Title

Statistical Projections for Multi-dimensional Visual Data Exploration

Authors

Nguyen, Hoa
Stone, Dáithí
Bethel, E Wes

Publication Date

2016-10-01

DOI

10.1109/ldav.2016.7874338

Peer reviewed

Statistical Projections for Multi-dimensional Visual Data Exploration and Analysis

Hoa Nguyen[1], Dáithí Stone[2], E. Wes Bethel[3] [1]University of Utah, Salt Lake City, UT, USA

[2]Lawrence Berkeley National Laboratory, Berkeley, CA, USA

October 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, 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.

Legal Disclaimer

Statistical Projections for 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

When working with large, multidimensional and multivariate data, science users are frequently interested in understanding variation in data, as opposed to the actual data values. Our work focuses on exploring how a simple statistical metric, the *Coefficient of Variation* (or C_v), can be used in several different ways to faciliate understanding variation in data. As a statistical measure, it offers a key advantage over more widely accepted measures like standard deviation, namely to its ability to capture local variation properties. As a multidimensional projection operator, C_v is an effective way of reducing data size while preserving the key variational signal. Visualizations produced from C_v that target conveying variation in data are highly informative, especially compared to those produced with more widely known methods. We demonstrate these ideas within the context of a two-part application case study focusing on understanding long-term trends in the the changes in precipitation and winds in large-scale climate model ensemble output.

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. Here, the objective for the visual data analysis and exploration process is to gain deeper insight into data by studying its variation, as opposed to studying absolute data values.

In data exploration and analysis, there are often instances where understanding variation in data, is of greater interest than the study of the absolute data values. For example, in analysis of ensemble collections of data, identifying regions in the domain that exhibit variability across ensemble members is often a primary objective.

Existing methods for visual exploration of ensemble collections of data often rely on measures like *standard deviation*, which is measure of global population variation. However, as a global measure, it can be impossible to interpret without additional information, such as the population mean. Such global measures can be less informative and useful than local measures. 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. Therefore, a metric sensitive to local variation in these transfers could be a more informative descriptor of how the climate is varying through time or space than a metric that focuses on global variation.

Using a two-part application case study focusing on understanding long-term trends in precipitation and wind in large-scale climate model ensemble output, 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 to show that C_v does a more effective job in capturing variation than $\bar{x}$ or σ. And that C_v is, as a derived scalar field, a more effective way to visually present variation than other commonly used methods. Second, we demonstrate 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. Collectively, these elements are a useful collection of properties in working with large, complex, multidimensional data.

The main contributions of this work are:

- A simple and widely applicable methodology of using a statistical measure of variation, C_v, for the purpose of visually conveying variational signal within large, complex, multidimensional data to identify variation of data.
- Demonstration of visual encodings and multidimensional projections that use C_v, which can help users quickly interact with data and efficiently perform visual data analysis and exploration tasks.
- A two-part case study that shows these methods in use to study long-term trends in precipitation and winds from large-scale climate model ensemble output.

2 BACKGROUND AND RELATED WORK

2.1 Computing and Visualizing Variation

There are several different approaches to computing and displaying variation in data. One of the earliest methods for displaying data population characteristics, including variation, is the *box plot*. 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 [17] extended this idea to depict variation in data features. One limitation of box plots is they are useful for presenting a small number of samples, and attempts to use this, or other glyph-based methods, on large collections of data will result in excessive visual clutter.

Another approach is to compute a scalar field that is representative of variation in a dataset, then use traditional techniques for display. Potter et al., 2009 [10] present a system, Ensemble-Vis, that uses this idea; it computes and displays variation in data, with a particular focus on ensemble collections of climate model output. While Ensemble-Vis uses several different linked views and interaction methods to facilitate user exploration, at the core of their approach is the use of mean and standard deviation as statistical

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

metrics. More recently, Demir et al., 2014 [5] use traditional scalar color mapping techniques in conjunction with brushing and linking to enable visual exploration of variation in ensemble collections of simulation output.

Pfaffelmoser also proposed method to visualize contour distributions in 2D ensemble data [9]. This paper makes no assumption about a stochastic uncertainty model, rendering it suitable for arbitrary ensemble distributions. It computes a statistical summary (probability density) of the ensemble over the spatial domain, including probability density values for arbitrary domain points. From this information, the uncertainty and topology of iso-contours can be determined, as well as the variations in gradient magnitude around these contours.

Previous use of C_v in the visualization community is quite limited. Shen et al., 1999 [12] use C_v within the context of creating a data structure to accelerate volume rendering. The idea is to use C_v as a measure of spatial and temporal regularity.

The approach we are taking our work is to focus on using C_v, rather than standard deviation, as the measure of variation in a collection of data for the purposes of visualizing data variation, as well as the basis of multi-variate data projection. After computing C_v measure of variation in a dataset, we have a scalar field, which is suitable for use with any traditional scalar field visualization method.

2.2 Projection and Data Reduction Methods

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 [19] 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. Here, the *average*, or $\bar{x}$ operator serves as the data reduction operator.

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 proven 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. 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 $m < n$. One approach to performing a projection is to extract a spatially constrained subsampling of data, like an orthogonal or arbitrary slice. Other approaches, like Principal Component Analysis (PCA) [1] or Isomap [13], 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. See Maaten et al., 2009 [15], for a comparative review of these methods. We are interested in different problems, namely presentation of variation and preserving variation across multiple scales and through different data-reducing projection operators. 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 our work, we use the term projection to refer to the process of converting a dataset of R^n dimensions to R^m dimensions, where $m < n$. Unlike subsampling approaches, such as those described above, we use one of several different projection operators—mean ($\bar{x}$), standard deviation (σ), and Coefficient of Variation (C_v)—to go from R^n to R^m.

3 Computing Variation

The concepts of variance, and variation, have deep roots in statistics. At its core, variance is a measure the of dispersion of data values in a population. It is computed as and is a measure of the degree to which individual data values, x, deviate from the population mean, $\bar{x}$. Closely related to variance, the *standard deviation* (σ) is a metric that also indicates the amount of dispersion. For datasets that follow a normal distribution, the size of the standard deviation indicates how tightly clustered the population data is about the mean. 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, the σ will be quite large.

The σ and Coefficient of Variation (C_v) (Eq. 1) quantities are also related; the C_v is essentially a normalized form of σ:

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

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.

With these three measures—V, σ, and C_v—what are the advantages and disadvantages of each as a measure of variation in data?

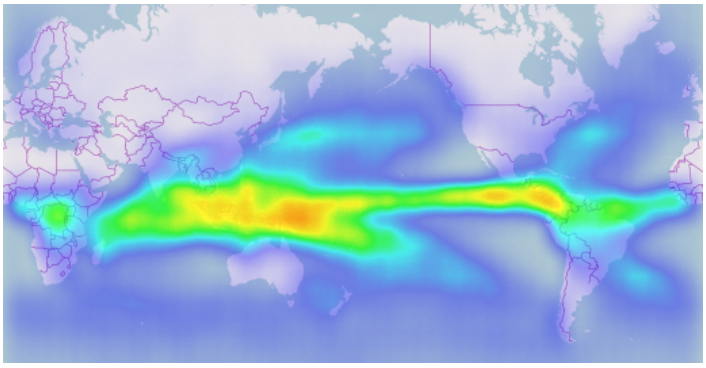
While both V and σ are numerically stable, they also require some knowledge about the underlying data to be useful. For example, if someone tells you $\sigma = 100$, is that a large or small value? To know the answer, you'd have to know $\bar{x}$. If, for example, $\bar{x} = 10^6$, then $\sigma = 100$ is a very small amount of variation. On the other hand, if $\bar{x} = 200$, then $\sigma = 100$ is a huge amount of variation.

On the other hand, C_v, being a normalized form of σ, does not require any knowledge of $\bar{x}$ to understand. Using the examples above, where $\bar{x} = 10^6$ and $\sigma = 100$, then $C_v = 0.0001$; and where $\bar{x} = 200$ and $\sigma = 100$, then $C_v = 0.5$. This example illustrates why simply knowing σ by itself is only of limited use.

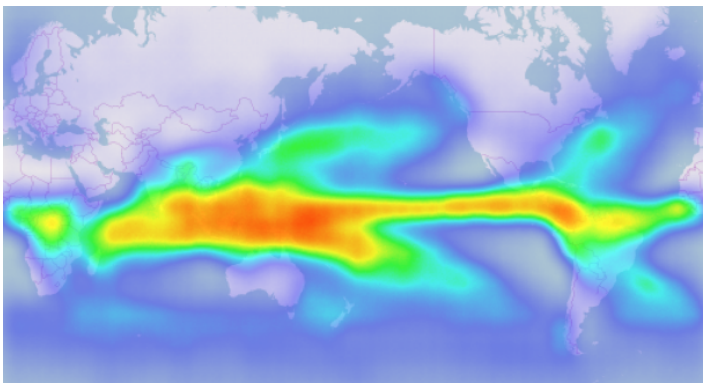
Despite its advantages, C_v does have a clear disadvantage: it becomes increasingly large as $\bar{x}$ goes to 0.0, and is undefined when $\bar{x}$ is 0.0. Among various workarounds one might consider would be detecting this condition and then adding some constant C to all data values, which would shift the mean away from 0.0 and also would cause any change to σ. That approach would have the effect of eliminating the effects resulting from a small, or zero, denominator, while leaving the underlying variation present in data unaffected.

4 Case Studies

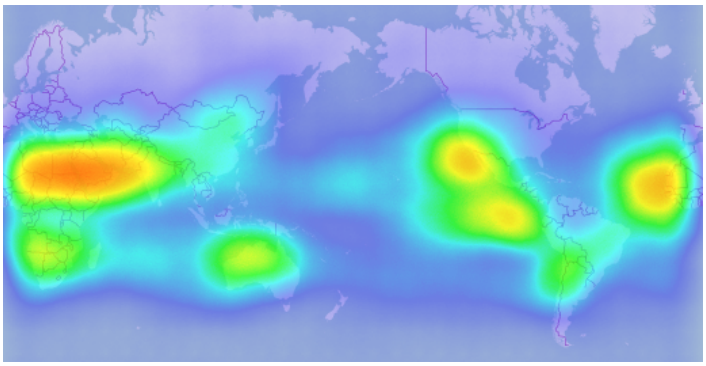
This case study focuses on exploring how the C_v can reveal features and characteristics that would otherwise not be visible using only $\bar{x}$



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



(b) σ, lat/lon projection through all ensemble members across all years.



(c) C_v, lat/lon projection through all ensemble members across all years.

Figure 1: Comparison of $\bar{x}$, σ, and C_v as the basis of a spatial projection of climate model output, where we go from 4D to 2D. The C_v projection shows specific features not visible in either the $\bar{x}$ or σ projections, which are both similar in appearance.

or σ in large-scale, complex climate model output. The case study consists of two parts, one focusing on precipitation (§4.1) and one focusing on winds (§4.2).

We use precipitation and wind speed data generated by the CAM5.1 global atmospheric climate model [7] run at approximately $1^\circ \times 1^\circ$ longitude-latitude resolution under observed boundary conditions from the period 1959-2014 [6]. Output from this run consists of multivariate, four-dimensional data: latitude, longitude, time, ensemble member. The size of precipitation data over 50 runs is 7.4 GBytes and the size of wind speed data is 122 GBytes.

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.

For both precipitation and wind studies, we are using the same general approach: produce different types of projections (spatial, temporal) using different projection operators ($\bar{x}$, σ, C_v), and make observations about the differences in science that emerge from each type of projection. In both cases, it turns out that C_v is able to reveal specific scientific features that are not present in the other two types of images, suggesting that C_v is quite useful in helping facilitate scientific knowledge discovery.

4.1 Precipitation

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 in some places 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 [8]. The C_v is closely related to the calculation of these fractional values.

This case study focusing on precipitation has two lines of exploration: space and time. The key idea in both investigations is that C_v reveals information that is not apparent using either $\bar{x}$ or σ.

4.1.1 Spatial Projections

To begin, we compare spatial projections of $\bar{x}$, σ, and C_v, shown in Fig. 1. Here, we are projecting climate data from a 4D space (latitude, longitude, time, ensemble member) to a 2D space (latitude, longitude). For each lat/lon point, we compute the projected value as $p = f(T,E)$ across all times T and ensemble members E, where $f \in [\bar{x}, \sigma, C_v]$.

The images of $\bar{x}$ and σ precipitation (Figs 1a and 1b) show 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 σ simply shows that areas with large precipitation amounts have freedom for large variability.

The image of C_v (Fig. 1c) 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 of higher 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.

4.1.2 Temporal Projections

The primary focus of this part of the case study is to facilitate visual comparison of the variability in climate model precipitation calculations with an observed measure of climate variability, the Oceanic Niño Index (ONI). The ONI is a metric of the shift between El Niño (warm) and La Niña (cool) events in the tropical Pacific [11]. 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.

There were major El Niño events during the years 1983 and 1998. Those major weather events resulted in substantial increases in precipitation in parts of the world, and are represented through exceptionally high ONI values during those years. We use this information in the examples that follow to look for visual correlation between precipitation variability as represented in different types of temporal projections and known major weather events.

We begin with temporal projections from a 4D space (latitude, longitude, time, ensemble) to a 1D space (time), shown in Fig 2. For each time value T, we compute $p = f(S_{lat}, S_{lon}, E)$ across all spatial locations (S_{lat}, S_{lon}) and ensemble members E, where $f \in [\bar{x}, \sigma, C_v]$. Since the data is computed and stored at monthly temporal resolution, our computations produce a yearly value from monthly values.

In the plot of $\bar{x}$ (blue bars in Fig. 2a), there is relatively little variation visible in the mean from year-to-year, with the main feature being a gradual long-term trend of increasing precipitation levels. Comparing these mean yearly values with the ONI and the the major El Niño events of 1983 and 1998, which are reflected with exceptionally high ONI values during those years, there is no visible correlation between yearly $\bar{x}$ and those high ONI values. Similar to the $\bar{x}$ plot, the σ plot (purple bars Fig. 2b) shows the little variation in the σ from year-to-year, and there is nothing remarkable about the σ during the major events of 1983 and 1998.

In contrast, looking at the C_v projection in Fig. 2c, these two major 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).

For the sake of completeness, we present a boxplot presentation of yearly precipitation values in Fig. 2d, along with annotation showing the major El Niño events of 1983 and 1998. Here, the box attributes are computed as yearly mean, min, max, and quartiles from the monthly precipitation model data, across all ensemble members. From this image, there is no visible evidence of anything remarkable happening in terms of precipitation variability associated with the major events of 1983 and 1998. The conclusion here is that visualization method, i.e., bar chart vs. boxplot, is not the key issue. The key issue is that C_v reflects data characteristics in a way not possible with either $\bar{x}$ or σ.

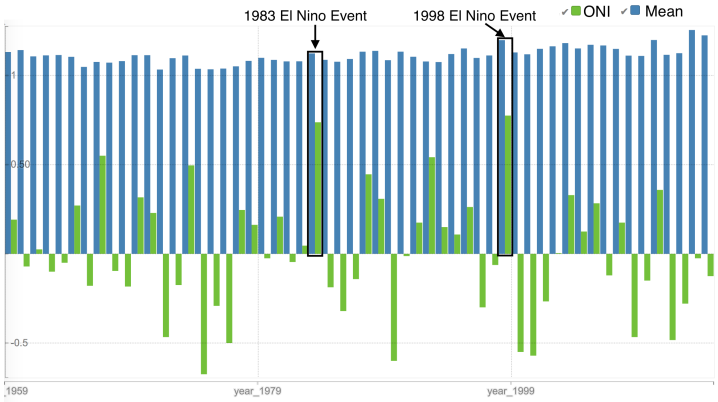
The properties of the C_v map in Fig. 1c help to explain the behavior of the yearly bar plots of C_v in Fig. 2c. In essence, C_v is acting as a combined index of the occurrence of El Niño events and of anomalous rainfall over subtropical regions. If data were only retained over the ocean, the C_v projection onto time would likely improve as an index of El Niño variability. On the other hand, if data were only retained over land (to mask out the El Niño aspect) then it would provide a metric of variations in subtropical deserts, without any parametric definition of what constitutes a subtropical desert. In contrast, the yearly bar plots of the mean and standard deviation are mostly reflecting activity in the ITCZ.

4.2 Winds

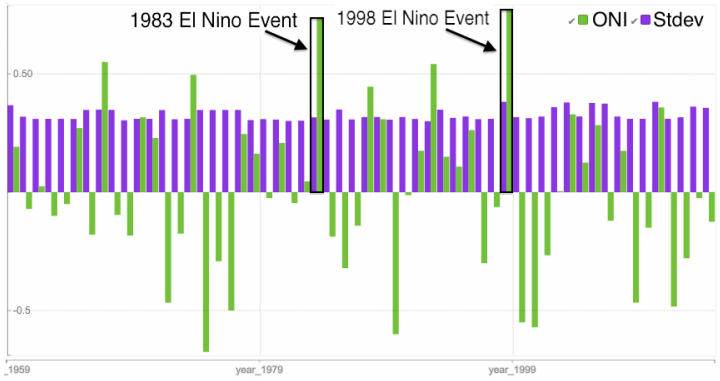
We now explore these projections for wind speed data from the climate model simulations. The data is the monthly average wind speed on the 500 hPa surface, the pressure surface that is about half the pressure at sea level and which lies approximately 5.5 km above sea level. The images in Fig. 3 are spatial projections, from a four-dimensional space—two spatial dimensions, time, and ensemble members—down to a two-dimensional latitude/longitude projection.

The most prominent features in the map of the mean winds are the mid-latitude jet streams. These winds are strongest over the ocean, flow from west to east in the 40°–50° latitude range, and extend down to the surface (hence named the "Roaring 40s" in the Southern Hemisphere). These appear as horizontally oriented regions of red in Fig. 3a, the projection of mean wind speed, $\bar{x}$.

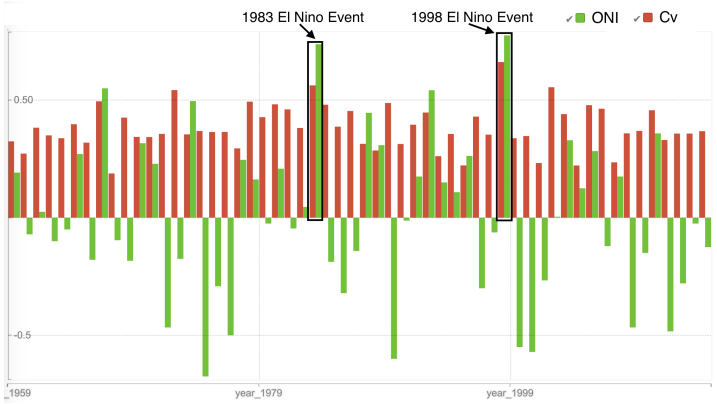
The map of σ (Fig. 3b) appears to show that the jets over the North Pacific and North Atlantic are variable, while the southern jet is instead quite steady except in the South Pacific. However, the C_v map (Fig. 3c) indicates that the spatial alignment of fea-



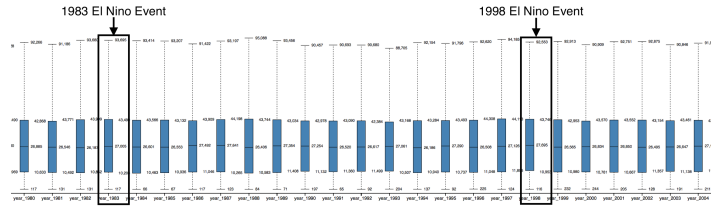
(a) $\bar{x}$.



(b) σ.

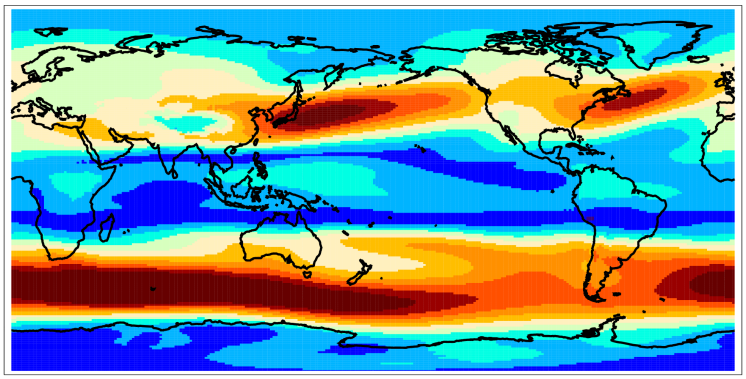


(c) C_v.

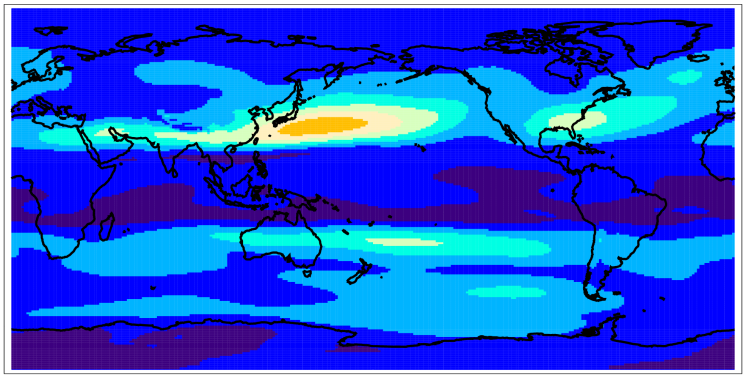


(d) Box plots for precipitation from year 1980 to 1994

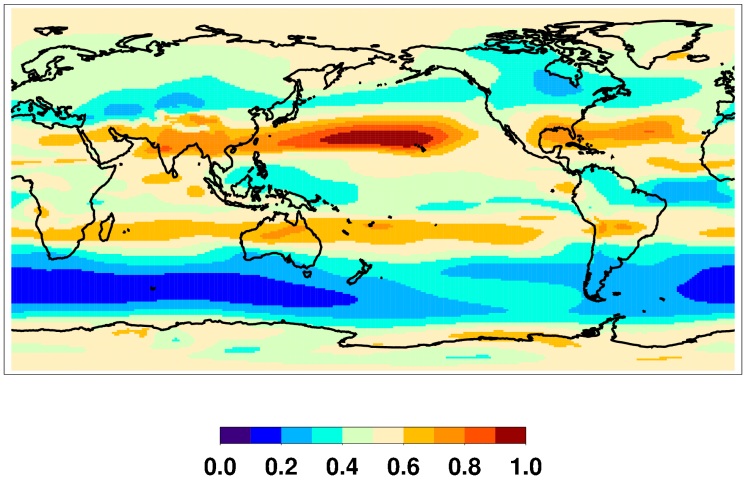
Figure 2: Comparison of $\bar{x}$, σ, and C_v as the basis for temporal projection operators, where we project from all spatial locations and ensemble members to yearly values. We show these temporal projections in comparison with the ONI. Of these projections, the C_v projection shows the strongest correlation with ONI, which is a known measure of climate variability.



(a) $\bar{x}$



(b) σ



(c) C_v

Figure 3: 2D spatial projections of 500 hPa wind speed from a 4D space. Comparison of the C_v map against the $\bar{x}$ map reveals that the strong mid latitude winds have a tendency to expand equatorward but not poleward, some thing that is harder to distinguish in the σ map.

tures is not perfect. In fact the jet cores over the North Pacific, North Atlantic, and Antarctic Oceans are all very steady. The variation instead comes from a tendency of the winds to expand toward the equator: there is little or no power on the poleward side of the jets. It is well known that the jets vary in their north-south position, with those variations so prominent that they form the leading Principal Components of extratropical variability, often termed the "Southern Annual Mode" (Antarctic), the "Pacific/North America Pattern" (North Pacific), and the "North Atlantic Oscillation" (North Atlantic); the northern two PCs are sometimes merged into the "Northern Annular Mode" [16, 14]. However, the point that these variations are manifest through an equatorward expansion of the winds, and not through a poleward expansion or through north-south shifts of the jet core, is not something that is readily apparent in the patterns associated with the PCs, which are themselves ignorant of the context of the mean base flow. Comparison of the σ and $\bar{x}$ maps can reveal the equatorward tendency, but it requires careful scrutiny. On the other hand, by stressing the differences between the $\bar{x}$ and σ maps, the high-value regions in the C_v map are more clearly displaced from those in $\bar{x}$ map, meaning that the asymmetry in the north-south movement of the jets is apparent even in just a casual comparison.

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. When looking at a science problem that focuses on studying variation in data, our approach is to focus on use of Coefficient of Variation (C_v) as a measure of variation, and show that it is capable of revealing information in data in a way not possible with the more commonly used *standard deviation*. Specifically, our case studies show that C_v reveals features in two different fields that are not visible using *standard deviation* or $\bar{x}$ as the basis for computing variation or the basis for dimension-reducing projections.

This idea, using C_v as the basis for computing variation and as the basis for doing multi-dimensional projection, is a simple one, but highly effective. Of the visual examples, the winds data results shown in Fig. 3 were produced by our climate science collaborator using the CDAT toolkit [18] and a small amount of custom Python code. We hope that this method will be useful to many other science applications, and that due to its simplicity, can be adopted and used by many different existing visualization tools.

While C_v does have some known shortcomings, these can be avoided or worked around. And C_v, as a normalized measure of variation, does hold promise as a vehicle for comparing variation in datasets having 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. A promising avenue for future work would be to explore this idea, comparison of variation across datasets having vastly different properties.

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.

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] M. New, M. Hulme, and P. Jones. Representing twentieth-century space-time climate variability. Part II: Development of 1901-96 monthly grids of terrestrial surface climate. *J. Climate*, 13:2217–2238, 2000.

[9] T. Pfaffelmoser and R. Westermann. Visualizing contour distributions in 2d ensemble data. In *EuroVis-Short Papers*, pages 55–59. The Eurographics Association, 2013.

[10] K. Potter, A. Wilson, P. T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. R. Johnson. Ensemble-vis: A framework for the statistical visualization of ensemble data. In *2009 IEEE International Conference on Data Mining Workshops*, pages 233–240, Dec 2009.

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

[12] H.-W. Shen, L.-J. Chiang, and K.-L. Ma. A fast volume rendering algorithm for time-varying fields using a time-space partitioning (tsp) tree. In *Proceedings of the Conference on Visualization '99: Celebrating Ten Years*, VIS '99, pages 371–377, Los Alamitos, CA, USA, 1999. IEEE Computer Society Press.

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

[14] D. W. J. Thompson and J. M. Wallace. Annular modes in the extratropical circulation. Part I: Month-to-month variability. *J. Climate*, 13:1000–1016, 2000.

[15] 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.

[16] J. M. Wallace and D. S. Gutzler. Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Wea. Rev.*, 109:784–812, 1981.

[17] 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.

[18] D. Williams, C. Doutriaux, J. Patchett, S. Williams, G. Shipman, R. Miller, C. Steed, H. Krishnan, C. Silva, A. Chaudhary, P.-T. Bremer, D. Pugmire, E. W. Bethel, H. Childs, M. Prabhat, B. Geveci, A. Bauer, A. Pletzer, J. Poco, T. Ellqvist, E. Santos, G. Potter, B. Smith, T. Maxwell, D. Kindig, and D. Koop. Ultrascale Visualization of Climate Data. *IEEE Computer*, 46(9):68–76, Sept. 2013.

[19] 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.