Geophysical Research Abstracts
Vol. 15, EGU2013-13082, 2013
EGU General Assembly 2013

An intelligent data model for the storage of structured grids

John Clyne and Alan Norton
National Center for Atmospheric Research, Boulder CO, United States (clyne@ucar.edu)

With support from the U.S. National Science Foundation we have developed, and currently maintain, VAPOR: a geosciences-focused, open source visual data analysis package. VAPOR enables highly interactive exploration, as well as qualitative and quantitative analysis of high-resolution simulation outputs using only a commodity, desktop computer. The enabling technology behind VAPOR's ability to interact with a data set, whose size would overwhelm all but the largest analysis computing resources, is a progressive data access file format, called the VAPOR Data Collection (VDC). The VDC is based on the discrete wavelet transform and their information compaction properties. Prior to analysis, raw data undergo a wavelet transform, concentrating the information content into a fraction of the coefficients. The coefficients are then sorted by their information content (magnitude) into a small number of bins. Data are reconstructed by applying an inverse wavelet transform. If all of the coefficient bins are used during reconstruction the process is lossless (up to floating point round-off). If only a subset of the bins are used, an approximation of the original data is produced. A crucial point here is that the principal benefit to reconstruction from a subset of wavelet coefficients is a reduction in I/O. Further, if smaller coefficients are simply discarded, or perhaps stored on more capacious tertiary storage, secondary storage requirements (e.g. disk) can be reduced as well. In practice, these reductions in I/O or storage can be on the order of tens or even hundreds.

This talk will briefly describe the VAPOR Data Collection, and will present real world success stories from the geosciences that illustrate how progressive data access enables highly interactive exploration of Big Data.