Challenges in the analysis of petascale climate model output datasets

Michael Wehner

Scientific Computing Group

Lawrence Berkeley National Laboratory

mfwehner@lbl.gov

Climate model output is a public resource

- ~25 international climate modeling groups have made data available to the scientific community for the IPCC AR4
  - "CMIP3 database"
  - www-pcmdi.llnl.gov
  - Various scenarios
    - 21st century futures
    - 20th century
    - Control (static atmospheric composition)
    - Global in spatial extent
    - Monthly, daily and 6 hourly temporal sampling
  - Strict metadata standards

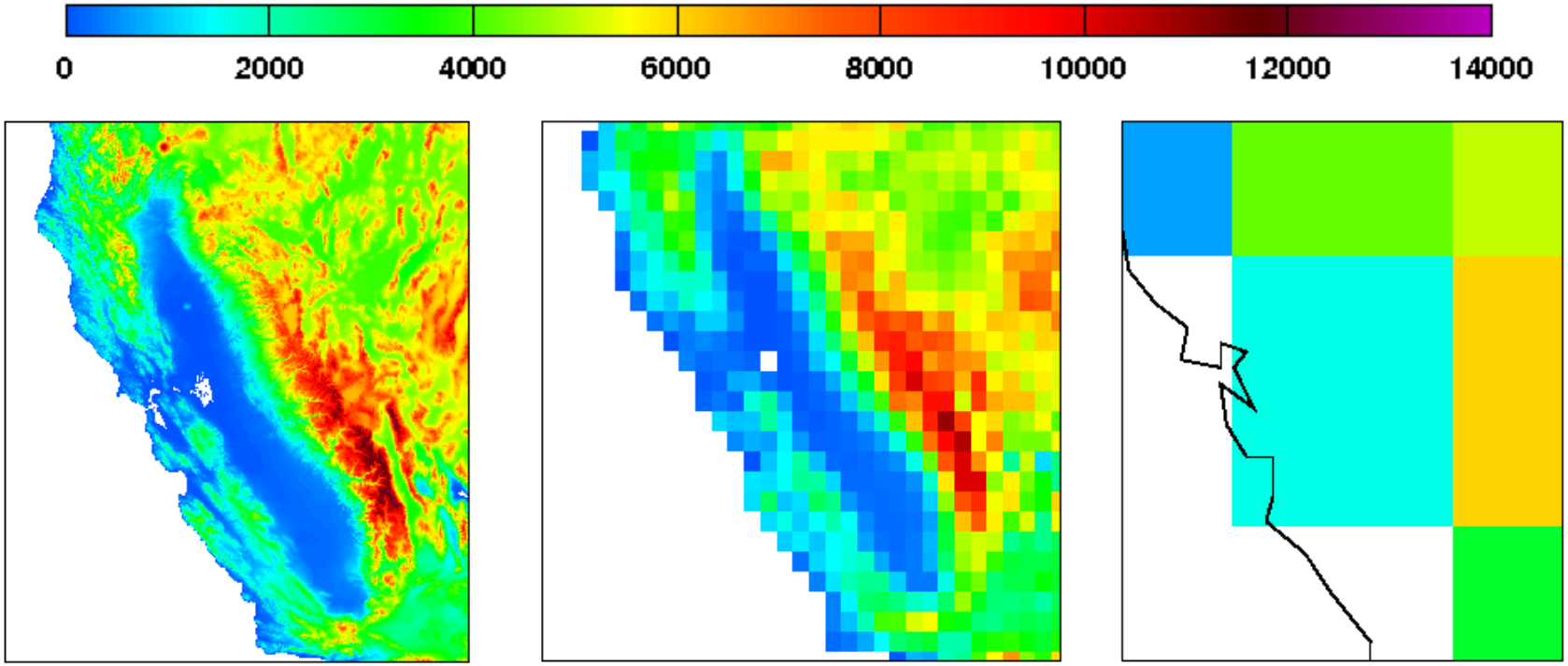
CMIP5 database for the IPCC AR5

- Data will not be centrally collected at LLNL
  - Too big
  - Simulations will be 4 to 16 times larger than AR4
- Distributed via the Earth System Grid
  - Major groups will host their own data
  - Third world will send to LLNL
  - ESG provides a unifying virtual location
  - ESG allows powerful search functions via the metadata
    - earthsystemgrid.org

Evolution of climate model resolution

- 2004 AR4
  - 200km
- 2008
  - 100km
- 2011 AR5
  - 50km
- 201x
  - 25km (approaching the limit of current cloud simulation)
- 2016+
  - 1km (A new class of model, cloud system resolving)

Surface elevation in feet



1km
Cloud system resolving models

25km
Upper limit of current climate models

200km
Typical resolution of IPCC AR4 models

How big are climate model datasets?

- **Large, but analysts generally use only 1 or 2 variables at a time.**
  - **Most are surface variables.**
  - **per century (GB):**

	200km	100km	50km	25km	1km
Monthly	0.1	0,5	2.0	8.0	6400
Daily	3.8	15.3	61.2	245.0	200000
6 hour	15.3	61.2	245	980	800000

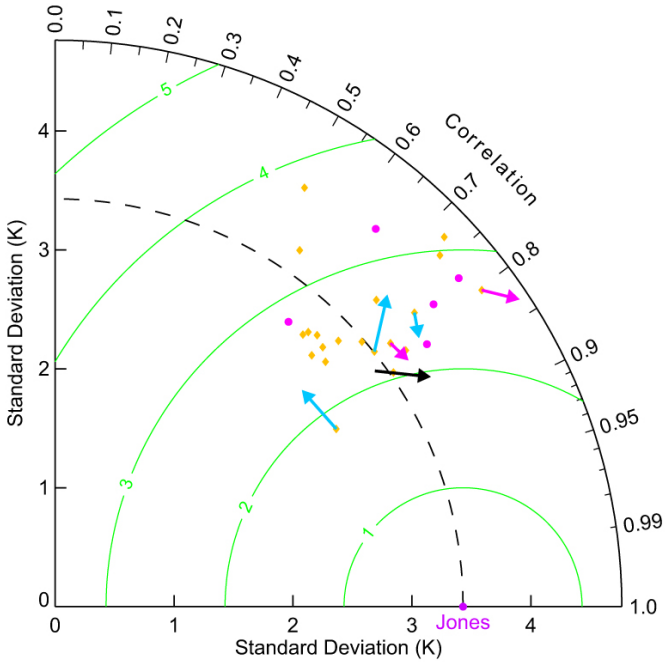
Tools

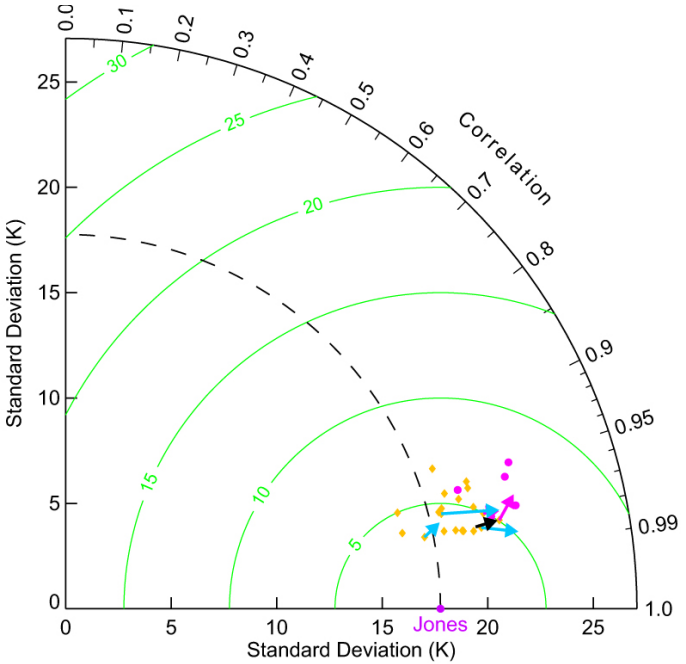
- Data can be generated faster than it can be analyzed!
  - Standard analysis tools are all serial
- But there is more parallelism in the analysis of model data than in producing it.
  - Time, space & realization are embarassingly parallel in many cases.
- Modest parallelism will suffice (a few hundred processors)
- Recommendation: DOE fund development of parallel climate model data analysis tools (CDAT)

A short survey of selected climate model analysis techniques

Ranking models based on performance

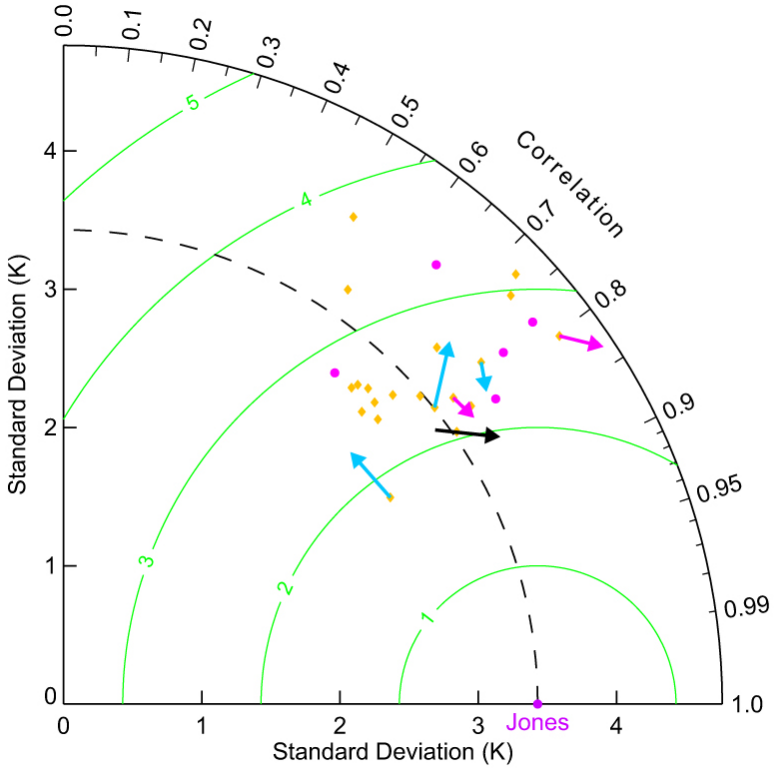
- "Taylor" Diagram
  - DJF Surface air temperature
- Models' ability to simulate the recent past varies.
- Skill varies according to:
  - Quantity of interest
  - Location
  - Season





Ranking models based on performance

- Can we use model skill to:
  - make better forecasts of future climate change?
  - more robustly detect and attribute current climate change?
- Issues:
  - How to quantify "skill"
  - How to weight models
    - Continuous weighting?
    - Binary weighting?
  - How good is "good enough"?



Detection and Attribution of climate change

- Detection
  - Did the climate change in the recent past?
- Attribution
  - Is the cause of climate change natural or human?

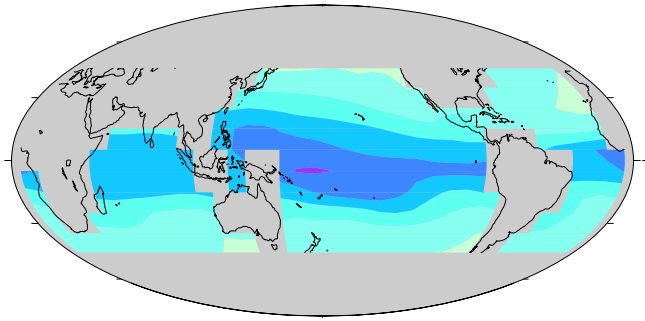
- A classic signal to noise problem
  - EOF analysis (PCA to you)

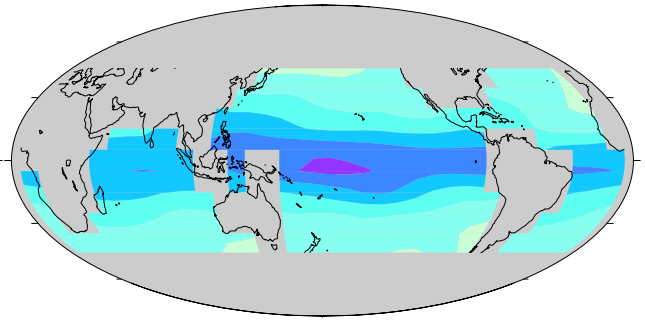
Detection and Attribution of climate change

- Quantify noise with a long control simulation
- Search for “fingerprint” of climate change in “forced” simulations
  - Forcings are natural and/or anthropogenic.
  - Calculate a set of Empirical Orthogonal Functions that describe the simulation of the recent past.
  - First EOF of the model is a guess for the fingerprint
  - Search through the observations for the “detection time”. I.e. that time when the fingerprint becomes consistently identifiable with 5% statistical certainty
  - Patterns of change are important

Patterns of water vapor change in 20CEN runs, control runs, and observations

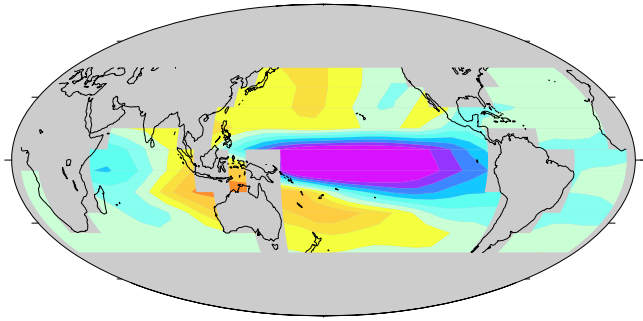
A) EOF1 of W_o
(V model 20CEN runs)

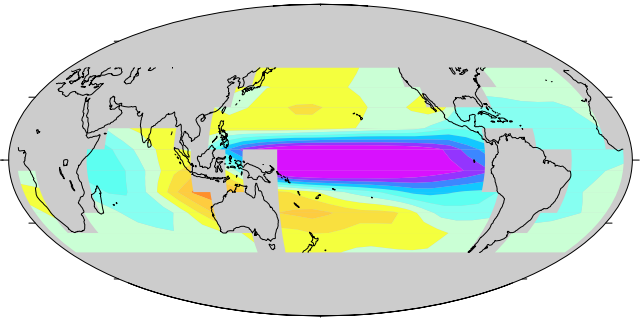




B) EOF1 of W_o
(No-V model 20CEN runs)

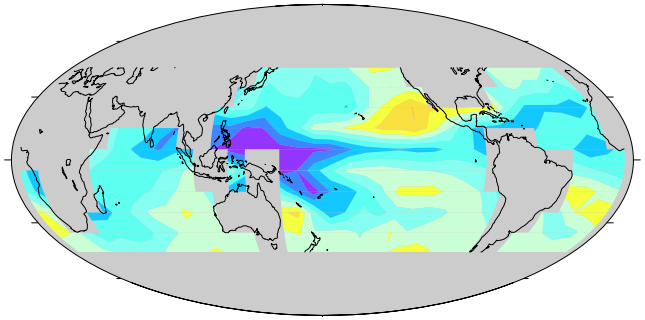
C) EOF1 of W_o
(V model CTL runs)

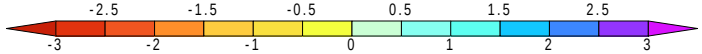




D) EOF1 of W_o
(No-V model CTL runs)

E) Total linear change in W_o
(1988-2006)

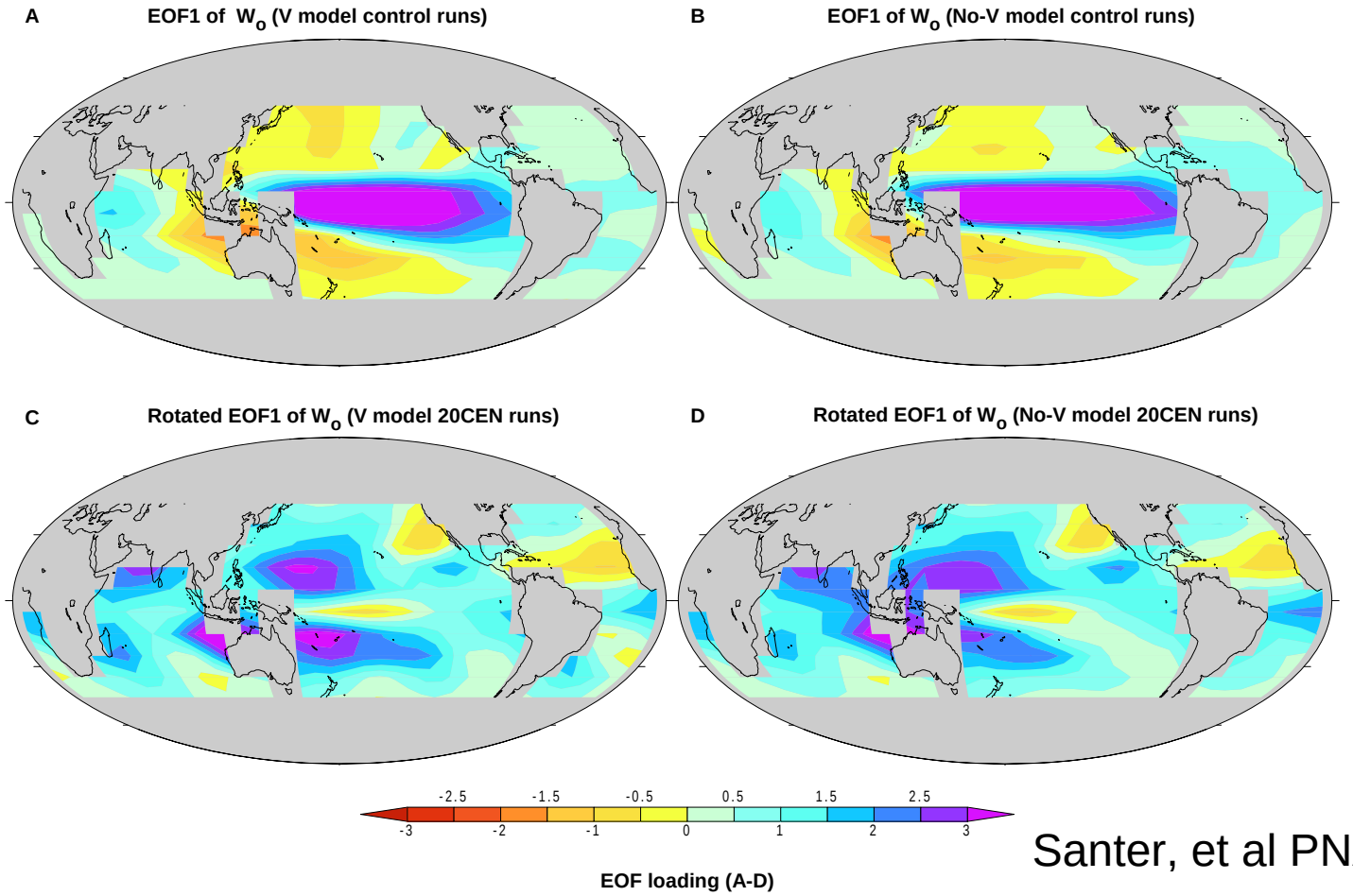




EOF loading (A-D) or total linear change in W_o in kg/m² (E)

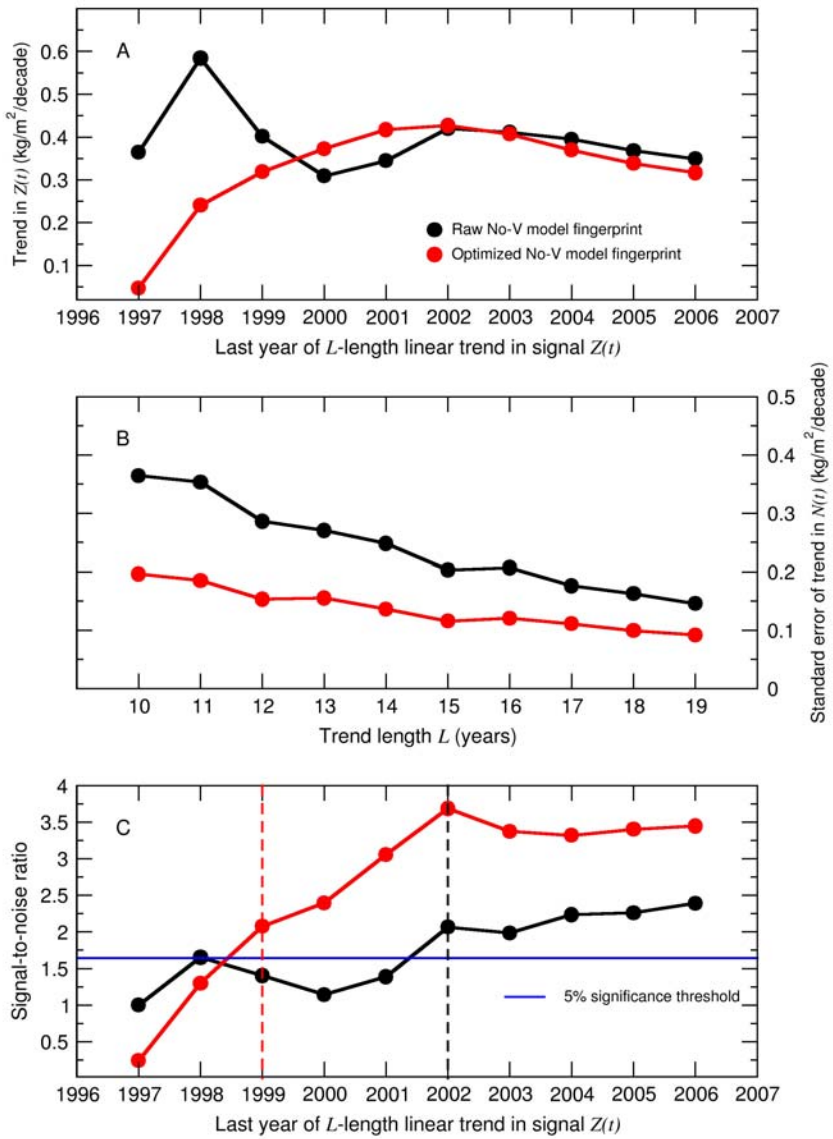
Santer, et al PNAS 2007

Optimization rotates the fingerprint away from high-ENSO noise directions



Santer, et al PNAS 2007

Detection time of human changes to atmospheric water vapor change



Santer, et al PNAS 2007

Detection time of human changes to atmospheric water vapor change

		Raw		*Optimized*	
		V fingerprint	**No-V fingerprint**	**V fingerprint**	**No-V fingerprint**
Mean included	**V noise**	**2002**	**2002**	**1999**	**1999**
	No-V noise	**2002**	**2002**	**1999**	**1999**
Mean removed	**V noise**	**N.D.**	**N.D.**	**2000**	**2000**
	No-V noise	**N.D.**	**N.D.**	**1999**	**1999**

Santer, et al PNAS 2007

N.D. = Not Detectable

Extreme weather in a changing climate

- The most severe consequences of climate change may be due to changes to
  - Hurricanes
  - Droughts
  - Heat Waves
  - Floods

Hurricanes

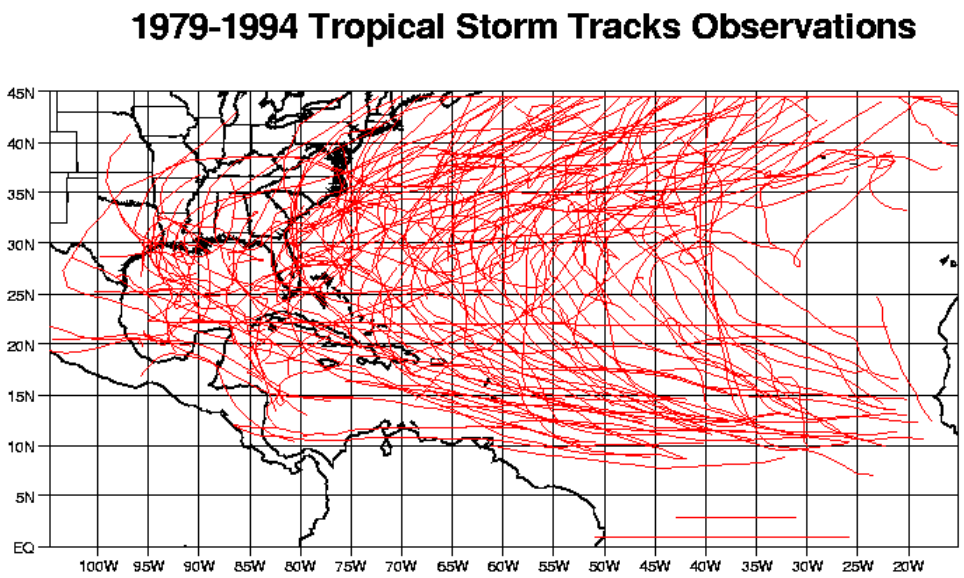
- Climate models finer than ~50km can simulate tropical cyclones reasonably well.
  - intensity of strong storms is sensitive to resolution
- How will the tropical cyclone cycle change as the planet warms?
  - Changes in frequency.
  - Changes in intensity.
  - Changes in duration.
  - Changes in areas vulnerable to tropical cyclones

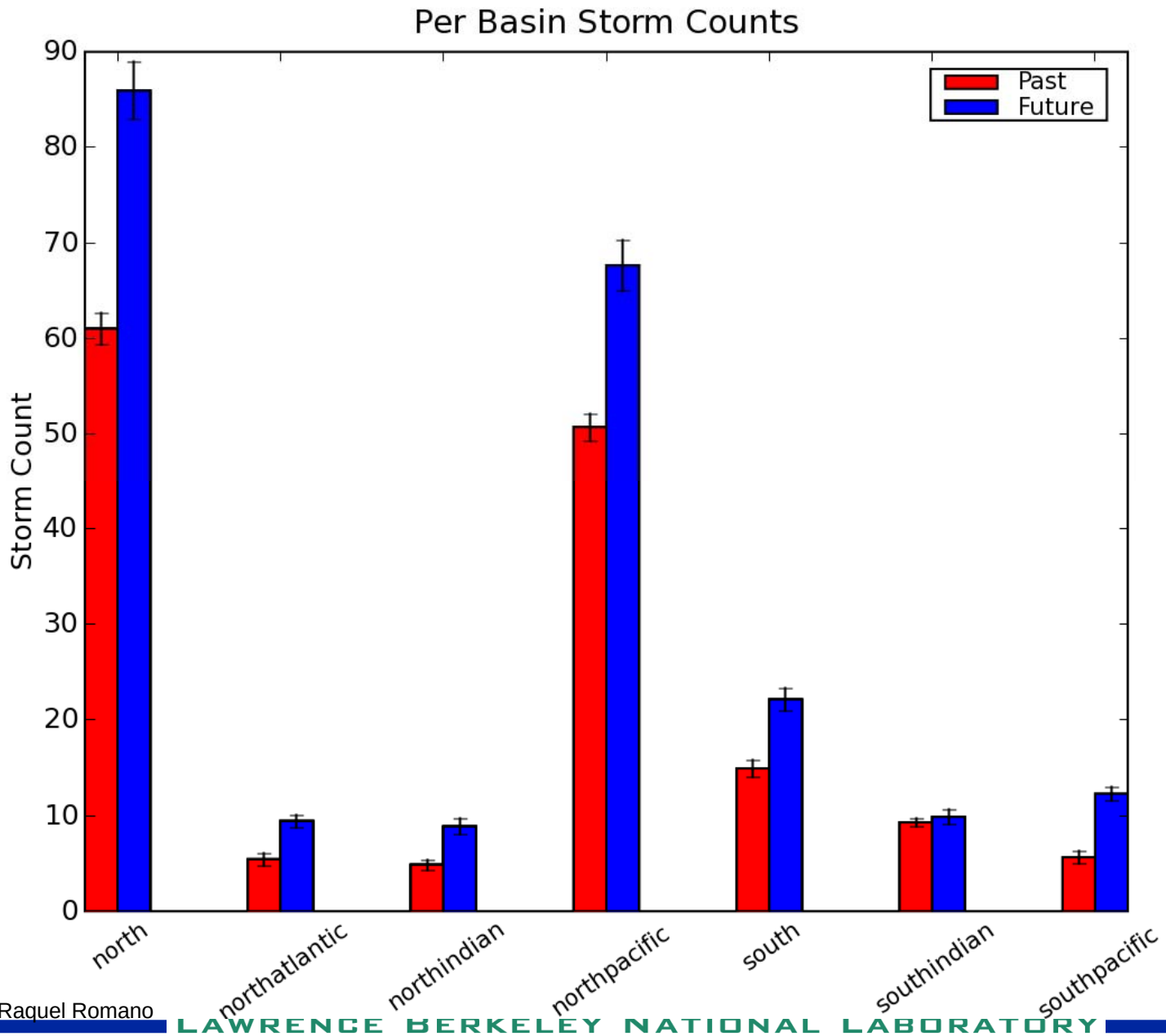
Hurricanes

- High resolution climate models can simulate the statistics of tropical cyclones, but do not reproduce individual named storms.
  - Similar studies could be made with other extreme weather events.
    - Tropical waves
    - Extratropical storms
- To quantify tropical cyclone statistics in model output datasets requires feature recognition and tracking.
  - High temporal resolution (3 or 6 hour)
  - Ensembles make the dataset very large.

Hurricanes

- GFDL tracking algorithm:
  - Find vorticity centers co-located with pressure minima
  - Thickened planetary boundary layer
  - Determine if 500mb air temperature is elevated
  - Look for continuity





Courtesy of Raquel Romano

Generalized extreme value theory

- **GEV distribution is demonstrated to be a good fit to the tails of the distribution of climate model output**

$$F(x) = \begin{cases} e^{-[1-k(x-\xi)/\alpha]^{1/k}} & k \neq 0 \\ e^{-e^{-(x-\xi)/\alpha}} & k = 0 \end{cases}$$

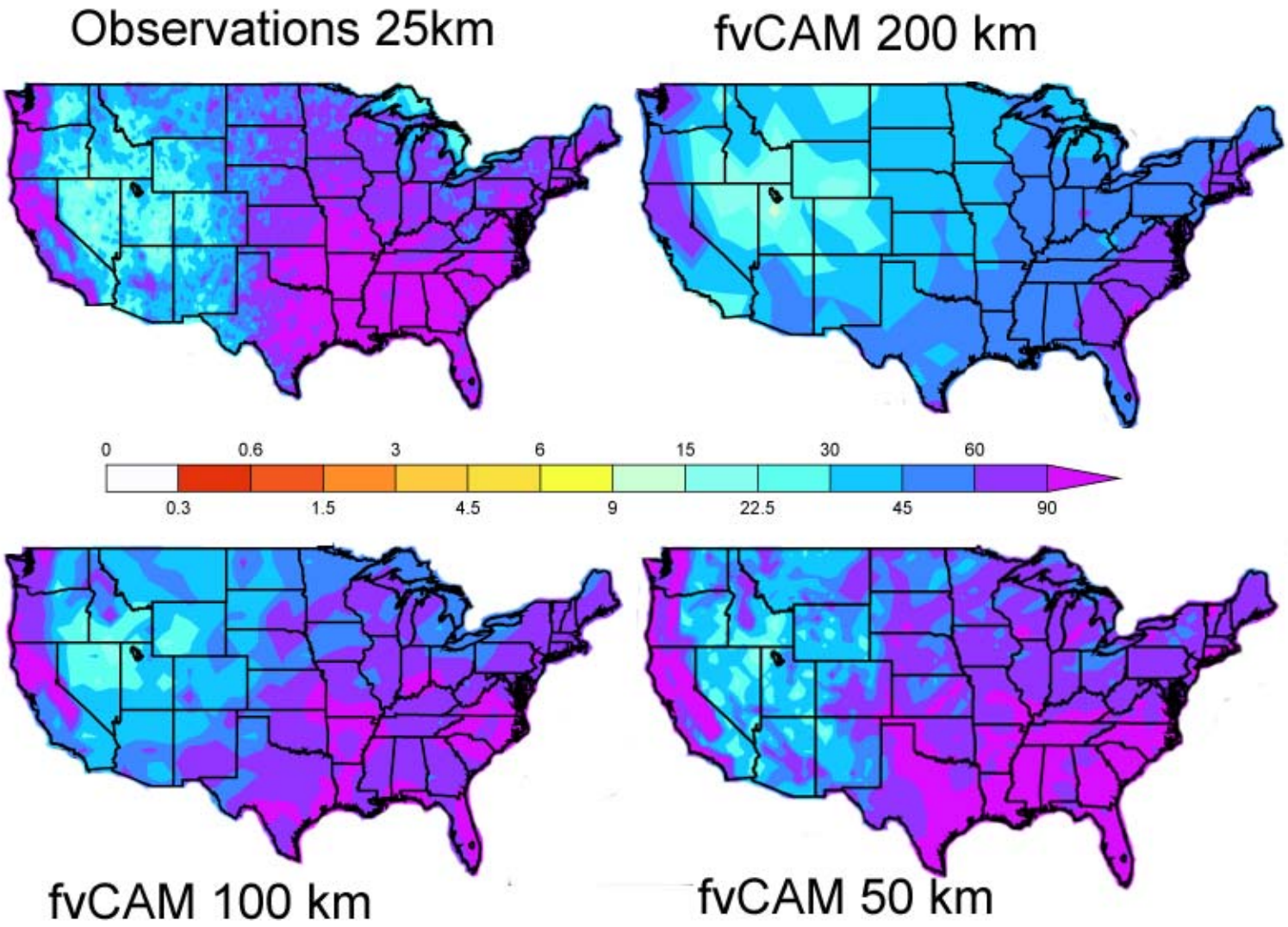
$$\xi = \text{location}$$

$$\alpha = \text{scale}$$

$$k = \text{shape}$$

- **Issue: GEV theory assumes stationary statistics. How to fit the three parameters in a time evolving setting?**

20 year precipitation return value



Climate mode output analysis summary

- Climate model output is a public resource
- Our ability to produce model output will soon exceed our ability to analyze it.
- There are multiple avenues to easy parallelism.
- Most (if not all) climate model output analysis is statistical.
- Powerful mathematical analysis tools are currently in use.
  - There is plenty of room for more!