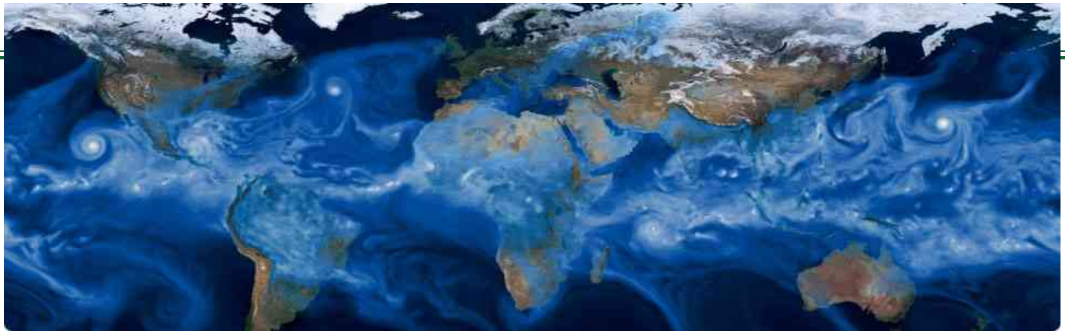


Visual Data Exploration and Analysis of Ultra-large Climate Data

E. Wes Bethel

22 September 2011

DOE/BER Program Meeting

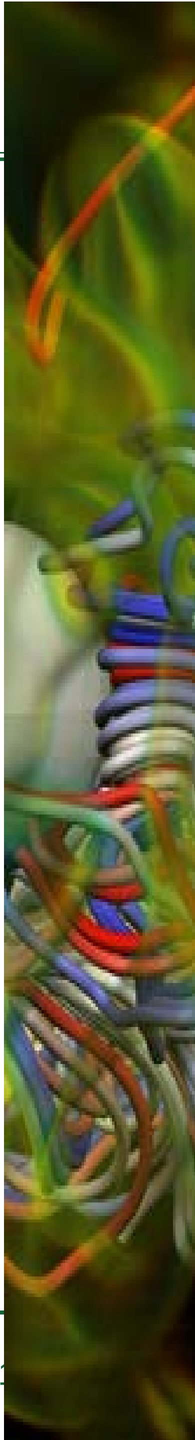
Washington, DC

Motivation/Problem Statement

- Earth system modeling: requires high-throughput processing capability for ultra-high resolution model output.
  - Legacy tools incapable of meeting throughput demands of AR5-class model output (0.25° or 0.1° spatial resolution).
- Regional and global climate modeling: analysis and process-based evaluation of multi-model climate change projections for the 21st century using innovative metrics.
  - Innovative capabilities like detecting/analyzing extreme events, etc. that are easily accessible to the climate science community and that can process high spatiotemporal resolution data.
- Deliver software to the climate science community that can provide high-throughput and advanced analysis capabilities.

How Big is Big? How Digested?

- Current CAM5.1 0.25° runs for a ~25 year period produce about 100TB of model output (Mike’s talk, Prabhat’s poster).
  - Analysis performed on 7000 cores of Hopper at NERSC.
- Data size multipliers:
  - Different models
  - Ensemble runs
  - Different experiments
- The technology we’re using has been proven to work on datasets consisting of trillions of grid points and run on tens of thousands of processors at NERSC, ORNL, ANL, TACC, and LLNL.

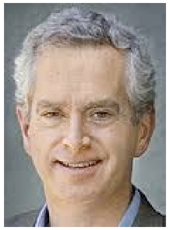


This Project

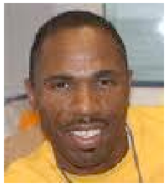
- Our team funded under BER LAB 10-05 to develop and deliver software that:
    - Supports high-throughput visual data analysis and exploration of ultra-high resolution (spatial, temporal) model output (and observational data)
    - Is proven to address pressing climate science questions:
        - Better understanding of extreme events and trends in a changing climate (near-ish term)
        - Regional-scale drivers, attributions, and consequences in a changing climate (longer term)

Our Team

- Climate Scientists
  - Collins (LBNL), Wehner (LBNL)
- Computer Scientists
  - (LBNL) Bethel, Childs, Prabhat, Weber, Wu,
  - (LLNL) Doutriaux, Williams
  - (LANL) Strelitz (no picture)
- Statisticians
  - Kaufman (UCB), Paciorek (UCB)







Those pictured here giving presentations or posters at this meeting.

Our Approach

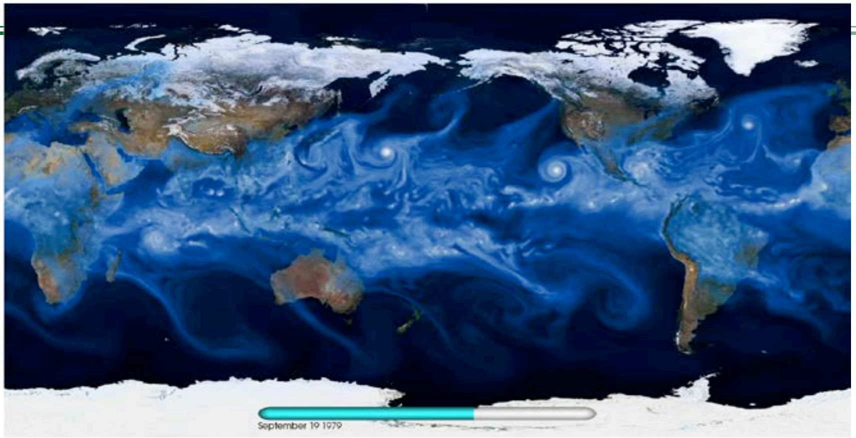
- Climate science:
  - Use “live” case studies as proxies to define requirements, field-test new technologies.
- Technology adaptation/evolution
  - Develop new methods to support detection and analysis of “features” (e.g., extreme events)
- Software
  - Build upon proven, production-quality visual data analysis and exploration software infrastructure
  - Close partnership with a Williams’ led 10-05 team aimed at delivering new capabilities from two 10-05 teams as part of the UV-CDAT system/infrastructure.

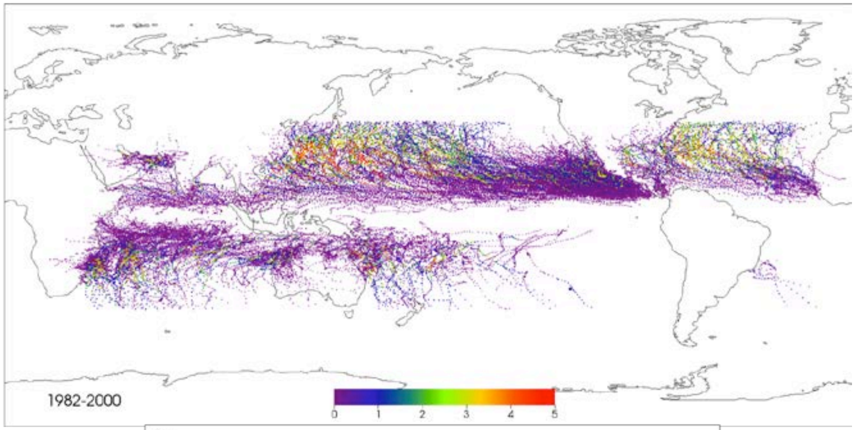
Progress

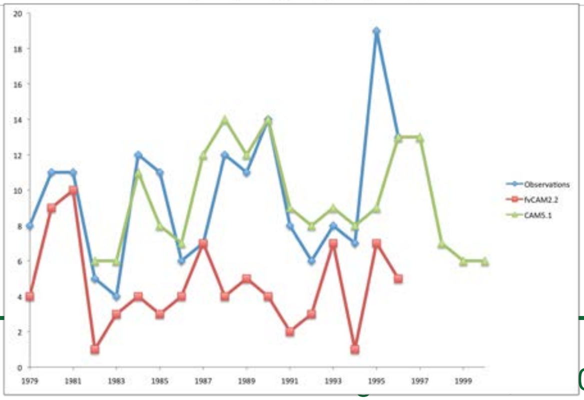
- Science case studies
  – Cyclone detection/analysis
  – Atmospheric rivers
  – Extremes in precipitation
- Technology adaptation and algorithmic advances
  – General Extreme Value theory (GEV) and its application to extreme event analysis
  – Uncertainty characterization and visualization (Strelitz talk)
  – Multivariate (ensemble) comparative analysis
- Software engineering and infrastructure (Hank's talk)
  – Temporal parallelization
  – Integration of R into VisIt, VTK for use in analysis studies at high concurrency.
  – Interfaces to UV-CDAT infrastructure.

Cyclone Detection

- Science objective: quantify hurricane/cyclone characteristics in a changing climate.
- Recent CAM5 0.25° runs at for 1982-2000 result in 100TB of model output.
- GFLD tracking code parallelized over time.
- Analysis time:
  - 2hrs wallclock on 7K CPUs.
  - Est. serial time: 583 days.
- See: Prabhat’s poster, Mike Wehner’s presentation.

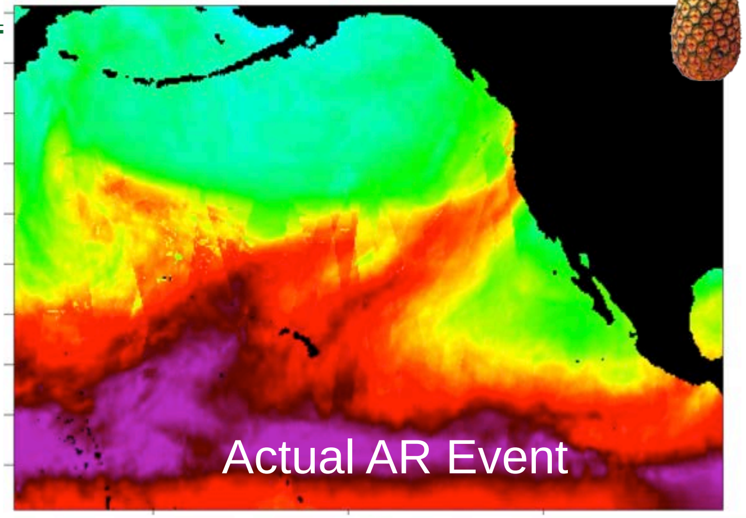


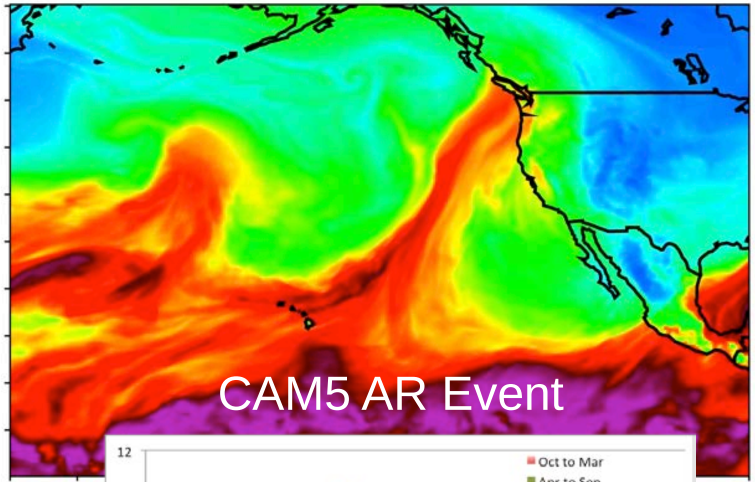


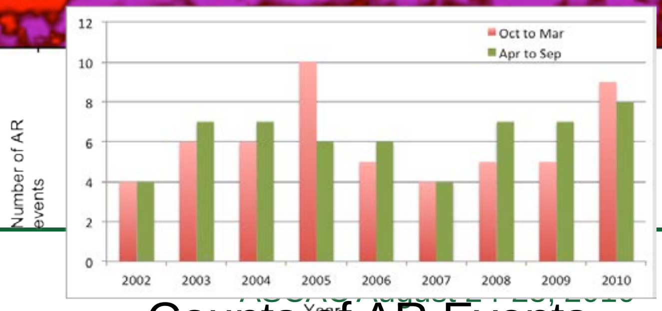


Atmospheric Rivers (aka “Pineapple Express”)

- Science objective: quantify AR characteristics in a changing climate.
- AMSR satellite data (140GB, 2002-2010), fvCAM (500GB, 1979-1993)
- Parallelize detection algorithm over time.
- Analysis time:
  - 3 seconds, 140GB satellite dataset 3400-way parallel (to 10K cores on fvCAM output)
  - Estimated serial time: 3 secs, linear scaling characteristics.
- See: Prabhat’s poster, Mike Wehner’s presentation.



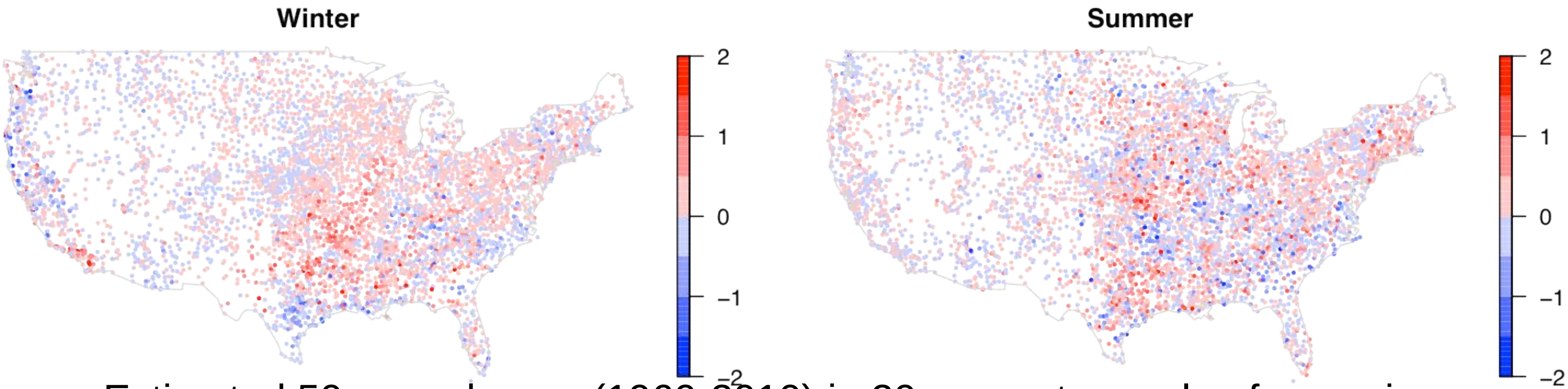




Counts of AR Events

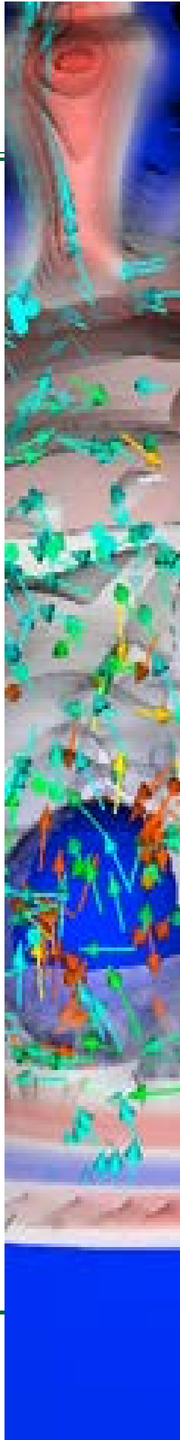
Generalized Extreme Value Analysis

- Science objective: estimate past and future changes in seasonal precipitation extremes using model output and observed data.
- Approach: local likelihood approach to spatial analysis of extremes
- Results: analysis of station data and model output.
- See Prabhat’s poster for more information.



Estimated 50 year change (1960-2010) in 20 year return value for maximum daily precipitation (inches) by season, based on observations.

Putting these Capabilities In Your Hands



- Questions you might be asking yourself:
  - Q: Sounds great. Is this a real, software implementation or an academic study?
  - A: this is real software.

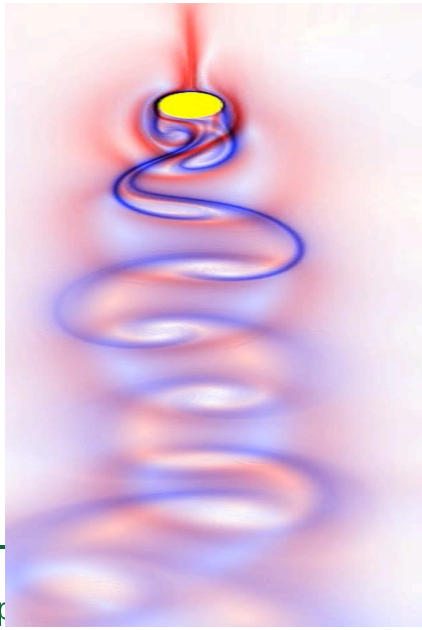
  - Q: How can I get my hands on it?
  - Q: Is it extensible?
  - Q: When will it be available?
  - Q: Will it work with My Big Problem?

Software Architecture and Engineering

- Most of the information about our s/w design, engineering, and distribution plans are in:
  - Hank’s talk (10:15) – large-scale capabilities of UV-CDAT.
  - Dean’s talk (10:45) – how this capability will be available through ESG gateways.
  - Claudio’s talk (11:30) – infrastructure behind UV-CDAT.

One Additional Observation

- These projects are successful – able to deliver an astounding amount of software capability on a relatively short time horizon – due to leverage.
- UV-CDAT contains several technologies that are the result of many years' worth of funding an effort:
  - VisIt – high performance parallel visual data analysis and exploration (SciDAC VACET, NNSA)
  - VisTrails – provenance management (SciDAC VACET)
  - ParaView – parallel visualization (NNSA)
  - ESG – climate data management (SciDAC)



The End

My contact information: ewbethel@lbl.gov