Statistical Projections for Multi-resolution, Multi-dimensional Visual Data Exploration and Analysis

Hoa Nguyen[2], Dáithí Stone[1], and E. Wes Bethel[1]

[1] Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[2] University of Utah, Salt Lake City, UT, USA

January, 2016

Acknowledgment

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Legal 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.

Statistical Projections for Multi-resolution, Multi-dimensional Visual Data Exploration and Analysis

Hoa Nguyen*
University of Utah

Dáithí Stone†
Lawrence Berkeley National Laboratory

E. Wes Bethel‡
Lawrence Berkeley National Laboratory

Abstract

An ongoing challenge in visual exploration and analysis of large, multi-dimensional datasets is how to present useful, concise information to a user for some specific visualization tasks. Typical approaches to this problem have proposed either reduced-resolution versions of data, or projections of data, or both. These approaches still have some limitations such as consuming high computation or suffering from errors. In this work, we explore the use of a statistical metric as the basis for both projections and reduced-resolution versions of data, with a particular focus on preserving one key trait in data, namely variation. We use two different case studies to explore this idea, one that uses a synthetic dataset, and another that uses a large ensemble collection produced by an atmospheric modeling code to study long-term changes in global precipitation. The primary findings of our work are that in terms of preserving the variation signal inherent in data, that using a statistical measure more faithfully preserves this key characteristic across both multi-dimensional projections and multi-resolution representations than a methodology based upon averaging.

Index Terms: G.3 [Statistics]: nonparametric statistics—visualizing data variation,H.5.m [Information Systems]: Information Interfaces and Presentation—miscellaneous: multi-variate, multi-resolution projection, I.6.6 [Computing Methodologies]: Simulation and Modeling—Simulation Output Analysis

1 Introduction

To facilitate knowledge discovery in the visual exploration and analysis of large, complex, multidimensional data, we examine the question of how to present meaningful information through a combination of data projections and summarization. Specifically, we focus on the use of a statistical measure, Coefficient of Variation (or C_V), which reflects the amount of variation in data.

There are often instances of data exploration and analysis where understanding variation in data is of greater interest than the absolute value of the data itself. For example, variability in the climate system consists primarily of transfers of energy, mass, and moisture between locations, rather than variations in the total energy, mass, or moisture globally; hence, a metric sensitive to these transfers could be a more informative descriptor of how the climate is varying through time than a metric that is insensitive to the transfers. This variation information is also important to help climate scientists predict future weather patterns. Therefore, we propose a methodology that will help users identify the variation information in data quickly and with high accuracy.

We explore two interesting properties of C_v in this paper within the context of complex multidimensional visual data exploration and analysis. The first is how C_v can convey the useful signal in data–the amount of variation–at multiple scales, which makes it useful within the context of multiresolution representations of data. This property, the ability to preserve useful signal (variation), is quite useful in creating reduced-sized, more manageable representations of large data. The second is the use of C_v as the basis for performing projection-based data reduction, where different views of a dataset are the result of projection from a higher-dimensional to lower-dimensional space. Together, these two properties facilitate understanding of variation in complex, multidimensional data.

In summary, the main contributions of this work are:

- A methodology of using a statistical measure of variation, C_v, for the purpose of visually conveying variational signal within complex, multidimensional, multiresolution data to identify variation of data.
- A visual exploration tool that uses C_v as the basis for multidimensional, multiresolution projections to help users quickly interact and efficiently perform visualization analysis tasks.
- Case studies that confirm our methodology provides more insight and better accuracy than the current methods those based on data averaging.

2 Previous Work

The issue of how to reduce large-sized datasets to ones that are more manageable is a topic that has been studied in many different forms over the years, though primarily within the context of focusing on data values, rather than data variation.

For image-based data, Williams, 1983 [13] introduced the concept of *mip maps*, which are multi-resolution forms of images. The process of constructing each successively coarser resolution of image involves a process by which four pixels are ”filtered,” or averaged together, into a single pixel. This approach, and those like it that use pixel-averaging, produces coarse-resolution datasets that appear ”blurred”; in effect, the high frequency component of the underlying original signal is lost through the repeated averaging process.

Wavelet-based representations of data, such as the Discrete Haar Wavelet Transformation [4], represent data as a combination of base values (averages) and differences. This approach has proved useful for addressing several problems of large-data visualization, including progressive data access and multi-resolution rendering (Clyne, 2012 [3]). Visually, the difference between a rendering of full- and reduced-resolution version of data appears as a loss of high-frequency detail.

Conceptually, a reduced-resolution, wavelet-encoded dataset represents averages (and differences) of data samples. At coarser and coarser resolutions, the effect is similar as for mip-map representations of images: the processing of averaging more and more data “washes out” the variational signal inherent in the underlying data. Hence, while methods that rely on computing data averages at multiple levels of resolution may be useful for representing data values, they are not promising for representing variation in data.

Other approaches for reducing the size multidimensional data center around the idea of projections. Simply stated, a projection is one that reduces a dataset from R^n dimensions to R^m dimensions, where $m < n$. Some approaches, like orthogonal or arbi-

[*] e-mail: hoanguyen@sci.utah.edu
[†] e-mail: dstone@lbl.gov
[‡] e-mail: ewbethel@lbl.gov

trary slice planes, are projections that are spatially constrained subsamplings of data. Other approaches, like Principal Component Analysis (PCA) [1] or Isomap [10], both examples of linear and non-linear dimension reduction, respectively, are essentially optimizations aimed at discovering lower-dimensional embeddings of higher-dimensional data that take into account the underlying characteristics of multidimensional data distributions. For example, PCA finds the projection that captures the most variance in data, but we are interested in different problems, namely presentation of variation and preserving variation across multiple scales. See Maaten et al., 2009 [11], for a comparative review of these methods. Whether or not methods like PCA or Isomap are useful when doing projections where the signal of interest is variation is an interesting one, but outside the scope of this paper.

In terms of visualizing variation in data, the *box plot* is a glyph-based method for displaying variation in data (Chambers, 1983 [2]). The box size reflects the distribution range in data in terms of quartiles, and the box glyph may include additional annotations to indicate the location of the median, and box "whiskers" indicate the full range of data to help show outliers. Whitaker et al., 2013 [12] extended this idea to depict variation in data features. There is a long history of using traditional scalar color mapping techniques to depict variation in data, and Demir et al., 2014 [5] use that visualization approach in conjunction with brushing and linking to enable visual exploration of variation in ensemble collections of simulation output. One limitation of box plots is they are useful for presenting a small number of samples. In contrast, with a single quantity, like C_v, we can show and perform operations, like visualization, projections, and multiresolution reductions, on entire fields.

Our approach for performing projections may be thought of as a hybrid of these two approaches. On the one hand, we use a subsetting-based approach, where a user can choose to project through one or more dimensions of a multi-variate dataset. For example, given a time-varying 3D structured dataset, a user could project across time, to produce a 3D dataset where each 3D value represents a projection across time. On the other hand, our approach is also like the subsampling method used by orthogonal or arbitrary slicing. But, rather than simply subsampling, or local weighted averaging (where a slice plane lies between grid points), we compute C_v during the projection across a subset of data to produce a single value that conveys variation within that subset.

3 Statistics Projection Methodology

We propose a method of data reduction by using statistical projections for multi-resolution or multi-dimensional data. Two statistical projections methods that reflect meaningful variation feature in data are standard deviation (σ) and Coefficient of Variation (or C_v).

σ is a statistic that indicates how tightly clustered all the various examples are around the mean in a set of data. When the examples are tightly bunched together and the bell-shaped curve is steep, σ is small. When the examples are spread apart and the bell curve is relatively flat, that indicates a relatively large standard deviation.

$$\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \qquad \text{and} \qquad C_v = \frac{\sigma}{\bar{x}} \tag{1}$$

In Eq. 1, n is the number of data points, and $\bar{x}$ is the mean, or average, of the set of n data points. C_v represents the ratio of σ to $\bar{x}$, and it is a useful statistic for comparing the degree of variation from one data series to another, even if the means are drastically different from each other.

The σ and C_v (Eq. 1) quantities are related; the C_v is essentially a normalized form of σ, which arguably makes it more useful in comparing the amount of variation in datasets where the range of data differs significantly.

Using the mean as the basis for multi-resolution reduction results in a "washing out" of the signal we are interested in, namely variation. We could use σ to show data variation. However σ by itself is only somewhat informative since it represents variation in the original data range, which may not be sufficient in understanding variation magnitude. Also, due to its property representing normalized variance, C_v is more useful in comparing variance across datasets having significantly different magnitudes.

4 Case Study

We applied our methods for two datesets, one is synthetic dataset and another is climate data.

4.1 Visualization for Synthetic Data: Multi-resolution Representation

To test the idea that use of a statistical measure like C_v is more effective at preserving variational signal across multi-resolution representations of data when compared to a method that uses averaging, we perform an evaluation using a synthetic dataset.

We created a synthetic 2D, $f(u,v)$ dataset at 1024^2 resolution where we begin with a base function value, $f = 0.5$, and add varying amounts of Gaussian noise to that base value. The standard deviation of the Gaussian distribution ranges from low to high across columns of v, so that one edge of the data appears to have relatively little noise, while the other edge appears to have significantly more noise (see Fig. 1a).

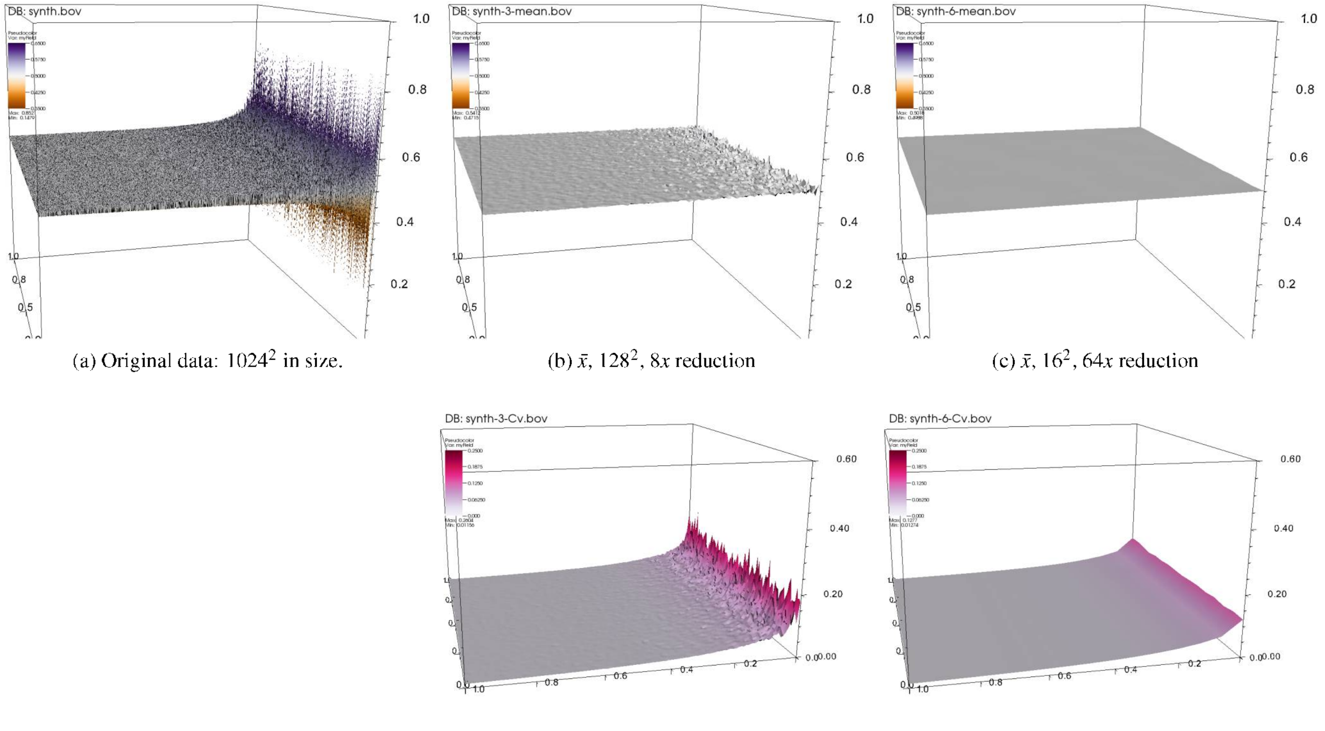
Beginning with this 1024^2 dataset, we generate multiple reduced-resolution versions using two different approaches. In the first approach, we compute each successively smaller dataset using averaging, so that 2×2 windows of input data samples are averaged to produce 1 output data sample. As we reduce the dataset further and further in size via averaging, it becomes increasingly "washed out" (Fig. 1b) and is eventually unrecognizable (Fig. 1c).

In the second approach, we use a similar method of reducing an $N \times N$ window of input samples to 1 output sample, but compute C_v over the input window rather than averaging. With this approach, the signature of variation in data is much better preserved across each successively smaller dataset. If we look at the variation across the original data, we see correspondence between the regions of high and low variation in the original data (Fig. 1a) and in C_v at 128^2 resolution (Fig. 1d). This signal, the high and low variation, is much better preserved across the multi-resolution versions of the C_v dataset (Fig. 1e). We see that the variation signature is still present with the low-resolution C_v image (Fig. 1e) but is completely "washed out" in the low-resolution $\bar{x}$ image (Fig. 1c).

This experiment shows that the C_v does a much more effective job at preserving the variation signal in data across multiple resolutions compared to using averaging. Use of an alternate unary function, like *maximum*, would produce different results than averaging, but would fail to represent the signal of variation inherent in the data. Similarly, boolean range queries, such as those forming the basis of query-driven visualization methods [9] would be ineffective in finding variation in data, as they focus on finding data that matches a given set of (compound) boolean criteria. The signal presented by variation is something that must be computed: it is not directly identifiable through unary operators nor compound range queries.

4.2 Visualization for Climate Data

Building upon the foundation in Sec. 4.1 that C_v is a better basis than averaging for preserving variational signal in data across multiple levels of resolution, we next explore extensions and applications of this idea to multi-dimensional data projections within the context of a climate data analysis problem and compare between visualization methods that used C_v.



(a) Original data: 1024^2 in size.

(b) $\bar{x}$, 128^2, $8x$ reduction

(c) $\bar{x}$, 16^2, $64x$ reduction

(d) C_v, 128^2, $8x$ reduction

(e) C_v, 16^2, $64x$ reduction

Figure 1: Comparison of $\bar{x}$ and C_v across multiple levels of resolution in a synthetic dataset.

Precipitation is one of the more visible and influential aspects of the climate system for society and ecological systems, and thus is a frequent topic of analysis. It represents one branch of the planet's hydrological cycle, wherein moisture evaporates over the ocean, is transported over ocean and land, precipitates out of the air, and then (if over land) returns to the ocean through rivers and groundwater.

Because precipitation amounts vary strongly across space (e.g. deserts versus rainforests) and sometimes across seasons, comparisons often require some form of normalization. A common way of doing this is by dividing by the mean, usually multiplying by 100 to get a percentage deviation from the historical mean. For instance, when generating gridded observational products of precipitation variations, point measurements at weather stations are converted to fractional anomalies, which are then interpolated; after the interpolation the fractional anomalies are multiplied with a spatially interpolated field of mean precipitation. The C_v is closely related to the calculation of these fractional values.

For our study, we used precipitation data generated by using the CAM5.1 global atmospheric climate model [7] run at approximately $1° \times 1°$ longitude-latitude resolution under observed boundary conditions from the 1959-2014 period [6].

The model was run 50 times with different initial states, thus producing an ensemble of 50 realizations of how the weather might have evolved. While the large number of simulations is unusual, the generation of multiple simulations in this manner is a standard approach for characterizing uncertainty in the climate system. Here we examine monthly mean precipitation output on the model's longitude-latitude grid.

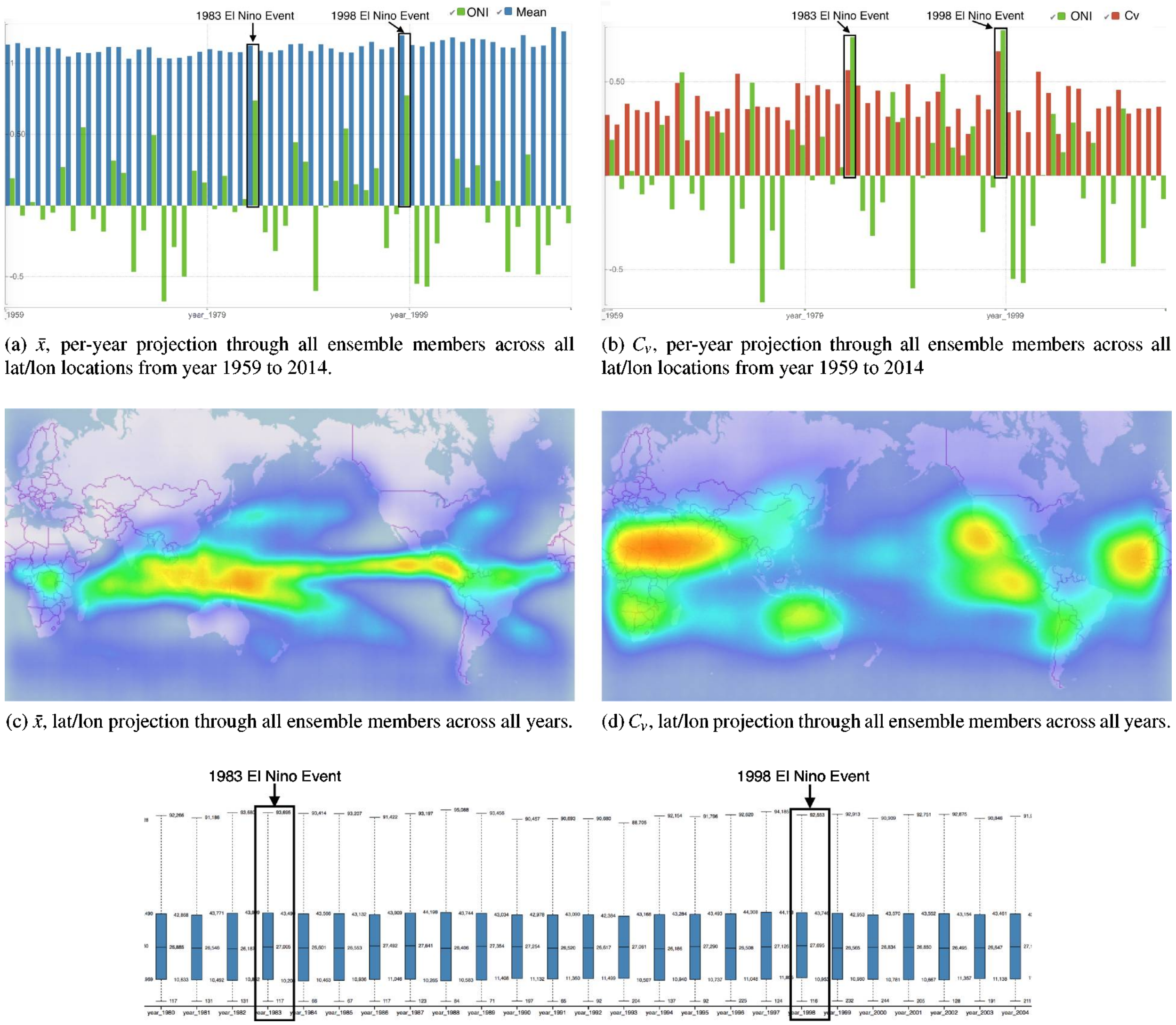
To aid in study of this collection of climate model output, the visualization of statistics projections using mean $\bar{x}$, and C_v are visualized in Fig. 2. Each visualization provides slightly a different visual representation. There are multiple tasks related to identifying the amount of variation in data visually such as identifying variational signal in yearly precipitation across all places or across different places and across all simulation runs or across different simulation runs. To answer these questions, climate scientists can easily interact with our visualization tool to switch between different views and find the best graph that represent this information most clearly.

One type of projection, a yearly projection, produces values for each year that convey either mean (Fig. 2a) or C_v (Fig. 2b) across all ensemble members and across all latitude/longitude grid points. We include a comparison with box plots to give a visual example/comparison of a glyph-based method (box plot) as shown in Fig. 2e with a field-based method (C_v).

To facilitate deeper insight, we also show in this visualization the Oceanic Niño Index (ONI), which is a metric of the shift between El Niño (warm) and La Niña (cool) events in the tropical Pacific [8]. This phenomenon is a well-documented driver of year-to-year variability in climate worldwide, representing a major shift of winds around the globe and providing the primary basis for forecasting on seasonal time scales. Fig. 2a shows both mean and ONI, while Fig. 2b shows both C_v and ONI.

In the mean plot (Fig. 2a), there is little variation in the mean from year-to-year, with the main feature being a gradual long-term trend. Looking at the mean values from year to year, the major El Niño events of 1983 and 1998, as indicated by high ONI values, are not particularly remarkable in this figure. In contrast, looking at the C_v in Fig. 2b, these two events correspond to the two years with the highest C_v values; the correspondence does not seem to hold for more moderate El Niño events, however (e.g. 1972).

We now explore a different type of projection, namely a spatial projection, which produces values at each latitude/longitude point that convey either $\bar{x}$ (Fig. 2c) or C_v across all ensemble members and



(a) $\bar{x}$, per-year projection through all ensemble members across all lat/lon locations from year 1959 to 2014.

(b) C_V, per-year projection through all ensemble members across all lat/lon locations from year 1959 to 2014

(c) $\bar{x}$, lat/lon projection through all ensemble members across all years.

(d) C_V, lat/lon projection through all ensemble members across all years.

(e) Box plots for precipitation from year 1980 to 1994

Figure 2: Comparison of $\bar{x}$ and C_V as the basis for multi-dimensional projections of precipitation in climate model output; comparison of field- and glyph-based (box plot) visualization methods.

across all years (Fig. 2d). The map of mean precipitation shows the band of rainfall that straddles the equator, known as the Intertropical Convergence Zone (ITCZ), along with the mid-latitude storm tracks that branch off from the ITCZ from the western sides of the major ocean basins; much less precipitation falls in higher latitude areas where the air is too cold to hold much water.

The map of C_V looks rather different. Generally it is highlighting the deserts in the subtropical areas to the north and south of the ITCZ. The air that has dried through precipitation while rising in the ITCZ moves poleward and descends here, leading to hot and dry conditions. The low mean precipitation means that the denominator of C_V is small, and the infrequent but substantial storms lead to a comparatively high numerator. The exception to this subtropical focus is the area with which C_V over the eastern tropical Pacific (i.e. against South America). Because the trade winds blowing from the east pull up cool water from the deep ocean here, the water at the surface is usually quite cool, does not evaporate much, and thus does not provide much moisture for subsequent rainfall. However, during El Niño years the winds reverse and temperature rises markedly, driving major thunderstorms.

The properties of the C_V map explain the behavior of the yearly bar plots of C_V in Fig. 2b. In essence it is acting as a combined index of the occurrence of El Niño events and of anomalous rainfall over subtropical regions. In contrast, the yearly bar plots of the mean are mostly reflecting activity in the ITCZ. While the mean reflects variations in rainfall where it is plentiful, application of a land filter to the global C_V (to mask out the El Niño aspect) means that C_V provides a metric of variations where precipitation is scarce. This has been achieved using the C_V without any parametric definition of what constitutes a subtropical desert.

5 Conclusion

In working with large, complex data, one key issue is how to effectively produce smaller-sized representations in a way that convey useful information. We focus on preserving variation in data across multiple resolutions and multidimensional projections. We demonstrate application of this approach on a synthetic dataset to show how C_V preserves variation across multiple data resolutions, and apply it to climate model output in multidimensional projections to faciliate deeper insight in global precipitation changes. A primary observation from this work is that the statistical metric C_v preserves the variation signal, which is "washed out" when using averaging.

Due to the fact C_v is a normalized measure of variation, this approach appears to be useful as the basis for seeing and comparing variation across datasets having vastly different ranges and scales. There are many potential applications and uses of this technique, from physical to social sciences. This approach lends itself to use of field-based visualization and analysis methods; it is easily incorporated into existing visualization tools and methodologies.

Acknowledgements

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231, through the grant "Towards Exascale: High Performance Visualization and Analytics," program manager Dr. Lucy Nowell. This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. We thank Dmitriy Morozov for his insightful comments on a draft of this paper, Vincent E. Beckner for discussions concerning the idea of using C_v as the basis for visualizing variation in data.

References

[1] S. Agarwal, B. Mozafari, A. Panda, H. Milner, S. Madden, and I. Stoica. Blinkdb: Queries with bounded errors and bounded response times on very large data. In *Proceedings of the 8th ACM European Conference on Computer Systems*, EuroSys '13, pages 29–42, New York, NY, USA, 2013. ACM.

[2] J. M. Chambers, W. S. Cleveland, B. Kleiner, and P. A. Tukey. *Graphical Methods for Data Analysis*. Wadsworth, 1983.

[3] J. Clyne. Progressive Data Access for Regular Grids. In E. W. Bethel, H. Childs, and C. Hansen, editors, *High Performance Visualization—Enabling Extreme-Scale Scientific Insight*, Chapman & Hall, CRC Computational Science. CRC Press/Francis–Taylor Group, Boca Raton, FL, USA, Nov. 2012. http://www.crcpress.com/product/isbn/9781439875728, LBNL-6466E.

[4] I. Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, June 1992.

[5] I. Demir, C. Dick, and R. Westermann. Multi-charts for comparative 3d ensemble visualization. *IEEE TVCG*, 20(12):2694–2703, 2014.

[6] C. Folland, D. Stone, C. Frederiksen, D. Karoly, and J. Kinter. The International CLIVAR Climate of the 20th Century plus (C20C+). *CLIVAR Exchanges*, 19:57–59, 2014.

[7] R. B. Neale, C. Chen, A. Gettelman, P. H. Lauritzen, S. Park, D. L. Williamson, A. J. Conley, R. Garcia, D. Kinnison, J. Lamarque, et al. Description of the NCAR community atmosphere model (CAM 5.0). *NCAR Tech. Note NCAR/TN-486+ STR*, 2010.

[8] G. G. W. Services. El Niño and La Niña Years and Intensities. http://ggweather.com/enso/oni.htm, last accessed December 2015.

[9] K. Stockinger, J. Shalf, K. Wu, and E. W. Bethel. Query-Driven Visualization of Large Data Sets. In *Proceedings of IEEE Visualization 2005*, pages 167–174. IEEE Computer Society Press, October 2005. LBNL-57511.

[10] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290:2319–2323, Dec. 2000.

[11] L. van der Maaten, E. Postma, and H. van den Herik. Dimensionality Reduction: A Comparative Review. Technical report, Tilburg University Technical Report, 2009. TiCC-TR 2009-005.

[12] R. T. Whitaker, M. Mirzargar, and R. M. Kirby. Contour Boxplots: A Method for Characterizing Uncertainty in Feature Sets from Simulation Ensembles. *IEEE Transactions on Graphics and Visualization*, 19(12):2713–2722, 2013.

[13] L. Williams. Pyramidal parametrics. In *Proceedings of the 10th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '83, pages 1–11, New York, NY, USA, 1983. ACM.