Smoky Mountains Computational Sciences and Engineering Conference

Forrest M. Hoffman
Oak Ridge National Laboratory

Presentation Title:
Big Data in the Geosciences: Data Mining Methods for Characterizing Ecoregions, Designing Sampling Networks, Detecting Forest Threats, and Understanding Climate Change Predictions

Abstract:
From field-scale measurements to global climate simulations and remote sensing, the growing body of very large and long time series Earth science data are increasingly difficult to analyze, visualize, and interpret. The size and complexity of these data exceed the limits of traditional analysis methods and tools. Data mining, information theoretic, and machine learning techniques are starting to be applied to problems of segmentation, feature extraction, change detection, and model–data comparison in the Earth sciences. In this presentation, I will describe analytics methods we have applied to various kinds of sparse in situ measurements, large-scale satellite observations, and copi- ous climate model output. Using a highly scalable cluster analysis technique based on the k-means algorithm, called Multivariate Spatio-Temporal Clustering (MSTC), we have segmented continen- tal and global land areas, creating customized ecoregions as a

framework for understanding plant and animal habitats and projecting species shifts under environmental change. Further, we have used this quantitative methodology to systematically delineate environmental sampling domains for informing site selection and determining the representativeness of measurement sites and networks for the U.S. Department of Energy’s (DOE’s) Ameriflux network, the National Science Founda- tion’s National Ecological Observatory Network (NEON), and DOE’s Next Generation Ecosystem Experiments (NGEE). For the U.S. Department of Agriculture Forest Service’s ForWarn system, we have simultaneously analyzed the very large record of NDVI (normalized difference vegetation index) data, available every eight days, for the continental United States at 250 m resolution from both of the National Aeronautics and Space Administration’s MODIS (Moderate Resolution Imag- ing Spectroradiometer) instruments aboard the Terra and Aqua satellites. Our results indicate the seasonal behavior of vegetation phenology and allow us to detect the effects of large scale distur- bances, such as wildfire, drought, pest infestation, severe storms, and changes in land use. Large volumes of output from today’s Earth system models, typically used to predict the effects of cli- mate change on centennial time scales, can be unwieldy and difficult to understand. Employing our MSTC methodology, we can easily show the large scale behavior of climate model projections and compare predictions from different models and different climate change scenarios. This collection of geoscience applications demonstrates the utility of data mining techniques and offers a look into the future of large scale climate data analytics enabled by high performance computing.

Bio:

Forrest M. Hoffman is a computational climate scientist in the Climate Change Science Insti- tute (CCSI) at Oak Ridge National Laboratory. With dual positions in the Computer Science & Mathematics and the Environmental Sciences Divisions, Forrest develops Earth system models for use on DOE’s Leadership Class computing resources and studies feedbacks between global biogeo- chemical cycles and Earth’s climate. He has a particular interest in developing and applying large scale data analytics methods for understanding the behavior of terrestrial ecosystems, characteriz- ing landscape dynamics, analyzing the global carbon cycle, and benchmarking Earth system model performance.