Multivariate Spatio-Temporal Clustering of Time-Series Data: An Approach for Diagnosing Cloud Properties and Understanding ARM Site Representativeness

Forrest M. Hoffman,* William W. Hargrove
Oak Ridge National Laboratory†

Anthony D. Del Genio
NASA Goddard Institute for Space Studies

1 MULTIVARIATE CLUSTERING

A multivariate statistical clustering technique—based on the iterative k-means algorithm of Hartigan (Hartigan, 1975)—has been used to extract patterns of climatological significance from 200 years of general circulation model (GCM) output. Originally developed and implemented on a Beowulf-style parallel computer constructed by Hoffman and Hargrove from surplus commodity desktop PCs (Hargrove et al., 2001), the high performance parallel clustering algorithm (Hoffman and Hargrove, 1999) was previously applied to the derivation of ecoregions from map stacks of 9 and 25 geophysical conditions or variables for the conterminous U.S. at a resolution of 1 sq km (Hargrove and Hoffman, 1999). Figure 1 describes this application of the k-means approach to Multivariate Geographic Clustering (MGC).

The left side of Figure 1 represents geographic space, while the right side illustrates the same map cells or observations in a multi-dimensional data space. The N characteristics of each map cell on the left are used as the N coordinates for that observation in data space on the right. In Figure 1, N is 3: temperature, organic matter, and rainfall. Having no information about the geographic coordinates of each observation, the iterative clustering algorithm finds k groups of observations based on their proximity, by simple Euclidean distance, in data space. Reassembling the map cells in geographic space and coloring them according to their cluster assignment yields a new map showing regions of approximately equal multi-variance with respect to the N characteristics used in the clustering process.

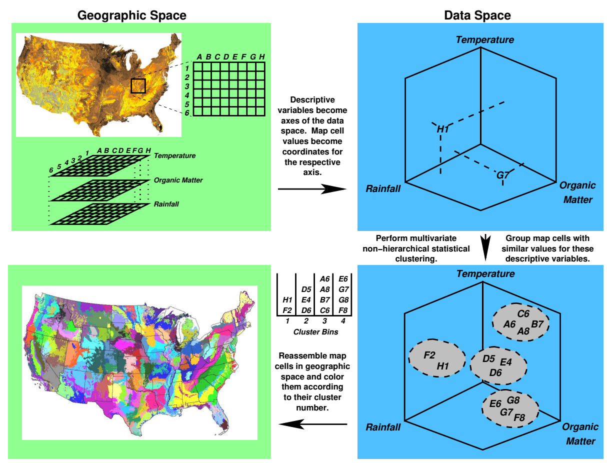

Figure 1: The Multivariate Geographic Clustering (MGC) procedure.

2 SPATIO-TEMPORAL CLUSTERING

Now applied both across space and through time, the clustering technique yields temporally-varying climate regimes from predictions of GCMs. A business-as-usual (BAU) scenario from transient runs of the Parallel Climate Model (PCM) (Washington et al., 2000) was clustered using three fields of significance to the global water cycle (surface temperature, precipitation, and soil moisture) from 1871 through 2098. An analysis of the five-year running average of cluster frequency (or regime land area) shows an increase in spa-

*Corresponding author address: Forrest M. Hoffman, Oak Ridge National Laboratory, P.O. Box 2008, Oak Ridge, TN 37831–6036 USA; e-mail: forrest@climate.ornl.gov

†Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract number DE–AC05–00OR22725.

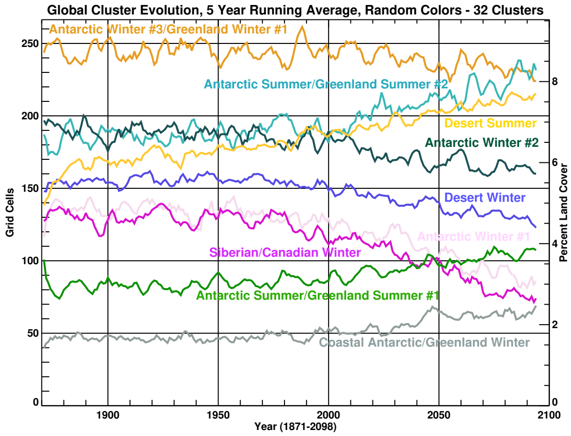

Figure 2: Trajectories of global environmental change can easily be identified as locations change among states. For example, the climate regime represented by Cluster 14 (Desert) increases through time while the regime represented by Cluster 10 (Siberian/Canadian Winter) decreases over the same period. Lines are plotted in the same random colors as their definitions shown in Figure 3.

tial area occupied by the cluster or climate regime which typifies summer-time desert regions (i.e., an increase in desertification) and a decrease in the spatial area occupied by the climate regime typifying winter-time high latitude permafrost regions (Figure 2). Additionally, significant changes are seen to occur in both Antarctica and Greenland due primarily to increasing temperatures over the 200 year time period. While desert-like conditions increase globally, the "desert winter" regime actually decreases in coverage indicating increasingly warmer winters.

Representative climate regimes were determined by taking three 10-year averages of the fields 100 years apart for northern hemisphere winter (December, January, and February) and summer (June, July, and August). The result is global maps of typical seasonal climate regimes for 100 years in the past, for the present, and for 100 years into the future. Figure 4 shows the past map and the future map for northern hemisphere winter. Reduction of complex multivariate data sets into a common set of clusters or regimes facilitates direct head-to-head comparison and the detection and quantification of long term climate change.

Cluster Number	Temperature [K]	Precipitation [$\times 10^{-7}$ kg m^{-2} s^{-1}]	Soil Moisture [Vol Frac]	Name
- 1	208.05	0.01	1.00	Antarctic Winter #1 (Coldest)
- 13	222.85	0.01	1.00	Antarctic/Greenland Winter #2
- 3	235.37	0.03	1.00	Antarctic Winter #3/Greenland Winter #1
- 10	241.29	0.05	0.23	Siberian/Canadian Winter
+ 6	246.71	0.05	1.00	Antarctic/Greenland Summer #2
+ 28	250.70	0.20	1.00	Coastal Antarctic/Greenland Winter
18	256.63	0.10	0.21	
4	257.42	0.12	0.45	
+ 12	262.40	0.09	1.00	Antarctic/Greenland Summer #1
2	271.05	0.12	0.28	
16	272.79	0.44	0.96	
25	272.80	0.28	0.28	
21	276.95	0.19	0.52	
17	277.47	0.48	0.28	
24	282.56	0.19	0.26	
- 5	283.39	0.03	0.14	Desert Winter
15	285.24	0.35	0.43	
11	291.96	0.67	0.32	
19	293.59	0.33	0.23	
29	294.43	0.84	0.32	
32	295.11	1.01	0.36	
27	295.77	1.19	0.37	
20	295.81	2.34	0.39	
31	296.20	0.50	0.28	
23	296.25	1.38	0.38	
9	296.29	2.01	0.40	
8	296.52	1.57	0.39	
26	296.56	1.77	0.40	
30	297.03	2.91	0.39	
7	297.22	0.03	0.27	Arid/Semi-Arid
22	297.28	0.18	0.21	
+ 14	298.55	0.02	0.10	Desert Summer (Hottest & Driest)

Figure 3: Each of the 32 climate regimes is quantitatively defined by the properties of its cluster centroid. This feature of clustering allows one to easily retrieve and understand the complex multivariate behavior of dynamic processes. Names can be ascribed to the more recognizable regimes. The random colors in the first column are the same as those used in Figures 2 and 4. The other colors are "similarity colors" obtained when each of the three variables is assigned to one of the RGB color guns.

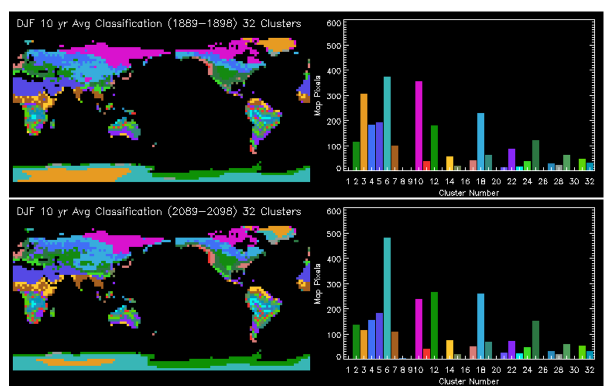

Figure 4: Comparison of different periods is facilitated by coloring maps according to their state space assignments. For example, both the gold colored region (Cluster 3) representing the coolest Antarctic summer and the magenta colored region (Cluster 10) representing the coldest Siberian/N. Canadian winter shrink from the beginning of the 200 year period to the end.

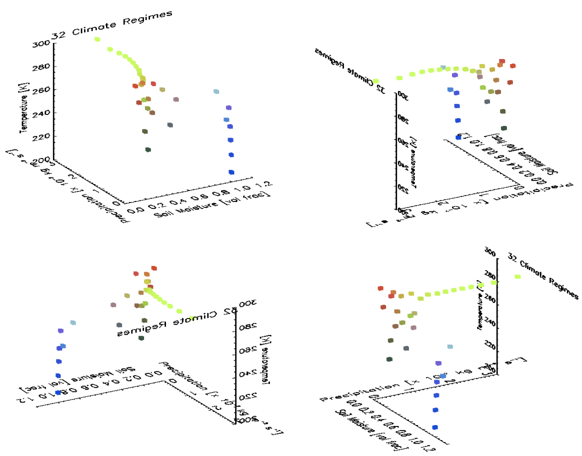

Figure 5: Climate regimes statistically defined in terms of three variables from 200 years of monthly global output of the Parallel Climate Model (PCM). These 32 regimes or states exhaustively indicate the subset of the climate state space occupied by PCM predictions, and can serve as basis states for intercomparision with measurements or other model predictions.

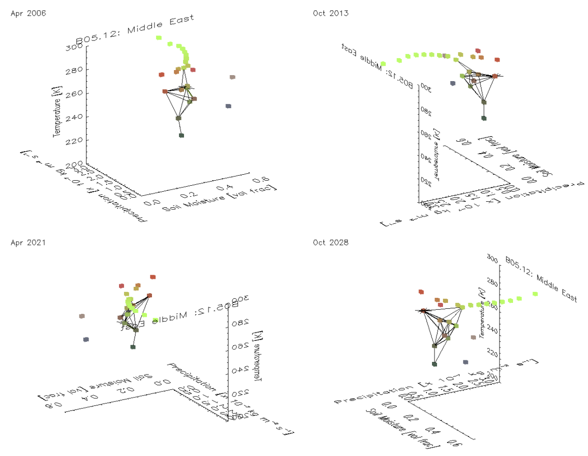

Figure 6: A trajectory among climate regimes is traced out through time as a single geographic location, in this case in the Middle East, experiences the conditions of these climate states.

3 PHASE SPACE REPRESENTATION

Using three-dimensional data or phase space representations of these climate regimes (i.e., the cluster centroids) allows one to observe the portion of this phase space occupied by the land surface at all points in space and time (Figure 5). Any single spot on the globe will exist in one of these climate regimes at any single point in time. By incrementing time, that same spot will trace out a trajectory or orbit between and among these climate regimes (or atmospheric states) in phase (or state) space (Figure 6). When a geographic region enters a state it never previously visited, a climatic change is said to have occurred. Tracing out the entire trajectory of a single spot on the globe yields a “manifold” in phase space representing the shape of its predicted climate occupancy. This sort of analysis enables a researcher to more easily grasp the multivariate behavior of the climate system and resulting impacts on the global water cycle.

4 APPLICATION TO ARM DATA

Cluster analysis is a powerful tool which can provide a common basis for comparison across space and through time for multiple climate simulations. Because it runs efficiently on a parallel supercomputer, the tool can be used to reveal long-term patterns in very large multivariate data sets. Given an array of equally-sampled variables, the technique statistically establishes a common and exhaustive set of approximately equal multi-variance regimes or states in an N-dimensional phase (or state) space. These states are defined in terms of their original measurement units for every variable considered in the analysis.

Clustering may be used not only to analyze and intercompare climate simulations, but also to analyze observations and intercompare them with each other and with model results. The area change graph in Figure 2 could show trends in cloud and climate states from ARM’s long time series measurements. When measurements are clustered in combination with model results, two trajectories among regimes—like the single trajectory shown in Figure 6—could be drawn simultaneously. These trajectories could be seen to diverge when models and measurements diverge and converge when models and measurements agree. By analyzing long time series observations with model or reanalysis results, the state space occupancy of a single ARM site could be plotted as a manifold in the “full” cloud/climate phase space yielding insights into the representativeness of individual sites or the entire ARM observation network.

Additional information, including color figures and 3-D animations, is available at http://climate.ornl.gov/.

References

Hargrove, W. W. and Hoffman, F. M. (1999). Using Multivariate Clustering to Characterize Ecoregion Borders. *Computing in Science & Engineering*, 1(4):18–25.

Hargrove, W. W., Hoffman, F. M., and Sterling, T. (2001). The Do-It-Yourself Supercomputer. *Scientific American*, 265(2):72–79.

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley & Sons, New York.

Hoffman, F. M. and Hargrove, W. W. (1999). Multivariate Geographic Clustering Using a Beowulf-style Parallel Computer. In Arabnia, H. R., editor, *Proceedings of the International Conference on Parallel and Distributed Processsing Techniques and Applications (PDPTA '99)*, volume III, pages 1292–1298, Las Vegas, Nevada. CSREA Press. ISBN 1–892512–11–4.

Washington, W. M., Weatherly, J. W., Meehl, G. A., Jr., A. J. S., Bettge, T. W., Craig, A. P., Jr., W. G. S., Arblaster, J. M., Wayland, V. B., James, R., and Zhang, Y. (2000). Parallel Climate Model (PCM) Control and Transient Simulations. *Climate Dynamics*, 16(10/11):755–774.