Subsetting, Visualization and Analysis of Cloud Resolving Model Data

Karen Schuchardt

Pacific Northwest National Laboratory

Thanks to Charlotte DeMott (CSU) for permission to re-use slides.

Outline

- CRM Project Overviews and Data
- Motivating Example
- Visualization Tools/Needs
- Collaboration Thoughts

Projects

- Design and Testing of a Global Cloud Resolving Model (GCRM) (Scidac SA, Randall)
- Center for Multi-scale Modeling of Atmospheric Processes (CMMAP) (NSF STC, Randall)
- Community Access to Global Cloud Resolving Model Data and Analyses (Scidac, Schuchardt)

Why Cloud Resolving Models

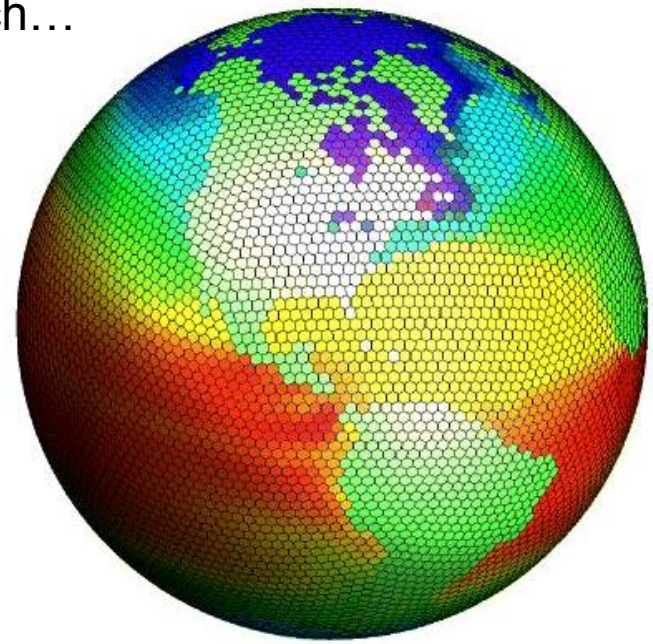
- [cirrus cloud] representation in Numerical Weather Prediction (NWP) models and General Circulation Models (GCMs) has been identified as one of the greatest uncertainties in weather and climate research.

- CRMs may be increasingly used in GCMs to replace the cumulus and stratiform cloud parameterizations (e.g., Khairoutdinov and Randall 2001). Thus, evaluating the representation of cirrus clouds in a CRM will soon be considered part of evaluating GCMs.

Yali Luo, Steven K. Krueger, and Gerald G. Mace, Kuan-Man Xu

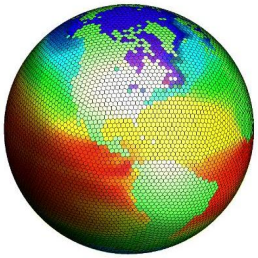
Global Cloud Resolving Model (GCRM)

Dave Randall (PI) (CSU), Akio Arakawa (UCLA) (CO-PI)
CSU Researchers: Heikes, Konor, DeMott, Dazlich…



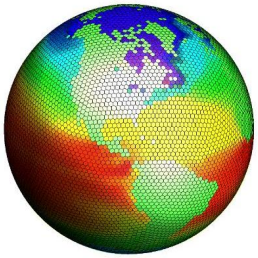
- prototype of **future-generation** cloud-resolving global atmospheric models for use in both **weather forecasting and climate** simulation
- **global atmospheric circulation** model with a grid-cell spacing of approximately 2-4 km, capable of **simulating the circulations associated with large convective clouds**
- Applications of the model will include development of improved versions of more conventional, less expensive models.

GCRM



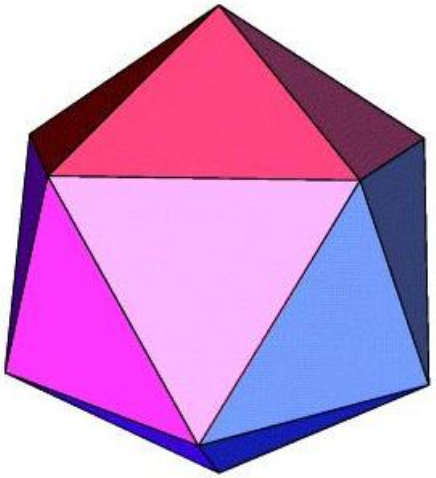
- Computationally **MOST** expensive to run
- Data intensive
  - Conservatively, 1 TB / hourly averages
- In 5 year time frame:
  - 1 month runs aimed at numerical weather predictions
  - Run >= two annual-cycle simulations
- Targeting primarily ORNL and NERSC facilities
- Production use: 20 year time frame

GCRM Data



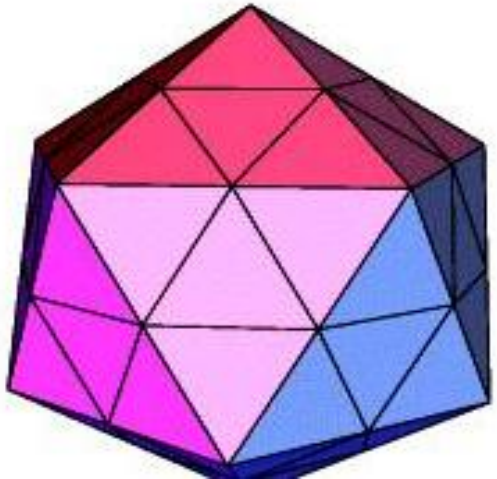
- Geodesic grid (to 2-4 km resolution)
- Time averaged global maps
  - 2d, 3d
- Time averaged region (more frequent data in some regions)
- Point data for pre-selected points (much more frequent data for some points)
- Restart data

Geodesic Grids

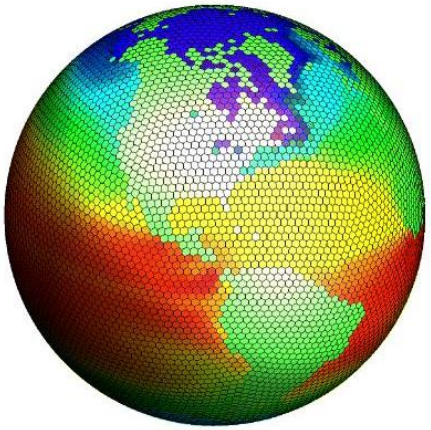


Regular Icosahedron
  Inscribed in a unit sphere
  20 triangular faces
12 vertices (grid points)

bisect each face



project each new vertex to the unit sphere



Properties of the geodesic grids at different resolution

R	Number of cells N_c	Number of cells along equator	Average cell area in km^2	Ratio of smallest cell to largest cell	Average distance between cell centers in km	Ratio of smallest to largest distance btn cell centers
0	42	10	1.21e7	0.885	3717.4	0.881
1	162	20	3.14e6	0.916	1909.5	0.820
2	642	40	7.94e5	0.942	961.6	0.799
3	2562	80	1.99e5	0.948	481.6	0.790
4	10242	160	4.98e4	0.951	240.9	0.789
5	40962	320	1.24e4	0.952	120.5	0.788

$$N_c = 5 \cdot 2^{2R+3} + 2 \;\; ; \;\; R \geq -1$$

CMMAP

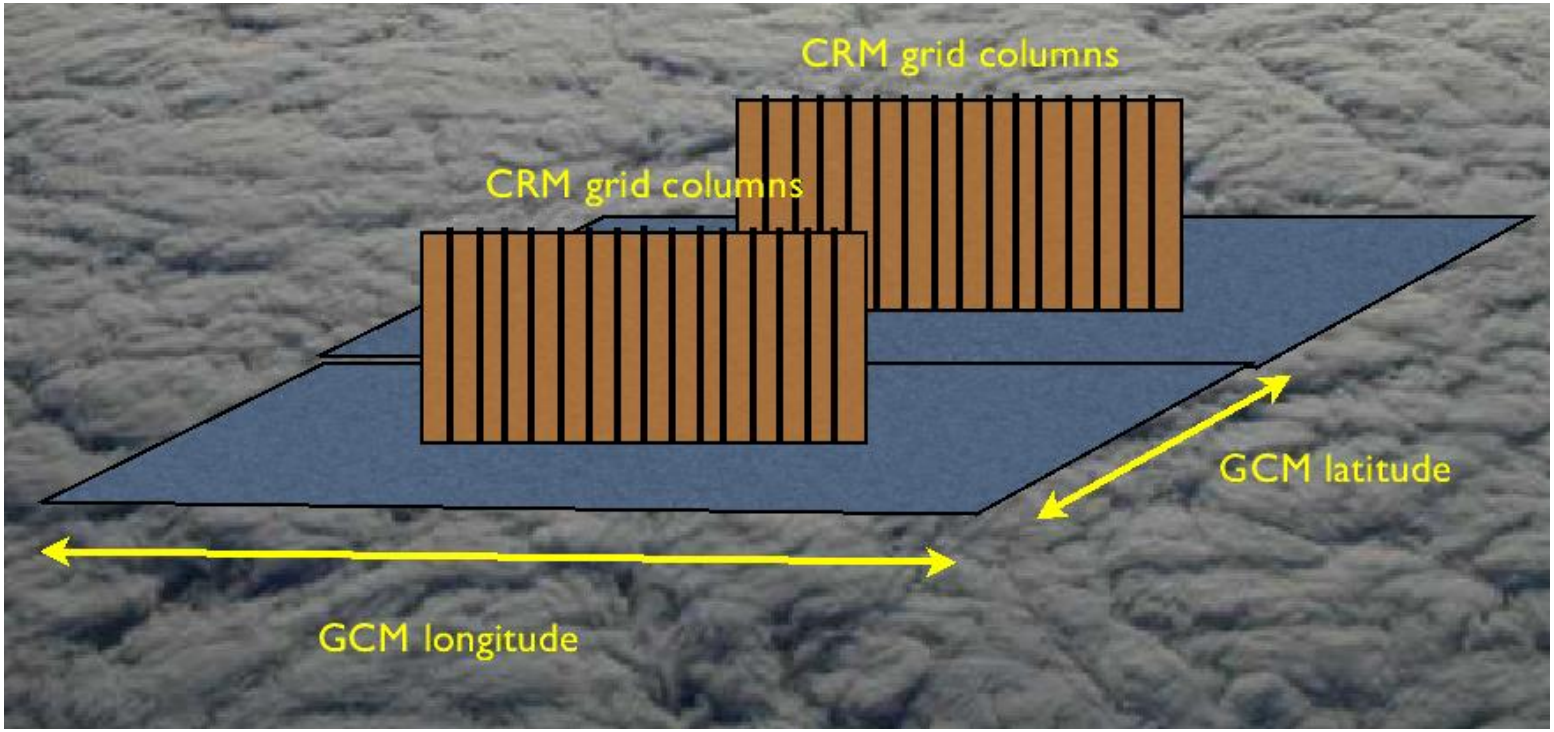
- $19 Million NSF Science and Technology Center (STC) to build climate models that will more accurately depict cloud processes, improving climate and precipitation forecasting.
- 100+ individuals at 29 organizations
- Roles:
  - Model developers
  - Model validators
  - Data Management
  - Meteorology
  - Students
  - Education outreach
  - Computer Company Outreach

Activities

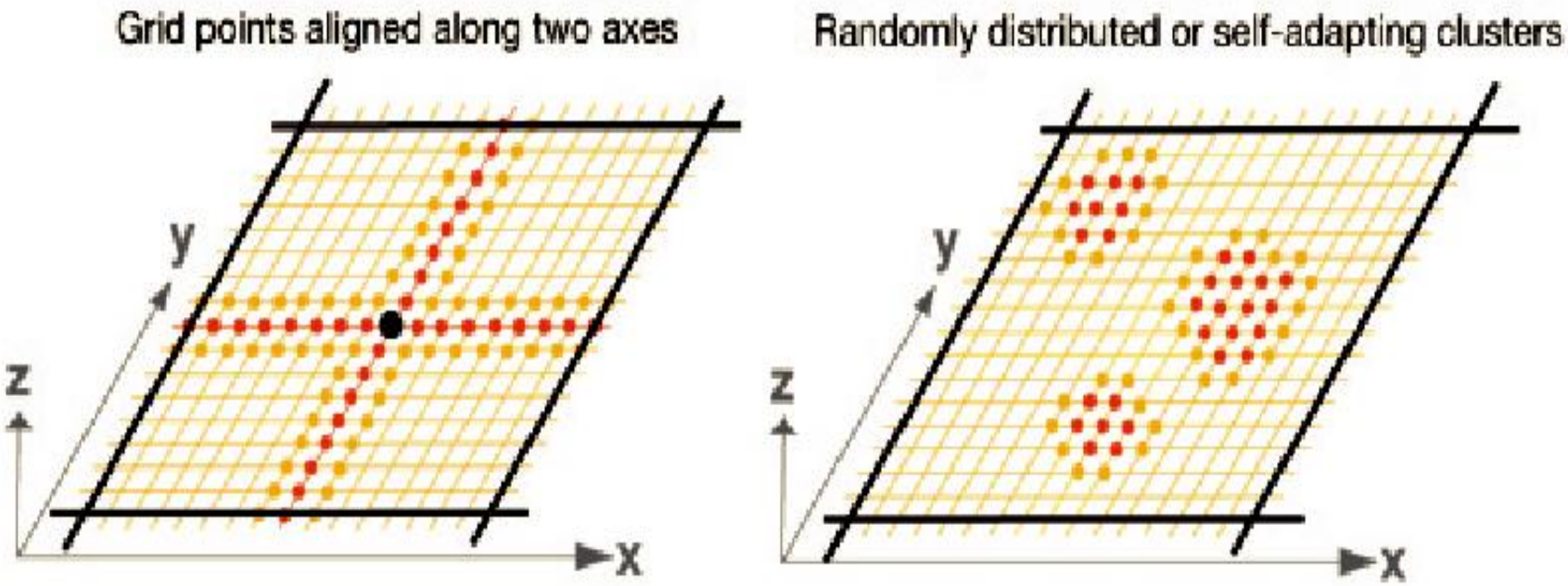
Research
Education, Outreach and Diversity
Knowledge Transfer

CMMAP

- Apply cloud resolving model to existing modeling frameworks (such as MMF) where the cloud model runs within each cell.
  - Less computationally challenging and data intensive than GCRM



CMMAP


Grid points aligned along two axes
y
z
x
Randomly distributed or self-adapting clusters
y
z
x

CMMAP

- Computationally, more expensive than GCM but less expensive than GCRM
- Research:
  - Heavy emphasis on validation against observed data
    - ARM surface data
    - CloudSat and Calypso satellite data
  - Performance optimization
  - Visualization
  - Data Management and Distribution of large data sets.

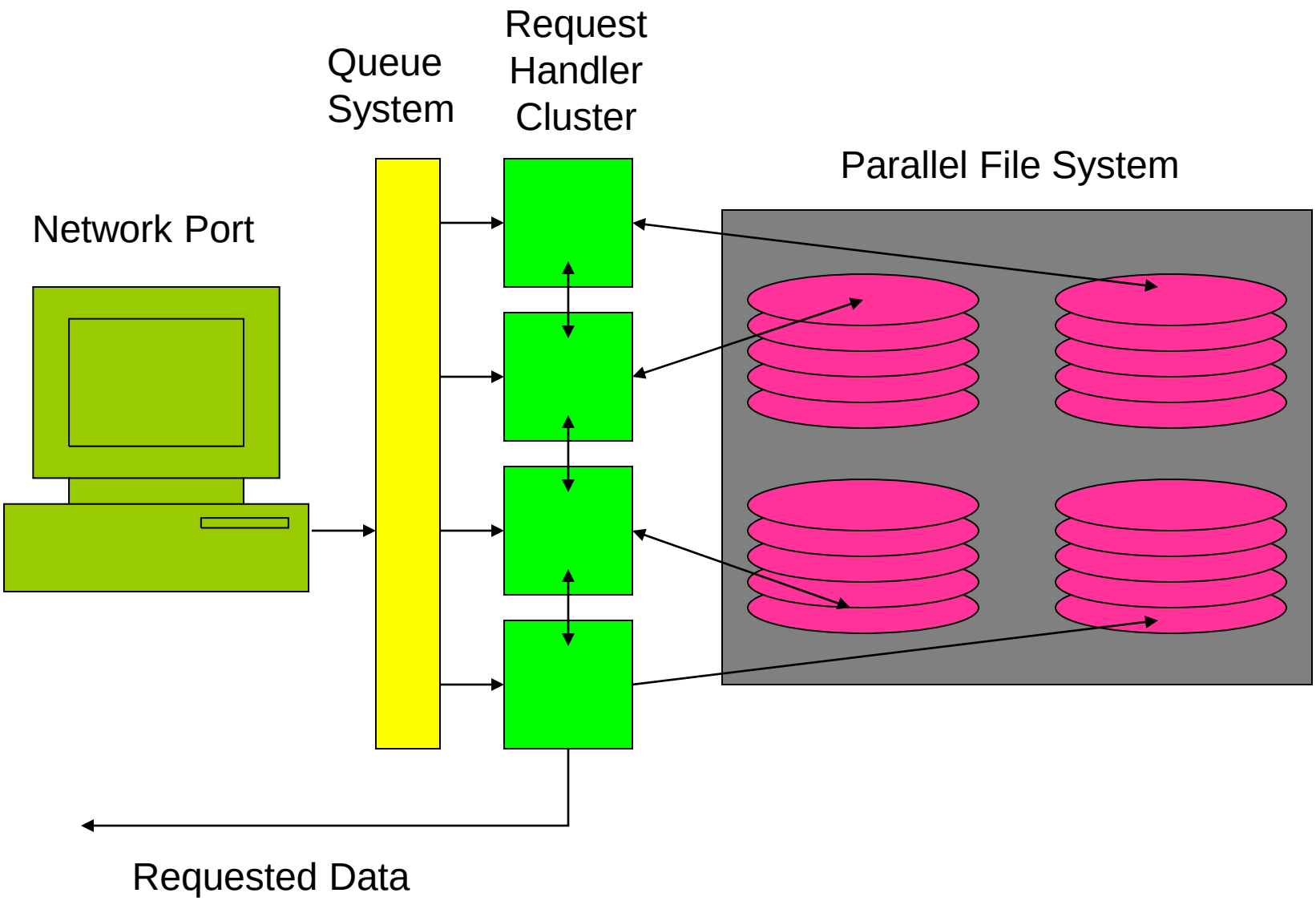
Petascale Subsetting, Analysis, and Visualization of Atmospheric Global Environment

Recall: data set sizes of ~ 1-10 TB/hour

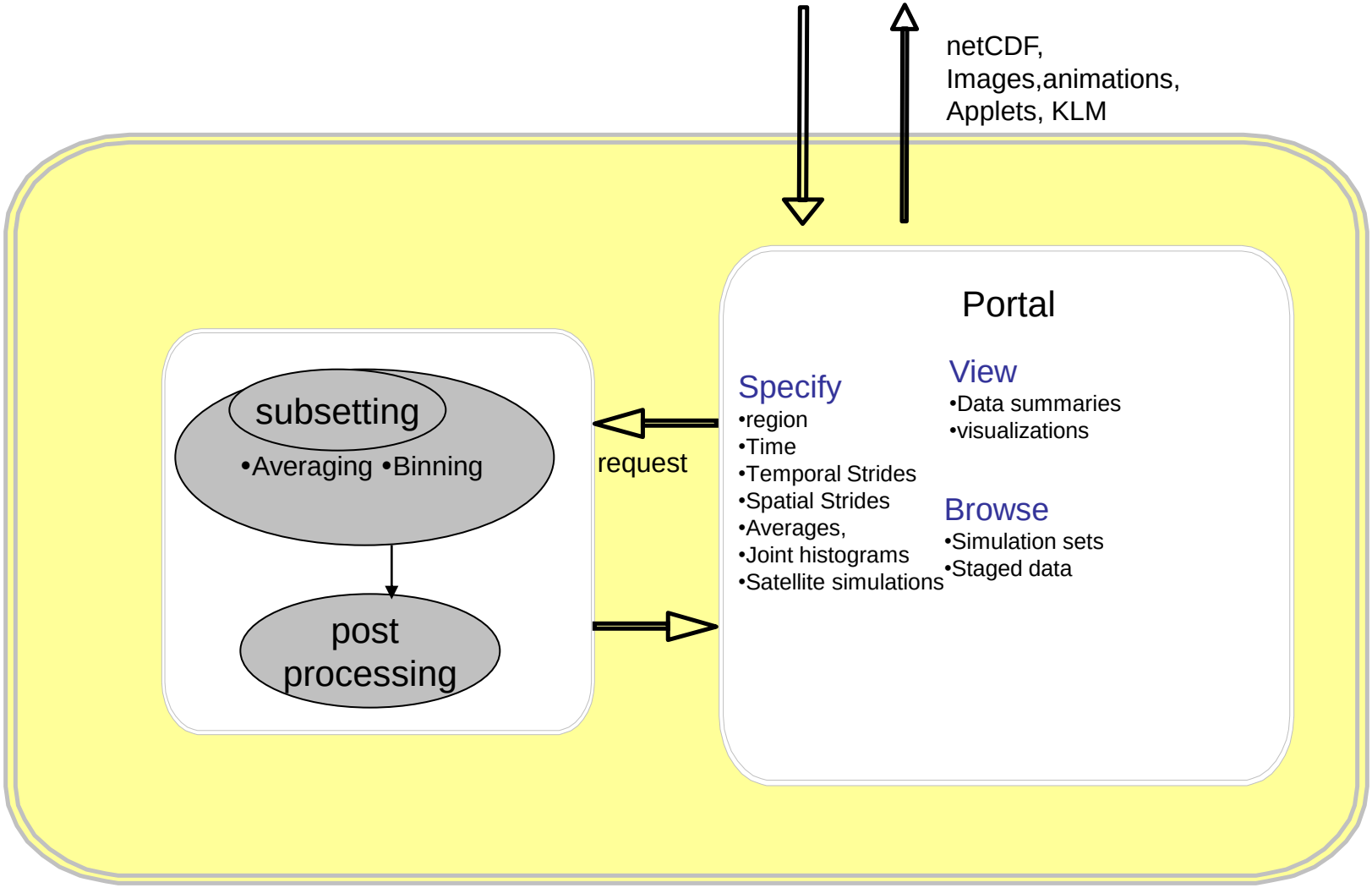
- .8 PB/month 8 PB/annual cycle

- IO Benchmarking to determine what is feasible
- Data subsetting and delivery
- Provide server-side analysis and visualization
- Bring the tools to the data

Architecture

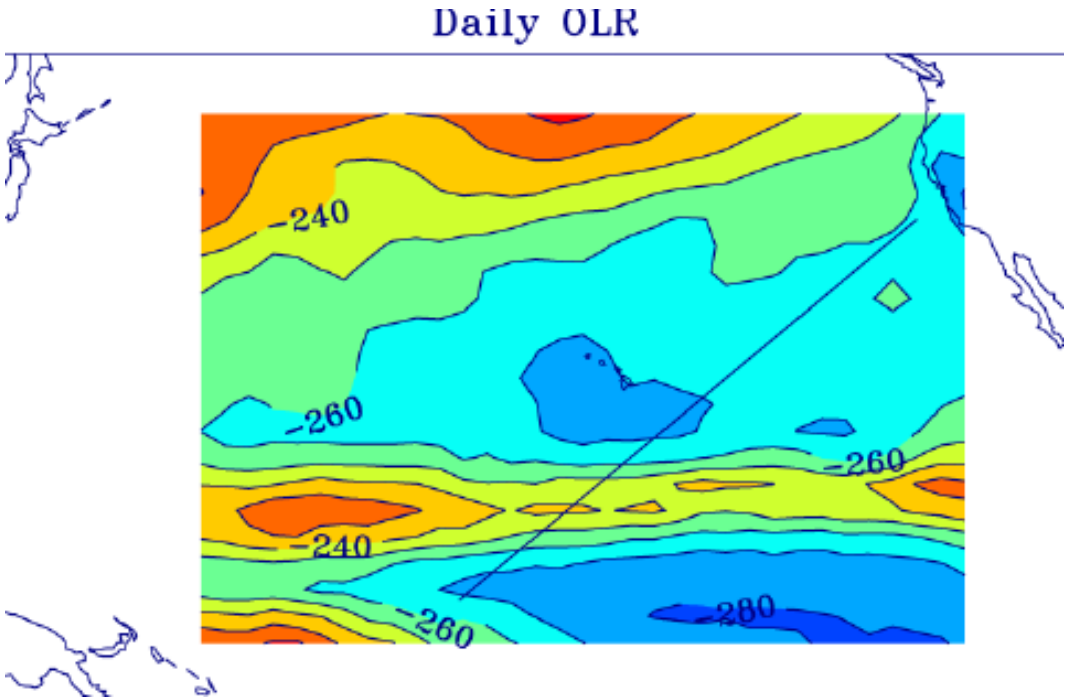


Software Architecture



Motivation: GEWEX Cloud System Study (GCSS)

- Can models reproduce the main properties of the diurnal cycle over subtropical oceans?
- Can models and satellites help characterize the humidity structure of the upper troposphere?



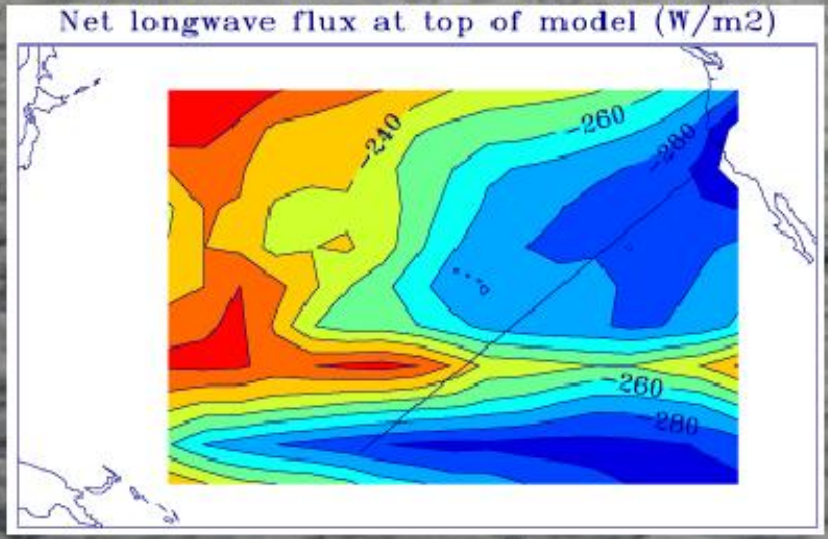
http://www.gewex.org/gcss.html

Questions Relevant to MMF

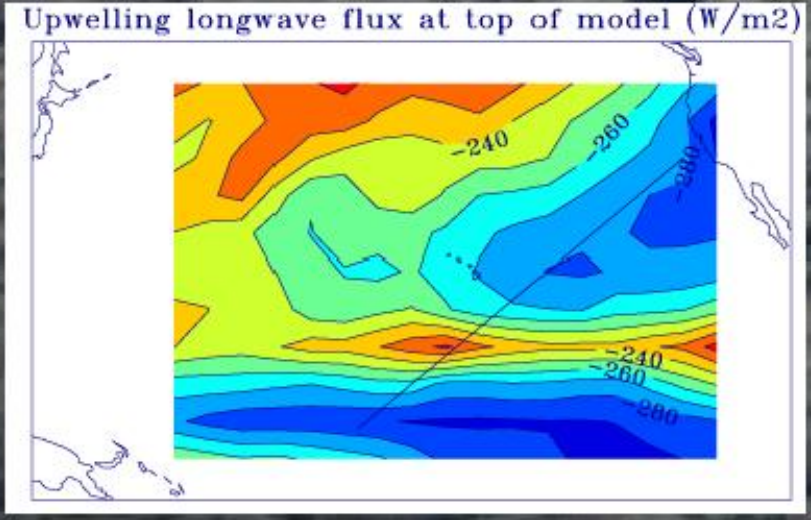
- Can MMF meet the challenges posed by GCSS
- Can the MMF produce boundary layer clouds and the transition to deep tropical convection?
- Does MMF really produce stratocumulus clouds?
- Do the MMF stratus clouds behave as they do in nature?
- Does the MMF produce a transition zone between Sc and Cu clouds?

Basic comparisons - OLR 1998

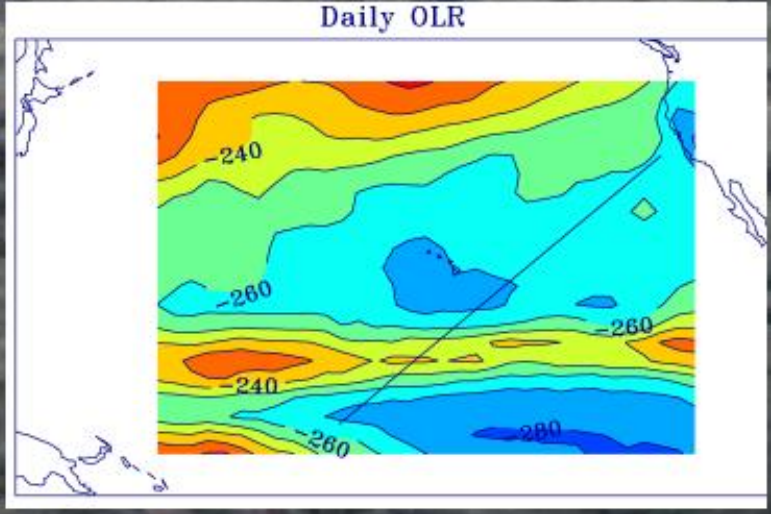
CAM3



MMF

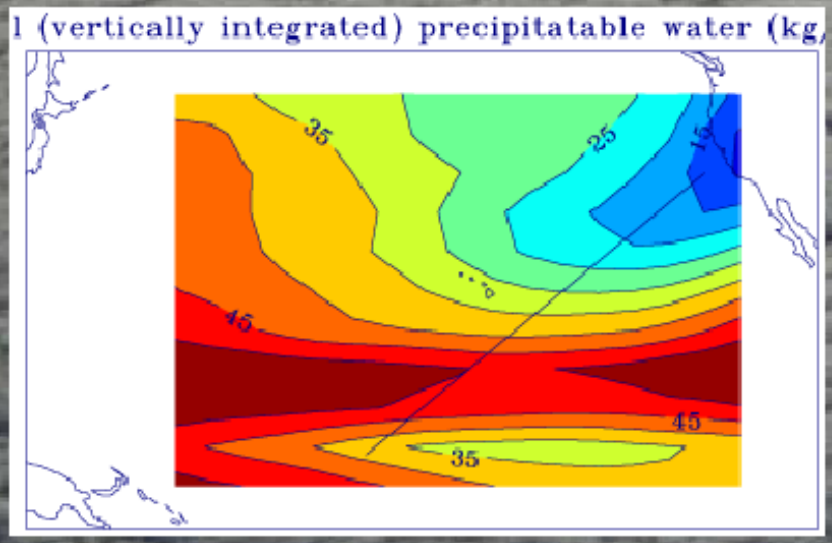


NOAA

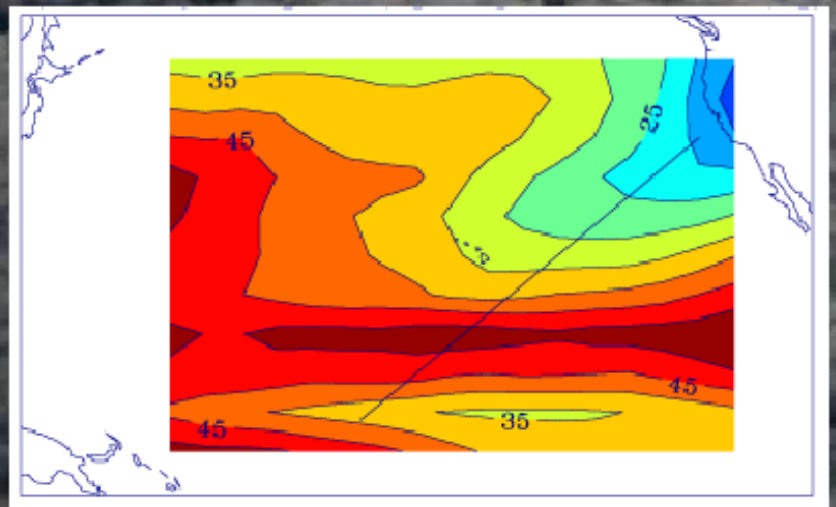


Precipitable Water (mm) - 1998

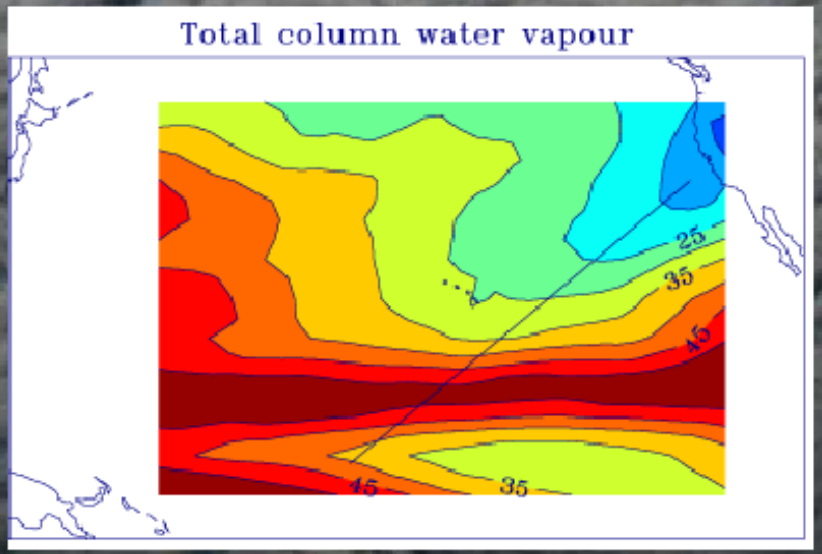
CAM3



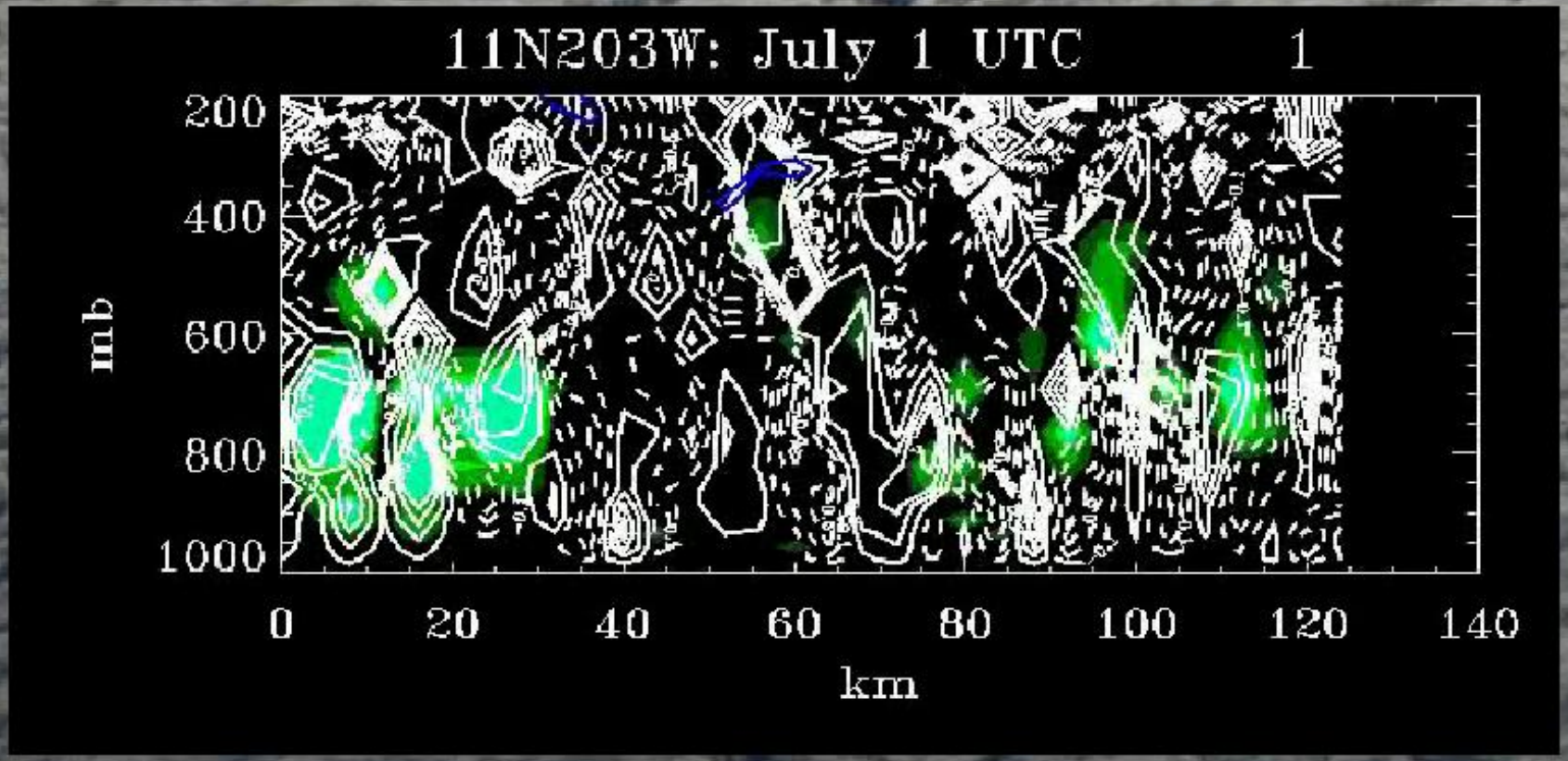
MMF



ERA40



ITCZ: Deep convection animation



Example for CRM on MMF

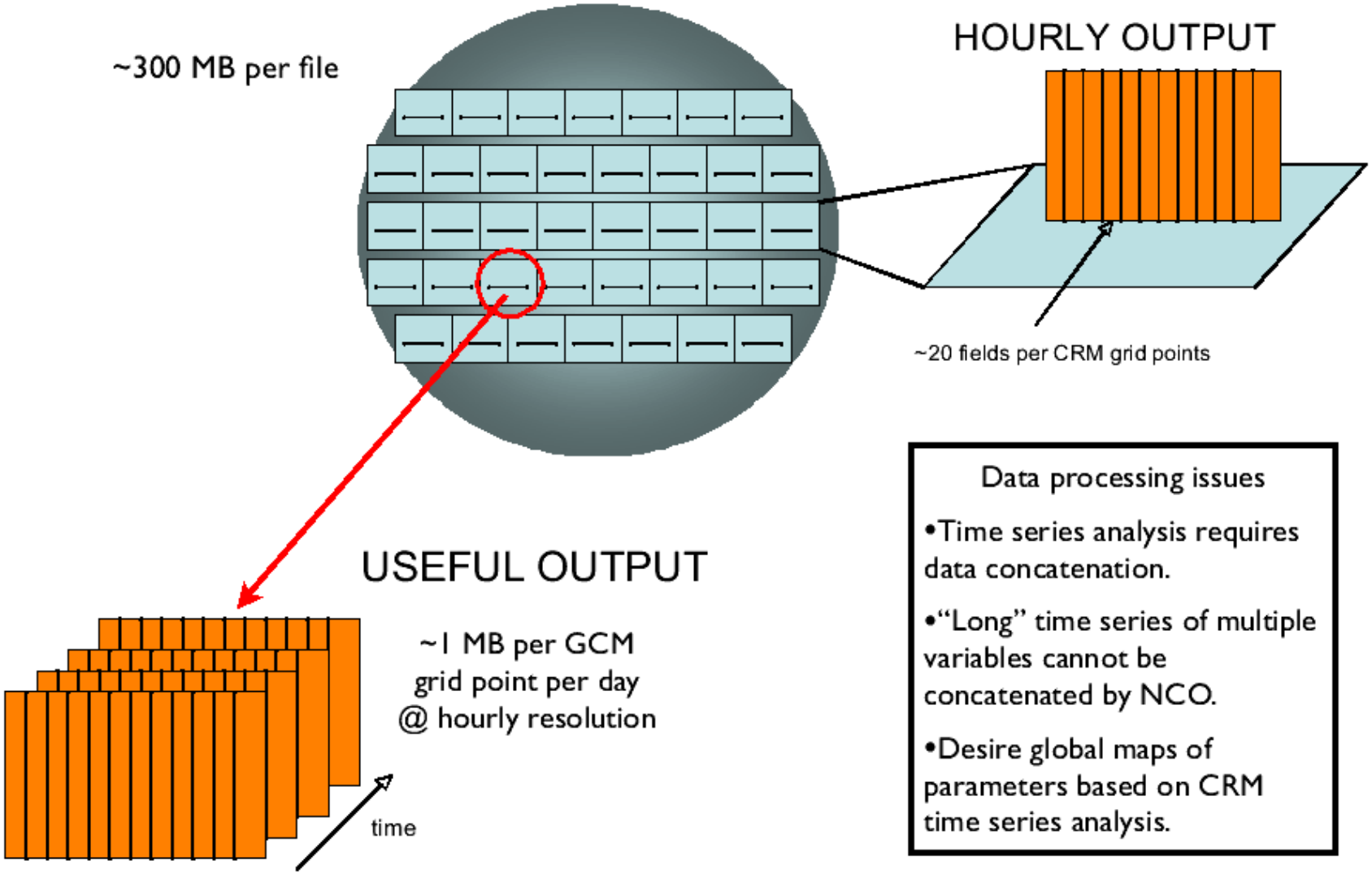
Data Output

- Lat x lon x level x variables (4D)
- Lon x lat x height x 32 CRM grid columns
- ~20 such arrays per hour

Data Use

- Extract a few GCM grid points
- Concatenate into a time series
  - For “short” time series, all fields and times into one files
  - For “longer” time series, must write one file per field

Data Issues - Schematically



Data processing issues

- Time series analysis requires data concatenation.
- "Long" time series of multiple variables cannot be concatenated by NCO.
- Desire global maps of parameters based on CRM time series analysis.

Bottlenecks – Lots of Data

- “The current process for moving output files to local storage where data can be extracted and subsequently analyzed by primarily serial tools breaks down”

- “…researchers expect to download one or more NetCDF files to local storage for analysis. The files typically contain more parameters than of interest for a particular analysis, thus creating unnecessary network traffic and increased processing time.”

Bottlenecks – Lots of Data

- In many cases, subsetting of data may still produce data sets that are too large; **server side analysis and visualization** capabilities are needed
  - Animations of a cell or region, or globe
- Parallel support (for subsetted local data) will be useful to take advantage of the multi processor desktop chips.

Current Visualization Tools

- Model Developers: Everybody has their own favorite
  - MatLab
  - IDL
  - NCAR Tools
- CMMAP Community
  - Ditto except on a wider scale

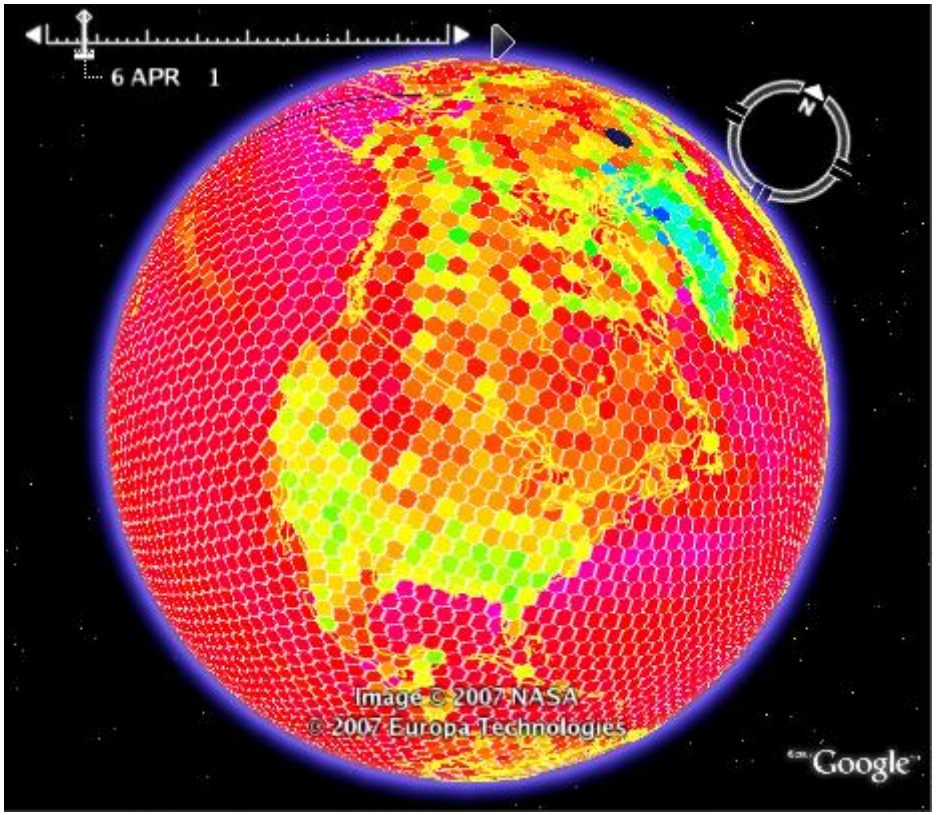
Platforms (GCRM): Linux, Mac

Future Capabilities

- Server side subsetting and analysis
  - Predefined subsetting, analysis, visualization
  - Plug in mechanism to allow users to add custom features
  - Examples with abstractions so users don’t have to know all the complexity of parallel tools

Capabilities (Continued)

- Effective use of tools such as GoogleEarth, Google maps, NASA’s World Wind
  - Course detail, zoom to fine detail
- Typical visualizations
  - Lines (as in GCSS)
  - Blocks
  - Region of time
  - Moving regions
  - Orographic maps
- “The skies the limit”

Collaboration Opportunities

- Model vs Observed comparisons
- Analysis / viz tools we can plug into back end services
- Fly throughs
- Direct support for geodesic grids