The Earth System Grid (ESG): Turning Climate Datasets into Community Resources

Presented by

The ESG-CET Team

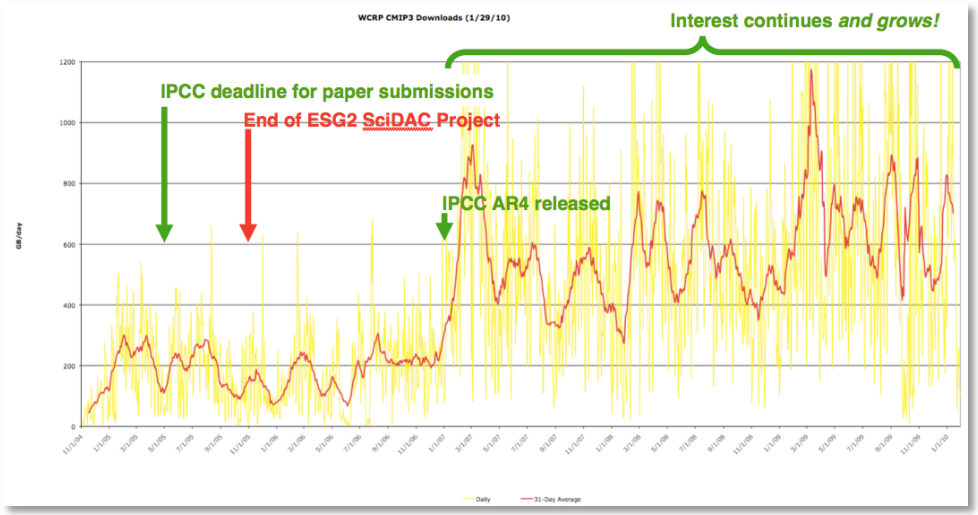
including

Lawrence Berkeley National Laboratory
Lawrence Livermore National Laboratory
Los Alamos National Laboratory
National Center for Atmospheric Research
National Oceanic and Atmospheric Administration
Oak Ridge National Laboratory
University of Southern California

www.earthsystemgrid.org

The growing importance of climate simulation data

- Broad investments in climate change research
  - Development of climate models
  - Climate change simulation
  - Model intercomparisons
  - Observational programs
- Climate change research is increasingly data intensive
  - Analysis and intercomparison of simulation and observations from many sources
  - Data used by model developers, policy makers, health officials, etc.
    - WG II – impacts, adaptation, and vulnerability
    - WG III – mitigation of climate change
- Interests in CMIP-3 data
  - SciDAC-1 experience illustrates timing of IPCC milestones and use of archive
  - Something similar will happen with CMIP-5
- Broad impact of ESG
  - Over 20K users
  - Over 600 published scientific papers

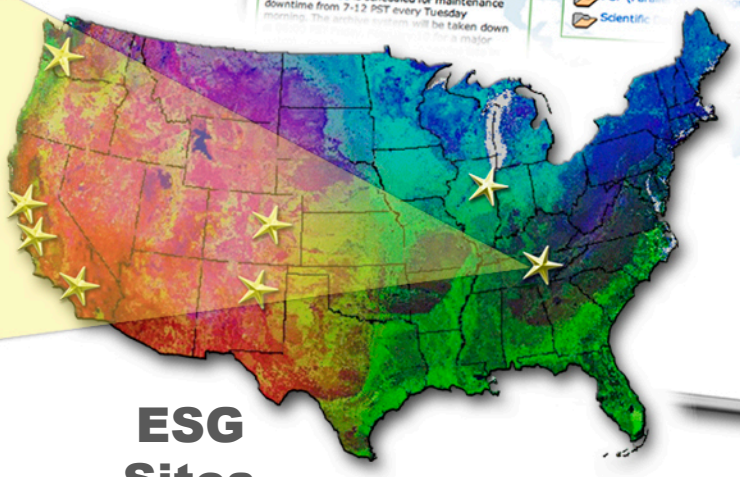


Courtesy: Robert Drach, LLNL

Earth System Grid objectives

To support the infrastructural needs of the national and international climate community, ESG is providing crucial technology to securely access, monitor, catalog, transport, and distribute data in today's grid computing environment

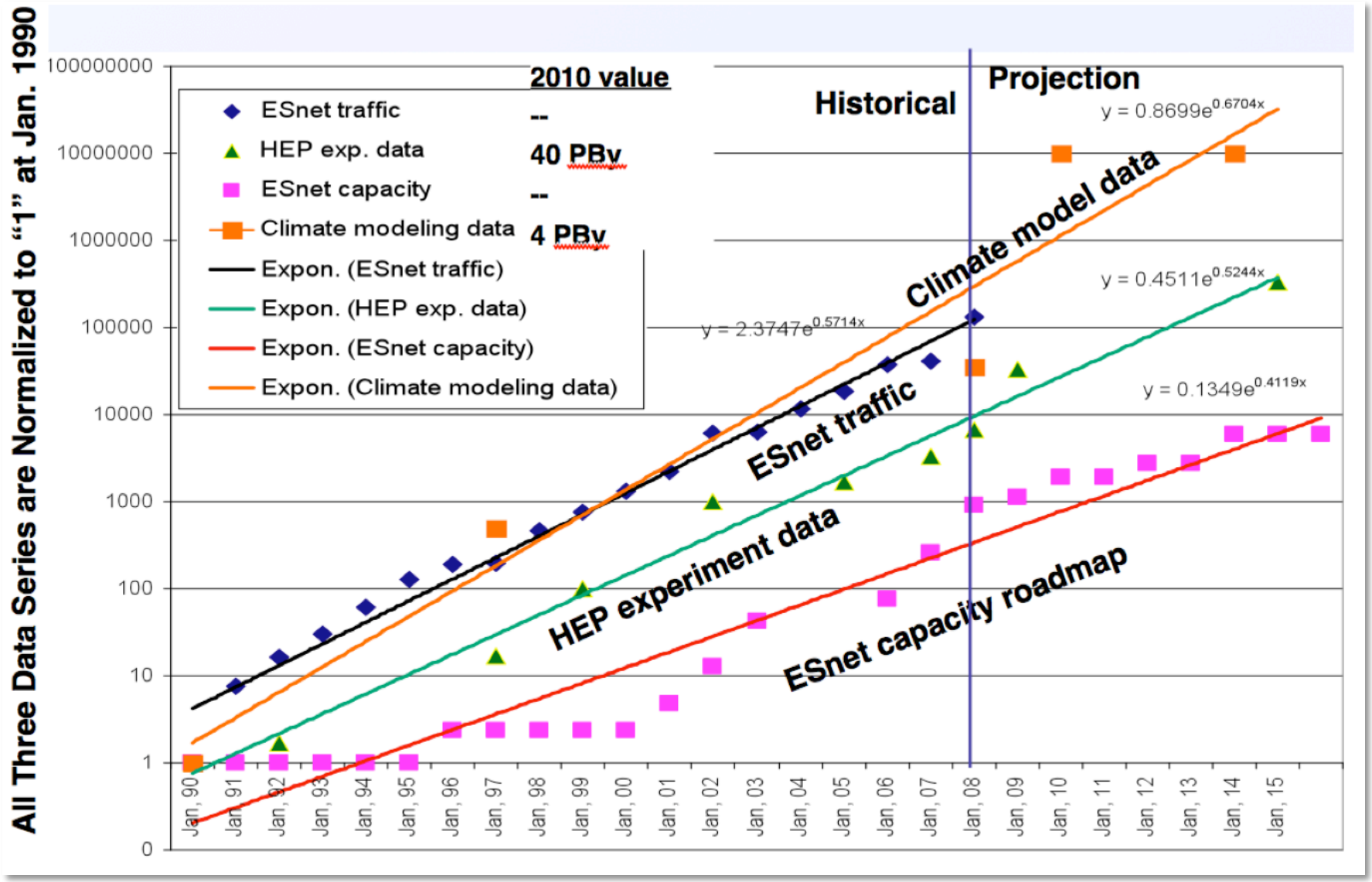
HPC hardware running climate models



ESG Sites

ESG Portal

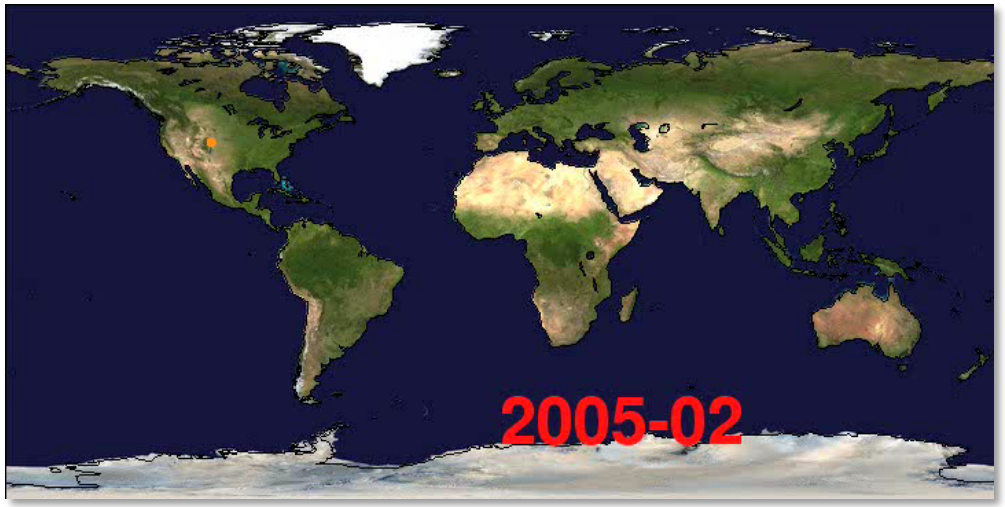
Climate data explosion (hundreds of XB by 2020)



ESG's current statistics

- LLNL CMIP-3 (IPCC AR4) ESG portal
  - 35 TB of data at one location
    - 83,337 files, model data from 13 countries
    - Generated by a modeling campaign coordinated by the Intergovernmental Panel on Climate Change (IPCC)
    - Over 600 scientific peer-review publications
- ORNL (C-LAMP, Legacy Data) ESG portal
  - 71 TB of data at one location
    - 270,378 files
    - Includes all Carbon Land Model Intercomparison Project (C-LAMP) data
- NCAR CCSM ESG portal
  - 237 TB of data at four locations (NCAR, LBNL, ORNL, LANL): 965,551 files
    - Includes the past 7 years of joint DOE/NSF climate modeling experiments
- Geographic distribution of the users that downloaded data from ESG web portals
    - Over 2,700 sites
    - 120 countries
    - 20,000 users
    - Over 1 PB downloaded

- Serving data to the community
  - Coupled Model Intercomparison Project, Phase 3 (CMIP-3)
  - Community Climate System Model (CCSM)
  - Parallel Climate Model (PCM)
  - Parallel Ocean Program (POP)
  - The North American Regional Climate Change Assessment Program (NARCCAP)
  - Cloud Feedback Model Intercomparison Project (CFMIP)
  - Carbon-Land Model Intercomparison Project (C-LAMP



Courtesy: Gary Strand - NCAR

Evolving ESG to petascale

ESG Data System Evolution

2006

Central database

- Centralized curated data archive
- Time aggregation
- Distribution by file transport
- No ESG responsibility for analysis
- Shopping-cart-oriented web portal

2009–2010

Testbed data sharing

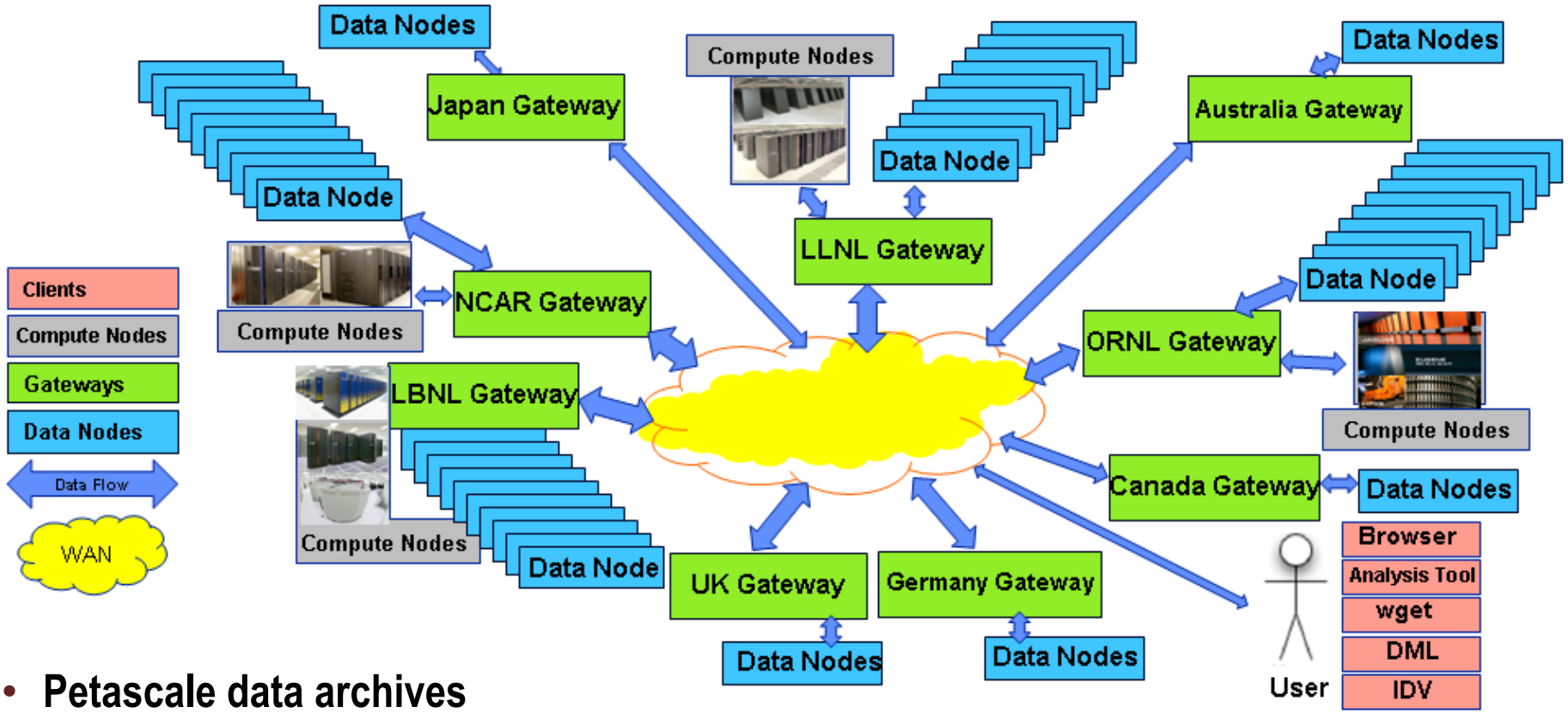
- Federated metadata
- Federated portals
- Unified user interface
- Selected server-side analysis
- Location independence
- Distributed aggregation
- Manual data sharing
- Manual publishing

2011

Full data sharing (add to testbed…)

- Synchronized federation
  - Metadata, data
- Full suite of server-side analysis
- Model/observation integration
- ESG embedded into desktop productivity tools
- GIS integration
- Model intercomparison metrics
- User support, life cycle maintenance

Next-generation ESG architecture



- Petascale data archives
- Broader geographical distribution of archives
  - Across the United States
  - Around the world
- Easy federation of sites
- Increased flexibility and robustness

ESG-CET gateways and data nodes

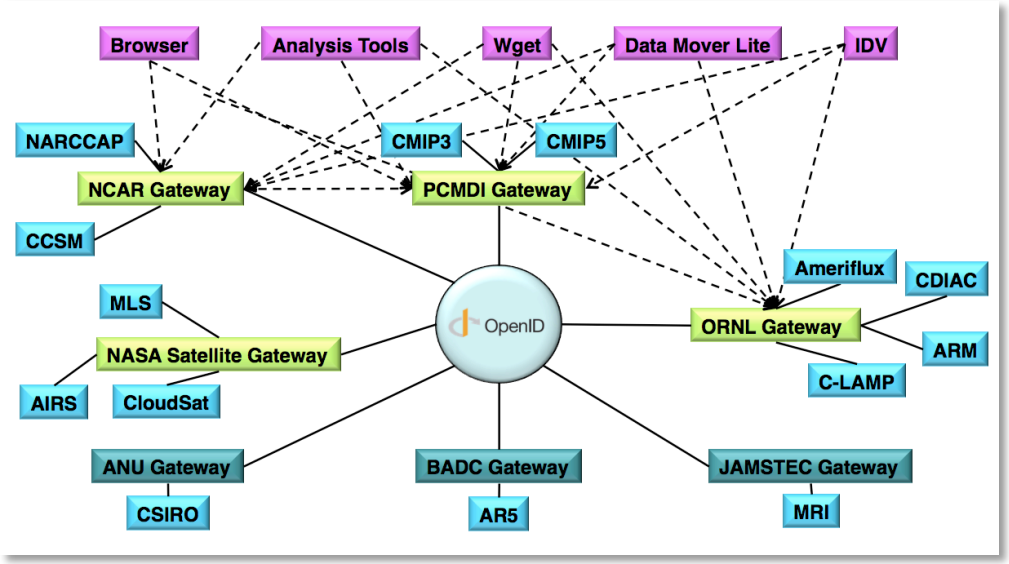
- **Federated architecture**

  Federation is **a virtual trust relationship among independent management domains that have their own set of services**. Users authenticate once to gain access to data across multiple systems and organizations.

- **Gateways**
  - Where data is discovered, requested
  - Portals, search capability, distributed metadata, registration, and user management
  - May be customized to an institution's requirements, topical focus
  - Fewer sites than data nodes
  - Currently: PCMDI, NCAR, ORNL, NASA; coming soon: GFDL, BADC, MPIM, JAMSTEC, ANU
- **Nodes**
  - Where data is stored and published
  - Data may be on disk or tertiary mass store
  - Each data node can publish to any gateway (facilitates topical gateways)
  - Data reduction/analysis
  - Complex architecture, including possible minimalist deployment w/o services
  - Anticipate ~20 data nodes for CMIP5, many others have expressed interest (over 50 sites)
- **Sites**
  - A site can be both a gateway and a data node

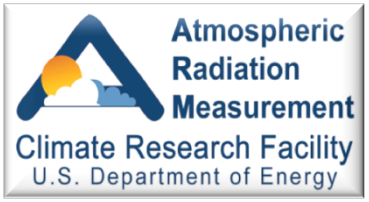
Data discovery from any gateway

- Distributed and federated architecture
- Support discipline-specific gateways
- Support browser-based and direct client access
- Single sign-on
- Automated GUI-based publication tools
- Full support for data aggregations
  - A collection of files, usually ordered by simulation time, that can be treated as a single file for purposes of data access, computation, and visualization
- User notification service
  - Users can choose to be notified when a data set has been modified

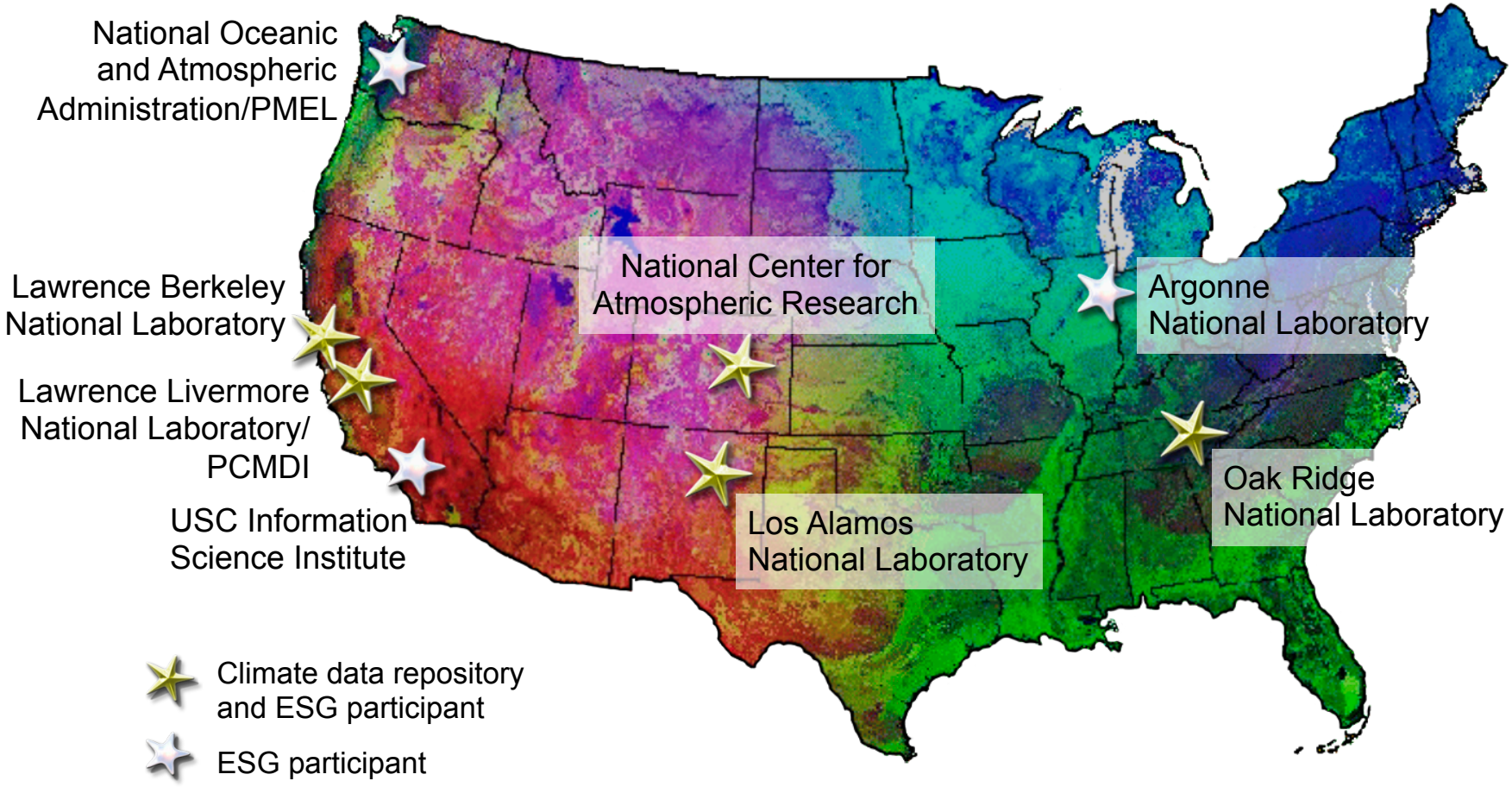


Observations for CMIP-5 simulations (ORNL)

- ORNL has established a pilot program to share observations to support model-to-data comparison
  - Goal is to demonstrate the feasibility of integration of high-value datasets within ESG
    - Atmospheric Radiation Measurement (ARM) Archive
    - Carbon Dioxide Information and Analysis Center (CDIAC)
    - ORNL Distributed Active Archive Center (DAAC)
  - Successfully demonstrated publication of AmeriFlux dataset (1,386 files)
- Goal: Provide access to a wealth of DOE observations via ESG
  - LLNL, NCAR, ORNL exploring ESG enhancements to support observations
  - Initial results are promising
  - Integration of these datasets will require substantial effort
    - Spatio/temporal projections
    - Metadata harvesting and ingestion
    - Observation/model data comparison tools



The team and sponsors



U.S. DEPARTMENT OF ENERGY | Office of Science

SciDAC Scientific Discovery through Advanced Computing

Contact

Earth System Grid

ORNL booth at SC2010

Galen Shipman (site PI), Ross Miller, Feiyi Wang

Internet

http://esg.ccs.ornl.gov