大规模集合模拟数据的耦合场线分析
Coupled Ensemble Flow Line Advection and Analysis

Hanqi Guo\(^1,2\), Xiaoru Yuan\(^1,2\), Jian Huang\(^3\), and Xiaomin Zhu\(^4\)

1) Key Laboratory of Machine Perception (Ministry of Education), and School of EECS, Peking University
2) Center for Computational Science and Engineering, Peking University
3) Department of Electrical Engineering and Computer Science, University of Tennessee in Knoxville
4) National Super Computing Center in Jinan
MAJOR HURRICANE KATRINA (AL12)

NCEP GFS Ensemble track guidance valid 1800 UTC, 27 August 2005

Current Intensity: 100 kt
Current Basin: North Atlantic

Courtesy of Dr. Jonathan Vigh, Colorado State University, and NCEP
Background – Ensemble Run Data

Run 1  Run 2  Run 3
Background – Ensemble Run Data

• Facets of ensemble run data
  – Multivariate, Multi-valued, Time-varying
  – Huge and complex: TB/PB/EB scales

• Tasks in ensemble data visualization
  – Interactive exploration
  – Visualization of uncertainty
  – Comparison between runs (scalar/vector field)
Eulerian- and Lagrangian-Specifications

• Eulerian:  \( \mathbf{v} = \mathbf{v}(\mathbf{x}, t), \ p = p(\mathbf{x}, t), \ T = T(\mathbf{x}, t), \)

• Lagrangian:  \( \mathbf{X} = \mathbf{X}(\mathbf{a}, t), \ p = p(\mathbf{a}, t), \ T = T(\mathbf{a}, t), \)

• Relationships between two specifications:

\[
\mathbf{v}(\mathbf{X}(\mathbf{a}, t), t) = \frac{\partial \mathbf{X}(\mathbf{a}, t)}{\partial t}. \quad \Phi : \mathbf{x} \mapsto \Phi_{t_0 + t}(\mathbf{x})
\]
Lagrangian-based Distance Metric and Variation Measurement

- The flexible distance metric as the distance of flow maps
  \[ d_{x,t}(U, U') = \mu(U(\Phi_{t+t}^t(x)), U'(\Phi_{t+t}^t(x))), \tau \in [t, t + t_0] \]
  - \( \mu \) can be maximum distance, Hausdorff distance, etc.
  - U can be the location, some scalar quantities, etc.
- In our application, we use accumulated difference as \( \mu \)
  \[ d_{x,t}(U, U') = \int_t^{t+t_0} \| U(\Phi_{t+t}^t(x)) - U'(\Phi_{t+t}^t(x)) \|^2 d\tau \]
- The variation values of the ensemble run
  \[ \mathcal{V}(x, t) = \frac{1}{N(N-1)} \sum_{i<j} d_{x,t}(U_i, U_j) \]
The Timeline View

• A series of 1D MDS projections to show the trends of differences in global/local statistical region

\[
\min \sum_{i<j} (||x_i(t) - x_j(t)|| - D_{i,j}(t))^2
\]
Pipeline in Concept

- Ensemble data (large)
- Field line data (much larger than ensemble data)
- Variation field (small)
- Filtered lines (even smaller)
Parallel System Design

• **The goal**: accelerate the following computations
  – The massive field line tracing
  – Field line comparison, a.k.a variation computation

• **The major challenge**: extraordinary memory requirements for intermediate results
  – usually 1,000x times larger than raw ensemble data

• **The solution**: an improved DStep system with scalable data management
Briefs on Dstep

The Pipeline of the Parallel System (1)

• Both data scale and problem size are often too large to handle in practice
• A streamed data management mechanism is used to make the system scalable, given the memory limits
The Pipeline of the Parallel System (2)

• The step (map) stage: <seed, partial_line>
  – Field line advection

• The reduce stage: <seed, partial_lines_all_runs[]>
  – Line merge
  – Line re-sampling
  – Variation comparison

• The scalable data management
  – Scalable I/O (BIL, parallel-netCDF/HDF5, MPI-IO, etc.)
  – Batch streaming of new queries
Benchmark Platform: NCSSJN

• ShenWei-based supercomputer
  – SW1600 processor, 1.0~1.1GHz
  – 1GB memory for each core
  – 40Gbps high-speed interconnection

• x86-based supercomputer
  – Intel Xeon E5675 hexa-core processor, 3.06GHz
  – 4GB memory for each core
  – QDR Infiniband interconnection

• Shared global filesystem: SWGFS
Scalability

• Strong scalability test in National Super Computer Center in Jinan (ShenWei and x86 architectures)
Applications

• GEOS-5 global climate model from NASA Goddard Space Flight Center
  – 8 run ensemble simulation
  – 1°x1.5° resolution, with 72 pressure levels
  – From Jan 2000 to Dec 2011

• WRF ARW model
  – base and no_urban runs to investigate the impact of urbanization
  – 100x100x27 resolution, East China
  – From 2012-7-1 00:00:00 UTC to 2012-7-10 18:00:00 UTC
WRF Simulation
Contemporary Systems

• Recent advances on parallel particle tracing
  – The scalability in our work is different due to the extraordinary memory use because of the nature of Lagrangian-based metric

<table>
<thead>
<tr>
<th>Method</th>
<th>Data Size</th>
<th># of particles</th>
<th># of processes</th>
<th>Machine</th>
<th>Year</th>
</tr>
</thead>
</table>
Conclusions

• We propose a novel approach to extract features as differences in ensemble data with Lagrangian-based distance metric
• A parallel system called eFLAA is required to support the scalable analysis
Acknowledgements

• Funding
  – NSFC Project No. 61170204
  – NSFC Key Project No. 61232012
  – “Strategic Priority Research Program - Climate Change: Carbon Budget and Relevant Issues” of the Chinese Academy of Sciences, Grant No. XDA05040205.

• People
  – Dr. Junfeng Liu (College of Urban and Environmental Science)
  – Dr. Xiaoguang Ma (IAP, CAS)