Editors: Cláudio T. Silva, csilva@nyu.edu | Daniel Weiskopf, weiskopf@visus.uni-stuttgart.de

Reducing the Analytical Bottleneck for Domain Scientists: Lessons from a Climate Data Visualization Case Study

Aritra Dasgupta | Pacific Northwest National Laboratory
Jorge Poco | University of Washington, Seattle
Enrico Bertini and Claudio T. Silva | New York University

Recent advances in high-performance computing have helped produce huge amounts of data at a very fast rate. But the relative speeds of data analysis and exploration, especially in scientific disciplines, have been disproportionately low. For example, using a modern supercomputer, it only takes a few days to produce terabytes of scientific simulation output data, but it then takes weeks or months of tedious scripting for domain scientists to perform analysis and exploration. Richard Hamming famously said, "The purpose of computing is insight, not numbers" (https://en.wikipedia.org/wiki/Richard_Hamming). But most current scientific data analysis tools have focused exclusively on the data scalability problem and not so much on the complexity and variety of the data space that affect the process of deriving insights from the data.

In domains such as biology and climate, many scientists still adopt manual, time-consuming data analysis processes or use tools that don't tightly integrate interactive and analytical capabilities. To let domain scientists analyze, explore, and synthesize insights from large simulation data, there's a need for new techniques and analytical abstractions that will significantly speed up the

 1521-9615/16/$33.00 © 2016 IEEE Copublished by the IEEE CS and the AIP

analysis through an iterative, human-in-the-loop process.

We term this discrepancy between the relative speeds of data generation and analysis as an analytical bottleneck for domain scientists (see Figure 1). In this article, we present evidence from cross-domain collaborations between visualization researchers and climate scientists about the successful use of interactive visualization to reduce this bottleneck. We also comment on the opportunities and open challenges that aren't just unique to the visualization of climate model data but generalizable across many disciplines.

Climate Science Background

Climate scientists generate mathematical models for simulating physical processes (see Figure 2).[1–4] There is inherent uncertainty and disagreement in the parameterization of these models and in understanding their effects on outputs. However, gauging consensus among model outputs is critical for achieving high accuracy about prediction of environmental events, climate change patterns, and so on. The Holy Grail in climate science is to understand why certain choices of parameterizations produce similar or different model outputs and how these parameterizations affect the quality of the outputs in terms of agreement with observation data. The work reported here is based on a collaboration among a group of visualization researchers (henceforth referred to as *we*) and climate scientists as part of the Multi-scale Synthesis and Terrestrial Model Inter-comparison Project (MsTMIP; http://nacp.ornl.gov) and the US National Science Foundation (NSF)–funded DataONE (www.dataone.org) initiative. All our collaborators have at least 10 years of experience in climate modeling. MsTMIP is a formal multiscale synthesis, with prescribed environmental and meteorological drivers shared among model teams and simulations standardized to facilitate comparison with other model results and observations through an integrated evaluation framework.

State of the Art

Climate scientists aim to pursue key scientific questions by running model simulations. By leveraging high-performance computing techniques, these simulation runs can produce terabytes of data very quickly. But using state-of-the-art techniques, data analysis usually takes a large amount of time as they don't allow for rapid iteration or provide analytical support for quickly finding patterns of interest. Many current tools are hypothesis-driven, meaning they allow scientists to pursue a specific hypothesis by performing certain computational or visualization tasks, but they lack the flexibility to provide different perspectives into the data—examples of such tools include UV-CDAT[5] and Paraview.[6] The large analysis time is thus caused not only by the scale but also by the complexity of the simulation data: these models generally produce tens of output variables varying over different scales of space and time, and the models' internal structure is defined by hundreds of parameter values. Scientists want to analyze where, when, and at what scales of space and time the model outputs are similar or different and then reconcile those similarities and differences in outputs with various parameterization choices.

The complex nature of climate modeling data thus necessitates cutting-edge analytical methods for interactive subsetting of the data and rapid exploration of alternative hypotheses on

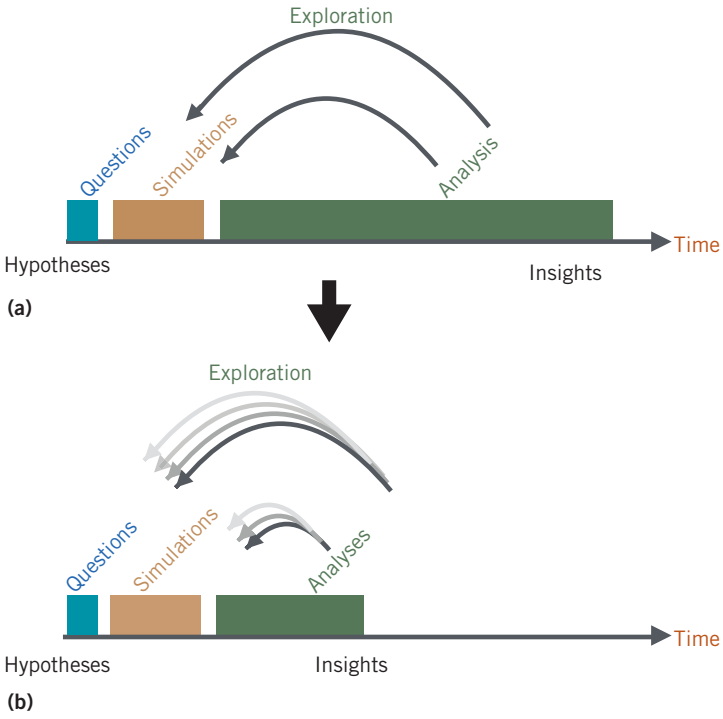

Figure 1. Reducing the analytical bottleneck in scientific data analysis through iterative exploration of large-scale data. In comparing (a) the state of the art and (b) a bottleneck reduction, we see the time taken for analysis and deriving insights being significantly reduced, leading to a richer exploration of alternative hypotheses on the fly, and consequently, a faster and greater return on investment of analysis time for domain scientists.

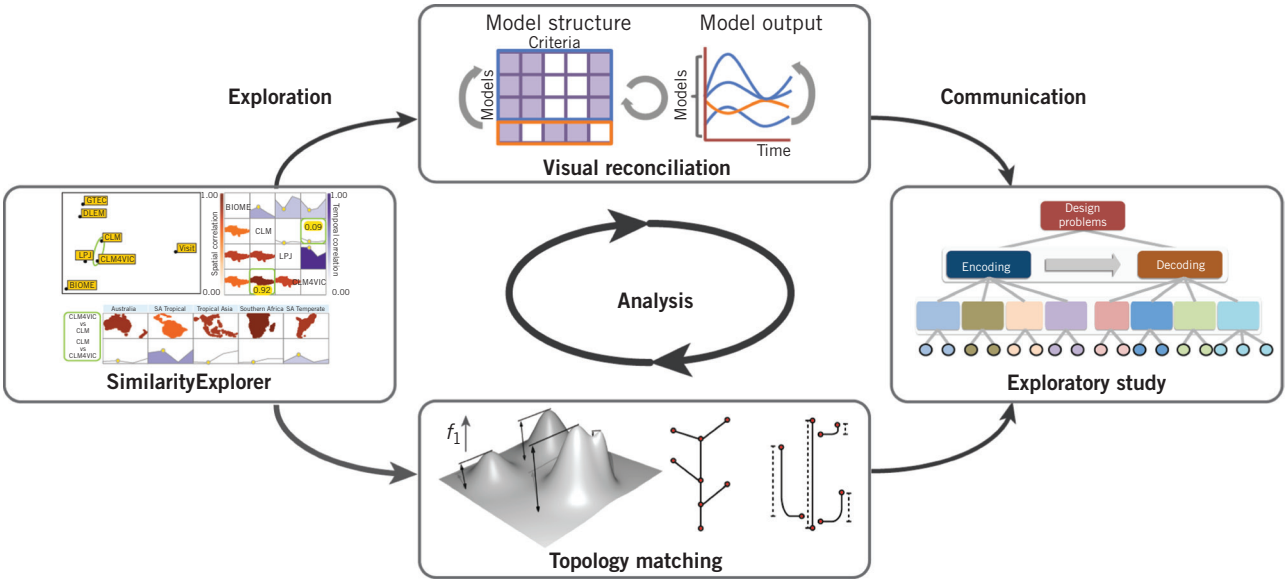

Figure 2. Our collaborative research on climate data visualization for exploration[1] and analysis[2,3] of model simulation data and visual communication[4] of scientific insights demonstrated how incorporation of interactive visualization in the scientific data analysis pipeline can help reduce the analytical bottleneck.

the fly, thereby reducing analysis time. Researchers have argued for the need to develop data-driven analytical methods[7] that can complement hypotheses-driven scientific analysis methods. It's in this context that we define the analytical bottleneck.

The Analytical Bottleneck

Current visualization tools are good at letting scientists pursue known questions by performing statistical analysis on interesting model outputs and parameters. However, even to reach a point where scientists find an interesting pattern, they have to navigate a large, complex search space. For example, aerosols, clouds, and atmospheric motions interact with each other through hundreds or thousands of different physical and chemical processes that span a wide range of spatial (from nanometers to the size of the Earth) and temporal (from fractions of a second to millennia) scales. To add to the complexity, the number of parameters in state-of-the-art climate models is large (on the order of hundreds), with varying effects on the outputs, which are often difficult to quantify.

The main analytical task is to sift through different sets and combinations of these data attributes or facets and detect patterns hidden in a subset of these facets. Traditional hypothesis-driven methods and tools are inadequate to handle the complexity of these multiple combinations, where scientific questions often depend on the exploratory analysis process. In scenarios where scientists pursue unknown unknowns,[8] these tools don't support rapid exploration of alternative hypotheses. The analytical bottleneck is caused by three main inadequacies in current analysis tools: the ability of scientists to easily convert their high-level analysis goals into the visualization interface through interactions, the lack of multiple perspectives into the data as scientists often need to look at different views before reaching their conclusions, and rapid, dynamic exploration of different hypotheses in which the system adapts to the interactions, proactively searches for interesting patterns, and helps scientists in narrowing down their visual search process.

Impact of Visualization

As Figure 2 shows, a climate scientist's workflow comprises three distinct stages: exploration of high-dimensional parameter and output spaces, analysis of causal relationships between inputs and outputs, and synthesis and communication of key insights about model performance. Here, we describe the impact of our research in these areas.

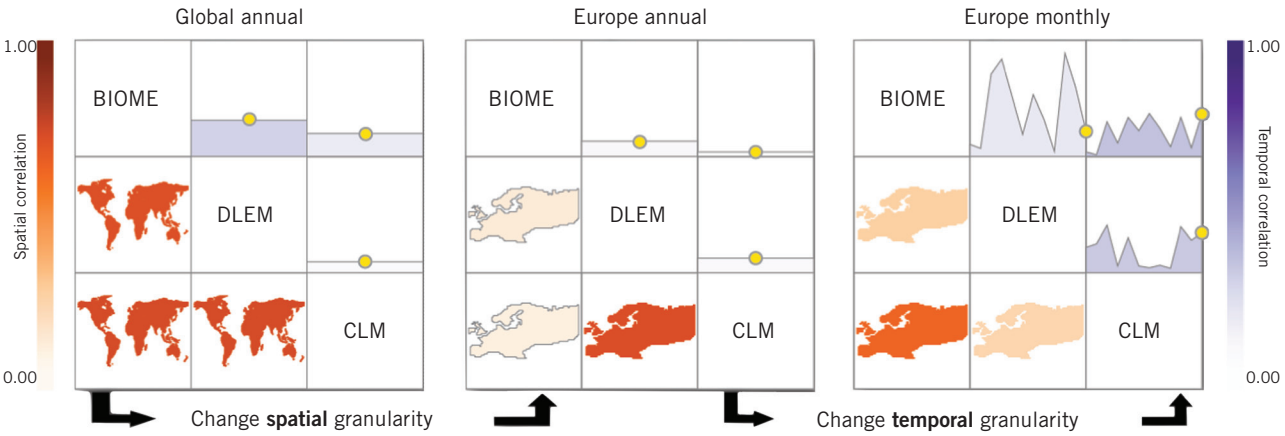

Figure 3. SimilarityExplorer enables multiscale exploration of spatiotemporal model similarity by letting scientists quickly and flexibly explore where and when multiple model outputs are similar and verify their hypotheses on the fly.

Exploration of Multifaceted Model Data

Data objects can often be described by multiple facets—for example, patient data can be described by demographics, disease symptoms, treatment and prognosis history, and so on. Similarly, climate models can be described by space, time, outputs, parameters, and so on.

Understanding similarity relationships across heterogeneous facets is challenging, so we developed a multifaceted visual exploration tool called SimilarityExplorer[1] to help scientists understand the similarity of model outputs at different scales of space and time (see Figure 3) and with different combinations of output variables. By explicitly encoding[9] spatial and temporal correlations among models, the tool lets scientists identify similar outputs at different scales of space and time. One of our collaborators commented that "this would allow them to develop hypotheses on performing additional experiments" and that "the free-style nature of the exploration lends well to shift from one variable to another and support root-cause analysis."

Analysis from Multiple Perspectives

We developed a visual reconciliation technique[2] to help scientists detect correlations across input parameters and temporal outputs. As Figure 4 shows, an iterative refinement strategy lets them dynamically create, modify, and observe the interaction among groupings, thereby making the potential explanations apparent. The strength of the visual reconciliation technique is in the tight integration of an underlying optimization component with a visual feedback mechanism that guides scientists toward potential groupings of interest. Regarding the effectiveness of the reconciliation technique, another collaborator observed that "one of the most valuable functions of the technique is to effectively remove from consideration the complications created from model structures that have little to no effect on outputs, and to effortlessly show and rank the differential effects on output created by seemingly related or unrelated model structures."

In addition, given that climate models can be seen as high-dimensional scalar functions, we developed a topology-based method to help climate scientists understand and explore the differences between models directly in the high-dimensional domain.[3] We introduced the concept of maximum topology matching to identify similarities and differences between a given pair of models. Furthermore, we designed a visualization interface that lets scientists explore models using their topological features to study the differences between pairs of models (see Figure 5).

Communication of Scientific Insights

Climate scientists commonly use various scripting languages to produce their own visualizations for either probing the data or using the visualization to publish and disseminate their results to a wider audience. However, our initial interactions with the scientists revealed that many existing visualization practices in their domain don't obey visualization best practices and perceptual design principles. This is especially true when these

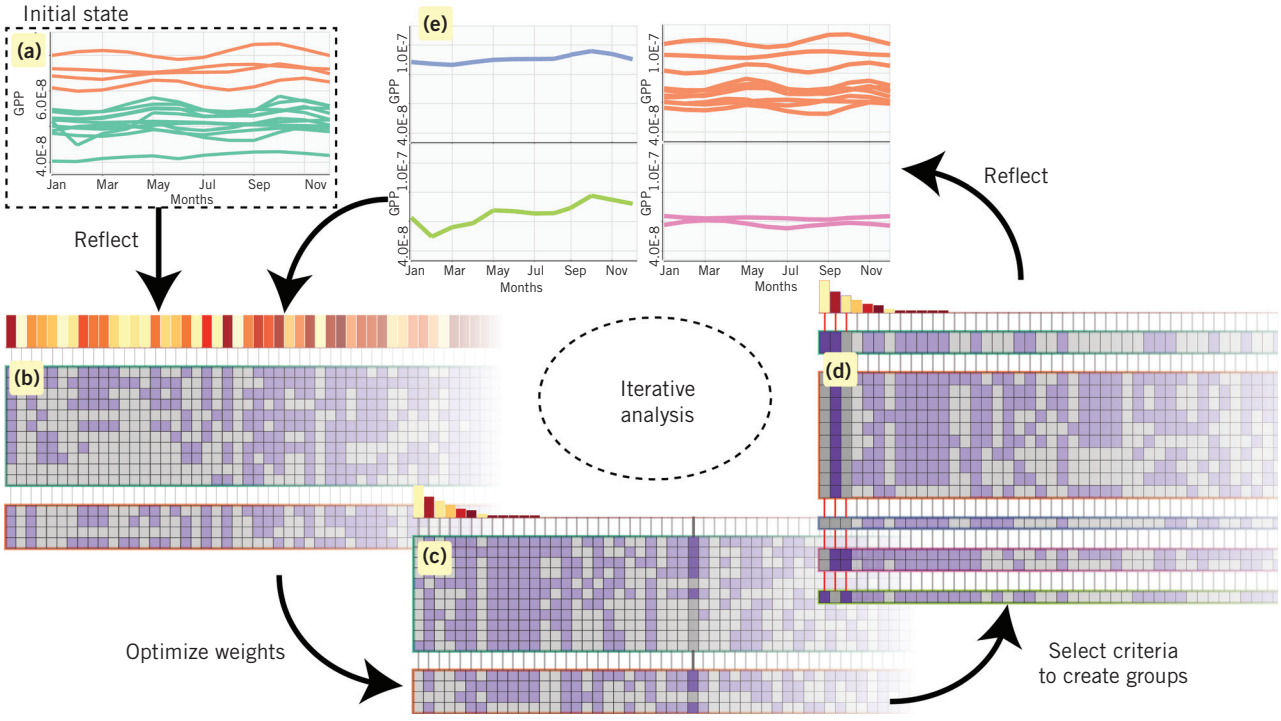

Figure 4. Visual reconciliation facilitates iterative refinement of heterogeneous climate model groups based on their output (time series) and structure (matrix). Starting from (a) a time-series of model outputs, (b) (c) (d) scientists can progressively manipulate the linked matrix for discovering the causal relationships between inclusion of different model structural parameters and their effect on the output.

visualizations are used to communicate scientific insights from exploration and analysis. To systematically investigate how the state of the art can be improved, we collected common examples of data visualizations (line charts, geographical maps, and scatter plots) designed by climate scientists, and in tight collaboration with them, we developed a taxonomy of design problems. Subsequently, we used the taxonomy to develop solutions for common design pitfalls and performed qualitative evaluation of our solutions by getting their feedback.

In the process, we also reflected on differences in opinion about design problems between visualization researchers and climate scientists, and the possible causes for such differences, before distilling a few design guidelines that can be used across different science disciplines where domain experts design visualizations as part of their daily routine.

Criteria for Reducing the Analytical Bottleneck

The tools and techniques we developed aim to address the analytical needs of climate scientists with the design principles of information visualization[10,11] and visual analytics.[12] We accomplished this through interview sessions, participatory design processes, and both qualitative and quantitative studies for evaluating our tools. Our collaborators consider our work as a significant contribution to the state of the art in climate science. The feedback and responses we received were largely positive and hold much promise for the future. From our collaboration, we distilled four criteria that visual interfaces should fulfill to reduce the analytical bottleneck. Here, we define these criteria and provide examples of how they can be practically implemented by providing examples from our tools.

Flexibility

A visualization tool for climate scientists should be flexible enough for interactive subsetting through different combinations of space and time scales and for providing multiple seeds as points for starting an analysis. Climate scientists often start with a specific hypothesis but require flexibility in slicing and dicing through the space-time data

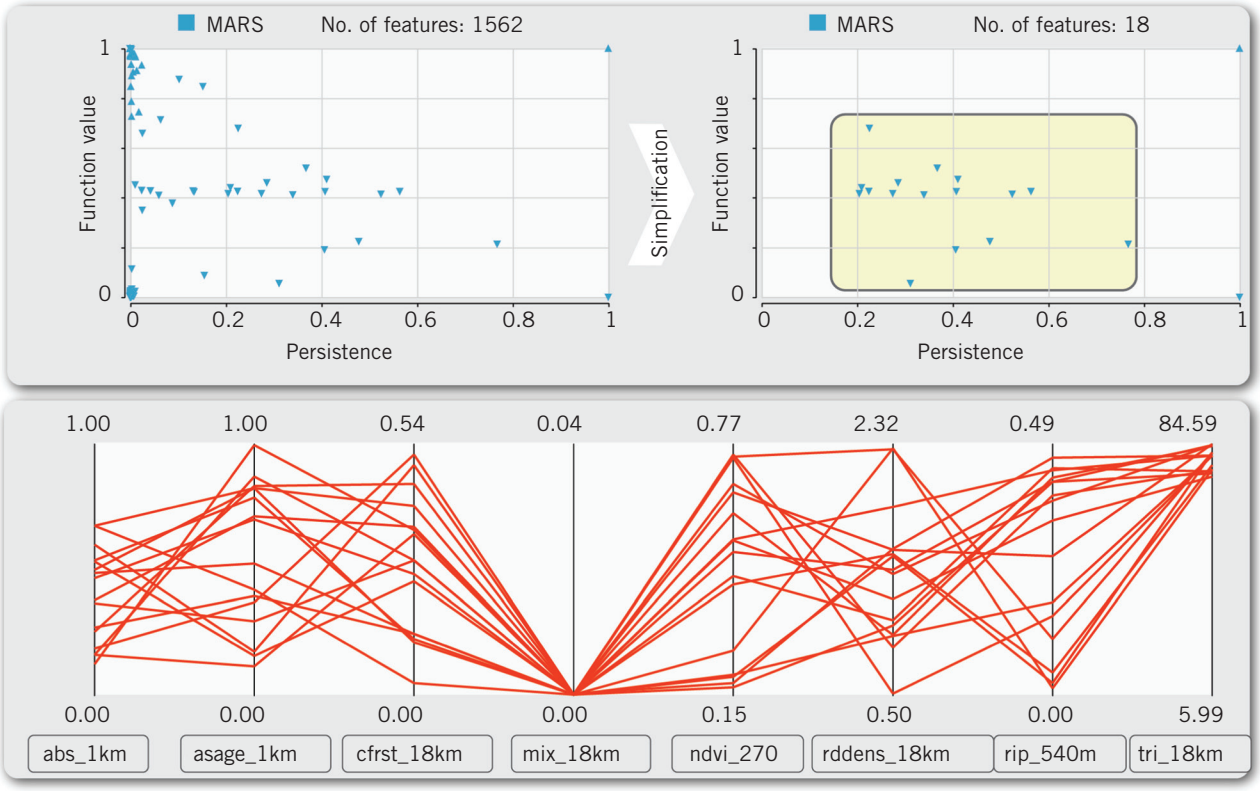

Figure 5. Exploring MARS model using a topology-based method. A scatter plot denotes the similar and dissimilar extrema of a given pair of functions, and parallel coordinates represent the location of extrema of interest in the high-dimensional predictor space.[3]

cubes that climate models represent. This is especially relevant when they want to compare global climate patterns with regional anomalies along different time scales, such as annual, monthly, or weekly patterns of interest. They also want to be flexible in the way they define relationships among different models.

Accordingly, SimilarityExplorer defines similarity among climate models based on both spatial and temporal correlations, enabling scientists to look at both pairwise and overall similarity across all models. Once scientists look at global similarity among models, they can drill down into the different space and time scales of interest dynamically through interaction. With the visual reconciliation technique[2] scientists could steer the analysis process from two seed points: similarities in temporal model outputs and characteristics of the parameter space. Such flexibility gave them the freedom to explore model facets from dynamic perspectives according to their changing hypotheses.

Efficiency

We define efficiency in terms of how fast insights can be generated from data, not in terms of how fast data can be processed. While the latter is certainly a precursor for high analytical efficiency, we assume that the analysis is being carried out in a high-data-throughput environment. Such efficiency is a result of scientists being able to quickly detect interesting patterns among a large number of possible combinations of data attributes. Even in very large datasets, meaningful patterns lie hidden only among a subset of combinations of records and attributes. For example, the climate model data we used has binary parameters. A model structure is a function of these parameters, so if there are c criteria, there can be $2c$ combinations of this function.

Climate scientists don't have an objective way of choosing one set of criteria over another, which can influence the output. The visual reconciliation technique enabled them to efficiently search through these exponentially large possible combi-

We observed that the effectiveness of the techniques we developed mainly depended on two factors: the degree of transparency to which the automated methods are integrated within visual analytics methods, and the optimality of the visualization design.

nations and judge their effects on outputs. While using SimilarityExplorer, the scientists could get a quick overview of the spatiotemporal behavior of the outputs, and by applying the reconciliation technique, they could diagnose the causes for the similarity and discrepancies among multimodel output behavior.

Serendipity

Many important scientific inventions are based on serendipitous discoveries. In this context, we posit that scientific data analysis tools should be designed with an explicit goal of facilitating such data-driven discovery by allowing scientists to pursue alternative hypotheses and steer the analysis process for discovering the unexpected[12] while the system proactively suggests interesting patterns. In the context of climate model data analysis, our goal was to design systems with an integration of analytical components and interactive visualization, such that scientists can verify their known hypothesis and also explore alternative hypotheses on the fly by exploring unknown unknowns.

Both the visual reconciliation and topological data analysis techniques allowed scientists to detect patterns that weren't possible using manual data analysis or state-of-the-art analysis tools. This utility was reflected in multiple positive comments such as, "Looking at models this way is interesting for me. With our current tools, we wouldn't have otherwise known that it was at high values of mean-summer that the models differed, and that might be of interest in an in-depth study." We're currently working with our collaborators in an ecological data analysis context that further investigates the implications of the results obtained.

Effectiveness

In our experience of collaborating with the climate scientists, we observed that the effectiveness of the techniques we developed mainly depended on two factors: the degree of transparency to which the automated methods are integrated within visual analytics methods, and the optimality of the visualization design.

In the first case, while developing the SimilarityExplorer tool and the reconciliation and topological data visualization techniques, scientists were skeptical of the analysis outputs when data transformation in an analytical method either wasn't known to them or they couldn't control the process. Therefore, in SimilarityExplorer, we used custom-defined distance functions such as spatial and temporal correlations and exposed the effects of the individual output variables, rather than combining them to produce a multidimensional projection.

In the second case, through our interviews and qualitative and quantitative studies, we found that scientists tend to use the same visualizations for communicating their results to a wider audience that they would normally use for their own analysis process. As we showed in a previous work,[4] common design pitfalls such as improper choice of visual variables, color maps, or comparison techniques affect the amount of insight that can be gained from visualizations. Currently, we're extending our work to conduct a quantitative study on the use of color maps for climate data visualization. The goal here is to study whether perceptually motivated color maps are more effective in common climate data analysis tasks than the more traditionally used color maps in the climate science domain.

Open Issues and Challenges

Several open issues and challenges weren't fully explored in the course of our collaboration. These issues not only apply to climate data visualization but can be generalized across other scientific disciplines.

Task Abstraction

One of the recurring challenges in our collaboration was to develop a task abstraction model that can be used to develop tools and techniques.

Although we used results from interviews and mapped them to existing task taxonomies, we can pursue other promising directions, such as formally establishing the role of a visualization liaison,[13] someone who can significantly accelerate the process of narrowing down a set of tasks via visual interfaces. Such task abstraction is also useful in cases where metrics must be developed to optimize the visual search process,[14] especially for complex scenarios such as multivariate, spatiotemporal, scientific data analysis.

Trust-Augmented High-Dimensional Data Analysis

Many automated pattern detection methods become ineffective when applied to data with high dimensionality, becoming black boxes that domain experts don't completely trust.[8] To facilitate a more transparent analysis process, we'll explore open areas of research in high-dimensional data visualization[15] by developing and applying promising techniques such as subspace search[16] and topology-based analysis of climate model data. Subspace clustering methods can be leveraged to suggest interesting variable combinations and let scientists visually search and detect relevant subspaces. Spatiotemporal behavior of these subspaces can be exposed through topological methods.

Uncertainty Handling

Visual representation of uncertainty along multiple dimensions is another open area of visualization research. We'll leverage and extend existing research on uncertainty visualization[17,18] by devising novel visual representations of parametric uncertainty integrated with analytical methods for iterative sensitivity analysis. A lot of progress has been made in uncertainty visualization, but this still remains an open issue across various domains.[19]

Dissemination of Scientific Knowledge

Despite their promise, visualization techniques haven't been used extensively to communicate scientific insights.[20] By leveraging the theoretical framework developed in our earlier work, and developing interactive visual narratives[21] about model fidelity, scientists will be able to complete the loop in their analysis process, exploring uncertainty and sensitivity of model parameters and outputs, evaluating model fidelity, and then tracing back the causes of infidelity to uncertainty levels and parameterization. This will lead to broader dissemination of scientific results through the use and adoption of visual narratives, leading to engagement across a wide range of audiences and potentially influencing climate change policies through such dissemination.

Although the definition of big data can vary across different stakeholders, we undeniably live in an era of complex heterogeneous data, and generating the value out of such data necessitates novel data science methods. In the course of our collaboration, we found that traditional and evolving data science domains are still disconnected: domain scientists are often skeptical of using automated data mining or machine learning methods, whereas data scientists are often unable to understand the analytical needs. We believe that interactive visualization and visual analytics will play a key role in bridging these diverse areas by bringing more transparency and flexibility into analysis processes. That said, more collaborative efforts are necessary to completely bridge the gap, and we hope our work inspires future collaborations by providing a starting point. ■

Acknowledgments

We extend our gratitude to members of the Scientific Exploration, Visualization, and Analysis working group (EVA) as part of the DataONE project for their feedback and support on the collaborative projects reported here. Thanks to Philip J. Rasch, Hui Wan, and Dustin L. Arendt (scientists at the Pacific Northwest National Laboratory) for the many discussions about the adoption of interactive visualization in the climate scientists' analysis workflow.

References

1. J. Poco et al., "SimilarityExplorer: A Visual Inter-Comparison Tool for Multifaceted Climate Data," *Computer Graphics Forum*, vol. 33, no. 3, 2014, pp. 341–350.
2. J. Poco et al., "Visual Reconciliation of Alternative Similarity Spaces in Climate Modeling," *IEEE Trans. Visualization and Computer Graphics*, vol. 20, no. 12, 2014, pp. 1923–1932.
3. J. Poco et al., "Using Maximum Topology Matching to Explore Differences in Species Distribution Models," to appear in *Proc. IEEE SciVis*, 2015.
4. A. Dasgupta et al., "Bridging Theory with Practice: An Exploratory Study of Visualization Use and

Design for Climate Model Comparison," *IEEE Trans. Visualization and Computer Graphics*, vol. 21, no. 9, 2015, pp. 996–1014.
5. E. Santos et al., "UV-CDAT: Analyzing Climate Datasets from a User's Perspective," *Computing in Science & Eng.*, vol. 15, no. 1, 2013, pp. 94–103.
6. J. Ahrens, B. Geveci, and C. Law, "Paraview: An End-User Tool for Large-Data Visualization," *The Visualization Handbook*, Elsevier, 2005, p. 717.
7. J.H. Faghmous and V. Kumar, "A Big Data Guide to Understanding Climate Change: The Case for Theory-Guided Data Science," *Big Data*, vol. 2, no. 3, 2014, pp. 155–163.
8. D. Sacha et al., "The Role of Uncertainty, Awareness, and Trust in Visual Analytics," to appear in *IEEE Trans. Visualization and Computer Graphics*, 2015.
9. M. Gleicher et al., "Visual Comparison for Information Visualization," *Information Visualization*, vol. 10, no. 4, 2011, pp. 289–309.
10. J. Bertin, *Semiology of Graphics: Diagrams, Networks, Maps*, University of Wisconsin Press, 1983.
11. S.K. Card and J. Mackinlay, "The Structure of the Information Visualization Design Space," *Proc. IEEE Symp. Information Visualization*, 1997, pp. 92–99.
12. J.J. Thomas and K.A. Cook, "A Visual Analytics Agenda," *IEEE Computer Graphics and Applications*, vol. 26, no. 1, 2006, pp. 10–13.
13. S. Simon et al., "Bridging the Gap of Domain and Visualization Experts with a Liaison," *Proc. Eurographics Conf. Visualization* (EuroVis), E. Bertini, J. Kennedy, and E. Puppo, eds., 2015, pp. 127–131.
14. A. Dasgupta, R. Kosara, and L. Gosink, "VIMTEX: A Visualization Interface for Multivariate, Time-Varying, Geological Data Exploration," *Computer Graphics Forum*, vol. 34, no. 3, 2015, pp. 341–350.
15. S. Liu et al., "Visualizing High-Dimensional Data: Advances in the Past Decade," *Proc. Eurographics Conf. Visualization* (EuroVis), R. Borgo, F. Ganovelli, and I. Viola, eds., 2015; https://diglib.eg.org/handle/10.2312/eurovisstar.20151115.127-147.
16. L. Parsons, E. Haque, and H. Liu, "Subspace Clustering for High Dimensional Data: A Review," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, 2004, pp. 90–105.
17. A.T. Pang, C.M. Wittenbrink, and S.K. Lodha, "Approaches to Uncertainty Visualization," *The Visual Computer*, vol. 13, no. 8, 1997, pp. 370–390.
18. A. Dasgupta, M. Chen, and R. Kosara, "Conceptualizing Visual Uncertainty in Parallel Coordinates," *Computer Graphics Forum*, vol. 31, no. 3, 2012, pp. 1015–1024.
19. S. Liu et al., "A Survey on Information Visualization: Recent Advances and Challenges," *The Visual Computer*, vol. 30, no. 12, 2014, pp. 1373–1393.
20. G.J. McInerny et al., "Information Visualisation for Science and Policy: Engaging Users and Avoiding Bias," *Trends in Ecology & Evolution*, vol. 29, no. 3, 2014, pp. 148–157.
21. M. Krzywinski and A. Cairo, "Points of View: Storytelling," *Nature Methods*, vol. 10, no. 8, 2013, pp. 687–687.

Aritra Dasgupta is a research scientist at Pacific Northwest National Laboratory. His research interests include high-dimensional data visualization, visual communication of scientific insights, and mixed-initiative visual analytics systems. Dasgupta received a PhD in computer science from UNC-Charlotte. Contact him at aritra.dasgupta@pnnl.gov.

Jorge Poco is a postdoctoral researcher at University of Washington, Seattle. His research interests include data visualization, visual analytics, and data science. Poco received a PhD in computer science from New York University. Contact him at jpocom@uw.edu.

Enrico Bertini is an assistant professor at the School of Engineering, New York University. His research interests include data visualization for high-dimensional data analysis and machine learning, evaluation of communication-oriented visualization, and visualization for text analytics. Bertini received a PhD in computer engineering from Sapienza University of Rome. Contact him at enrico.bertini@nyu.edu.

Claudio T. Silva is a professor at the School of Engineering, New York University. His research interests include big data and urban systems, visualization and data analysis, and sports analytics and visualization. Silva received a PhD in computer science from the State University of New York at Stony Brook. Contact him at csilva@nyu.edu.