Visualization-Driven Structural and Statistical Analysis of Turbulent Flows

Kenny Gruchalla[1], Mark Rast[2], Elizabeth Bradley[1], John Clyne[3], and Pablo Mininni[4]

[1] Department of Computer Science, University of Colorado, Boulder, Colorado
[2] Laboratory for Atmospheric and Space Physics, Department of Astrophysical and Planetary Sciences, University of Colorado, Boulder, Colorado
[3] Computational and Information Systems Laboratory, National Center for Atmospheric Research, Boulder, Colorado
[4] Departamento de Física, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Argentina and Geophysical Turbulence Program, National Center for Atmospheric Research, Boulder, Colorado

Abstract. Knowledge extraction from data volumes of ever increasing size requires ever more flexible tools to facilitate interactive query. Interactivity enables real-time hypothesis testing and scientific discovery, but can generally not be achieved without some level of data reduction. The approach described in this paper combines multi-resolution access, region-of-interest extraction, and structure identification in order to provide interactive spatial and statistical analysis of a terascale data volume. Unique aspects of our approach include the incorporation of both local and global statistics of the flow structures, and iterative refinement facilities, which combine geometry, topology, and statistics to allow the user to effectively tailor the analysis and visualization to the science. Working together, these facilities allow a user to focus the spatial scale and domain of the analysis and perform an appropriately tailored multivariate visualization of the corresponding data. All of these ideas and algorithms are instantiated in a deployed visualization and analysis tool called VAPOR, which is in routine use by scientists internationally. In data from a 1024^3 simulation of a forced turbulent flow, VAPOR allowed us to perform a visual data exploration of the flow properties at interactive speeds, leading to the discovery of novel scientific properties of the flow, in the form of two distinct vortical structure populations. These structures would have been very difficult (if not impossible) to find with statistical overviews or other existing visualization-driven analysis approaches. This kind of intelligent, focused analysis/refinement approach will become even more important as computational science moves towards petascale applications.

1 Challenges to data analysis

A critical disparity is growing in the field of computational science: our ability to generate numerical data from scientific computations has in many cases

exceeded our ability to analyze those data effectively. Supercomputing systems have now reached *petaflop* performance [1], supporting numerical models of extraordinary complexity, fidelity, and scale. In supercomputing centers, terabyte data sets are now commonplace and petabyte data sets are anticipated within a few years. However, analysis tools and the computational machinery that supports them have not been able to scale to meet the demands of these data. For many computational scientists, this lack of analysis capability is the largest barrier to scientific discovery.

The imbalance of scale between numerical simulation and data analysis is largely due to their contrasting demands on computational resources. Large-scale numerical simulation is typically a *batch* processing operation that proceeds without human interaction on parallel supercomputers. Data analysis, in contrast, is fundamentally an *interactive* process with a human investigator in the loop, posing questions about the data and using the responses to progressively refine those questions[2]. While some data analyses certainly can be performed in batch mode, this is only practical for completely predetermined investigations. Exploratory analysis depends on hypothesis generation and testing, which requires an interactive environment that can provide timely and meaningful feedback to the investigator. Unfortunately, this kind of interactive workflow is not well-suited to batch access on a parallel supercomputer. Another key bottleneck in the analysis process is data storage. If the data exceeds the size of the available random access media, one must manage its storage and exchange across different media. Disk transfer rates are generally inadequate to support interactive processing of large-scale computational data sets.

These dilemma can be addressed by data reduction, and a variety of schemes have been developed to reduce data volumes while maintaining essential properties. Their methods and results depend on the scientific goals of the simulation and analysis. For example, in the investigation of turbulent flows, analysis of strictly statistical or spectral properties can enable significant reduction in data dimensionality, while analysis of local flow dynamics, thermodynamics, or stability does not. In the later cases, the only solution is to reduce the physical volume under analysis. There are two general classes of methods for this. First, one can isolate and extract local sub-regions from the global domain. The success of this strategy depends on locating those regions in the solution that are of particular scientific importance or interest, which is a real challenge to intelligent data analysis. There has been some very interesting work in the IDA community on dimensional reduction for this purpose[3, 4]. The visualization-driven approach described in this paper extracts regions of interest using an iterative interactive filtering technique that employs a combination of global and local flow statistics. The second class of data-volume reduction techniques uses a coarsened global approximation of the discrete solution to subsample the data over the entire domain. The obvious challenge here is selecting an appropriate coarsening (method and scale) to maintain accuracy — and confidence in the results.

Both of these data-reduction techniques have been implemented in VAPOR, an open-source desktop visualization-driven analysis application (available at

http://www.vapor.ucar.edu). It closely couples advanced visualization with quantitative analysis capabilities, and it handles the complexities of large datasets using a hierarchical data model. It is designed to support a multi-phase analysis process, allowing the investigator to control a speed-and-memory versus a locus-and-quality trade off. This is an ideal context within which to explore intelligent data reduction. Using the ideas described in the previous paragraph, we have extended VAPOR's capabilities to include definition, manipulation, and refinement of feature sub-spaces based on multi-scale statistics of turbulent structures contained within the data. The user can explore the data volume at a coarsened resolution to gain a qualitative understanding and identify regions/structures of interest. Those regions/stuctures can then be investigated at progressively higher resolutions with computationally intensive visualization and analysis performed on progressively smaller sub-domains or structure populations. The base functionality of VAPOR provides data coarsening in the form of multi-resolution access via wavelet decomposition and reconstruction [5]. VAPOR's data coarsening approach coupled with a simple a simple sub-domain selection capability and has a successful track record in the analysis of large-scale simulation data using only modest computing resources [6]. With the addition of intelligent region-of-interest extraction, VAPOR can now provide interactive, scientifically meaningful explorations of tera-scale data volumes.

The following Section describes VAPOR's visualization-driven analysis capabilities; Section 3 demonstrates their power using data from a 1024^3 forced incompressible hydrodynamic simulation.

2 VAPOR: A desktop analysis and visualization application

Many applications have been developed specifically for the visualization and analysis of large-scale, time-varying numerical data, but all of them have significant limitations in terms of visualization, analysis, and/or scalability. In many of these tools, the emphasis is on the algorithms and the generation of aesthetic images, rather than on scientific discovery [7]. Visualization-oriented applications, like Paraview [8], Visit [9], and Ensight, lack quantitative analysis capabilities, and many of them demand specialized parallel computing resources. High-level, fourth-generation data languages such as ITT's IDL and Mathworks's Matlab are on the opposite end of the analysis-visualization spectrum. They provide a rich set of mathematical utilities for the quantitative analysis of scientific data but only limited visualization capabilities, and they do not scale well to very large data sets.

The goal of the VAPOR project was to address these shortcomings. It provides an integrated suite of advanced visualization capabilities that are specifically tailored to volumetric time-varying, multivariate numerical data. These capabilities, coupled with the intelligent data reduction strategies introduced in the previous sections, allow investigators to rapidly identify scientifically meaningful spatial-temporal regions in large-scale multivariate data. VAPOR's design

— both its functionality and its user interface — is guided by a steering committee of computational physicists to ensure that it truly meets the needs of the end-user community. Scalability is addressed through a multi-phase analysis process that is based on a combination of region-of-interest isolation, feature extraction (with our extensions), and a hierarchical data model. Finally, VAPOR interfaces seamlessly with high-level analysis languages like IDL, allowing its user to perform rigorous quantitative analyses of regions of interest.

2.1 Visualization

VAPOR incorporates a variety of state-of-the-art volume rendering and flow-visualization techniques, including both *direct* and *indirect* volume rendering [10]. Direct volume rendering describes a class of techniques, which generate images directly from volumetric data without any intermediate geometric constructions, while indirect volume rendering constructs geometric *isosurfaces*. To support the visualization of vector fields, VAPOR provides both *sparse* and *dense* particle-tracing methods [11]. The former render the geometry of individual trajectories of particles seeded in a flow field, and can support both steady (time-invariant) and unsteady (time-varying) trajectory integration. Dense particle-tracing methods synthesize textures that represent how the flow convolves input noise.

VAPOR's integrated design allows these volume rendering and flow-visualization techniques to be used in different combinations over a single analysis run, in concert with the intelligent data reduction strategies described in Section 2.3, as the investigator progressively isolates and refines scientifically meaningful regions of the data. The hierarchical data model that supports this is described in the next Section. By utilizing the optimized data-parallel streaming processors of modern *graphics processing units* (GPUs), VAPOR can effectively work with volumes of the order of 1536^3 [6].

2.2 Hierarchical data model

The VAPOR data storage model is based on wavelet decomposition [5, 12]. Data are stored as a hierarchy of successively coarser wavelet coefficients; each level in this hierarchy represents a halving of the data resolution along each spatial axis, corresponding to an eight-fold reduction in data volume. In this manner, VAPOR maintains a series of useful coarsened approximations of the data, any of which can be accessed on demand during an analysis run, without an undue increase in storage requirements. Wavelet data are organized into a collection of multiple files: one binary file containing the wavelet coefficients for each time step, each variable, and each wavelet transformation level, and a single metadata file that describes the attributes of the field data (e.g., the grid type, the time steps, the spatial resolution, the field names, etc.).

This storage model naturally supports intelligent, interactive data decomposition. It allows VAPOR to operate on any subset of time steps, variables, and wavelet transformation levels, which has a variety of important advantages, including iterative focus and refinement of the analysis effort. An investigator

can control the level of interactivity by regulating the fidelity of the data, first browsing a coarsened representation across the global spatial-temporal domain to identify regions or features of interest and then examining those reduced domains in greater detail. The hierarchical data/metadata combination also allows VAPOR to work with very large data collections, as data components can be stored off-line until required, handling *incomplete* data sets smoothly.

2.3 Multivariate feature extraction

The VAPOR volume rendering capability forms the basis for the multivariate feature extraction technique we have implemented to isolate structures of interest in large data sets. A multidimensional *transfer function* is used to define a mapping from data values to the color and opacity values in the volume rendering. The opacity component of this function visually separates the volume into opaque features. VAPOR users can construct and refine these functions iteratively, and use them in different thresholding schemes to visually separate the volume into opaque regions.

Once these regions have been visually identified using the transfer function, the individual structures are extracted and tagged using a connected-component labeling algorithm [13], an image-processing technique that assigns groups of ϵ-connected data points[5] a unique feature label. Once the features have been identified in this manner, they can be visualized and analyzed as individual features oppose to a set of opaque voxels. The individual features can be visualized in isolation or as members of sub-groups, and the data points and geometry of each can exported to external analysis packages for further analysis, as described in Section 2.4. This allows any user-defined physical property to be computed on the associated data points contained in each feature, over any field or combination of fields in the data. VAPOR presents the resulting values and distributions to the user as a table of feature-local histograms and property values, as shown in Figure 1. Using this table, the set of features can be culled based on the central moments of their distributions to further focus the study. The entire reduction process—including the transfer function design and feature definition—can be iterated to progressively refine the analysis, providing insight into the multivariate properties of structures across multiple scales.

2.4 Coupled visual, quantitative, and statistical analysis

Understanding large-scale simulation data is an exploratory process that can be greatly facilitated by combining highly interactive, qualitative visual examination with quantitative numerical analysis. Visualization can be used to motivate analysis through the identification of structures in the data, giving rise to hypotheses that can be validated or rejected through numerical study. Likewise, the analysis can be used to drive the visualization, identifying salient quantitative characteristics of the data through numerical study, and then visualizing their associated geometric shapes and physical properties. VAPOR's design

[5] i.e., those that are connected by an ϵ chain

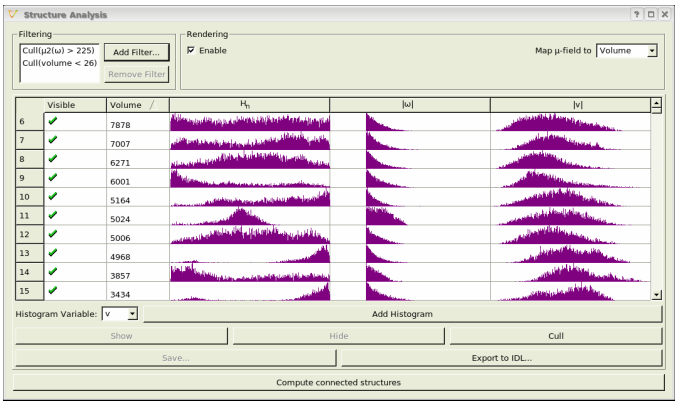

Fig. 1. The VAPOR structure analysis dialog, which displays feature-local histograms of user-selected field distributions.

seamlessly combines qualitative visual and quantitative numerical investigation, enabling its users to interactively transition between the two. Its multi-resolution visualization and region-of-interest isolation capabilities, in conjunction with its hierarchical data representation, allow its users to cull data intelligently and pass appropriate subsets to an external quantitative analysis package.

Smooth integration of all of these capabilities required some interesting design decisions. VAPOR performs GPU-accelerated visualization natively, as described in Section 2.1, and hands numerical analysis off to IDL. VAPOR and IDL sessions are run simultaneously; after regions of interest are identified in the former, the associated data volumes are exported via metadata descriptors to the latter for further study. The tight coupling between IDL and VAPOR is accomplished by a library of data-access routines, which allow IDL access to the wavelet-encoded data representation. (This approach is readily generalizable to other analysis packages, complementing and enhancing existing user capabilities.) The qualitative/quantitative tandem is very effective: IDL, as mentioned at the beginning of Section 2, does not scale well to large data sets [12], but VAPOR's ability to focus the study neatly bypasses that problem, and the results of IDL analysis on focused regions can be seamlessly imported back into the VAPOR session for visual investigation. By repeating this process, very large data sets can be interactively explored, visualized, and analyzed without the overhead of reading, writing, and operating on the full data volume.

3 Application to vortical structures in Taylor-Green flow

As an illustration of the power of the ideas described in the previous sections, we use VAPOR to explore data from an incompressible Taylor-Green forced turbulence simulation with a microscale Reynolds number of $R_\lambda \sim 1300$ [14]. The particular structures in this data that are of scientific interest involve vortic-

ity, but the volume contains so many of these structures, of different sizes and strengths, as to pose a truly daunting analysis problem. The small-scale structures are particularly hard to isolate, so that is what we set out to analyze with VAPOR.

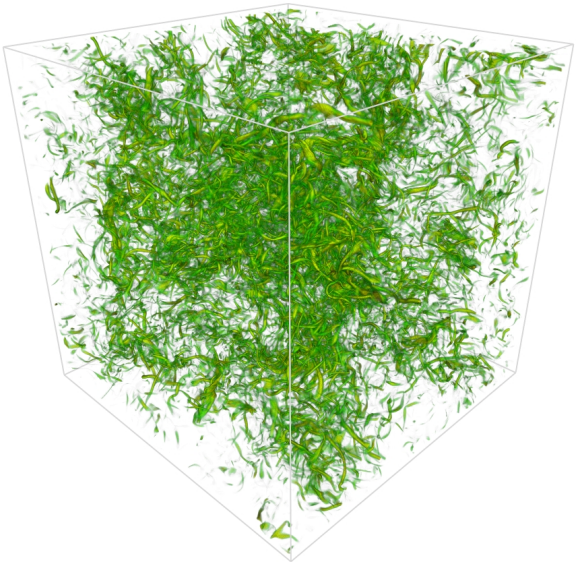

Fig. 2. A volume rendering of areas of strong vorticity in Taylor-Green turbulence isolates tens of thousands of vortical structures.

3.1 Global vorticity and structure identification

Vortices play important roles in the dynamics and transport properties of fluid flows, but they are surprisingly hard to define, which complicates the task of designing a vortex extraction method. Jiang et al. [15] provide an extensive survey of current techniques. As a working definition, we treat a vortex filament, tube, or sheet as *a connected region with a higher relative amplitude of vorticity than its surrounding* [16]. Many vortex detection and visualization methods use the same definition, and most of them operationalize it by thresholding the magnitude of the vorticity. This is the starting point for our analysis, but VAPOR's capabilities allowed us to add other scientifically meaningful analysis steps and iteratively focus the process. In this particular case, it allowed us to investigate the correlation between vorticity and helicity across multiple scales and discover important structural properties that were previously unknown.

The first step in the process is to threshold the vorticity of the Taylor-Green data using the opacity contribution of the multidimensional tranfer function. The fields in the data include the simulated velocity vector field and two derived

fields: a vorticity vector field and a normalized helicity field. Vorticity is defined as the curl of a velocity field, $\boldsymbol{\omega} = \nabla \times \boldsymbol{v}$, characterizing the pointwise rotation of fluid elements. Helicity is a scalar value, $H_n = \frac{\boldsymbol{v} \cdot \boldsymbol{\omega}}{|\boldsymbol{v}||\boldsymbol{\omega}|}$, the cosine of the angle between velocity and vorticity. An initial vorticity threshold was chosen to begin separating the tube-like vortical structures in the data volume. This step isolates tens of thousands of vortical structures, as shown in Figure 2. Using VAPOR's iterative refinement capabilities, we focus the study by further considering the helicity within these structures. A global analysis across the entire data volume, Figure 3, shows that both helicity and its pointwise correlation with vorticity are distributed in a nearly uniform fashion—i.e., that all angles between velocity and vorticity vectors occur with similar frequencies across all values of vorticity. While this is a useful result, it lumps the whole data volume together, possibly obscuring important local differences. Using VAPOR to generate feature-local histograms, we find that different high-vorticity regions do indeed have distinct helicity distributions, Figure 3(c). Three populations of structures are conspicuously evident: those whose helicity distributions span the full range with no distinct peak, those with a peak at high absolute values of helicity (i.e., dominated by nearly aligned or anti-aligned velocity and vorticity vectors), and those whose helicity distributions peak near zero (i.e., dominated by nearly orthogonal velocity and vorticity).

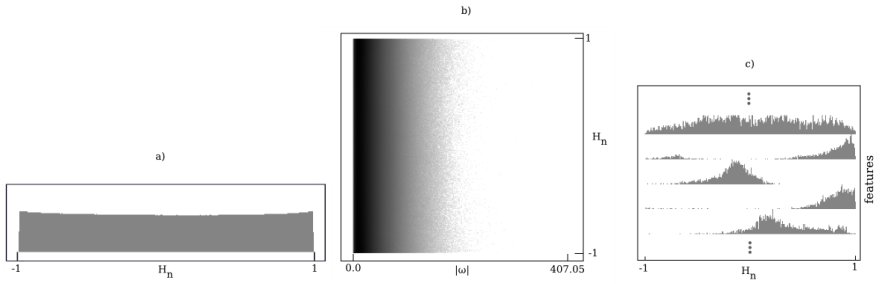

Fig. 3. The relationship between vorticity and helicity in Taylor-Green Turbulence a) the histogram of global normalized helicity, indicating helicity, measured point wise in the domain, has a nearly uniform distribution; b) the scatter plot of vorticity magnitude versus normalized helicity, showing that helicity has a nearly uniform distribution across all values of vorticity; c) a selected subset of the feature-local helicity histograms from features defined by high vorticity that show individual regions of strong vorticity have distinct helicity distributions.

Using VAPOR's intrinsic capabilities, we have thus effectively differentiated the regions of strong vorticity into three structure populations, based on their helicity distributions. In order to investigate the statistics and local dynamics of these structures, we next extend the analysis through a combination of visualization in VAPOR and focused study in the coupled quantitative analysis

package. By visualizing individual features in isolation, we find that the wide noisy distributions belong to composite structures that were not well separated into individual components by the original vorticity thresholding, while the other two populations are those of individual tube-like structures. This result allows us to further cull the dataset and focus on the tube-like structures with either high or low helicity magnitude. Both populations have similar geometries, but streamlines seeded in these regions, as shown in Figure 5, reveal that their flow properties are quite different. In the low-helicity tubes, the streamlines twist around the core; in the high-helicity tubes the streamlines more closely follow the writhe of the tube.

Further interactive analysis of these distinct vortex structures can proceed either by examining the statistical properties of the population or the detailed dynamics of any one of them. Looking first at statistics of the population of vortical structures as a whole, we note that, while structures with all values of helicity exist, there seems to be a small deficit of those with high absolute mean value compared to the point-wise helicity distribution (Figure 4a). Moreover, the helicty of any given structure is well defined and symetrically distributed about its mean value (Figure 4b and 4c). The helicity distribution within a great majority of the structures has both small variance and skewness.

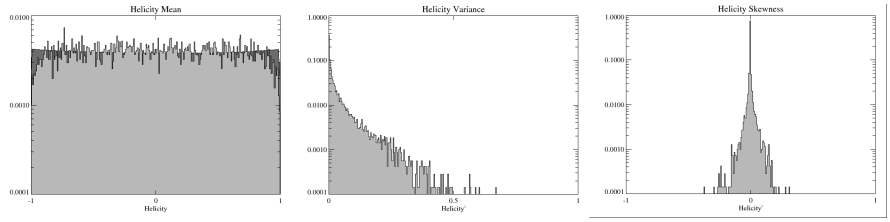

Fig. 4. Distributions of the first three central moments of the feature-local helicity distributions.

The detailed dynamics underlying any single vortex structure is also accessible. By exporting planar cross sections through tubes using VAPOR's cross-section capability, average radial profiles of the helicity and vorticity can be constructed (Figure 5c & 5d). Distinct differences between the maximally and minimally helical structures are apparent. The maximally helical structure has one sign helicity throughout, while the minimally helical twisted structure shows a change in the sign of helicity near its border (Figure 5c). This appears to be associated with inward (toward the pinched section midway along the tube) axial flow surrounding the outside of the vortex tube and outward (diverging from a pinched section midway along the tube) axial flow in its core, (Figure 6). A temporal history of these flows would be critical in confirming what looks to be a significant vorticity amplification mechanism in this minimally helical vortex filament. Also critical in future analysis would be the ability to combine the statistical and dynamical analyses presented here to determine how common

this mechanism is and whether particular dynamical processes are statistically linked to specific structure populations.

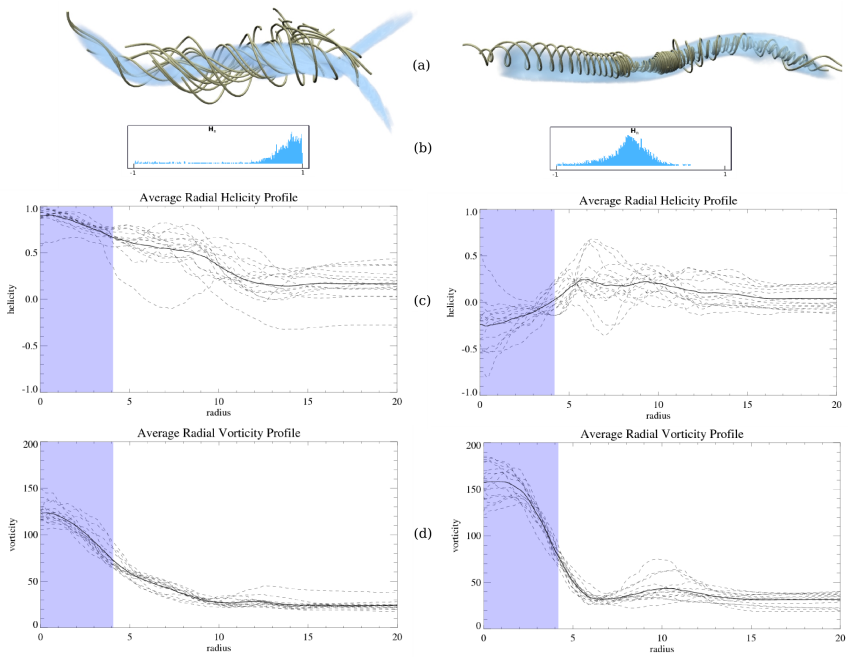

Fig. 5. Local dynamics of two strucutures with different helicity distributions showing: a) streamlines seeded within segmented region; b) the feature-local helicity histogram; c) an average radial helicity profile; d) an average radial vorticity profile. The shaded region of the radial profiles represents the inside of the visualized structure.

The primary advantage of coupling visual data investigation with a data analysis language is the ability to defer expensive calculations of derived quantities until they are needed and then perform them only over sub-domains of interest. The computational requirements for computing such variables in advance, across the entire domain, is often impratical, overwhelming the available analysis resources. Furthermore, some quantities, as was shown by our analysis of the Taylor-Green flow, can only be computed with reference to the location of a flow structure and are therefore not in principle a priori computable. The coupling between VAPOR and IDL facilitates the calculation of derived quantities as needed over sub-regions of the domain, realizing considerable savings in storage space and processing time.

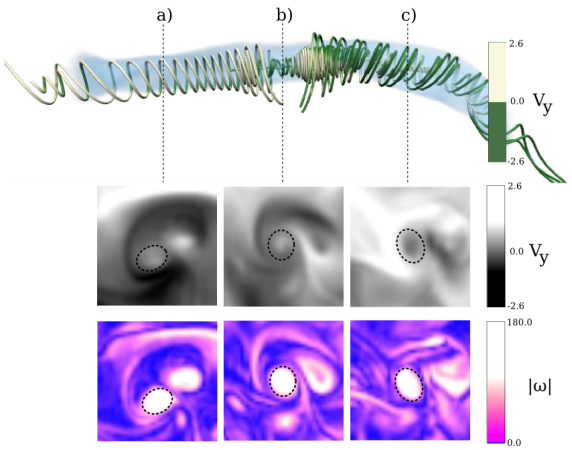

Fig. 6. Top: Streamlines colored by y-component velocity, which approximates the axial velocity. Bottom: vorticity magnitude and Y-component velocity cross-sections taken at positions a, b, & c (extents of the structure are bounded by the dotted line)

4 Conclusions

VAPOR's tight integration of visualization with traditional techniques like statistics, fourth-generation data languages, and effective information-management strategies meets the challenges that are inherent in visual exploration of complex turbulence data. This will only become more important as data volumes increase. The Taylor-Green flow simulation described in the previous section, which has 1024^3 degrees of freedom, can be readily computed on today's *teraflop* supercomputing platforms. The emergence of *petaflop*-capable machines will enable simulations at vastly greater scales, resulting in substantially larger data volumes. 4096^3 simulations have already been conducted on existing supercomputers [17] and the recent NSF Track 1 Petascale computing solicitation calls for a system capable of executing a homogeneous turbulence simulation with $12,288^3$ degrees of freedom [18]. The interactive analysis model in this paper, with its reliance on progressive data refinement, visual data browsing, and region/structure-of-interest isolation, is intrinsically highly scalable. We have described our experiences with this analysis model in the context of investigating numerically simulated turbulence. However, we believe that these techniques have applicability across a broad spectrum of data-intensive sciences.

5 Acknowledgements

We wish to thank the National Science Foundation and the National Center for Atmospheric Research for their computational support.

References

1. Barker, K.J., Davis, K., Hoisie, A., Kerbyson, D.J., Lang, M., Pakin, S., Sancho, J.C.: Entering the petaflop era: the architecture and performance of roadrunner. In: 2008 ACM/IEEE conference on Supercomputing, Austin, Texas, IEEE Press (2008) 1–11
2. Keim, D., Ward, M.: Visualization. In Berthold, M., Hand, D., eds.: Intelligent Data Analysis: An Introduction. 2 edn. Springer-Verlag (2000)
3. Yang, L.: 3D Grand Tour for multidimensional data and clusters. In: Proceedings of the Third International Symposium on Intelligent Data Analysis (IDA-99). (1999) 173–184
4. Rehm, F., Klawonn, F., Kruse, R.: Mds-polar: A new approach for dimension reduction to visualize high-dimensional data. In: Proceedings of the Sixth International Symposium on Intelligent Data Analysis (IDA-05). (2005) 316–327
5. Clyne, J.: The multiresolution toolkit: Progressive access for regular gridded data. (2003) 152–157
6. Clyne, J., Mininni, P.D., Norton, A., Rast, M.: Interactive desktop analysis of high resolution simulations: application to turbulent plume dynamics and current sheet formation. New Journal of Physics **9** (2007)
7. Lorensen, B.: On the death of visualization. In: NIH/NSF Fall 2004 Workshop Visualization Research Challenges. (2004)
8. Ahrens, J., Brislawn, K., Martin, K., Geveci, B., Law, C.C., Papka, M.: Large-scale data visualization using parallel data streaming. IEEE Computer Graphics and Applications **21** (2001) 34–41
9. Childs, H., Brugger, E., Bonnell, K., Meredith, J., Miller, M., Whitlock, B., Max, N.: A contract based system for large data visualization. In: Proceedings of IEEE Visualization. (2005) 191–198
10. Engel, K., Hadwiger, M., Kniss, J.M., Lefohn, A.E., Salama, C.R., Weiskopf, D.: Real-time volume graphics. A K Peters, Ltd, Los Angeles, CA (2006)
11. Weiskopf, D., Erlebacher, G.: Overview of flow visualization. In Hansen, C., Johnson, C., eds.: Visualization Handbook. Academic Press (2005)
12. Clyne, J., Rast, M.: A prototype discovery environment for analyzing and visualizing terascale turbulent fluid flow simulations. In Erbacher, R.F., Roberts, J.C., Grohn, M.T., Borner, K., eds.: Visualization and Data Analysis 2005. Volume 5669., San Jose, CA, USA, SPIE (March 2005) 284–294
13. Suzuki, K., Horibia, I., Sugie, N.: Linear-time connected-component labeling based on sequential local operations. Computer Vision and Image Understanding **89** (2003) 1–23
14. Mininni, P.D., Alexakis, A., Pouquet, A.: Nonlocal interactions in hydrodynamic turbulence at high reynolds numbers: the slow emergence of scaling laws. Physical review. E, Statistical, nonlinear, and soft matter physics **77** (2008)
15. Jiang, M., Machiraju, R., Thompson, D.: Detection and visualization of vortices. In Hansen, C., Johnson, C., eds.: Visualization Handbook. Academic Press (2005)
16. Wu, J.Z., Ma, H.Y., Zhou, M.D.: Vorticity and Vortex Dynamics. 1 edn. Springer (May 2006)
17. Kaneda, Y., Ishihara, T., Yokokawa, M., Itakura, K., Uno, A.: Energy dissipation rate and energy spectrum in high resolution direct numerical simulations of turbulence in a periodic box. Physics of Fluids **15** (2003) L21–L24
18. : Leadership-class system acquisition - creating a petascale computing environment for science and engineering NSF solicitation 06-573.