Title

Applying Model-Based Clustering for Analysis of Community Atmospheric Model Output

Permalink

https://escholarship.org/uc/item/062248fn

Author

Chen, Wei-Chen

Publication Date

2012-02-29

Applying Model-Based Clustering for Analysis of Community Atmospheric Model Output

Wei-Chen Chen[1], George Ostrouchov[1], David Pugmire[1], Mr Prabhat[2], and Michael Wehner[2]
1: Oak Ridge National Laboratory, Oak Ridge, TN.
2: Lawrence Berkeley National Laboratory, Berkeley, CA.

DISCLAIMER: This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor the Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or the Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or the Regents of the University of California.

Acknowledgements:

This work is partially supported by the Director, Office of Laboratory Policy and Infrastructure Management of the U.S. Department of Energy under Contract No. AC02-05CH11231. The funding for this work was provided by US DOE BER program.

Applying Model-Based Clustering for Analysis of Community Atmospheric Model Output

Wei-Chen Chen[1], George Ostrouchov[1], David Pugmire[1], Mr Prabhat[2], and Michael Wehner[2]

1: Oak Ridge National Laboratory, Oak Ridge, TN. 2: Lawrence Berkeley National Laboratory, Berkeley, CA.

Introduction

Modern Community atmosphere models (CAM5) run at high spatial and temporal resolution, effectively producing hundreds of terabytes of climate output. The size of these datasets impedes application of advanced statistical analysis such as model-based clustering, which can be used for extracting data features based on finite mixture models. We develop a parallel version of expectation and maximization (EM) algorithm to enable model-based clustering on such large datasets. We apply the technique to CAM5 output, isolate clusters corresponding to weather systems, and use a variety of visualization techniques to display the clusters and the correlations between selected climate variables.

Dataset

CAM5 produces about 200 terabytes of simulating weather and climate from 1979 to 2004. Data selected for clustering analysis include extreme precipitations and relevant covariates (Q, T, and OMEGA) which are high dimensional multivariate data consisting of daily averaged variables in 3D physical space.

Parallel Model-Based Clustering

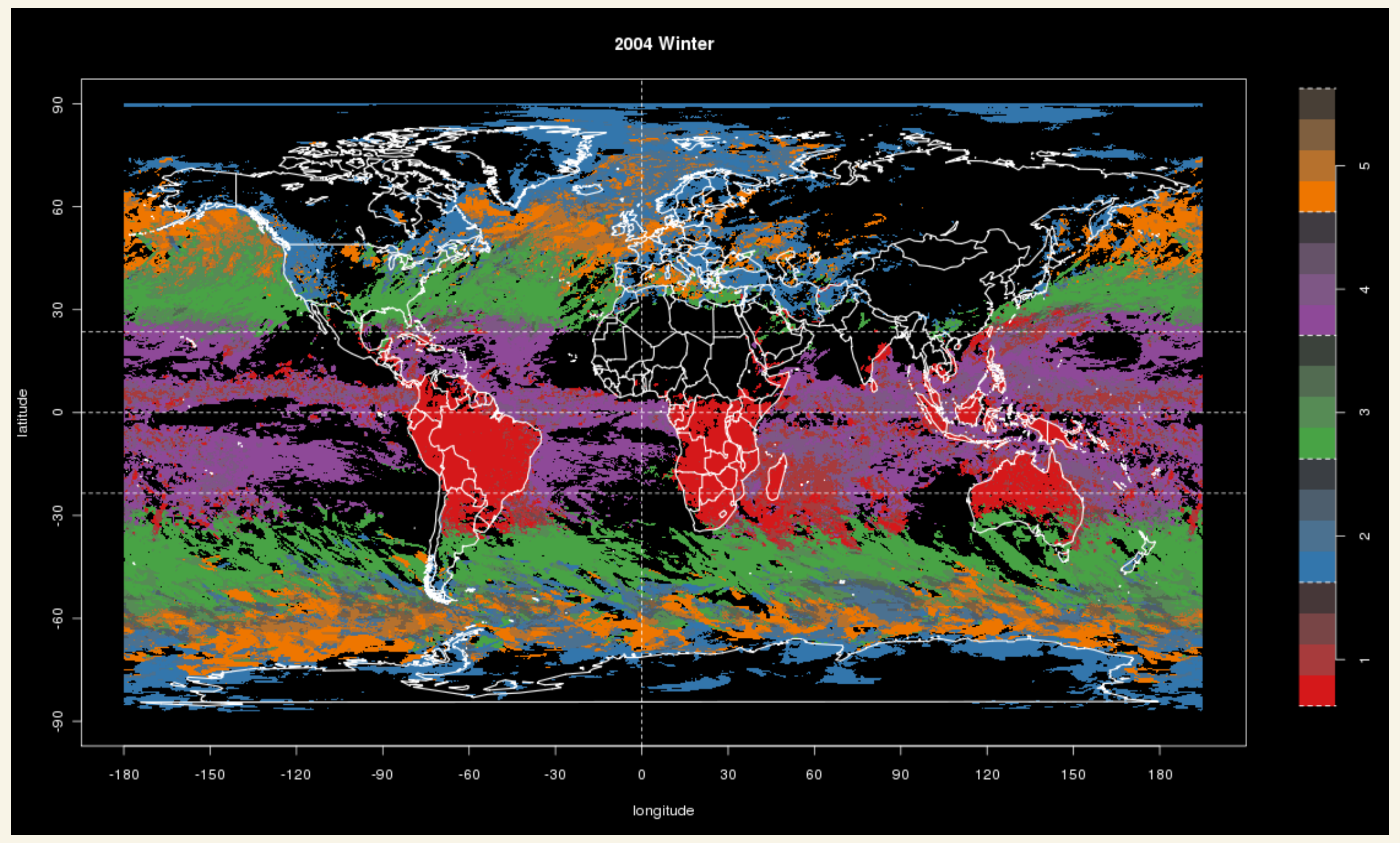
The model-based clustering utilizes finite mixture modes and assumes multivariate Gaussian distributions with unstructured dispersions for each component. The model parameters are estimated by parallel EM algorithms which are implemented in a released R package `pmclust`. In order to deal with large distributed datasets, the package is designed in a single program multiple data (SPMD) programming model. Several efficient EM-like algorithms are also implemented in the package to reduce computing time and convergent iterations.

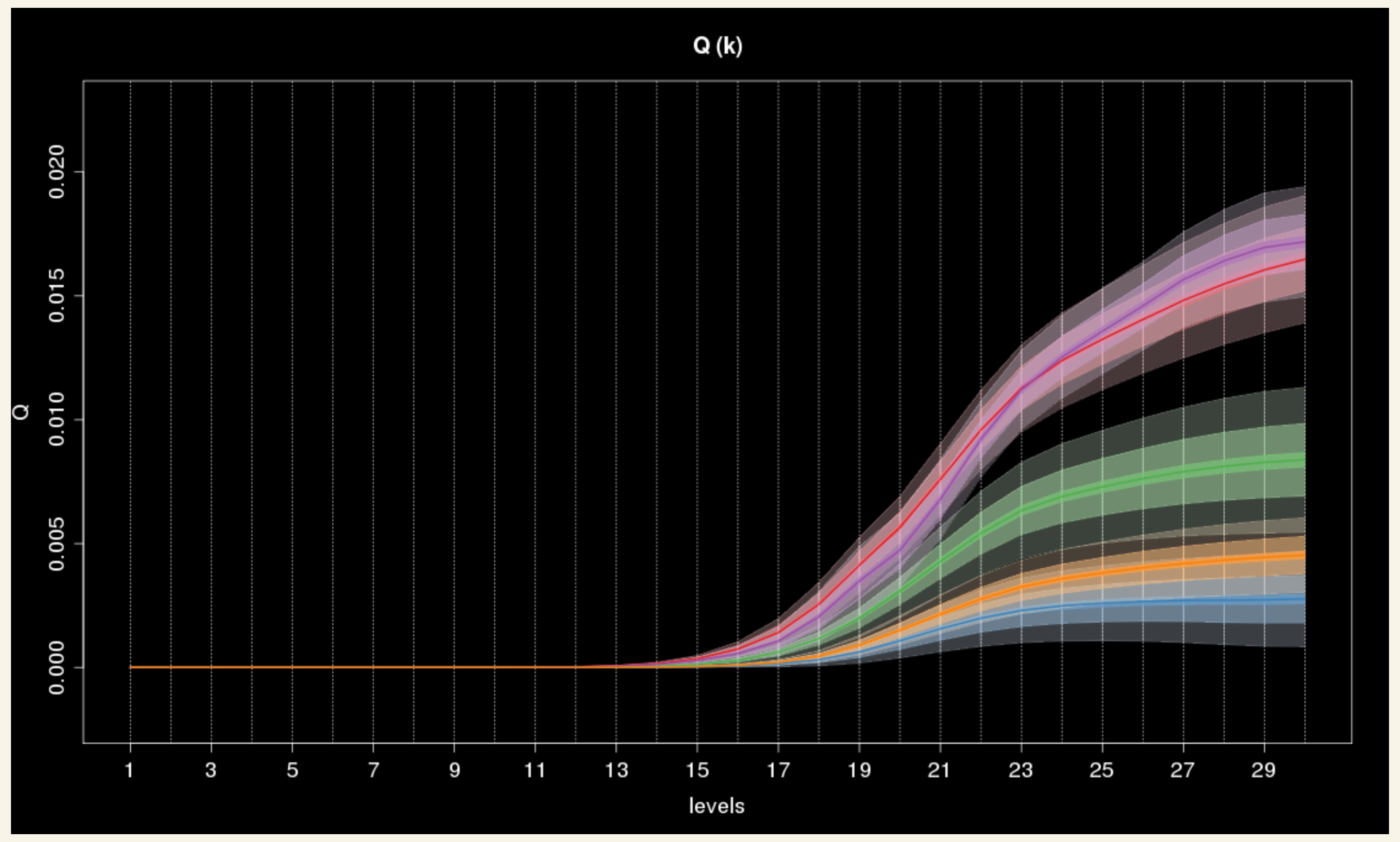
Result

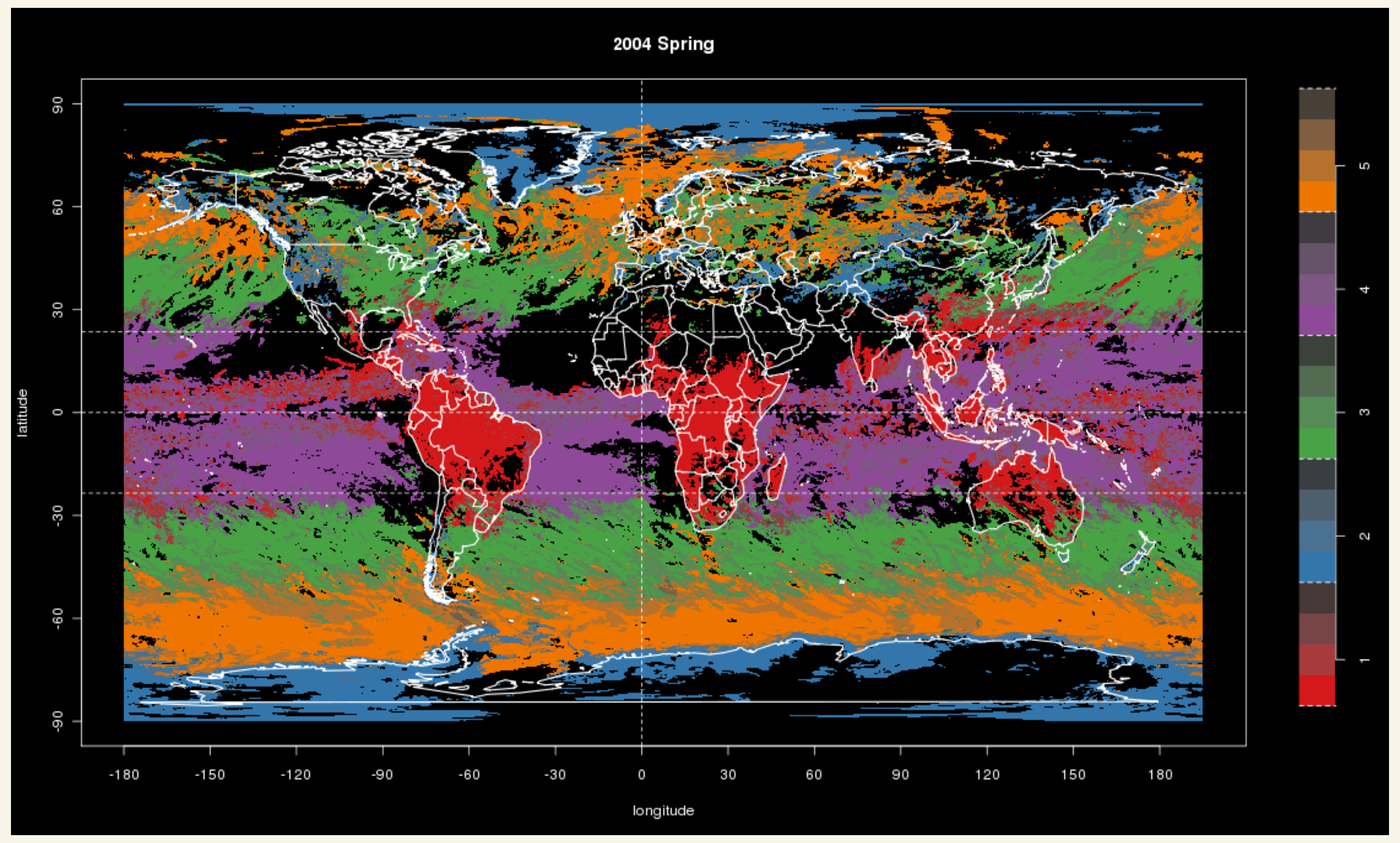
We cluster data in the feature space, and display cluster areas of extreme precipitation by seasons on maps. The relevant variables associated with clusters are also illustrated on parallel coordinate plots. We determine the number of clusters by the consistence analysis based on adjusted Rand indices comparing different number of clusters and random initializations.

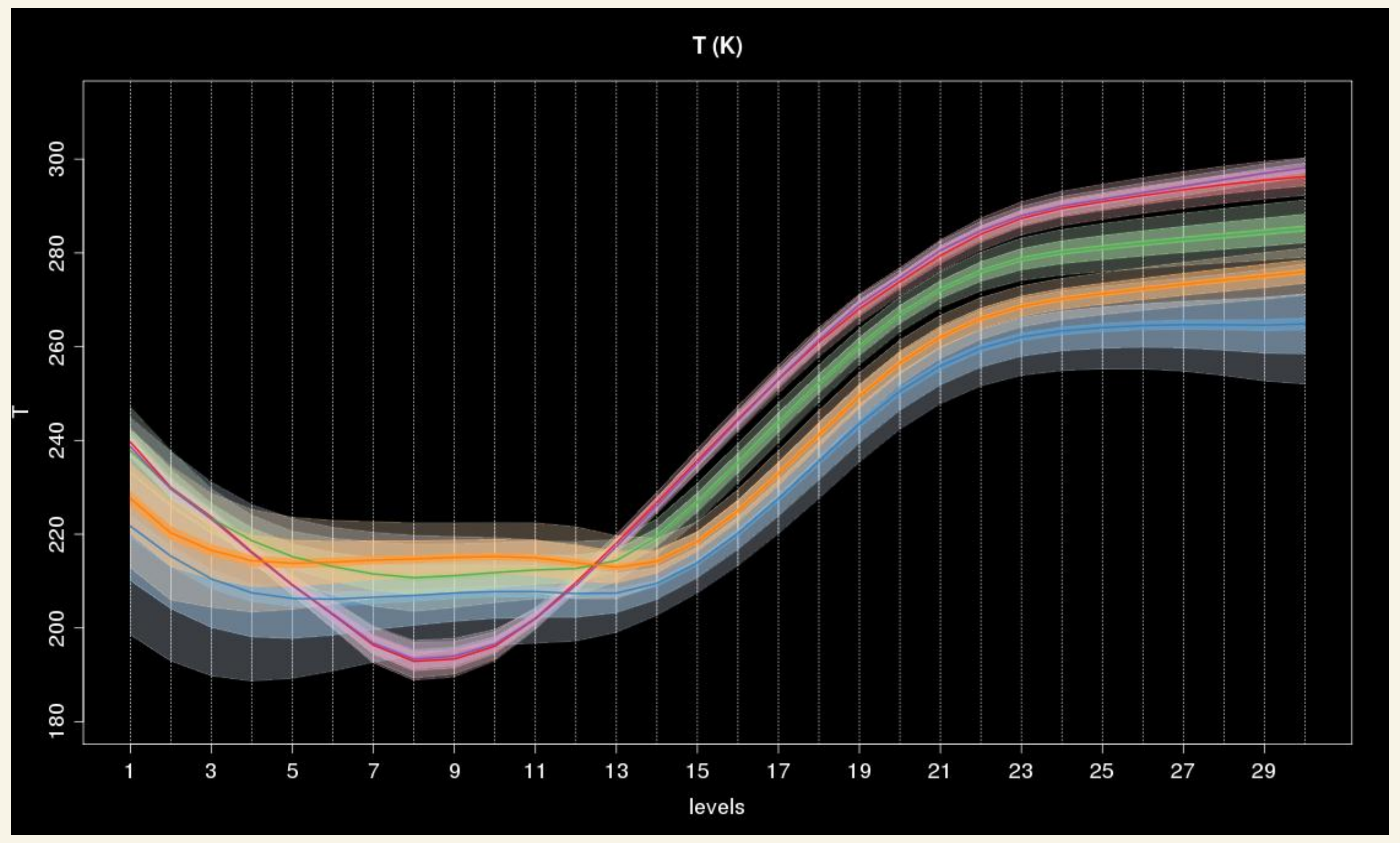
Reference

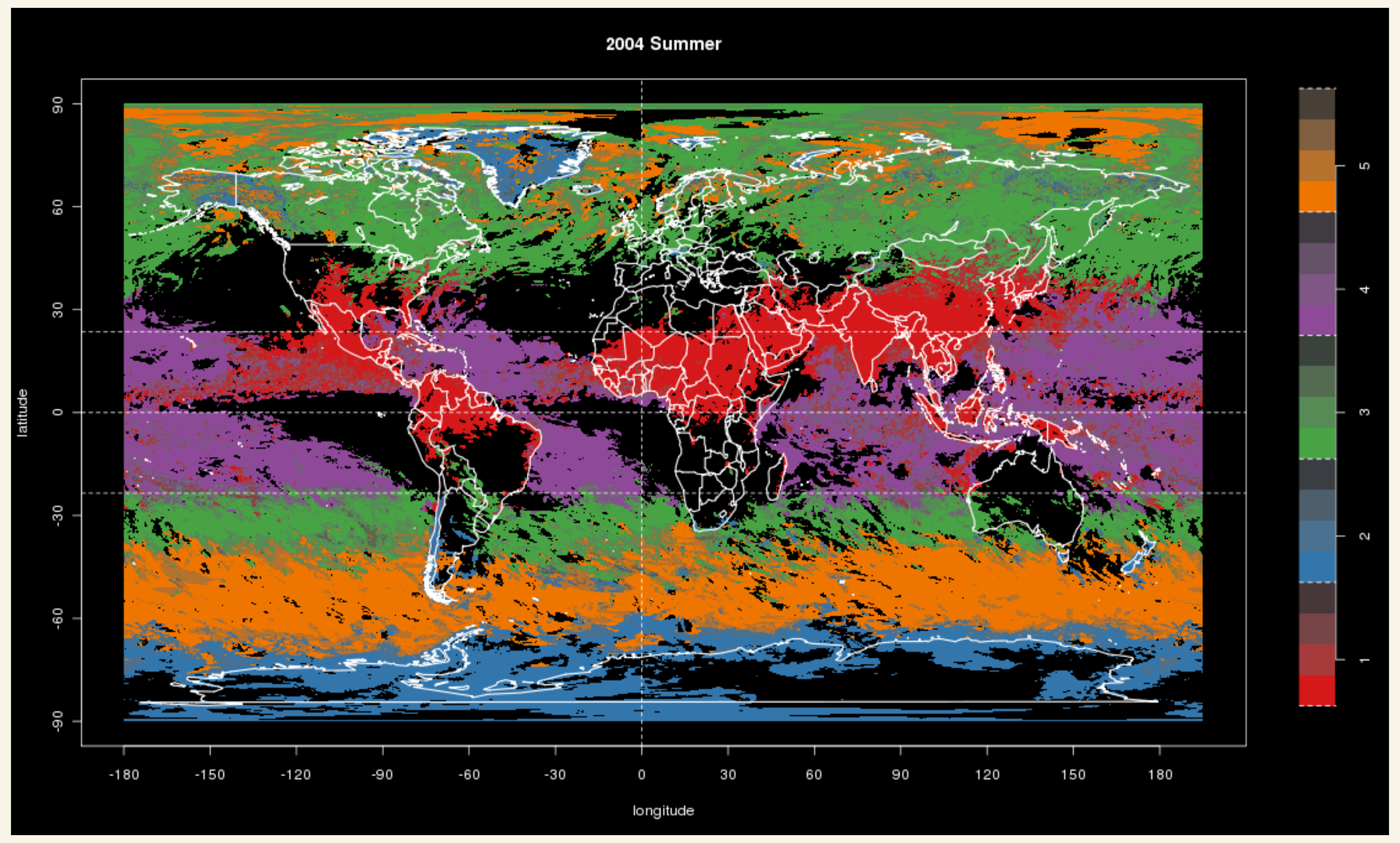
1. Community Atmosphere Model (CAM5). http://www.cesm.ucar.edu/models/cesm1.0/cam/
2. pmclust R package on CRAN. http://cran.r-roject.org/web/packages/pmclust/index.html
3. High Performance Statistical Computing website. http://thirteen-01.stat.iastate.edu/snoweye/hpsc/

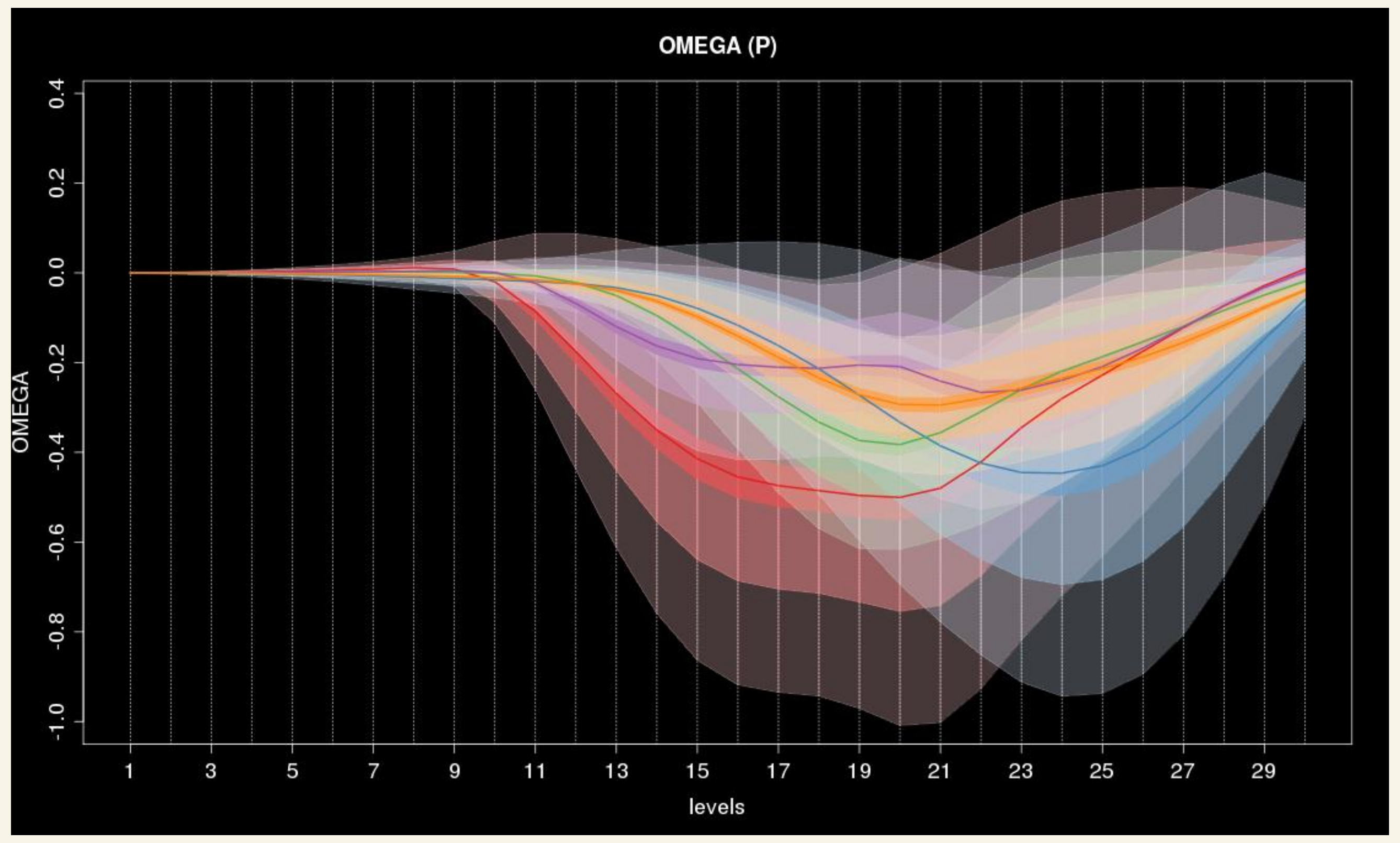


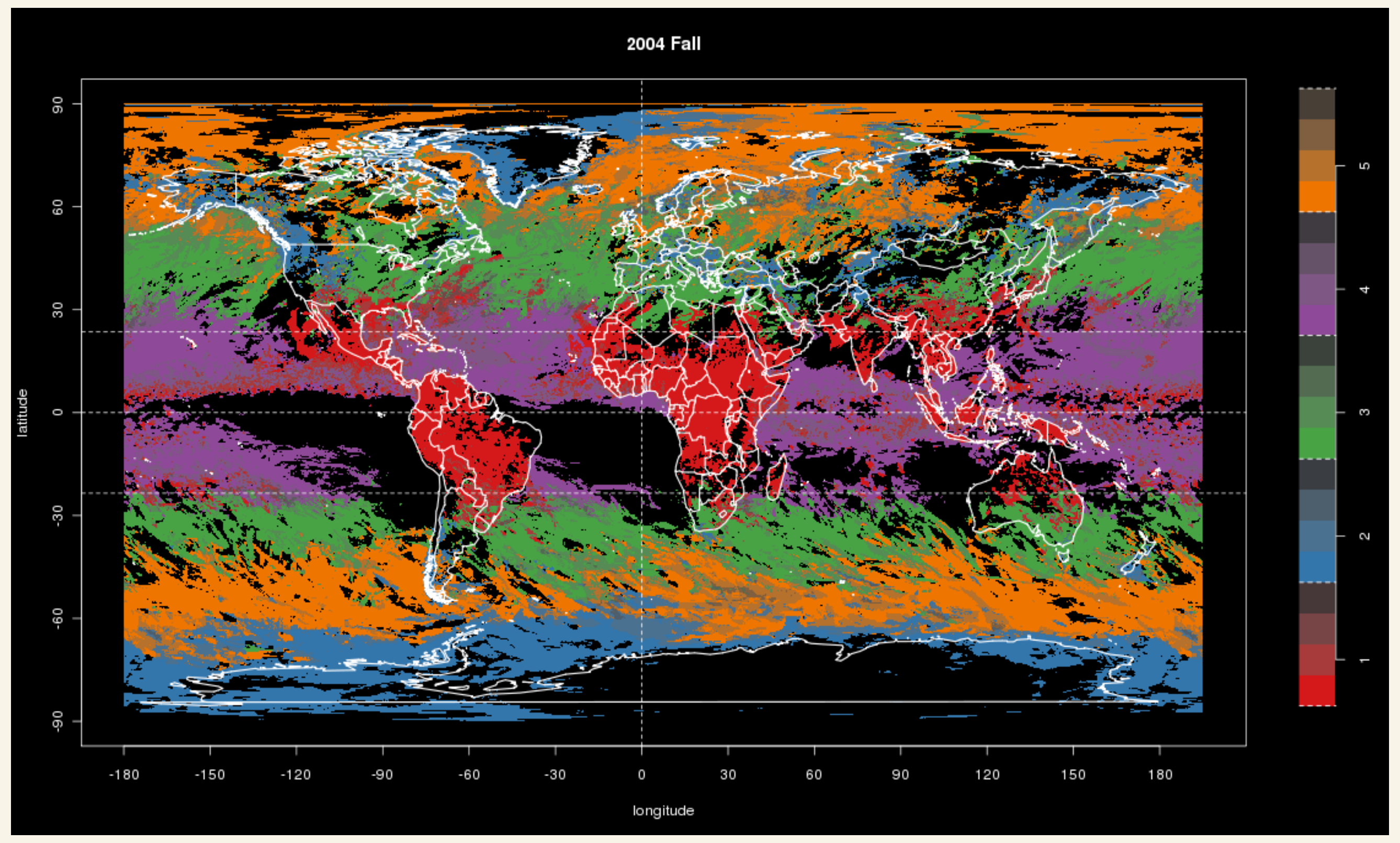


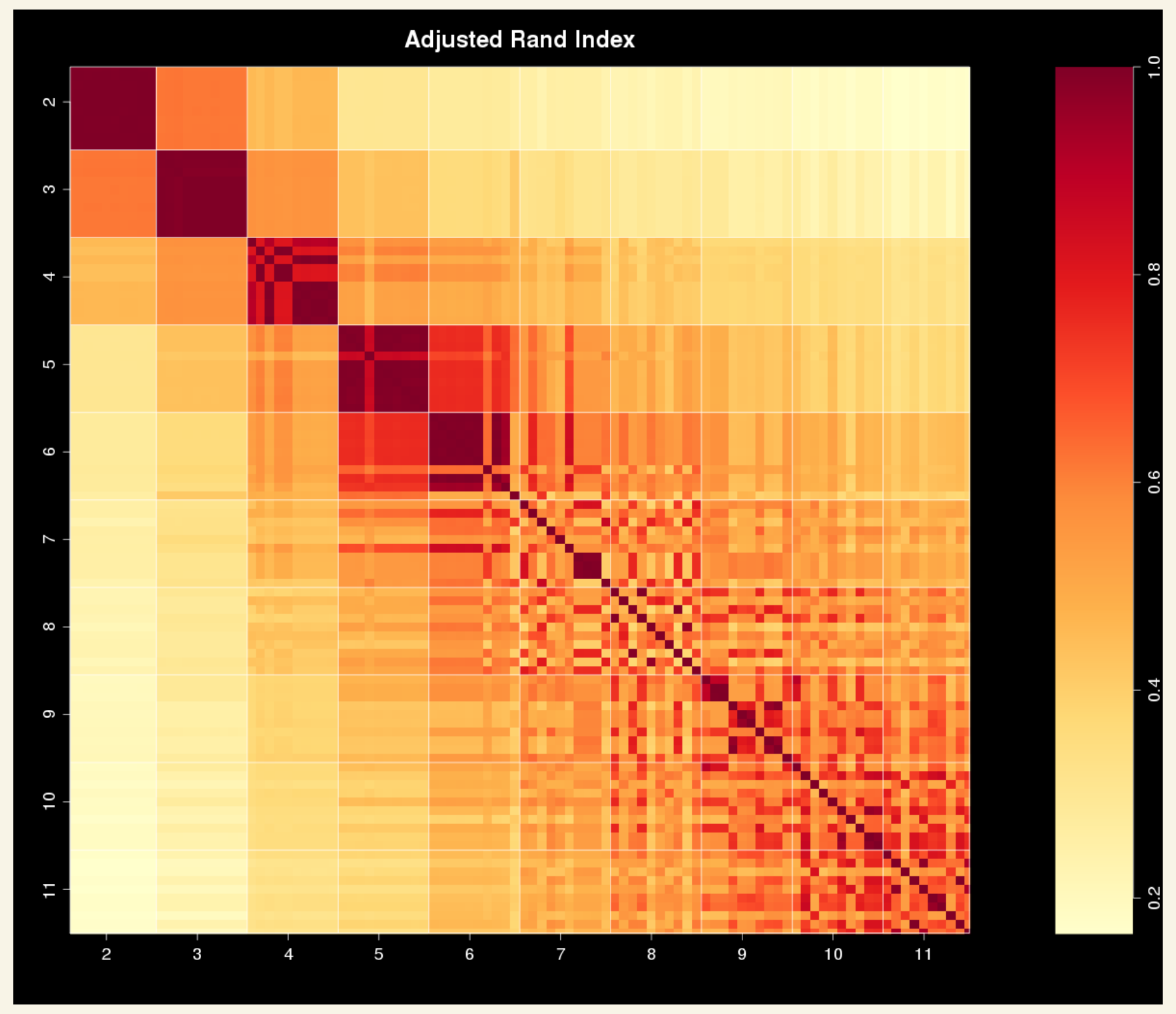












ACKNOWLEDGEMENT

The authors are grateful to Wayne McKinney and Peter Takacs for very useful discussions. The Advanced Light Source is supported by the Director, Office of Science, Office of Basic Energy Sciences, Material Science Division, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 at Lawrence Berkeley National Laboratory.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.