U.S. DEPARTMENT OF ENERGY | Office of Science

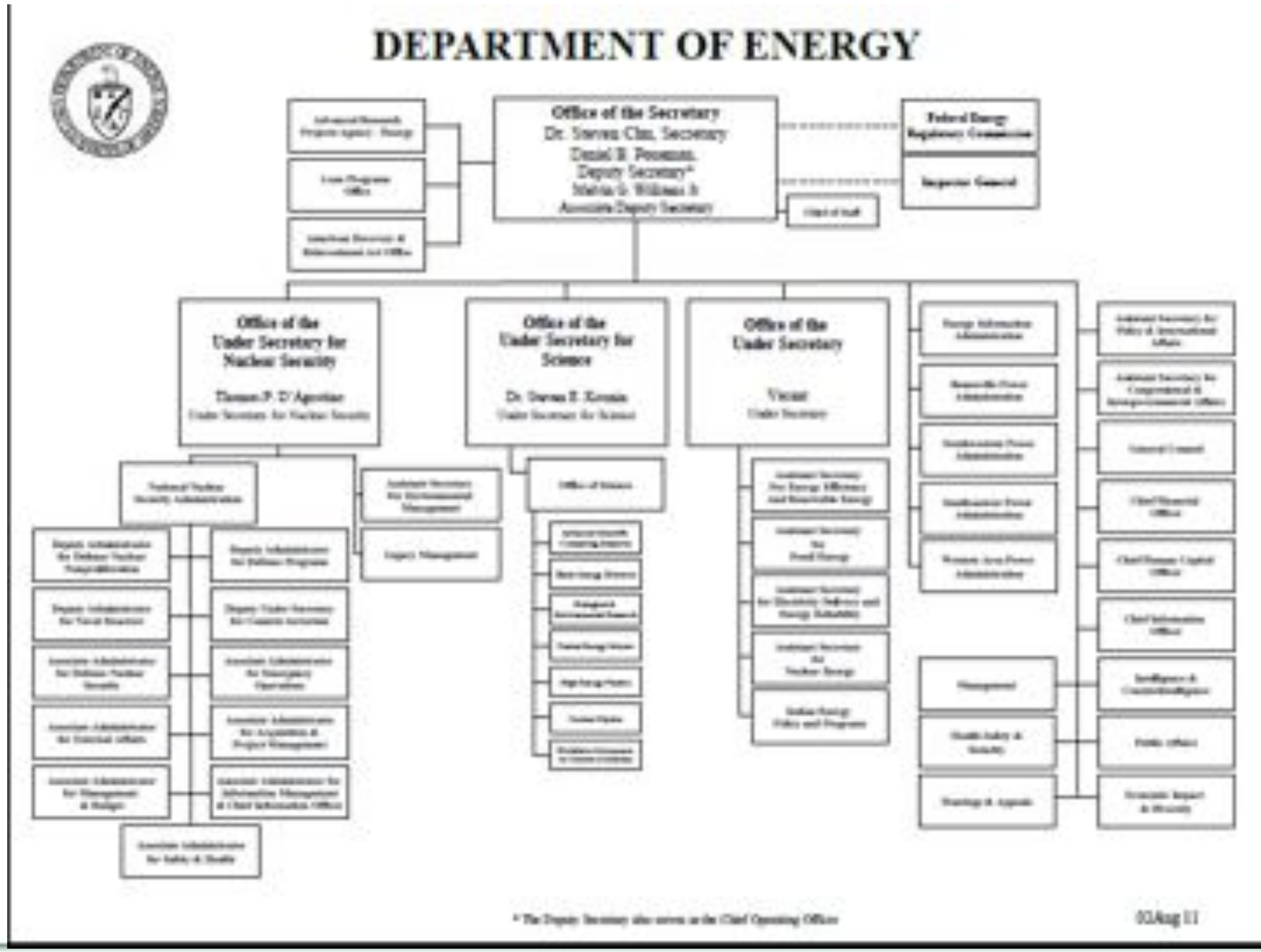
Scientific Data Challenges

October 24, 2011
Large Data Analysis and Visualization (LDAV) Symposium
IEEE VIsWeek
Providence, RI

Lucy Nowell, PhD
Computer Scientist & Program Manager
Advanced Scientific Computing Research
Office of Science
U.S. Department of Energy
www.science.energy.gov

U.S. Department of Energy

The mission of the Energy Department is to ensure America's security and prosperity by addressing its energy, environmental and nuclear challenges through transformative science and technology solutions.



Office of Science

Science to Meet the Nation’s Challenges Today and into the 21st Century

The Frontiers of Science

- Supporting research that led to over 100 Nobel Prizes during the past 6 decades—22 in the past decade alone
- Providing 45% of Federal support of basic research in the physical sciences and key components of the Nation’s basic research in biology and computing
- Supporting over 27,000 Ph.D.s, graduate students, undergraduates, engineers, and support staff at more than 300 institutions

21st Century Tools of Science

- Providing the world’s largest collection of scientific user facilities to over 26,000 users each year

Basic Energy Sciences

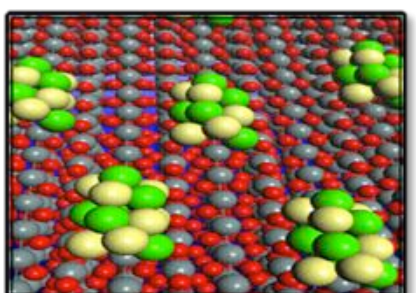
Understanding, predicting, and ultimately controlling matter and energy flow at the electronic, atomic, and molecular levels

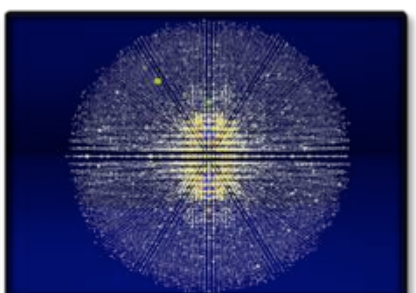
The Scientific Challenges:

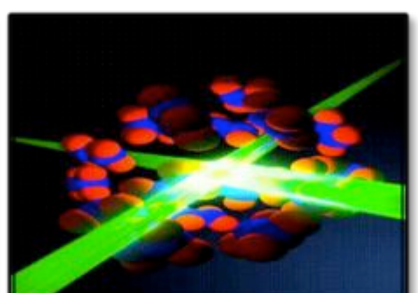
- Synthesize, atom by atom, new forms of matter with tailored properties, including nano-scale objects with capabilities rivaling those of living things
- Direct and control matter and energy flow in materials and chemical assemblies over multiple length and time scales
- Explore materials functionalities and their connections to atomic, molecular, and electronic structures
- Explore basic research to achieve transformational discoveries for energy technologies

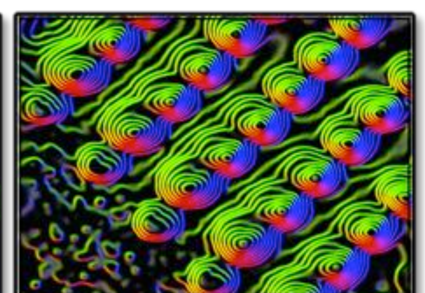
FY 2012 Highlights:

- Science for clean energy
  - Batteries and Energy Storage Hub
  - Interface sciences for high efficiency PV & next-generation nuclear systems; molecular design for carbon capture and sequestration; enabling materials sciences for transmission and energy efficiency; predictive simulation for combustion
- Computational Materials and Chemistry by Design and Nanoelectronics research with inter-agency coordination.
- Enhancements at user facilities:
  - LCLS expansion (LCLS-II); NSLS-II EXperimental Tools (NEXT); APS Upgrade (APS-U); TEAM II (aberration-corrected microscope); upgraded beamlines and instruments at the major facilities









Biological and Environmental Research

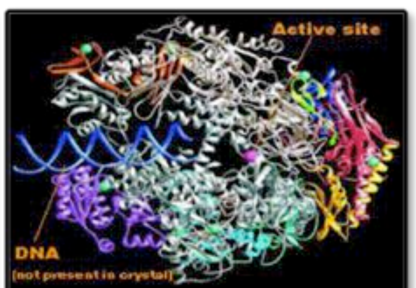
Understanding complex biological, climatic, and environmental systems across vast spatial and temporal scales

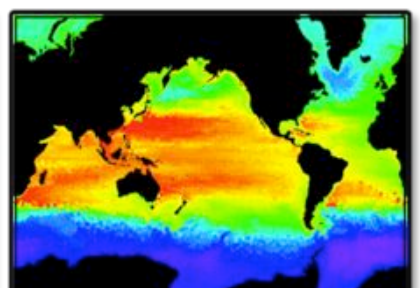
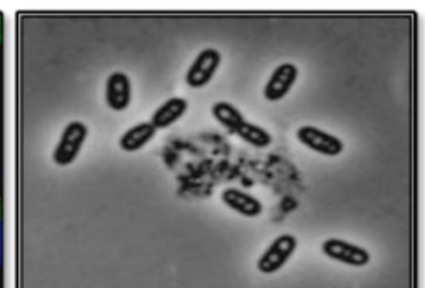
The Scientific Challenges:

- Understand how genomic information is translated with confidence to redesign microbes, plants or ecosystems for improved carbon storage, contaminant remediation, and sustainable biofuel production
- Understand the roles of Earth's biogeochemical systems (atmosphere, land, oceans, sea ice, subsurface) in determining climate so we can predict climate decades or centuries into the future, information needed to plan for future energy and resource needs.

FY 2012 Highlights:

- Clean energy biodesign on plant and microbial systems through development of new molecular toolkits for systems and synthetic biology research.
- Research and new capabilities to develop a comprehensive Arctic environmental system model needed to predict the impacts of rapid climate change.
- Continue support for the three DOE Bioenergy Research Centers, and operations of the Joint Genome Institute, the Environmental Molecular Sciences Laboratory, and the Atmospheric Radiation Measurement Climate Research Facility.

Fusion Energy Sciences

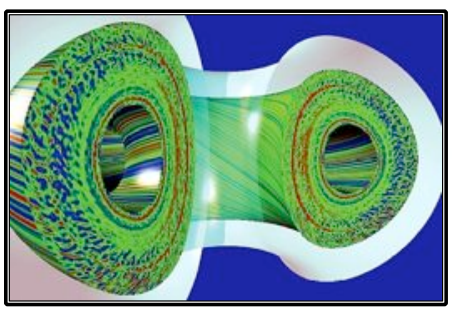
Understanding matter at very high temperatures and densities and building the scientific foundations for a fusion energy source

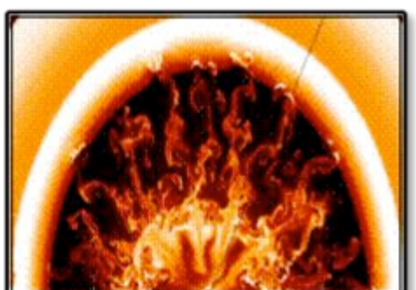
The Scientific Challenges:

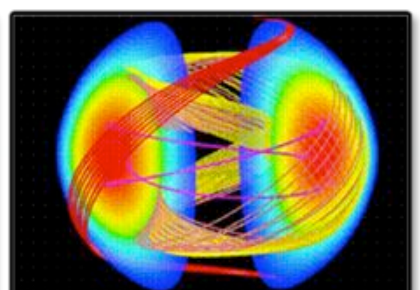
- Control a burning plasma state to form the basis for fusion energy
- Develop materials that can withstand the harsh heat and neutron irradiation in fusion facilities
- Manipulate and control intense transient flows of energy and particles
- Control the interaction of matter under extreme conditions for enabling practical inertial fusion energy

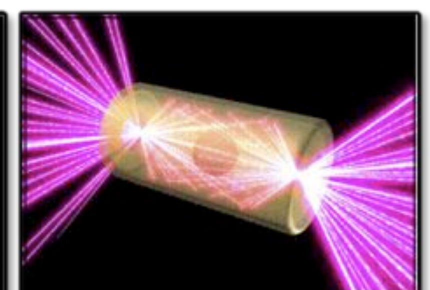
FY 2012 Highlights:

- ITER construction is supported
- DIII-D, Alcator C-Mod, and NSTX operate and investigate predictive science for ITER
- HEDLP investments continue in basic research on fast ignition, laser-plasma interaction, magnetized high energy density plasmas, and warm dense matter
- International activities are increased
- SciDAC expands to include fusion materials
- The Fusion Simulation Program pauses to assess now-completed planning activities









Nuclear Physics

Discovering, exploring, and understanding all forms of nuclear matter

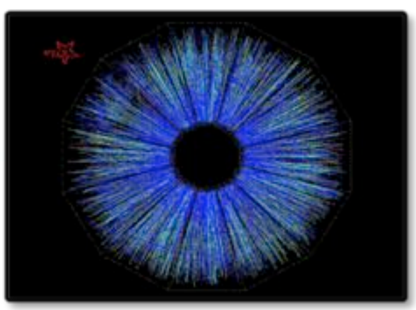
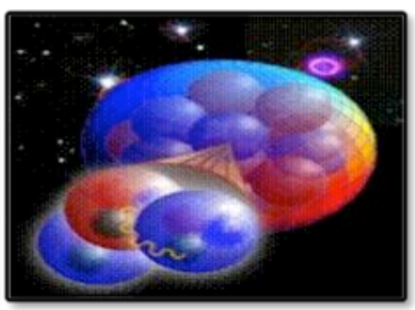
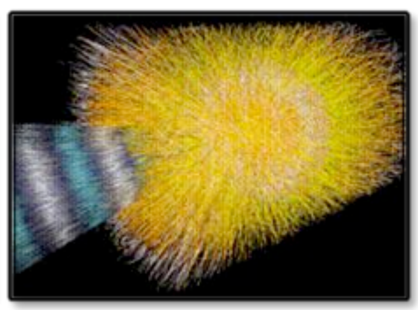
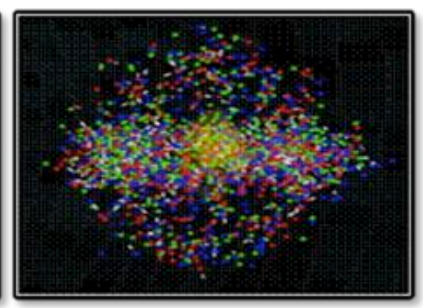
The Scientific Challenges:

Understand:

- The existence and properties of nuclear matter under extreme conditions, including that which existed at the beginning of the universe
- The exotic and excited bound states of quarks and gluons, including new tests of the Standard Model
- The ultimate limits of existence of bound systems of protons and neutrons
- Nuclear processes that power stars and supernovae, and synthesize the elements
- The nature and fundamental properties of neutrinos and neutrons and their role in the matter-antimatter asymmetry of the universe

FY 2012 Highlights:

- 12 GeV CEBAF Upgrade to study exotic and excited bound systems of quarks and gluons and for illuminating the force that binds them into protons and neutrons.
- Design of the Facility for Rare Isotope Beams to study the limits of nuclear existence.
- Operation of three nuclear science user facilities (RHIC, CEBAF, ATLAS); closure of the Holifield Radioactive Ion Beam Facility at ORNL.
- Research, development, and production of stable and radioactive isotopes for science, medicine, industry, and national security.

High Energy Physics

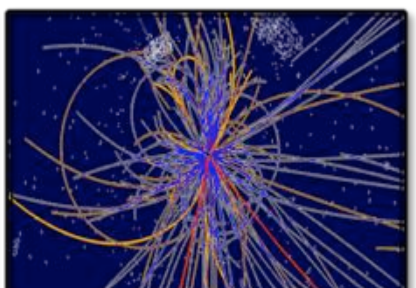
Understanding how the universe works at its most fundamental level

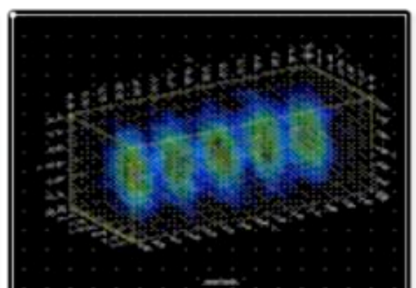
The Scientific Challenges:

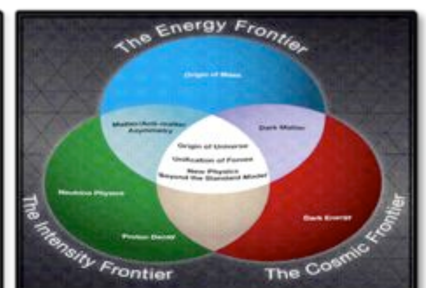
- Determine the origins of mass in terms of the fundamental particles and their properties
- Exploit the unique properties of neutrinos to discover new ways to explain the diversity of particles
- Discover new principles of nature, such as new symmetries, new physical laws, or unseen extra dimensions of space-time
- Explore the “dark” sector that is 95% of the Universe (Dark Matter and Dark Energy)
- Invent better and cheaper accelerator and detector technologies to extend the frontiers of science and benefit society

FY 2012 Highlights:

- Support for U.S. researchers at the LHC
- Research, design, and construction for NOvA, LBNE, and Mu2e experiments as part of a program of high energy physics at the intensity frontier
- Research in accelerator technologies including superconducting radio frequency and plasma wakefield acceleration.
- U.S. participation in international collaborations pursuing dark matter, dark energy and neutrino physics; the Reactor Neutrino Experiment in China and the Dark Energy Survey in Chile begin operations in FY 2012







Advanced Scientific Computing Research

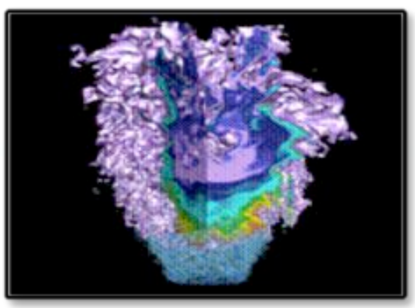
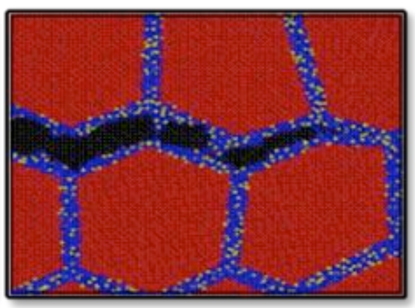
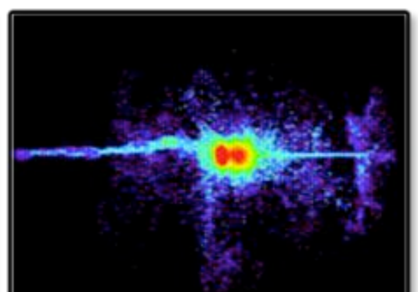
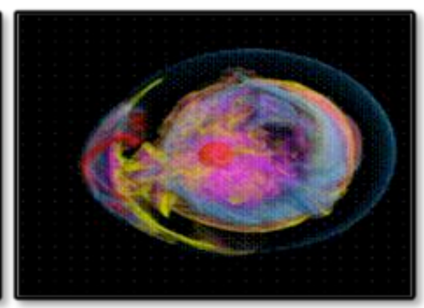
Delivering world leading computational and networking capabilities to extend the frontiers of science and technology

The Scientific Challenges:

- Deliver next-generation scientific applications using today's petascale computers.
- Discover, develop and deploy tomorrow's exascale computing and networking capabilities.
- Develop, in partnership with U.S. industry, next generation computing hardware and tools for science.
- Discover new applied mathematics and computer science for the ultra-low power, multicore-computing future.
- Provide technological innovations for U.S. leadership in Information Technology to advance competitiveness.

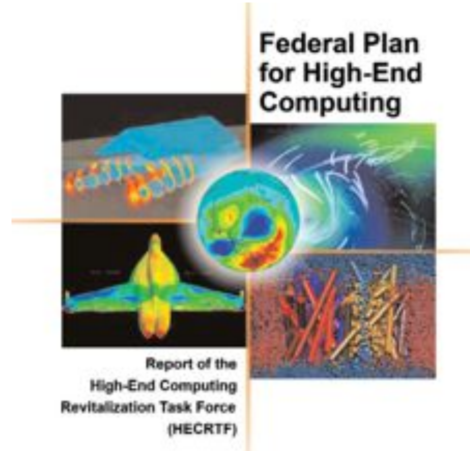
FY 2012 Highlights:

- Research in uncertainty quantification for drawing predictive results from simulation
- Co-design centers to deliver next generation scientific applications by coupling application development with formulation of computer hardware architectures and system software.
- Investments in U.S. industry to address critical challenges in hardware and technologies on the path to exascale
- Installation of a 10 petaflop low-power IBM Blue Gene/Q at the Argonne Leadership Computing Facility and a hybrid, multi-core prototype computer at the Oak Ridge Leadership Computing Facility.

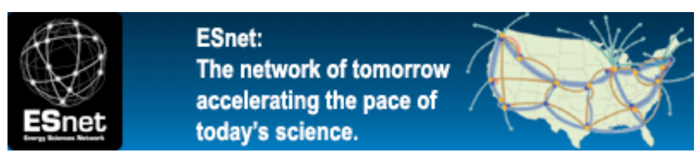
ASCR Facilities

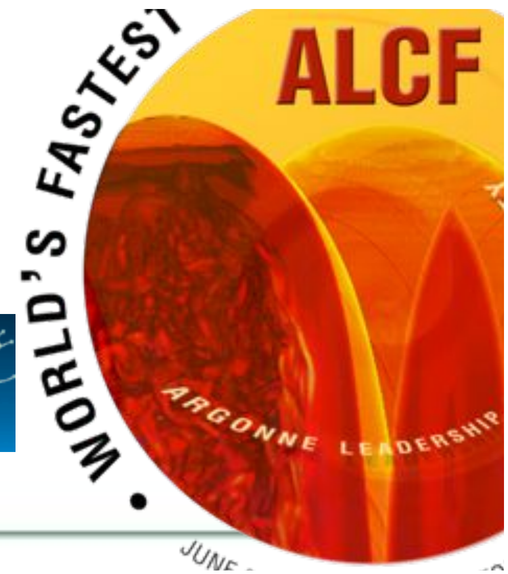
- Providing the Facility – **High-End and Leadership Computing**
- Investing in the Future - **Research and Evaluation Prototypes**
- Linking it all together – **Energy Sciences Network (ESnet)**



OLCF
Oak Ridge Leadership Computing Facility

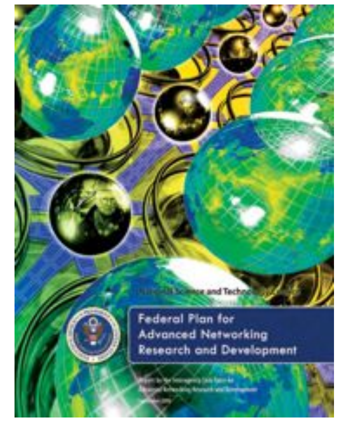
NERSC





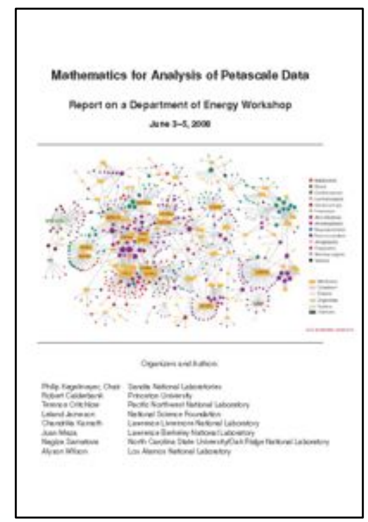
ASCR Research

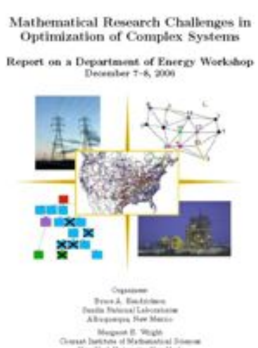
Substantial innovation is needed to provide essential system and application functionality in a timeframe consistent with the anticipated availability of hardware

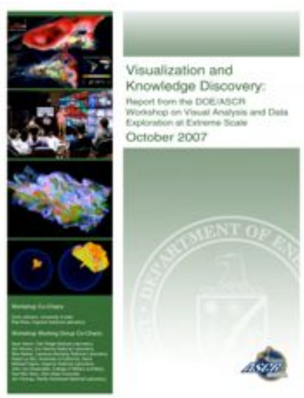


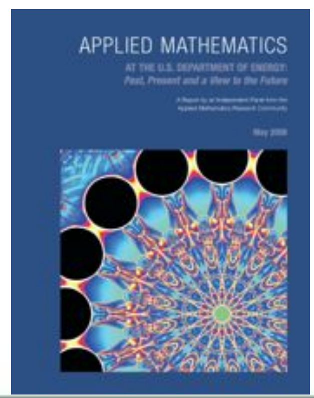
Provide forefront research knowledge and foundational tools:

- **Applied Mathematics**
- **Computer Science**
- **SciDAC**
- **Next Generation Networking for Science**

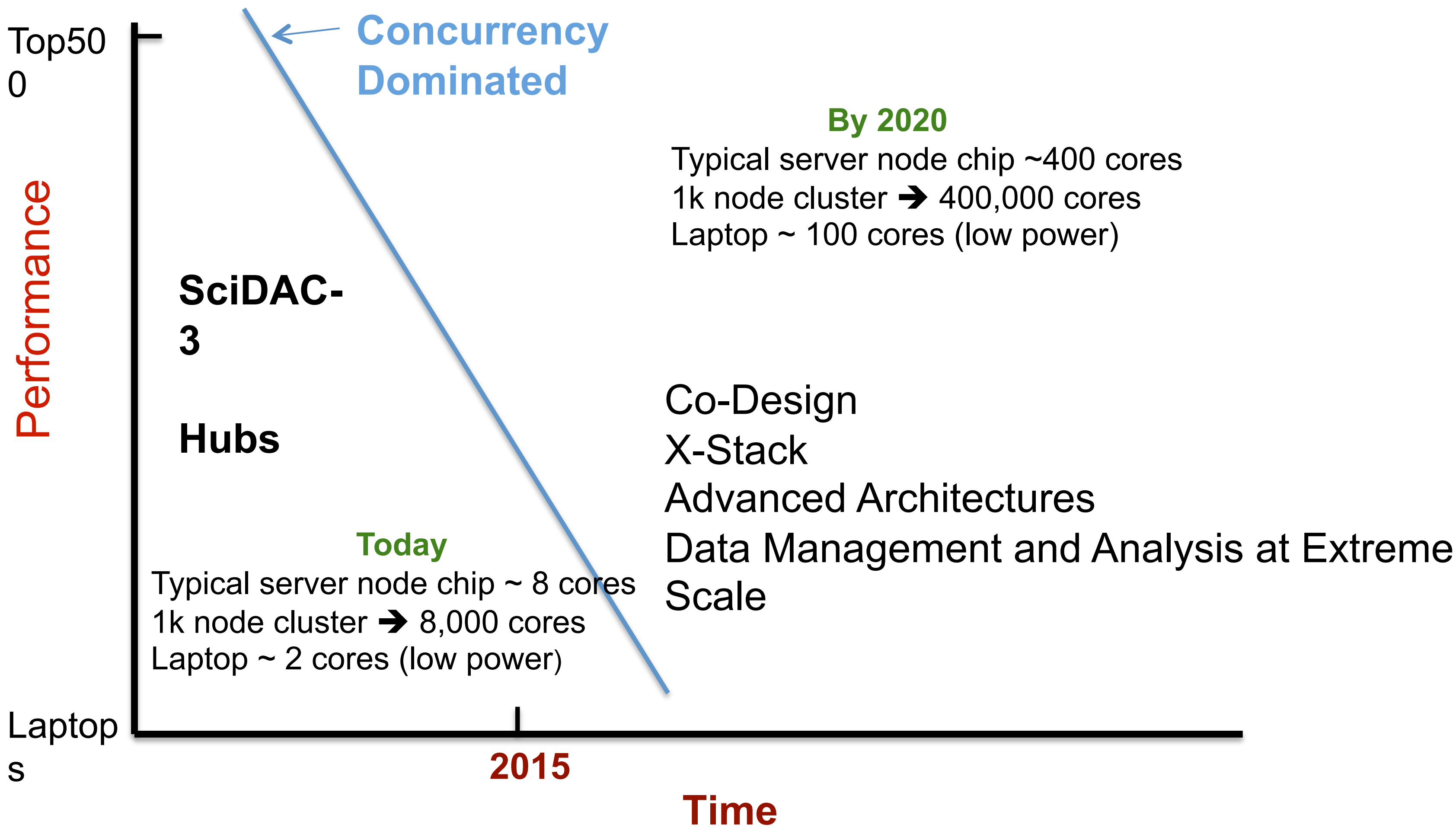






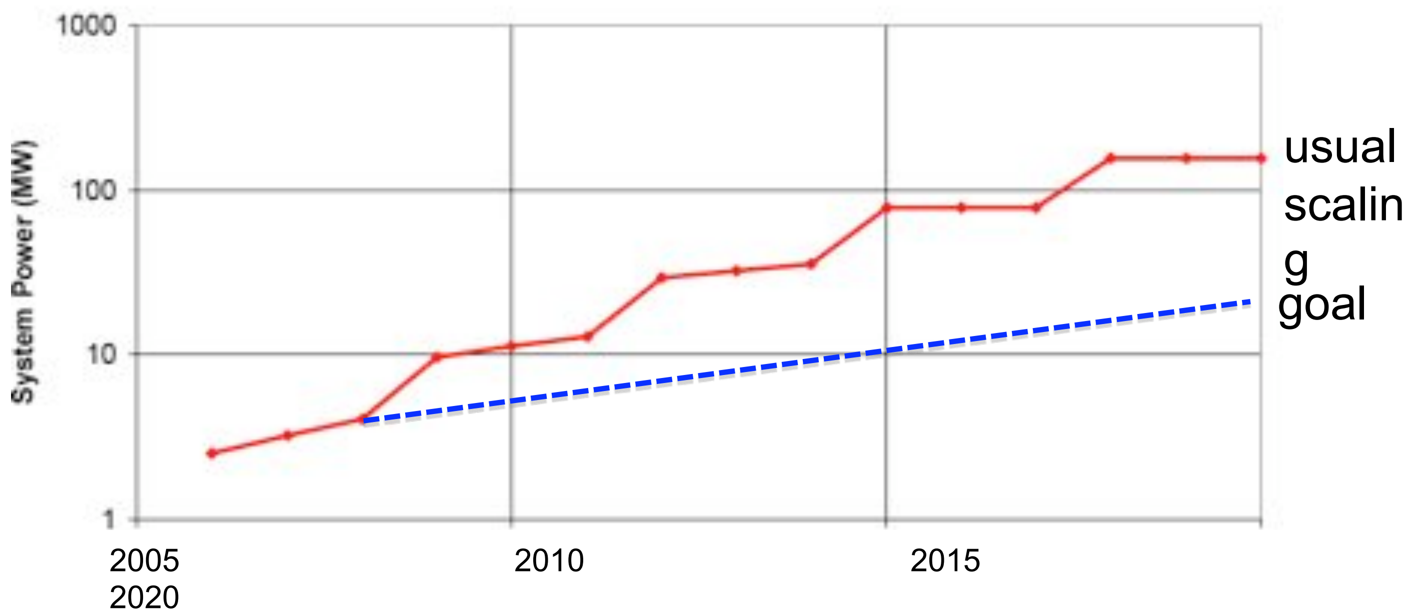


Why Computing is Changing



The Future is about Energy Efficient Computing

- **At $1M per MW, energy costs are substantial**
- **1 petaflop in 2010 will use 3 MW**
- **1 exaflop in 2018 at 200 MW with "usual" scaling**
- **1 exaflop in 2018 at 20 MW is target**



Approaches

- **Locality, Locality, Locality!**
- **Billion Way Concurrency;**
- **Uncertainty Quantification (UQ) including hardware variability;**
- **Flops free data movement expensive so:**
  - Remap multi-physics to put as much work per location on same die;
  - Include embedded UQ to increase concurrency;
  - Include data analysis if you can for more concurrency
  - Trigger output to only move important data off machine;
  - Reformulate to trade flops for memory use.
- **Wise use of silicon area**

Future of Data Driven Science

- All of these hardware trends impact data driven science (in many cases more than compute intensive);
- Data from instruments still on 18-24 month doubling because detectors on CMOS feature size path;
- 100 gigabit per second per lambda networks on horizon;
- Disk read and write rates will fall further behind processors and memory;
- Significant hardware infrastructure needed to support this which probably will not be replicated at users' home institution (i.e. launching a petabyte file transfer at a users laptop is not friendly)

BES ESnet Requirements Workshop Highlights

- **Instrument upgrades are driving dramatic changes in science process for BES facility users**
  - **Current paradigm is to take data home on portable USB hard drives**
  - **Current paradigm tractable because data sets are small, and end sites have poor network performance characteristics**
  - **Instrument and facility upgrades are increasing data set sizes by up to 100x in the next few years (individual beamlines at some facilities, significant portion of NSLS-II at BNL in 2014)**
  - **New paradigm: data sets too large, not portable, must be analyzed at facility and/or transferred to home institution**
- **Remote control will become more important as experiments become increasingly automated**
  - **Science Data Network (SDN) has a role to play here (we have seen this coming)**
  - **Need to vet operational models with facilities**
- **BES facilities and users are going to need a lot of help from ESnet, site networks, and tools providers – ESnet already planning pilot projects (ALS, NSLS)**

Other BES-ESnet Workshop Highlights

- **Multiple scientists expressed a need for remote visualization – this will increase as data sets increase in size (data too large to transfer home → analyze data at remote institution → remote visualization need)**
- **Significant knowledge gap in science community regarding appropriate tools, proper setup for data transfer, etc**
  - **http://fasterdata.es.net is helping, but there is more to do**
  - **ESnet working with facilities (network tuning talk at workshop generated significant interest from attendees)**
- **Cybersecurity is a major obstacle**
  - **SSH/SCP are used for data transfer (installed by default and approved by cybersecurity) and are the worst performers for data transfer**
  - **Appropriate tools (e.g. GridFTP) need to be deployable in order to be useful – if cybersecurity prevents their deployment, they do not exist as useful scientific tools**
- **Portals and other data services (e.g. NERSC Science Gateways) were of interest to several participants**

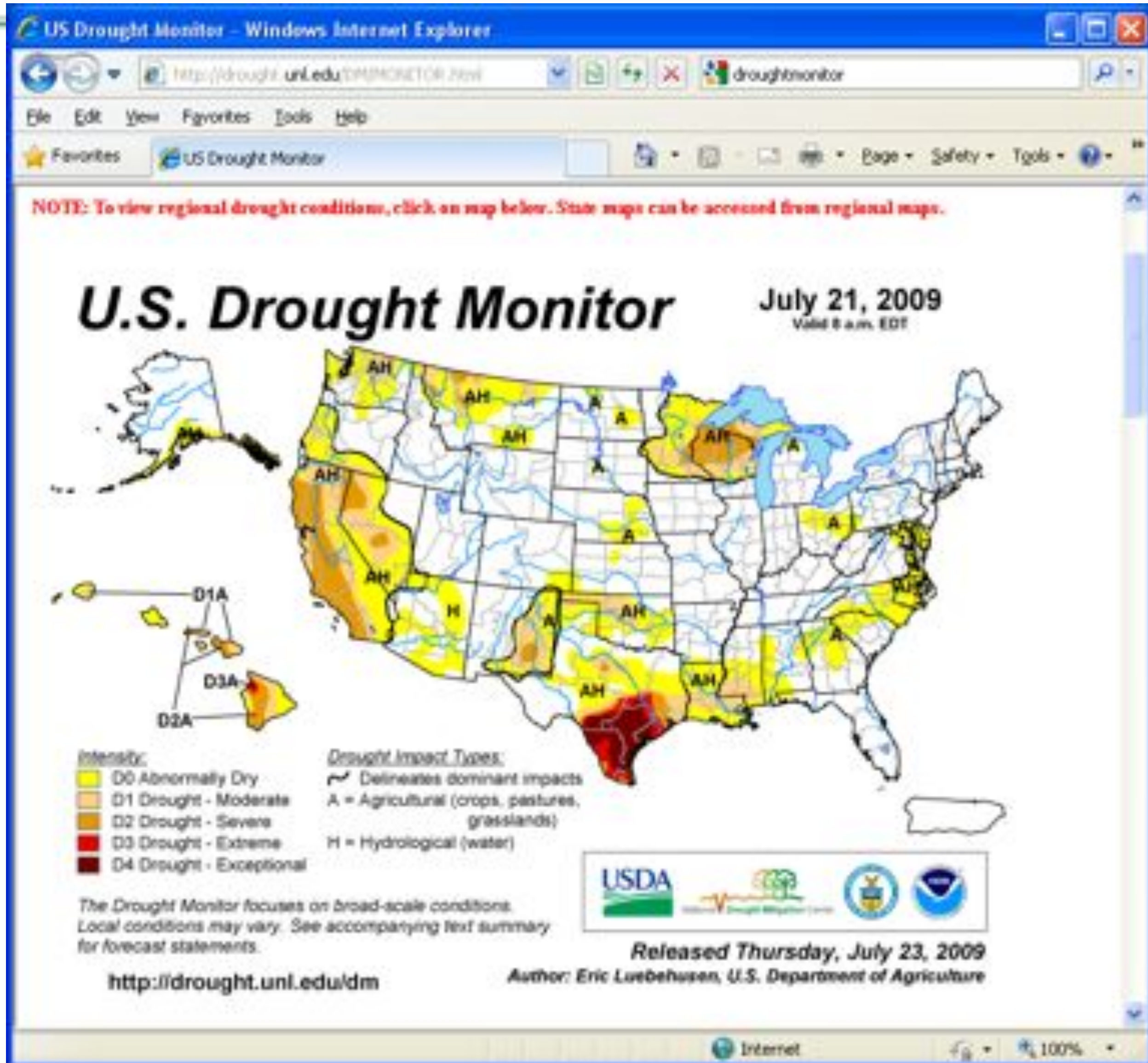
Research Challenges

- **Managing, moving, analyzing data;**
- **Transforming data into knowledge;**
- **Defining metadata and methods for automatic capture;**
- **Impact of distributed research teams;**
- **Changing scientific workflows; and**
- **Understanding the strengths and weaknesses of standards.**

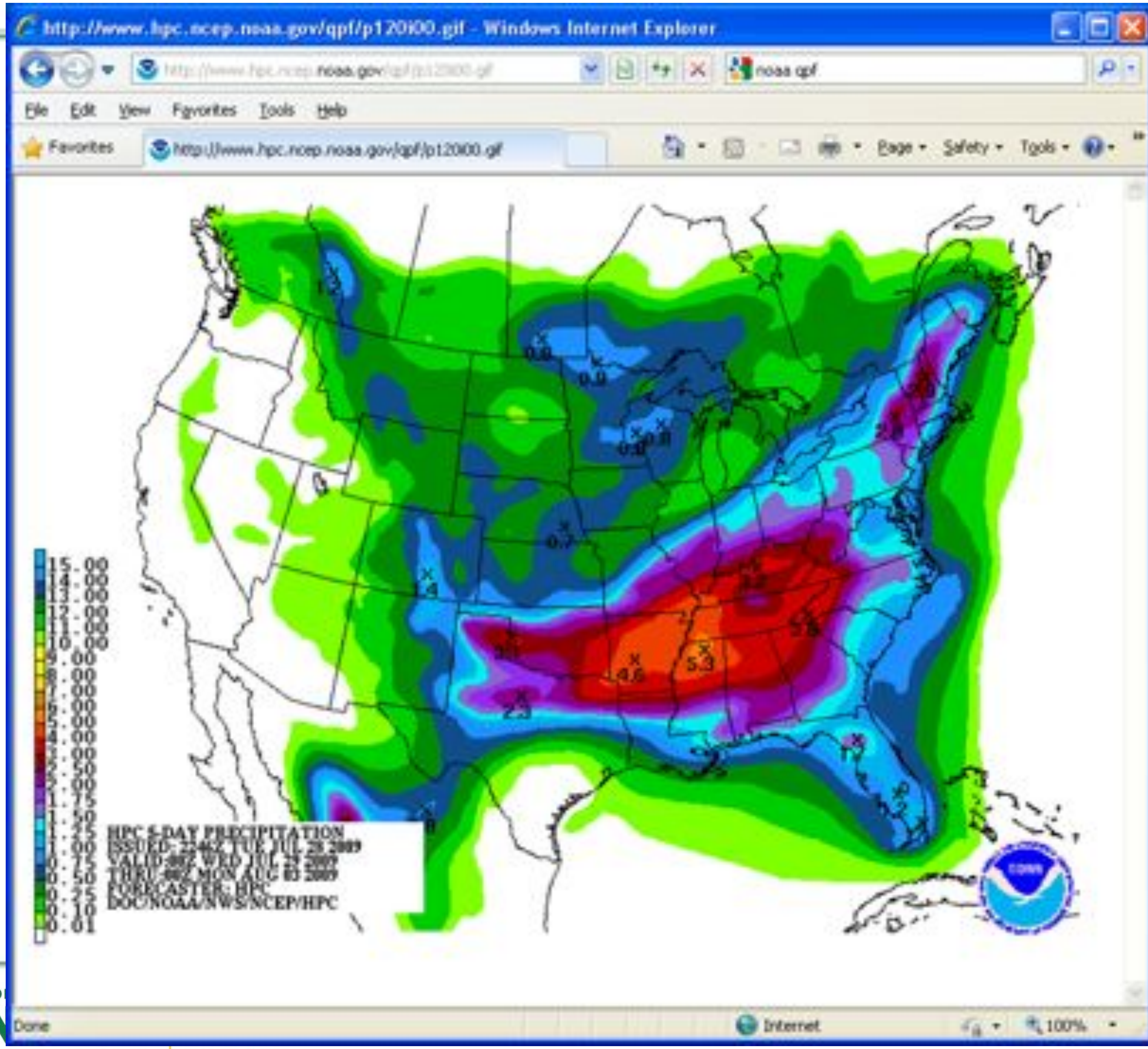
Disclaimer and Caveats

- **I am not a climate scientist.**
- **I am not an expert in modeling and simulation.**

A Visual Analytics Task

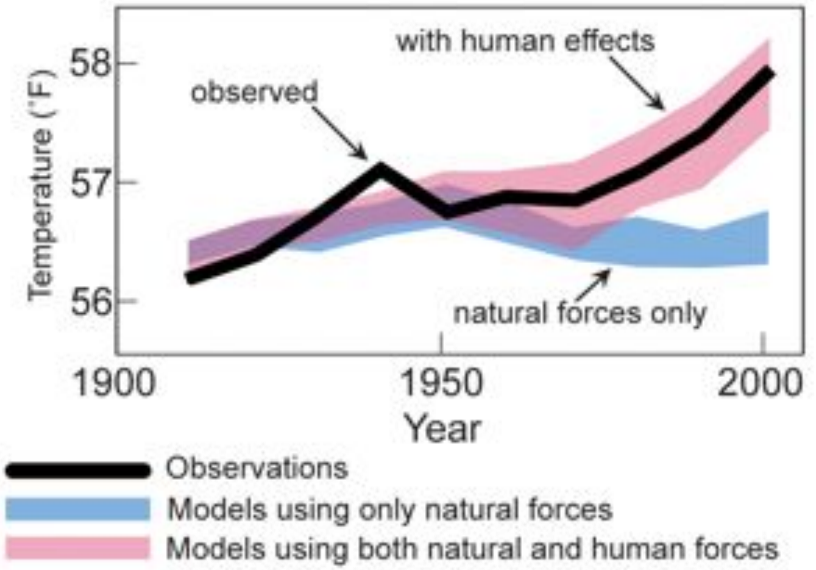


http://www.hpc.ncep.noaa.gov/qpf/day1-5.shtml



Analytical Challenges?

- Visual comparison task vs. human limitations of visual memory
- Change blindness
- Need for integration of information
- Unavailability of precise data
- Limits of prediction
- Cognitive biases



Hegerl *et al.*[49]

The blue band shows how global average temperatures would have changed due to natural forces only, as simulated by climate models. The red band shows model projections of the effects of human and natural forces combined. The black line shows actual observed global average temperatures. As the blue band indicates, without human influences, temperature over the past century would actually have first warmed and then cooled slightly over recent decades.[58]

Global Climate Change Impacts in the United States

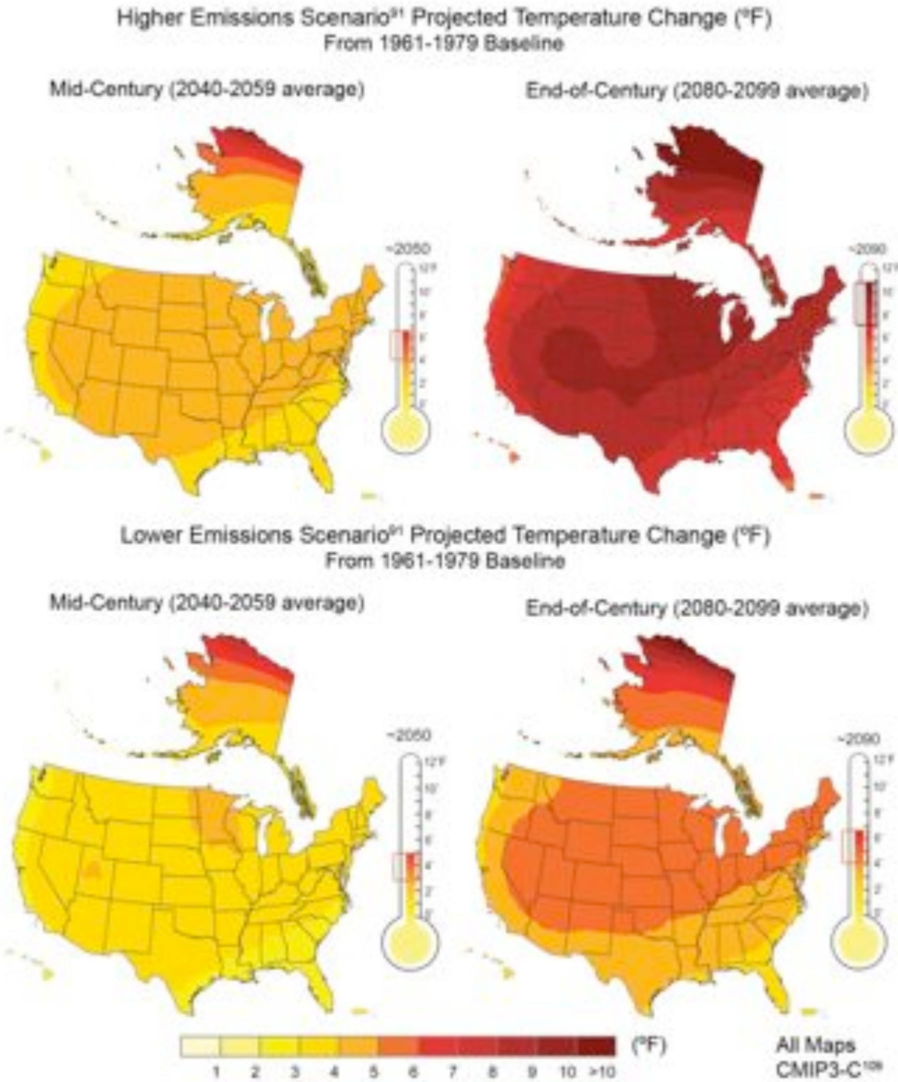
http://www.globalchange.gov/publications

Global Trends 2025: A Transformed World

- **The whole international system—as constructed following WWII—will be revolutionized. Not only will new players—Brazil, Russia, India and China (the BRICs) — have a seat at the international high table, they will bring new stakes and rules of the game.**
- **The unprecedented transfer of wealth roughly from West to East now under way will continue for the foreseeable future.**
- **Unprecedented economic growth, coupled with 1.5 billion more people, will put pressure on resources —particularly energy, food, and water—raising the specter of scarcities emerging as demand outstrips supply.**
- The potential for conflict will increase owing partly to political turbulence in parts of the greater Middle East.
  - http://www.dni.gov/nic/NIC_2025_project.html

Implications of Climate Change

- **Changing Patterns of Precipitation**
  - More-frequent flooding
  - More-frequent and more-severe droughts
- **Tension over Resources**
  - Water, Food Supplies, Energy, etc.
- **Impacts on Human Health and Behavior**
- **Sea-Level Rise and Coastal Impacts**
  - Transportation, Industry, Ecosystems, etc.

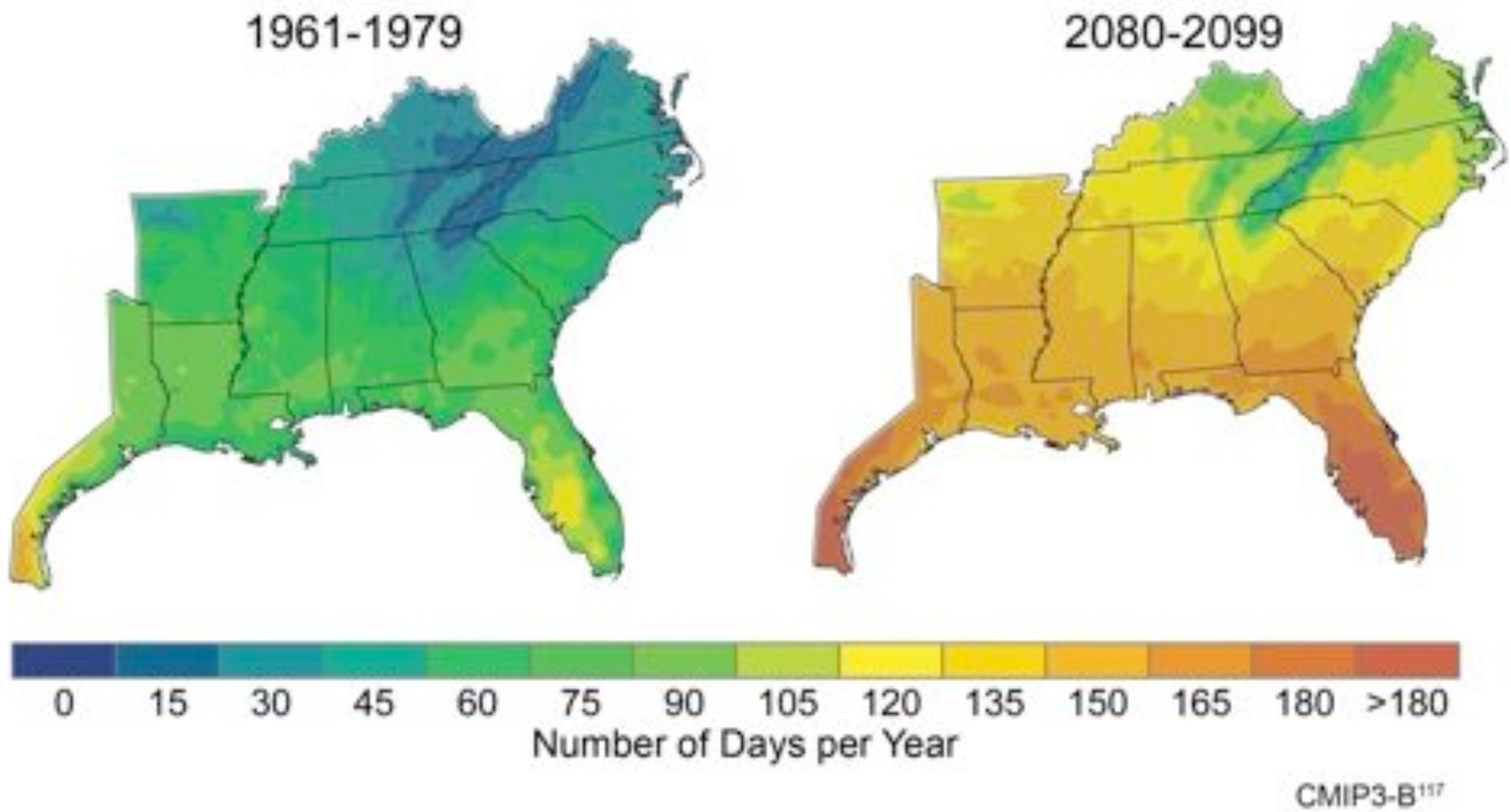


The maps on this page and the previous page are based on projections of future temperature by 16 of the Coupled Model Intercomparison Project Three (CMIP3) climate models using two emissions scenarios from the Intergovernmental Panel on Climate Change (IPCC), *Special Report on Emission Scenarios* (SRES).91 The "lower" scenario here is B1, while the "higher" is A2.[91] The brackets on the thermometers represent the likely range of model projections, though lower or higher outcomes are possible. Additional information on these scenarios is on pages 22 and 23 in the previous section, *Global Climate Change*. These maps, and others in this report, show projections at national, regional, and sub-regional scales, using well-established techniques.[110]

- **The global average temperature since 1900 has risen 1.5 degrees F. By 2100, it is expected to rise 2 to 11.5 degrees F.**
- **The U.S. average temperature has risen by a comparable amount and is very likely to rise more than the global average over this century.**
  - *Global Climate Change Impacts in The United States, p. 9*

More Than Doubling the Number of Days Over 90°

- **xxx**

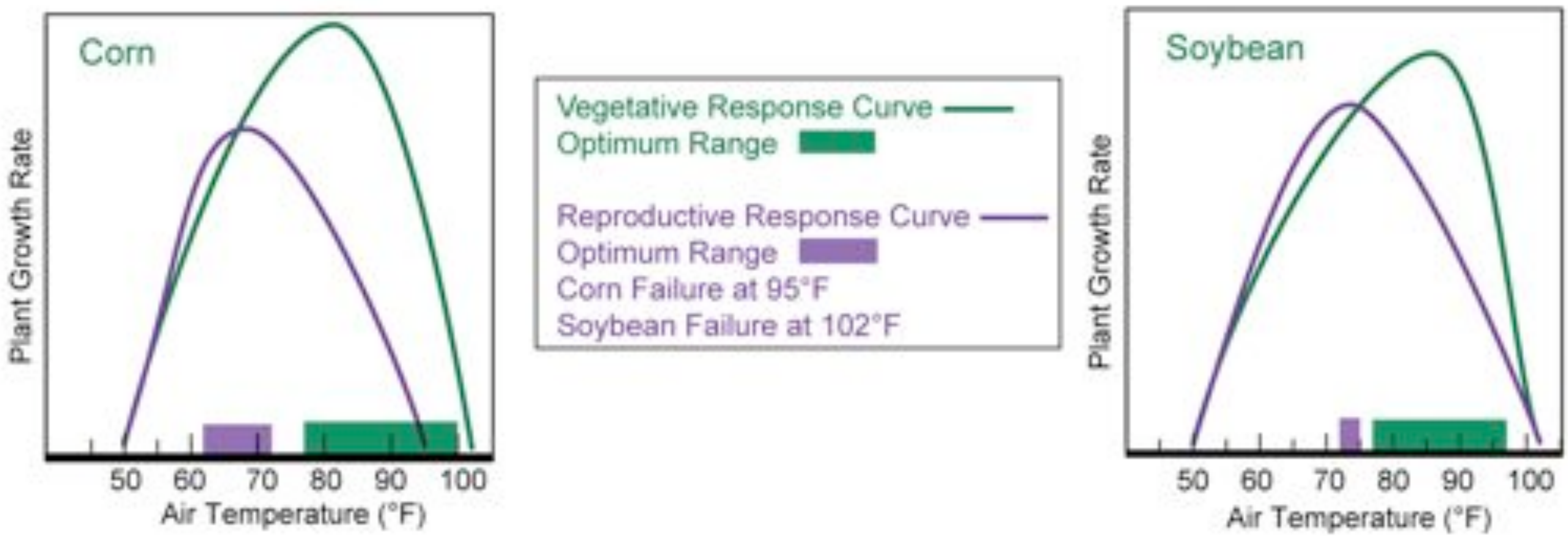


The number of days per year with peak temperature over 90°F is expected to rise significantly, especially under a higher emissions scenario[91] as shown in the map above. By the end of the century, projections indicate that North Florida will have more than 165 days (nearly six months) per year over 90°F, up from roughly 60 days in the 1960s and 1970s. The increase in very hot days will have consequences for human health, drought, and wildfires.

How to Feed a Hungry World Population?

- **The World Food Bank estimates that demand for food will rise by 50% by 2030, as a result of growing world population, rising affluence, and the shift to Western dietary preferences by a larger middle class.**
- **p. viii**

Rising Temperatures Threaten Key Crops



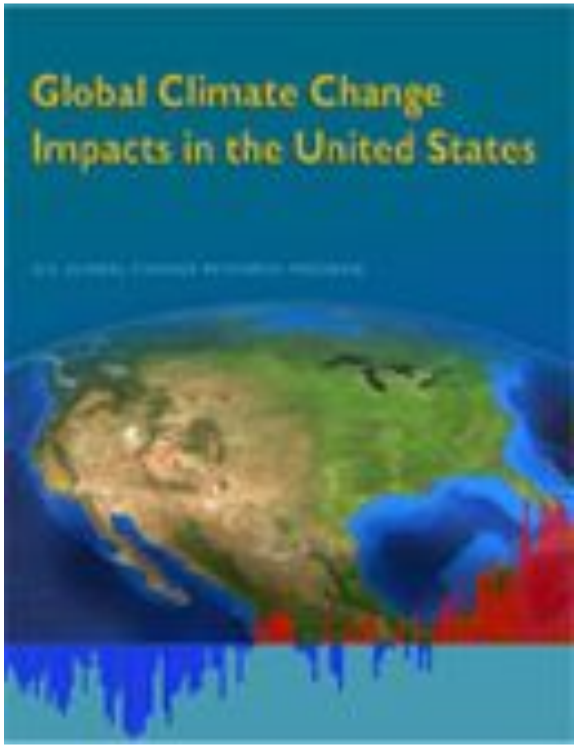
ARS USDA

For each plant variety, there is an optimal temperature for vegetative growth, with growth dropping off as temperatures increase or decrease. Similarly, there is a range of temperatures at which a plant will produce seed. Outside of this range, the plant will not reproduce. As the graphs show, corn will fail to reproduce at temperatures above 95°F and soybean above 102°F.

Potential Climate Change Impact on National Security

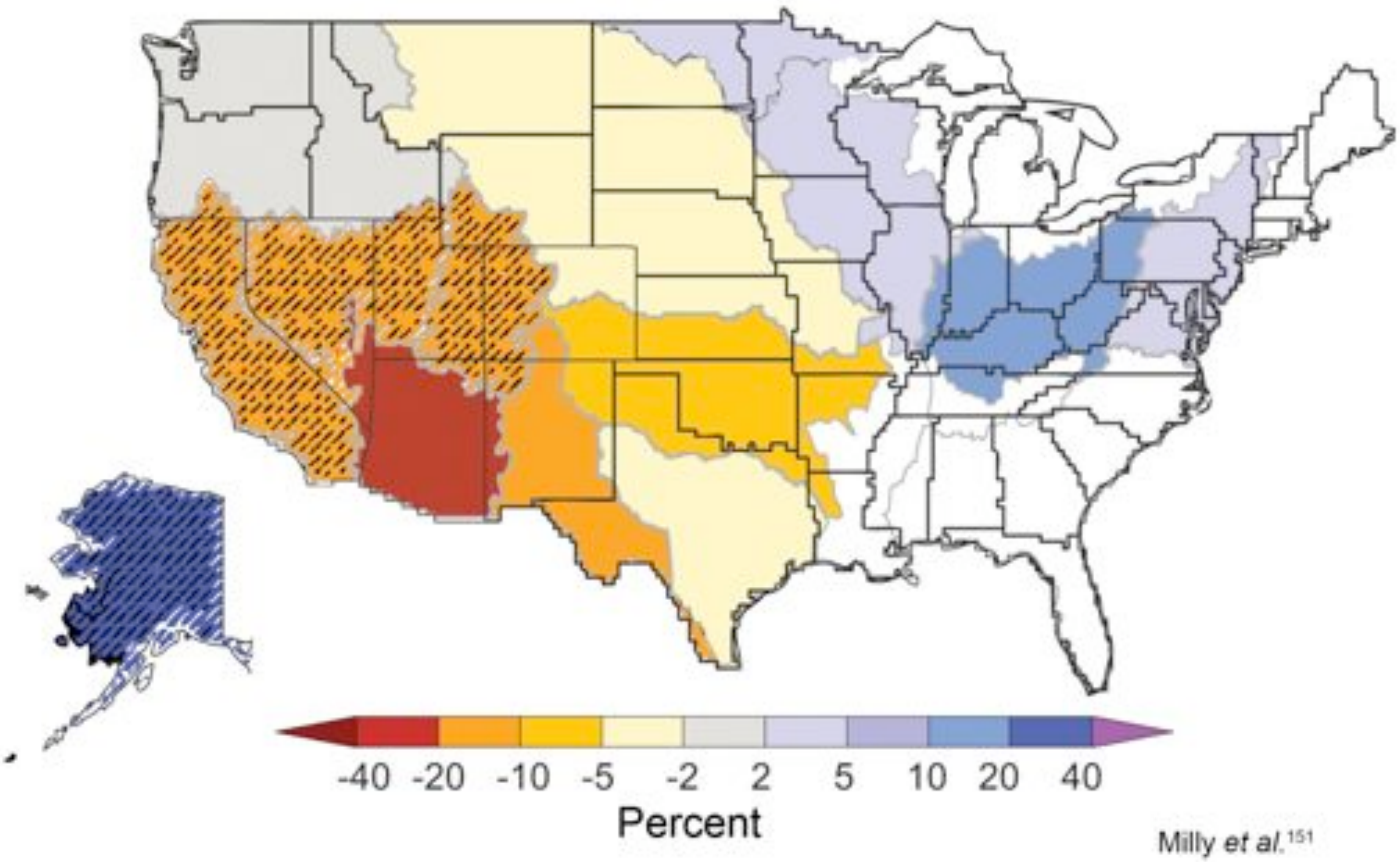
- **Conflicts or mass migrations of people resulting from food scarcity and other resource limits, health impacts, or environmental stresses in other parts of the world could threaten U.S. national security.**

  » *Global Climate Change Impacts in The United States, p. 11*



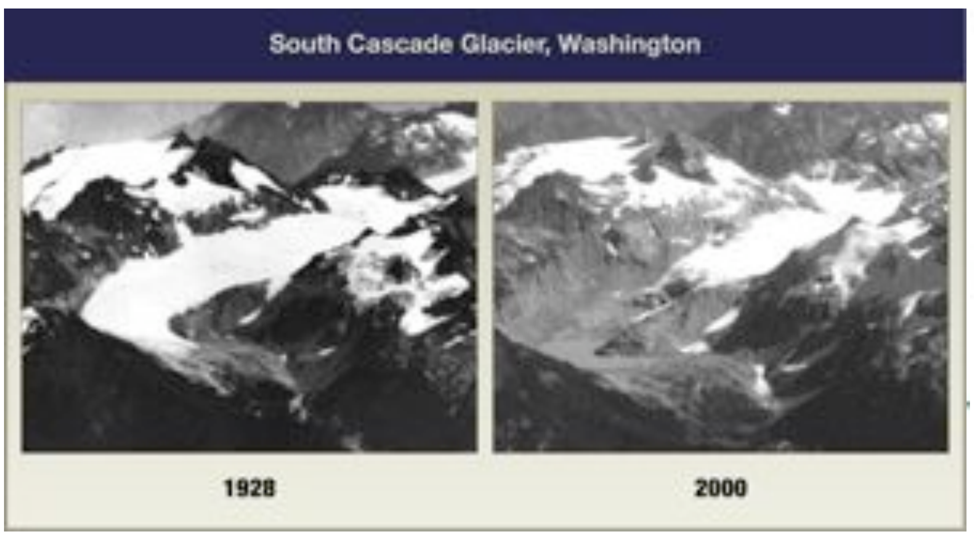
Measured Changes in Water Resources

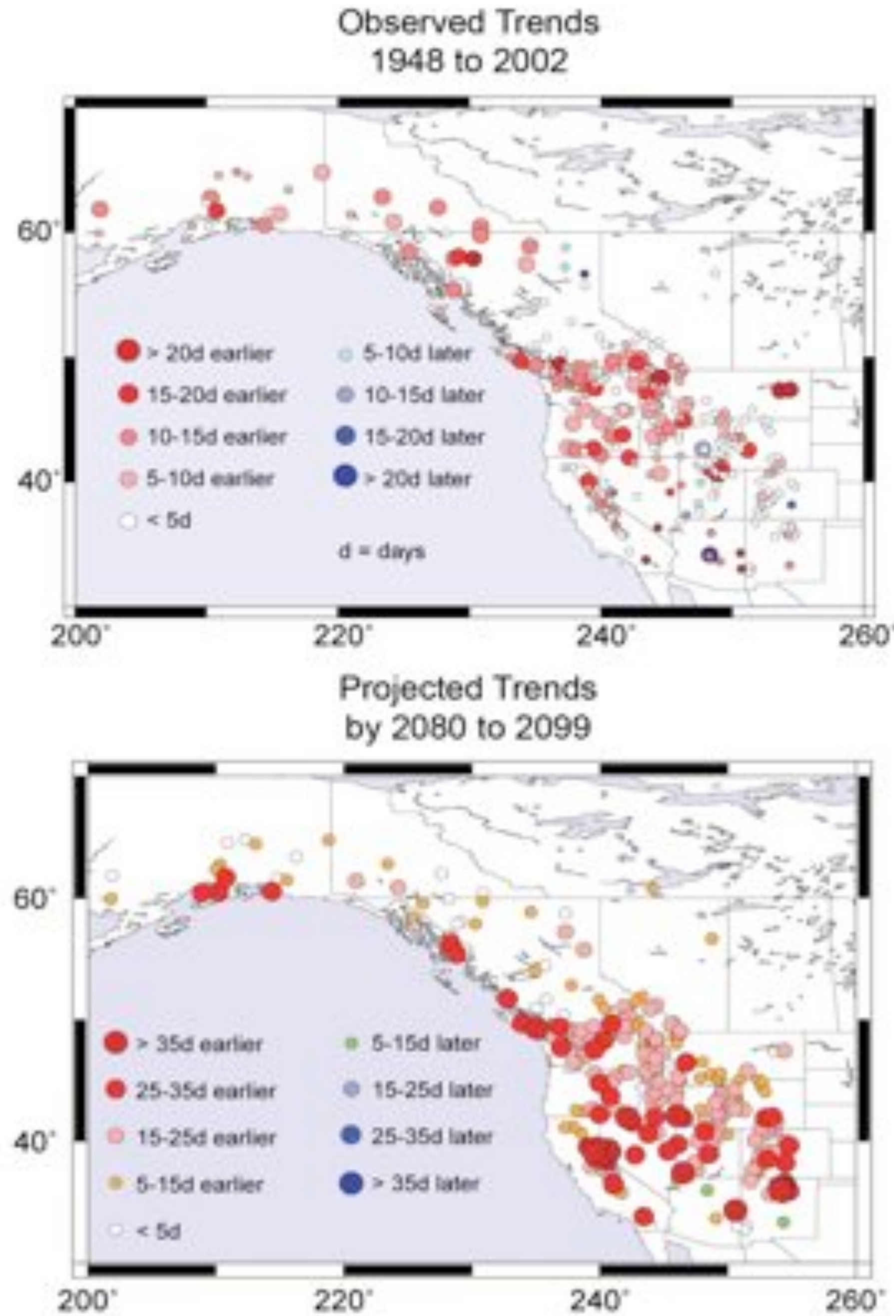
- ssss



Projected changes in median runoff for 2041-2060, relative to a 1901-1970 baseline, are mapped by water-resource region. Colors indicate percentage changes in runoff. Hatched areas indicate greater confidence due to strong agreement among model projections. White areas indicate divergence among model projections. Results are based on emissions in between the lower and higher emissions scenarios.[91]

- **Changing Water Runoff Timing Intensifies Water Management Challenges**



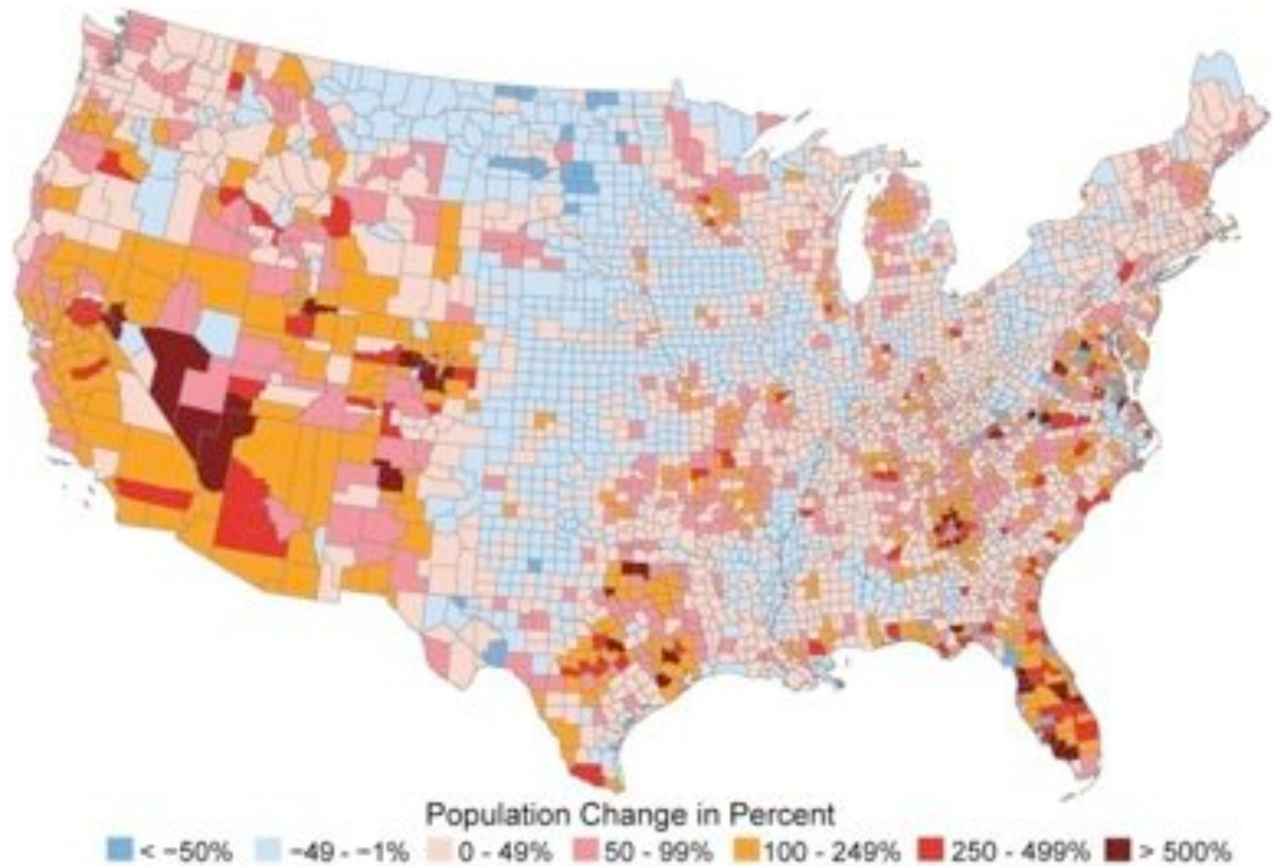


Stewart *et al.*[157]

Top map shows changes in runoff timing in snowmelt-driven streams from 1948 to 2002 with red circles indicating earlier runoff, and blue circles indicating later

Increasingly Bicoastal Population

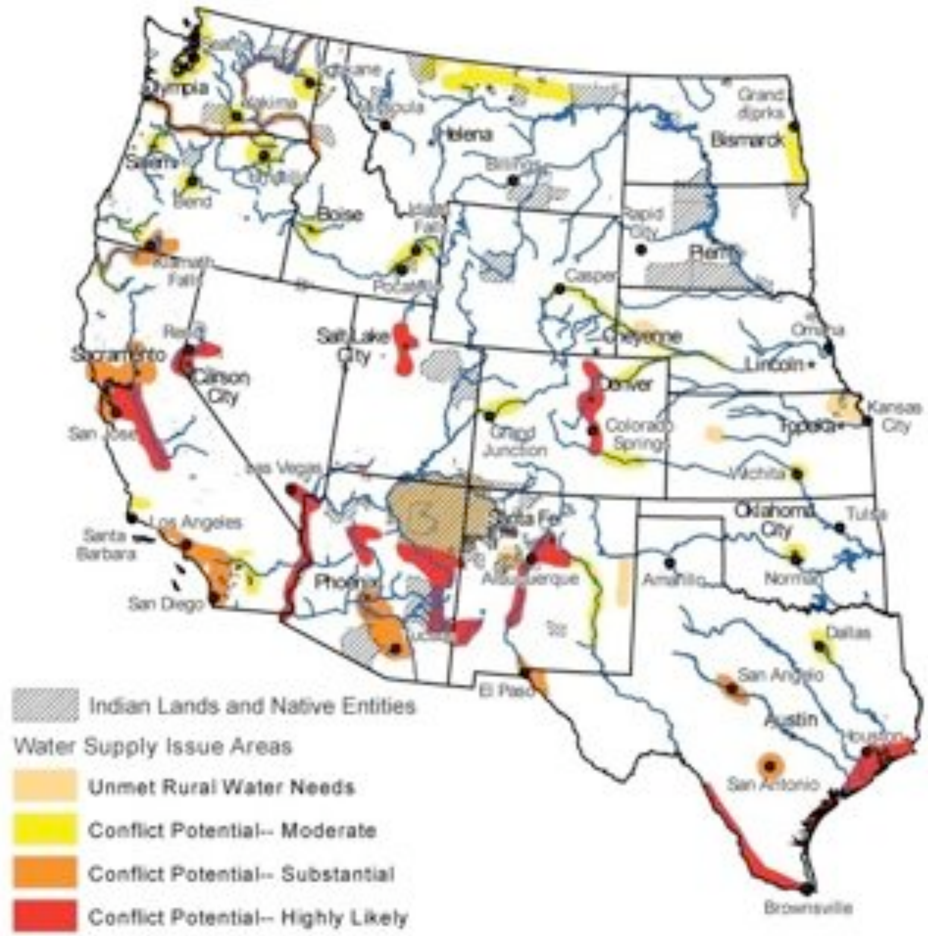
Population moving to drought-prone regions of the West.



U.S. Census

The map above, showing percentage changes in county population between 1970 and 2008, graphically illustrates the large increases in places that require air conditioning. Areas with very large increases are shown in orange, red, and maroon. Some places had enormous growth, in the hundreds of thousands of people. For example, counties in the vicinity of South Florida, Atlanta, Los Angeles, Phoenix, Las Vegas, Denver, Dallas, and Houston all had very large increases.

Increasing Domestic Conflict Over Water Rights by 2025



USBR[171]

The map shows regions in the West where water supply conflicts are likely to occur by 2025 based on a combination of factors including population trends and potential endangered species' needs for water. The red zones are where the conflicts are most likely to occur. This analysis does not factor in the effects of climate change, which is expected to exacerbate many of these already-identified issues.[171]

Scarcity of Water and Food

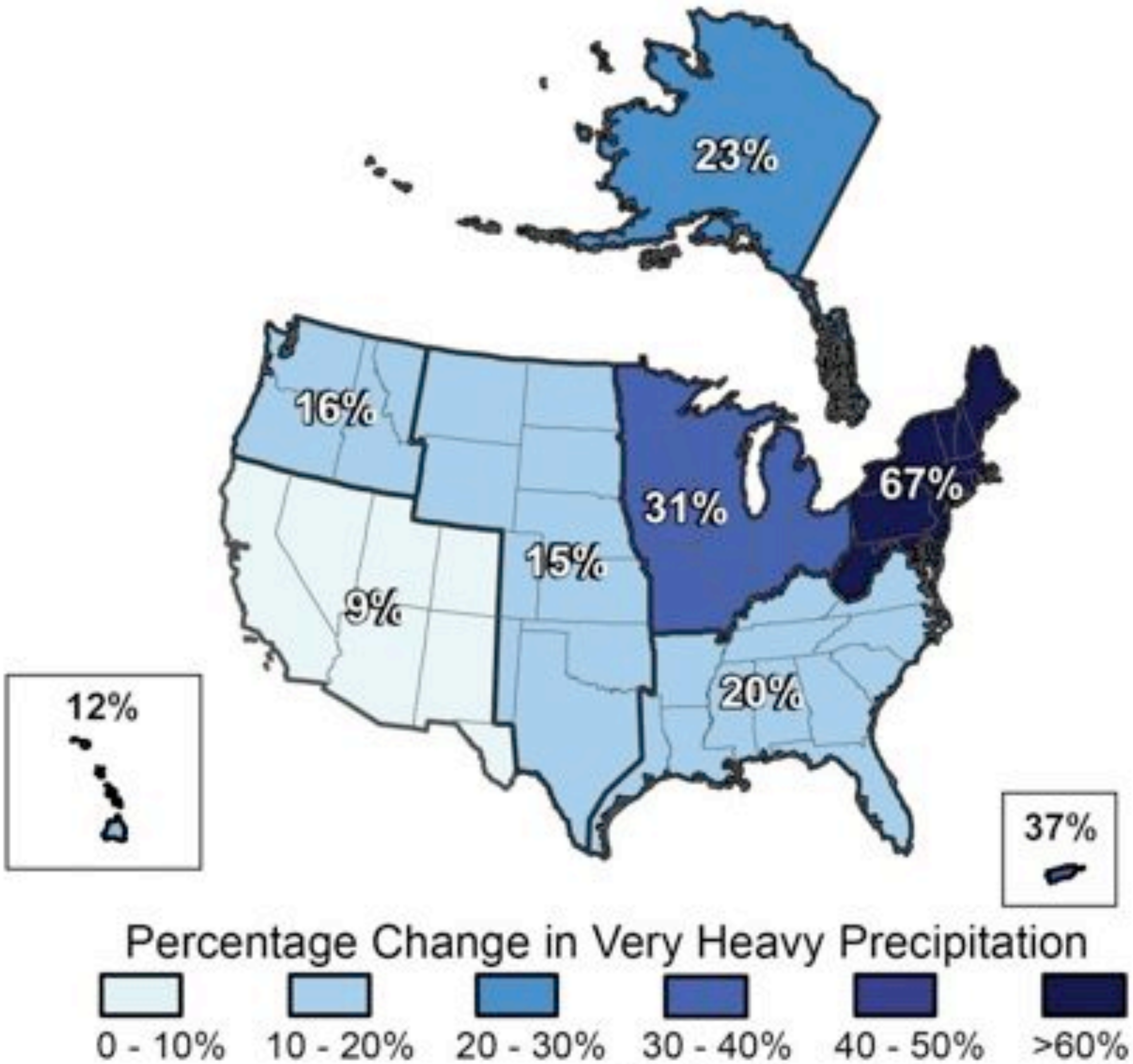
- **Lack of access to stable supplies of water is reaching critical proportions, particularly for agricultural purposes, and the problem will worsen because of rapid urbanization worldwide and the roughly 1.2 billion persons to be added over the next 20 years.**

- **Today experts consider 21 countries, with a combined population of about 600 million, to be either cropland or freshwater scarce. Owing to continued population growth, 36 countries, with about 1.4 billion people, are projected to fall into this category by 2025.**
  - p. viii

Conflict Over Resources

- **Types of conflicts we have not seen for awhile – such as over resources – could reemerge. Perceptions of energy scarcity will drive countries to take actions to assure their future access to energy supplies…**

- **With water becoming more scarce in Asia and the Middle East, cooperation to manage changing water resources is likely to become more difficult within and between states.**
    - p. x



Updated from Groisman *et al.*[113]

The map shows the percentage increases in very heavy precipitation (defined as the heaviest 1 percent of all events) from 1958 to 2007 for each region. There are clear trends toward more very heavy precipitation for the nation as a whole, and particularly in the Northeast and Midwest.

Increasing Frequency of Very Heavy Precipitation Events

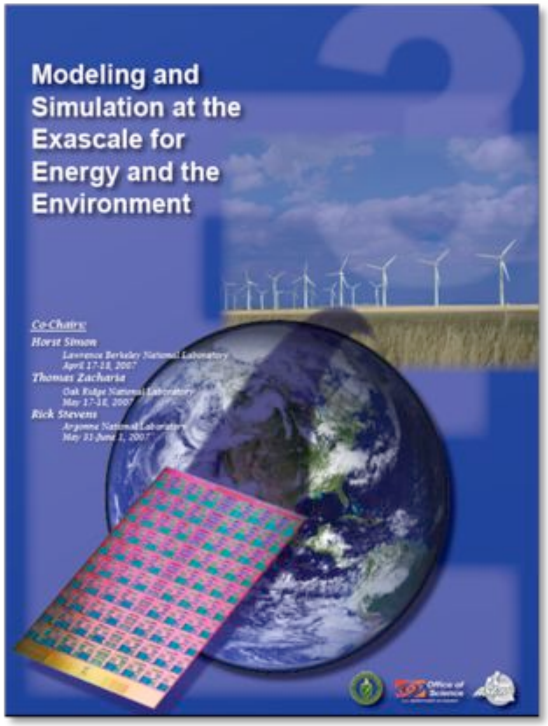
Increased Flooding of Coastal Cities and Infrastructure

Increasing area of coastal cities subject to flooding

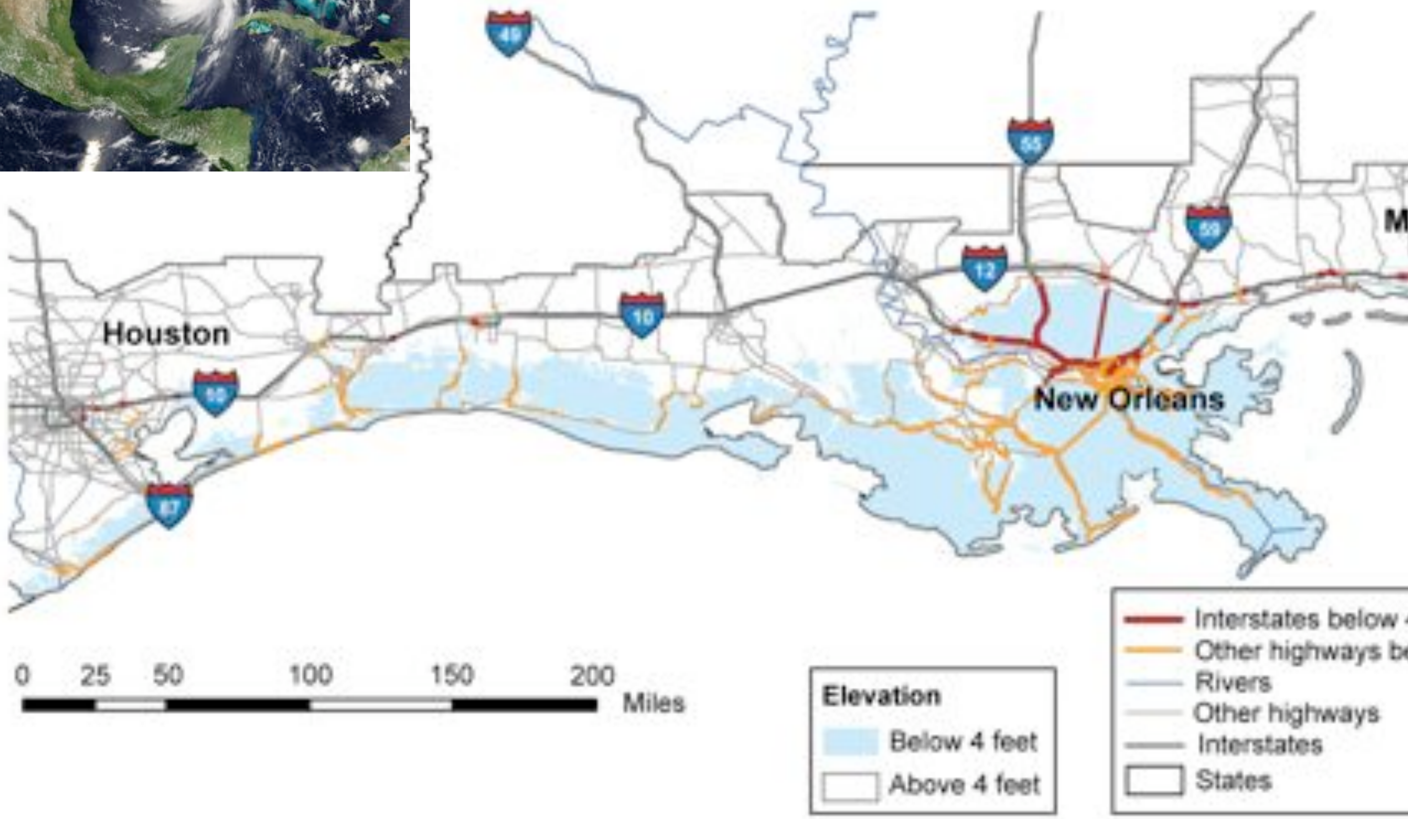
The light blue area above depicts today's FEMA 100-year flood zone for the city (the area of the city that is expected to be flooded once every 100 years). With rising sea levels, a 100-year flood at the end of this century (not mapped here) is projected to inundate a far larger area of New York City, especially under the higher emissions scenario.[91] Critical transportation infrastructure located in the Battery area of lower Manhattan could be flooded far more frequently unless protected. The increased likelihood of flooding is causing planners to look into building storm-surge barriers in New York Harbor to protect downtown New York City.[234,370,371]

Sea Level Rise Threatens Coastal Infrastructure

- **Critical infrastructure – including oil refineries (80% of U.S. refining capacity is is at <= 1.5m above sea level), nuclear power stations, ports, and industrial facilities – is often concentrated around coastlines.**
  - *Modeling and Simulation at the Exascale for Energy and the Environment (2007), p.9*



Gulf Coast Vulnerability: 2400 Miles of Roadway Below 4 Feet Elevation



CCSP SAP 4.7[217]

Within 50 to 100 years, 2,400 miles of major roadway are projected to be inundated by sea-level rise in the Gulf Coast region. The map shows roadways at risk in the event of a sea-level rise of about 4 feet, within the range of projections for this region in this century under medium- and high-emissions scenarios.[91] In total, 24 percent of interstate highway miles and 28 percent of secondary road miles in the Gulf Coast region are at elevations below 4 feet.[217]

Sea Level Rise

- **During the last interglacial period (~120 thousand years ago, ka) with similar CO2 levels to pre-industrial values and arctic summer temperatures warmer than today, sea level was 4-6 m above present.**

- **During the last two deglaciations sea level rise averaged 10-20 millimeters per year with large "meltwater fluxes" exceeding sea level rise of 50 millimeters per year lasting several centuries.**
  - USGS Report, Abrupt Climate Change, Chapter 2. Rapid Changes in Glaciers and Ice Sheets and their Impacts on Sea Level
  - http://downloads.climatescience.gov/sap/sap3-4/sap3-4-final-report-ch2.pdf – p. 2

U.S. DEPARTMENT OF ENERGY | Office of Science

Ice Sheets and Glaciers

- **If the entire Greenland Ice Sheet melted, it would raise sea level by about 20 feet.**
- **The West Antarctic ice Sheet…contains enough water to raise global sea levels by about 16 to 20 feet.**
- **If the East Antarctic Ice Sheet melted entirely, it would raise global sea level by about 200 feet.**
- **Complete melting of these ice sheets over this century is thought to be virtually impossible, although past climate records provide precedent for very significant decreases in ice volume, and therefore increases in sea level.**

  – *Global Climate Change Impacts in The United States*, p. 18

How Much Will the Sea Rise?

- **There is some evidence to suggest that it would be virtually impossible to have a rise of sea level higher than about 6.5 feet by the end of the century.**
  - *Global Climate Change Impacts in The United States*, p. 25

Research Challenges

- **How can a visual analysis system represent data sets for which there is not even one pixel per data point without falsely conveying a sense of uniformity in data? How can small anomalies in petascale to exascale data be made perceptible?**
- **How can data from multi-scale, multi-physics simulations and experiments be represented so that the complexity of the science is accessible while also maintaining comprehensibility of the displays?**
- **Can visual analytic systems support collaborative data analysis and communication across distributed multi-disciplinary teams of scientists, without loss of disciplinary semantic context in the visual representation?**
- **Can visual analysis systems help scientists understand and manage the uncertainty that results from exascale system characteristics?**
- **Can visual analysis systems provide insight to developers and application scientists about the behavior of applications on exascale computation platforms and support application optimization?**
- **How can the accuracy of communication with the human users of visual analysis systems be validated and maintained across multiple data sets and potentially vast user communities?**

Workshop Report

- ***Scientific Discovery at the Exascale: Report from the DOE ASCR 2011 Workshop on Exascale Data Management, Analysis and Visualization*, February 2011, Houston, TX**
- ***http://www.olcf.ornl.gov/wp-content/uploads/2011/01/Exascale-ASCR-Analysis.pdf***
- **Organizer: Sean Ahern, ORNL; Co-Chairs: Arie Shoshani, LBNL, and Kwan-Liu Ma, UC Davis**
- **Principal Finding: "The disruptive changes posed by a progressive movement towards the exascale in HPC threaten to derail the scientific discovery process. Today's success in extracting knowledge from large HPC simulation output are not generally applicable to the exascale era, and simply scaling existing techniques to higher concurrency is not sufficient to meet the challenge." – p. 1**

ASCR-BES Workshop on Data

"Data and Communications in Basic Energy Sciences: Creating a Pathway for Scientific Discovery"

October 24-25, 2011 **Bethesda, MD**

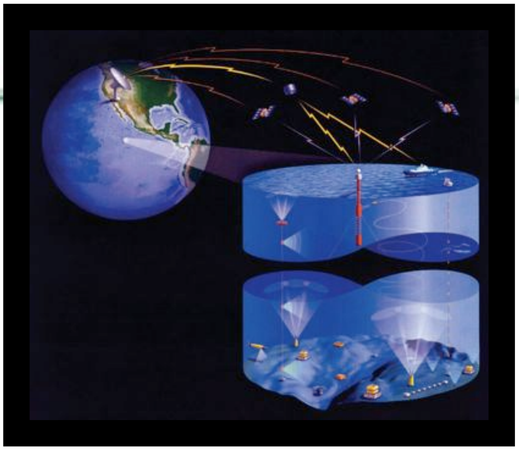
Goals & Objectives

- **Identify and review the status, successes, and shortcomings of current data (including analysis and visualization) and communication pathways for scientific discovery in the basic energy sciences;**
- **Ascertain the knowledge, methods and tools needed to mitigate present and projected data and communication shortcomings;**
- **Consider opportunities and challenges related to data and communications with the combination of techniques (with different data streams) in single experiments;**
- **Identify research areas in data and communications needed to underpin advances in the basic energy sciences in the next ten years;**
- **Create the foundation for information exchanges and collaborations among ASCR and BES supported researchers, BES scientific user facilities and ASCR computing and networking facilities.**

Co-Chairs

- **Peter Nugent, NERSC, & J. Michael Simonson, SNS**

The Power of Visualization



- **You can *do* science without graphics. But it's very difficult to *communicate* it in the absence of pictures. Indeed, some insights can only be made widely comprehensible as images. How many people would have heard of fractal geometry or the double helix or solar flares or synaptic morphology or the cosmic microwave background, if they had been described solely in words?**

- **To the general public, whose support sustains the global research enterprise, these and scores of other indispensable concepts exist chiefly as images.**

NSF Science and Visualization Challenge 2007, Special Report

http://www.nsf.gov/news/special_report/scivis/index.jsp?id=challenge

Thank You!

Lucy Nowell

lucy.nowell@science.doe.gov

http://science.energy.gov/ascr/index.html