The Center for Ocean-Land-Atmosphere Studies (COLA) at George Mason University (GMU) is focused on the predictability of climate variations from days to decades. The research conducted by COLA is largely based on global, coupled models of the Earth system such as are used to produce recent past climate simulations, predictions of intra-seasonal to decadal climate variations, and projections of future climate.

Recent Computing Innovation: A recent computing advance that has significantly altered the landscape for these types of studies is the substantial increase in high-performance computing (HPC) resources, e.g. those funded by the former Office of Cyberinfrastructure of NSF and managed by the former TeraGrid and current XSEDE network. Concurrently, Earth system model codes have been developed that can make use of those HPC resources. The result is that the climate models that have been used for the past several decades recently have been substantially improved, particularly in terms of spatial resolution, resulting in an explosive uptick in the volume and complexity of model output data. The fields of interest include both monthly mean and higher frequency (e.g. daily) values of a wide range of simulated/predicted four-dimensional atmospheric, oceanic and land surface variables. The focus on predictability requires ensembles of tens of equi-probable simulations. As such, analyses require rapid access to petabytes of data. These data are generated on the HPC systems that are available for civilian Earth science research, but only small subsets are transported to the local computing facility due to resource limitations.

For example, a question of high interest in climate research is whether or not the incidence, frequency, magnitude, seasonality, intermittency and remote impacts of El Niño events will change in the future as the global climate warms. A related question is whether or not the predictability of these aspects of El Niño will change. In order to answer these questions, the daily and monthly mean values of sea surface temperature, thermocline depth, precipitation, and 3D atmospheric state variables (temperature, pressure, humidity and 3D components of the wind) and cloud amount must be analyzed for historical simulations and scenarios of future climate, from all available models and all ensemble members. In order to assess the changes in the remote response to El Niño, global grids must be evaluated and several other variables must be analyzed, especially land surface quantities such as soil moisture, leaf area index, and snow depth. The upshot is that, in order to answer a question that is simply posed about recent past, near-term future and long-term projected changes in the climate system, hundreds of daily 2D grids have to be obtained for several thousands years of model simulations. On grids of 1° or finer spatial resolution, the data volume is of O(1 PB). For grids with the higher spatial resolution that is now being employed for some of these simulations (global Earth grids with 15-30 km cell spacing), there is an order of magnitude more data, hence the data volumes of interest are O(10 PB). The data represent quantities characterizing the state of the atmosphere, oceans, land surface, sea ice, suspended aerosols, and biota, an enormously complex set of data types.

Science Challenge: The volume and complexity of data that must be quantitatively analyzed in a probabilistic framework have reached levels that cannot be addressed with conventional data analysis and visualization tools. Furthermore, it is no longer feasible for each of O(10) researchers at each of O(100) institutions to have a local copy of the O(10 PB) data. Therefore, innovations in intelligent information systems – machine learning, data analytics, advanced visualization, and automated, distributed workflow – are needed to enable the discovery of features, the quantitative comparison and the analytic synthesis of data that are complex, voluminous and widely distributed among archive and processing centers.