Some big data challenges in observing, monitoring and modeling the earth system

Valentine Anantharaj <anantharajvg@ornl.gov>
National Center for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, TN 37831

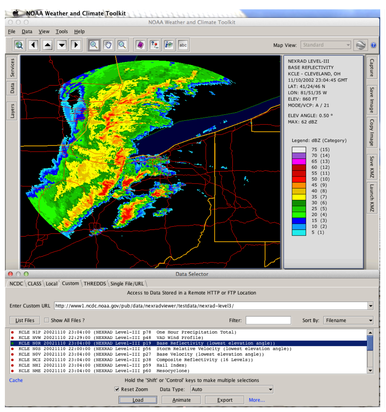
Earth observing systems

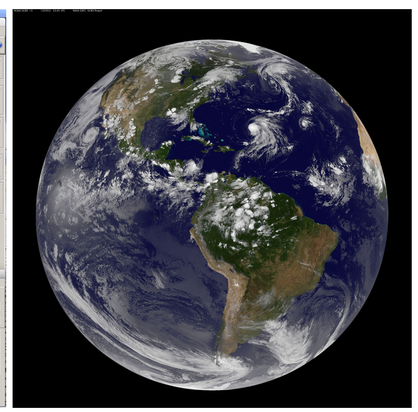
GOES-R Satellite – Advance Baseline Imager [2015]

- Full disk (FD) scans every 5 or 15 minutes
- ~ 8 GB FD image set (Scan mode 3) => ~ 280 TB / Yr
- 20+ "derived" products [~ 670 TB / Year archive]
- Near-realtime operational use [~ 55 Mbps ABI only]

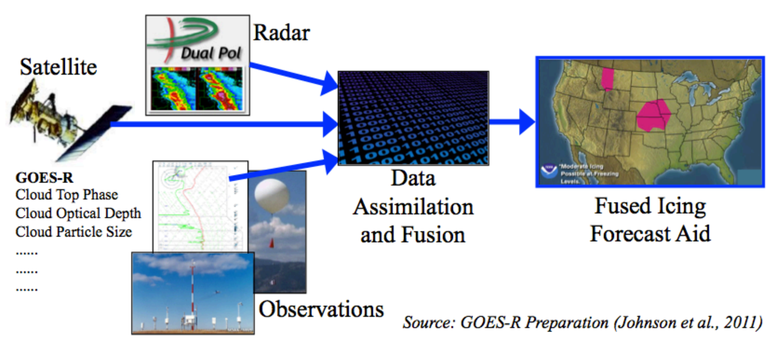
US NEXRAD

- Operational *doppler* RADAR
- Volume of data varies based on scan mode
- Typical NEXRAD II volume scan ~ 1.5 GB / day
- 58 systems in operation => ~ 32 TB / year [estimated]
- Numerous NEXRAD III products



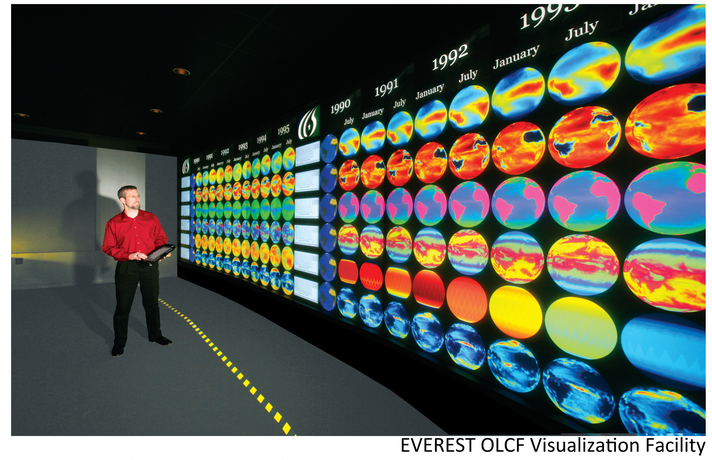


GOES-R data fusion (example)



Source: GOES-R Preparation (Johnson et al., 2011)

Numerical Modeling



EVEREST OLCF Visualization Facility

Ultra High Resolution Project (J. Hack, PI)

- 70M hours using the *Community Earth System Model*
- 35,000 files, ranging in sizes from 21 MB to 110 GB
- The total volume of model output will be 300+ TB

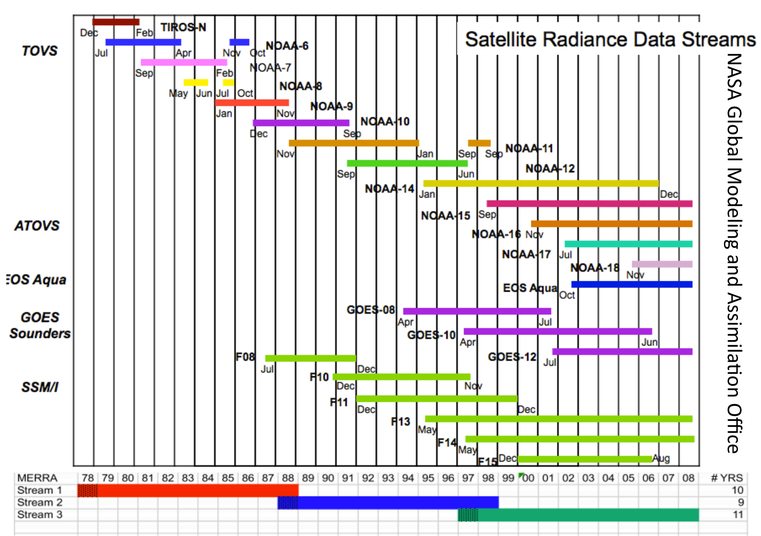
Early Science using CAM-SE (J-F. Lamarque, PI)

- *CAM-SE* with *MOZART* Chemistry (106 tracers)
- 450M hours on Titan using GPU accelerators.
- 250+ TB of simulation results

Climate End Station (W. Washington, PI)

- ~ 200M hours in 3 years and 200+ TB data

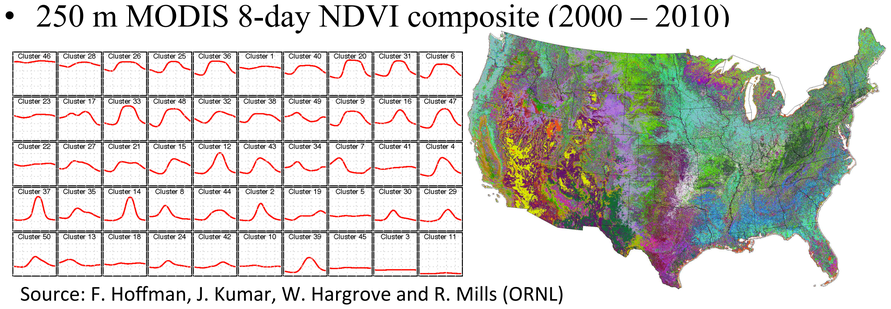
Satellite streams for reanalysis (Example: MERRA)



New insight via analytics

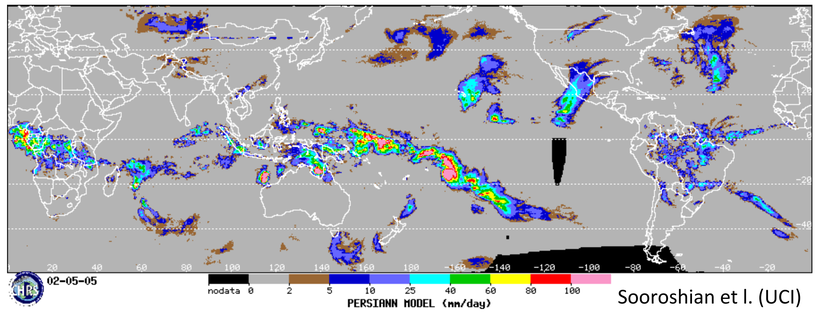
Phenoregions

- Clustering and time series analysis of multiyear data
- 250 m MODIS 8-day NDVI composite (2000 – 2010)



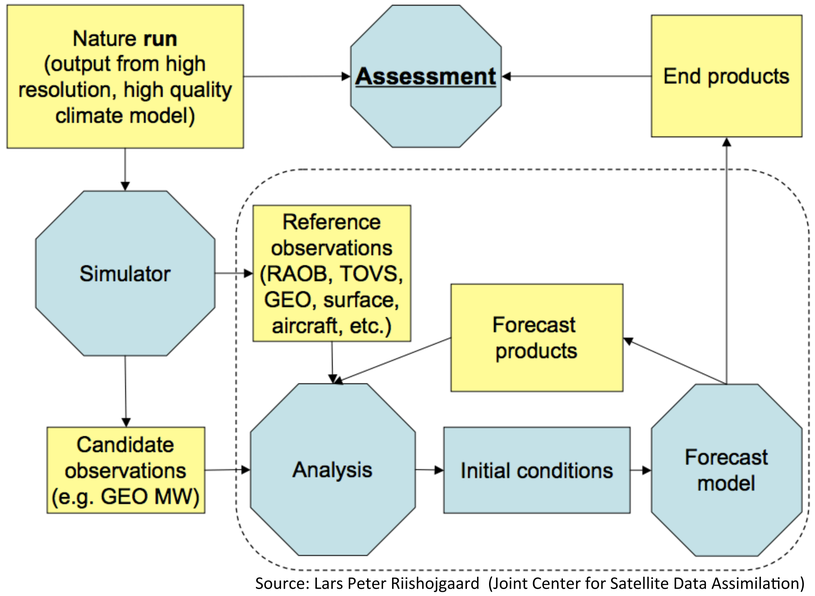
Source: F. Hoffman, J. Kumar, W. Hargrove and R. Mills (ORNL)

Using ANN to merge multisatellite data (PERSIANN)



Sooroshian et l. (UCI)

Observing System Simulation Experiments (OSSE)



Source: Lars Peter Riishojgaard (Joint Center for Satellite Data Assimilation)

Office of Science
U.S. DEPARTMENT OF ENERGY
SciDAC
ASCR