# ORNL Statistics and Data Sciences

Presented by

George Ostrouchov

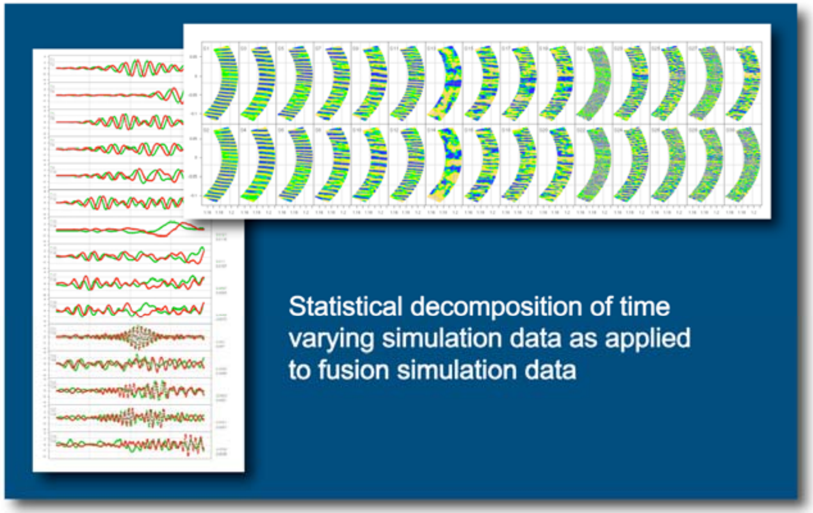
Statistics and Data Sciences
Computer Science and Mathematics Division

## Understanding Variability and Bringing Rigor to Scientific Investigation

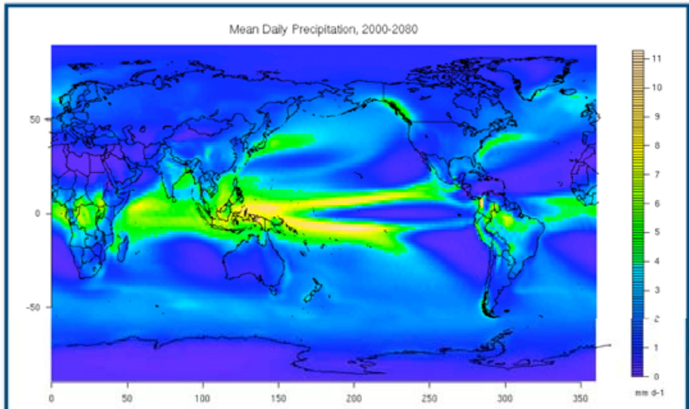
# Common evolutionary steps:
## Experimental Science and Computational Science

- Early computational science relies largely on intuitive design and visual validation.
  - Computational experiments are expensive.
  - Petascale data sets are nearly as opaque as real systems—statistical analysis must select what to visualize.
  - Uncertainty analysis is in its infancy.
- Statistics is a major partner in bringing computational science to the rigor and efficiency standards of experimental science.
  - Methods to see through, examine, and classify variability.
  - Uncertainty quantification.
  - Statistical design of experiments.
  - Fusion of data and computational experiment.

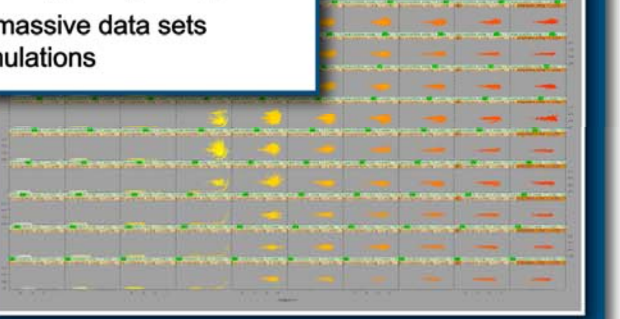
# ORNL Statistics and Data Sciences Group:
## Cutting through the fog of high-dimensional variability



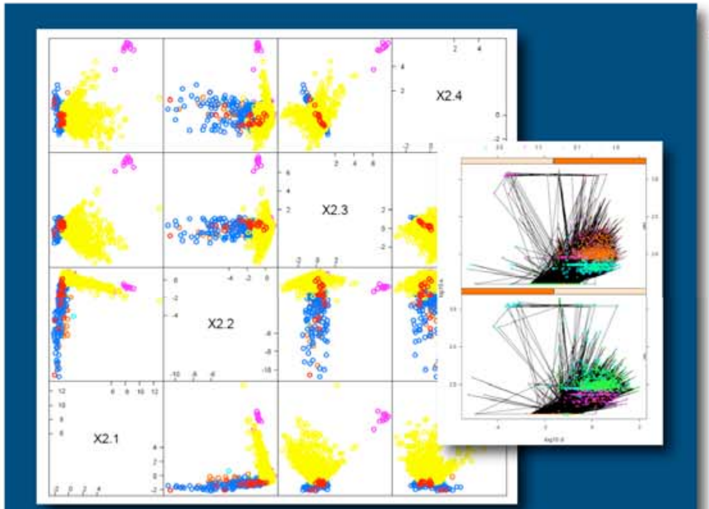
Statistical decomposition of time varying simulation data as applied to fusion simulation data



Data-parallel analysis of massive data sets from climate simulations

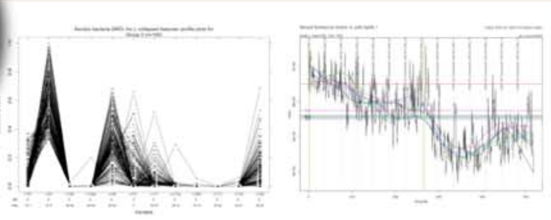


Visualization of high-dimensional relationships



Network intrusion detection without content analysis. High-dimensional density space classification

High-dimensional multivariate classification for chem-bio mass spectrometer



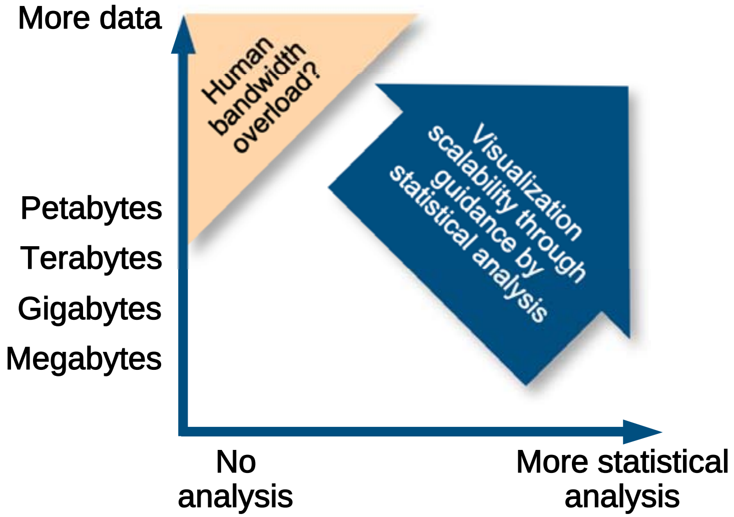
D Wolf: wolfdg@ornl.gov

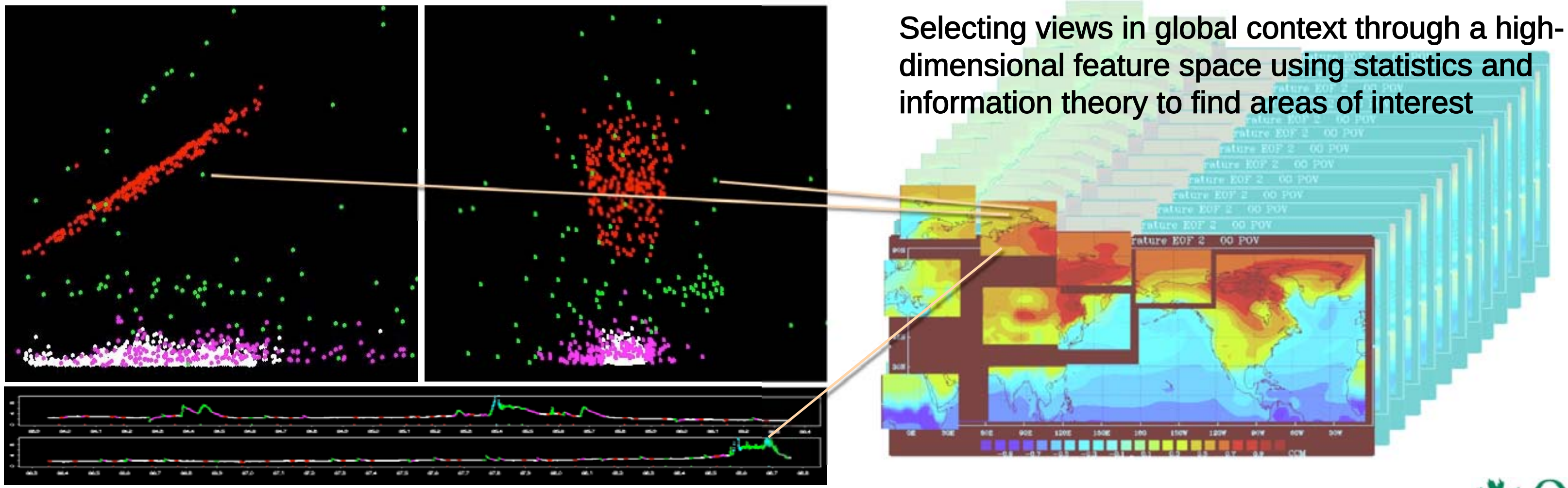
# A statistical framework for guiding visualization of massive data sets

Browse a petabyte? Not humanly possible!

To see 10% of a petabyte at 100 MB per second takes 35 work days!

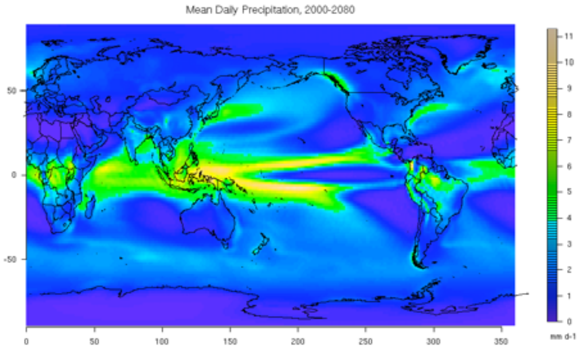
Statistical analysis must select views or reduce to quantities of interest in addition to fast rendering of data.

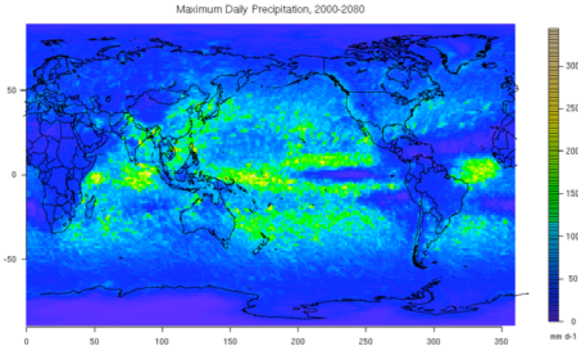

More data
Petabytes
Terabytes
Gigabytes
Megabytes
No analysis
More statistical analysis
Human bandwidth overload?
Visualization scalability through guidance by statistical analysis



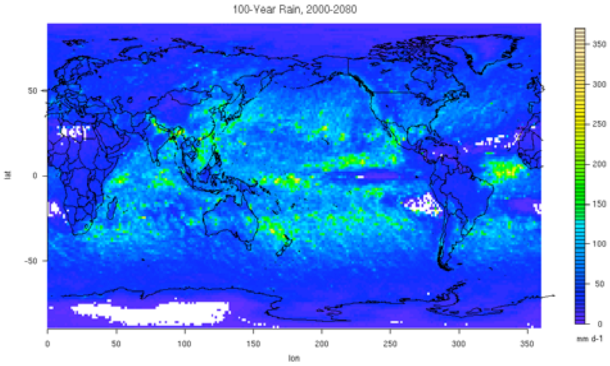
Selecting views in global context through a high-dimensional feature space using statistics and information theory to find areas of interest

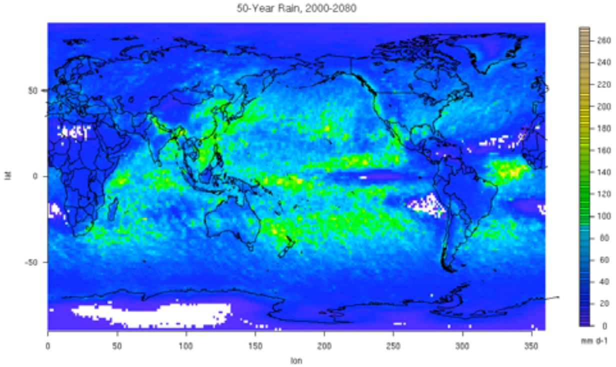
# Interactive climate analysis with data-parallel R


Mean Daily Precipitation, 2000-2080

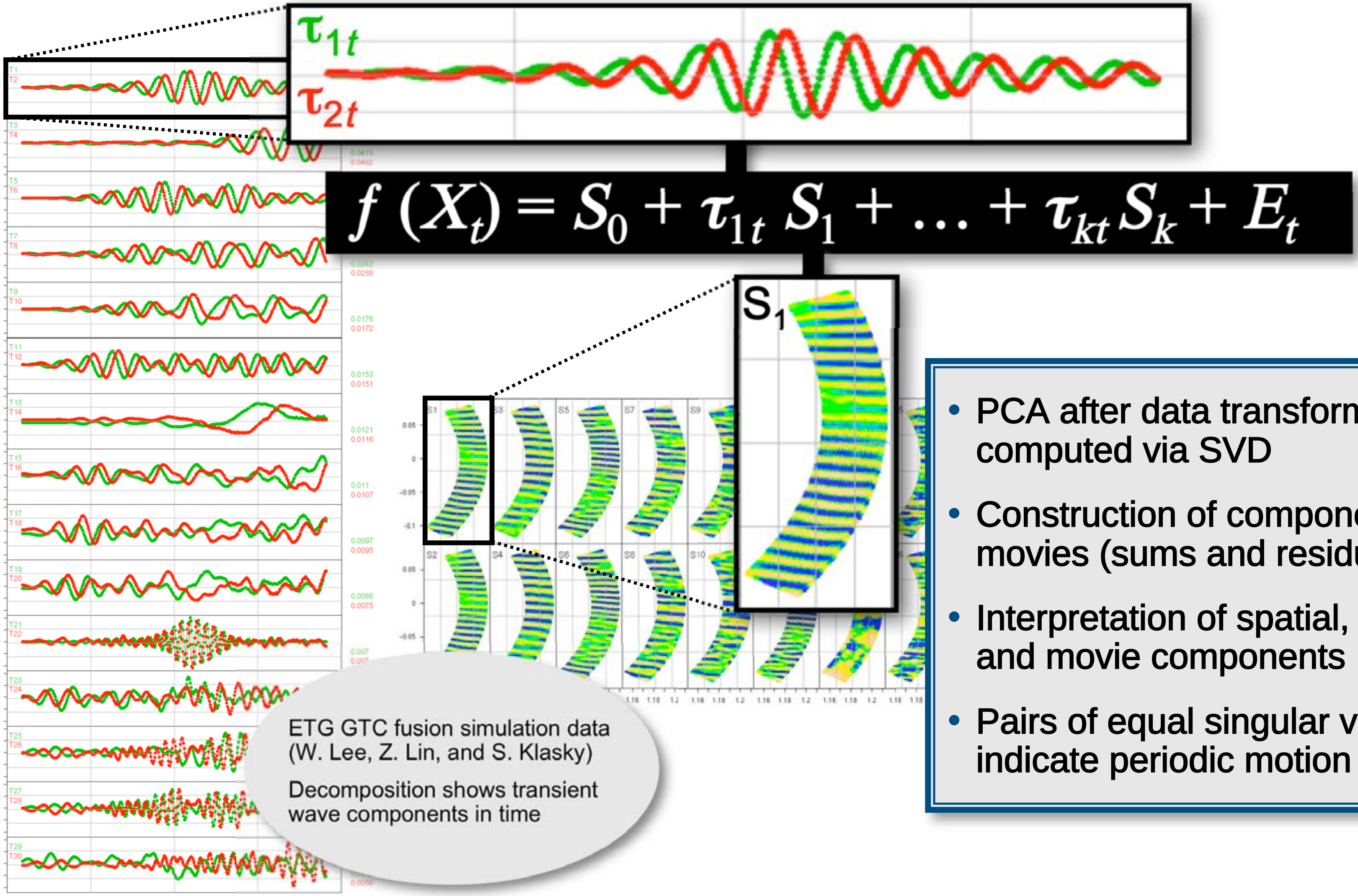

Maximum Daily Precipitation, 2000-2080

- Data-parallel R interactive runtime environment:
  - NetCDF data-parallel readers then R/Rmpi operations on distributed data
  - Extremely broad range of analysis methods:
    - So far, binning, subsetting, univariate statistics, regression methods, and extreme value methods tested
    - Other analyses being tested
- CCSM3 daily precipitation data 2000–2080, 10 GB (climate science M. Branstetter)
  - Comments counterclockwise from top:
    - Where it rains and where it doesn’t rain
    - Where the biggest daily rain event extremes happen
    - Generalized extreme value (GEV) prediction of 50-year rain event (insufficient data in dry areas)
    - GEV prediction of 100-year rain event (insufficient data in dry areas)


100-Year Rain, 2000-2080


50-Year Rain, 2000-2080

# Statistical decomposition of time-varying simulation data



$$f\,(X_t) = S_0 + \tau_{1t}\, S_1 + \ldots + \tau_{kt}\, S_k + E_t$$

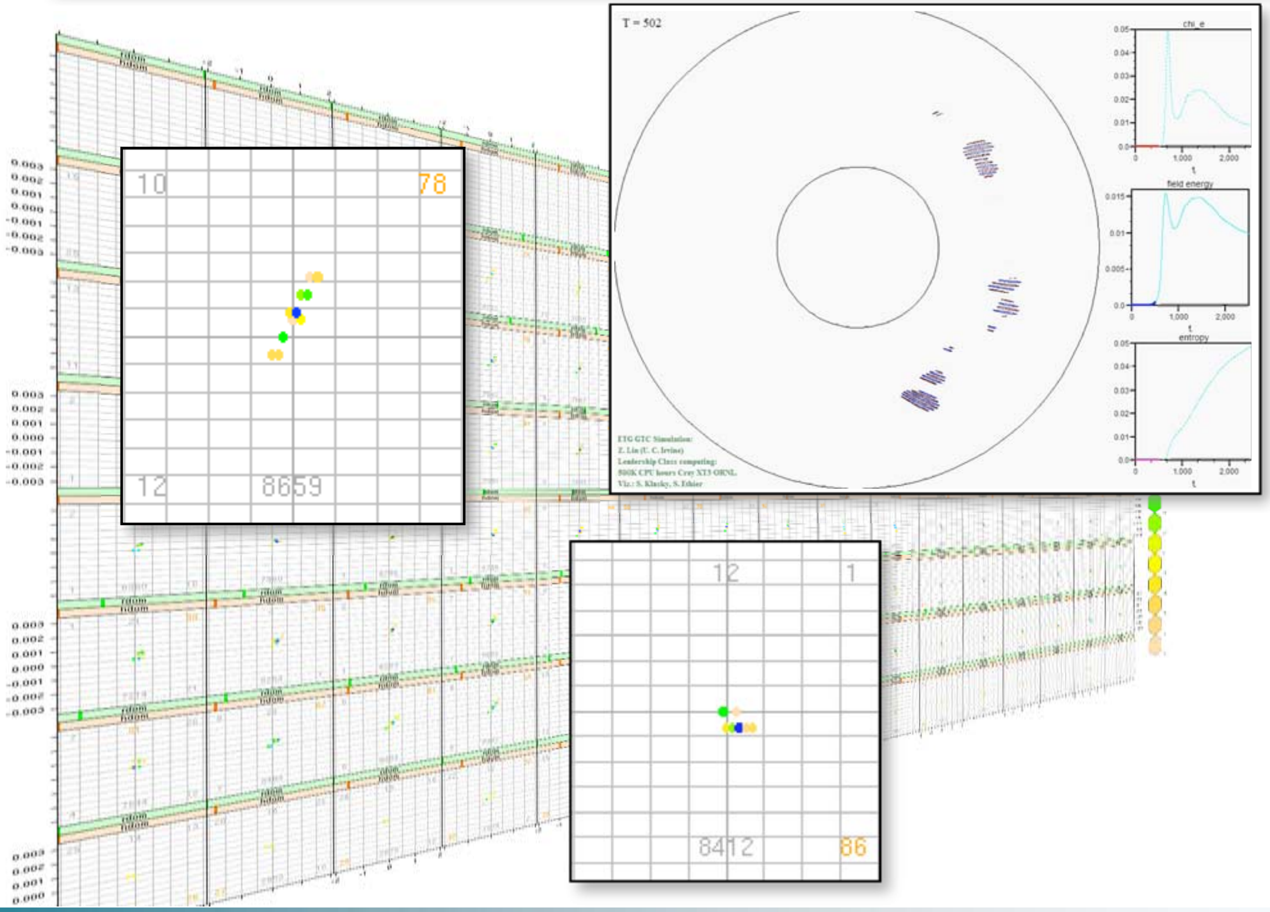
ETG GTC fusion simulation data (W. Lee, Z. Lin, and S. Klasky)

Decomposition shows transient wave components in time

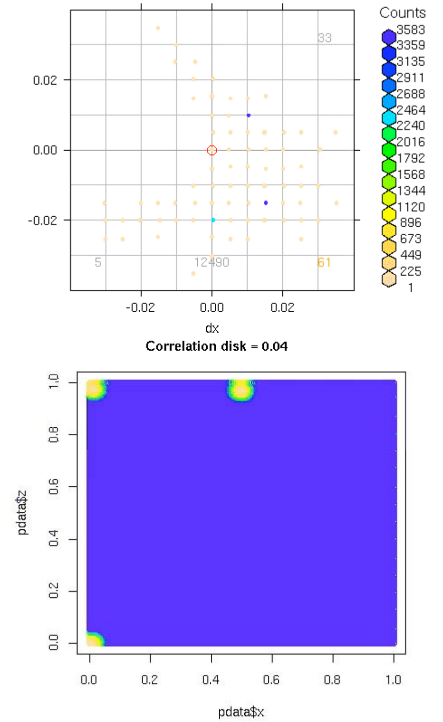
- PCA after data transformations, computed via SVD
- Construction of component movies (sums and residuals)
- Interpretation of spatial, time, and movie components
- Pairs of equal singular values indicate periodic motion

# Local feature motion density analysis

- Computes statistical density of all motion within a scene. (Computation and graphics written in R)
- An array of small multiples displaying 128 local densities over a Tokamak slice is best viewed on an ultra-high-resolution device like ORNL's EVEREST facility
- Developed for GTC Electronic Temperature Gradient microturbulence data (Zhihong Lin) to explore wave motion directions



Calibration on three features with constant but different velocities



# Conditional small multiples provide access to high-dimensional relationships in climate

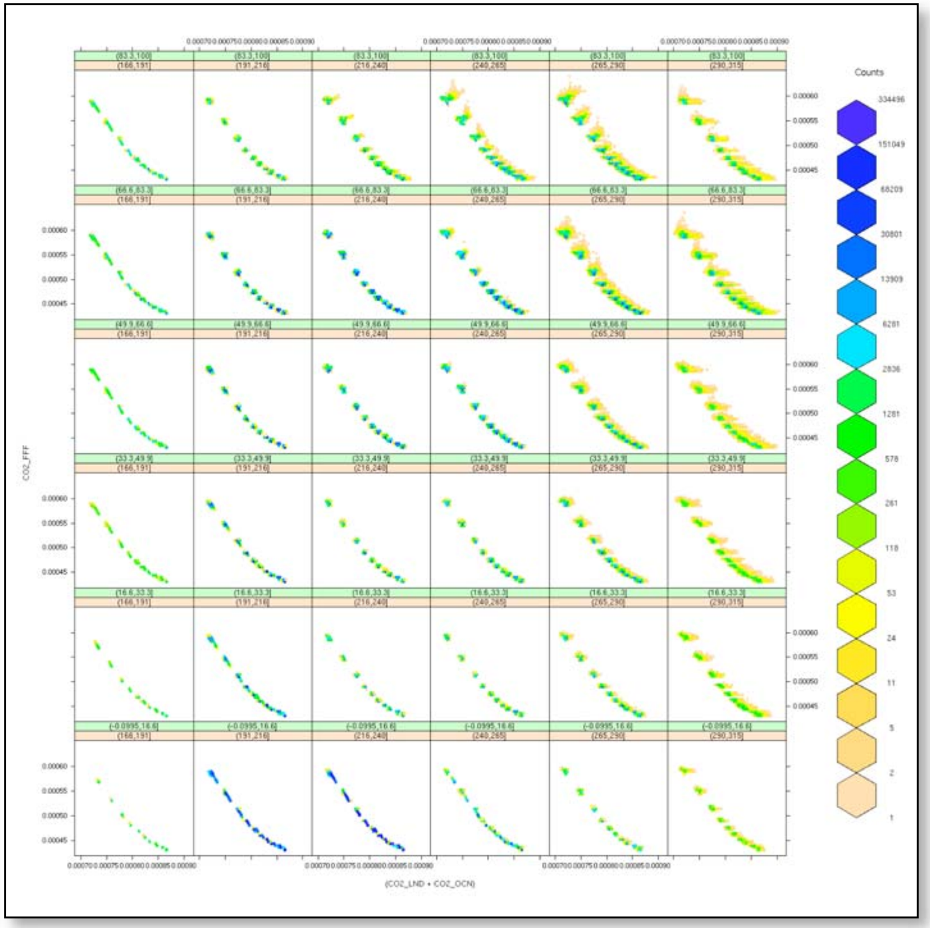
Statistical density of world atmospheric $CO_2$ over 5 dimensions:

→x: $CO_2$ from land & ocean sources
↑ y: $CO_2$ from fossil fuels
►across row: Cloud fraction
▲ up column: Humidity percent
Superposition: 1900, 1910, …, 1990, clusters

Color: Density (count of grids) conditional onvalues of 5 above variables on log scale

Produced interactively with hexagonal binning in R Environment for Statistical Computing and Graphics (Data source: CCSM on Cray X1E at ORNL/NCCS)

Ideal for high resolution displays like ORNL's EVEREST facility



# Contact

**George Ostrouchov**

Statistics and Data Sciences
Computer Science and Mathematics Division
(865) 574-3137
ostrouchovg@ornl.gov