An introduction to VAPOR: A desktop environment for exploration of Earth & Space sciences CFD data

John Clyne, Dan Lagreca, Alan Norton
National Center for Atmospheric Research
Boulder, CO USA

CScADS Scientific Data and Analytics for Petascale Computing Workshop
July 26-29, 2010

This work is funded in part through U.S. National Science Foundation grants 03-25934 and 09-06379, and through a TeraGrid GIG award.

VAPOR Project Goals

- Improve scientific productivity by facilitating interactive analysis and exploration of the largest numerical simulation outputs without the need for Herculean interactive computing resources

And…

- Change role of Advanced 3D Visualization in sciences

From: a scientific finale

- Pictures for publication and presentation
- Performed by visualization experts

To: an integral part of the scientific discovery process

- Visual data analysis aiding investigation
- Performed by scientists

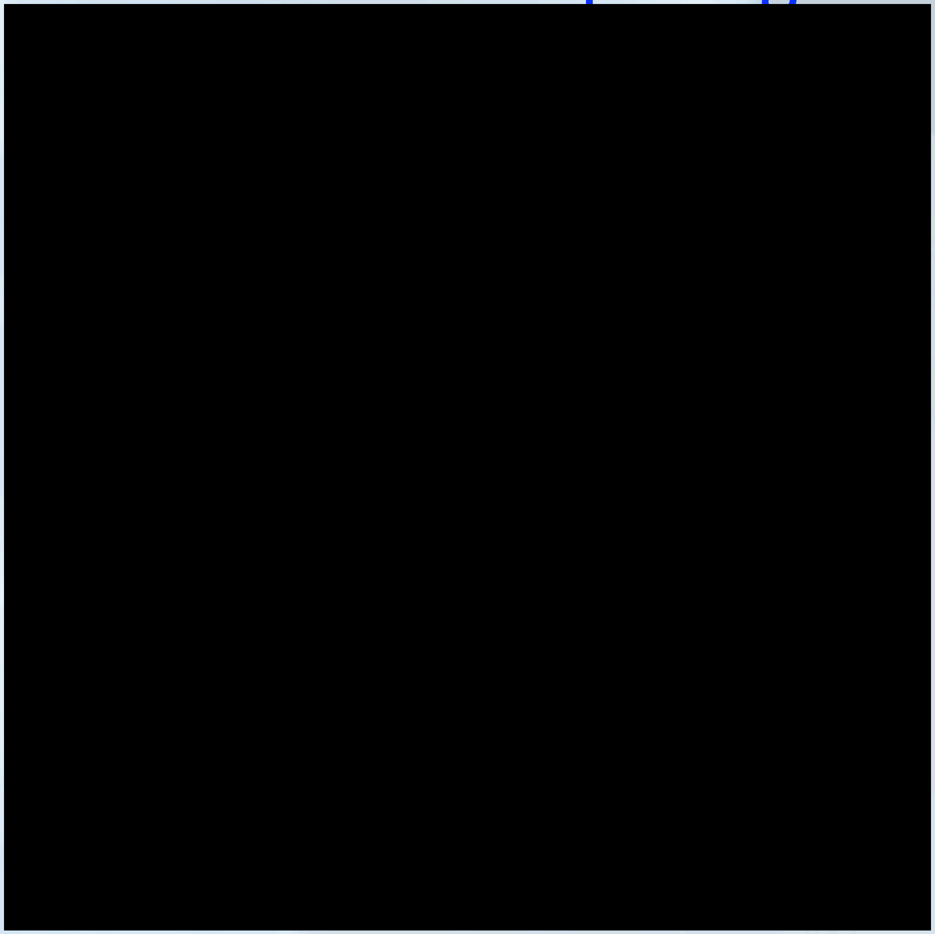
Outline

- Problem motivation
- VAPOR overview – what makes VAPOR unique?
  - Data model
  - Earth and space sciences focus
  - Analysis capabilities
- Laptop demonstration
- Future directions

Solar thermal starting plume
Computed at the dawn of *terascale* computing

- 2003 - Simulation
  - 6 months run time
  - 504x504x2048 grid
  - 5 variables (u,v,w,rho,temp)
  - ~500 time steps saved
  - 9 TBs storage (4GBs/var/timestep)
  - 112 IBM SP RS/6000 processors
- 2004 - Post-processing
  - 3 months
  - 3 derived variables (vorticity components)
- 2004 - Analysis
  - **Abandoned!!!**

- 2006 - Analysis Resumed
- 2007 - *New Journal Physics* publication



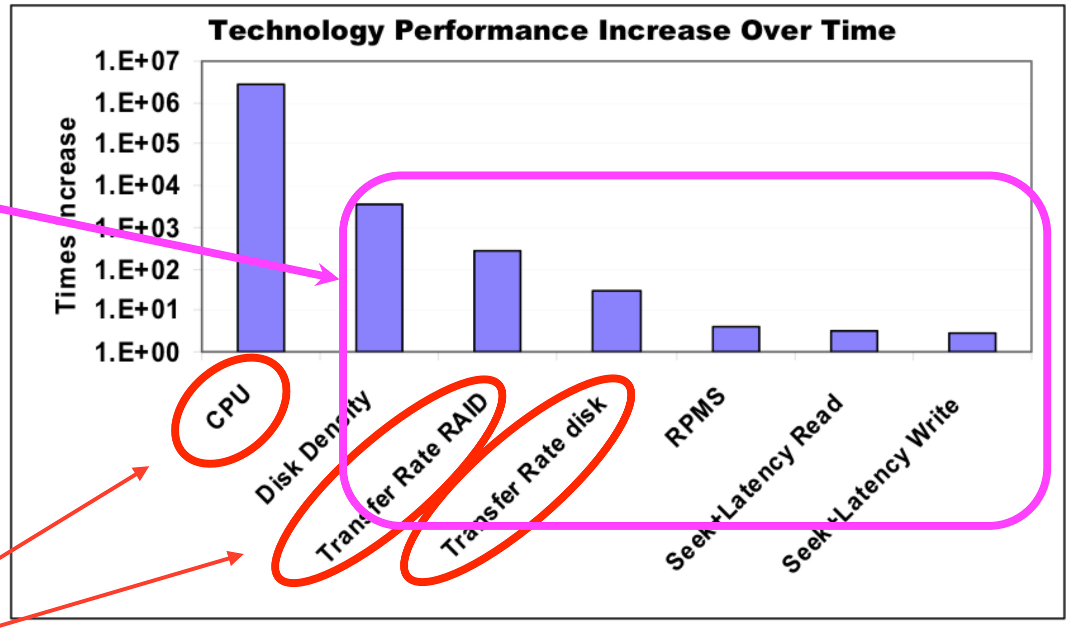
Mark Rast, NCAR/CU, 2003

Computing technology performance increases from 1977 to 2006

Moore's Law does not apply to these!!!

Orders of magnitude difference between improvements in CPU speed and IO bandwidth

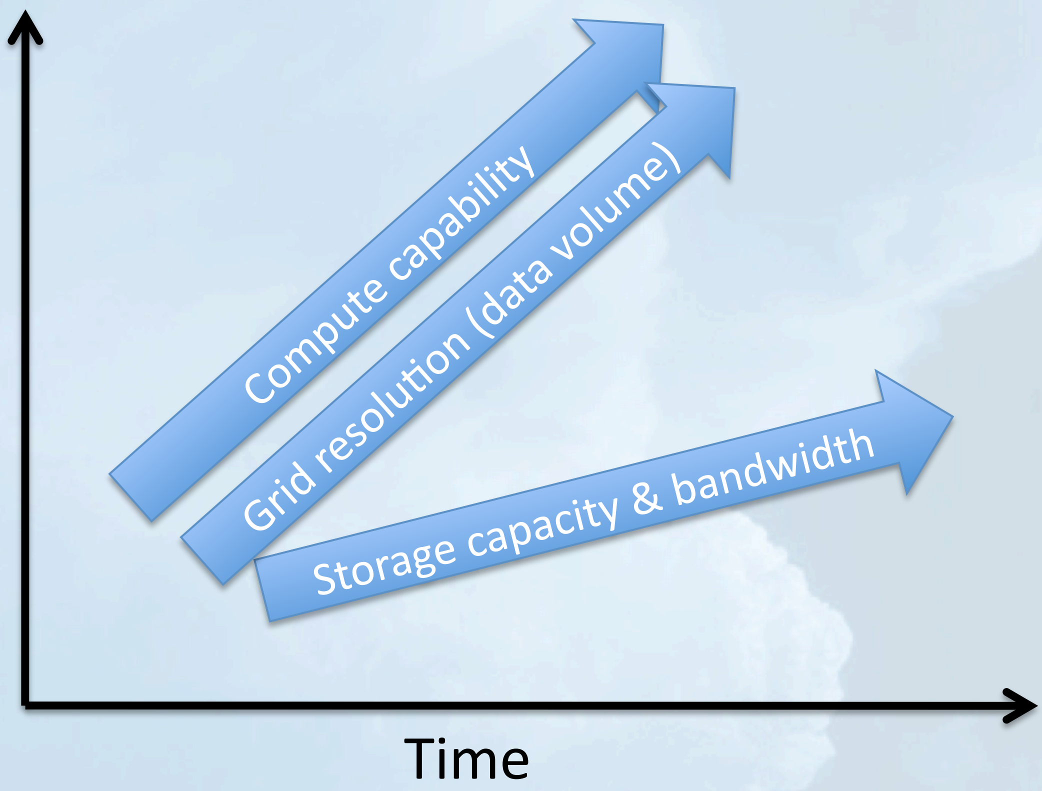
Balance between compute and IO is changing rapidly



Increases in processor speed and disk density have both grown at alarming rates while disk transfer rates have only grown modestly and disk agility has hardly improved at all.

High End Computing Revitalization Task Force (HEC-RTF), Inter Agency Working Group (HEC-IWG) File Systems and I/O Research Workshop 5

What does this mean for data analysis and visualization?



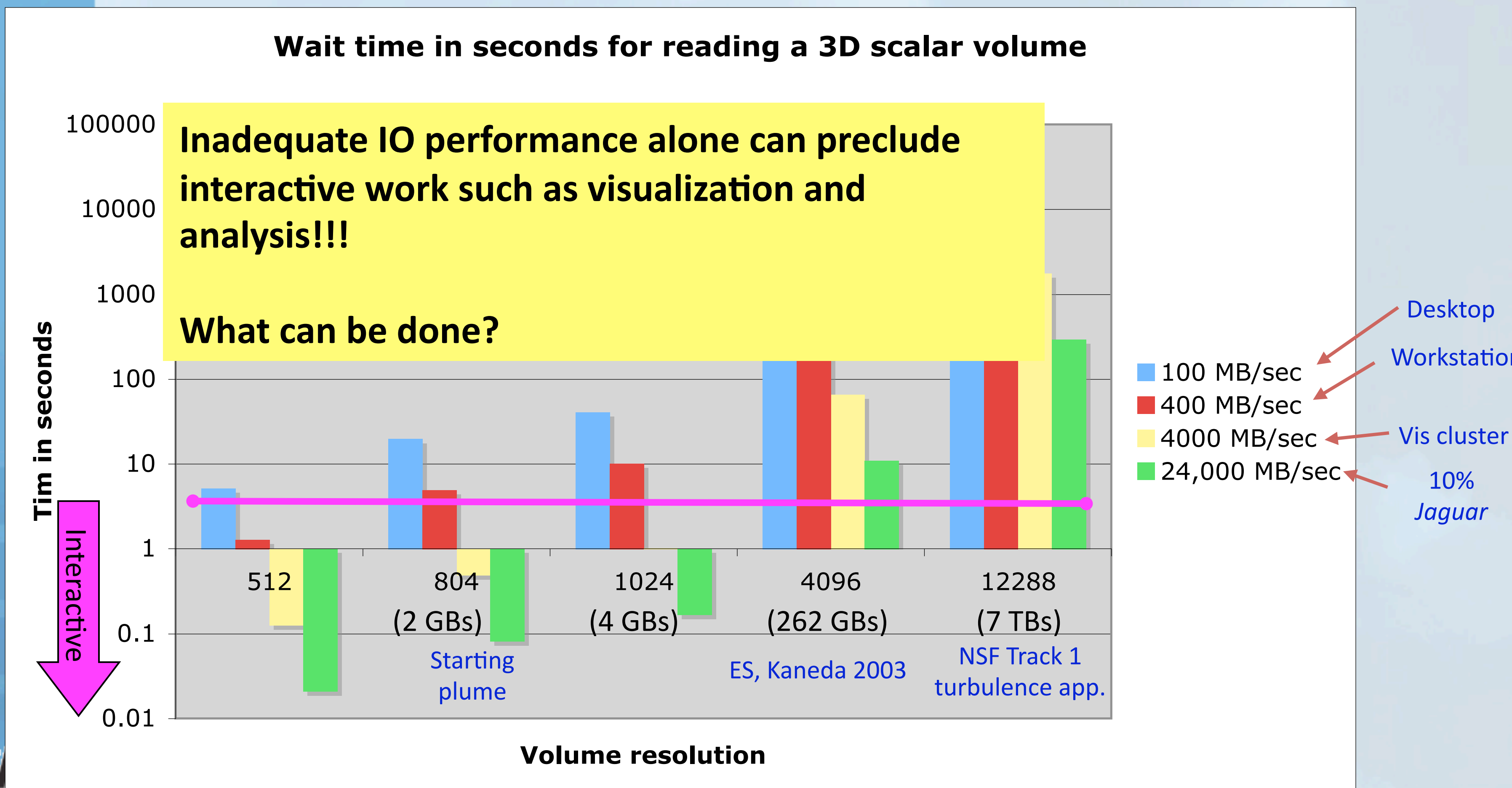
Definition: A system is *interactive* if the time between a user event and the response to that event is short enough maintain my full attention

If the response time is…

- 1-5 seconds : I'm engaged
- 5-60 seconds : I'm tapping my foot
- 1-3 minutes : I'm reading email
- > 3 minutes : I've forgotten why I asked the question!

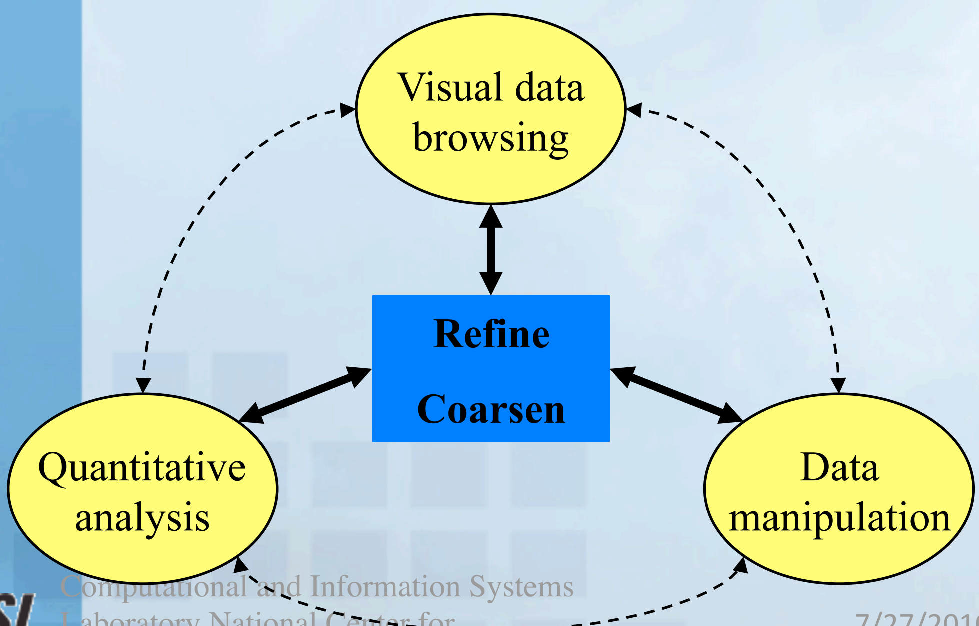
What is meant by *interactive* analysis?

Mark Rast, University of Colorado, 2005



VAPOR Key Components

1. **Domain specific**: earth and space sciences CFD
2. **Data analysis**: qualitative and quantitative data interrogation and manipulation capabilities
3. **Terabytes from the desktop**: operates on terascale sized simulations with only desktop computing
   - Multiresolution data representation to permit speed/quality tradeoffs
   - Region of Interest (ROI) identification and isolation

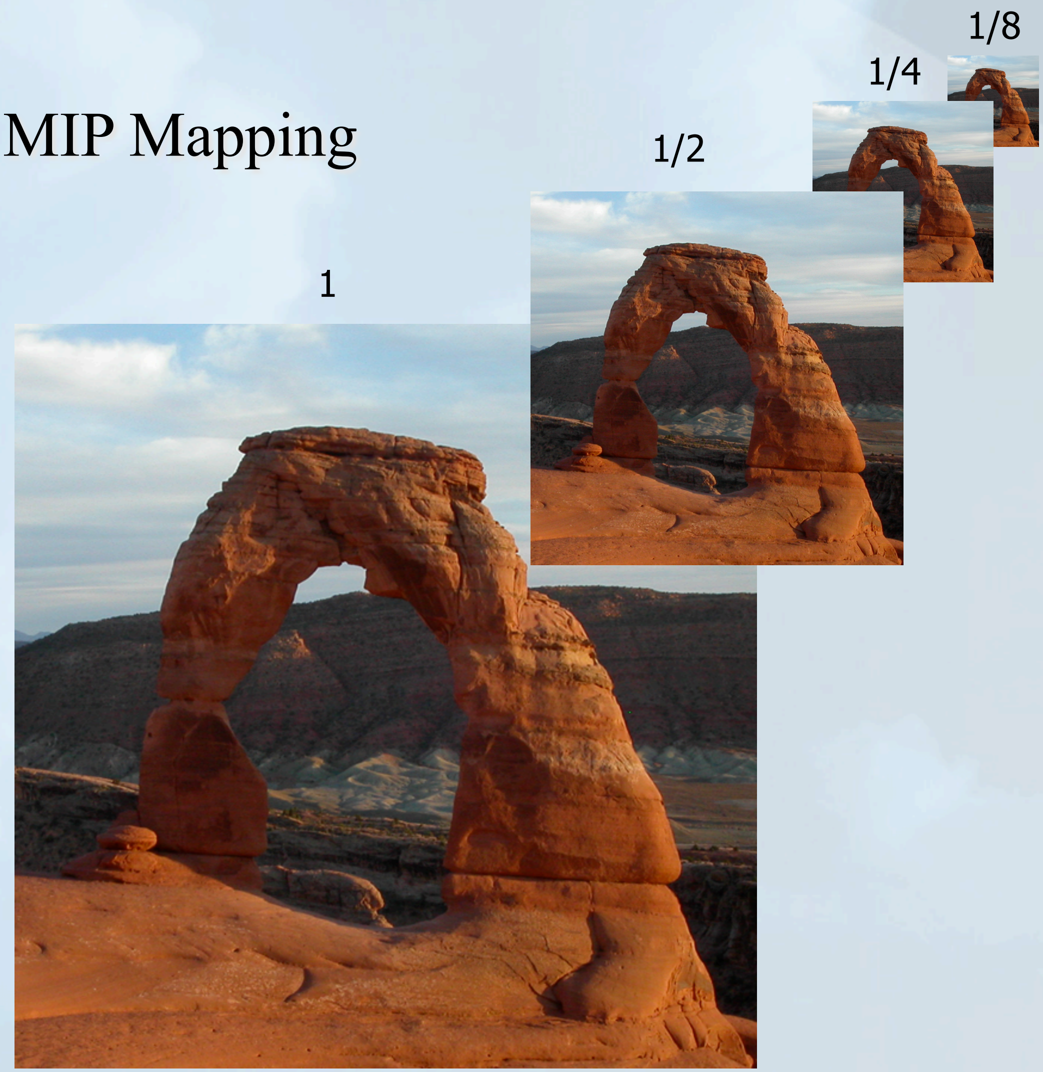


Combination of visualization, ROI isolation, and multiresolution data representation that provides sufficient **data reduction** to enable interactive work

Think *Google Earth*!!!

Key component (1)
Speed/quality tradeoffs with multiresolution data

- 2D Example: Texture MIP Mapping



Multiple copies of data at varying power of two resolutions

Storage costs :

$$\sum_{l=0}^{L} 1/2^{dl} = 1 + 1/2^{d} + 1/2^{2d} + 1/2^{3d} \ldots$$

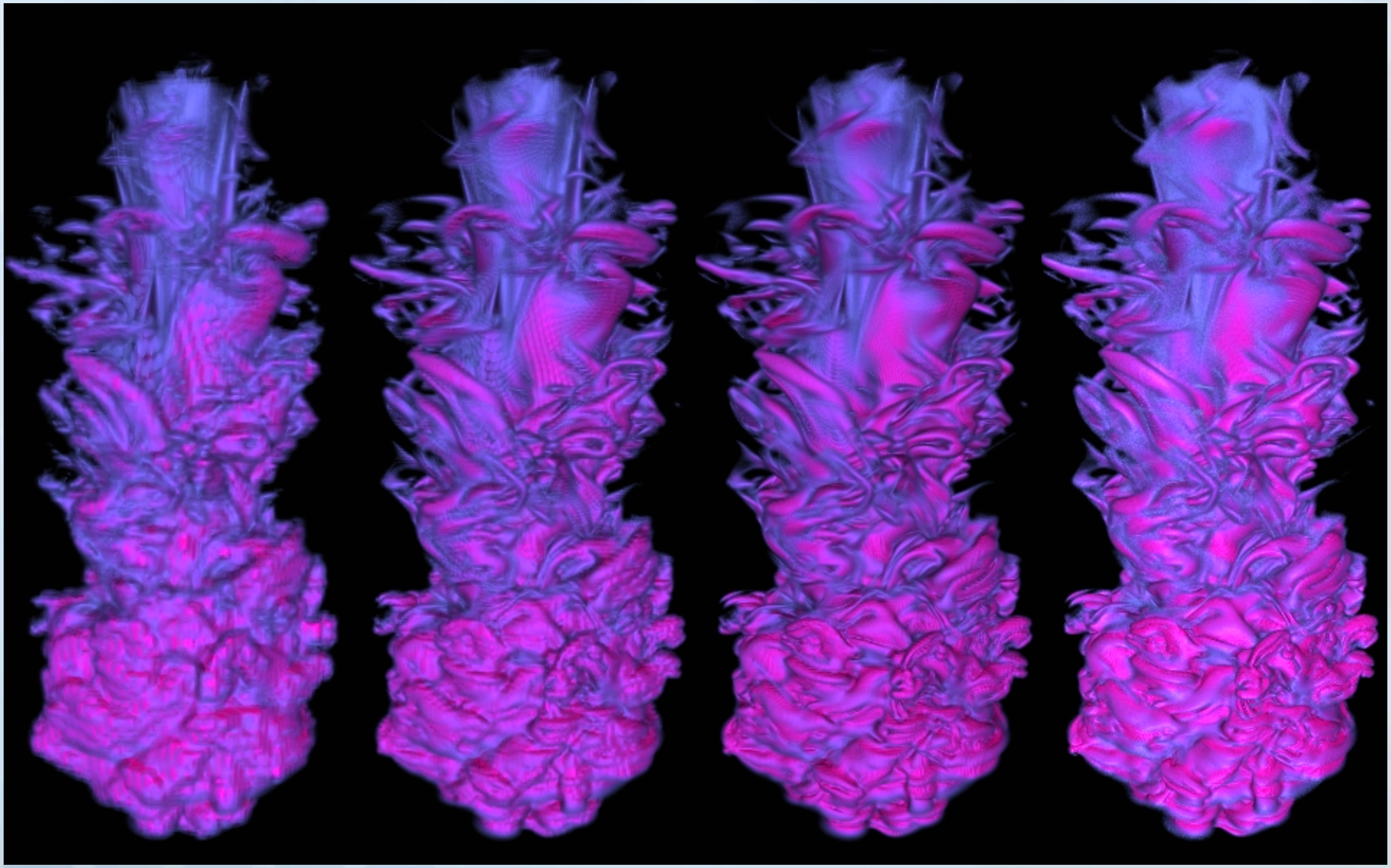
Wavelet transforms for 3D multiresolution data representation

NCAR

- Some wavelet properties:
  - Permit hierarchical data representation
  - Invertible and lossless (subject to floating point round off errors)
  - Numerically efficient (O(n))
    - forward and inverse transform
  - No additional storage cost

Solar thermal plume at varying resolutions [M. Rast, 2006]

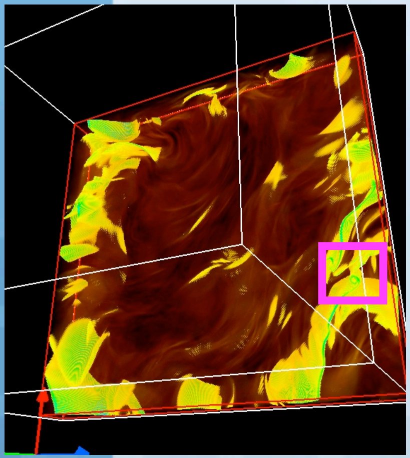
What have we lost???

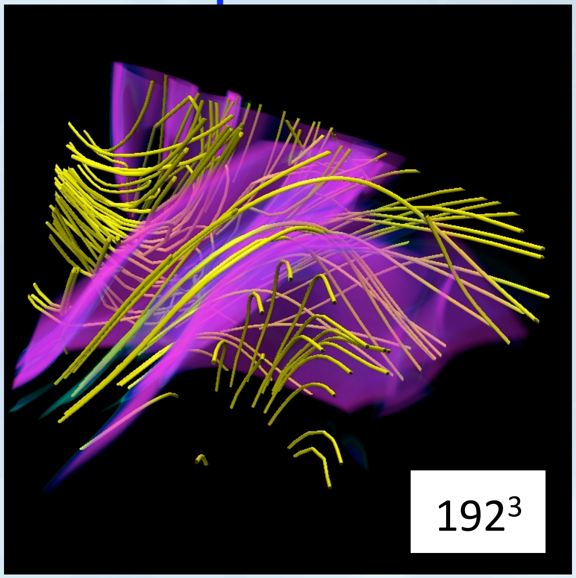


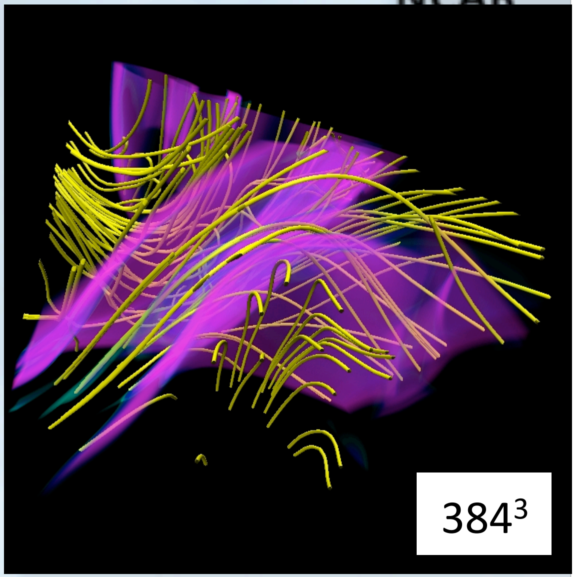
63^2x256 | 126^2x512 | 252^2x1024 | 504^2x2048 (native)

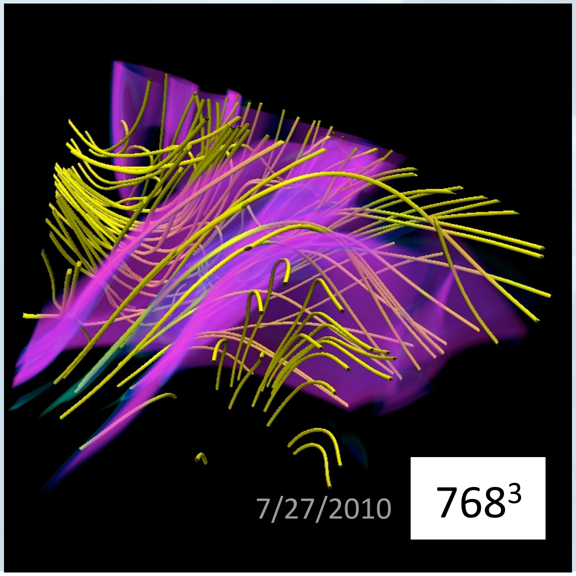
Magnetic field line integration resolution comparison

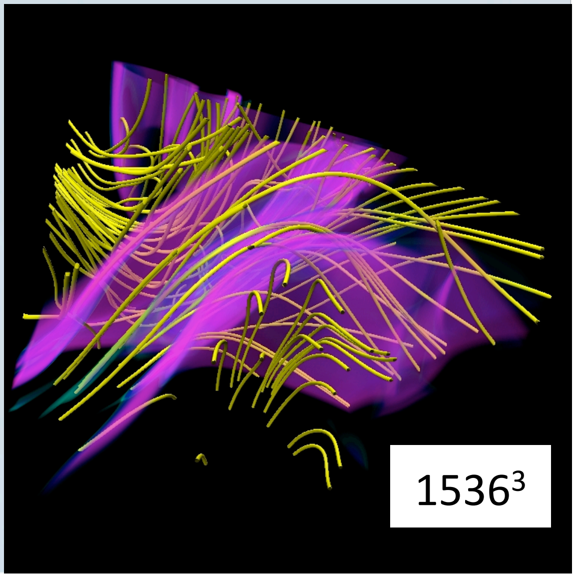
- 1536^3 MHD Simulation
- 4th order Runge-Kutte
- Mininni et al. (2007)









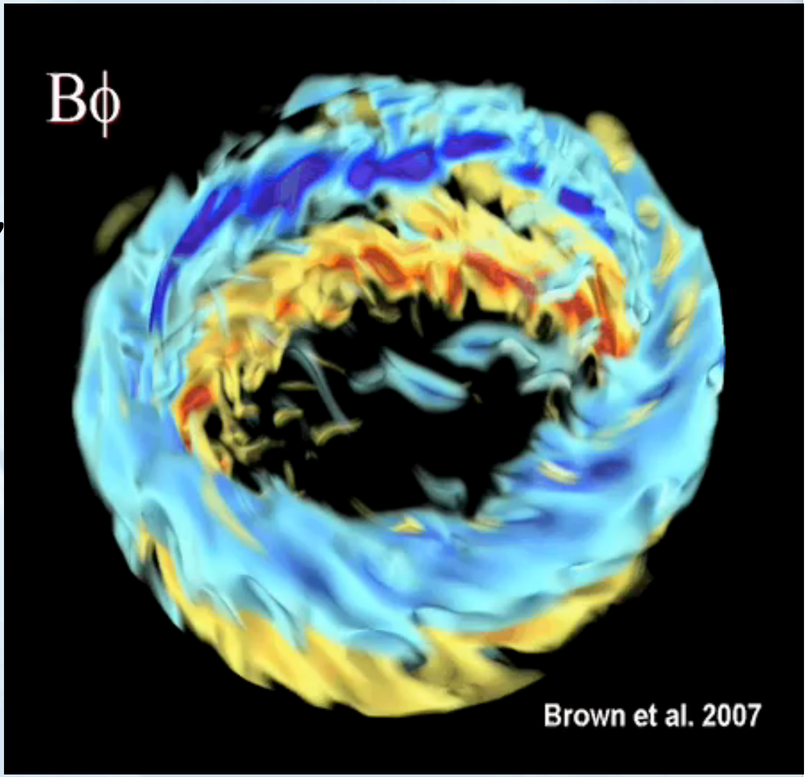


Key component (2)
Earth sciences CFD focus

- Scientific steering committee guides development
- Algorithms
  - General purpose
    - E.g. volume rendering, isosurfaces
  - Specialized for CFD
    - E.g. steady and unsteady flow visualization, field line advection
    - Geo-referenced data
    - Physically based feature tracking
- Data types and grids supported
  - Cartesian, AMR, terrain following, staggered, and spherical (prototype)
  - Temporal data with non-uniform sampling
- Domain specific Graphical User Interface
  1. Features you need are there (hopefully!)
  2. Features you don't need are not there
     - => improved ease-of-use

Algorithms:
Spherical shell data volume rendering

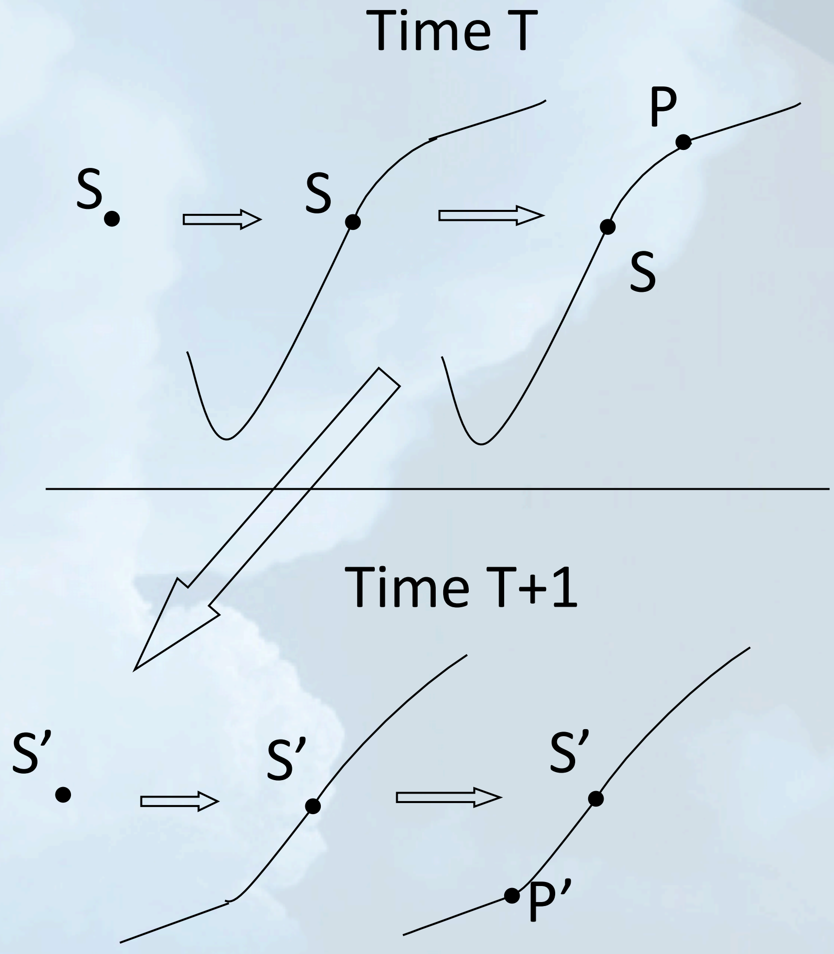
- Simulation of deep convection in convection zones of solar-like stars



- Grid geometry is a spherical shell, covering all latitudes and longitudes and spans a depth of 0.72-0.96 solar radii
- Non-uniform grid spacing in latitude and radial axes
- Image courtesy of Ben Brown, CU

Algorithms: Magnetic field line advection

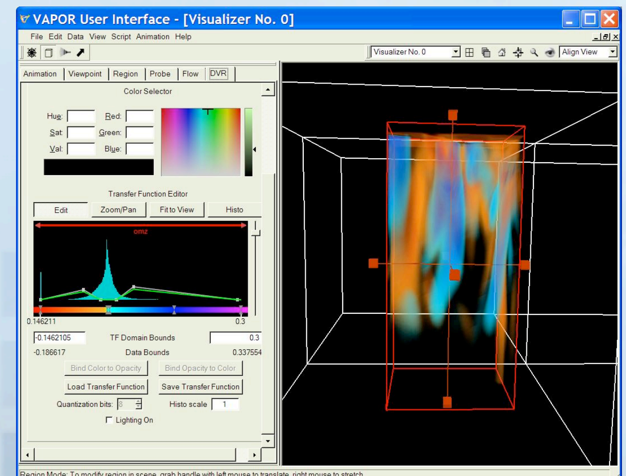
- Combines steady and unsteady flow integration to advect field lines in a time-varying velocity field
  - Algorithm proposed by Aake Nordlund, Neils Bohr Institute and Pablo Mininni, U. Buenos Aires [Mininni et al, 2008]

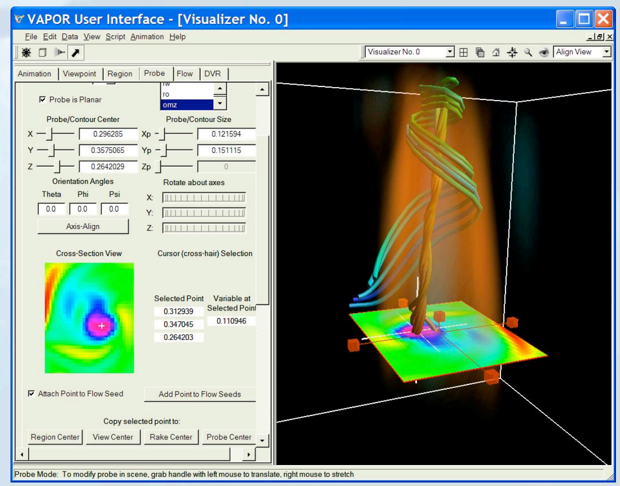


Data courtesy Pablo Mininni

Key component (3)
Data analysis

1. Quantitative information available throughout GUI
   – E.g. histograms, probes, annotation, user coordinates
2. GUI supports *visualization-aided* analysis
3. Coupled with IDL® to calculate and visualize derived quantities in region-of-interest
   – Immediate analysis applied to data identified in visualization
   – Immediate visualization of derived quantities calculated in IDL
      - Identify region of interest
      - Export to IDL session
      - Import result into visualization
4. Coming soon: Integrated Python (numpy/scipy) calculation engine

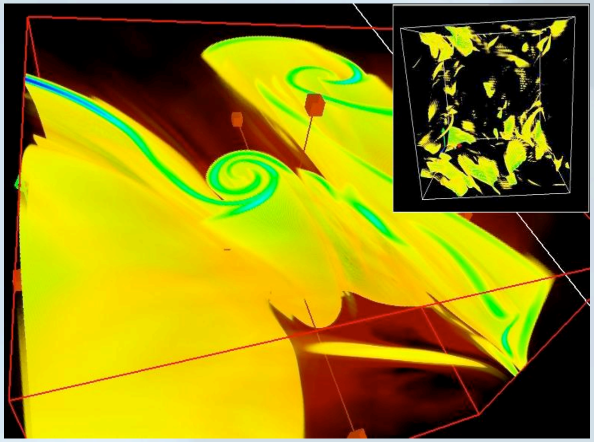




Live demos on a laptop

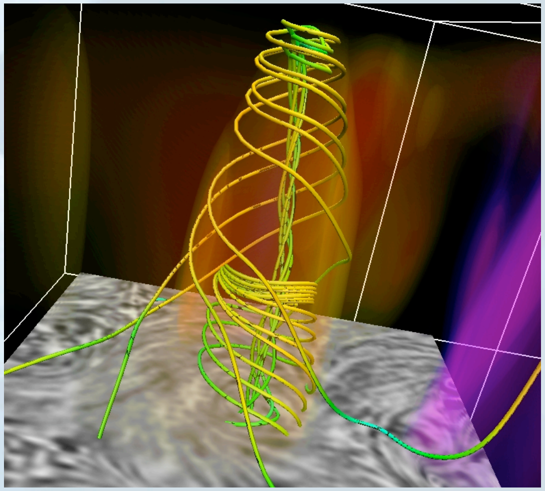
MHD decay [*Mininni et al., PRL **97**, 244503 (2006)*]

- 1536^3, pseudo-spectral method
- 12GBs/variable/time-step
- Exhibits new finding in MHD: current "folding" and "roll-up"



Compressible convection simulation [Rast, et al, 2000]

- 512^2x256
- horizontally periodic, dimensions 6x6x1(deep) constant heat flux into the bottom, constant temperature on top



Future directions

- Broadening scientific end user community
  - E.g. Weather researchers, ocean modelers
    - Geo-referenced 2D & 3D data
    - Nested grids
    - Missing data
- Expanding analysis capabilities
  - Integration with numpy/scipy
- Hierarchical data model
  - VAPOR data importers for VisIt and ParaView
  - Fortran-callable, distributed memory (MPI) API
- Extensible architecture

- Preparing for petascale computing

Wavelet based hierarchical data representation has been shown to enable powerful speed/quality tradeoffs in VAPOR. Data sets up to 2048^3 can effectively be analyzed with modest computing resources. But…

- Power-of-two reductions are limiting
- The current model may not scale to petascale data sets

More aggressive data reduction required for petascale applications

Discrete Wavelet Transforms

- Discrete Fourier transform

$$f(t) = \frac{1}{N}\sum_{n=0}^{N-1} a_t e^{j2\pi nt/N} \qquad (0 \le t \le N-1)$$

- Discrete Wavelet Transform

Scaling term (coarse representation of signal)

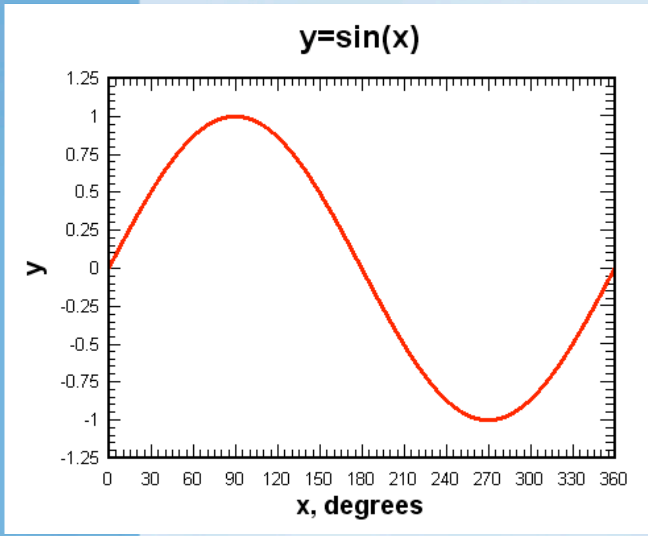
$$f(t) = \sum_k c(k)\phi_k(t) + \sum_k \sum_{j=0}^{\log_2 N} d_j(k)\,\psi_{j,k}(t)$$

Detail term (high frequency components of signal)

$$\phi(t) = \sum_k h_\phi(k)\sqrt{2}\phi(2t-k), \quad k \in Z \quad \text{scaling function}$$

$$\psi(t) = \sum_k h_\psi(k)\sqrt{2}\phi(2t-k), \quad k \in Z \quad \text{wavelet function}$$

  - Properties
    - Multiresolution representation
    - Efficient: Linear time complexity
    - Adaptable: Can represent functions with discontinuities, bounded domains, and arbitrary topology
    - Time frequency localization: Many coefficients are zero or close to zero



Fourier transform basis function: sine, cosine

A very small sampling of wavelet transform basis functions

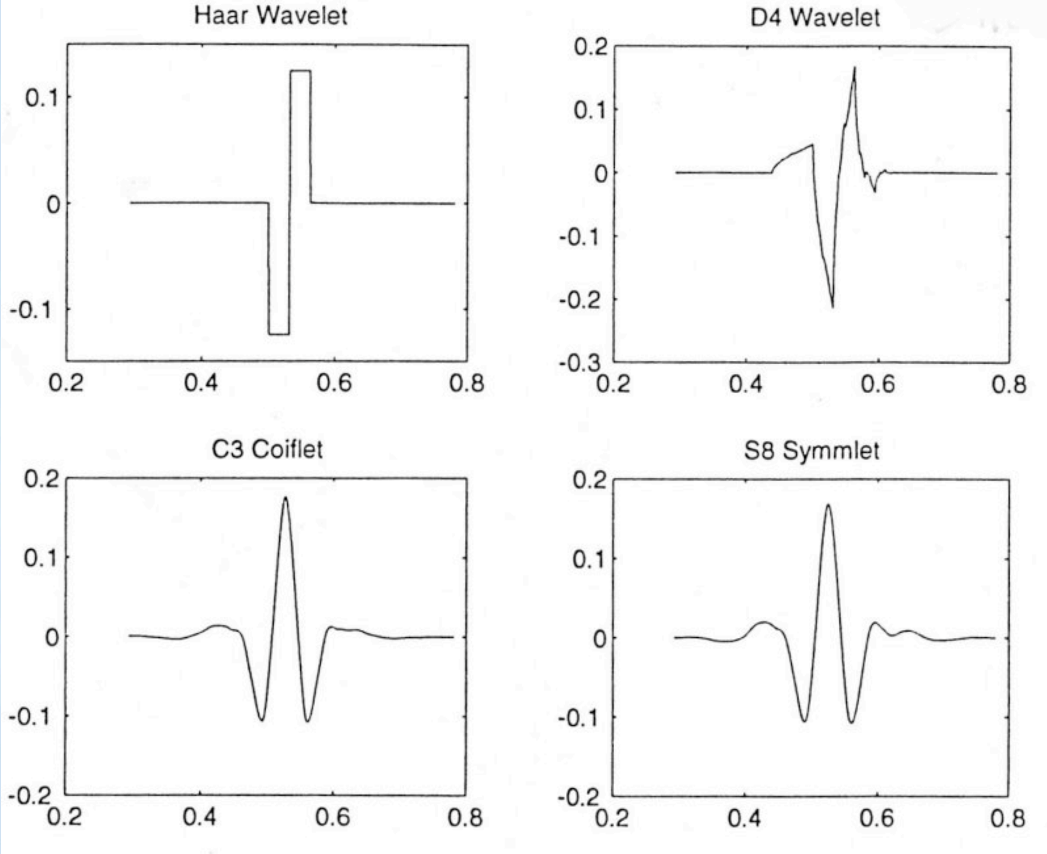


Image credit: K.H. Parker

Many wavelet families and parameterizations within each family to choose from. Best choice is often far from obvious.

Wavelet compression and progressive access (1)

Frequency truncation method

- Truncate "*j*" parameter of expansion:

$$f(t) = \sum_k c(k)\phi_k(t) + \sum_k \boxed{\sum_{j=0}^{\log_2 N} d_j(k)\psi_{j,k}(t)}$$

- Provides coarsened approximation at power-of-two increments
- Good
  - Simple
  - Fast
  - **Maintains structure of original grid**
- Bad:
  - Limited to power-of-two reductions
  - Compression quality

Wavelet compression and progressive access (2)

Coefficient prioritization method

- Goal: prioritize coefficients used in linear expansion

$$f(t) = \sum_{n=0}^{N-1} a_n u(t), \quad \text{original } f(t) \qquad \hat{f}(t) = \sum_{m=0}^{M-1} a_m u(t), \quad (M < N), \quad \text{compressed } f(t)$$

L^2 error given by: $L^2 = \left\| f(t) - \hat{f}(t) \right\|_2^2$

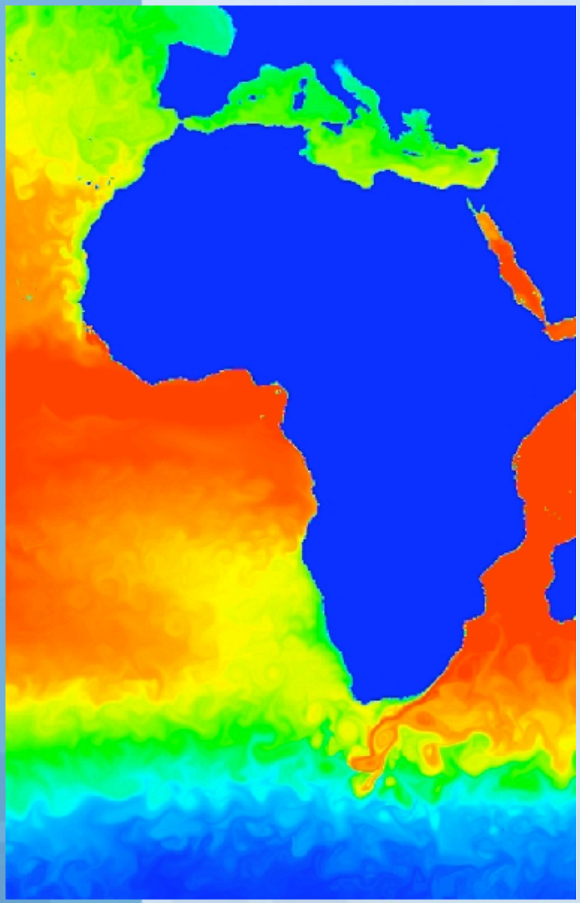
If *u(t)* (*ϕ(t)* and *ψ(t)* in case of wavelet expansion functions) are *orthonormal*, then

$$\text{orthonormal: } \langle u_k(t), u_l(t) \rangle = \int u_k(t) u_l(t)\, dt = \begin{cases} 0, & k \neq l \\ 1, & k = l \end{cases}$$

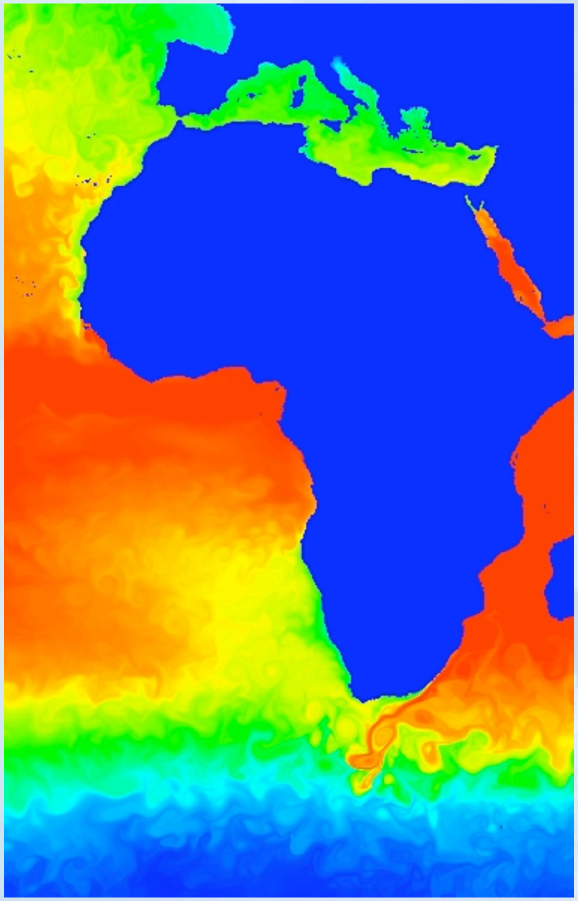
$$L^2 = \sum_{i=M}^{N-1} \left(a_{\pi(i)} \right)^2 = \left\| f(t) - \hat{f}(t) \right\|_2^2, \text{ where } a_{\pi(i)} \text{ are discarded coefficients}$$

- The error is the sum of the squares of the coefficients we leave out!
- So to minimize the L^2 error, we simply discard (or delay transfer) the smallest coefficients!
- If discarded coefficients are zero, there is no information loss!

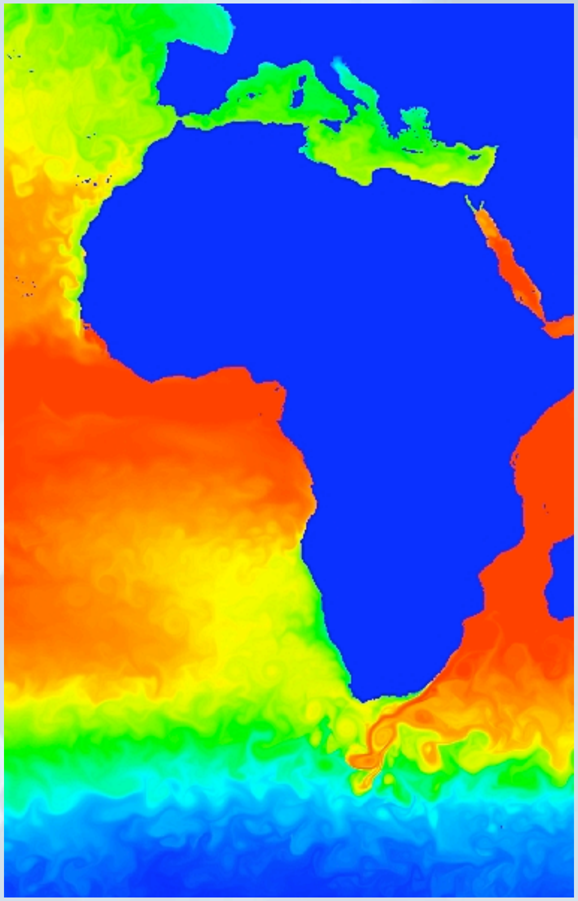
8:1 Compression
Global POP 1/10 degree ocean model [F. Bryan, 2006]



Frequency truncation



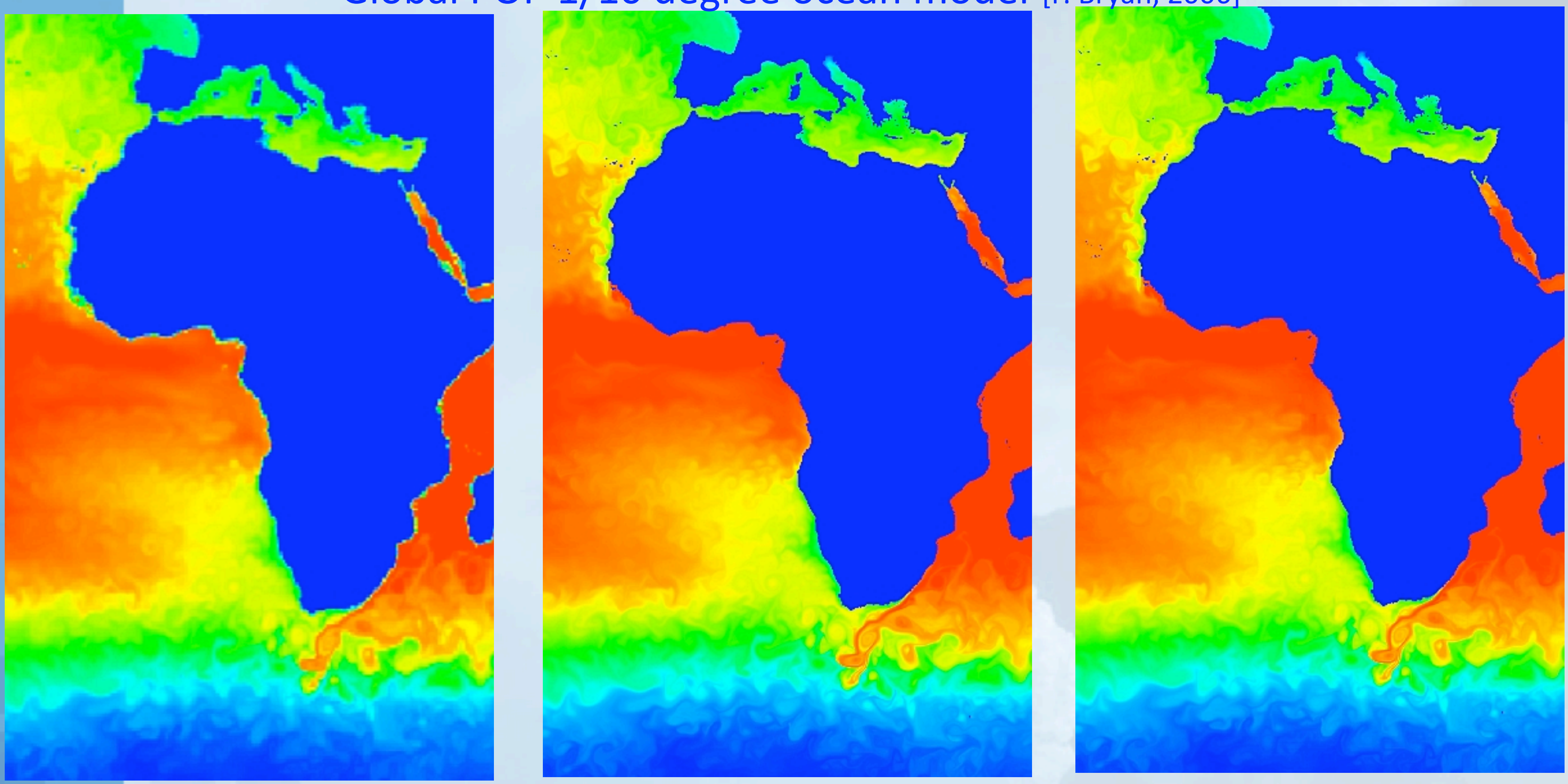
No compression



Coefficient prioritization

64:1 Compression

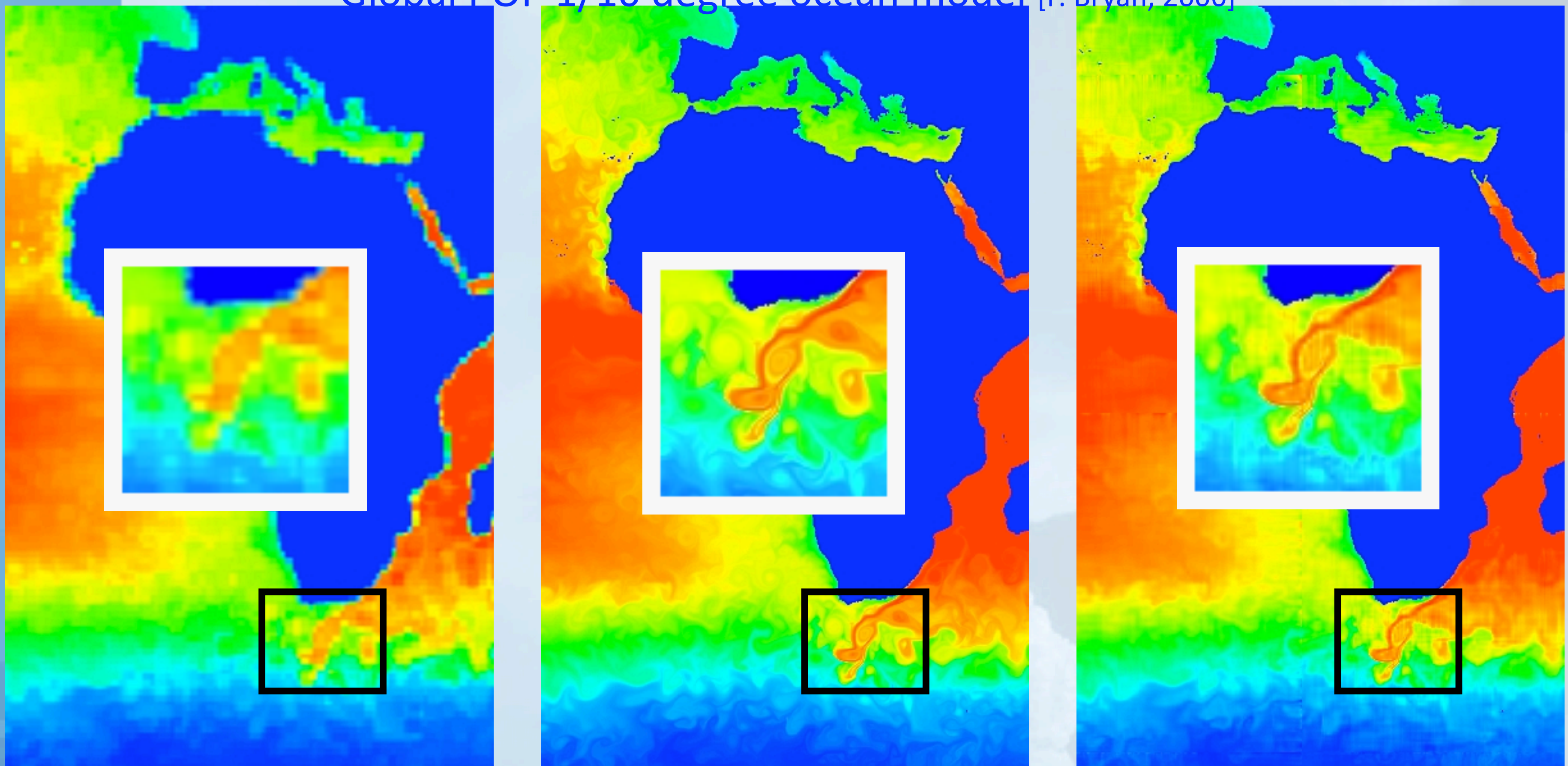
Global POP 1/10 degree ocean model [F. Bryan, 2006]



Frequency truncation | No compression | Coefficient prioritization

512:1 Compression

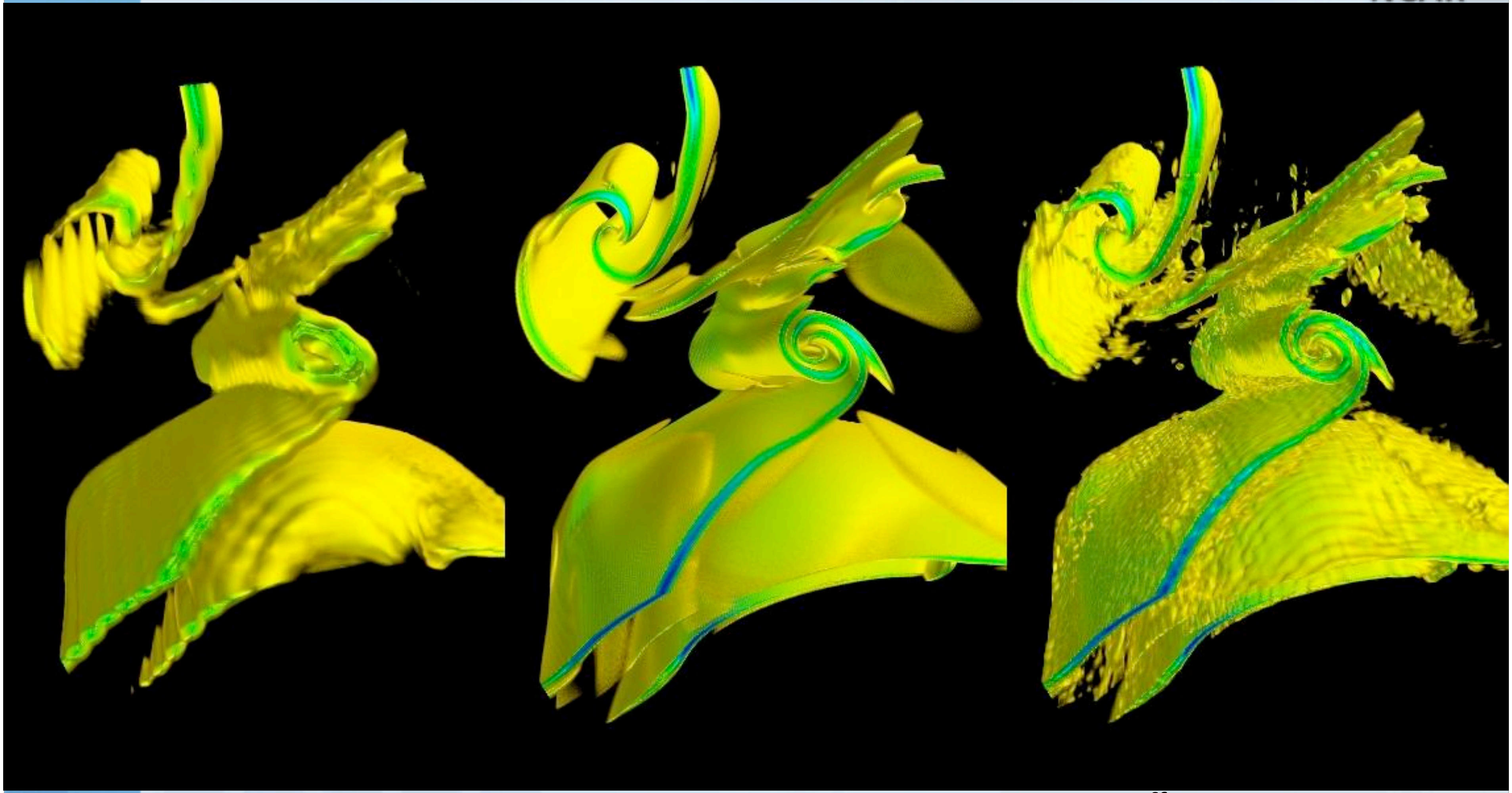
Global POP 1/10 degree ocean model [F. Bryan, 2006]



Frequency truncation | No compression | Coefficient prioritization

512:1 Compression

1536^3 MHD Decay Simulation [P. Mininni et al, (2006)]



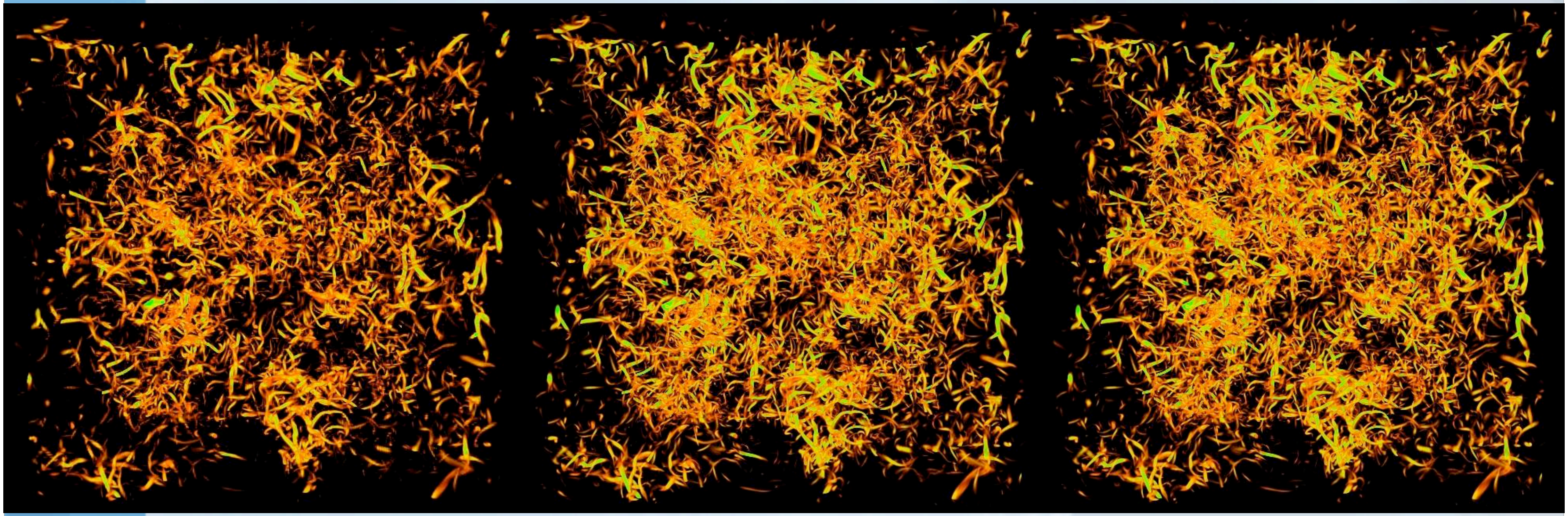
Frequency truncation

No compression

Coefficient prioritization

8:1 Compression

1024^3 Taylor-Green turbulence (enstrophy field) [P. Mininni, 2006]



Frequency truncation

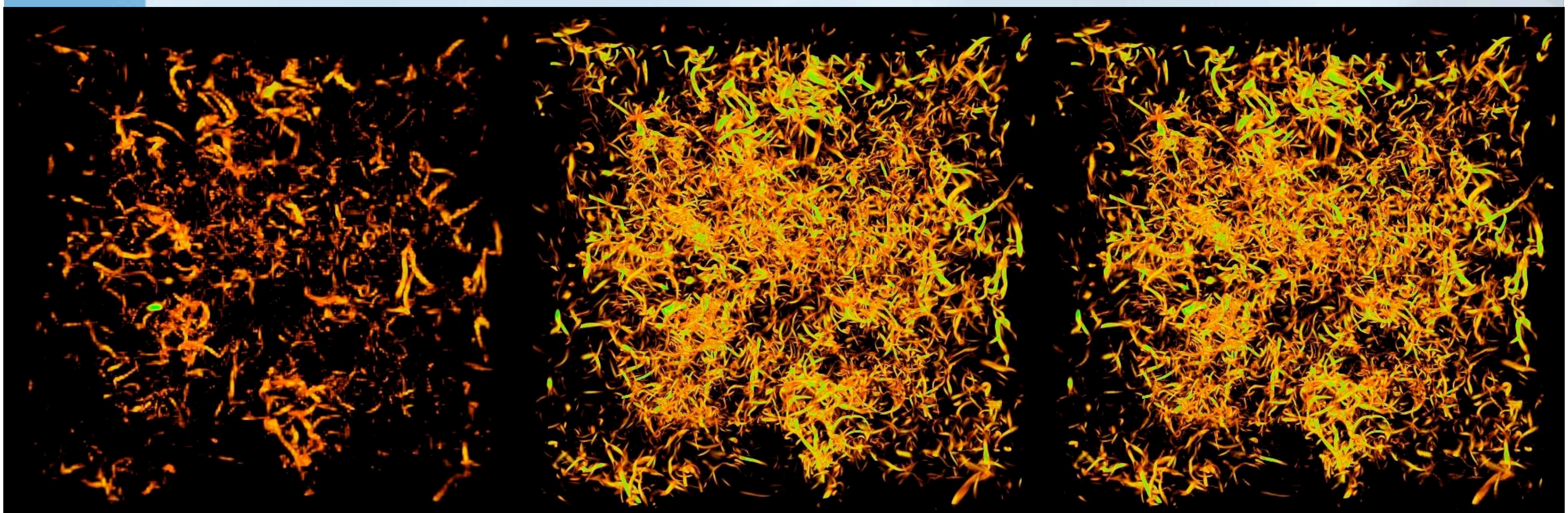
No compression

Coefficient prioritization

Coiflet-12 wavelet
No blocking

64:1 Compression

1024^3 Taylor-Green turbulence (enstrophy field) [P. Mininni, 2006]

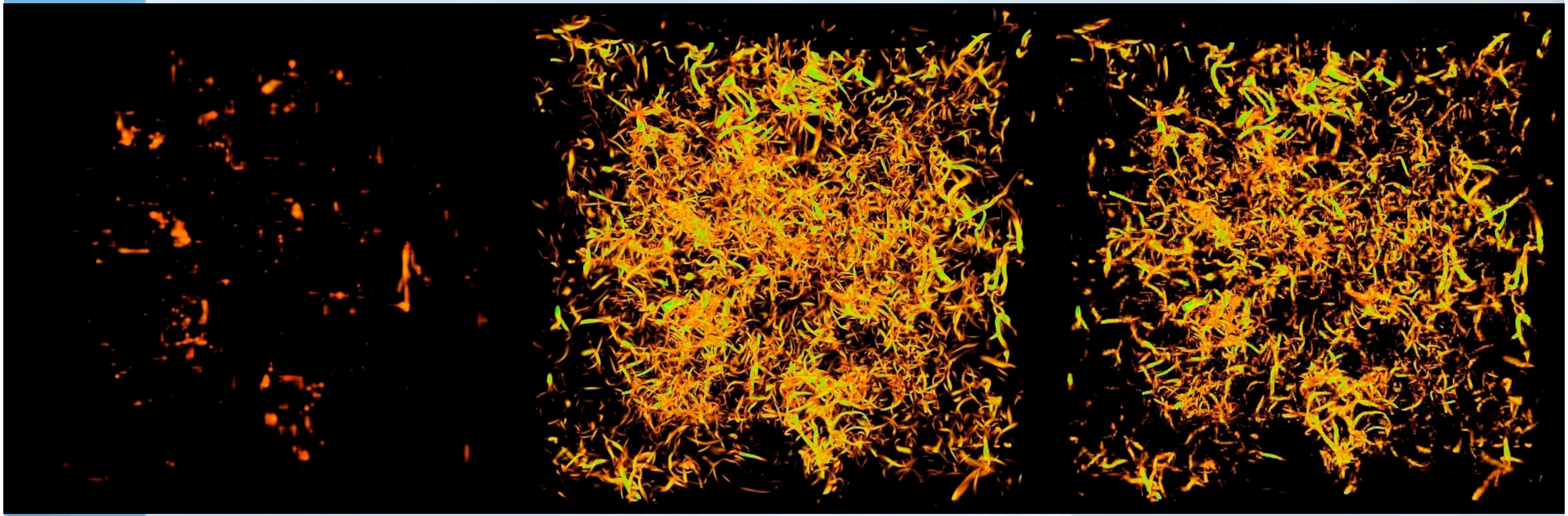


Frequency truncation | No compression | Coefficient prioritization

Coiflet-12 wavelet
No blocking

512:1 Compression

1024^3 Taylor-Green turbulence (enstrophy field) [P. Mininni, 2006]

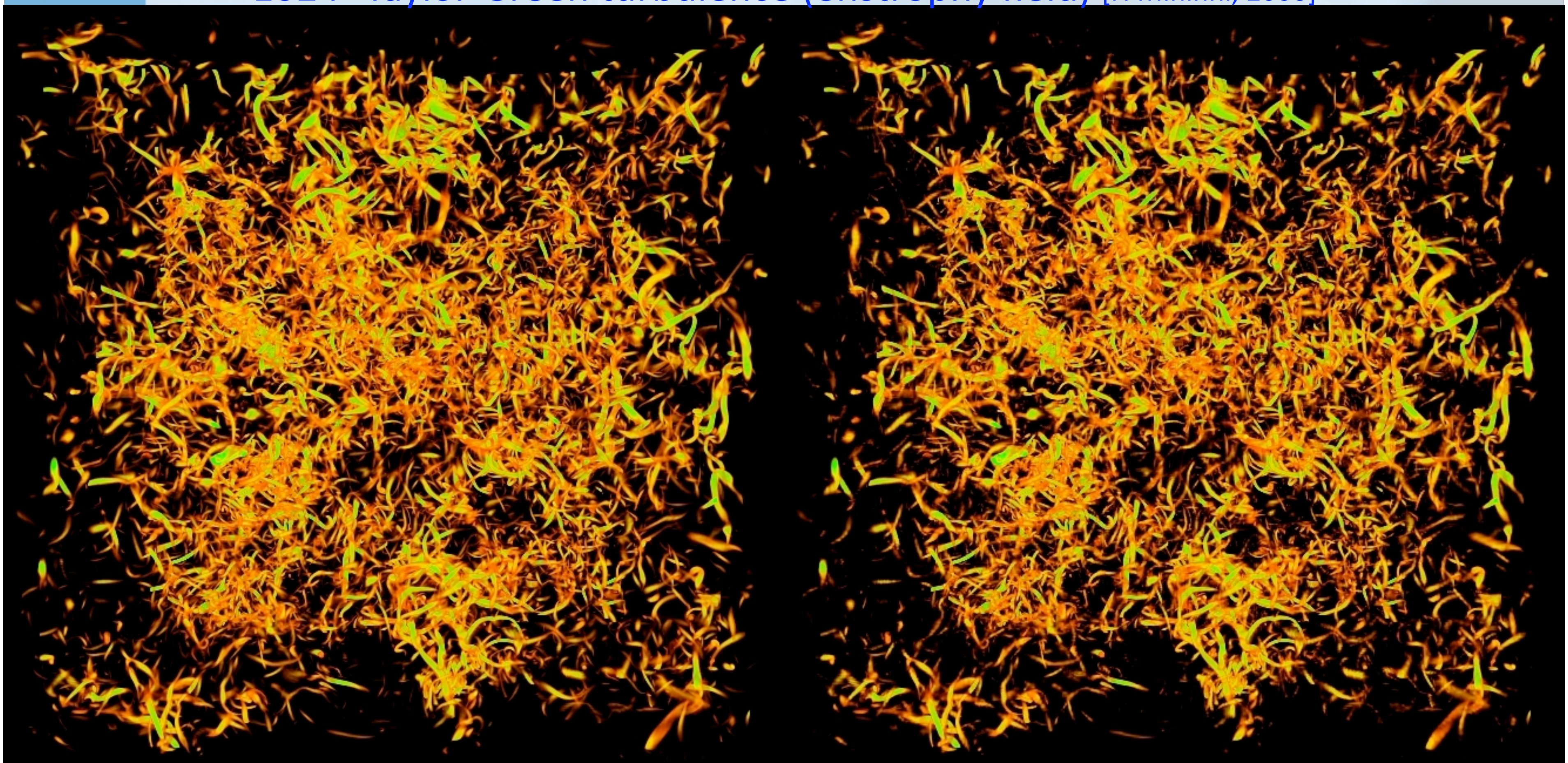


Frequency truncation | No compression | Coefficient prioritization

Coiflet-12 wavelet
No blocking

100:1 Compression

1024^3 Taylor-Green turbulence (enstrophy field) [P. Mininni, 2006]



No compression

Coefficient prioritization

Coiflet-12 wavelet
No blocking

VAPOR Summary

- Multiresolution data representation
  - Enables interactive access to massive datasets
  - **Hypothesis may be interactively explored with coarsened data and later validated (perhaps non-interactively) with native data**
- Visualization aided data analysis
  - Intended to be used by scientists, not visualization specialist
  - Requirements defined by a steering committee of scientists
- Narrow focus: Earth & space CFD simulations
  - Algorithms
  - Data types
- Emphasis on desktop/laptop platforms, not on visualization supercomputers

Acknowledgements

- Steering Committee
  - Benjamin Brown – U. of Wisconsin
  - Nic Brummell – CU
  - Gerry Creager – Texas A&M
  - Yuhong Fan - NCAR, HAO
  - Aimé Fournier – NCAR, IMAGe
  - Pablo Mininni - NCAR, IMAGe
  - Aake Nordlund - University of Copenhagen
  - Leigh Orf - Central Michigan U.
  - Yannick Ponty - Observatoire de la Cote d'Azur
  - Thara Prabhakaran - U. of Georgia
  - Annick Pouquet - NCAR, ESSL
  - Mark Rast - CU
  - Duane Rosenberg - NCAR, IMAGe
  - Matthias Rempel - NCAR, HAO
  - Geoff Vasil, CU
- Developers
  - John Clyne – NCAR, CISL
  - Dan Lagreca – NCAR, CISL
  - Alan Norton – NCAR, CISL
  - Kenny Gruchalla – NREL
  - Victor Snyder – CSM
  - Kendal Southwick – NCAR, CISL
- Research Collaborators
  - Kwan-Liu Ma - U.C. Davis
  - Hiroshi Akiba - U.C. Davis
  - Han-Wei Shen - Ohio State
  - Liya Li - Ohio State
- Systems Support
  - Joey Mendoza - NCAR, CISL
  - Pam Gilman - NCAR, CISL

NCAR

Questions???

www.vapor.ucar.edu

vapor@ucar.edu