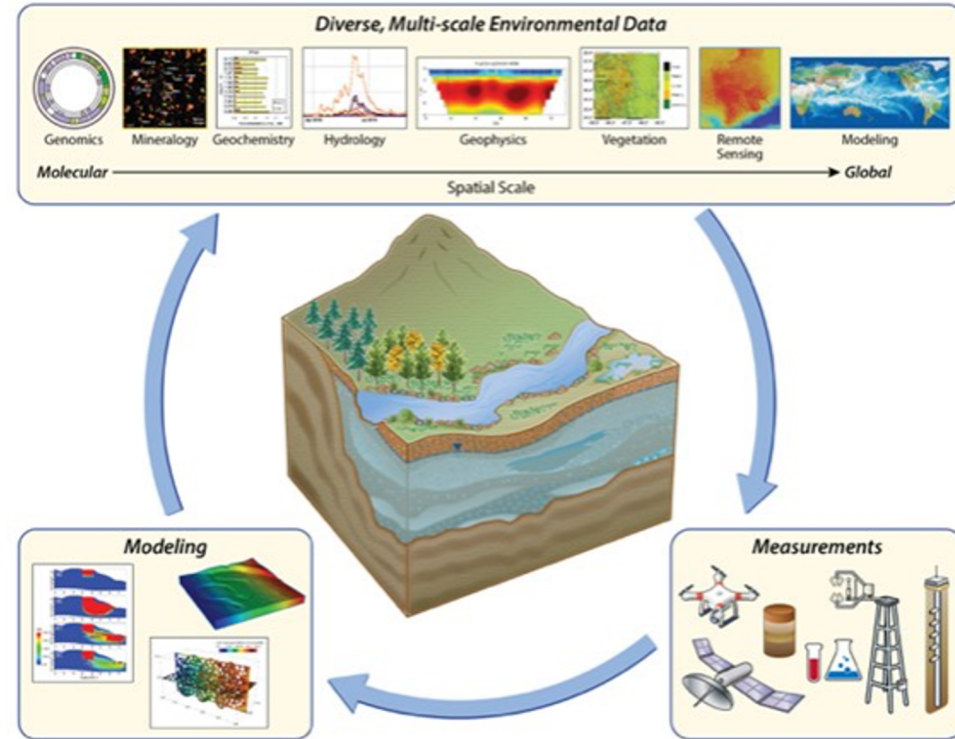
ILAMB

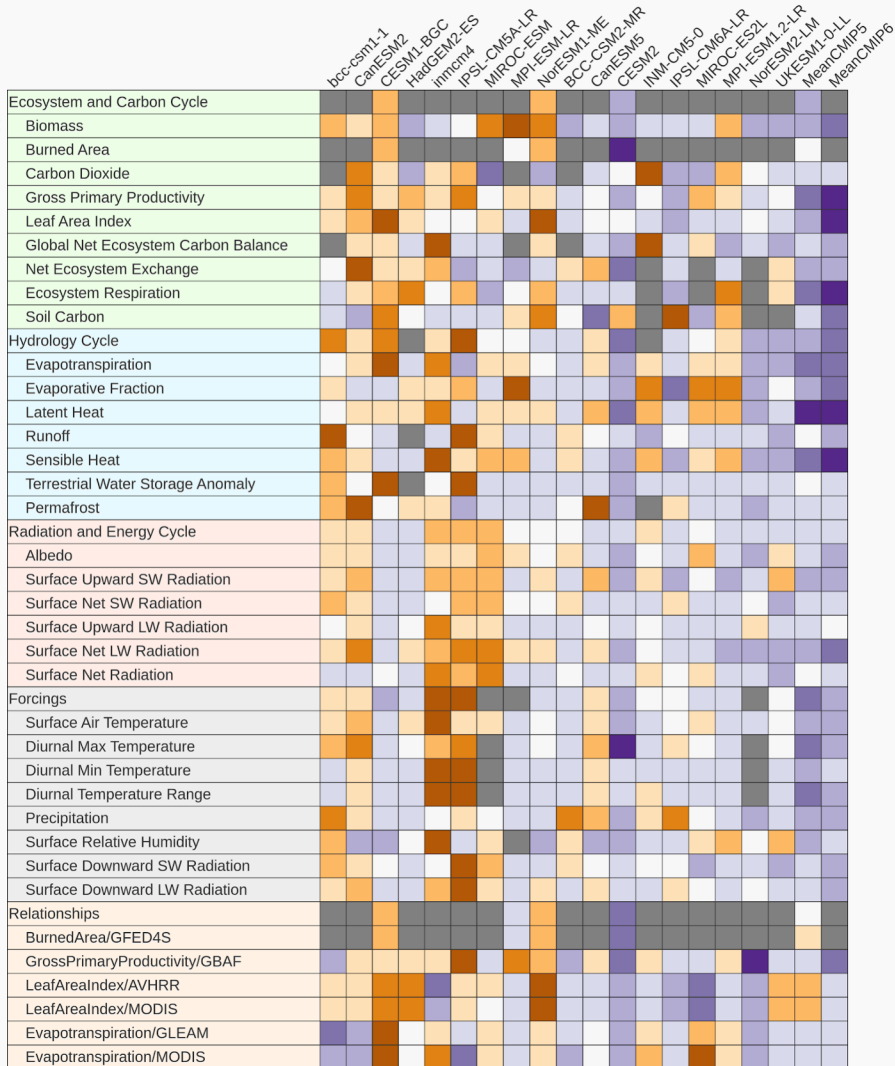International Land Model Benchmarking (ILAMB) community coordination activity was designed to

- Develop internationally accepted benchmarks
- Strengthen linkages between experimental, remote sensing, and modeling communities
- Support design & development of open source benchmarking tools, like the ILAMB Package

CMIP5 vs. CMIP6 Land Models

- The CMIP6 suite of land models (right) has improved over the CMIP5 suite of land models (left)
- The multi-model mean for CMIP*x* outperforms any single CMIP*x* model
- The mean CMIP6 land model is the "best model" overall

(Hoffman et al., in prep)

Relative Scale

Worse Value   Better Value

Missing Data or Error

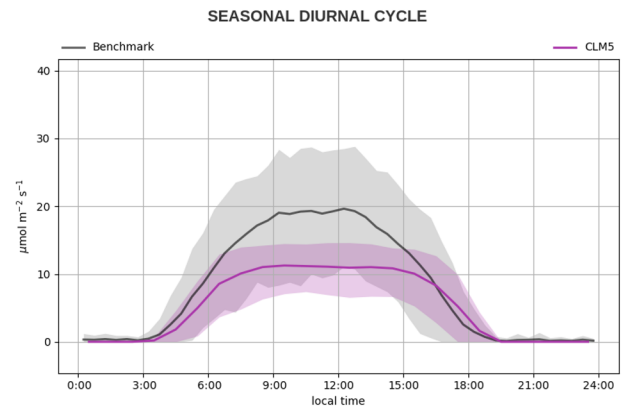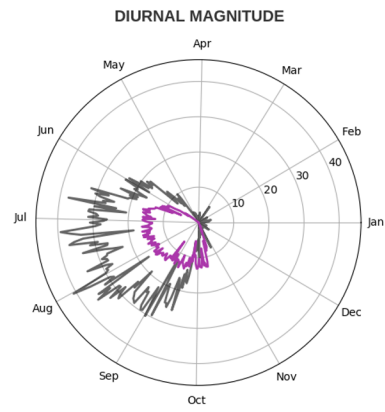U.S. DEPARTMENT OF **ENERGY** | Office of Science

PEcAn-ILAMB Integration

ILAMB Graphical Diagnostics

New PEcAn-ILAMB site-level diagnostics

GrossPrimaryProductivity / AMF_US-WCr / global / CLM5

		Mean State		All Models				Data Information			

	Download Data	Number of Years [1]	Season Length [d]	Diurnal Peak Timing [h]	Mean Season Uptake [1e-6 mol m-2 s-1]	Season Beginning [d]	Season Ending [d]	Diurnal Peak Timing Score [1]	Diurnal Uptake Score [1]	Season Beginning Score [1]	Season Ending Score [1]	Season Strength Score [1]	Overall Score [1]
Benchmark	[-]	14.0	120.	10.6	8.07	144.	264.						
CLM5	[-]	14.0	153.	10.4	5.62	138.	291.	0.896	0.756	0.681	0.361	0.668	0.672
ED2a	[-]	2.00	140.	12.8	3.77	138.	278.	0.750	0.570	0.712	0.616	0.601	0.650
ED2b	[-]	6.00	161.	10.8	4.35	120.	281.	0.910	0.610	0.370	0.517	0.630	0.607
SIPNETa	[-]	7.00	136.	9.79	6.86	145.	281.	0.908	0.818	0.801	0.510	0.835	0.774
SIPNETb	[-]	2.00	178.	4.50	5.76	104.	282.	0.521	0.670	0.205	0.400	0.850	0.529
SIPNETc	[-]	7.00	128.	8.64	8.81	144.	273.	0.830	0.769	0.811	0.716	0.736	0.773

⊕ 1999

⊖ 2000

DIURNAL MAGNITUDE

SEASONAL DIURNAL CYCLE

U.S. DEPARTMENT OF ENERGY | Office of Science

DOE CMIP6 Hackathon

- The **IPCC Sixth Assessment Report** had an aggressive schedule for publication of CMIP6 analysis research
- To support RGMA scientists doing multi-model research and benchmarking, **EESSD RGMA** & **Data Programs** are coordinating & sponsoring
  - Staging **CMIP6 output from ESGF** plus **reanalysis** & **observations**
  - Series of **tutorials** on CMIP6 organization, Jupyter notebooks, and (V)CDAT
  - **RGMA CMIP6 Hackathon** via videoconferencing at multiple hubs
- Lab & university researchers are co-organizing activities
  - *Forrest Hoffman (ORNL, RUBISCO), Jialin Liu (NERSC), Paul Ullrich (UC Davis, HYPERFACETS), Michael Wehner (LBNL, CASCADE), Wilbert Weijer (LANL, HiLAT)*
- NERSC: **4 PB disk storage** & **interactive computing resources**
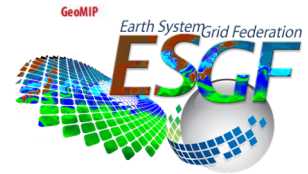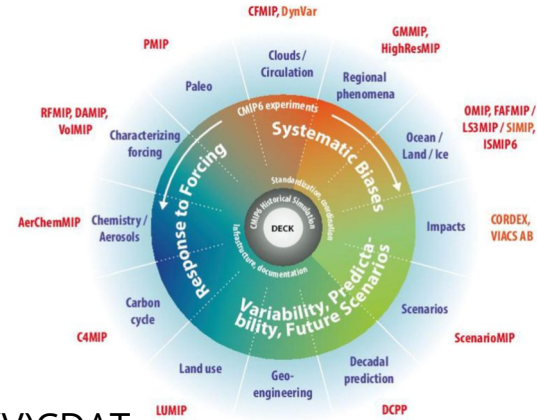
/global/cfs/projectdirs/m3522/cmip6

COV Report: Model integration across scales and EESD

"... develop a strategy for model integration across scales… to encourage new interdisciplinary model science that spans different existing program areas."

Specific recommendations:

- Develop a concrete overarching vision … including engagement with BSSD for microbe-to-Earth System scale initiatives;
- Increase community engagement (broad audiences such as AGU) around this vision to foster development of new strategies and research topics;
- Ensure that observational and modeling components of the scientific program are more tightly aligned; and
- Ensure synergies are optimally benefiting broader scientific objectives.

U.S. DEPARTMENT OF ENERGY | Office of Science

Model–Data Integration Grand Challenges and Priorities

We are seeking *your contributions* to update the 2016 white paper that described a framework for a Virtual Laboratory

We will hold **two Breakout Meetings this week**. Please respond to the Doodle Poll at https://www.doodle.com/poll/u89vfev3rgfzfsu5 **now**, and dates/times for the meetings will be announced today

To prepare, consider these questions for Model–Data Integration:
1. Are the goals and objectives in the white paper still relevant?
2. What new goals should be identified?
3. Should the priorities for development and deployment of cyberinfrastructure be changed based on recent resource development or technological advances?
4. What new priorities should be identified?

Towards a Shared ESS Cyberinfrastructure:
Vision and First Steps

Report from the ESS Executive Committee Workshop on Data Infrastructure
August 29-30, 2016.
DOE Headquarters, Germantown, MD

https://doesbr.org/BERfiles/Towards_a_Shared_ESS_Cyberinfrastructure.pdf

Model–Data Integration Grand Challenges and Priorities

Read the *Quick Start Reading Guide* in the "Reading" Google Drive area or directly at
https://docs.google.com/document/d/1x_w6VViNDM_MzxL3rgimWgGPM8bwpJaX_SHLsCRqAd0/edit?usp=sharing

We will **crowdsource writing & editing** of an updated chapter on Model–Data Integration during and after **two Breakout Meetings** this week. At **third Breakout Meeting on Monday, May 18**, we will integrate feedback and outline chapter

We will use the Google Doc *ESS-CI_MDI_Discussion* in the "Model-Data Integration" Google Drive area or directly at
https://docs.google.com/document/d/1qnzGZj9x58EjqDeCn4VwmeabaBgAgcDE3cTt5ZGMqbl/edit?usp=sharing

- Model–Data Integration Crowdsourcing Feedback and Discussion *– Enter ideas/feedback anytime!*
- Notes from Model–Data Integration Breakout Meeting #1
- Notes from Model–Data Integration Breakout Meeting #2
- Model–Data Integration Writing/Editing Assignments and Schedule