Toward Visual Analysis of Ensemble Data Sets

Or,

You want to render *what*?

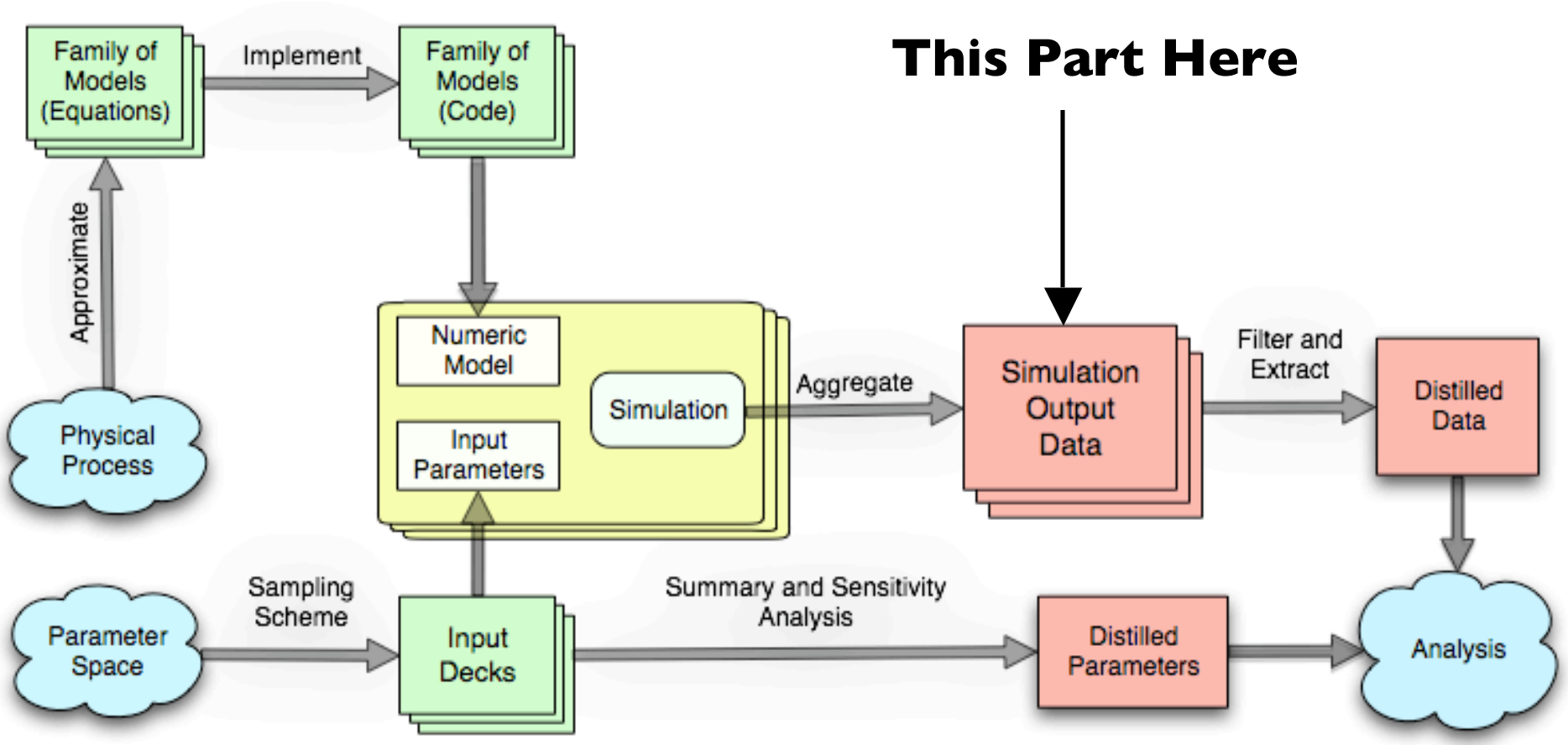
2009 Ultrascale Visualization Workshop

November 16, 2009

Andy Wilson
Sandia National Laboratories
Kristi Potter
SCI Institute, Univ. of Utah

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

What is an ensemble data set?



Sandia National Laboratories

Ensembles are…

- **Large**
  - Tens of gigabytes to hundreds of terabytes
- **Multivariate**
  - Typically 10s to 100s of state variables
- **Time-varying**
- **Multivalued**
  - Think of it as PDF-valued instead of scalar-valued
- **Awkward**
  - Raw data is frequently discarded in favor of an Excel spreadsheet

Ensembles help mitigate uncertainty

- **Multiple models**
  - Incorporate strengths of different approaches
- **Multiple runs**
  - Sample an input space of uncertain parameters
  - Perturb measured inputs to mitigate model/measurement error
- **Multiple grids**
  - Evaluate and demonstrate convergence
- **Multiple values**
  - Reason about the most likely simulation outcomes

A Few Examples

- **NOAA/NCEP Short-Term Reference Ensemble**
  - Weather for North America
  - 4 models, 21 members, 624 state variables, 30 timesteps, 36GB/run, 3 runs/day
- **Climate Simulations (Earth System Grid)**
  - Worldwide climate over millenia
  - 30TB+ repository at LLNL, lots more elsewhere
  - Varying simulation domains
- **Parameter studies for uncertainty quantification**
  - Engineered systems under stress
  - Weather/climate data makes a good proxy

Sandia National Laboratories

Driving Questions

- **What conditions are predicted by this ensemble?**
- **Where and when do those conditions occur?**
- **What is the relative probability of some outcome?**

Sandia National Laboratories

Major Research Issues: Data Management

- **Key Insight: The user only ever needs a tiny subset of the data -- but that subset changes frequently.**

- **Examples:**
  - **What phenomena does the ensemble predict?**
    - This is usually derived or inferred from the data
  - **Where and when will it happen?**
    - Moral equivalent of an SQL WHERE clause
  - **What is the relative probability of X?**
    - Derive this from "where and when" by aggregating over ensemble members

This calls for data stores with database-like access and query semantics.

#include <bill_howe_cloud_vis.ppt>

Sandia National Laboratories

Major Research Issues: Many-Valued Data

- **Spatial PDF visualization does not (yet) appear necessary**
  - Summary statistics + drill-down suffices
  - Even that much is difficult
- **The world is often not Gaussian**
  - Beware of mean + standard deviation!
  - Watch out for multimodal distributions
- **There's Just Too Much Data**
  - "Display it all and let the analyst browse" doesn't work
  - Query-driven visualization becomes very important here
  - We may not be able to use the supercomputers "just for vis"

Our Approach: Data Store

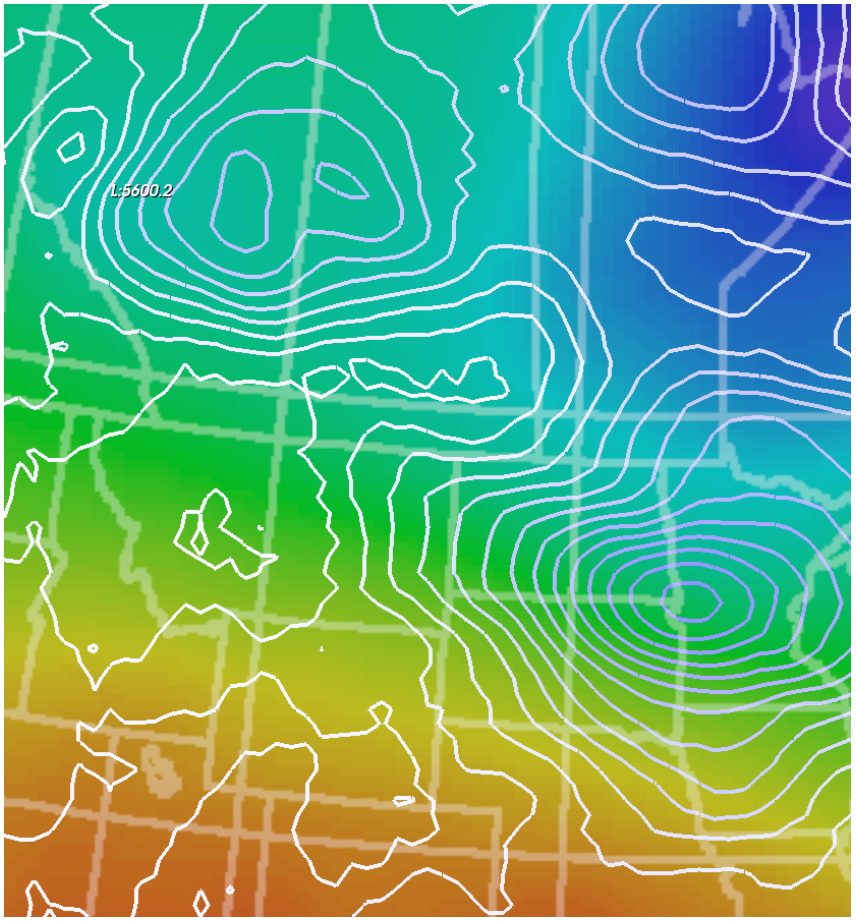
- **Netezza NPS data warehouse appliance**
  - Parallel database with CPU right next to disk
  - Schema exposes data values directly to the database
  - All numeric queries go through SQL
- **Research into data stores continues**
  - Column-store database?
  - Numeric index like FastBit?
  - Reordering data for multiresolution access?
  - Cloud storage and processing?
- **Assumption: We can move/index/reorganize the data at least once (possibly as it's being generated)**

Sandia National Laboratories

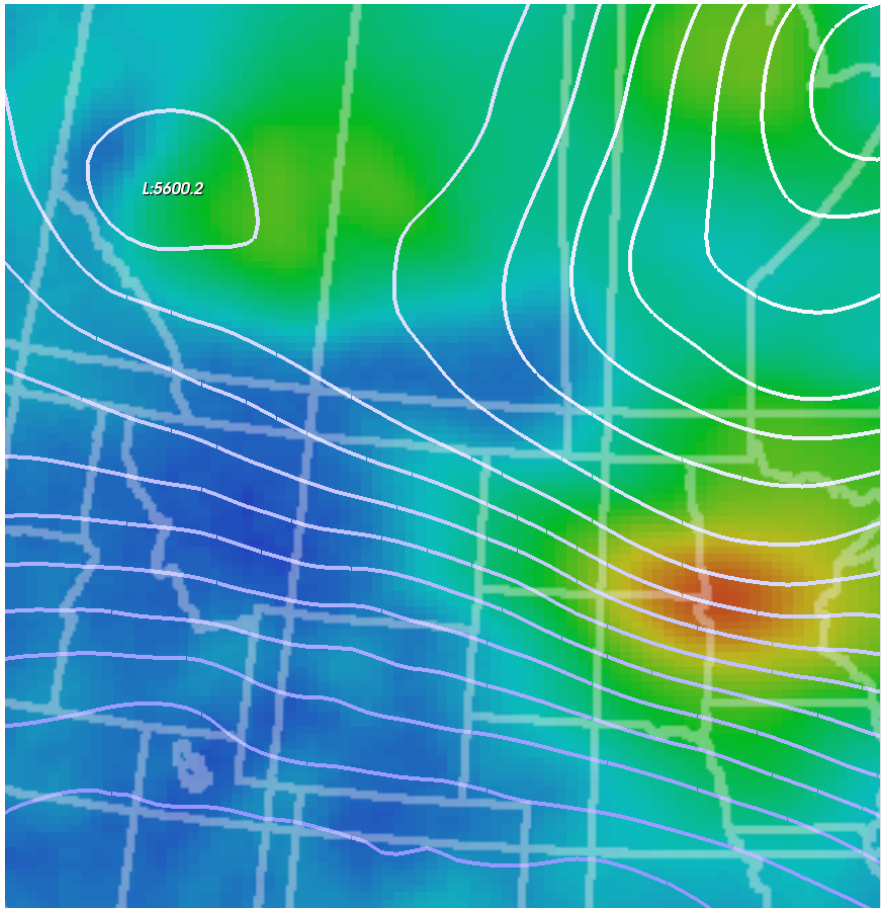
Our Approach: Visualization

- **There Is No Perfect Display**
  - Many displays, all linked
- **An ensemble data set has several dimensions:**
  - Time (1)
  - Space (2, 2.5 or 3)
  - State Variable (1 to 1000)
  - Ensemble Member (1 to 1000)
- **Collapse dimensions to yield 2- or 3D display**
  - "Collapse" means "extract or aggregate" here

Summary Display



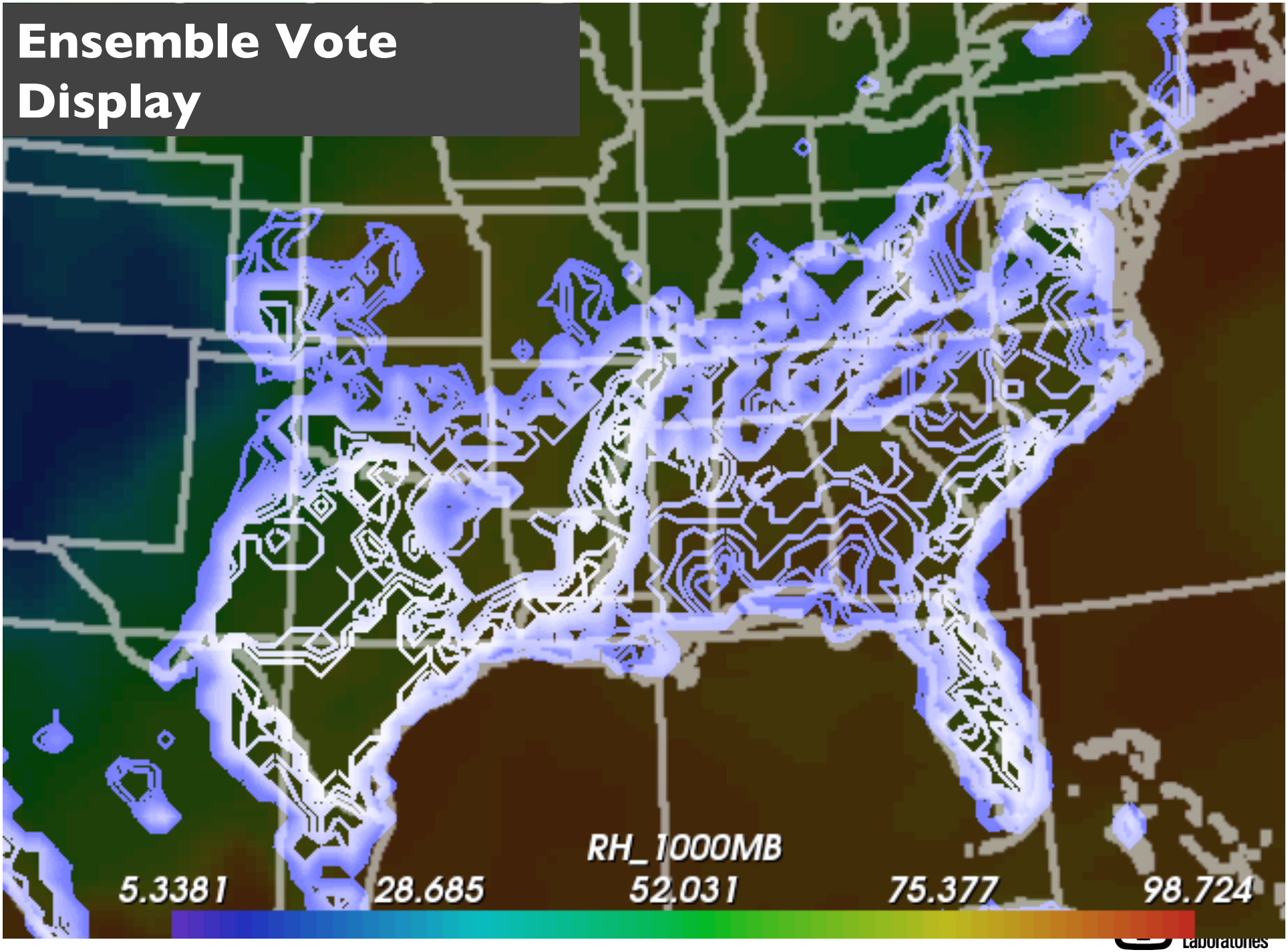
Color = Mean



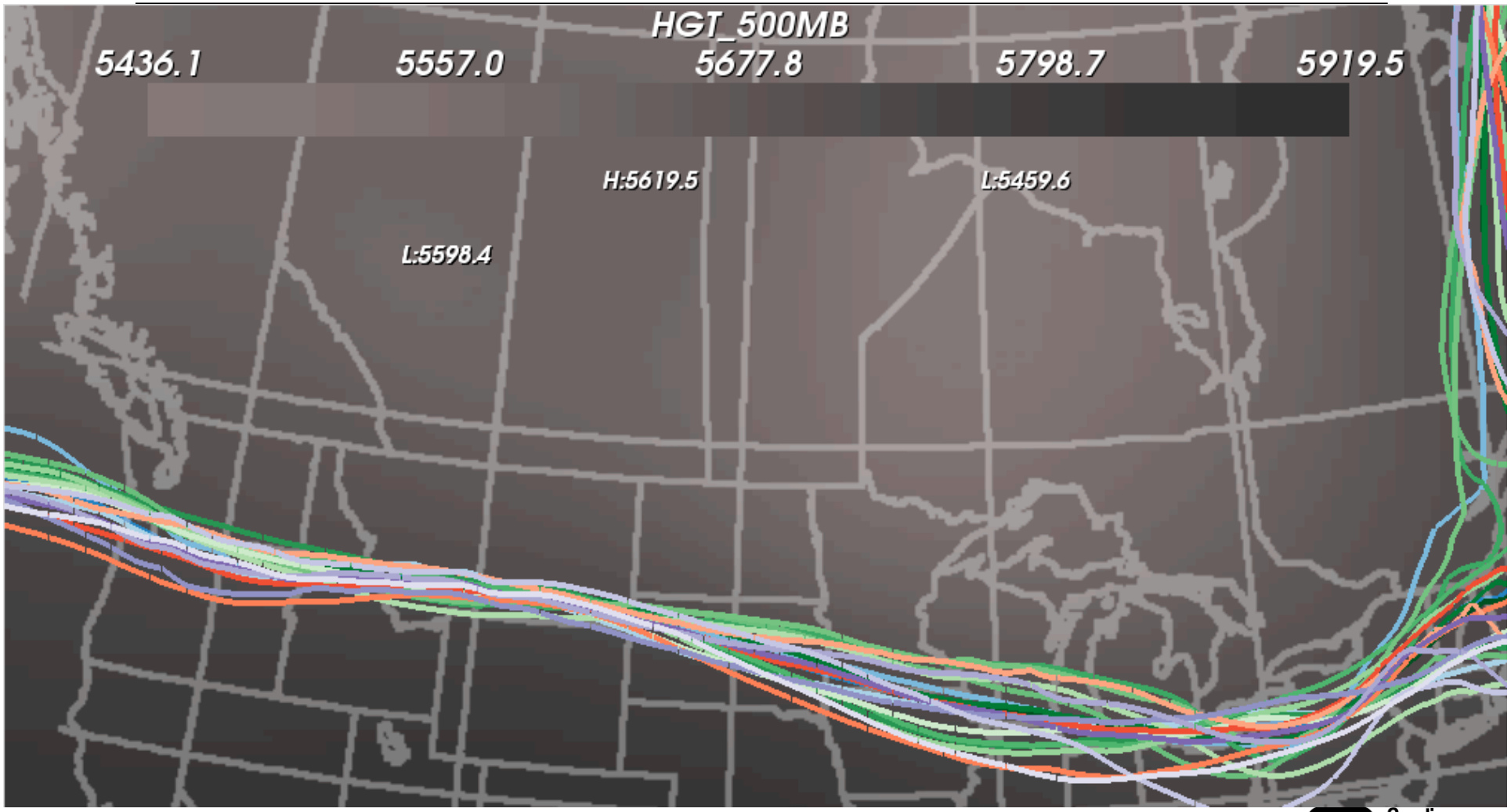
Color = Std. Dev.

Sandia National Laboratories

Ensemble Vote Display

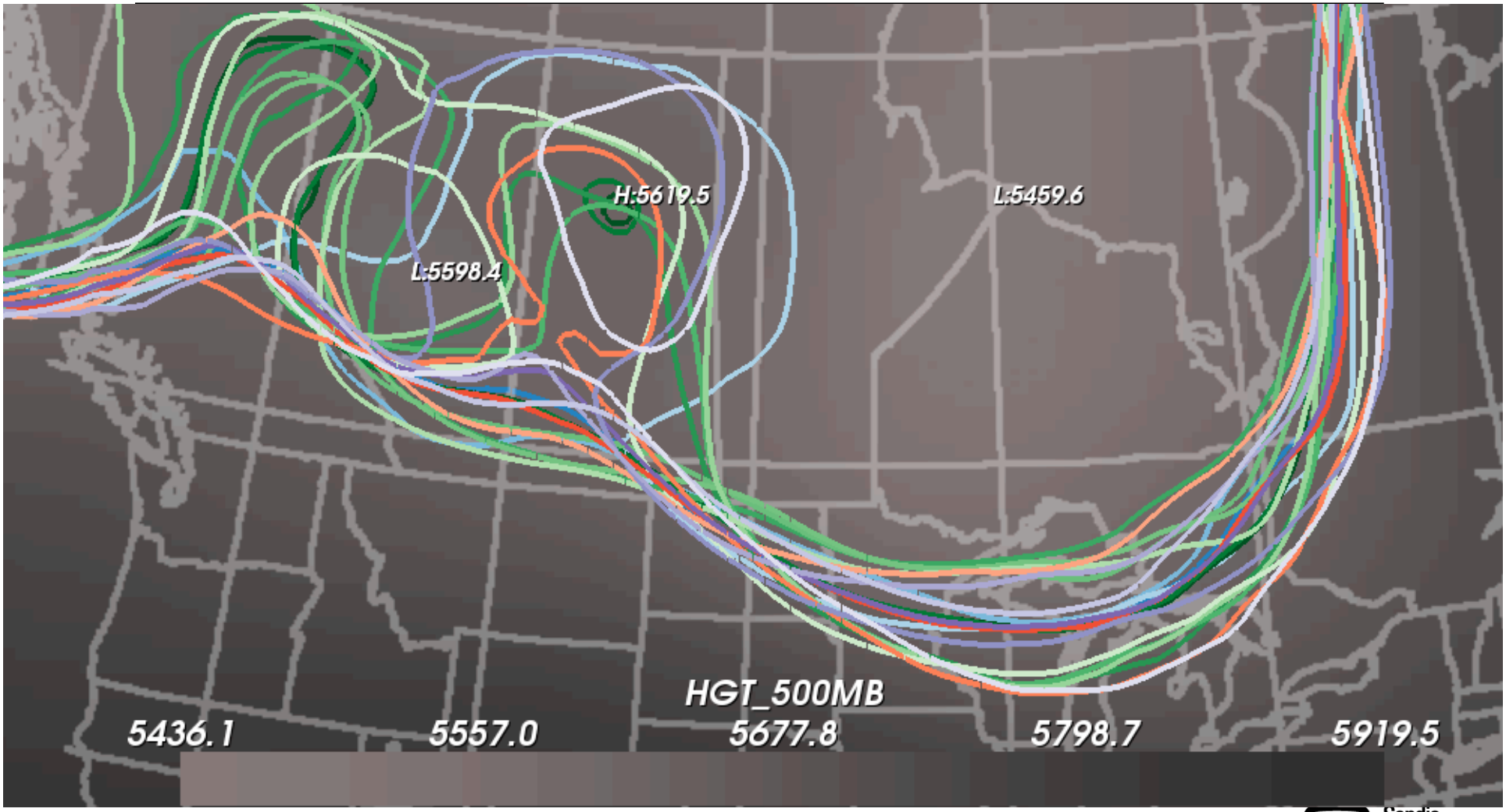


Spaghetti Plot (I)



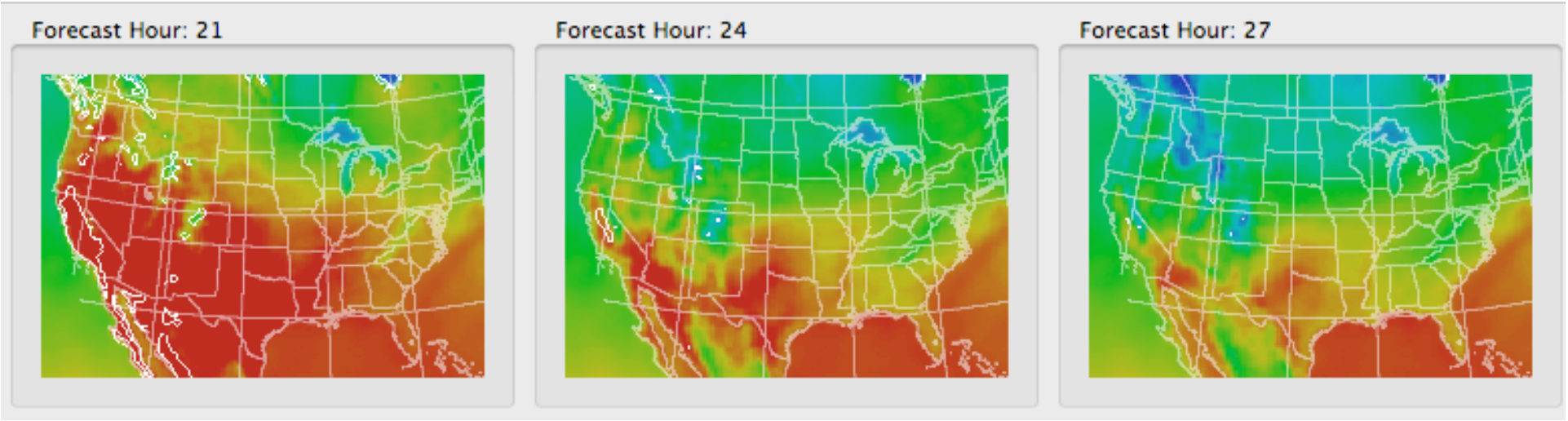
Sandia National Laboratories

Spaghetti Plot (2)



Sandia National Laboratories

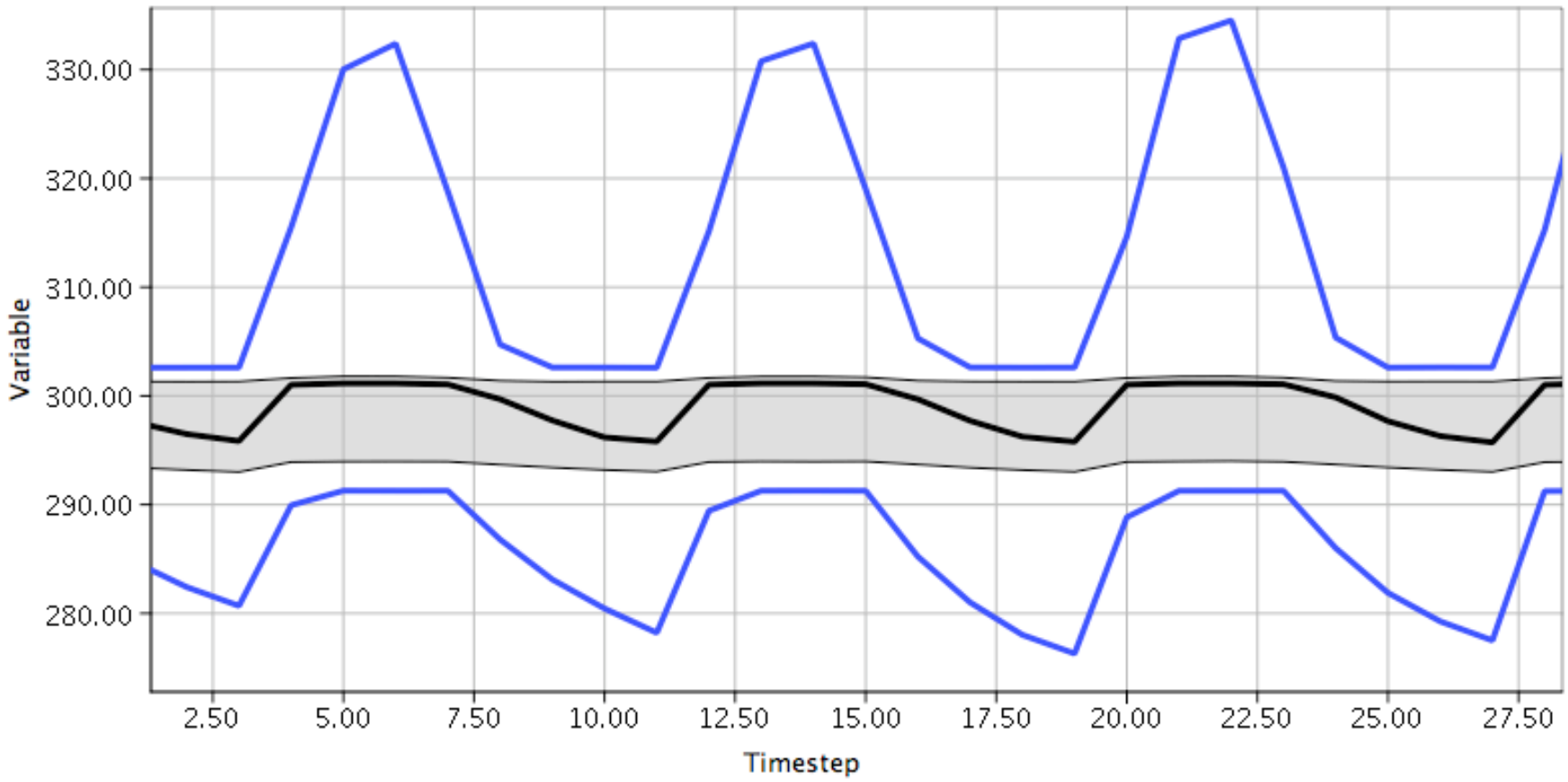
Filmstrip Display



- **Small multiples of summary display**
- **Show one variable over many timesteps at low resolution**
- **Cameras linked with other frames**
- **Selection linked with other frames and other displays**

Sandia National Laboratories

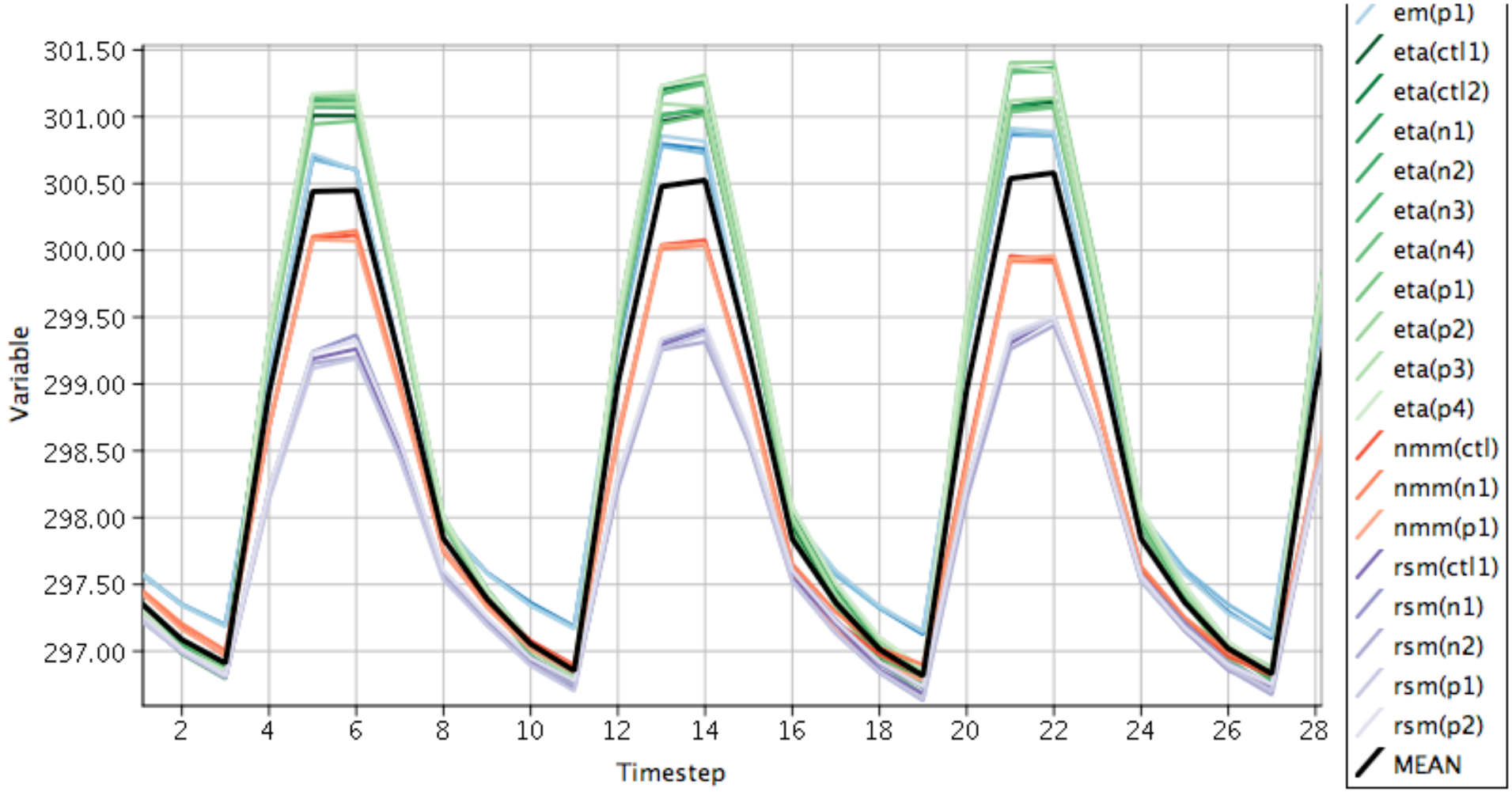
Quartile Chart



- **Shows quick overview of distribution and clustering of data values over time**

Sandia National Laboratories

Trend Chart



Sandia National Laboratories

Discussion

- **Exactly what should we display?**
  - How should we display it?
- **What statistics are appropriate?**
  - Do we have enough data to support them?
- **How can we indicate missing data?**
- **How do we store and access the data?**

Future Work

- **3D** (not as difficult as it appears)

…once that's done…

- **Better display metaphors**
- **Data fusion across different grids and time domains**

Sandia National Laboratories

Acknowledgements

- Kristi Potter, Univ. of Utah
- NOAA/NCEP
- Dean Dobranich, SNL
- Valerio Pascucci, Univ. of Utah
- Chris Johnson, Univ. of Utah
- Peer-Timo Bremer, LLNL

Sandia National Laboratories

Questions?

Sandia National Laboratories