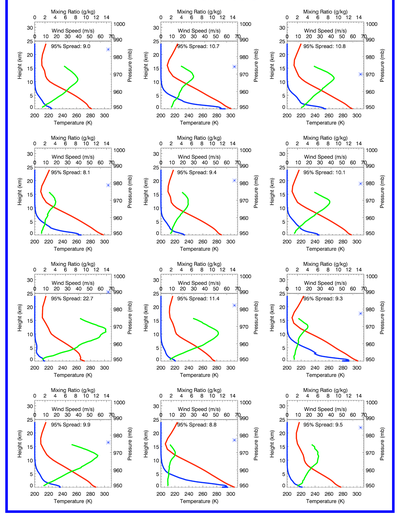
A Cluster Analysis Approach to Comparing Atmospheric Radiation Measurement (ARM) Data with Global Climate Model (GCM) Results

Forrest M. Hoffman[1], Salil Mahajan[2], William W. Hargrove[3], Richard T. Mills[1], and Tony Del Genio[4]
[1]Oak Ridge National Laboratory, [2]Texas A&M University, [3]USDA Forest Service, [4]NASA GISS

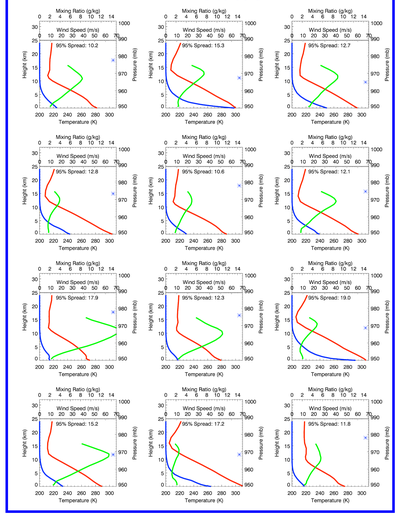
Introduction

Cluster analysis was employed to compare ARM observational data at the Southern Great Plains (SGP) site with corresponding 6-hourly output from an integration of the Community Climate System Model Version 3 (CCSM3) run under the IPCC SRES A2 scenario for the current decade. Cluster analysis is a technique for classifying multivariate data into distinct regimes or states based on Euclidean distance in a phase space formed from the variables under consideration. A three-way process was used for the comparison: 1) CCSM output was projected onto states derived from ARM observations, 2) ARM observations were projected onto states derived from CCSM output, and 3) both ARM observations and CCSM output were projected onto states derived from the combination of the two datasets. A parallel clustering algorithm developed at ORNL has been improved by adding an acceleration technique and a method for handling empty clusters. Both serve to significantly reduce the time-to-solution. In addition, a parallel principal components analysis (PCA) tool has been developed to reduce the dimensionality of the analysis phase space while preserving most of the variance contained in the data.

ARM Observations Projected onto ARM States



CCSM Results Projected onto ARM States



ARM Observational Data

For this study, we used the AERIPROF3FELTZ Value Added Products (VAPs) for the SGP site for the time period April 2002–April 2007. Derived from observations from the Atmospherically Emitted Radiance Interferometer (AERI) instrument, the data used were temperature and water vapor mixing ratio vertical profiles at 48 levels in the atmosphere. In addition, wind speed at 62 levels from the NOAA wind profiler and surface pressure, both from the WPDNMET.X1.b1 VAP, we used in the analysis. ARM observations were averaged over 6 hours to correspond to CCSM results.

CCSM Model Results

Corresponding data over the SGP site were extracted from eight years of CCSM output from a single ensemble member performing an integration with the IPCC SRES A2 scenario for the 21st century. This particular ensemble member had a large number of atmospheric fields saved from the Community Atmosphere Model (CAM) as 6-hourly averages at a spatial resolution of about 1.4°×1.4°. CCSM model results were interpolated onto the vertical levels of ARM observations to facilitate comparison.

Cluster Analysis Methodology

A high performance parallel cluster analysis tool, developed by Hoffman and Hargrove at ORNL, employing an iterative *k*-means clustering algorithm was used to group multivariate atmospheric column data comprised of 159 factors into 12 distinct atmospheric states. We applied a three way methodology to comparing ARM observations with CCSM output as follows:

1) ARM observations were clustered into 12 states, and CCSM results were projected onto those states;
2) CCSM results were clustered into 12 states, and ARM observations were projected onto those states; and
3) ARM and CCSM data were combined and clustered into 12 states, and both sets of data were projected onto those states.

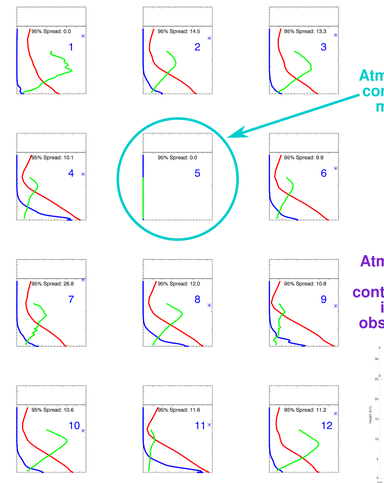
References

Mahajan, S., F. M. Hoffman, W. W. Hargrove, S. W. Christensen, and R. T. Mills. December 2007. "A Cluster Analysis Approach to Comparing Atmospheric Radiation Measurement (ARM) Data and Global Climate Model (GCM) Results." *Eos Trans. AGU*, 88(52), Fall Meet. Suppl., Abstract A41A-0010.

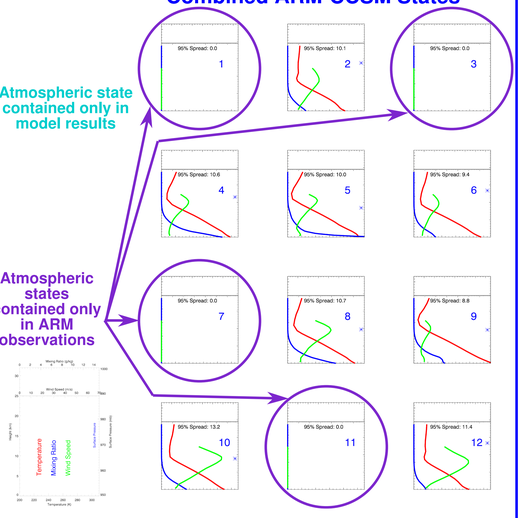
Hoffman, F. M., W. W. Hargrove, D. J. Erickson, and R. J. Oglesby. August 3, 2005. "Using Clustered Climate Regimes to Analyze and Compare Predictions from Fully Coupled General Circulation Models." *Earth Interactions*, 9(10): 1-27. doi:10.1175/EI110.1.

Hargrove, W. W. and F. M. Hoffman. 2004. "Potential of Multivariate Quantitative Methods for Delineation and Visualization of Ecoregions." *Environmental Management*, 34(5): s39-s60. doi:10.1007/s00267-003-1084-0.

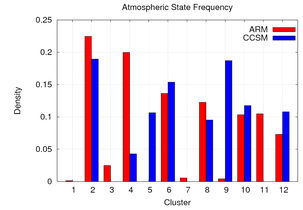
ARM Observations Projected onto Combined ARM-CCSM States



CCSM Results Projected onto Combined ARM-CCSM States

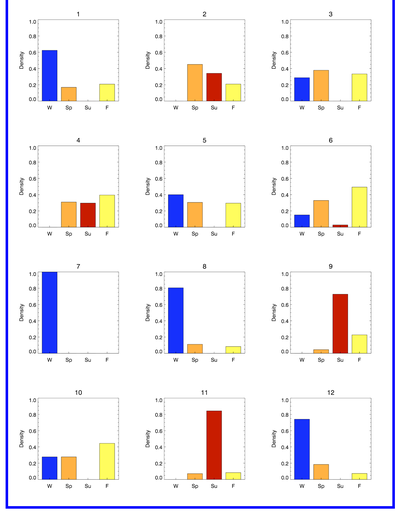


Comparisons of 12 atmospheric states derived from the combination of ARM observations and CCSM results indicate that distinct singular states exist in each dataset. State #5 is characterized by very high humidity and temperature at the surface, and it has no analog in the ARM observations. States #1, #3, and #7 have very low frequency in the observations (see plot at right), so their absence from model predictions does not suggest a serious problem. However, state #11, which is characterized by high humidity and temperature with very low wind shear, is never predicted by CCSM. In addition, CCSM predicts an over-abundance of state #9 (low humidity and high temperature conditions) while under-representing state #4 (moderate humidity, temperature, and shear conditions). Misrepresentation of atmsopheric states in CCSM over the SGP site could have impacts on predictions of cloud formation and hence the local radiation budget.

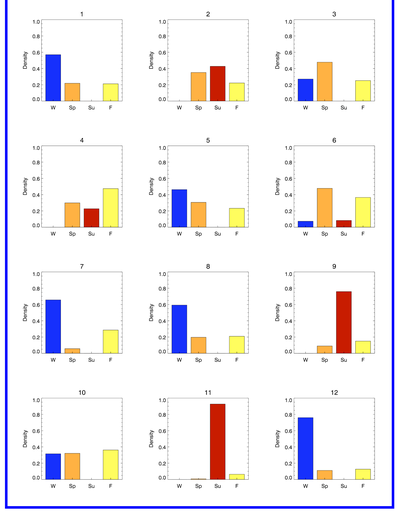


See it on the Web at
http://www.climatemodeling.org/arm

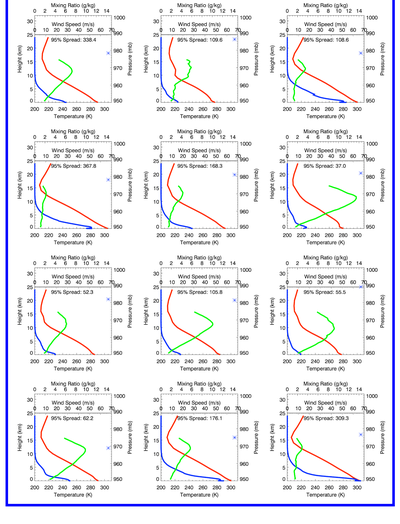
Seasonal Distribution of ARM Observations on ARM States



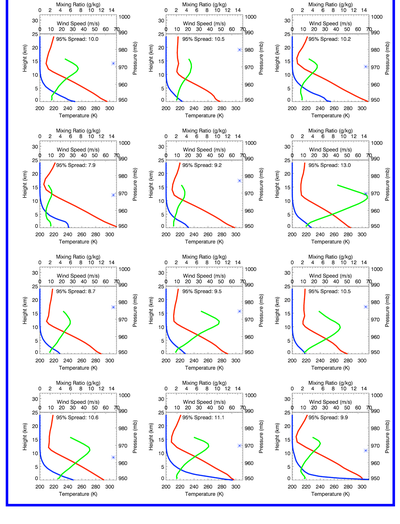
Seasonal Distribution of CCSM Results on ARM States



ARM Observations Projected onto CCSM States



CCSM Results Projected onto CCSM States

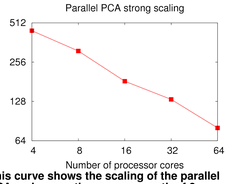


Model-Observation Intercomparison as a Data Mining Problem

The wealth of atmospheric observational and model data sets are posing new challenges for analysis and intercomparison. Data mining techniques, like cluster analysis, are proving useful for extracting patterns from the long time series data that ARM collects and that global models produce. Tools for performing these kinds of analysis need to be just as scalable and extensible as the models themselves in order to handle these very large, long time series data.

Our recent software engineering efforts have included 1) implementing an acceleration technique that uses the triangle inequality to reduce the number of comparisions which must be performed in clustering iterations, 2) implementing a method for handling centroids that lose all cluster membership in an iteration (called "empty clusters"), and 3) developing a parallel and scalable principal components analysis (PCA) code.

No publicly available tool can perform PCA on data sets as large as those we use. Our new PCA tool can operate on these very large data sets and effectively capture the variance in the data using a reduced number of dimensions, both speeding up cluster analysis and allowing for analysis of larger datasets. Parallelism is exploited in the reduction of the data matrix to bidiagonal form via Householder reflections. The parallel PCA tool can process massive data sets: 1 year of 6-hourly global 3D atmosphere data (48M cells × 131 variables), a 48 GB file, can be processed in only 1533 seconds on 40 nodes of Cray XT4 at ORNL.



This curve shows the scaling of the parallel PCA code operating on one month of 6-hourly data on the Cray XT4.