HPC and Big Data: COLA’s Experience in the Advanced Scientific Discovery Program

Jim Kinter, COLA

Computing in the Atmospheric Sciences
Annecy, France 12 September 2013
COLA News

COLA Moves to GMU Fairfax Campus

We are pleased to announce that the Center for Ocean-Land-Atmosphere Studies (COLA) will become an integral part of George Mason University (GMU) in 2013-14. COLA staff and the COLA computing facility will be collocated with Climate Dynamics faculty and students and the AOES Department on the GMU main campus in Fairfax, Virginia.

James L. Kinter III,
Director, COLA

Barry A. Klinger,
Graduate coordinator, AOES

Jagadish Shukla,
Director, Climate Dynamics Program

David M. Straus,
Chair, AOES

GMU Research Hall – Home of AOES and COLA.

CAS2K13 September 2013 – Jim Kinter
GMU Ph.D. Program in Climate Dynamics

Affiliated with the Department of Atmospheric, Oceanic, and Earth Sciences

Faculty

• T. DelSole; Ph.D., Harvard Univ.
• P. Dirmeyer; Ph.D., Univ. of Maryland
• E. Jin; Ph.D., Seoul National Univ.
• B. Huang; Ph.D., Univ. of Maryland
• V. Krishnamurthy; Ph.D., M.I.T.
• J. Lu; Ph.D., Dalhousie Univ.
• J. Kinter; Ph.D., Princeton Univ.

B. Klinger; Ph.D., M.I.T./Woods Hole Ocean. Inst.
E. Schneider; Ph.D., Harvard Univ.
P. Schopf; Ph.D., Princeton Univ.
J. Shukla (director); Ph.D., B.H.U.; Sc.D., M.I.T
C. Stan; Ph.D., Colorado State Univ.
D. Straus (chair, AOES); Ph.D., Cornell Univ.
5 Myths About Big Data

1. There is a clear definition of Big Data. (I don’t know what it is, but I’ve got it!)

2. Big Data is new.
   - Science has been using Big Data for a long time, e.g., Kepler “mining” the obs of Brahe.
   - Statisticians: Big Data = Statistics, albeit sexier, more broadly applied

3. Big Data is revolutionary.
   - More likely to have modest, gradual impact.
   - Large effects are easy to recognize (small data), but handling subtleties require Big Data

4. Bigger data is better.
   - Big data sets are hard to work with, even using automated methods.
   - Bias can still be present in big data sets.

5. Big Data means the end of science.
   - Can’t go fishing for correlations and explain the world, e.g., spurious correlations or conflated cause and effect
   - Still need hypotheses, ideas and theories: “If you don’t ask good questions, your results can be silly and meaningless”

“Having more data won’t substitute for thinking hard, recognizing anomalies, and exploring deep truths.”

Predictability* of the Physical Climate System

Overarching Scientific Questions

What limits predictability at all time scales from days to decades? Is there a fundamental limit? What is the role of model error? Initial conditions error?

How do the initial state, the coupling of system components, and the changes in external forcing contribute to predictability at different time scales?

What aspects of the total climate system (troposphere, stratosphere, world oceans, land surface, vegetation, sea ice, land ice, snow) are predictable in which geographic regions, for which seasons, and how does that change in the future? For the current and future generation of climate models and observing systems?

What is the optimal combination of models to predict means? Extremes? No current models are perfect, e.g. for regional water cycle

* Note: Predictability is a necessary (but not sufficient) condition for attribution
Why does climate research need HPC and Big Data?

• Societal demand for information about weather-in-climate and climate impacts on weather
• Seamless days-to-decades prediction & unified weather/climate modeling
• Multi-model ensembles and Earth system prediction
• Requirements for data assimilation
Driver: Societal Demand for Climate Information

- America’s Climate Choices
- (USGCRP) National Climate Assessment
- Intergovernmental Panel on Climate Change
Regional Climate Change – Beyond CMIP3 Models’ Ability?
Driver: Seamless Prediction, Unified Modeling

• **Seamless prediction**: Viewing weather and climate prediction as initial-boundary value problems that share common processes and dynamics and that can be addressed using unified models with common methods across a broad range of time scales and spatial resolutions.

* Note: Prediction implies starting from an observed initial state, which in turn implies data assimilation
Driver: Multi-Model Ensembles & Total Climate System Prediction

Total Climate System – Earth System

Physical Climate System
- Atmospheric Physics/Dynamics
- Ocean Dynamics
- Terrestrial Energy/Moisture
- Global Moisture
- Soil
- CO2
- Marine Biogeochemistry
- Terrestrial Ecosystems
- Tropospheric Chemistry
- Biogeochemical Cycles

Climate Change

External Forcing
- Sun
- Volcanoes

NMME Forecast for Niño 3.4 IC=201308

CO2

CO2

Pollutants

Human Activities
Balancing Demands on Resources

Computing Resources

Resolution

Data Assimilation

Complexity

Duration and/or Ensemble size

CAS2K13 September 2013 – Jim Kinter
HPC & Big Data at COLA

Representative Projects

**Project Athena**: An International, Dedicated High-End Computing Project to Revolutionize Climate Modeling (Dedicated XT4 at NICS)

→ Update on CAS2K11 briefing from Martin Miller

**Project Minerva**: Exploring High Spatial Resolution for Seasonal Climate Prediction (Dedicated Advanced Scientific Discovery on NCAR Yellowstone)

**PetaApps Team**: Climate Models’ Representation of Unpredictable Noise in the Atmosphere, Ocean or Sea Ice (TeraGrid Ranger and Kraken)
Origins of *Project Athena*

- 2008 World Modeling Summit: dedicate petascale supercomputers to climate modeling
- U.S. National Science Foundation offered to dedicate the Athena supercomputer for 6 months in 2009-2010 as a pilot study
- An international collaboration (*Project Athena*) was formed by groups in the U.S., Japan and the U.K. to use Athena to take up the challenge
Project Athena

Collaborating Groups

COLA - Center for Ocean-Land-Atmosphere Studies, USA (NSF-funded)
ECMWF - European Center for Medium-range Weather Forecasts, UK
JAMSTEC - Japan Agency for Marine-Earth Science and Technology, Research Institute for Global Change, Japan
University of Tokyo, Japan
NICS - National Institute for Computational Sciences, USA (NSF-funded)
Cray Inc.

Codes

NICAM: Nonhydrostatic Icosahedral Atmospheric Model
IFS: ECMWF Integrated Forecast System
NICS Resources for Project Athena

• The Cray XT4 – **Athena** – the first NICS machine in 2008
  – 4512 nodes w/ AMD 2.3 GHz quad-core CPUs + 4 GB RAM
  – **18,048 cores** + 17.6 TB aggregate memory
  – **165 TFLOPS peak** performance
  – Dedicated to this project during October 2009 – March 2010 \(\rightarrow\) 72 million core-hours!

• Other resources made available to project:
  – **85 TB Lustre file system**
  – **258 TB auxilliary Lustre file system** (called **Nakji**)
  – **Verne**: 16-core 128-GB system (data analysis) during production phase (2009-2010)
  – **Nautilus**: SGI UV with **1024 Nehelem EX cores**, 8 GPUs, 4 TB memory, 960 TB GPFS disk (data analysis) in 2010-11

Many thanks to NICS for resources and sustained support!
# Project Athena Experiments

<table>
<thead>
<tr>
<th>Model/Exp.</th>
<th>Resolution</th>
<th># Cases</th>
<th>Period</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>NICAM / Hindcasts</td>
<td>7 km</td>
<td>8</td>
<td>103 days</td>
<td>21 May - 30 Aug 2001 - 2009</td>
</tr>
<tr>
<td>IFS / Hindcasts</td>
<td>125 km</td>
<td>39 km</td>
<td>48</td>
<td>395 days</td>
</tr>
<tr>
<td></td>
<td>16 km</td>
<td></td>
<td></td>
<td>1 Nov - 30 Nov (following year) 1960 - 2007</td>
</tr>
<tr>
<td></td>
<td>10 km</td>
<td>20</td>
<td></td>
<td>1 Nov - 30 Nov (following year) 1989 - 2007</td>
</tr>
<tr>
<td>IFS / Hindcasts</td>
<td>125 km</td>
<td>39 km</td>
<td>9</td>
<td>103 days</td>
</tr>
<tr>
<td></td>
<td>16 km</td>
<td></td>
<td></td>
<td>21 May - 30 Aug 2001 - 2009 NICAM analogs</td>
</tr>
<tr>
<td></td>
<td>10 km</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFS / Summer</td>
<td>39 km</td>
<td>6</td>
<td>132 days</td>
<td>21 May - 30 Sep selected years</td>
</tr>
<tr>
<td>Ensembles</td>
<td>16 km</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFS / Winter</td>
<td>39 km</td>
<td>6</td>
<td>151 days</td>
<td>1 Nov - 31 Mar selected years</td>
</tr>
<tr>
<td>Ensembles</td>
<td>16 km</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFS / AMIP</td>
<td>39 km</td>
<td>1</td>
<td>47 years</td>
<td>1961 - 2007</td>
</tr>
<tr>
<td></td>
<td>16 km</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IFS / Time Slice</td>
<td>39 km</td>
<td>1</td>
<td>47 years</td>
<td>2071 - 2117</td>
</tr>
<tr>
<td></td>
<td>16 km</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

http://wxmaps.org/athena/home/
Project Athena Publications

Sample Results

- Basics of model climate
- South Asian monsoon
- Resolution dependence of snow
- Diurnal cycle of precipitation
- Projection of climate change
- Tropical cyclones
- Tornadoes in climate simulation
Europe Growing Season (Apr-Oct)
Precipitation Change: 20\textsuperscript{th} C to 21\textsuperscript{st} C

"Time-slice" runs of the ECMWF IFS global atmospheric model with observed SST for the 20\textsuperscript{th} century and CMIP3 projections of SST for the 21\textsuperscript{st} century at two different model resolutions.

The continental-scale pattern of precipitation change in April – October (growing season) associated with global warming is similar, but the regional details are quite different, particularly in southern Europe.
Future Change in Extreme Summer Drought Late 20\textsuperscript{th} C to Late 21\textsuperscript{st} C

10\textsuperscript{th} Percentile Drought: Number of years out of 47 in a simulation of future climate (2071-2117) for which the June-August mean rainfall was less than the 5\textsuperscript{th} driest year of 47 in a simulation of current climate (1961-2007).
Sample Results

• Basics of model climate
• South Asian monsoon
• Resolution dependence of snow
• Diurnal cycle of precipitation
• Projection of climate change
• **Tropical cyclones**
• Tornadoes in climate simulation
Athena – Clouds and Precipitation

Boreal Summer 2009
Brian Doty
COLA
Project Athena: Summary

- **Good news**: Extreme spatial resolution improves many of the qualitative features of large-scale climate simulation
- **As expected**: High spatial resolution provides higher fidelity representation of features sensitive to orography or geography
- **Unexpected**: Nonlinear dynamical effects can alter simulation changes due to spatial resolution improvements much more and possibly in different ways than we might have expected
- **Bad news** (as expected?): Large biases remain in hard-to-simulate fields like tropical precipitation → still need to understand and properly represent the effects of subgrid-scale physical processes
Project Minerva

• Explore the impact of increased atmospheric resolution on model fidelity and prediction skill in a coupled, seamless framework by using a state-of-the-art coupled operational long-range prediction system to systematically evaluate the prediction skill and reliability of a robust set of hindcast ensembles at low, medium and high atmospheric resolutions

• NCAR Advanced Scientific Discovery Program to inaugurate Yellowstone (72 K-core IBM iDataPlex)

• Allocated 21 M core-hours on Yellowstone

• Used ~28 M core-hours (Our jobs squeaked in under core size that “broke” the system)

Many thanks to NCAR for resources and sustained support!
Project Minerva

ECMWF team:
- Frederic Vitart, lead
- Roberto Buizza
- Erland Kallen
- Franco Molteni
- Tim Stockdale
- Peter Towers
- Nils Wedi

COLA team:
- Ben Cash, lead
- Rondro Barimalala
- Paul Dirmeyer
- Mike Fennessy
- V. Krishnamurthy
- Julia Manganello
- David Straus

University of Oxford:
- Tim Palmer
- Andrew Dawson
<table>
<thead>
<tr>
<th>System</th>
<th>Atmosphere model cycle</th>
<th>Atmosphere spectral truncation</th>
<th>Atmosphere vertical levels</th>
<th>Ocean model</th>
<th>Ocean horizontal res, equatorial refinement</th>
<th>Ocean vertical levels</th>
</tr>
</thead>
<tbody>
<tr>
<td>MINERVA</td>
<td>IFS cy 38r1</td>
<td>T319 / T639 / T1279</td>
<td>91 levels, top = 1 Pa</td>
<td>NEMO v 3.0/3.1</td>
<td>1 degree, ~ 0.3 deg. Lat</td>
<td>42 levels</td>
</tr>
<tr>
<td>System 4</td>
<td>IFS cy 36r4</td>
<td>T255</td>
<td>91 levels, top = 1 Pa</td>
<td>NEMO v 3.0/3.1</td>
<td>1 degree, ~ 0.3 deg. Lat</td>
<td>42 levels</td>
</tr>
<tr>
<td>ENS (current)</td>
<td>IFS cy 38r2</td>
<td>T639 (d 0-10), T319</td>
<td>62 levels, top = 5 hPa</td>
<td>NEMO v 3.0/3.1</td>
<td>1 degree, ~ 0.3 deg. Lat</td>
<td>42 levels</td>
</tr>
<tr>
<td>ENS (end 2013)</td>
<td>IFS cy 40r1</td>
<td>T639 (d 0-10), T319</td>
<td>91 levels, top = 1 Pa</td>
<td>NEMO v 3.4</td>
<td>1 degree, ~ 0.3 deg. Lat</td>
<td>42 levels</td>
</tr>
</tbody>
</table>

**System 4**: Operational seasonal prediction system  
**ENS**: Operational medium-range/monthly prediction system

*Courtesy Franco Molteni & Frederic Vitart, ECMWF*
## ECMWF Coupled Ensemble Systems

<table>
<thead>
<tr>
<th>System</th>
<th>Coupler</th>
<th>Time range of ocean-atmosphere coupling</th>
<th>Coupling frequency</th>
<th>Unperturbed initial cond. for re-forecasts</th>
<th>Atmospheric perturbations</th>
<th>Ocean perturbations</th>
<th>Stochastic model perturbations</th>
</tr>
</thead>
<tbody>
<tr>
<td>MINERVA</td>
<td>OASIS-3</td>
<td>from start</td>
<td>3 hours</td>
<td>ERA-Interim + ORA-S4</td>
<td>SV, EDA from 2011 dates</td>
<td>5 ocean analyses + SST perturbations</td>
<td>3-timescale SPPT + KE backscatter</td>
</tr>
<tr>
<td>System 4</td>
<td>OASIS-3</td>
<td>from start</td>
<td>3 hours</td>
<td>ERA-Interim + ORA-S4</td>
<td>SV</td>
<td>5 ocean analyses + SST perturbations</td>
<td>3-timescale SPPT + KE backscatter</td>
</tr>
<tr>
<td>ENS (current)</td>
<td>OASIS-3</td>
<td>from day 10</td>
<td>3 hours</td>
<td>ERA-Interim + ORA-S4</td>
<td>SV, EDA from current or recent date</td>
<td>generated by ENS member fluxes during day 1 to 10</td>
<td>2-timescale SPPT + KE backscatter</td>
</tr>
<tr>
<td>ENS (end 2013)</td>
<td>sequential, single executable</td>
<td>from start</td>
<td>3 hours</td>
<td>ERA-Interim + ORA-S4</td>
<td>SV, EDA from current or recent date</td>
<td>5 ocean analyses</td>
<td>2-timescale SPPT + KE backscatter</td>
</tr>
</tbody>
</table>

**ORA-S4**: Ocean Re-Analysis for ECMWF System-4  
**SV**: Singular Vectors of 48-hour linear propagator  
**EDA**: Ensemble of Data Assimilations (low-res 4D-var)  
**SPPT**: Stochastic Perturbation of Physical Tendencies scheme  

Courtesy Franco Molteni & Frederic Vitart, ECMWF
## Minerva Prediction Experiments

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Years</th>
<th>Ens. Size</th>
<th>Initial Months</th>
<th>Duration (mon)</th>
</tr>
</thead>
<tbody>
<tr>
<td>T319_base</td>
<td>1980-2011</td>
<td>51</td>
<td>May, Nov</td>
<td>7</td>
</tr>
<tr>
<td>T319_2_year_extension</td>
<td>1980-2011</td>
<td>15</td>
<td>May</td>
<td>24</td>
</tr>
<tr>
<td>T639_base</td>
<td>1980-2011</td>
<td>15</td>
<td>May, Nov</td>
<td>7</td>
</tr>
</tbody>
</table>
| T639_extended_ensemble    | 1980-2011     | 36        | May, Nov       | May: 5 mo
                |               |           | Nov: 4 mo      |                |
| T639_2_year_extension     | 1980-2011     | 15        | Nov            | 24             |
| T1279_base                | 2000-2011     | 15        | May            | 7              |
Minerva: Coupled Prediction of Tropical Cyclones

11-12 June 2005 hurricane off west coast of Mexico: precipitation in mm/day every 3 hours (T1279 coupled forecast initialized on 1 May 2005)

The predicted maximum rainfall rate reaches 725 mm/day (30 mm/hr)

Based TRMM global TC rainfall observations (1998-2000), the frequency of rainfall rates exceeding 30 mm/hr is roughly 1%
Minerva vs. Athena – TC Frequency (NH; JJASON; T1279)

9-Year Mean (2000-2008)

<table>
<thead>
<tr>
<th></th>
<th>OBS</th>
<th>Athena</th>
<th>Minerva</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>49.9</td>
<td>59.1</td>
<td>48.9</td>
</tr>
</tbody>
</table>

IBTrACS (red), Athena (green), Minerva (black)
Individual Forecast Anomalies

MAY 2013

IC: 01 May 2013
Ens: 2 members
May Mean

Precipitation

Soil Wetness
Individual Forecast Anomalies

MAY-JULY 2013

IC: 01 May 2013
Ens: 2 members
May-July Mean

Precipitation

Soil Wetness

CAS2K13 September 2013 – Jim Kinter
Peta-Apps Team (2010-2012)

- Kinter, COLA (PI)
- Collins, UC Berkeley (co-PI)
- Kirtman, U. Miami (co-PI)
- Loft, NCAR (co-PI)
- Yelick, LBL (co-PI)
- Ahearn, NCAR
- Bitz, U. Washington
- Bryan, NCAR
- Dennis, NCAR
- Min, U. Miami
- Nolan, UC Berkeley
- Siquiera, U. Miami
- Stan, COLA

Many thanks to TACC and NICS for resources and sustained support!
Correlation
High-Pass
SST vs. \(|V_{srf}|\)

Bryan et al. 2010 J. Climate

0.5° atm
1.0° ocn

Satellite obs

0.5° atm
0.1° ocn
PetaApps: Rainfall Simulation

Rainfall: HRC (shades) and LRC (contours)

Observed Rainfall
Challenges and Tensions

- Making effective use of large allocations – takes a village
- Exaflood of data
- Resolution vs. parameterization
- Sampling (e.g. extreme events)
- Climate scientists are being forced to think about data & code issues

TENSIONS

| HPC capability | Automation/abstraction | Data-driven development | Small, portable code | Tight, local control of data | Data analysis capacity | Human control | Science-driven development | End-to-end tools | Distributed data |

"Having more data won’t substitute for thinking hard, recognizing anomalies, and exploring deep truths.”

Exaflood: Challenge and Opportunity

• In January 2007, Bret Swanson of the Discovery Institute coined the term **exaflood** for the **impending flood of exabytes** that would cause the Internet's congestive collapse.

• Hay et al., 2010: *The Fourth Paradigm* →
Data Volumes

• **Project Athena:** Total data volume
  Spinning disk
  
  1.2 PB (~500 TB unique)*
  40 TB at COLA
  0 TB at NICS (was 340 TB)

* no home after April 2014

• **Project Minerva:** Total data volume
  Spinning disk
  
  0.9 PB (~700 TB unique)
  100 TB at COLA
  500TB at NCAR (for now)

• That much data breaks everything: H/W, systems management policies, networks, apps S/W, tools, and shared archive space

• **NB:** Generating 700 TB using 28 M core-hours took ~3 months; this would take about a day on a system with 1M cores!
Athena and Minerva: Harbingers of the Exaflood

- Familiar diagnostics are hard to do at very high resolution
- Have we wrung all the “science” out of the data sets, given that we can only keep a small percentage of the total data volume on spinning disk? How can we tell?
- Must move from ad hoc solutions → systematic, repeatable solutions (transform Noah’s Ark → a Shipping Industry)
- “We need exaflood insurance.”
  - Jennifer Adams