VISUAL INTERCOMPARISON OF MULTIFACETED CLIMATE DATA

DISSERTATION

Submitted in Partial Fulfillment of

the Requirements for

the Degree of

DOCTOR OF PHILOSOPHY (Computer Science)

at the

NEW YORK UNIVERSITY
POLYTECHNIC SCHOOL OF ENGINEERING

by

Jorge L. Poco Medina

September 2015

VISUAL INTERCOMPARISON OF MULTIFACETED CLIMATE DATA

DISSERTATION

Submitted in Partial Fulfillment of

the Requirements for

the Degree of

DOCTOR OF PHILOSOPHY (Computer Science)

at the

NEW YORK UNIVERSITY
POLYTECHNIC SCHOOL OF ENGINEERING

by

Jorge L. Poco Medina

September 2015

Approved:

____

Department Head Signature

____

Date

Copy No. ____________________

University ID#: N14822882

Approved by the Guidance Committee:

Major: Computer Science

Cláudio T. Silva
Professor of
Computer Science and Engineering

Juliana Freire
Professor of
Computer Science and Engineering

Enrico Bertini
Assistant Professor of
Computer Science and Engineering

Jean-Daniel Fekete
Senior Research Scientist

UMI Dissertation Publishing
ProQuest CSA
789 E. Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

Vita

Jorge Poco is from Arequipa, Peru. He received the BE in System Engineering from the National University of San Agustin, Peru, in 2008, and the MS in Computer Science from the Institute of Mathematics and Computer Science at the University of So Paulo, Brazil in 2010. Since September 2010, he has started his PhD under the supervision of professor Claudio Silva, first at the University of Utah, latter at the New York University Polytechnic School of Engineering. As part of his professional life he worked in zAgile Inc as a software engineer on 2008. He did internships at Google Inc. (2008 and 2010), Kitware Inc. (2011), Oak Ridge National Laboratory (2012) and Xerox Research (2013).

His research has focused on data visualization. He has participated in projects on information visualization, scientific visualization, and visual analytics. He has also been involved in interdisciplinary collaborations that focused on the development of novel visualization methods to enable both climate and urban data analysis.

Acknowledgements

First, I would like to thank my family, specially my parents. All the support they have provided me over the years was the greatest gift anyone has ever given me.

I would like to thank my advisor Cláudio Silva for giving his guidance throughout the entire PhD program. The last five years have changed my life and I consider myself to be very fortunate to get an opportunity to work with him during this time.

I would like to thank all of my committee members: Cláudio Silva, Juliana Freire, Enrico Bertini, and Jean-Daniel Fekete, for their valuable comments and insightful discussions to make this dissertation possible.

I would also like to acknowledge my co-authors. In particular, Aritra Dasgupta, Harish Doraiswamy, Nivan Ferreira, Yaxing Wei, Robert Cook, and William Hargrove. A special thanks to Aritra, with whom I had the opportunity to work with in most of the research presented in this dissertation.

A special thank to that person who has been next to me during the last events in my life. Her support and lovely words were very important.

Jorge Poco
September 2015

To my parents, with affection.

ABSTRACT

VISUAL INTERCOMPARISON OF MULTIFACETED CLIMATE DATA by

Jorge L. Poco Medina

Advisor: Prof. Cláudio T. Silva, Ph.D.

Submitted in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy (Computer Science)

September 2015

Gauging consensus among predictions and outputs of multiple simulation models is a critical problem for understanding global climate change patterns. This requires similarity analysis of climate models which typically involve multiple data facets like space, time, input parameters, output variables, etc. Such model inter-comparison enables scientists to explore and develop different hypotheses about ecosystem processes and climate change indicators. While it is widely accepted that interactive visualization can enable scientists to better explore model similarity from different perspectives and different granularities of space and time, currently there is a lack of such visualization tools.

To fill this gap, the main contributions of this dissertation are grouped in three stages: *Design Space Analysis*, *Visual Exploration*, and *Visual Analytics Approaches*. In the first stage for *Design Space Analysis*, we understood the state-of-the-art of static visualizations that climate scientists use. Based on this exploratory study, we derived a design problem taxonomy of static plots. After analyzing the results of this study, as a follow-up, we set up another study on color map usage by climate scientists.

By reflecting on the inadequacies of the static visualizations, and because analysis of similarity and dissimilarity is a complex problem given the multiple *facets* involved in such comparisons. We designed a *Visual Exploration* tool. SimilarityExplorer is an exploratory visualization tool which facilitates visual intercomparison of climate model data and its multiple facets, like, space, time, similarity, output variables, etc.. Making it easier for climate scientists to explore model relationships from multiple perspectives.

Even with exploration tools, it is still difficult to analyze the whole dataset or explore the complete parameter space. That is why, in the third stage *Visual Analytics Approaches*, we analyzed how multiple descriptors of these models, namely, their structural characteristics and their outputs can be reconciled using a novel visual analytics paradigm 'visual reconciliation'. Then, we proposed a topology-based framework to help study the differences in various models directly in the high dimensional data domain.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Climate scientists have made substantial progress in understanding the earth's climate system, particularly at global and continental scales. Climate research is now focused on understanding climate change over wider ranges of time and finer-space scale, which generates ultra-scale datasets. At such scales, a single snapshot of data will result in a terabyte or more of data, and modest time scales will result in petabytes of data. An insightful analysis in climate science depends on using software tools to discover, access, manipulate, and visualize the datasets of interest. These data exploration tasks can be complex and time-consuming, and they frequently involve many resources from both the modeling and observational climate communities. However, currently there is a lack of flexible visual analytics techniques to support such complex exploration tasks, and this thesis aims to fill that gap.

In general, climate simulations refer to one or more output variables (*e.g.*, temperature, precipitation, gross primary productivity). These simulations are run using multiple models, initial conditions, or parameterizations in order to gain confidence in the results and bound understanding. Consensus among model results is an important metric used for judging model performance. Analysis of model output similarity and dissimilarity is a complex problem because of the multiple *facets* involved in such comparisons: space, time, output variables, and model similarity. Thus, novel visualization techniques that integrate space, time, and similarity, are needed to let climate scientists efficiently explore models relationships from multiple perspectives. At the same time, the visualization techniques need to be augmented with automated analytical models for guiding the domain experts in their exploration, since manual exploration of the large parameter spaces is cumbersome.

The ever-growing data deluge has made visualization an important medium for intuitively portraying and communicating complex information, cutting across various disciplines such us climate sciences. However, creating visualizations demand significant time and effort, which often creates a bottleneck for domain experts [1]; and creating effective visualizations requires knowledge about visualization design principles and best practices. That is why, a systematic analysis of how climate scientists use and design visualizations is required for reflecting upon the causes and effects of design problems. It is important to follow-up this work with multiple user studies to understand the mismatch between visualization principles and the state-of-the-art in the climate science domain.

In this dissertation we tackle the problem of intercomparison of multifaceted climate data from three fronts: i) *design space analysis*, ii) *visual exploration tools*, and iii) *visual analytics approaches*.

The detailed discussion about these contributions is preceded by a background on climate modeling and model intercomparison goals which are relevant for this dissertation.

1.2 Climate Models

Climate scientists and ecologists (henceforth, we use the term "climate scientists" or "ecologists" interchangeably) build computer-based models to simulate, understand and predict climate systems. These models are based on mathematical representations that can incorporate the physics, chemistry, and other processes of the atmosphere, oceans and land. In this dissertation, we focus on two types of climate models:

Terrestrial Biosphere Models (TBM). TBMs simulate terrestrial ecosystem processes and the terrestrial-atmosphere carbon exchange in relation to prescribed boundary conditions: vegetation cover, soil properties, climate, etc. They have become an integral tool for extrapolating local observations and understanding to much larger terrestrial regions, as well as for testing hypotheses about how ecosystems will respond to changes in climate and nutrient availability [2]. TBMs can be used to attribute carbon sources (*e.g.*, fires, farmlands) and sinks (*e.g.*, forests, oceans) to explicit ecosystem processes.

Species Distribution Models (SDM). SDMs combine observations of species occurrence or abundance with environmental layers. They are used to gain ecological insights and to predict distributions across various landscapes including terrestrial, freshwater, and marine realms [3]. They help ecologists answer questions about the relationship between the environmental variables.

Model Intercomparison. A key approach for climate modeling is to use multiple models as a way to gain confidence in the results and bound understanding. Therefore, intercomparison of a suite of climate models over space, time, and different land cover types is an important research area. Thus, researchers want to know which models are similar, and why, when, and where they are similar. But the volume and complexity of model outputs present many challenges for analysis and visualization. Furthermore, to gain additional confidence in model output, researchers compare observations with model simulations in a benchmarking activity.

1.3 Thesis Statement

Effective understanding of similarities and differences among multiple climate models requires the combination of novel visual exploration techniques with automated analytical methods for enabling the climate scientists to identify salient patterns, and generate and validate hypotheses about climate phenomena.

1.4 Contributions

This dissertation proposes the use of novel visual analytics techniques for the purposes of exploration and analysis of climate data. The related contributions not only advance the scientific understanding of relationships among climate models, but also address important research challenges in the visualization community. These include multi-scale geospatial data exploration, correlating the effect of high-dimensional parameter spaces with model outputs, and finally, bridging the gap between the domain experts' analysis goals and effective visualization techniques through participatory design processes.

Based on the three fronts we mentioned before, our contributions can be summarized as follows:

In *Design Space Analysis.*

- An Exploratory Study of Visualization Use and Design for Climate Model Comparison [4].

1. We propose a classification scheme that categorizes the design problems in the form of a descriptive taxonomy. The taxonomy is a first attempt for systematically categorizing the types, causes, and consequences of design problems in visualizations created by domain experts;
2. We demonstrate the use of the taxonomy for: i) identifying problem consequences and their trade-offs, ii) a detailed analysis of causes of matches and mismatches about design problems between visualization experts and climate scientists, and iii) feedback on redesigned solutions for a representative sample of problem instances;
3. We provide a summary and analysis of the findings for enabling scientists in designing improved visualizations, and for reflecting on the gaps and opportunities for visualization research.

- Perceptual Evaluation of Color Scales for Climate Model Comparison [5].
  1. We characterize geospatial data comparison tasks performed by climate scientists. These are (i) judging overall magnitude, (ii) evaluating differences in spatial variation, and (iii) identifying regions of maximal difference;
  2. We measure the performance of climate scientists in each of these tasks using different color scales;
  3. We compare the scientists' quantitative performance against their perceived performances and preferences;

In *Visual Exploration Tools.*

- SimilarityExplorer: A Visual Intercomparison Tool for Multifaceted Climate Data [6].
  1. We propose a domain characterization for the TBM community by systematically defining the domain-specific intents for analyzing model similarity and characterizing the different facets of the data;
  2. We define a classification scheme for combining visualization tasks and multiple facets of climate model data in one integrated framework, which can be leveraged for translating the tasks into the visualization design;
  3. We present *SimilarityExplorer*, an exploratory visualization tool that facilitates similarity comparison tasks across both space and time through a set of coordinated multiple views;
  4. We present two case studies from climate scientists, who used our tool for a month for gaining scientific insights into model similarity.

in *Visual Analytics Approaches.*

- Visual Reconciliation of Alternative Similarity Spaces in Climate Modeling [7].
  1. We introduce a novel visual analytics paradigm: *visual reconciliation* as the problem of reconciling multiple alternative similarity spaces through visualization and interaction;
  2. We apply visual reconciliation to help climate scientists understand the dependency between alternative similarity spaces for climate models;
  3. We facilitate iterative refinement of groups with the help of a feedback loop and optimization techniques to guide the exploration;
  4. We present case studies that demonstrate the usefulness of our technique in the area of climate science.
- Using Maximum Topology Matching to Explore Differences In Climate Models [8].
  1. We introduce the concept of maximum topology matching that computes a locality-aware correspondence between similar extrema of two scalar functions.
  2. We design a visualization interface that allows ecologists to explore Species Distribution Models using their topological features and to study the differences between pairs of models found using maximum topological matching.
  3. We demonstrate the utility of the proposed framework through several use cases using different data sets and report the feedback obtained from ecologists.

1.5 Outline

In order to understand the common problems in climate data visualizations, in Chapter 2 we describe an exploratory study, developed closely with our collaborators. Based on this study, in Chapter 3 we explain the results of a user study to understand the mismatch between the visualization principles and the ubiquitous uses of rainbow colormap in the climate community. Next, in Chapter 4 we depict the SimilarityExplorer, a visual intercomparison tool for multifaceted climate data. Then, in Chapter 5 we introduce the visual reconciliation technique. In Chapter 6 we explain the topology-based framework to explore differences in various models directly in the high dimensional space. Finally in Chapter 7 we conclude the dissertation along with future work.

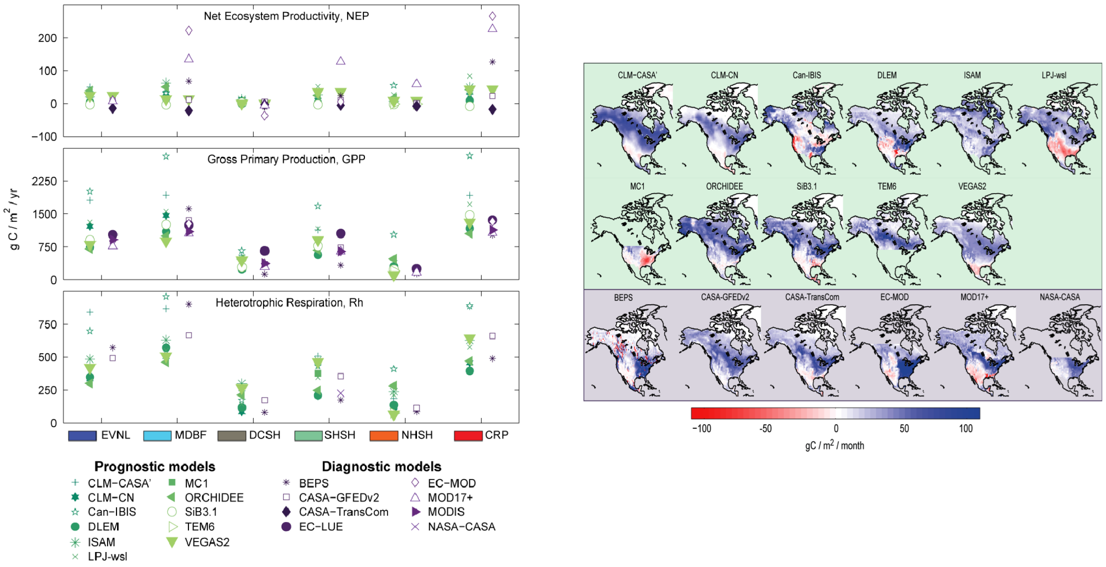
Chapter 2

An Exploratory Study of Visualization Use and Design for Climate Model Comparison

Creating visualizations demands significant time and effort, which often creates a bottleneck for domain experts [1]; and creating *effective* visualizations requires knowledge about visualization design principles and best practices. However, there has been little work on systematically judging the quality of visualizations used and created by non-experts in visualization. While authors like Tufte and Few [9, 10] have critiqued visualization examples and offered guidelines for better design, very few academic attempts exist for classifying types of design problems and judging their consequences, especially when domain experts design visualizations.

To fill this gap, in this chapter we describe a systematic analysis of how climate scientists use and design visualizations for reflecting upon the causes and effects of design problems. The data that we analyzed comprises of a series of semi-structured interviews with climate scientists, about visualizations collected from research papers and presentations.

The benefits of such an exploratory study are two-fold. First, it allows domain scientists to better critique their visualization designs and incorporate that knowledge into building more effective visual representations. Second, reflecting on the analysis of visualization design problems is an opportunity for the visualization community to investigate how the state-of-the-art in visualization meets the analysts' needs, and introspect how design principles can be better applied to suit the evolving challenges in data presentation and communication. In this work we judge how well domain experts and visualization researchers agree on design problems, based on which we



(a) **Design problems in a stacked scatter plot** stemming from over plotting and use of many different symbols.

(b) **Design problems in the multiple maps** stemming from poor encoding of relative similarity.

Figure 2.1: **Illustrating two common visualization use case scenarios and their associated visualization design problems**, for comparing terrestrial biospheric models (figures adapted from [2]). In **(a)** stacked scatter plots with multiple visual symbols lead to an ineffective visual search for models and inefficient comparison of spread among their output variables. In **(b)** outliers indicated by red regions are clearly visible but similarity analysis among 17 different maps is difficult without any encoding that reflects relative similarity among the models.

redesigned some of their existing visualizations and judged the effectiveness of the solutions from their feedback.

In our study, we focus on comparison of terrestrial biospheric models. Typical visualization usage and design by climate scientists for such comparisons is shown in Figure 2.1. Figure 2.1(a) shows the use of scatter plot for comparing output variables for multiple models. Figure 2.1(b) shows the use of multiple maps for analyzing similarity of models over different spatial regions. The challenges for concise visual representation in these cases is non-trivial because of the underlying diversity and complexity of the data. The aim of this exploratory study was to find, for these complex analysis tasks, what are some recurring design problems. While we also found some examples of optimal visualization designs, our goal in this chapter was not to comment on the general state-of-the-art in visualization practice in climate science, but to focus on the problematic visualization designs and devise a model for describing those problems.

Our high-level analysis questions for understanding visualization design problems were: do the chart types address the goals of visual representation? Are there design flaws specific to those chart types or are there generalizable problems cross-cutting chart types? Does the literature on visualization design offer solutions to those problems? We collected a representative sample of 15 research papers from our collaborators that used visualizations for comparing terrestrial biospheric models. The over-arching goal was to create a taxonomy that systematically answers the aforementioned analysis questions. In summary, the contributions of this chapter are three-fold:

1. A systematic classification of visualization design problems in the climate science domain resulting in a descriptive taxonomy of types, causes, and implications of such problems.

2. Application of the taxonomy for: i) identifying problem consequences and their trade-offs, ii) a detailed analysis of causes of matches and mismatches about design problems between visualization experts and climate scientists, and iii) feedback on redesigned solutions for a representative sample of problem instances.

3. Summary and analysis of the findings for enabling scientists in designing improved visualizations, and for reflecting on the gaps and opportunities for visualization research.

2.1 Related Work

We discuss the related work with respect to the existing studies on visualization usage and design, and the relevant theoretical models that have been proposed for characterizing the visualization process and design.

2.1.1 Studies on Usage and Design of Visualization

In recent times, there has been some progress towards studying how people outside the visualization community use, design, and reason about visualizations. This body of research is critical in diversifying the field of visualization by gaining insight into the potential roadblocks that people from different communities face, while designing perceptually effective visualizations and subsequently using the interpretations for their benefit.

To this end, Grammel et al. attempted to understand how people who are unfamiliar with visual data analysis, *i.e.* InfoVis novices, construct visualizations and the potential roadblocks in doing so [11]. They found that constructing effective visual mappings was the most significant roadblock, which was consistent with the findings of Heer et al. [12]. Somewhat related to this, researchers [13] studied the problems low-literacy users face while retrieving information online, and how interactive visualization can help in the process. Researchers have also explored the problem [14] from the point-of-view of existing visual analytics tools: they found that while conducting investigative analysis, several roadblocks exist in understanding, choosing, using, and reading views properly.

Some of the studies also focus on collaborative environments. Walny et al. [15] studied how pen and touch interactions on interactive whiteboards facilitate reasoning and understanding of visualizations. Isenberg et al. [16] discussed the role of the tabletop for visual analytics tasks and derive design implications for future co-located collaborative tabletop problem solving systems. Most of these studies focus on the usage patterns of visualizations for novice users. In our work, the focus is on domain experts who have compiled the data to specifically address their research questions but who do not have detailed expertise to design the most effective visualizations. There is a lack of studies that characterize the types of problems that arise when domain experts design visualizations.

2.1.2 Models Characterizing Visualization Process and Design

Among the many theoretical models that exist in visualization, the ones that are relevant for our work fall into two broad categories: i) models which characterize the visualization process, starting from data transformation to human perception and cognition, and ii) models that capture the different aspects of a visualization design and its implications, especially from an end user's perspective. One of the earliest instances of a process model, was the data-state reference model proposed by Card [17], which was later extended by Chi's pipeline model [18] for representing different data transformation stages and the intervening operations. This was further extended by Ware [19] whose model focused more on the visual representation and its perceptual implications. For the visualization design models, we find instances where researchers have studied the use and creation of visualizations from a designer's point of view [20] or as the product of a collaboration between designers and end users [21]. Heer et al.[12] proposed a model for providing guidelines to novice users on the encoding type used. We propose a taxonomy model, which is similar in its characteristics with the visual uncertainty model [22] that combines both visualization process and design in one holistic framework. The functionality of our model is similar in spirit with the work of Walny et al. [23], who generated a taxonomy by studying how visualizations on white boards are typically produced, what their purposes are and how people from outside the visualization community use visual thinking for solving their problems.

2.2 How Climate Scientists Use Visualization

Our collaborators are climate scientists specializing in Terrestrial Biosphere Models (TBMs). As visualization researchers, we worked closely with them in the Multi-Scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP) as part of the DataONE Scientific Exploration, Visualization, and Analysis (EVA) Working Group. In this section we describe briefly what domain-specific problems they aim to address by designing visualizations.

TBMs simulate terrestrial ecosystem processes and the terrestrial-atmosphere carbon exchange in relation to prescribed boundary conditions: vegetation cover, soil properties, climate, etc. They have become an integral tool for extrapolating local observations and understanding to much larger terrestrial regions, as well as for testing hypotheses about how ecosystems will respond to changes in climate and nutrient availability [2]. TBMs use complex analysis scenarios and generate diverse and large volumes of multidimensional data. Visualization thus constitutes an integral component of most model output analysis processes not only for understanding and representing the data, but also for subsequent dissemination of the scientific knowledge. Visual representations in the form of images and charts used in academic publications and presentations play a critical role in communicating the scientific findings to a broader community.

A key approach for environmental modeling is to use multiple models as a way to gain confidence in the results and bound understanding. Therefore, intercomparison of a suite of terrestrial biospheric models over space, time, and different land cover types is an important research area. But the volume and complexity of model outputs present many challenges for analysis and visualization.

To gain additional confidence in model output, researchers compare observations with model simulations in a benchmarking activity. Furthermore, modelers want to know which models are similar, and why, when, and where they are similar. Linking model structures with model output can be used to understand why models are different from benchmarks and each other. Visualization plays a crucial role in all

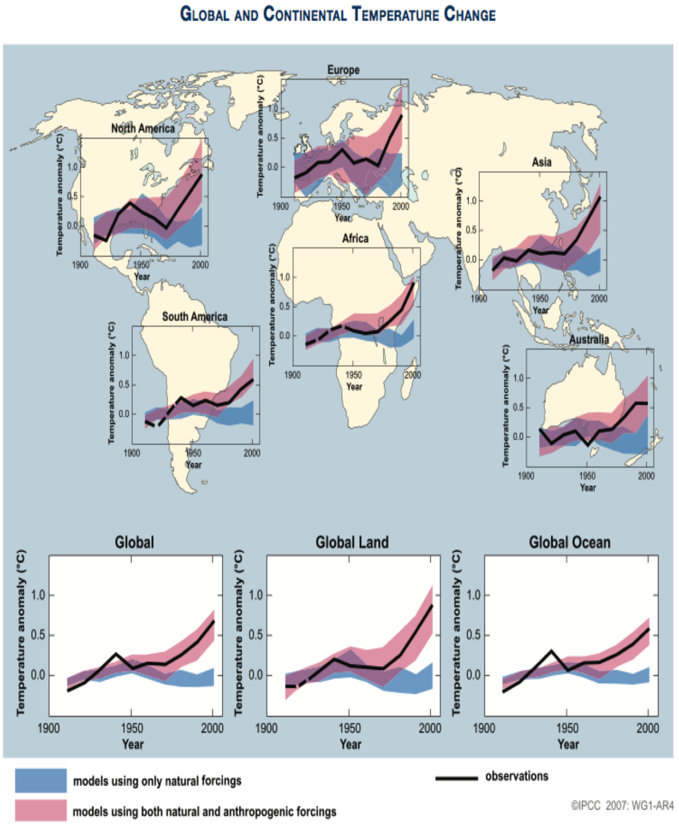


Figure 2.2: **One of the few examples of optimal visualization design from our collected sample [24].** where the intent was effectively captured and communicated. Here small multiple line charts are used in conjunction with maps for showing region-wise temperature variation of two classes of climate models.

of these steps for understanding model characteristics and visually representing the scientific findings.

2.3 Methodology

The goal of our study was to diagnose design problems in visualizations created by climate scientists. While in course of our research we also discovered visualization examples which adhered to the best practices (Figure 2.2), our aim here was to focus exclusively on the causes and consequences of the problems. We followed a descriptive approach where we could provide useful guidelines for climate scientists and discover challenges for visualization researchers. To achieve this purpose we adopted a qualitative methodology featuring in-depth analysis of climate visualization examples, generation of descriptive classifications schemes, as well as multiple interviews, workshops, surveys, and focus groups. In the following section, we describe our methodology in details.

2.3.1 Participants

In the course of our project we collaborated with 20 climate scientists, with 5 of them being direct collaborators from the Multi-Scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP). Most of them have over ten years of experience in climate modeling. The overall goal of MsTMIP is to provide feedback to the terrestrial biospheric modeling community to improve the diagnosis and attribution of carbon sources and sinks across regional and global scales. Our group of collaborators were mostly climate modelers working as part of the EVA working group under the DataONE initiative; and they spanned across different national labs and universities within the United States.

With our direct collaborators, we interacted over a six month period via semi-structured interviews, which were both in-person and through teleconferences, and three workshops where we exchanged knowledge about our respective domains and conducted interviews. With the indirect collaborators, we attended their presentations at workshops, took note of their visualization designs and received their feedback on our findings through teleconferences.

2.3.2 Evaluators

The group of evaluators who were involved in data collection, analysis, and synthesis; comprised of four data visualization experts: one doctoral student, one research scientist and two faculty members. All the evaluators have at least four (and for two of them more than ten) years of research and practical experience in visualization. The coding part of our work loosely follows the tradition of expert-based evaluation of user interfaces like *heuristic evaluation*, where it has been demonstrated that a small number of experts can reliably detect most of the problems [25].

Also, following-up on the same tradition, rather than relying exclusively on the personal judgment of the evaluators, we created guidelines and support material to inform and guide their work. Since a single established set of visualization heuristics does not exist yet, we decided to: i) review the few initial attempts to create visualization heuristics we found in the literature [26, 27]. and ii) create our own synthesis of visualization principles drawn from the visualization design and research literature. We provide more details about the synthesis of visualization principles in Section 2.3.5.

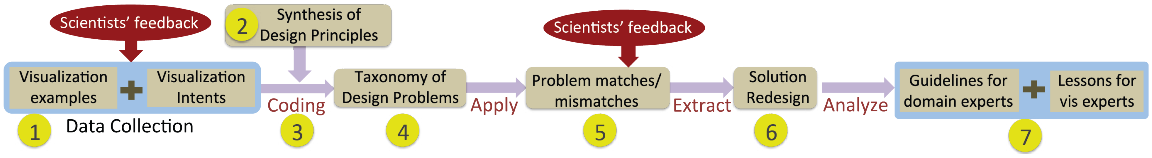


Figure 2.3: **Workflow for our qualitative study** comprising of seven different stages as annotated in yellow. The workflow highlights the tight interaction with climate scientists that led to the taxonomy of design problems, its subsequent refinement and application for finding solutions. The problems and solutions were further analyzed and reflected upon for providing general design guidelines to domain experts and highlighting lessons learnt for visualization researchers.

2.3.3 Workflow and Goals

The workflow we adopted for our exploratory study is outlined in Figure 2.3 and is characterized by seven distinct stages we performed to gather the necessary data and perform our research.

In (1) **Data collection** we interacted with our collaborators, through in-person meetings and teleconferences to collect *visualization examples* and *intents* that are representative of the typical tasks performed by the climate scientists. In (2) **Synthesis of design principles** we reviewed the existing literature on visualization design principles and organized them into a reference list we used to inform and guide our critique of visualization examples. In (3) **Coding** we used the reference list to manually annotate the collected examples and generate descriptive codes that captured potential design problems. In (4) **Taxonomy generation** we systematically and iteratively refined and organized the codes to generate a design problem taxonomy. In (5) **Problem matches and mismatches** we discussed representative examples of the collected design problems with the climate scientists to gather instances of diverging opinion between visualization experts and the domain experts. This phase allowed us to refine the taxonomy and to build a much richer view on how visualization design principles can and should be instantiated in practice. In (6) **Solution Redesign**, based on the suggestion of our group of collaborators, we extended our analysis to include discussions of solutions. We redesigned some selected examples and gathered additional feedback from the scientists. Finally, in (7) **Guidelines and lessons learned** we reflected on the output generated by our analysis and interactions and came up with a set of general guidelines, pitfalls and lessons learned.

For data collection, the qualitative analysis of the examples, and the following derivation of a design problems taxonomy; we followed the grounded theory methodology [28]. Grounded theory is a systematic methodology used in social sciences to derive classification schemes from the analysis of large quantities of qualitative data and guides the researchers through iterative phases of data collection, code generation, and their organization into descriptive categories. This approach recently gained some momentum in the visualization research community and it has been successfully adopted in a growing number of studies to analyze visualization artifacts and their use in real-world scenarios [29, 30, 31, 23, 32]. This methodology allowed us to critique and analyze visualizations created by scientists without any pre-formed hypothesis, thereby allowing the data to dictate the taxonomy that emerged. The aspects we imbibed from grounded theory were alternating between data collection and analysis, refining the conceptual relationships within the data, and subsequent generation of a theoretical paradigm for structuring our findings.

2.3.4 Data Collection

During our data collection phase we first interacted with the climate scientists to generate a representative sample of visualizations to be used for our analysis.

2.3.4.1 Visualization Examples

The visualization examples were collected in consultation with our collaborators, from a set of 20 presentations in two workshops, 15 research papers from the climate science domain, and four interview sessions. Our effort was to ensure that the collected sample represents the state-of-the-art in visualizations used for comparing climate model data. From these sources we generated a total of 105 images which we used as the basis for our study. Given the high experience level of our collaborators we were confident from our interactions, that these images constituted a representative sample for our study. Among the examples we collected, 80% of the visualizations comprised of geographical maps, scatter plots and variants of line charts. The remaining 20% was a heterogeneous set of examples which could not be organized into any consistent group or description. For this reason we decided to exclude them from the analysis and focused on 80% consistent group of images, which comprised of 40 line charts, 30 geographical maps, and 15 scatter plots.

2.3.4.2 Visualization Intents

In a preliminary coding pass we realized it was hard to judge the merit of the collected examples without first knowing the visualization intents. Rather than evaluating the collected example exclusively under an abstract set of principles, we preferred to ground our analysis on the following main questions:

Q1. Does the chart represent the intent correctly?

Q2. Does the chart convey the main message efficiently and effectively [33]?

These high-level questions guided the latter stages of the evaluation pipeline, such as the synthesis of design principles and coding.

As we felt we could not derive the intent and message of every single example without the help of our domain experts, we conducted surveys and teleconferences to come up with a more reliable set of intents. To achieve this purpose we performed the following steps: i) two investigators formed an initial idea of the intents from the descriptions in the collected material and collected them in a document containing pairs of images and intents; ii) we distributed a survey with the intents to a group of 2 scientists asking them to mark whether they agreed or disagreed with the stated intent and to add their own version of the intent where necessary; iii) for those cases where the scientists disagreed (with us or between them) we performed an interview session for further clarification and collected the right intents.

This phase was crucial in understanding the motivation and context behind the creation of the visualizations. In some cases, we initially misread the intent, which was subsequently clarified in these interview sessions. For example, as shown in Figure 2.1(a), we initially deduced that the initial intent was to identify the models which showed high or low values for the carbon flux output variables `NEP`, `GPP`, `Rh` (organized vertically in the three stacked panels). But in an interview session it was clarified that the primary intent is to show the overall variability of prognostic (green) and diagnostic (purple) models and also to show the which models belonged which ecoregions for the different variables.

An analysis of the collected intents showed that in the context of climate model intercomparison, the general intents of scientists were mainly to identify similar models and compare their spatial or temporal variability over different granularity of space and time. In previous work, we had designed an interactive interface that helped scientists realize these intents through a rich exploratory visualization tool [6]. In this work, we judged the design problems of the static visualizations with respect to these intents.

2.3.5 Synthesis of Design principles

The basis for our judgment of design problems was the rich literature of visualization design principles that have evolved over time. While we observed that there is no single conceptual framework which can be applied for such analysis, various theoretical principles, starting from Bertin's seminal work, to the most recent research on information graphics, have guided visualization design over the years. These design principles come from different areas of visualization research, and address different but often complementary issues such as: principles for optimal visual design, criteria for design based on data type, matching the visual properties to best support human perception and cognition, and also more recently, how best to communicate the message by properly structuring the information. All these categories subsume the two general questions **Q1** and **Q2** on which we eventually based our judgment of the visualizations. Before starting the coding phase, we worked on collecting design principles and synthesizing them in a way that codes based on those principles could be used for the study. Note that these principles were applied only in the context of the scientists' intents for any visualization. Here, we provide a brief summary of the main sources we used to inform our research.

2.3.5.1 Optimality of design

From Tufte's seminal work [9] and Tukey's research on statistical graphics [34] we adopted the principles of graphical excellence and integrity. Graphical excellence encompasses a number of criteria for design by maximizing the *the data/ink ratio* (*i.e.*, information and data density) and avoiding accessory elements and embellishments. Graphical integrity refers to a truthful representation of the data (*related to* **Q1**) for avoiding potential misinterpretation due to scaling issues or distortion.

2.3.5.2 Criteria for design based on data type

From the early work of Mackinlay [33] we adopted the criteria of *expressiveness* and *effectiveness* to qualify visualizations in terms of the encoding parameters and degree of salience of visual attributes. From Bertin's seminal work [35] we derived principles of effective visual encoding of data features into visual variables. We also considered the work of Card et al. [36] and MacEachren et al. [37] which revise and extend the early of work of Bertin respectively in information visualization and geo-visualization. These enabled us to judge the appropriateness of the visualization parameters (*related to* **Q1**) based on the data type they represented.

2.3.5.3 Perceptual Implications of Design

Bertin's seminal work on visual variables such as position, size, shape, color, orientation, and texture; formed the most important basis of our judgment of the appropriateness of the design parameters. Bertin focused on defining the possible visual variables and reflected on their perceptual implications for visualization design. The work of Bertin was extended by Cleveland and McGill [38] and Ware [19], who provided much needed empirical evidence of the perceptual effectiveness of visual variables through controlled user studies. Together with Bertin's work, such experimental research form the core of the science of visualization. *For addressing* **Q2**, we utilized the following concepts inspired from these threads of research: the importance of visual encoding keeping pre-attentive processing in mind, ranking of visual variables based on different tasks, perceptual effects and properties of color, importance of spatial organization of visualization design, etc.

2.3.5.4 Design for more effective visual communication

Finally we also considered recent approaches from data visualization practitioners. These mainly address the concern of how visualization should not only support exploration and analysis, but should also be able to visually communicate the data. *For addressing* **Q2**, we utilized design guidelines from Stephen Few's book "Show Me the Numbers" [10] and from "The Functional Art" [39] a data narrative-oriented book written by data journalism expert Alberto Cairo.

2.3.6 Coding

In the coding phase we analyzed all the image instances for potential problems with respect to the visualization intents that were collected in the initial phase. The codes we used for describing the problems were based on our synthesis of design principles. For each example we collected codes describing design problems and relevant issues. For instance, the scatter plot example shown in Figure 2.1(a), was coded with: *clutter, chart selection, and color map.* Wherever more clarity was needed, we resolved our doubts by asking further questions to the scientists. We met at regular intervals to share, compare, merge, and refine the set of codes; and after several iterations we reached a stable set. Halfway into this process, during our discussions we realized design problems sometime have non-trivial implications and solutions. For this reason we started collecting, together with problems, descriptions of design consequences and their trade-offs, which are presented in Sections 2.4.3 and 2.7.

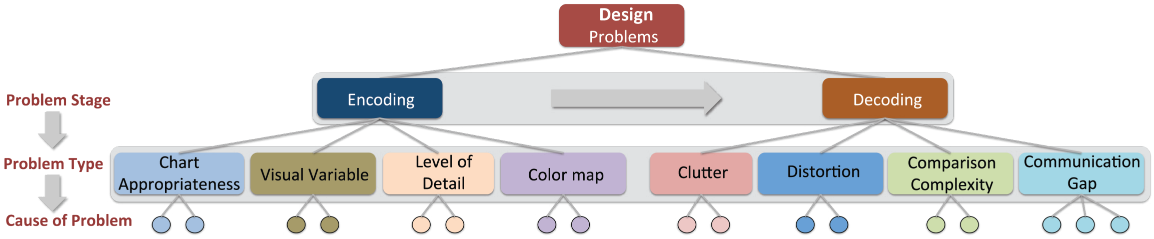


Figure 2.4: **Different levels in the design problem taxonomy**. Problems are categorized according to the stage in the visualization pipeline, the type and the cause. The leaf nodes for problem cause are shown in Figures 2.5 and 2.7. The most frequently occurring problems were the visual variable problem (37% cases), communication gap (30% cases), and clutter (29% cases); followed by color map choice (28% cases) and distortion (20% cases). Some of the less frequent problems were level of detail (17% cases), comparison complexity (15% cases), and chart appropriateness (13% cases).

2.3.7 Generating a Taxonomy of Design Problems

After collecting the codes we moved to the *axial and selective coding* [28] phase where we merged, grouped and structured the codes into a full taxonomy. During this phase we went through several refinements by having one of the investigators mainly working on the classification scheme and another investigator testing the scheme with the library of examples, while discussing inconsistencies collaboratively. We stopped the process when we felt that we reached a stable and satisfactory description of all the problems.

One of the issues during this phase was to choose an agreed upon level of abstraction for categorizing the design problems. For this we used a bottom-up approach by analyzing which problems are similar in terms of: which stage of the visualization process they were introduced and what effect they had on the visual representation and the perception of patterns. Accordingly, we came up with a three-level taxonomy, that helped us categorize the causes and implications of the problems.

2.4 Taxonomy of Design Problems

The taxonomy we have derived is a classification scheme where a visualization example can be associated with multiple design problems that are described by different nodes of the taxonomy tree (Figure 2.4). For deciding a classification that captures the causes and effects of design problems, we took inspiration from the taxonomy of visual uncertainty [22] based on the traditional information visualization pipeline [18]. The latter can be regarded as a visual communication channel [40] and thought of being composed of two distinct phases: encoding, that is associated with mapping the data on to the screen-space; and decoding, that is associated with perceptual and cognitive processes on the user's side. The classification scheme, based on encoding and decoding stages as the first level, enables us to systematically analyze different dimensions of the design problems. The levels are described below:

i) Problem Stage. The first level decides whether it is in the encoding or the decoding stage that a design problem is found. Encoding deals with problems that mostly depend on the choices the designer makes when deciding how to transform data points to visual features. Problems on the encoding side enable us to address if the encoding strategy (*e.g.*, choices involving chart type, visual variables, color map) adheres to visualization best practices. Decoding captures design problems that go beyond the specific scope of visual encoding and the deliberate encoding choices a designer makes. These problems may be due to the effect of the limited screen resolution, perceptual implications of the visual parameters and auxiliary elements (grids, legends, annotations). They mainly affect the effectiveness with which the user can decode the presented information.

ii) Problem Type. An encoding problem or a decoding problem can be classified into multiple types, depending on the problem characteristic. The type level separates the encoding and decoding problems into different classes, which encapsulate the low-level causes of these problems. The classes belonging to the encoding side reveal the gaps in fulfilling the necessary conditions for a good encoding. For example, it could reveal if the chart type and visual mapping were appropriate. Fulfilling the necessary conditions for a good encoding might not always be sufficient for a visualization to be useful. The classes belonging to the decoding side reveal if the necessary conditions were also sufficient, by revealing if there was too much clutter, or if the there was distortion of information or the visual complexity was too high.

iii) Cause of problem. This is the level at which leaf nodes of the taxonomy tell us the precise cause of the problem. These causes reveal the low-level details of the problem types. For example, from this level we know the cause of a color map problem or the cause of a distortion problem.

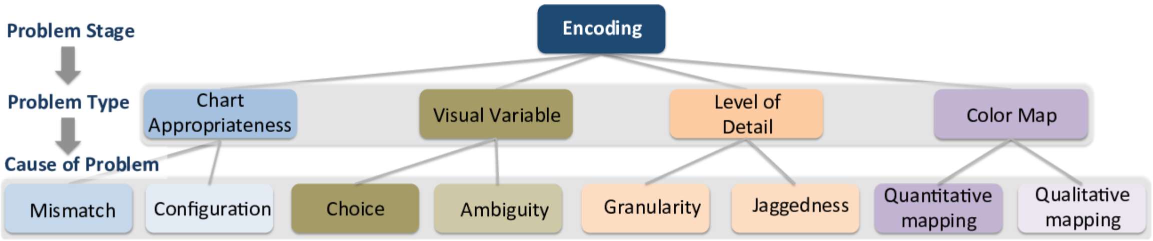


Figure 2.5: **Design problems found at the encoding stage of the visualization pipeline** reflected how well the input parameters such as chart type, visual variables, level-of-detail and color map were chosen by the scientists. Darker color at the lowest level indicates higher frequency of a particular cause (*e.g.* Choice) within a problem type (*e.g.* Visual Variable problem).

2.4.1 Encoding Problems

The encoding side of the taxonomy helps us ask questions such as: "*Do the encoding parameters such as chart selection reflect the intent?*", "*Can the visual mapping and color map choice be improved?*", "*Given the intent, is the data shown at an appropriate level of detail?*". In this section we describe the causes of design problems during the encoding stages (Figure 2.5).

2.4.1.1 Chart appropriateness

The first design decision that the scientists have to make for reflecting their intent, is which chart type to use. The chart appropriateness issue deals with whether the charts selected by the scientists appropriately reflected their intent. For judging this problem, we analyzed if any inherent limitation of a chart type, or the resulting configuration of the visual representation interfered with the intent. The two causes for the appropriateness problem were as follows:

Mismatch. Mismatch captured cases in which the chosen visual representation was not the best option for conveying the intent due to its inherent limitations. This issue was observed mainly in scatter plots. For example, one of the intents in the scatter plot in Figure 2.1(a) was to find which models belonged to which ecoregions. The author attempted to convey this intents through a variant of a traditional scatter plot, where the X-axis represents a categorical attribute (the ecoregions) rather then a numeric value, as is usually expected. This unexpected configuration created confusion among the scientists and made the chart difficult to interpret. A scatter plot is unable to clearly show the models that belong to a particular ecoregion due to over plotting. An additional problem is the use of the many different symbols for representing each model, which leads to an inefficient conjunctive visual search. In Section 2.6 we discuss a solution to this problem.

Configuration. Configuration problem deals with the arrangement of multiple charts in a common canvas. We found several examples where scatter plots, line charts and maps were stacked together for comparing climate model behavior. This can be observed in Figure 2.6. The intent in this case was to compare the temporal variation of annual cycles. But the horizontal stacking of the line charts, where time was represented on the X-axes, made it difficult to compare the Y-axes values. This problem was also observed in Figure 2.1(b) where multiple maps where stacked together without any ordering based on similarity.

2.4.1.2 Visual Variable

Visual variable problem captures cases in which the designers made poor choices in the mapping data attributes to visual variables. This is one of the most important design decisions in the visualization process [35, 36]. Since the number of data attributes generally outnumbers the number of visual variables (position, shape, size, color, orientation, etc.) by far, effective utilization of the latter is crucial in designing effective visualizations. The two causes for the visual variable problem were as follows:

Choice. The *choice* of visual variables was one of the leading causes of problems we found in our collection. While the choice affects all the subsequent visualization stages of human perception and cognition, here we focus on how visual mapping can "*above all show the data*" as suggested by Tufte's principle of graphical excellence [9]. The different classes of problems due to choice of visual variables were: representation of discrete data attributes in scatter plots and line charts using a combination of visual variables, and use of color as a quantitative channel for comparing averages and differences on geographical maps. For example in the scatter plot in Figure 2.1(a), one of the main problem was the representation of discrete data attributes, *i.e.* climate models, using visual variables such as shape, texture, and orientation concurrently. The use of multiple symbols causes conjunctive visual search [19] which is inefficient and not a good use of the pre-attentive capabilities of the human vision system. Moreover, combination of shape (different symbols), texture (filled and unfilled shapes), and orientation (triangles pointing in different directions), do not adhere to the rule of integral and separable visual dimensions [19]. In line charts, a recurring problem was

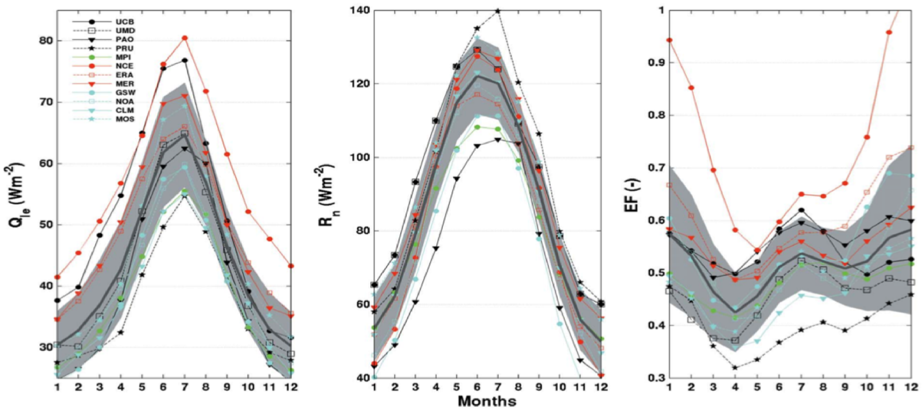


Figure 2.6: **Problems due to clutter: color mixing, visual variable problem: ambiguity, and chart appropriateness: configuration** The intent behind this multiple line chart figure [41] is to enable readers to analyze the variation of annual cycles over time in terms of the ensemble mean, the standard deviation and the individual values. It is difficult to compare temporal trends due to the side-by-side placement. Color mixing among the lines causes clutter.

the use of dots, solid lines, and dotted lines which would create difficulty in recognizing and tracing the different items.

Ambiguity. Another category of problems with visual variables, was ambiguity, where the use of visual variables was inconsistent: either different visual variables were used to represent the same data attribute, or the same visual variable was used to represent different data attributes. While the choice of visual variables reflects how well the latter reflects the different data properties, ambiguity reflects if even after a correct choice was made, there were additional inconsistencies. For example, in the line chart example in Figure 2.6, there are only a few different colors used for representing the different categories. It almost seems there is a relationship among them, although none is explicitly mentioned in the text. The same problem was observed in maps, where a white or grey was used to represent two different factors: absence of data and lack of correlation among values that are represented. As evident ambiguity can lead to misinterpretation of the data where a relationship can be deduced even if there is none and if there are multiple relationships, only one of them might have been conveyed.

2.4.1.3 Level-of-Detail

For the visual encoding of data attributes, it is important to choose an appropriate level-of-detail that would not only preserve the fidelity of the data as much as possible, but also effectively communicate the intent. The two causes of the level-of-detail problem were as follows:

Granularity. This problem was observed in cases where either a coarser or finer granularity level could better reflect the intent. For example, a recurring issue with the maps was that pixel-based representation was used for mapping quantitative variables and enabling model comparison. While this led to high fidelity, comparison across multiple maps was difficult because of the low-level details that readers had to classify and compare. As shown in Figure 2.1(b), this would be inefficient as the lowest level of granularity would not facilitate a high-level overview of the salient patterns. It would instead cause a sequential search for finding similarities and dissimilarities among the maps.

Jaggedness. charts representing time series. The salient peaks and crests in the time series were occluded because of the jaggedness. The main source of the problem was the tendency of the scientists to plot daily or monthly data, even when the intent was to show the annual variability of any given entity. In those cases smoothing could be used by computing an average and that would highlight the main trends. In the redesign Section (Section 2.6), we present a line chart example (Figure 2.12) that shows these jagged patterns.

2.4.1.4 Color Map

Choosing an appropriate color map is essential for the effective display and analysis of data. Based on fundamental human perceptual principles and the type of data being displayed (sequential, diverging, or categorical), there are formal and systematic ways to make an appropriate color choice based on the task at hand. Color maps for quantitative attributes are important for making an accurate judgment, while those for qualitative attributes are important for distinguishing among different categories efficiently. Since the implications of these two types of color maps are different, we treated these two problems separately:

Quantitative mapping. For quantitative color maps, the rainbow color map was used in most cases. As extensively documented in the visualization literature [42, 43], the lack of perceptual ordering and isoluminance in case of rainbow color map can cause inaccurate interpretation of the data. It has also been shown in case of scientific data, the crucial role that a perceptually motivated color map plays, for example in

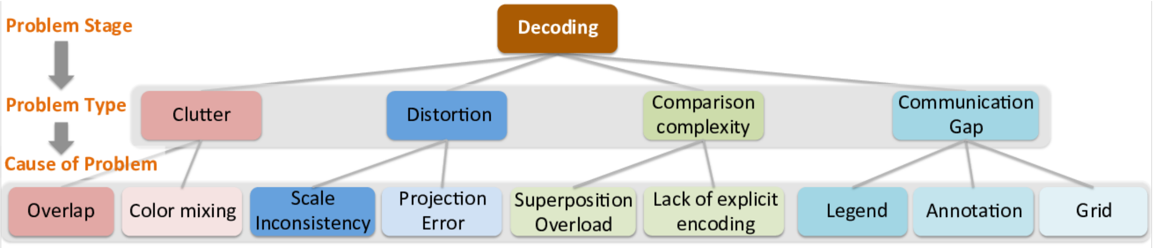


Figure 2.7: **Design problems found at the decoding stage of the visualization pipeline** reflected how the communication of the intended message was affected by the design choices. Darker color at the lowest level indicates higher frequency of a particular cause (*e.g.* Overlap) within a problem type (*e.g.* Clutter).

case of diagnosing heart conditions [44]. From the examples we collected, geographical maps suffered from this problem the most. In many cases we found that scientists are only interested in recognizing the extreme values, and the colors red and blue are associated with the semantics of temperature: red signifying hot regions and blue signifying cooler regions. But in many of those cases, all the hues of the rainbow are used for encoding the data. A divergent color map with only a luminance variation [45] would be appropriate in that case.

Qualitative mapping. For qualitative color maps, the problem was when representing discrete variables with color (Figure 2.12). If the hues are not separable enough, visual search for the variables would be inefficient. We found that it is a common requirement for the climate scientists to represent more than 10 discrete variables (in the form of regions or climate models) in a single chart. If color is the chosen visual variable, the choice of hue then becomes critical. The Tableau 20 color palette can be used in that case. ColorBrewer [45] only offers about 11 distinct colors.

2.4.2 Decoding Problems

Once the encoding parameters are chosen in the design process, to the judge the quality of the visualization, we have to judge its perceptual implications. Analysis of problems at the decoding stages of the visualization, that is the perception and cognition stages, enables us to evaluate a visual representation by asking questions such as: *"Is it perceptually confusing?" "Does it represent the patterns without distorting it or being too complex?" "Does it emphasize the intended message clearly enough?"*. In this section we describe the problems caused during these stages of the pipeline (Figure 2.7).

2.4.2.1 Clutter

We adopt the definition of clutter which relates the degradation of a display with the number, representation, and organization of items [46]. Many of the visualization examples, across maps, line charts, and scatter plots were cluttered and there were different reasons for clutter. The two causes of clutter were as follows:

Color mixing. Color mixing (Figure 2.6) was one of the causes for clutter. This is different from the color map problem, because color mixing mainly occurred between the the chart elements and the background or among the different sets of symbols. For this case, the color map could have been appropriate, but there needed to be another extra degree of caution for avoiding color mixing. Color mixing was observed mainly in maps and line charts. For example in the line chart in Figure 2.6, the color mixing between the grey band, the grey mean line and the chosen colors for the other lines cause clutter.

Overlap. Overlap encompasses over-plotting of data items in scatter plots and maps, and large number of crossings in line charts. In some instances of line charts, the thickness of individual lines made it difficult to identify and trace the paths of individual lines. Over plotting of different visual variables on a scatter plot (Figure 2.1(a)) made it difficult to recognize and visually search for the individual data points. While over plotting and overlap are artifacts of the representation, and are often unavoidable, the key question here is whether these artifacts interfered with the intent. For example, in case of the scatter plot example, over plotting interferes with the intent, as identifying each model is one of the intents of the author.

2.4.2.2 Distortion

Distortion of the data in a visual representation is a serious problem that can either lead to potential misinterpretation or an inaccurate perception of the data, especially when quantitative attributes are involved. The causes for distortion are as follows:

Scale inconsistency. Choosing different scales for the same variable leads to inconsistent representation of the patterns. This is a decoding problem, because the chosen scale is appropriate for a given variable, but when multiple variables are involved, lack of attention to consistency can mislead readers. We found this to be a problem in some geographical maps, where a single rainbow color map was used to represent data which were at different scales, leading to misrepresentation of the patterns. In some

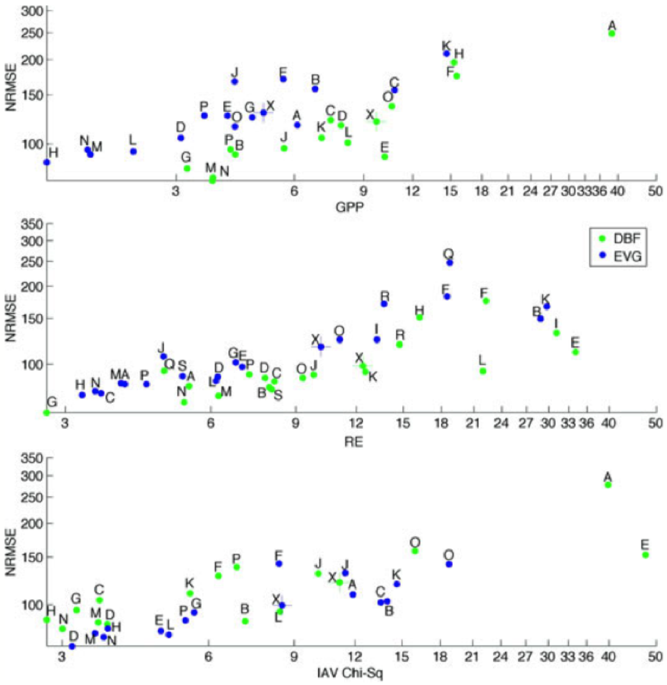


Figure 2.8: **Problems due to chart appropriateness: mismatch, distortion: scale inconsistency, comparison complexity: superposition overload, communication gap: legend, annotation.** The intent in this figure [47] was to compare the errors of two plant functional types (`DBF`, `EVG`), color-coded in green and blue, and quantified in two ways: NRMSE and Chi-Sq. The X-axis in all three scatter plots represent the Chi-Sq statistic. It is hard to separate the patterns between the two functional types: a regression line and annotation of key trends would more clearly communicate the message.

scatter plots and line charts (Figure 2.8) we found that the scale of one or more of the attributes are different from others. In this case the X-axis represents the chi-squared statistic and in the topmost line chart, the tick placement is different from the others, signifying a different log-scale for encoding the data than others.

Projection error. This problem is typical of maps, and due to the inherent mapping from a 3D sphere to a 2D surface, we observed projection error in some of the map examples. We observed that while some error is unavoidable, use of better projection techniques could reduce the amount of the error. For example, an equal-area projection will be more appropriate in displaying area-sensitive data like fire-burnt area.

2.4.2.3 Comparison complexity

The main goal in TBM domain being model intercomparison, the primary intent behind most visualizations was to facilitate comparison at different levels. We found that the comparison complexity in terms of number of data points per chart, or number of charts per views, or their placement, led to some design problems. We take inspiration from Gleckler et al.'s taxonomy of visual comparison methods [48] for categorizing these problems. We classified this as a decoding problem because even a correct encoding choice could lead to comparison complexity and influence the communication of the main message. The two causes of comparison complexity were as follows:

Superposition overload. This category deals with the case where the number of entities in a chart are far too many for facilitating an effective comparison. This issue was observed mainly in line charts and scatter plots. As opposed to a small multiple display, a large single [49] was often required by climate scientists for comparing models to observations, or comparing individual values to ensemble mean. In some of those cases, superposition overload of too many elements led to clutter (Figure 2.12) and in some case, although clutter was not caused, superposition overload interfered with the intent (Figure 2.8). In the first case, the drawbacks of the superposition are obvious and we discuss a small multiple solution to this problem in Section 2.6.3. We found small multiples being used by scientists in case of maps, but we found only one line chart example where a small multiple was used for reducing overloading.

Lack of explicit encoding. This case with the issue where explicit encoding of relationship among the compared entities would have led to better design. It has been observed that small multiples are important while visualizing multiple variables [9] but care should be taken to position and sequence the individual charts appropriately so that visual search is optimized [10]. For example, in case of the multiple maps (Figure 2.1(b)), the intent here is to deduce the degree of similarity among the different models. However a random arrangement does not immediately show how much similar, the maps are. It requires visual inspection almost on a pixel-by-pixel basis for judging similarity. In that case, extracting some summary statistic about the degree of similarity and using that for positioning the maps seemed to be a good solution. We will discuss this solution and its evaluation in the case study section.

2.4.2.4 Communication Gap

In many of the visualization examples we collected, we found problems with factors which do not directly interfere with the intent, but might create problems with communication of patterns. These are auxiliary information about charts, which were categorized as follows:

Grids. Grids can be used for chunking the important pieces of information, which might not be intuitive immediately. Human brains are good at picking out patterns. Often, fairly small changes to a graphic layout that strengthen the appearance of grouping or other types of patterns will add to the ability of the graphic to deliver an instant impression or overview of the message being communicated. While unnecessary grid lines must be avoided in keeping with the idea of minimizing non-data ink proposed by Tufte [9], judicious use of grids help in capturing the reader's attention to the salient portions of the chart. For example, use of column-wise grid lines in Figure 2.1(a) could separate the ecoregion-wise patterns for the different models. In our collected examples, we found two scatter plots were grids were used to chunk the information space, for denoting groups of data points belonging to a model or a year.

Legend. In some visualization examples, we found that the charts are not self-contained: lack of legends for different symbols or relationships meant one has to either browse through the captions or the textual description for making sense of what the symbols mean. This was especially difficult where lots of different symbols are used on a chart, for example, the scatter plot shown in Figure 2.8.

Annotation. In different visualization examples, we observed that annotation of salient patterns or data points on the chart could communicate the intent or some other critical aspects more effectively. For example, in case of the multiple maps (Figure 2.1(b)), the white color on the maps denotes a lack pf spatial extent, but that is not documented within the image itself. An annotation would clearly communicate this important aspect of the chart.

2.4.3 Problem consequences

After we created the final version of the taxonomy, we realized that while the taxonomy enabled us to categorize the problems and their causes, it did not capture their severity, and most importantly, their impact. In light of the numerous trade-offs

Causes of Problem	**Consequence**
Visual variable problem: *ambiguity*	Misinterpretation
Distortion: *scale inconsistency*	
Distortion: *projection error*	Inaccuracy
Distortion: *scale inconsistency*	
Color map: *quantitative mapping*	
Chart appropriateness: *mismatch*	Lack of expressiveness
Chart appropriateness: *configuration*	
Visual variable problem: *choice*	
Level-of-detail: *jaggedness*	
Visual variable problem: *choice*	Inefficiency
Level-of-detail: *granularity*	
Color map choice: *qualitative mapping*	
Clutter: *color mixing*	
Clutter: *overlap*	
Comparison complexity: *superposition overload*	
Comparison complexity: *lack of explicit encoding*	
Communication gap: *grids*	
Comparison complexity: *lack of explicit encoding*	Lack of emphasis
Communication gap: *grids, legend*	
Communication gap: *annotation*	

Table 2.1: **Connecting design problems to problem consequences sorted by severity**. *Misinterpretation* has the highest degree of severity owing to the misrepresentation of the intent. *Lack of emphasis* is least severe as the problems are dependent on the inefficiency of the visual communication process, and not the incorrectness of the representation.

a visualization practitioner has to face when creating a visualization, it would be useful to have: guidance on how severe a visualization problem could be, and a categorization of consequences it may lead to. To solve this problem we went through our list of problems again and consulted our synthesis of design principles which were based on the two high-level questions: **Q1**, about correctness; and **Q2**, about effectiveness and efficiency of visual representations. Based on these questions and inspired by the seminal work on graphical integrity [9] and the criteria of expressiveness and efficiency by Mackinlay [33], we created a list of potential problem consequences. These consequences bridged the low-level causes of design problems to high-level effects, which were more comprehensible from a domain scientists' point-of-view. The association of design consequences with design problems is shown in Table 2.1, sorted by their level of severity. The level of severity is defined by the graphical integrity principle [9], according to which the most important criteria for a visual representation is to represent the data correctly and accurately. The different problem consequences are described below.

2.4.3.1 Misinterpretation

Certain design problems could lead to misinterpretation of the data. Since this consequence directly interfered with the correctness of the interpretation, and violated the principles of graphical integrity [9], it had the highest level of severity. As shown in Table 2.1, ambiguity of visual variables and distortion due to scale inconsistency (Figure 2.8) could lead to the misinterpretation of the data.

2.4.3.2 Inaccuracy

In scientific data analysis, an important requirement for visual representations is to allow scientists to deduce accurate estimates from the display. When certain design problems could lead to an inaccurate interpretation of the data with respect to the original intent, they would cause inaccuracy. The most prevalent design problems that caused this issue were distortion due to projection error and quantitative color maps in the form of rainbow color maps. Problems like chart mismatch could also cause an inaccuracy problem. For example in case of the scatter plot example (Figure 2.1(a)), one has to mentally compute the spread of the different output variables, and therefore inaccurately perceive the differences in the spread.

2.4.3.3 Lack of expressiveness

The expressiveness [33] criteria dictated whether the visual representation matched with the properties of the data attributes. A lack of expressiveness condition would not clearly convey the intent as the certain aspects of the representation would not match the intent. The problems leading to lack of expressiveness from our taxonomy were chart mismatch, chart configuration, visual variable problem due to choice, level-of-detail due to jaggedness, and lack of explicit encoding.

2.4.3.4 Inefficiency

Efficiency of algorithms are measured in terms of speed. Inefficiency in visualization design could also be traced to the slowness of the interpretation on the part of the reader. When certain design problems did not directly interfere with the interpretation of the data with respect to the original intent in terms of its correctness or accuracy, but affected the speed and efficiency, they led to inefficiency. This category encompasses the principles of effectiveness [33], use of pre-attentive features [19] and visual variables for efficient search for patterns [38]. The problems leading to this consequence based on our taxonomy were level-of-detail due to granularity, qualitative color maps, clutter due to color mixing, superposition overload, and communication gap due to lack of grids and legend.

2.4.3.5 Lack of emphasis

In static charts it is often important to draw the reader's attention to salient portions of patterns that have higher priority than the rest. This can be done by highlighting different aspects and organizing the information in a structured way [10]. While these do not directly correspond to the data being shown, the emphasis on the key aspects of a chart affects that message that readers decode from a chart. The problems leading to a lack of emphasis consequence based on our taxonomy were lack of explicit encoding, grids and annotations. Since this consequence does not directly interfere with the intent, it has the lowest level of severity.

2.5 Matches and mismatches

At this point of our study, we realized we had the opportunity to get back to our group of climate scientists and get feedback on our categorization of the design problems. We realized this would not only be a useful way of validating our work, but it would also be interesting to observe how visualization problems compiled by a group of visualization expert would be received by a group of domain scientists. We realized that while extensive research exists on reporting design problems when evaluating visualization and, as we have seen above, on providing visualization guidelines, there's very little understanding or even exploration of how criticism and guidance is received and used by domain experts. We were interested in spotting cases where visualization experts and domain scientists disagree and dig deeper into why and how this kind of disagreement happens.

2.5.1 Interview Procedure

Before conducting the interview, for avoiding redundancy, we made a pass through all the problem categories, in an attempt to filter out the images which are very similar to one another. For example, in the case of a rainbow color map problem, we only showed a few examples which expressed the problem.

We arranged for a face-to-face interaction with our direct collaborators, as part of a workshop, and the entire interaction lasted for about *four* hours. We described the taxonomy along with problems from the collected examples. Since some of the scientists did not have a background in visualization, we first gave examples of best practices in the choice of visual variables, color maps, chart selection, etc. We exercised caution in not using too many technical terms, but explained the problems as illustratively and simply as possible. We asked the scientists to fill up a spreadsheet where they had to write if they agreed or disagreed with a design problem. They also added in a comments section, the reasons for their disagreement.

We realized that there could be disagreement among the climate scientists themselves, about design problems. We did not want an apriori settlement of their disagreement, but instead wanted to collect raw data about the same and see if there is a majority disagreement. Therefore we requested our collaborators to record their feedback independently. After collecting all the responses, we separated cases which had majority agreement (more than half the people agree about a design problem) and majority disagreement (more than half the people disagreed among themselves in acknowledging a design problem).

2.5.2 Cases with Majority *Agreement*

Figure 2.9 shows the distribution of the percentage of majority agreement and disagreement, sorted by high percentage of disagreement majority from top to bottom. We can observe that the scientists were generally in agreement with problems that lead to the most severe consequence, *i.e.*, misinterpretation: scale inconsistency and ambiguity of visual variables. Also, there was a high percentage of agreement for comparison complexity problems and communication gap problems caused by lack of legend and annotation. As observed from Table 2.1, these categories lead to lack of emphasis consequence whose degree of severity was low and did not directly interfere with the communication of the intent.

In course of our interaction with the scientists we could reason with this apparent dichotomy, that is, they tended to agree with problems with lead to consequences with both highest and lowest degrees of severity. One of the reasons was that, the scientists could immediately recognize why certain problems led to misinterpretation as the visual representation in those cases misrepresented the data. There were other cases, like the comparison complexity problem, where majority of the scientists agreed with the problem, but they were not aware of the solution. It took a while for us to illustrate how lack of explicit encoding or superposition overload hindered their main intended task, which was comparison of models, and which solutions could work better. We showed them sketches and examples of how these problems could be solved by

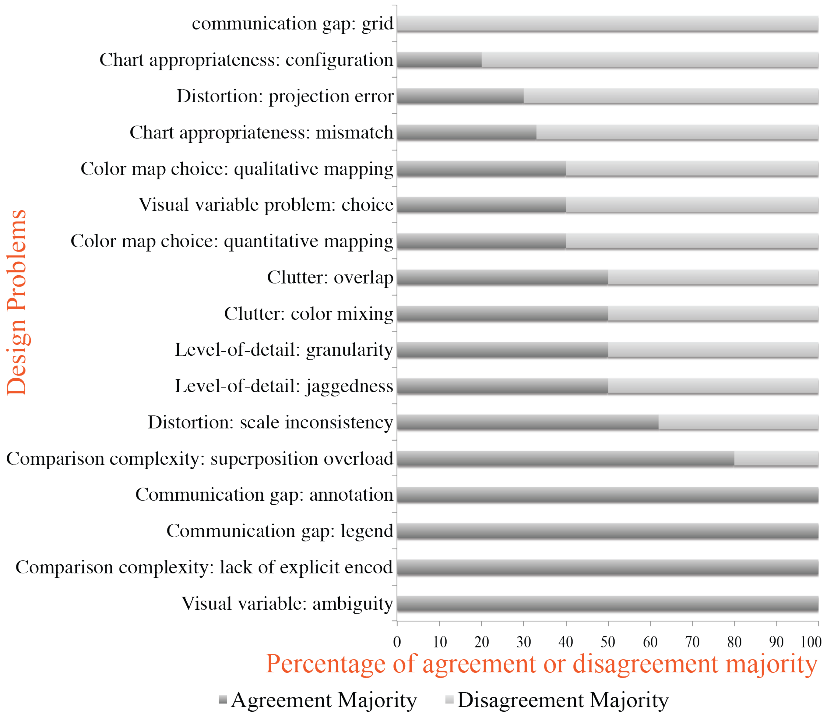


Figure 2.9: **Causes of problems sorted by high percentage of majority disagreement** from top to bottom along the Y-axis. Scientists were mostly in agreement with the most severe problems. They were generally in disagreement with problems which required significant knowledge about visualization best practices and those which were in conflict with the domain conventions.

making the decoding process more efficient and emphasizing the salient patterns; after which they agreed with the problem. Understanding the effect of the communication gap problem caused by lack of legend or annotation did not require much visualization expertise. In several cases the scientists commented:

> *"This figure desperately needs a legend, it is so difficult to flip back and forth to know what the symbols mean."*

For the annotation problem, the scientists sometimes acknowledged that annotation of the main trends would help them to focus directly on the main message instead of searching for it.

2.5.3 Cases with Majority *Disagreement*

During our interview, we found several instances where it was difficult for the visualization experts to convince the climate scientists about design problems, including well-known pitfalls like rainbow color map and $3D$ views. The cases with majority disagreement are shown in Figure 2.9. These were, especially cases where the problems did not directly interfere with the intended tasks. In other words, the problems did not have any misleading consequence, however they interfered with accuracy, expressiveness, efficiency, and emphasis. Some of the comments we got for problematic charts based on these consequences were:

> *"Improvements are subjective."*
> *"Minor problem."*
> *"This might be a problem but I am ok with this plot."*

Next we describe categories of major disagreements we have found during our interactions with the scientists. They are categorized in three main classes:

2.5.3.1 Domain Conventions

We found that the source of some of the design problems were existing conventions that the climate scientists followed. In some examples of line charts and scatter plots, we found them to be too cluttered or superposed with too many details to make sense of. For example, in many cases we found the use of dots along with lines on line charts, which cluttered the display and which we coded for both clutter due to overlap and choice of visual variables. However scientists explained that the observed data is by convention encoded by black dots, and the simulated data is encoded otherwise to distinguish them, and enable comparison between the two categories, *i.e.* observed and simulated.

For the superposition overload problem we found a group of line charts, that were similar to the one in Figure 2.12, but with the additional complexity of multiple dots in addition to the lines, representing observation data. This was an obvious candidate for overload. However, some of the scientists asserted that this was more of a convention in the climate science community for representing observation data on top of lines for comparing simulation data, and they were comfortable with such a representation.

2.5.3.2 Loss Aversion

Similar to the tendency to avoid losses rather than acquire gains, which is popularly known as a *loss aversion* problem [50], climate scientists tended to focus more on avoiding loss of data in their visualizations, than on tuning the chart parameters for gaining insight from them. The recurring level-of-detail problem with line charts and maps exemplified this tendency. In examples where multiple maps were compared for understanding similarity of models, we suggested that a coarser granularity would facilitate more effective comparison. This was because if the number of maps being compared is more than three or four, it becomes difficult to perform a pixel-by-pixel comparison, where the data is encoded at the finest level of granularity.

For the granularity problem, were multiple maps were compared using a pixel-by-pixel mapping (Figure 2.1(b)), although there was largely a consensus, some scientists said:

> *"I would have to see the coarser version to really know if it is better, though."*

While one of them said:

> *"Difficult to get a widespread trend. Personally, My eye tends to be drawn to red dots so I may be missing much of the information presented. A seldom problem is that I have trouble mentally overlaying the different plots. Each plot is a different variable/model. From this visualization, I have trouble comparing the locations and extremism of values at the same geography."*

This showed that the comparison mechanism is not effective due to the low granularity but they were not convinces unless alternate solutions showed them the real benefit.

In many line charts, daily data were plotted where monthly or annual were being compared. Scientists observed that there is a need or tendency to show all the data because the time spent in extracting the data is significant. Also they believe that there might be some anomalies that might be missed by aggregating the data. Though all of them did not agree, there was general consensus about this fact. When shown alternatives with monthly averages computed, one of the scientists commented:

> *"I agree with you on this. But the situation is that people, especially scientists, they tends to show data as raw as possible. I think if this figure is used in scientific publications, it's fine."*

2.5.3.3 Awareness about visualization best practices

We found that in many cases scientists were not aware of what the visualization best practices are and why they should adhere to them. The categories which led to a lot of discussion between the scientists and visualization experts, involved use of $3D$

plots (coded under chart mismatch), use of color maps and choice of visual variables. It has been well documented in the literature that $3D$ plots lead to distortion and ineffective reading of the data [51]. However, since climate scientists are already oriented towards reading $3D$ volume visualizations, they did not think that a $3D$ layout for abstract data could create a problem.

The same applies to the color map problems. In most cases they felt that the since they are already used to reading data from rainbow color maps, a more perceptually motivated color map would not make a difference to the goal of the intent. In a few cases they commented that:

> *"I agree that the color map can be better but that would be a cosmetic changes and won't affect the intent."*

The effective use of pre-attentive features was also another category where climate scientists did not agree with most of the problems.. For example, the scatter plot in Figure 2.1(a) encodes all the models by using different symbols. Even in absence of over plotting, the different symbols would cause an inefficient, conjunctive visual search. We discuss later in Section 2.6 how we could avoid this problem.

2.6 Solution Redesign

During our interactions with the climate scientists suggested that they needed to look at some solutions for better understanding the consequences of problems and how to avoid those. We agreed while it was useful to directly see why some problems should be avoided, and it was also important to see if the redesigned visualizations solved their problems better. We believed this phase would useful for visualization experts, because we got additional inputs which were not explicitly revealed in the previous phases. However, it is worth noting here that we follow a descriptive approach rather than a prescriptive one [52] and the final decision to judge the merit of a solution is left to the scientists.

For selecting images for our solution redesign, it was necessary to select a sampling of cases where scientists disagreed with the problem, or they agreed with the problem but were unaware of the solution. This would potentially demonstrate the effectiveness of optimal visualization designs to the climate scientists, if they found the solutions to be beneficial. As described in Section 2.5.3, there was a high level of agreement for problems that led to misinterpretation, inaccuracy, and lack of emphasis. Therefore, we selected examples for which the problems mainly led to lack of expressiveness and inefficiency, given by Table 2.1. We also selected examples where there was a high degree of agreement about the problem, but they were unaware of the solution, like, the problem due to comparison complexity caused by superposition overload and lack of explicit encoding.

To get feedback on the solutions, for each of the images, we specifically asked them if the redesigned solution conveyed the original intent better, and also if there were additional information they could gather which was not possible in the original visualization.

2.6.1 Scatterplot

Intent(s). The context of analysis here was comparison of models with respect to their output variables and respective ecoregions (Figure 2.1(a)). The primary intent here was to look at the spread of the prognostic and diagnostic models with respect to the different output variables, and a secondary task was to identify, for each variable, which models belonged to which ecoregion.

Design Problems and Consequences. The design problems were chart mismatch, choice of visual variables, granularity, and lack of grid lines. Chart mismatch happens because for the scatter plot to convey the first intent, readers have to mentally compute the spread, which is avoidable with a different representation; and more seriously, the second intent is very difficult to convey on a scatter plot due to over plotting. The granularity problem is caused by plotting every data point in a scatter plot, whereas the intent was to look at the spread of the models. Visual variables with different colors, shapes and orientation cause users to perform an inefficient, conjunctive visual search for the models. Absence of grid lines leads to a lack of emphasis of where the users should focus their attention: the chart should be read column-wise, which can be emphasized by use of grid lines. Grid lines were used by the climate scientists in two other examples, but as we observed from Figure 2.9, most of them did not agree that adding grids could be beneficial in several examples.

Solution. A box plot is a more appropriate solution for conveying the first intent, *i.e.*, allowing the users to readily understand the different patterns of spread. This is shown in Figure 2.10. Since there are only a few outliers, we label them directly on the plot. For showing the membership of models in an ecoregion, we use a tabular representation of the models in the form of a heat map, eliminating the need of additional visual variables that might lead to confusion. The heat map is basically a

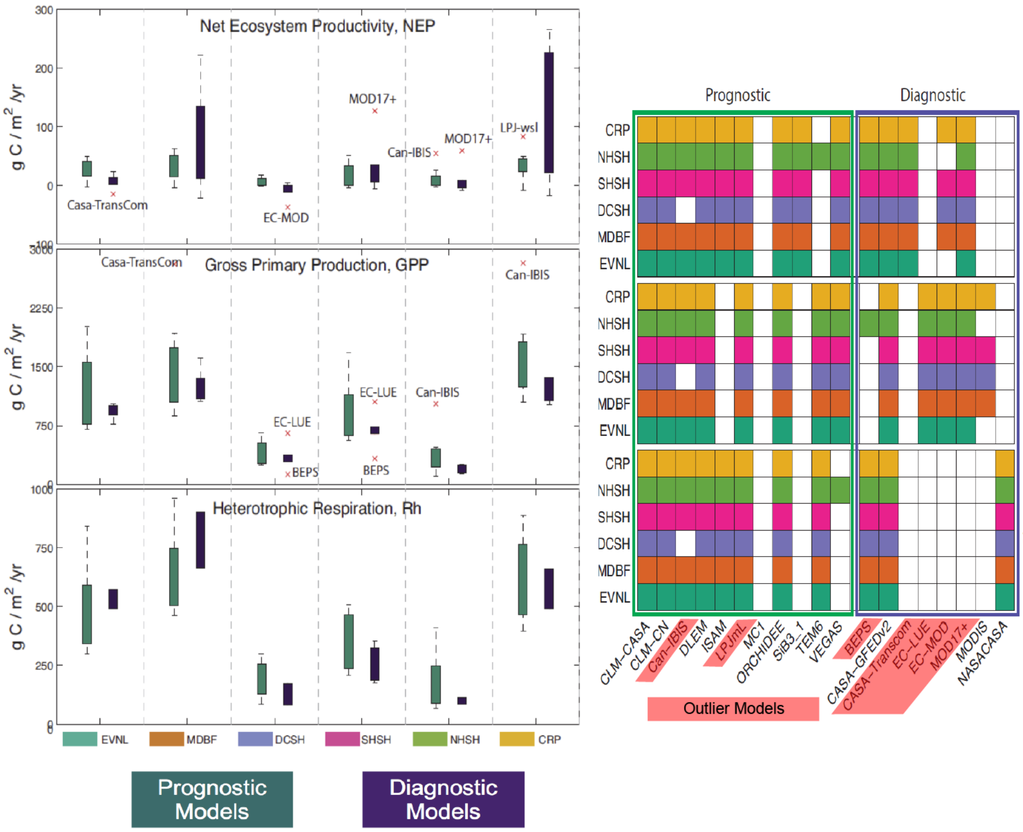


Figure 2.10: **Solution redesign for improving the scatter plot shown in Figure 2.1(a)**. A box plot conveys the intent of showing spread among the output variables of prognostic and diagnostic models, and the accompanying heatmap alleviates the problem of having to visually search multiple symbols for knowing which models belong to which ecoregions.

presence-absence chart where a cell is colored if a model belongs to an ecoregion, and left white, if the model is absent for that ecoregion.

Using the box plot, one can immediately recognize the much higher than average spread for diagnostic models, for the `CRP` ecoregion for the `NEP` variable. One can also compare across different variables and models; for example the less average spread for both classes of models for `GPP` and `NEP` variable, where there are a few outlier models. This trend is however absent for the `Rh` variable. From the heatmap one can also immediately detect that most of the outlier models are diagnostic models, and find that the `Rh` variable is contained by much less number of diagnostic models as opposed to `NEP` and `GPP`.

Scientists' Feedback. Initially scientists were not convinced that the box plots are an improvement as they thought the scatter plot showed more information, like the ability to spot the models directly on the visualization. However when we added the heat map, showing their ecoregion membership much more efficiently, they were convinced that the combination of box-plot and heat map eliminated the design problems. This is evident from the following comment by one of the scientists:

> *"Initially, I was inclined to reply that I liked seeing the model scatter on the first graph,* i.e., *I can think that I can see skew, bi-modality, etc from the scatter plot and if we just slightly offset the points horizontally and with only* 20 *points, then the overlap would not be too bad and I could glean more information. But upon examining the plot, well, you convinced me. What you provided does as good of a job as what I had imagined I would have preferred. In particular, showing the box and stem AND the outliers gave a good bit of information, as did the heat map."*

They were also convinced about the utility of the dual view:

> *"one can immediately detect that most of the outlier models are diagnostic models. This was very difficult to achieve using the original scatter plots. One can also see that most of those models are for* `NEP` *and* `GPP` *and those are not present for the* `Rh` *variable."*

2.6.2 Map

Intent(s). The intent here was to identify similarities and dissimilarities among models for summer months during the period 2000-2005, based on the spatial distribution of the `NEP` variable.

Design Problems and Consequences. The design problems was mainly a lack of explicit encoding as the positioning of the maps do not represent the degree of similarity among the models. Thus, the scientists have to sequentially search and compare models to get insight into their relative similarity. It is thus hard to find pairwise similarity between maps and find groups of similar maps.

Solution. We aimed to improve the visualization by deriving a summary statistic about similarity that climate scientists use, and manipulate the layout for encoding similarity. In discussions with the climate scientists, we used statistical information about the models, that is, root mean squared difference (RMSD) which is widely used in the climate science domain. Using the pairwise computation of RMSD between models, we applied multidimensional projection for displaying the maps in a two-dimensional space (Figure 2.11) using the ISOMAP [53] technique. RMSD was used as the distance function and in the two-dimensional space the proximity of the models denote similarity.

For representing the maps directly based on the projection view, we adjusted the layout using an optimization algorithm [54], so that maps did not overlap and spatial information was retained. Displaying the maps directly was important as the spatial extents of the models were different and the scientists wanted to see them on the geographical map. The projection view shows clear patterns. Note the point representing the `MC1` model is far away from the rest of points, it means that its corresponding map is very different than the others. Another example are the maps `SibB3` and `VEGAS`, their points are near meaning that they are similar (confirmed by the looking at the maps). Another similar group is formed by `CLM-CASA` and `ORCHIDEE` maps.

Scientists' Feedback. There was consensus among the scientists that the resultant figure not only conveyed the original intent but also showed additional information, like quantifying the degree of similarity or dissimilarity of the models based on a metric they were familiar with. They observed that this is a new visualization approach than what they are used to, and one of them expressed caution about the abstraction being used:

> *"I have to be cautious about the MDS method used. I agree that placing maps in different locations will be beneficial. But the MDS method is only one way to represent the similarity among those maps from one certain specific aspect."*

However, they were convinced about the utility of the approach and its benefit in expressing model similarity:

> *"Shows the outliers, and their degree of outlying, more clearly than the original. This is a great solution to a very commonplace visualization in climate modeling."*

2.6.3 Line Chart

Intent(s). The intent here was to compare the temporal variability of multiple models with respect to each other and also with respect to the ensemble mean.

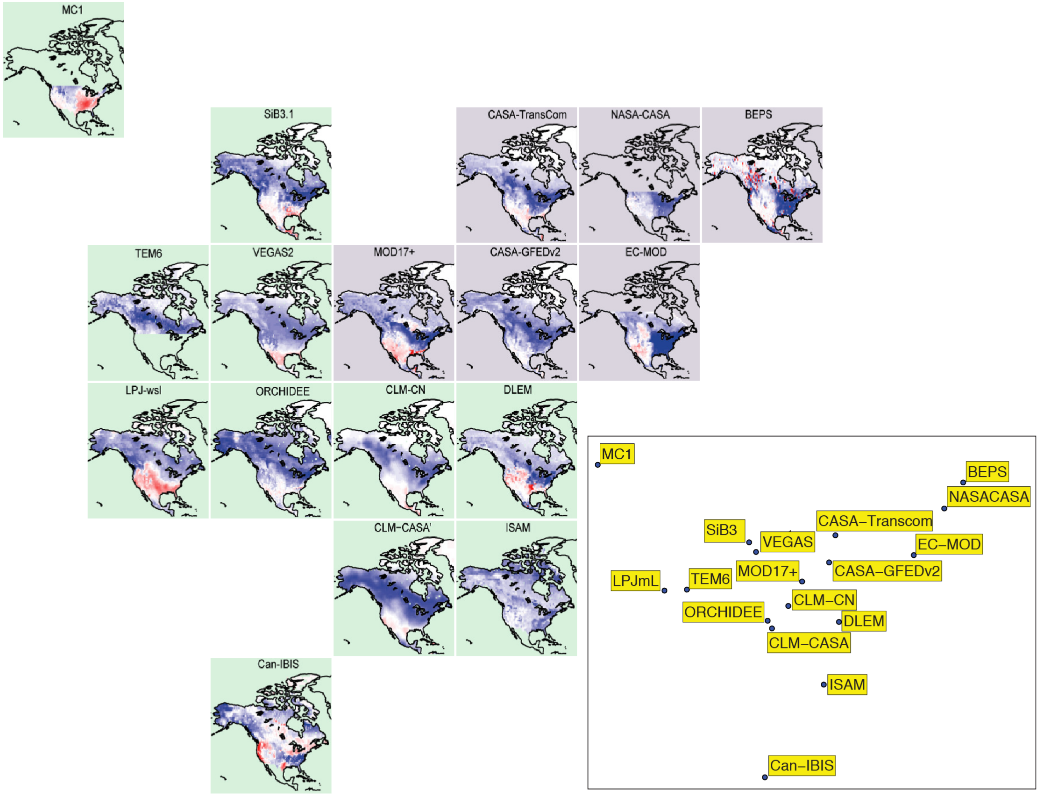


Figure 2.11: **Solution redesign for the multiple maps in Figure 2.1(b)**. Explicit encoding of similarity through multidimensional projection, where higher proximity signifies greater similarity among the models. The layout of the maps is based on the projection shown at the inset. Subsequent layout optimization for ensuring no overlap enables efficient comparison among multiple models.

Design Problems and Consequences. The problems with the spaghetti plot [55] as shown in Figure 2.12 were comparison complexity due to superposition overload, level-of-detail due to jaggedness of lines; and clutter due to overlap. Superposition overload and overlap led to an inefficient comparison of the temporal patterns. Jaggedness was caused by plotting of monthly data and this lead to a lack of expressiveness as the salient annual peaks and crests were occluded.

Solution. We aimed to solve this problem by converting the large single [49] or the spaghetti plot, into a series of small multiples. As shown in Figure 2.13, we converted the individual lines into a band for showing the range of variation, and plotted the ensemble mean in each of the plots, shown by the black line. Each line plot now belongs to a model, and the line is highlighted in red. We avoid using different colors for each model, as the labels are sufficient for identification of a model. We found the small multiple approach being used the scientists mostly in case of maps, and in one

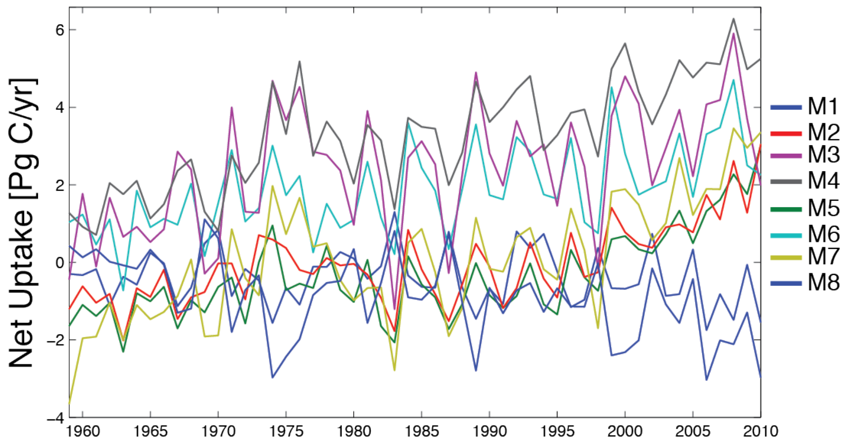


Figure 2.12: **Original representation of the spaghetti plot** for comparing temporal variation of net uptake of 8 models, indicated by the colored lines.

more example, where maps and line charts were used for linking the spatiotemporal trends. One of the problems we did not address was the jaggedness of lines. Since we are losing resolution by using small multiples, one option was to compute an average by combining several years, and smooth out the time series. However due to information loss, scientists were not comfortable with the idea of smoothing by computation of average.

Scientists' Feedback. The scientists unanimously felt that the resulting small multiple display overcomes the problems that are traditionally present with a spaghetti plot. One of them commented:

> *"The new plots are definitely better than the original one. It's difficult to identify each model line in the original plot due to over-plotting."*

They appreciated the minimalist design by using few colors and also the fact that temporal variation could be compared quickly and intuitively with both the multi-model range, and the lines representing different models. One of the scientists also observed that:

> *"If the goal is to visualize model similarity then we can apply the same layout optimization as applied to the multiple maps example to rearrange similarly behaving models together."*

They were confident that this would be an exemplary visualization which will be emulated in model comparison scenarios and preferred over the traditional spaghetti plot.

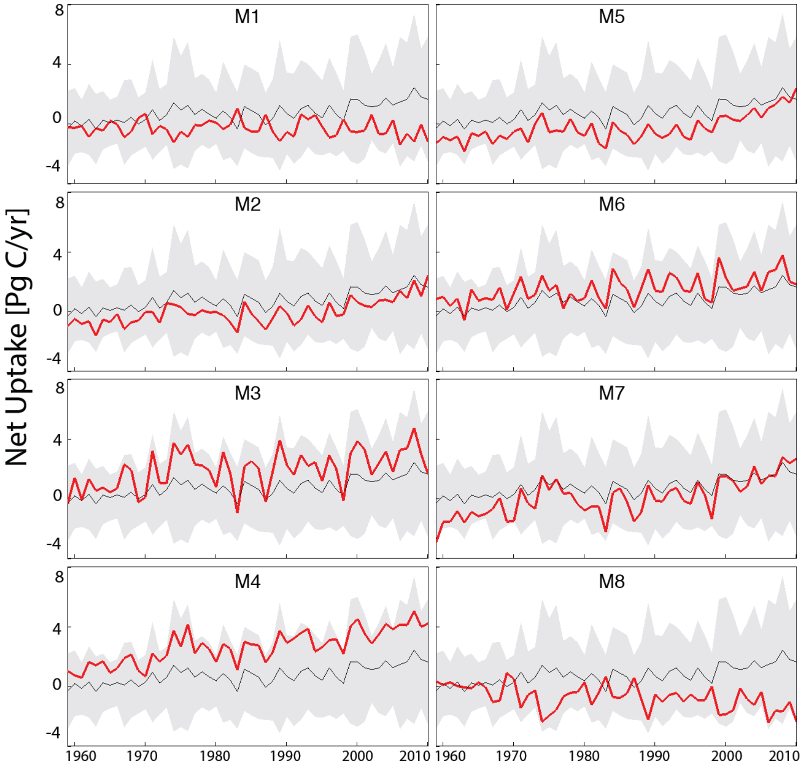


Figure 2.13: **Solution redesign for improving the spaghetti plot.** Separating the model representations into small multiples of line charts enables efficient comparison of each model trend with the ensemble mean and range of variance.

2.7 Design Problem Trade-Offs

In this section we present one of the the key findings of our study, which is a reflection on the trade-offs among the different problem consequences. Many design problems and consequences cannot be simultaneously avoided. An awareness of the trade-offs is necessary for the scientists to judge how best to configure a visualization. The first decision that scientists have to take, is to weigh which design consequence more, and accordingly decide which potential design problems to avoid the most. A perfect visualization is hard to achieve and there is no one-size-fits-all formula for generating one. Visualization design is heavily parameterized by the scientists' intent, which needs to take into account the different trade-offs. Following were the different trade-offs we found in our analysis.

2.7.1 Lack of Expressiveness vs Inaccuracy

This trade-off was observed in cases where scientists' intent for visualizing the data at the finest level of detail, led to a lack of expressiveness of the salient patterns. For example, in case of jagged lines, an average could be computed to reduce the number of steps, and that would lead to higher expressiveness, but at the cost of inaccuracy. Another similar example of this trade-off was choice of visual variable and granularity problem in maps due to pixel-by-pixel representation on maps. As acknowledged by some of the scientists, a coarser representation would have lead to better expressiveness of the data.

2.7.2 Inefficiency vs Inaccuracy

This trade-off was observed in cases where an accurate representation was achieved at the cost of an efficient one. Superposition of multiple lines and points for comparison with observation data is a common practice with climate scientists. While in many cases superposition facilitates accurate multiway comparison, fulfilling the expressiveness criteria, in some cases this also leads to clutter leading to inefficiency. Especially during in publications and broader dissemination, these criteria are important. In these cases, small multiples and use of explicit encoding of relationships should be considered.

2.7.3 Lack of Emphasis vs Inefficiency

This trade-off was observed in cases where scientists' intent of keeping charts free of clutter, for achieving more efficiency, came at the cost of a lack of emphasis. Charts should be self-contained by use of proper labelling, grids and annotations if necessary, which help emphasize the intended message. Improper use of these auxiliary information however can clutter charts and make the decoding process inefficient. This trade-off is also echoed my Few's mantra of minimizing non-data ink [10].

2.8 Guidelines for Avoiding Design Problems

The design problems and consequences enable a visualization expert to reflect on design-trade-offs and formulate solutions based on the intents that consider those trade-offs. Climate scientists are not familiar with all visualization best practices, which we demonstrated in Section 2.5.3. It was thus necessary to abstract the problems, consequences and their solutions in the form of guidelines, that are more comprehensible from a scientist's perspective. Our objective was to distill the general problem trends and provide guidelines that can enable climate scientists to avoid those problems. The following guidelines should be understood in the context of the classes of visualizations we collected, which were maps, scatter plots, and line charts; and the scientific intents, which mainly centered around understanding and expressing similarity of climate models in those visualizations. The following guidelines are discussed with the context of the related design problems, so that scientists are able to bridge the gap between the design recommendations and current practices, and embrace the best practices in visualization.

2.8.1 Keep Audience in Mind

A recurring issue cutting across different design problems was the tendency of scientists to use the visualizations designed for their own analysis, for publication and dissemination of their results as well. This was triggered by an implicit assumption about the familiarity of the audience with what to look for in the data. On the other hand, to cater to a broad audience, whether internal or external to the climate science community, the visualization itself should be expressive enough to convey the intent, without overwhelming the audience with the details.

In those cases, rather than representing all the data (level-of-detail problem), it is more important to show and highlight the trends by abstracting or aggregating some of the data. As we had pointed out in Section 2.5.3, the problem of loss aversion was a leading cause of problems in maps and line charts. There were many cases where the old adage of 'less is more' held true for the visualization designs. For example, if a line chart is too jagged because all time-steps are represented, it can obscure the message. In cases where the intent is to visually express similarity of multiple models, scientists can choose to represent the data at a coarser granularity by choosing a visual variable (choice of visual variable problem) other than color, like orientation of lines or glyphs which have been successfully adopted in the geographical data visualization domain [37].

2.8.2 Guide Users Attention to Salient Visual Objects

A critical requirement of any visualization design is to explicitly guide user's attention to the salient patterns. Enabling visual comparison of similar and dissimilar models was the underlying intent of the images we collected. To facilitate such comparison, key elements of the visual representations should serve as indexes for visual search for finding models that are similar or dissimilar.

We did not find effective use of Gestalt laws of grouping, which can be an effective visual cue in these cases. This led to the comparison complexity problem (Figure 2.7). For example, as we had shown in Figure 2.11 absence of explicit encoding the message about similar or dissimilar models is not fully expressed. Another examples is the problem of superposition overload (Figure 2.12). This causes clutter which can either disinterest the audience or cause trouble in finding the patterns. Scientists should avoid relying on the audience's mental operations to make those visual comparisons, which can be both inefficient and ineffective in absence of any visual cue. As we had elaborated in Section 2.4.3, explicit encoding of relationships and emphasis of the key message can alleviate these problems.

2.8.3 Focus on the Message and Make it Self-Contained

In many cases, the scientists' intents were not fully conveyed as the message of a chart was incomplete due to either lack of emphasis of the take-away message or a lack of synergy between the auxiliary information and visual representation. In complex visualizations, it is often necessary not just showing the data, but also explaining what the visualization conveys through highlights and texts. To make the message clear, scientists can use size, color or orientation, that is substantially different from that of the other objects in the visualization, than the one which is most important. This is especially true of outlier objects.

The design problems related to this category were mostly those associated with a lack of emphasis. Although auxiliary information about charts only help when a chart is effective in the first place, they can help focus human attention very quickly to the salient portions of a chart. In cases of complex charts with multiple messages, this aids the user in decoding the intended message very efficiently. Charts should also be self-contained, without the audience having to search for the meaning of the legends in some other table or graph, which was true of some examples we collected from the research papers.

2.8.4 Tie Color Selection with Data Semantics

We observed in Figure 2.9 that choice of color was one of the problem categories where there was a lot of disagreement between visualization experts and climate scientists. We found that our collaborators generally considered the use of color as more of an aesthetic issue than it being tied to the data semantics. While there were some good examples of choice of color maps, in majority of the cases we found that the choice of color map was not appropriate.

Apart from color maps, we also found some other inappropriate choices of colors. For example, in some cases, we found the most important visual object, such as the mean or the trend line being encoded in gray, in which case it would be hidden in clutter and not be emphasized. Apart from different color scales for different data types, we also recommend the effective use of color in emphasizing certain objects (*e.g.*, red can create a pop-out effect) or muting certain aspects of the data, like using gray to de-emphasize points in a scatter plot that create noise and use color only to encode certain salient points.

2.8.5 Be Mindful of Defaults

For several design problems we had assessed, one of the precursors for the problems was the defaults of the tool that the scientists were using for creating those visualizations. One of the infamous defaults in many tools is the rainbow color map. The other one is the selection of random symbols for showing discrete data. In our taxonomy, these led to the color map choice and visual variable problems (Figure 2.5). The consequences of these problems can be as severe as misinterpretation, or lead to lack of emphasis for salient patterns (Table 2.1). In these cases, it is necessary for the scientists to look beyond the defaults and introspect if the defaults impede data analysis and visual communication. Such introspection might ultimately require scientists to manually configure visualizations for overcoming the problems with the default settings.

2.9 Scope and Impact

In this section we discuss the scope of our work and the impact in terms of the generalizability and utility of mapping design principles to domain-specific, static visualization designs.

2.9.1 Limitations

Our work has some important limitations to take into account; first of all is its subjective nature due to the qualitative methodology, While use of grounded theory and bottom-up approach to building visualization usage models are gaining ground [23, 29], it is an acknowledged fact that subsequent research needs to be done to develop prescriptive solutions to the problems. Accordingly, our focus in this work has been to reflect on the design problems through the descriptive taxonomy, which can be expanded in scope through further research to build prescriptive, broadly applicable solutions. Moreover, there are multiple ways of describing the design problems we found. We are not claiming that this is the only way to classify visualization design problems we found. However, we are confident that through our collaboration with a broad group of climate scientists and our understanding of the state-of-the-art in visualization practices, our classification provides a good starting point for bridging the gap between visualization best practices and existing climate data visualizations. The guidelines should be understood in the context of the sample of images we collected. These guidelines still need to be validated through empirical evaluations.

2.9.2 Generalizability

Although the sample of visualizations we collected was limited by their type (maps, line charts, and scatter plots) for us to build prescriptive solutions, we believe many aspects of our study are generalizable. First, although we used only three types of visualizations, they represent a broad set of usage scenarios in climate science: understanding spatial patterns, temporal patterns and looking at bivariate relationships among variables. The tasks mainly involved visual comparison of distributions, correlations, and variability, which are common analysis tasks cutting across climate science and even other domains. From that perspective, we are confident that the problem classification will be applicable to different domains and usage scenarios.

Second, the problem classification itself follows a mapping between general design principles and visualization examples from a domain. Even if some of the problems we found in the climate science domain do not exist in other domains, the same principles and classification scheme based on encoding and decoding *problem stages* would still apply. The same would apply for the *problem type* level, only the causes of the problems might be different. For example, there can be different causes for a level of detail problem, or the problem or clutter or distortion, but these problem types are still applicable for judging the quality of visualizations. As mentioned before, to the best of our knowledge, our work is a first step towards bridging the gap between general design principles and how they are realized in practice.

Third, from a visualization perspective, some problems we found are symptomatic of general gaps in research involving static visualizations. First, while much research has focus on judging effectiveness of interactive visualizations, many mediums such as publications and presentations are constrained by their static nature. We found that representing multivariate or multi-model relationships and effectively visually communicating their relationships have non-trivial challenges. Second, use of bad defaults has been widely talked about, but rarely addressed in the tools available today, with a few exceptions, like the Paraview tool where the rainbow color map was changed to the perceptually effective divergent color map [56]. Our findings should encourage such changes in the visualization tools, which will ensure better designs by domain experts.

2.9.3 Utility

A large body of research focuses on interactive visualization and it is a general assumption that good interactive visualization design can be easily and directly turned into good visual presentation design. But our work points to the fact there are different challenges and gaps and we need to better understand and research this difference. Visual presentation is not just taking pictures from our interactive tools and placing them into our papers and presentations. The design has to tell a compelling story about the findings of the scientists to the non-technical stakeholders, and in visualization, the presentation and story-telling aspect has received much less attention till date [57]. Some well-defined best practices like harmfulness of rainbow color maps [43] need more empirical validation, especially in the science community [44] for establishing the objective reasons behind recommendation of perceptually motivated color maps. A survey of existing visualization tools, investigating the quality of the defaults, will be helpful in identifying these issues [58] and will enable visualization non-experts like domain scientists design visualizations more efficiently.

2.10 Summary

In this chapter, we have presented a comprehensive study of visualizations designed by climate scientists and classified their shortcomings by categorizing the causes and consequences of design problems in the form of a taxonomy. In the process, we have investigated the cross-domain agreement and disagreement about design problems and highlighted their reasons. Further, we have demonstrated the utility of our taxonomy by getting feedback on redesigned solutions, which the scientists found to be beneficial for their practical use.

During our interactions with the scientists we found multiple mismatches between visualization best practices and the state-of-the-art in the climate science domain. For instance, we noticed that the *rainbow* colormap, despite ample research advocating against its use, is considered the de-facto standard for encoding scalars on geographical maps. To study this phenomenon we set up a user study which will be explained in the next chapter.

Chapter 3

Perceptual Evaluation of Color Scales for Climate Model Comparison

Color-coded geographical maps are an integral part of geospatial data analysis. In many domains, different color scales are used to represent the magnitude of continuous variables (*e.g.*,, amount of rainfall, population density) on these maps.

Based on the study presented on Chapter 2, we found that the rainbow color scale is de-facto standard for geospatial data analysis tasks related to climate model comparison. However, ample research evidence [42, 59, 43] has demonstrated the perceptual limitations of the rainbow color scale for making quantitative judgments. Given this mismatch between the visualization best practices and the state-of-the-art in the climate science domain, the goal on this work was to compare the analytical effectiveness of rainbow color scales with that of perceptually-guided color scales on climate data analysis tasks.

Our study addresses certain gaps in the field. First, we observed that perceptual experiments that tested theoretical hypotheses usually involved artificial laboratory tasks and stimuli, and were conducted on untrained undergraduate students or colleagues. Studies that focused on tasks tended to select tasks that were artificial distillations of real-world analyses. And, even in studies where real-world practitioners performed real-world tasks, the task was domain specific, and not easy to generalize to other analysis environments. To address these limitations, we designed a study using real-world climate-model data, analyzed by practicing climatologists, performing tasks that not only captured their everyday analysis problems, but could be used to generalize to other situations involving the analysis of spatial data.

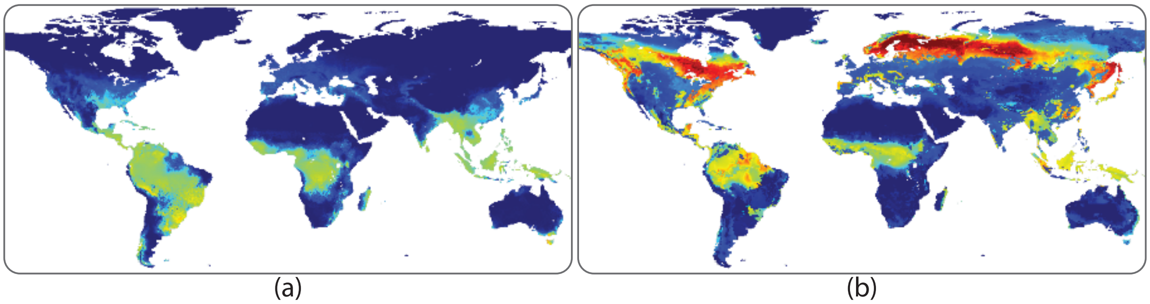


Figure 3.1: **Pairwise comparison of color-coded geographical maps representing climate model outputs encoded with the rainbow color scale** Climate scientists use geographical maps for visually comparing the *magnitude* (*e.g.*, average temperature) and the continuous *spatial distributions* of model output variables.

To approach these problems, we developed a long-term collaboration with climate scientists, and used this collaboration to characterize their tasks. Typically, this involves comparing pairs of color-coded geographical maps, each containing the output of a different model or time step (Figure 3.1). The climate scientists study these pairs, to understand the underlying reasons for agreement or disagreement among the models. Our first contribution was to characterize their work into three tasks: (i) making quantitative visual estimations about the differences in overall magnitude, (ii) estimating the spatial variation of model outputs, and (iii) identifying regions of maximal magnitude difference. Once we had identified a set of representative tasks, our second contribution was to evaluate performance of scientists on these tasks using different color scales (Figure 3.2). We compared performance on the rainbow color scale and systematically selected two other color scales which had been shown to have perceptual advantages over rainbow color scales. To create a set of stimuli that would allow us to generalize our results to results to other domains where similar tasks are performed using color-coded geographical maps. Since much of the scientific evidence demonstrating drawbacks of the rainbow color scale stems from experiments using artificial stimuli, we generated experimental stimuli that controlled the spatial and magnitude characteristics of the real-world geospatial maps. We used these stimuli to measure the performance of climate scientists in the the three different tasks.

At the end of the study we also conducted a survey, where we asked the study participants to record their preference, confidence level, and perceived accuracy of their judgments using the different color scales. Our third contribution is a comparison of the scientists' quantitative performance against their perceived performances and preferences. This is especially important as the use of rainbow color scale is nearly ubiquitous in climate modeling. Differences between scientists' perceived and actual performance could help create greater awareness about perceptual principles of visualization, which could potentially lead to the adoption of visualization best practices in real-world tasks such as those reported in this study.

3.1 Related Work

We discuss the related work with respect to three main threads of research which are relevant in the context of our work: i) theoretical principles of color perception in visualization; ii) color scales for different visualization tasks and (ii) the empirical evaluation of color scales for synthetic and real-world tasks. Many of these principles are reviewed by Silva et al. [60].

3.1.1 Theoretical principles of color perception

Human color perception is three-dimensional, which means that any color a human can see can be represented by three independent dimensions. A color scale can be represented as a trajectory in a 3-dimensional color space. The Rainbow color scale (*RBW*), for example, is a linear interpolation from (0,0,255) to (255,0,0) in an RGB color space. Several researchers have explored different trajectories in different color spaces in an effort to identify effective color scales, taking advantage of knowledge about human color perception. For example, Robertson [61] explored the idea of creating color scales that spanned the greatest range of discrimination steps in a 3-D color space. Important advances have been made by focusing on color scales that are based on the three perceptual dimensions of hue, luminance and saturation. Rogowitz et al. [62] pointed to earlier psychophysical scaling work by Stevens [63] which had shown that monotonic variations in luminance and saturation produced monotonic perception of variations in perceived magnitude. Stevens had also found that monotonic differences in hue did not produce monotonic differences in perceived magnitude, which led them to predict that spectral hue-based color scales (like the Rainbow color scale) would not faithfully represent changes in data magnitude. To test these hypotheses, this group [64] conducted psychophysical increment detection experiments using hue, luminance, and saturation scales, constructed in several color spaces. They found that with luminance and saturation-varying color scales, increment detection for Gabor patches produced consistent, and sensitive, increment detection for luminance and saturation color scales, however for the rainbow color scale equal steps along the color scale did not produce equal steps in perceived magnitude.

Other characteristics of human vision are also important in considering the construction of color scales. For example, Rogowitz and Treinish [42, 59] observed that the human luminance system has higher spatial-frequency sensitivity than the opponent color system, suggesting that color scales designed to represent the magnitude of fine resolution detail should contain a monotonic luminance component. Using a novel technique, Rogowitz and Kalvin [65] tested rainbow, luminance, iso-luminant and heated-body color scales, and found that magnitude perception was driven by the luminance component. Based on this research, Kindlmann et al. [66] developed a luminance-matching technique which could be used to create color scales that contained a range of hues, with monotonically varying luminance. Since hue provides categorical information, it was posited that such color scales could both effectively carry magnitude information while also providing segmentation information.

3.1.2 Adapting Color Scales to Different Visualization Tasks

The selection of an appropriate color scale depends on the visualization task. Brewer [67, 45] has proposed three classes of color scales for spatial data representation, with a focus on representing data in choropleth maps for geospatial analysis [68, 69, 37]. The “sequential” color scale has a monotonically varying luminance component, as described in previous section. The “qualitative” color scale, developed for categorical and ordinal data, uses a fixed number of hue steps. The “diverging” scale has one hue component transitioning to another by passing through an unsaturated value in the middle. For scientific data visualization, Moreland [56] has developed a version of the diverging color scheme, which has recently been accepted as the default color scale in ParaView [70]. Bergman et al. [71] introduced a rule-based system that suggested appropriate color scales based on the data type (ordinal, interval, ratio), spatial frequency, and on the task. For “isomorphic” tasks, color scales with a monotonic luminance or saturation are suggested; for segmentation tasks, binned color scales are offered; for highlighting tasks, color scales with highlighted ranges are proposed. Tominski et al. [72] extended these ideas by proposing a task taxonomy and appropriate color scales comparison, localization, and data value identification tasks.

3.1.3 Empirical evaluation of color scales for synthetic and real-world tasks

In order to isolate and study specific experimental variables, many empirical studies in this field have relied on using synthetic stimuli. This involved the detection of Gabor patches [64]. Ware [73] used artificial stimuli to explore users' ability to read magnitude information from a region on a visualization and map it onto a value on a color scale for five different color scales. To emulate real-world medical imaging situations, Tajima et al. [74] and Levkowitz and Herman [75] used the detection of artificial phantom "blobs" in medical images to reveal advantages of the luminance grayscale over other color scales, including the heated-body scale. Recently, Borkin et al. [44] studied visual performance using the rainbow and a heated-body color scale in a real-world setting, with cardiologists. In a task involved identifying arterial blockages, they found a very large and significant advantage of the heated body color scale (monotonic luminance and a small hue variation) over the rainbow color scale.

Although these task-based experiments make significant contributions to our knowledge, most of them were either performed with artificial stimuli or were performed using a very specific domain-based task. In the experiments reported here, we study the performance of working climatologists on three real-world tasks, using three color scales. Unlike previous studies, like the one reported by Borkin et al. [44], where they let participants look a single image and judge regions of highest magnitude, we explore the situation where scientists make judgments based on comparing pairs of color-coded map. We study three different tasks, which are representative of their analysis process: detecting relative magnitude difference, judging similarity, and identifying regions of maximum difference. Also, to increase the generalizability of our results, we develop a method for selecting pairs of maps for comparison that characterizes their differences parametrically. Our comparison stimuli vary in overall magnitude and in spatial similarity. This allows us to extend our conclusions beyond the domain of climate science, to other domains where analysts judge magnitude differences, evaluate spatial similarity, and identify regions of maximal difference between spatial representations of interval data.

3.2 Task Analysis

We interacted with our direct collaborators over a period of six months for collecting examples of maps and their corresponding analysis goals. Through in-person and online semi-structured interviews, we refined and iterated upon our understanding of these goals. We organized several face-to-face meetings (and many remote follow-up meetings) with our collaborators to understand in details what specific questions they ask and judgments they make using color-coded geographical maps. During these meetings we presented examples taken from the scientists' work and asked to describe what kind of questions and visual operations they would perform when looking at them.

Climate Model Comparison Tasks. For comparing model outputs, climate scientists typically produce color-coded geographical maps representing distribution of model outputs (Figure 3.1). Scientists generally use these maps for visually estimating the variation in model outputs and testing their hypotheses. In this study, we focus on an important output variable in climate modeling, which is Gross Primary Productivity (`GPP`). `GPP` is arguably the most important health indicator of the eco-system as it captures the relationship between the carbon cycle and impact of climate change. Scientists generally perform visual comparison tasks through juxtaposition of these maps [48] in a small multiple setting [9]. While there have been previous attempt to build a general model of visualization tasks [76, 77] and apply those models in climate science [6], we look at much narrower analysis scenarios and formulate tasks that take into account specific and relevant scientific intents.

Task Types. The geospatial data analysis goals therefore fall into two main classes: comparing magnitude of the encoded variable (like `GPP`) and comparing spatial distributions. We finally narrowed down the list of tasks to two types of judgments and three relevant task the scientists identified as frequent and important.

***What* and *How Much* Judgments.** We iteratively refined our task set and realized most of their judgments fall into two categories: *what* and *how much* judgments, cross-cutting magnitude and similarity of distributions. For the *what* category, scientists are generally interested in identifying the areas with similar and dissimilar spatial distributions. For the *how much* category, scientists usually make a quantitative judgment of the differences and degree of similarity or dissimilarity between the models. Making quantitative judgments from maps are usually expensive visual operations, since they involve an elementary reading level [35], that is, at a pixel level. Scientists can overcome this difficulty using their experience. Moreover, generally they perform this operation when there is a table depicting the total. In that case they visually estimate the magnitude for building confidence into the numbers and also for looking at the spatial patterns.

Task 1 - Magnitude Estimation. In this task the scientist aims at estimating the difference in terms of global mean `GPP` between a pair of color-coded geographical maps (A and B). In these maps `GPP` is encoded through color so that one can see how `GPP` values are distributed across the globe. The task requires the scientist to look at the distribution of `GPP` values across the map and mentally estimate the mean and the compare this value between two maps.

Task 2 - Similarity Estimation. In this task the scientist aims at comparing the distribution of values across two maps and provide an estimate of how similar they are on a range $[1, 5]$. Note that similarity estimation depends both on the magnitude of values as well their spatial distribution.

Task 3 - Identification of Regions of Maximum Difference. In this task the scientist aims at identifying areas of the two maps where the two models differ considerably. Which in turn translate into identifying corresponding areas of the two maps where the color values differ considerably.

3.3 Choice of Color Scales

In addition to the rainbow color scale we chose two color scales based on our goal correcting the luminance and variation of hues.

3.3.1 Rainbow Scale ($\boldsymbol{RBW}$)

This color scale is shown in Figure 3.2. It is the default colormap in many prominent systems, such as Matlab. It is created by linearly interpolating between (0,0,255) and (255,0,0) in RGB Colorspace. This colormap provides highly-saturated colors, from blue, through cyan, green, yellow, and orange to red. Luminance is not monotonic, so that equal steps in the data are not perceived to be equal perceptual steps, and for this reason, it is not a good candidate for representing magnitude information [44, 43, 42]. However, the luminance of RBW is monotonic at the lower end of its range, so magnitude information would be expected to be successfully carried over this range [65]. Although the scale is colorful, the colors are not perceived as a smooth variation from hue to hue. Instead, we see bands of colors of unequal

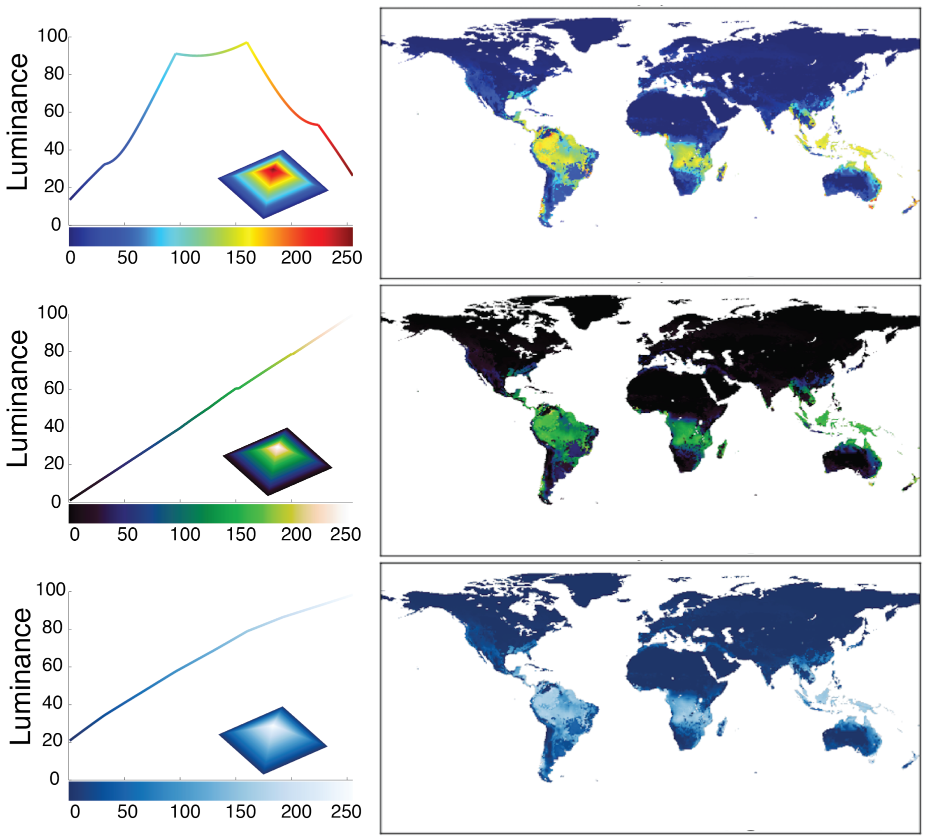


Figure 3.2: **Three different color scales used in our study and the corresponding luminance plots**. The *RBW* (top) has a non-monotically varying luminance which is overcome by the *KIN* (middle) and the *BLU* scales (bottom). We describe the selection of these color scales in Section 3.3.

sizes. Magnitude variations within a color band will not be easily discriminated [42], and since the bands are of unequal sizes, these regions of low discriminability will be unevenly spaced over the data range.

3.3.2 Blue Scale (*BLU*)

If accurate quantitative information has to be extracted out of color, a popular alternative choice is a scale that maps data values to the luminance parameter without changing other parameters. The *BLU* (Figure 3.2) is a popular selection from the Brewer Library [45], and has its roots in geographical map design. It is one hue (blue), and its luminance increases monotonically over its whole range. One problem of single-hue color scales however is that the single hue representation makes it hard to segment, and thus label, areas of uniform color intensity. For instance, while in a map that uses the *RBW* it is possible to identify the red or green or blue areas as areas with uniform or similar values, the same is not possible with single-hue color scale. From our interactions we learn that climate scientists prefer the *RBW* as it is easier to name the different regions based on the different hues.

3.3.3 Kindlmann Scale (*KIN*)

One solution to this problem which does not compromise on perceptual orderability of the values in the color scale is to use a multi-hue color scale that also has a uniform increase of luminance. The *KIN* (Figure 3.2) is one such color scale which was suggested by Kindlmann [66] as an alternative to the *RBW*. It uses vibrant, saturated colors, while also providing a monotonic luminance. The scale runs from dark violet, through blue, to green, to yellow to white. Since the hue variations cover a small range of hues, with monotonic increases in luminance, the color banding effect present in the *RBW* is eliminated. It has also been posited that color scales that add a hue component to the luminance variation would provide additional discrimination steps, owing to the hue variations [78, 65] .

3.3.4 Other Considerations

We also decided *not* to modify the color scales we selected. Based on perceptual principles, it would be possible to modify these scales to potentially achieve better magnitude and similarity judgments. For example, attempts have been made to modify the *RBW* to reduce the color banding. The *KIN* could be modified to explicitly vary the saturation along the scale, and the *BLU* could start at lower luminance levels to increase its dynamic range. We also decided not transform the data from its original form that is used by climate modelers. In particular, the data values for `GPP` are densely distributed at the low end, and a logarithmic transform would help spread the data values more evenly over their range. However, modelers mostly visualize the data in its original scale, and in a comparative setting, they do not prefer transforming the data.

We also considered testing a divergent color scale [56]: a scale in which the mean value in the color scale is a neutral color, and hue increases in saturation, and often decreases in luminance, as the values move to the extrema. We did not test this color scale because the data we are representing are interval data, ranging monotonically from the low end of the scale to the highest values. The climate scientists we worked with felt that the divergent color scale would imply that there was a neutral "zero" with variations above and below this imputed "mean", which did not match the structure of the data.

3.4 Hypotheses Generation

The numerous interactions we had with our collaborators and the subsequent task analysis process allowed us to generate a number of hypotheses that we tested in our experiments. For mapping the performance on these tasks using the color scales, our first aim was to detect if color scales have any effect on the three tasks at all.

Based on the established principles, we evaluate if luminance correction of the rainbow scale will affect the performance. Luminance is not monotonic, so equal steps in the data are not perceived to be equal perceptual steps, and for this reason, it is not a good candidate for representing magnitude information [44, 43, 42]. However, the luminance of the *RBW* is monotonic at the lower end of its range, so magnitude information would be expected to be successfully carried over this range [65]. Luminance monotonicity is critical for representing equal steps in the data as equal perceptual steps, so we expect the *KIN* and *BLU* to afford better judgments than the *RBW* in general for both tasks. The *KIN* has the highest luminance dynamic range, which we expect will provide the user with more discrimination steps, and, combined with luminance monotonicity, will provide higher accuracy for fine spatial variations.

For the judgment about spatial distributions, we hypothesized the color scales which provide higher discriminability by helping in the segmentation of the map, will be beneficial. Although the rainbow scale is colorful, the colors are not perceived as a smooth variation from hue to hue. Instead, we see bands of colors of unequal sizes. Magnitude variations within a color band will not be easily discriminated [42], and since the bands are of unequal sizes, these regions of low discriminability will be unevenly spaced over the data range.

In **judgment of magnitude**, if the hypothesis about hue variations providing additional information is correct [73], we expect the *KIN* to lead to better performance than the *BLU*, which provides no hue variations. The *RBW* also has a high dynamic range, but it is not monotonic therefore we expect it to perform worse than both scales.

As for **judgment about spatial distributions**, since the *RBW* does not provide smooth transitions in perceived hue across the range, we predict that this will limit the observers' ability to make magnitude discriminations within ranges of equal perceived hue. We predict that the color banding will have particularly deleterious effects in the similarity judgment, where the observer is comparing spatial structures across models.

The color banding may create apparent regions of similarity which may not be present in the data. We also predict the *BLU* will perform worse than the *KIN* in this task due the lack of hues, which can clearly segment areas of uniform values and also due to the lower dynamic range.

As mentioned before we were also interested in comparing objective performance to subjective assessment, therefore here we also hypothesize that despite the *RBW* will perform worse than the *BLU* and *KIN* in all tasks, it will be perceived, after using all color scales multiple times in the study, as more effective and accurate by our group of collaborators.

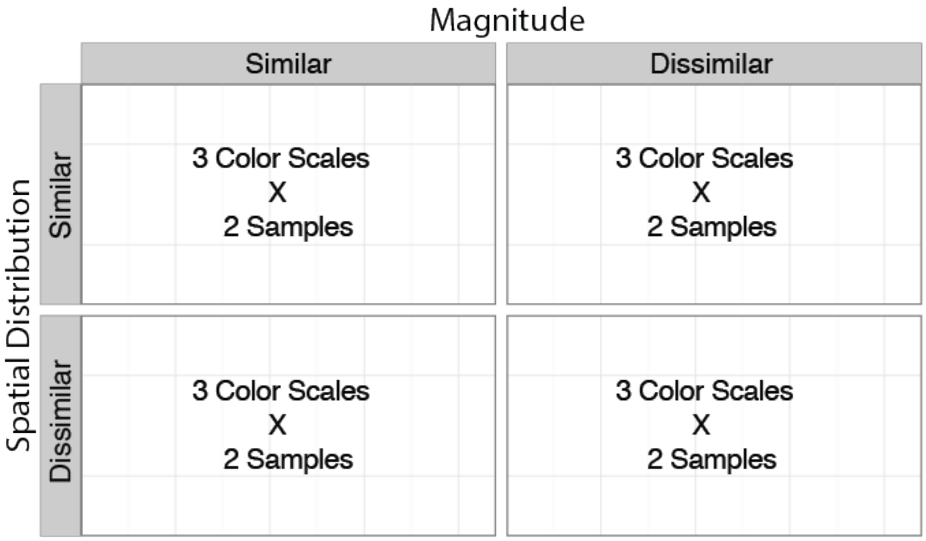
3.5 Methodology

In this section we present the details of our methodology for the study, namely, data generation, the choice of stimuli, participants, settings, and evaluation metrics.

3.5.1 Data Generation

Based on our general hypotheses, we aimed to design the stimuli based on two main principles: (1) testing variabilities in the data that may have an effect on performance and (2) using the real data our collaborators use in their research and are familiar with to make the task as realistic as possible. Since all our tasks are comparisons we aimed at generating pairs of maps that differ along two main dimensions: *magnitude* and *spatial distribution*. In our data generation process, we ensured that there is sufficient variability and coverage across the possible conditions.

Ensuring Variability and Coverage. Our maps pairs are generated using the `GPP` variable from 6 models (`BIOME, GTEC, SIB3, CLM, CLM4VIC, LPJ`). Each model has a spatial resolution of 360 × 720 and monthly temporal resolution of 360 time steps (20 years). The greatest variability in the model outputs is generally found across different seasons be it a same or different year. However we did not want to pick and choose the data from seasons of a particular year, as some events might affect that `GPP` for a region in a particular year, and we would not be able to account for that. Instead, to ensure variability we selected 10 random time steps for each model and compare against all the time steps of all the other models. Thus, we have in total 108000 pairs (6 models × 5 models × 10 random time steps × 360 time steps). This not only ensured variability in the data, but also a coverage of the data points.



		Magnitude	
		Similar	Dissimilar
Spatial Distribution	Similar	3 Color Scales X 2 Samples	3 Color Scales X 2 Samples
	Dissimilar	3 Color Scales X 2 Samples	3 Color Scales X 2 Samples

Figure 3.3: **Selection of trials based on data bins**: For each bin in the data we randomly chose 2 sample map pairs which were represented using 3 different color scales. For each of the four bins we show examples of map pairs in Figure 3.4.

Eventually the map pairs for our study were selected from these pairs, based on our definition of stimuli as described below.

Controlling the Parameters. The parameters for representing the model output using geographical maps are the following: data range, projection type, weights to different areas, etc. In many cases, depending on the data range, analysts would choose a non-linear (like a log transform) mapping between the data and the visual variable, which is color in this case. However, in course of our discussion with the scientists, we found that transformation of the data in terms of aggregation or using a different scale is not something they prefer during their analysis. There is obvious information loss (in case of aggregation) or different representation of the data (in case of log transform), which they want to avoid. Therefore, we assume a linear mapping between the data and the color scale. In the tasks we selected, since they were performed in a comparative setting, projection errors would not affect the results.

3.5.2 Selection of Stimuli

As shown in Figure 3.3 and illustrated later in Figure 3.4, we selected stimuli for the experiment by grouping pairs of maps into four bins according to the scheme: low/high difference in magnitude and low/high difference in spatial distribution. For instance, two maps can have a similar distribution of values across the maps but

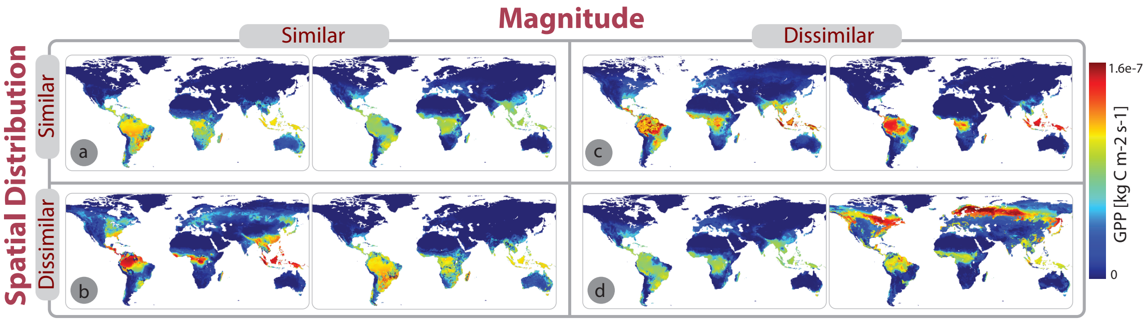


Figure 3.4: **Examples of map pairs generated based on similar and dissimilar magnitude, and similar and dissimilar spatial distribution**. In our study we controlled for these two factors and aimed at finding how these variations in the data affect visual comparison tasks of climate scientists using different color scales.

different overall magnitude. As show in Figure 3.4, it is possible to have maps with similar spatial distribution but different magnitude (top right) as well as maps with different spatial distributions but similar magnitude (bottom left). It is important to notice that while these differences may seem hard to understand by a non-expert, climate scientists are highly trained to derive this information from the color-coded maps.

In order to automatically generate map pairs that fall into the four groups outlined above, we had to devise metrics that capture the amount of difference between two maps in terms of magnitude and spatial distribution. For this purpose, we asked the scientists to assist us with this problem and subsequently derived two measures: *Root Mean-Squared Difference* (RMSD) to quantify the difference between two spatial distributions and *Absolute Magnitude Difference* (AMD) to quantify the difference between two global mean GPP.

RMSD is obtained comparing corresponding intensity values pixel-by-pixel between the two maps using Euclidean distance. Both of these metrics were area-weighted as equatorial regions have higher climatological weight than tropical regions. Maps A and B have similar global mean GPP when AMD is low and similar spatial distributions when RMSD is low.

Figure 3.5 shows the distribution of these two metrics in the data that we generated. In order to create effective stimuli we selected, for both measures, map pairs in the lower quartile, to generate cases of high similarity, and those in the upper quartile to generate case with low similarity. Accordingly we have four bins in the data: similar global mean GPP and similar spatial distribution (Figure 3.4(a)), similar global mean GPP and dissimilar spatial distribution (Figure 3.4(b)), dissimilar global mean GPP

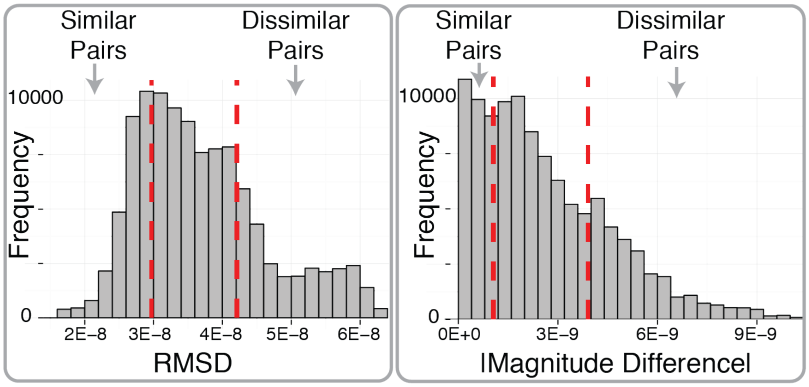


Figure 3.5: **Data Generation:** The histogram for Root Mean-Squared Difference (RMSD) is shown that quantifies the distance between the spatial distributions. On the right, histogram for the absolute value in magnitude difference is shown, where magnitudes are the global mean `GPP` values for maps A and B.

and similar spatial distribution (Figure 3.4(c)), and dissimilar global mean `GPP` and dissimilar spatial distribution (Figure 3.4(d)).

3.5.3 Trials and Participants

In order to allow all participants to be exposed to all color scales, we decided to design the experiment as a repeated measures design. For all tasks we selected 2 samples from each of the 4 combinations discussed above with a total of 24 trials for each task.

Each participant was exposed to all tasks and trials. The tasks where ordered sequentially and the trials where randomized to get rid of learning effects.

- For **Task 1** the participants were asked to answer the following: *"Given the global mean `GPP` based on one map (A), what is the global mean `GPP` of map (B)?"*. For providing their answer, participants had to adjust a slider, the range of which was set from the overall minimum to the overall maximum of mean `GPP`.
- For **Task 2**, they were given two maps and asked the question: *"how similar are the spatial distributions of the two maps?"* They were provided with a Likert scale, the range of which was 1 (most dissimilar) to 5 (most similar).
- For **Task 3**, they were given the same pair of maps as in Task 2 and asked to answer the question: *" identify the region with maximum difference between the two maps"*. In this case, they had to select a particular point on the map which they thought was the roughly the center of the region.

Since both Task 2 and Task 3 deal with the problem of identifying spatial distributions we decided to share the same trials between the two tasks, that is, the questions were asked based on the same map-pair. Since Task 2 and Task 3 were executed on the same trial, each participant had to answer questions for a total of 48 trials.

We selected our participants anonymously through mailing lists of climate scientists, and 39 participants completed the study. Among the 24 were male and 15 were female. Since 3 of them reported for color-blindness, we included the responses of the rest 36 in our study. The participants were from 24 to 65, with the median experience being 10 years in climate science and 6 years in using color scales with maps.The range of their overall experience was between 0 and 33 years. *The total number of trials was thus* $48 \times 36 = 1728$.

3.5.4 Study Setting

The experiments reported in this study were all web-based. This setting was necessary as all our participants in the study are climate scientists spread across different academic institutions and research labs across United States and Europe, implying the necessity to conduct this study remotely. One of the critical issues with our study is to ensure reliability and minimization bias in the results. In our experimental set-up we took several measures to address these. First, we took care of the case if a participant did not understand the question or if he/she is ready for the test. To this effect, we showed them example questions and let them quit the study if they did not understand the question. They could not go back to check the answers or get a feedback on the correctness of their responses. The IP address of the participants are recorded, so we know if the same participant has responded twice. Even if they stopped the study and took a break, they would not be allowed to start from the beginning. They had to start from where they left off. This prevented unintentional repetition of the tasks by a participant.

3.5.5 Metric for Correctness of Magnitude Judgment (Task 1)

In case of Task 1 our ground truth is the true value of global mean `GPP`. However, this judgment was made based on the reference map. To capture the comparative nature of the task, where the subjects had to judge a value relative to one, we needed a metric which treated the judged magnitude as a fraction of the given value. For computing the relative error we take inspiration from metric proposed by Cleveland and McGill [38], which can be formulated as:

$$\begin{aligned}\text{Judged Per Cent} &= \frac{\text{Estimated } GPP_B}{GPP_A} \times 100\\ \text{True Per Cent} &= \frac{\text{True } GPP_B}{GPP_A} \times 100\\ \text{Relative Error} &= |\text{Judged Per Cent} - \text{True Per Cent}|\end{aligned}$$

By normalizing with respect to the true GPP value, we are accounting for the different ranges in the data which can affect the amount of difference in the GPP values, that can be overestimated and underestimated.

3.5.6 Metric for Correctness of Similarity Judgment (Tasks 2)

The ground truth for similarity is computed using the *RMS* distance between two maps as we had discussed before. However, this is a computed measure of similarity which might be different from the perceived similarity. Determining correctness based on precise classification can be problematic as there can be individual differences in perceiving the degree of similarity. For a more adaptive metric to participants' performance, we wanted to select a *similarity threshold* based on the distribution of the responses.

Similarity Threshold. For measuring the error in judgment, we split the responses into two parts: those responses which are greater than 3 for pairs with dissimilar magnitude and those which are less than 3 for similar magnitude. These are the cases with errors.

3.5.7 Metric for Correctness of Identification of Most Dissimilar Regions (Tasks 3)

For evaluating the correctness of the responses we compute the precision of the scientists' click by counting the number of clicks in a dissimilar region divided by the total number of clicks. However, dissimilarity between two maps is a continuous function, and we need to define a *dissimilarity value* (d) for which regions are most different in those maps. At d=0, every click is correct as maps are totally different as all differences are greater than 0. When we increase d, maps gradually become less and less different and the precision of clicks become less and less. Accordingly we plot the precision function and select our threshold to be at the 50 *percent* level, which is

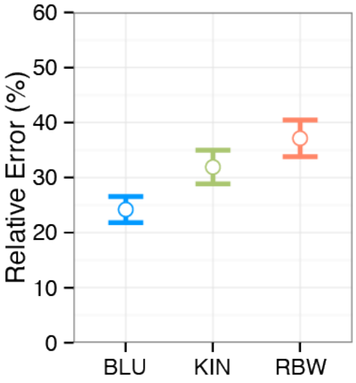


Figure 3.6: **Overall Relative Error in Task 1.** We can observe that for Task 1, accuracy was higher on average in the *BLU* than the *KIN* or the *RBW* color scales.

a measure of central tendency, as half the scores are above and half are below. Based on this threshold, we select the corresponding dissimilarity value (d) and evaluate the performance of the color scales for that value of dissimilarity between a pair of maps.

3.6 Results

In this section we report the significant results ($p < 0.05$) for all three tasks. For all our results we computed the 95% confidence intervals using the bootstrapping method.

3.6.1 Task 1: Judgment of Magnitude

As described in Section 3.2, Task 1 was about magnitude judgment: judging the global mean GPP from one map, where the same was given for another map.

3.6.1.1 Overall Effect

Figure 3.6 plots overall performance , across all four conditions, using the three color scales, *RBW*, *KIN* and *BLU*. The dependent measure is the relative metric which we had described earlier in Section 3.5.5. Across all conditions, users had a significantly higher error rate with the *RBW* (37%), and significantly fewer errors with *KIN* (32%) and the *BLU* (24%). That is, the two monotonic luminance scales were more effective in helping the analysts make correct judgments about the global mean GPP than the *RBW*, and the *BLU* was superior to the hue-enhanced *KIN* scale. These results were significant at the $p < .001$ level (Friedman ($\chi^2(2) = 39.38$). A post-hoc Nemenyi pairwise test ($p = .05$) revealed that performance with the *BLU* was

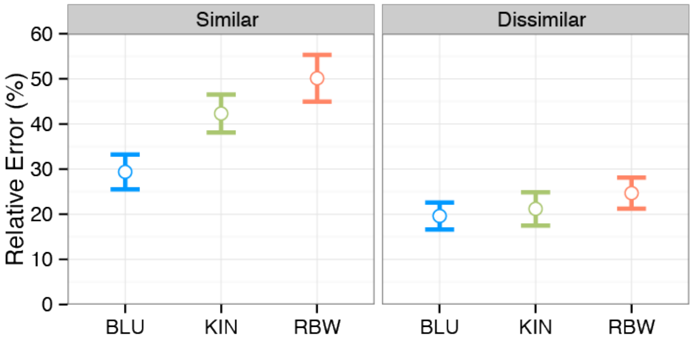


Figure 3.7: **Effect of Spatial Distribution on Task 1.** One of the surprising findings in Task 1 was that scientists committed more errors across color scale when spatial distributions were similar (on the left) than when they were dissimilar. Also color scales have less effect on the judgment.

significantly better than *RBW* ($p < .001$), and *KIN* ($p < .001$), and that performance with the *KIN* map was significantly better than with the *RBW* ($p < .001$).

3.6.1.2 Effect of Spatial Distribution

Figure 3.7 drills down to examine relative error for pairs of maps that are either similar (left panel) or dissimilar (right panel) in their spatial distribution. The ordering of results for the three color scales is the same in both conditions, that is, *RBW* produces the highest relative error, followed by *KIN*, followed by *BLU*. All these differences are significant when the maps being compared are spatially similar. Friedman ($\chi^2(2) = 46.4$, $p < .001$), with all differences between color scales significant ($p < .001$) in the Nemenyi pairwise test ($p = .05$). There was a significant, but weaker, main effect of color scales when the maps being compared were spatially dissimilar, Friedman($\chi^2(2) = 8.72$, $p < .05$), with only the difference between the *RBW* and the *BLU* scale being significant in the Nemenyi pairwise test ($p = .05$). Thus, the *RBW* color scale affords less accurate comparisons of magnitude, whether the spatial distributions are similar or dissimilar, but the degree to which the monotonic luminance scales outperform is much greater when the maps are similar. These statistical results are summarized in Table 3.1. This also shows clearly that the task of judging `GPP` is much harder when the maps have similar spatial distributions. The relative error using the *RBW* is almost twice that when using the *BLU*, and it is only that *BLU* that is in the same range as the dissimilar comparisons.

		Relative Error Mean and 95% C.I.			
		Rainbow	Kindlmann	Blues	p-value
distribution	overall*	37.0 [33.7, 40.3]	31.9 [28.9, 34.9]	**24.3 [21.8, 26.7]**	<0.0001
	similar *	49.6 [44.7, 54.5]	42.9 [38.7, 47.1]	**29.0 [25.2, 32.7]**	<0.0001
	dissimilar*	24.9 [21.3, 28.5]	21.6 [17.8, 25.4]	**19.5 [16.7, 22.3]**	<0.05
magnitude	similar*	31.5 [26.3, 36.6]	22.4 [18.9, 25.8]	**16.1 [13.7, 18.6]**	<0.0001
	dissimilar*	43.7 [39.2, 48.3]	41.6 [37.1, 45.9]	**32.2 [28.5, 35.8]**	<0.0001

Table 3.1: **Relative Error Mean and 95% C.I. in Task 1**: Significant results are indicated by *. Higher accuracy is indicated in bold. Overall error and those for similar and dissimilar distribution are shown in each row.

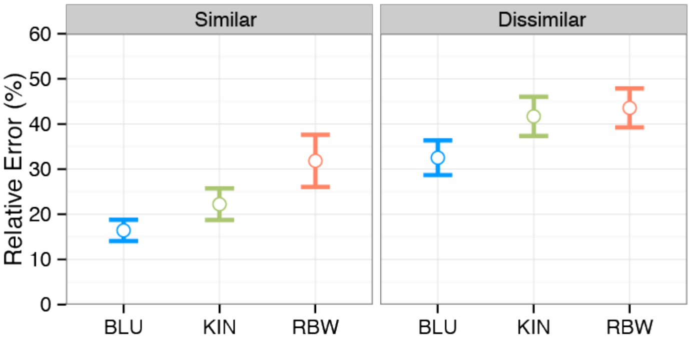


Figure 3.8: **Effect of Magnitude on Task 1.** There was a consistent trend of less errors with *BLU* for both similar and dissimilar magnitude cases. As expected, scientists commit higher errors when magnitudes are dissimilar.

3.6.1.3 Effect of Magnitude

Figure 3.8 drills down to examine relative error rates for pairs of maps that are either similar (left panel) or dissimilar (right panel) in their magnitude, that is the global mean `GPP`. For similar magnitude, there is a significant difference in performance between *RBW* and both the color scales. Friedman ($\chi^2(2) = 25.7$, $p < .001$), with all differences between color scales significant ($p < .001$) in the Nemenyi pairwise test ($p = .05$), $p < .001$. A post-hoc Nemenyi pairwise test ($p = .05$) revealed that performance with the *BLU* was significantly better than *RBW* ($p < .001$), and *KIN* ($p < .01$), and that performance with the *KIN* map was significantly better than with the *RBW* ($p < .01$). The variability in the estimates also much higher in case of *RBW*.

However, for dissimilar magnitude the errors are higher across all color scales, and the difference between *RBW* and *KIN* is less pronounced. These results are also significant. Friedman ($\chi^2(2) = 36.22$, $p < .001$), with all differences between color scales significant ($p < .001$) in the Nemenyi pairwise test ($p = .05$), $p < .001$. A post-hoc Nemenyi pairwise test ($p = .05$) revealed that performance with the Blue scale was significantly better than *RBW* ($p < .001$), and *KIN* ($p < .001$), and that performance with the *KIN* map was significantly better than with the *RBW* ($p < .01$).

3.6.1.4 Analyzing the Effects by Drilling Down

A full breakdown of the data is shown in Figure 3.9. In this case, instead of representing the errors, we show the comparison between the judged magnitude and the true magnitude. The rows in this 2 × 2 quadrant represent data for maps that were either spatially similar (top row) or spatially dissimilar (bottom row). The columns show the conditions where the maps were either similar (left column) or dissimilar (right column) in overall magnitude. There are two sets of data within each quadrant, which show the results for the two test cases. Unlike Figures 3.7 and 3.8, we are plotting the judged percent difference between the comparison map and the standard, not the relative error. Ground truth, rather than being a normalizer, is shown explicitly as a short vertical line associated with each data set. The figure of merit in this graph is the degree to which the judged `GPP` value approximates the value of the vertical line (the ground truth). The closer the data points to the ground-truth line, the better the performance.

As we saw in Figure 3.7, the biggest difference between color scales occurred when the spatial distribution between the comparison map and the standard was similar. These results are broken out in the two top quadrants of Figure 3.9. When the magnitude of the `GPP` difference is low (top left quadrant), the ordering of the color maps observed in Figure 3.7 is maintained in all trials. *RBW* is farthest from ground truth, with *KIN* second, and *BLU* providing the best vehicle for capturing ground truth. This effect is stronger in the first trial (left set). The top right quadrant shows the case where the maps had similar spatial distributions but dissimilar magnitudes. In one trial, the real percentage difference in `GPP` was low (right set); in the other, the real percentage difference in `GPP` was high (left set). In both cases, performance using *BLU* is closest to the ground truth, with greatest departure from ground truth with the *RBW*.

We also learned in Figure 3.7 that the differences between color scales was weaker when the two maps being compared had dissimilar spatial distributions. These are broken out in the two bottom quadrants in Figure 3.9. When the spatial distributions of the two maps were dissimilar, but the magnitudes were similar, the observers were able to judge ground truth accurately, independent of the color scale. When the spatial distributions were dissimilar and the magnitude was also dissimilar, there were significant departures from ground truth, and in one of the trials, the lowest level of

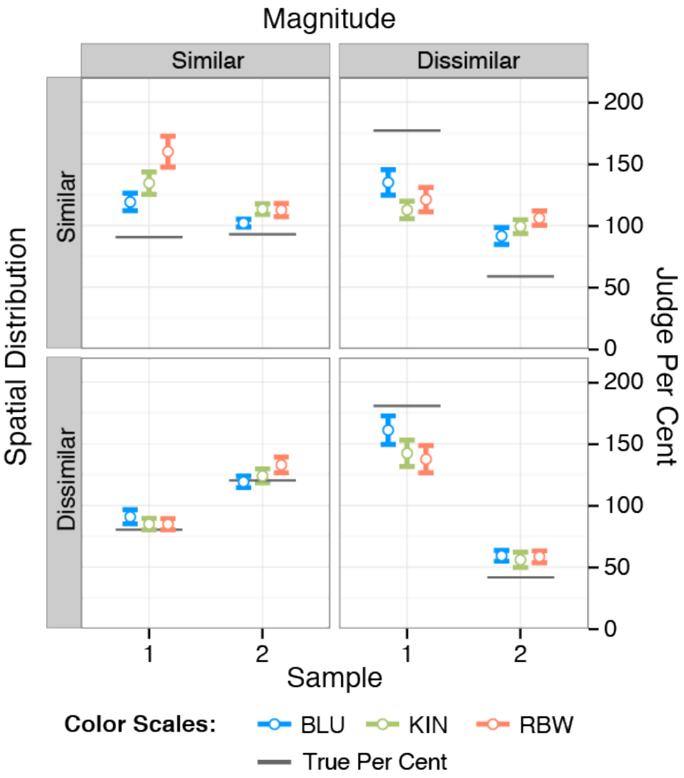


Figure 3.9: **Task 1: Confidence intervals showing estimates of** `GPP` **B with respect to the four quadrants**: Participants were much more consistent for dissimilar spatial distribution across similar and dissimilar magnitudes. On the other hand, the performance was the worst for similar distribution and dissimilar magnitude.

performance was afforded by *RBW*.

Another interesting result in Figure 3.9 comes from looking at the individual trials within each quadrant. When `GPP` is low, (*e.g.*, the ground truth line is toward the bottom of the quadrant), the observers tended to overestimate the `GPP` level of the comparison stimulus. When `GPP` was high, (*e.g.*, the ground truth line is to the top), the observers tended to underestimate the level. This effect is seen in three of the four quadrants. In the fourth quadrant (dissimilar spatial/similar magnitude) the observers' judgments were very accurate.

3.6.1.5 Confidence vs. Relative Error

First we analyzed if confidence levels vary across color scales. We found that there are significant differences in confidence levels as revealed under a Friedman test: $(\chi^2(2) = 13.5, p < 0.01)$. We found that scientists were more confident on average with the *RBW* than the *BLU*, and more confident on average with the *KIN* than

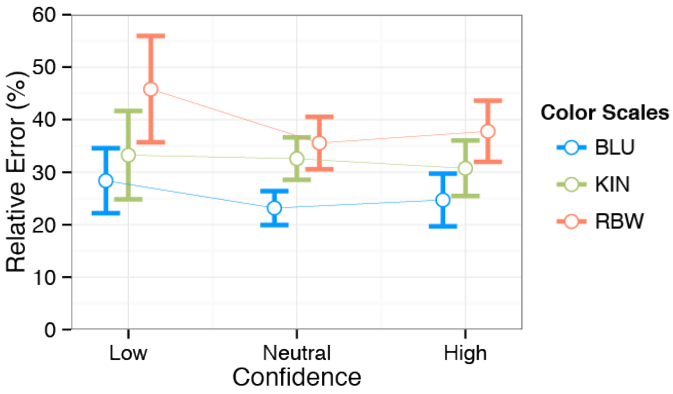


Figure 3.10: **Confidence vs Error for Task 1.** We can observe that even with high confidence the error was less in the *BLU* as compared to the *RBW*.

the *BLU*. A posthoc Nemenyi pairwise test ($p = 0.05$) revealed these differences were significant with $p < 0.001$. We expected the scientists to be more confident with the *RBW* as compared to the other ones.

However, we found that regardless of the confidence level, they were more likely to be more inaccurate with the *RBW* as opposed to the *BLU* as observed in Figure 3.10. Whether they were low on confidence or high on confidence, they committed more errors with the *RBW* color scale.

3.6.1.6 Summary of Findings

For the magnitude judgment task, one of our hypotheses as we had described in Section 3.4 was that accuracy with multi-hue color scales would be higher due to greater discriminability among the hues. This was also hypothesized by Ware [73]. However, based on the task and conditions of our study, we did not get an evidence to support this hypothesis. The most interesting finding was the fact that similar spatial distributions not only showed greater error in judgment, but also a greater variability in the effect of the color scales on the task. We had expected scientists to perform better in magnitude judgment tasks when the distributions were similar. On the contrary, we found that color scales did not affect the performance significantly when the spatial distributions were different. This finding needs further research for diagnosis of the cause. Possibly, the *what* judgment is made easier by dissimilar distributions, where it is easy to distinguish between different shapes at different locations on the map. But in case of similar distributions, the *what* judgment is difficult as most shapes are similar. The magnitude judgment then occurs at an elementary level [35] and hence is error-prone. Another interesting finding was the case of *similar magnitude*

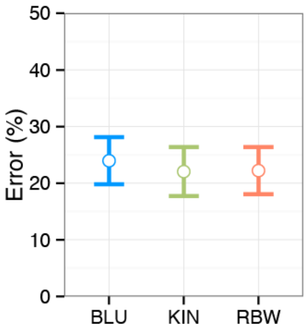


Figure 3.11: **Task 2 Results:** For the task on comparison of pairwise similarity of maps, we failed to detect any effect of the color scales. Even after drilling down by differences in distribution and magnitude, there was no significant effect detected.

and *dissimilar spatial distribution*, which was the only condition under which not only error of judgment was low, but the effect of the color scales was also not significant. A major finding was that although scientists were confident with their judgments, they consistently commit more errors with that color scale. Also, when they were not confident the error with the *BLU* was not only low, but they agree more, as given by the small confidence interval was also low.

3.6.2 Task 2: Judgment of Spatial Distribution

As described in Section 3.2, Task 2 was about judging the degree of similarity between a pair of maps. We evaluate the performance on Task 2 based on the similarity threshold we set, as described earlier in Section 3.5.6. The similarity threshold was necessary for comparing the measured similarity in the data to the degree of perceived similarity as indicated on the Likert scale. As observed in Figure 3.11 we failed to detect significant difference in performance across all the color scales. These results were not significant under a Friedman test: $(\chi^2(2) = 0.97, p = 0.6)$. We found similar results while drilling down into similar and dissimilar spatial distribution. This was a surprising finding as we expected at least *KIN* to perform significantly better than *RBW* or *BLU*.

3.6.3 Task 3: Identification of Most Dissimilar Region

As described in Section 3.2, Task 3 was about identifying the region of maximal difference, and as explained in Section 3.5.7, we evaluate the scientists' performance on identification of dissimilar region based on computing the precision of their judgments. In 3.12 we show an example of the variance in the clicked regions across different color scales.

We can observe in Figure 3.13(a) that the precision of a correct click for the *KIN* is higher than that of the *BLU* or the *RBW*. Based on the chosen $threshold = 6.6e^{-8}$, we observe that the *KIN* performed the best in helping the scientists identify the most dissimilar regions, as shown in Figure 3.13(b). Although this trend was found for a particular dissimilarity threshold, we found the results to be consistent across any chosen threshold. Using the Friedman's test we get statistical differences across the color scales ($\chi^2(2) = 15.75$, $p < 0.001$), a pairwise analysis revels that there were differences across the three color scales. The largest difference was between the *BLU* and the *KIN* with a difference of 11% ($p < 0.001$), followed by the *KIN* and the *RBW* with 8% ($p < 0.05$) and finally the *BLU* and the *RBW* with a difference of 3% in their correctness ($p < 0.05$).

There was a statistical difference in confidence across the color scales ($\chi^2(2) = 6.15$, $p < 0.05$). A pairwise comparison shows that the difference was between the *BLU* and the *KIN* ($p < 0.05$) and the *KIN* with the *RBW* ($p < 0.05$). In average the less confident color scale was the *BLU* with an average of 3.5.

3.6.3.1 Summary of Findings

To our surprise we failed to get support for our hypothesis that a multi-hue color scale would let better performance on Task 2, that is judging the degree of similarity between two maps. This is an interesting finding which could mean that the domain knowledge of the scientists are able to overcome the shortcomings of the *RBW*, but we need to conduct more experiments to test this hypothesis. Our hypothesis that *RBW* would perform worse than *KIN* in the *what* judgment of identifying regions of highest dissimilarity was proved correct as the *KIN* performed better for selection of dissimilar regions. Surprisingly, scientists were equally confident with all color scales in both

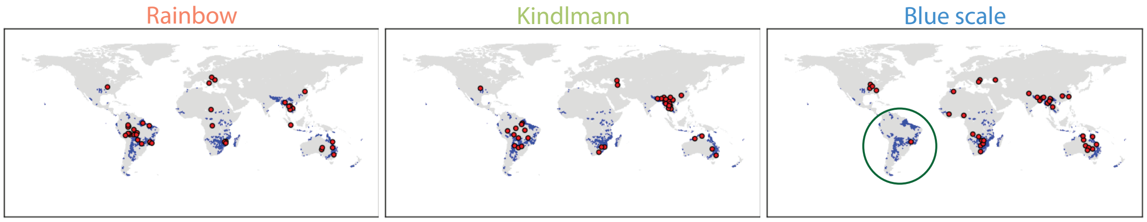
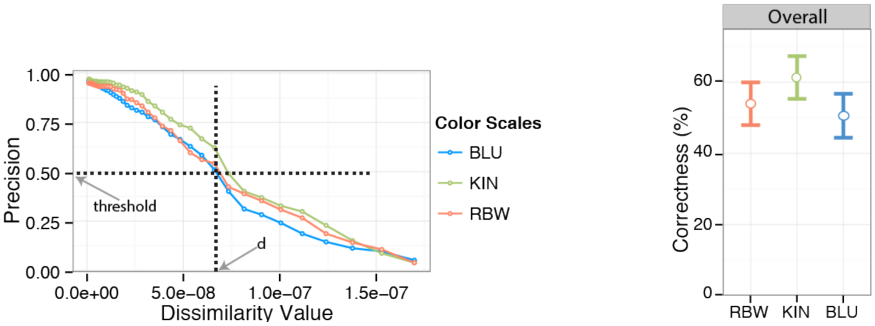


Figure 3.12: **For Task 3, difference maps showing click spots** at a *dissimilarity value*=$6.6e^{-8}$ selected based on the 50 per cent threshold as shown in Figure 3.13(a). We can observe that for the marked region, the number of clicks in South America is far less in the *BLU* as compared to the *RBW* or the *KIN*.



(a) **Precision curves** for different color scales showing the probability of a click to be in a dissimilar region and the confidence intervals for correct responses above the selected dissimilarity value.

(b) **Performance of color scales** As observed, the *KIN* is the most accurate in identifying most dissimilar regions and *BLU* is the least accurate.

Figure 3.13: **Task 3 Results:** We found the *KIN* was more effective in letting scientists identify the most dissimilar regions between maps.

Tasks 2 and 3, although their post study reveals they were much more confident on average with the *RBW*.

3.6.4 Survey of Participants' Perceived Performance

One of the goals of our study was to compare the perceived accuracy and confidence of the scientists with the objective measures from the study. To this effect, we collected subjective feedback from our participants in the last section of the study. We collected feedback about their familiarity, preference, confidence, perceived accuracy and ease of use of the color scales, by asking questions such as: "which color scale did you prefer the most", "which color scale were you most confident with", etc. The results are shown in Figure 3.14.

3.6.4.1 Perceived Accuracy and Confidence vs. Familiarity

Since the *RBW* is the *de facto* standard in climate science, it was not surprising that over 90% rated that they were most familiar with it. Despite the familiarity with the *RBW* among an overwhelming majority of them, nearly 25% of the participants felt more accurate or confident with ether the *KIN* or the *BLU*.

3.6.4.2 Familiarity vs. Preference

Second, comparing familiarity to preference, we observe a drop of nearly 40% for the *RBW*, which is compensated by more participants preferring either the *BLU* or

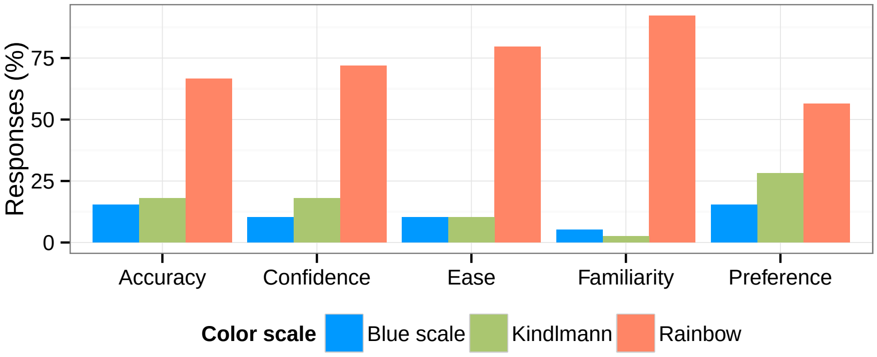


Figure 3.14: **Participant ratings** in per cent based on different subjective categories revealed despite their over all familiarity with the *RBW*, about 43% of the participants preferred the *KIN* or the *BLU* and 33% felt they were more accurate with them.

the *KIN*. As a corollary of these two findings, we can comment that despite the long history of the *RBW* usage, the study convinced many climate scientists about the efficacy of perceptually motivated color scales for their tasks.

Following are some of their comments that highlight their higher preference for the *KIN*:

> *"Kindlmann works best because it has both good tone contrast AND value contrast across the spectrum, whereas rainbow has good tone contrast but little value contrast and blues has little color contrast and not great value contrast."*

Another participant remarked:

> *"It was easier to see magnitude of change with rainbow, and especially hotspots in red. My concern was that I was overestimating the red areas and not paying enough attention to changes at the other end of the spectrum. I thought my first sense of overall global pattern change was easier with blues but it was much harder to compare changes in spatial pattern or magnitude between different regions. Kindlmann was therefore a compromise for me...not as dramatic, did not highlight the hotspots as much, but allowed me to compare differences more easily across regions."*

3.6.4.3 Perceived vs. Actual Accuracy

Comparing the accuracy of the responses of the scientists to their preferences from the subjective feedback, we observe a big disparity between their preferences and performance. Majority of them were familiar with the *RBW* and also preferred that scale for the study, thinking they are more accurate. But this is in contrast to their performance, where the average correctness for the *RBW* and the *KIN* are much greater. Also, their high confidence with the *RBW* did not translate into higher accuracy. Their introspection about how well they performed were out of sync with their performance.

3.6.4.4 Perceived Accuracy vs. Agreement

We would expect scientists to agree more on their estimations for rainbow color scale than any other color scale because of their familiarity and preference. However, as we had shown in Figure 3.10, there was surprisingly high variability in error in case of rainbow color scale, as opposed to the blue color scale, which they preferred the least, across all confidence levels. This further demonstrates that there is a clear discrepancy between what the scientists believe they are more accurate and comfortable with, and what they are actually accurate with on a given task.

3.7 Discussion

In this section, we reflect on the key findings in the context of the experimental settings and implications for future work. The goal of these experiments was to examine the effectiveness of different color scales in performing complex real-world analytical tasks. A major focus of our study was to develop test stimuli that spanned a wide range of climatological conditions and capture commonly performed tasks with color-coded geographical maps. To this effect, we constructed maps with the color scale representing the magnitude of the scalar variable, i.e., `GPP`. We expect that our results would generalize to the representation of any scalar variable on across a geographical map, at least at the spatial resolutions we studied.

3.7.1 Effect of Color Scales on Judgment of Magnitude

Across most of the tasks, the *BLU* outperformed the others. A simple grayscale ramp, consisting of one hue, enabled the best decisions about the magnitude and spatial variations between maps, and where the maps most differed. We believe this result owes largely to the monotonic luminance profile for the *BLU*, since luminance has been shown to be an effective “carrier” of magnitude information, especially for high spatial-frequency information. The *KIN* was also more successful than the *RBW*, and also includes a monotonic luminance component.

3.7.2 Effect of Multiple Hues

The *KIN* provides a variation in hue as the luminance increases along its range. It has been posited that these additional hue variations would provide additional levels of discriminability, and it has even been suggested that hue changes are a requirement for judging magnitude [73]. We found, to the contrary, that these additional hue variations detracted from performance in magnitude judgment and had little effect on similarity judgment.

3.7.3 Effect of Color Scales on Judgment about Spatial Distributions

The additional hue variations in the *KIN* as compared to the *BLU* were helpful when the scientists had to identify most dissimilar regions. Like the *RBW*, the *KIN* has ranges of highly saturated colors, and although luminance is monotonically increasing, there are still distinct hue regions within the scale. We think that these ranges may interfere with the observers’ ability to judge smooth variations in magnitude, if they occur within one of these color bands.

3.7.4 Need for Color Scales that are Adaptive to the Tasks

For these experiments we selected color scales that not only embody important perceptual variations, but which are also commonly used. Based on our current results, we are eager to develop color scales that best exemplify the characteristics we feel are most important. For example, a *BLU* with an even greater dynamic range may be provide even greater advantage, and a *KIN* that has less saturated colors, and therefore less banding, may provide all the advantages of a monotonic luminance profile plus the extra advantage of a subtle hue variation.

3.8 Summary

The study presented in this chapter showed that despite preference and familiarity with the rainbow color scale, of a large group of domain scientists, their performance was almost always consistently better with perceptually motivated color scales.

The most important contribution of this study is that it identifies analytical tasks that scientists perform, and explores how to augment that analysis and decision-making process with appropriate color scales. We helped the scientists understand the phenomena under study, to appreciate differences in magnitude and spatial variation across models. We believe the outcome of this study will inspire further experiments related to effects of color scales in the domain of climate science and beyond, which in turn can lead to greater and wider adoption of visualization best practices.

As result of the studies from this chapter and Chapter 2, we understood the common problems on visualization created by domain experts. Moreover, by reflecting on the inadequacies of the static visualizations, we decided to continue our work by designing a novel *Visual Exploration Tool* to interactively analyze similarities and differences among climate models which will be explained in the next chapter.

Chapter 4

SimilarityExplorer: A Visual Intercomparison Tool for Multifaceted Climate Data

As a result of our study presented on Chapter 2, we noticed the limitations of static-visualization in model comparison tasks. An insightful analysis in climate science depends on using software tools to discover, access, manipulate, and visualize the datasets of interest. These data exploration tasks can be complex and time-consuming, and they frequently involve many resources from both the modeling and observational climate communities.

Consensus among model results is an important metric used for judging model performance. Analysis of model output similarity and dissimilarity is a complex problem because of the multiple *facets* involved in such comparisons: space, time, output variables, and model similarity.

The goal of this work is to provide an interactive visualization tool that integrates space, time, and similarity, making it easier for climate scientists to explore model relationships from multiple perspectives.

The output of our chapter is a result of a six-month-long interaction between visualization researchers and climate scientists. Modelers generally perform their analyses by looking at spatial and temporal aspects in isolation, by running scripts, such as *MATLAB* and *R* on the data and by manually setting parameters. The first step during the iterative development of our tool was to provide the scientists with an interactive interface for selecting parameters and filtering the data. This was not sufficient as our interactions revealed that modelers needed a tool for analyzing both space and time within a single interface in order to judge multi-model similarity.

Existing visualization tools are only capable of integrating one or two facets as pointed out by Kehrer and Hauser [79]. Multifaceted data analysis is inherently challenging on two counts: i) preserving the mental model about the different facets, like space, time, and model similarity, necessitates an encoding strategy that preserves visual symmetry, and ii) exploring these facets at multiple levels of granularity and understanding their relationships necessitates a systematic interaction strategy. To address these challenges, in this chapter we present SimilarityExplorer tool which enables mutifaceted visual analysis of climate models, specifically, Terrestrial Biosphere Models (TBMs). Using our tool, climate scientists were able to get an overview of model similarity across space and time, and then drill down to further explore *where*, *when*, and by *how much* models were similar or different. A seamless integration and exploration of these facets in SimilarityExplorer let them generate and explore new hypotheses about model similarity which was not possible before.

This chapter consists of three key contributions: i) As part of the domain characterization [80] of climate model intercomparison, we present a systematic classification of the domain-specific intents of climate scientists, and that of the underlying data facets (Section 4.3), ii) we bridge the intents and facets with the visualization tasks and design through a classification scheme (Section 4.4); and iii) SimilarityExplorer is a tool that implements this classification. Our interactions with climate scientists were conducted before, during, and after the implementation phase for iterative refinement of the tool based on their feedback. In light of this, we present two case studies which helped elucidate and validate the benefits that scientists obtained when using SimilarityExplorer (Section 4.5).

4.1 Related Work

In this section we discuss the relevant related work with respect to spatiotemporal and multifaceted data visualization and tools available for climate data.

4.1.1 Simultaneous Encoding of Spatial and Temporal Relationships

Visualization of spatiotemporal data has witnessed a lot of research over the years. Peuquet [81] had introduced the popular triad representation framework which is a general formalization of temporal dynamics in geographic information systems. In our tool we imbibe the concepts of *when*, *where*, and by *how much* models are similar. The need to integrate space and time through an exploratory analysis tool was also proposed by Andrienko et al. [82]. They devised a visual analytics [83] framework for exploring spatiotemporal data through spatially referenced time series. Similarly, visual analytics approaches for event detection [84, 85, 86] have been proposed where spatial representation of the data is provided in conjunction with features for observing temporal trends and anomalies. While most of this work focused on direct encoding of the data, either spatially or temporally, Andrienko et al. applied self-organizing maps [87], for providing complementary perspectives on spatial and temporal relationships which is the guiding principle in SimilarityExplorer. The complexity in our work evolves from the fact that the scientists needed to understand the evolution of both spatial and temporal relationships simultaneously. This necessitated that the visualization provided an overview of spatial and temporal relationships, and then also allowed flexible interaction for exploring these relationships over both space and time.

4.1.2 Integration of Spatial and Non Spatial Data

There exists other approaches towards building visualizations for integrating spatial and non-spatial data [88]. Guo et al. [89] proposed a generalizable visual analytics approach for integrating techniques from cartographic, visualization techniques and machine learning. That methodology is general and can be applied to spatiotemporal data. Most of the existing tools only integrate one or two different facets [79]. In the SimilarityExplorer we integrate four different facets: *space*, *time*, *multiple variables*, and *model similarity*, which are crucial for visual comparison of the different properties of models. Our technique is similar in principle with Kehrer et al.'s work on visual analysis of heterogeneous data, multi-model scientific data with examples from climate research data [90]. Kehrer et al. focus on providing multiple perspectives into statistical relationships between multi-run and spatially aggregated simulation data through different interactive views. In SimilarityExplorer, similar to multi-run data, we focus on multi-model data; and in addition to spatial relationships and patterns, we consider time, multiple variables and different visual approaches towards encoding similarity and facilitating visual comparison through the use of small multiples [9].

4.1.3 Visualization solutions for climate data

For addressing the needs of the climate research community, there has been some work on hypothesis generation [91], task characterization [92], and tool development [93]. Steed et al. introduced EDEN [94], a tool based on visualizing correlations in an interactive parallel coordinates plot. Their focus is on a single model and analysis of the interdependence among variables. There also exists some general visualization tools such as Paraview [70], Visit [95] and VisTrails [96] which offer some specialized climate visualizations but almost all of them only present the data without supporting any analysis. Those specialized packages were integrated in a provenance-enabled climate visualization tool UV-CDAT [97]. However, like most other tools, UV-CDAT does not support multi-model analysis. It also does not support multivariate analysis and dynamic linking between the views. Through a closely-knit collaboration with climate scientists we were able to address the need for tools that emerge from genuine and interdisciplinary collaboration [98, 99], for solving the problems with such complex data.

4.2 Background of Model Intercomparison

We collaborated with 3 climate scientists from the Oak Ridge National Lab as part of the MsTMIP [1] Project. Each of them have at least ten years of experience in climate modeling and model intercomparison.

4.2.1 Data

The data consist of simulations from 7 different TBMs for over 20 years at monthly temporal resolution, collected over a spatial resolution of 0.5 degree. Each produces multiple output variables, of which *three* are relevant for the analysis presented here. For segmenting the globe, the scientists use 11 different eco-regions. The temporal granularity of interest to them were annual, seasonal, and monthly. As shown in Figure 4.1 each model can be represented by a spatiotemporal volume over latitude, longitude, and time. Since each model is associated with multiple output variables, each model can be thought of as being a vector of such volumes. The basic goal of climate scientists is to efficiently subset this array of cubes along multiple dimensions, in order to understand model similarity based on multiple facets: *when* are models similar, with respect to seasons and months, *where* are models similar, with respect to regions, *why* are models similar, with respect to the output variables.

[1] MsTMIP: Multi-Scale Synthesis and Terrestrial Model Intercomparison Project

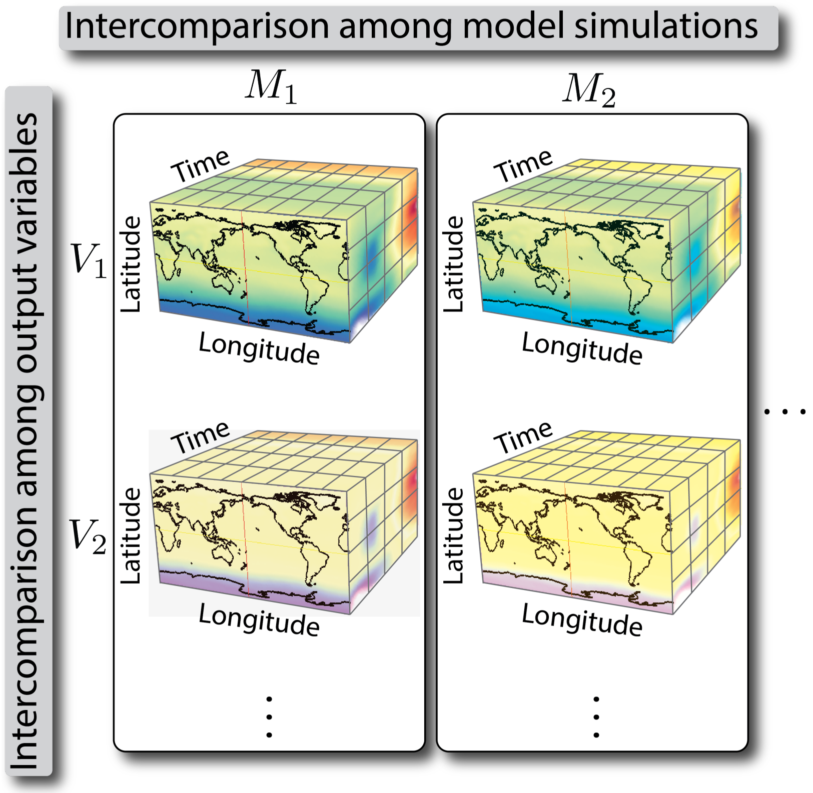

Intercomparison among model simulations
M_1
M_2
Intercomparison among output variables
V_1
V_2
Time
Latitude
Longitude
...

Figure 4.1: **Visualizing the complexity of multifaceted climate data** in terms of models, regions, time and variables.

4.2.2 Model Similarity

As a first step in our design study [100], we discussed with the climate scientists about their existing approaches for understanding model similarity. To reduce complexity of the data, they are used to compressing space and time. It emerged that, from a temporal aspect they are mostly interested in comparing model behavior for seasons or months aggregated across all years. In this context, they perform two distinct operations on the data for analyzing similarity from the spatial and temporal perspectives. These operations are sketched in Figure 4.2 and described below.

Spatial Correlation. For this step, as shown in Figure 4.2(a) the data is preprocessed in such a way that temporal information is aggregated but spatial granularity is preserved. For each point on the map, the average value for a time period is computed. Temporal granularity can range from *long-term mean* (value at one point is the average for all months and all years within the time period), *long-term monthly mean* (12 monthly maps, with each map representing an average month for the time period), and *seasonal* mean (four maps with each map representing an average season for the

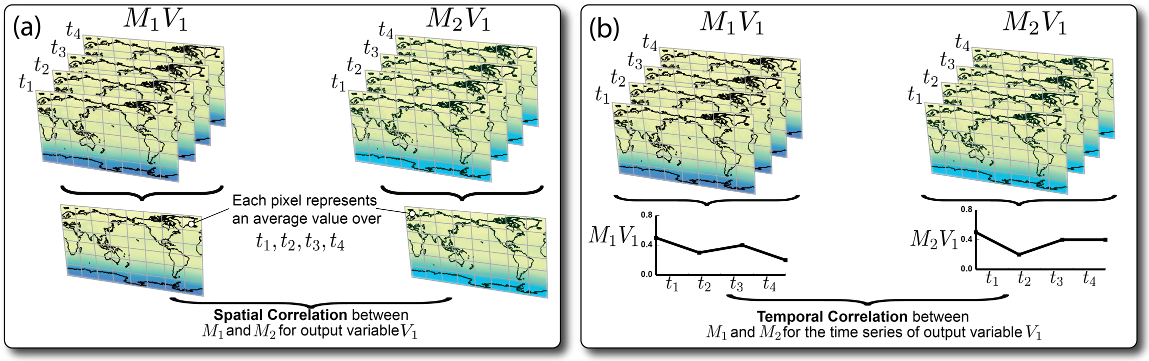

(a)
M_1V_1
M_2V_1
t_1
t_2
t_3
t_4
Each pixel represents an average value over t_1, t_2, t_3, t_4
Spatial Correlation between M_1 and M_2 for output variable V_1
(b)
M_1V_1
M_2V_1
0.8
0.4
0.0
Temporal Correlation between M_1 and M_2 for the time series of output variable V_1

Figure 4.2: **Similarity computation**. Illustration of spatial and temporal correlations are computed between models M_1 and M_2 after aggregating the temporal information. The spatial granularity is preserved at the cost of temporal information, and vice versa.

time period). Next, correlation between maps of two models is computed using the Pearson correlation coefficient.

Temporal Correlation. In this case, the data pre-processing helps aggregate spatial information but preserves temporal granularity (Figure 4.2(b)). For the map at each time step, a spatially averaged or summed value is computed. Next we compute a time series, which varies based on the temporal granularity: one value for *long-term mean*, 12 values for *long-term monthly mean* and four for *seasonal mean*. At the end the models are represented by their time-series signatures. While there are multiple ways for comparing time-series signatures of two models, in discussion with the scientists, we chose correlation as the measure for temporal similarity.

4.3 Domain Characterization

The initial discussion about the data characteristics was followed by an analysis of the domain-specific intents through face-to-face interactions and conference calls. In this section, we present the first contribution of our work, which is a characterization of the domain-specific intents of the climate scientists and the underlying data facets.

4.3.1 Domain Specific Intents

We identified four major intents of the climate scientists in the context of model intercomparison, which are as follows:

Questions	Tasks	Facets				Visualization Design	
		Space	Time	Variables	Similarity	Views	Comparison method
$\mathbf{Q_1}$	$\mathit{identify}(p)$ $\mathit{identify}(t)$ $\mathit{identify}(p,t)$	g/r	$a/s/m$	single	pairwise pairwise multi-way	matrix (maps) matrix (area graph) projection	explicit encoding
$\mathbf{Q_2}$	$\mathit{compare}(p,v)$ $\mathit{compare}(t,v)$	g,r	a,s,m	multiple	pairwise	matrices (map) *smlt*: maps matrices (area graph) *smlt*: area graph	juxtaposition
$\mathbf{Q_3}$	$\mathit{associate}(p)$ $\mathit{associate}(t)$	r	s,m	single	multi-way, pairwise pair-wise	parcoords, matrix time-series, matrix	juxtaposition superposition
$\mathbf{Q_4}$	$\mathit{distribution}(p,v)$ $\mathit{distribution}(t,v)$	r	s,m	multiple	multi-way, pairwise pairwise	parcoords time-series	juxtaposition superposition

p: Space t: Time v: Variables g: global r: regional a: annual s: seasonal
m: monthly *smlt*: small multiples

Table 4.1: **Translating tasks into visualization design through a classification scheme**. The visualization design is based on the different views and comparison methods required for reflecting the tasks that can be performed on the multifaceted data. Note that space and time have different levels of granularity. The symbol '/' reflects an OR operation and ',' reflects an AND operation.

$\mathbf{Q_1}$: In general, modelers would like to know the degree of spatial and temporal correlation of models with respect to any output variable.

$\mathbf{Q_2}$: With multiple models, they would additionally like to know which models are similar, and when, where, and why they are similar.

$\mathbf{Q_3}$: They would want to understand if different sub-regions agree or disagree with the global temporal or spatial correlations, or with the same for other sub-regions.

$\mathbf{Q_4}$: Scientists do not always trust the level of abstraction at which similarity is deduced, as there can always be anomalies that are not captured. Thus they wanted to look at the original distribution of the data to verify their hypotheses and validate their findings.

4.3.2 Facets: Space, Time, Variables, Similarity

The inherent complexity involving intercomparison of climate models stems from the multifaceted data underlying the climate models. The facets [79] relevant for the climate model data are space, time, variables, and similarity as shown in Table 4.1. *Space* and *time* also involve different levels of granularity. The different levels granularity for spatial data are global (g) and regional (r) and that for temporal data are annual (a), seasonal(s), and monthly (m), as shown in Table 4.1. Additionally, there are three output variables for each model. *Similarity* among models is the other facet which can be classified based on the following perspectives: i) pairwise: in this case scientists are interested in observing similarity between each pair of models and

ii) multi-way: in this case scientists are interested in observing similarity among all models taken together, and iii) one-to-many: in this case scientists might choose one model as a reference. Our collaborators revealed that the third option is rarely used in comparison of TBMs, since no model is known a priori to be any better as a "reference" than any other. As a result we did not implement this option in the tool.

4.4 Visualization Tasks and Design

The next step in our study was to connect the intents and facets though concrete visualization tasks and subsequently translate the tasks to visualization design. This led to our second contribution: a classification scheme for integrating tasks, facets, and design (Table 4.1).

4.4.1 Tasks

For identifying the tasks, we took inspiration from Zhou and Feiner's taxonomy [101], among which *identify*, *compare*, *associate*, and *distribution* are relevant here. Notably, the transition from $\mathbf{Q_1}$ to $\mathbf{Q_4}$ also indicates increasing complexity of the visualization tasks, which we describe below. In Table 4.1 the abbreviation after task name indicates the facet they operate upon.

4.4.1.1 Identify

The intent $\mathbf{Q_1}$, that is understanding model-model similarity is reflected in SimilarityExplorer by three variants of the identification tasks: finding the degree of spatial correlations among models ($\mathit{identify}(p)$), finding the degree of temporal correlation among them ($\mathit{identify}(t)$), and also finding the degree of overall spatiotemporal correlation ($\mathit{identify}(p,t)$). While the first two tasks reflect pairwise similarity, the last one expresses multi-way similarity. In Table 4.1, the symbol / reflects an OR operation. So in case of the identification tasks any granularity of space (g/r) and time (a/m/s) can be selected using different filters.

4.4.1.2 Compare

The intent $\mathbf{Q_2}$, that is understanding output-output similarity is reflected in SimilarityExplorer by the comparison tasks: comparing the degree of spatial correlation ($\mathit{compare}(p,v)$) and temporal correlation ($\mathit{compare}(t,v)$) among multiple output variables. These tasks can involve multiple selections of granularity of space and time indicating an AND operation as shown by the *comma* (g,r and a,s,m). For example, global correlation of models with respect to one output variable can be compared with the regional correlation.

4.4.1.3 Associate

The intent $\mathbf{Q_3}$ involves combining the understanding of similarity by analyzing the region-wise anomalies and trends for the models. This task applies to both spatial ($associate(p)$) and temporal correlation ($associate(t)$) for which different views are instantiated. These involve mainly drill-down and brushing operations and are performed at the regional granularity of space and monthly or seasonal granularity of time.

4.4.1.4 Distribution

The intent $\mathbf{Q_4}$ is reflected by the distribution task that helps provide a multi-way perspective on behavior of regions with respect to multiple models ($distribution(p, v)$), and on pairwise model-model relationships for all regions. Scientists could also get additional information about outlying regions and models using this task, which allows exploration at a greater level of detail than the other tasks. This task also involves drilling down to the temporal distribution of a pair of models ($associate(t, v)$).

4.4.2 Visual Encoding Challenges

The challenges in translating the tasks to different aspects of visual encoding were met by integrating the iterative feedback from the scientists' on our intermediate prototypes. We justify our key design choices with respect to the following aspects.

4.4.2.1 Separating Space and Time

The tasks described above required us to separate as much as possible, the facets of space and time, although in the final analysis, they are inextricably linked. A climate scientist remarked that he wanted *no time* in his analysis, but wanted to see only space. Upon reflection, we realized that what this user really wanted was more like *all time*, i.e., spatial correlations which had been composited over the entire time interval, with no temporal subsetting. In this sense, then, the spatial correlations shown are composited over time, and the temporal correlations are composited over space. This had to be reflected in the visual representation by having a separation between spatial and temporal encodings.

4.4.2.2 Facilitating Systematic Interaction

Both spatial and temporal relationships could vary over space (e.g., regions) and time (e.g., seasons). This decomposition needed to be reflected through brushing over space and time and selections of regions and time-steps. These operations also had to be associative: any spatial operation could adapt the temporal similarity to reflect the selected region and any temporal operation could adapt the spatial side to represent the correlation for a particular time step. Another role of interaction was to allow scientists explore different granularity of space and time. This was facilitated by interaction operations such as filtering and drill-down to additional views showing different levels-of-detail.

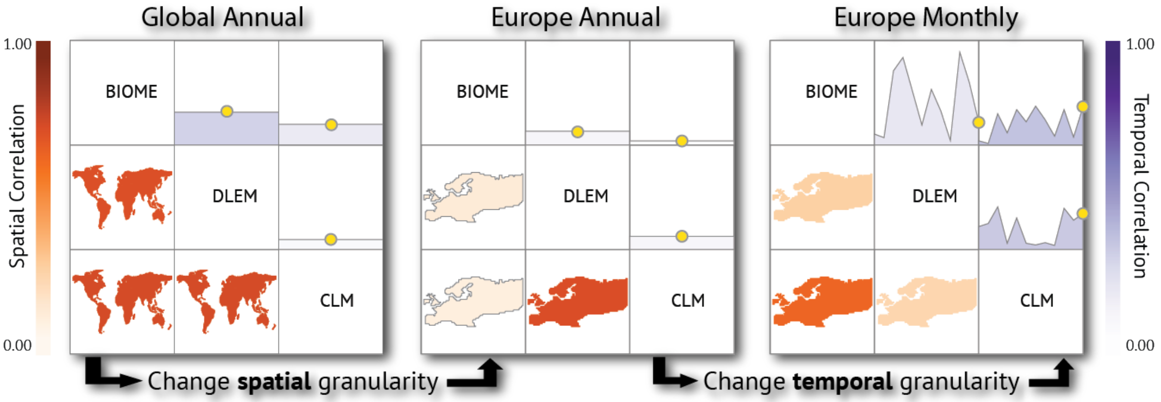


Figure 4.3: **Preserving the mental model and symmetry about spatial and temporal similarity** through use of *maps* for representing space and use of *area graphs* for representing time, and by reflecting the change in granularity on both sides.

4.4.2.3 Preserving the Mental Model

This was a critical design issue due to the interplay between space and time, and the need to *associate* them in a holistic view [87]. Both geographical maps and time-series could be used to represent variation of either the spatial or temporal correlation. In one of the interactive sessions we presented mock-ups that used time-series to represent the variation of both spatial and temporal correlation. But without consistent visual cues linking the representation to space or time, they were confused:

> *"I like this but I have to wrap my head around what the visualization is telling me: is it space or is it time? It will be much better if I don't have to process this in my mind."*

We resolved this issue by collectively taking a design decision: for temporal correlation we would display the variation of the correlation over time by displaying a time-series that adapts to the temporal granularity (annual, months, seasons). On the spatial side, we would display maps showing spatial correlation for the selected time step. Thus we use consistent spatial cues in the form of maps and temporal cues in the form of time series (Figure 4.4(b)(d)). By brushing over time, we would see the change in spatial correlation as the displayed map adapts to the selected time step.

4.4.2.4 Retaining symmetry while drilling down

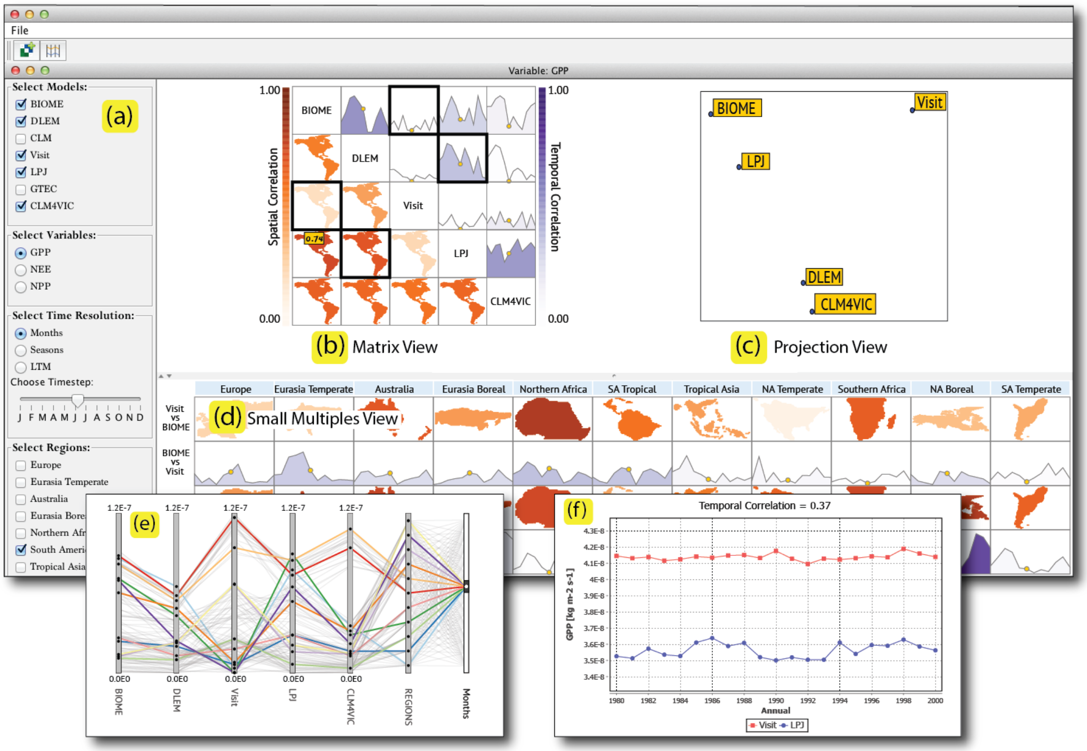
Preserving a symmetrical relationship among the different granularity of space and time through consistent visual representation was essential for scientists to keep track of any change that occurred. The change of spatial granularity is reflected by transforming the maps to represent the selected regions. The change of temporal granularity is reflected by transforming the number of steps in a time series (Figure 4.3).

4.4.3 Comparison methods

Facilitating visual comparison among the models and output variables is one of the main goals of this work. We followed Gleicher et al.'s taxonomy [48] of visual comparison methods for guiding the representation of the different aspects of similarity and the eventual placement of the different views. As shown in Table 4.1, the three comparison methods that are used are explicit encoding, juxtaposition and superposition. Explicit encoding is used to encode the degree of similarity among the different views with the help of correlation metrics. For comparison tasks multiple views are juxtaposed next to each other. We represent multiple time series by superposing them in the same view (Figure 4.4(f)). Different interaction mechanisms like filtering, brushing, linking, and drilling-down allow scientists to browse through the multiple perspectives of similarity.

4.5 SimilarityExplorer

Our third contribution is the design of the SimilarityExplorer, an exploratory visualization tool for analyzing multifaceted, multi-granularity, climate model similarity. This design was guided by: the domain characterization presented in Section 4, and the classification scheme described in Section 5. The scientists' analysis needs motivated our design decision of using multiple linked views [102], a visualization approach



Select Models:
☑ BIOME
☑ DLEM
□ CLM
☑ Visit
☑ LPJ
□ GTEC
☑ CLM4VIC

Select Variables:
● GPP
○ NEE
○ NPP

Select Time Resolution:
● Months
○ Seasons
○ LTM

Choose Timestep:
J F M A M J J A S O N D

Select Regions:
□ Europe
□ Eurasia Temperate
□ Australia
□ Eurasia Bore
□ Northern Afr
☑ South Ameri
□ Tropical Asia

Figure 4.4: **SimilarityExplorer is composed of a set of filters (a), similarity views (b, c, d) and data views (e, f).** The *similarity views* are **(b)** a matrix view for showing pairwise similarity, **(c)** a projection view for showing multi-way similarity, and **(d)** a small multiples view for showing region-wise spatiotemporal similarity. The *data views* are: **(e)** a parallel coordinates view for showing multi-model distribution of each variable, and **(f)** a time series for showing temporal distribution of any pair of models.

that is appropriate for flexible analysis of multifaceted data. There is an implicit hierarchy [103] in the type of views in SimilarityExplorer, which are *similarity views* and *data views*.

4.5.1 Similarity Views

With the help of similarity views, we explicitly encoded spatial and temporal correlation between models, based on the computation we had described in Section 4.2.2. The different similarity views are described below:

4.5.1.1 Matrix View

A model is a primary unit of comparison. Our collaborators needed a view that would show both spatial and temporal correlation for the models in one integrated view, that would be flexible enough to adapt to different granularity of space and time. We took inspiration from the multi-form matrix [104] designed by MacEachren et al. and designed a matrix view that reflects pairwise similarity between models (Figure 4.3). In keeping with the idea of preserving the mental model about space and time, it is divided into two halves across the diagonal: the cells in the lower triangle represent the pairwise spatial correlation through color-coded maps and the cells in the upper triangle represents the temporal correlation between two models. The color coding uses a continuous color map [45] and reflects the degree of correlation, with orange for correlations on the spatial side and purple for correlations on the temporal side. The color map adapts to the range of correlation values: if there are negative correlations, a divergent color map is used.

Scientists can perform the following tasks using the matrix as shown in Table 4.1: i) *identification* tasks by filtering the view by different regions or time and ii) comparison tasks launching multiple matrices of different variables (Figure 4.6). For the latter case, we could have encoded a derived statistic that would explicitly encode the average correlation based on multiple variables, in a single matrix. However, the scientists were interested in analyzing the high or low correlations for the individual variables. Thus we use the option of juxtaposing multiple matrices for the different variables.

The effect of changing spatial and temporal granularity are shown in Figure 4.3. The initial view is for showing global, annual correlation. On selection of a sub-region, *i.e.*, Europe, maps for Europe are shown on the spatial side, while the temporal side gets updated to show the annual average correlation for Europe. On selection of seasonal granularity, the area graph gets updated to a time-series representing the four seasons and shows the maps for the selected season. Thus spatial and temporal operations are symmetrical: they affect both sides of the matrix and the color-coding reflects the correlation for the selected time step.

4.5.1.2 Projection View

After presenting the matrix view to our collaborators, they felt the need for representation which gave a high-level overview of *all* models with respect to each other. This prompted us to design the projection view (Figure 4.4(c)) that shows multi-way similarity among models. Thus, it overcomes the limitation of the matrix view, which is only able to show pairwise patterns. As mentioned in Table 4.1, the projection view is used to mainly *identify* which models are more similar, triggering the subsequent analysis steps for exploring the reason for similarity. The projection view is generated by using the spatial or temporal correlation between models as the distance metric and then using multidimensional scaling (MDS) for mapping the data points onto a two-dimensional scatter plot. The physical proximity of models encodes their overall similarity. Initially, some of our collaborators were confused by the projection view but on seeing the merits of getting a multi-way overview of similarity they became more appreciative of its utility. One of them commented:

> *"The axes have no meaning here and we are not used to seeing this, but I really like the all-way comparison we can perform which we could not do before."*

This view adapts to different selections of time steps or regions.

4.5.1.3 Small Multiples View

The small multiples [49, 9] view as shown in Figure 4.4(d) supports drilling down into the correlation patterns for each individual region. The drill down operation can be initiated from both the spatial and temporal sides of the matrix: drill down from the spatial side shows a map representing spatial correlation for a region and a selected time step; and that from the temporal side shows time series representing variation of temporal correlation for a region. One of the design options was to show a global map for the spatial drill down, with individual regions being color-coded based on spatial correlation between two models. However, this would not have been symmetrical with the temporal side, as there would be a map for each time-step and it is visually complex to represent so many maps, and still preserve the mental model about the relationships.

Using this small multiples view, scientists can perform several comparisons: i) by selecting a cell within a matrix the region-wise spatial and temporal correlation for that pair is shown, which lets them compare anomalies between global and regional patterns, ii) by comparing across space and time, scientists can understand the cause of anomalies , and iii) by comparing these small multiples for different variables, scientists can hypothesize about which output variables affect similarity of models across different regions.

4.5.2 Data Views

Using the data view scientists can drill down to the distributions of different variables and gain information about outliers which the similarity views might not show. Below we describe the data views:

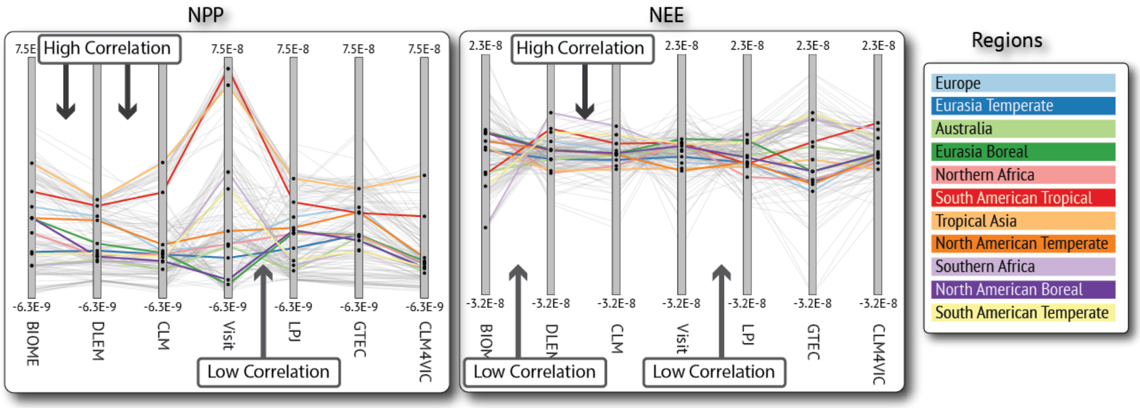


Figure 4.5: **Data View: Parallel Coordinates**. The ability to examine the region-wise range and distribution of variables enables climate scientists to relate the meta views to the patterns in the data view, i.e., parallel coordinates, and additionally, find clusters and outliers. For `NPP`, we can see a cluster of polylines for the regions South American Tropical and Tropical Asia for all models, indicating multi-model similarity for those regions.

4.5.2.1 Parallel Coordinates

For each output variable, we use parallel coordinates (Figure 4.5) for enabling scientists to analyze the multi-model similarity based on the region-wise distribution of the variable. In discussion with the scientists, we found that multivariate relationships among the different output variables are not of interest in their analysis. Instead of modeling parallel coordinates conventionally, where variables are mapped on to the vertical axes and data objects are mapped to polylines, we use one parallel coordinates plot per variable. We use each vertical axis to represent a model and a polyline connecting the different axes represents the value of a variable for a given region. We compute a global scale across all models, for mapping the values so that they are comparable. The regions are represented by a categorical color scale. The number of data points, that is the number of polylines, depends on the temporal granularity selected. For annual correlation, there is only one polyline per region, for seasons there are four, and in the case of the lowest level of temporal granularity, months, there are twelve polylines for each region.

Brushing by time and region allows the scientists to look at only specific instants of time, a few regions, or both. By observing the trajectory of polylines, scientists could perform a multiway comparison of region-wise distribution across models. By linking the parallel coordinates with the matrix view, they can also *associate* the degree of correlation among models with the data distribution across the different regions. In

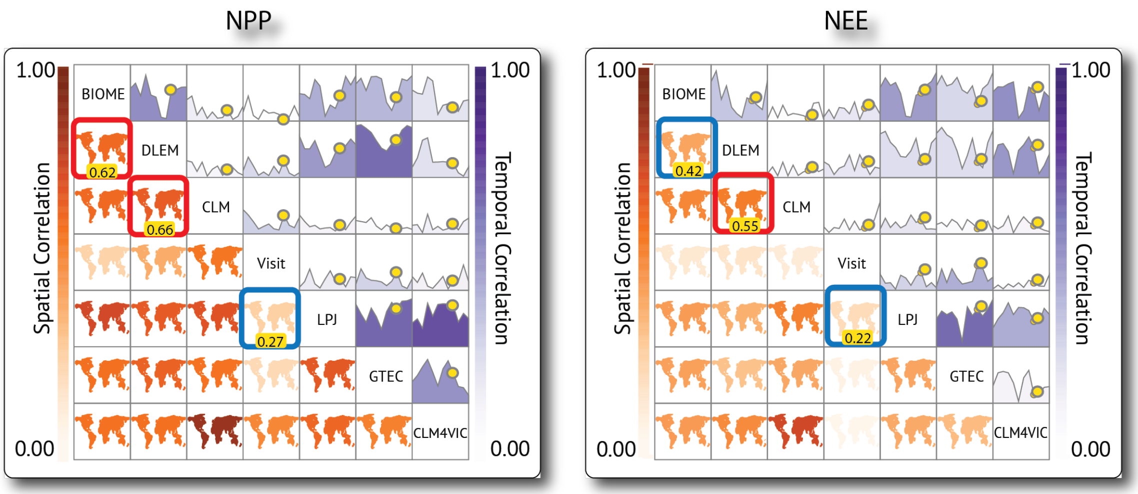


Figure 4.6: **Comparing multiple output variables** for different months and analyzing their distribution ($\mathbf{Q_2}$, $\mathbf{Q_4}$).

case of comparison of multiple variables, multiple parallel coordinates plots can be instantiated.

4.5.2.2 Time Series

The temporal correlation represented by the area graph in the matrix is based on a pair of time-series for each time-step. Since correlation is just one of the ways of representing the relationship between two time-series, the scientists were also interested in looking at the original time-series to find any additional information, like the high or low temporal distribution, or any anomalies. Based on this requirement, we designed a time series view that shows the temporal distribution of any variable for a pair of models. The view is instantiated when any cell on the temporal side of the matrix is selected (Figure 4.4(f)).

4.6 Case Studies

We describe the features of the SimilarityExplorer with two different scenarios that our climate scientist collaborators used for analyzing model similarity.

4.6.1 Understanding Output-Output Similarity ($\mathbf{Q_2}$,$\mathbf{Q_4}$)

The climate scientists wanted to compare how models behave with respect to two output variables: Net Primary Productivity (NPP) and Net Ecosystem Exchange

(NEE) for the month of September. Considered to be two of the most important "vital statistics" of ecosystems, NPP represents the amount of productivity that is available for growth, while NEE reflects the input/output balance of carbon to and from the ecosystem. Both output variables are critical for understanding the atmospheric carbon cycle. As shown in Figure 4.6, all the models seemed to be more spatially correlated with respect to NPP (on the top) than NEE (on the bottom). This prompted the scientists to look at the region-wise distribution of the variables for confirming this. The parallel coordinates plot for NPP (Figure 4.5, on the left) showed a high number of parallel lines between highly correlated models like BIOME-DLEM and DLEM-CLM. But the high correlation for BIOME-DLEM is absent for NEE (Figure 4.5, on the right), where lines are more scattered in different directions, reflecting the different input/output balance points for carbon across ecosystems in different regions. By using parallel coordinates plot, the scientists found that NPP (Figure 4.5, on the left) shows higher spread among the values than NEE (Figure 4.5, on the right). The high spread and high values of NPP for the Visit model appear to be outliers. The scientists concluded that these outlying regions were causing the Visit model to be quite different from the rest. This can also be seen in the matrix plots, by the consistently low spatial correlation between Visit and most of the other models, for both variables. However, for NEE, the distribution for Visit is identical to the distribution for the other models: in this case the lack of correlation causes Visit to be different from the rest.

The outlier regions, Tropical Asia and South American Tropical, appeared to be similar for all the models, as shown by the clustered polylines for NPP. The scientists confirmed that this was an expected pattern for tropical regions for NPP; such a pattern was expected to be absent for NEE, which was also confirmed by observing the parallel coordinates plot.

By using SimilarityExplorer the climate scientists were thus able to discover that the models had better agreement for tropical areas where there is little seasonality in growing conditions, like temperature. The models had lower agreement for temperate and boreal ecosystems that have distinct and more variability in growing conditions. 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."*

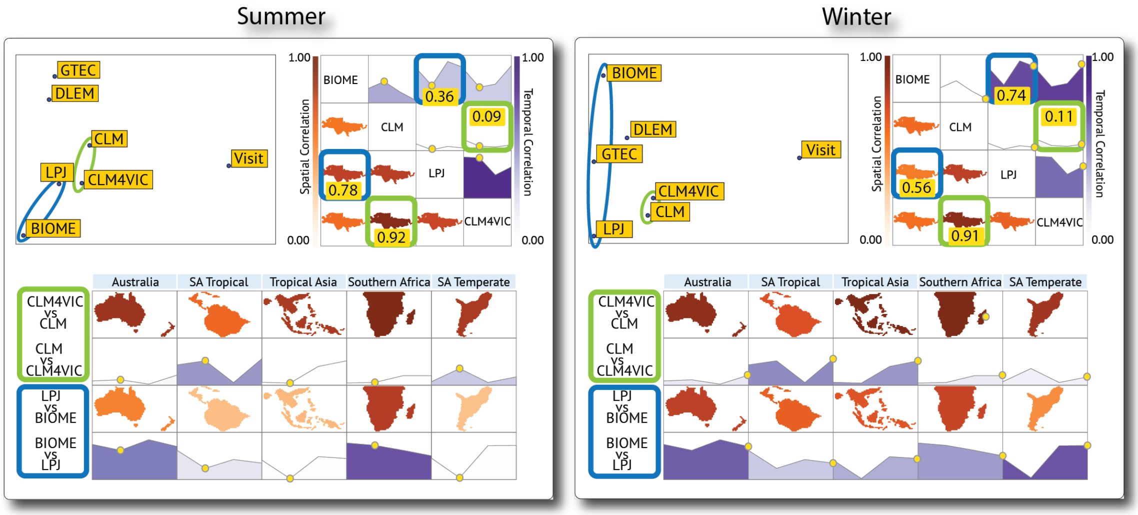


Figure 4.7: **Comparing model similarity for GPP and analyzing spatiotemporal anomalies for winter and summer (Q1, Q3).** Using the projection view, scientists were able to select similar models; using the matrix view they could compare spatial and temporal correlation (indicated by the numbers); and identify anomalies using the small multiples view.

4.6.2 Exploring Model-Model similarity (Q_1, Q_3)

Gross Primary Productivity (GPP) is arguably the most important ecosystem variable, indicating the total amount of energy that is fixed from sunlight, before respiration and decomposition. Climate scientists need to understand patterns of GPP in order to predict rates of carbon dioxide increases and changes in atmospheric temperature. The motivation for this scenario was to compare multiple models with respect to GPP by exploring model similarity for the Europe and Eurasia sub-regions; for the summer and winter seasons, and compare those trends with the correlations for tropical and temperate regions. As shown in the summer view in Figure 4.7, the model pairs of CLM-CLM4VIC and BIOME-LPJ appear to be similar, based on their relative proximity in the projection view. They selected these models and instantiated the matrix view (Figure 4.7). This showed high spatial correlation but low temporal correlation for the CLM-CLM4VIC model pair for summer, as well as for winter season. For comparing the trends with the temperate and tropical regions, they used the small multiples view. The notable deviations were i) SA tropical which showed higher temporal correlation across summer and winter for this model pair, and ii) Tropical Asia which showed higher temporal correlation than Europe and Eurasia sub-regions for the winter season. For the BIOME-LPJ pair, the models appeared to be more similar during summer than winter based on the projection view. The drop in spatial correlation during winter was confirmed by the matrix views. However, the temporal correlation was higher in winter than during summer. From the small multiples view, the scientists found that during summer the `SA Tropical`, `Tropical Asia` and `SA Temperate` regions had lower spatial correlation than `Europe` and `Eurasia` sub-regions; while `Tropical Asia` and `SA Temperate` had lower temporal correlation compared to the same. Both spatial and temporal correlation for this model pair seemed to increase for the winter season for the `SA Tropical`, `Tropical Asia` and `SA temperate` region. This trend was contrary to the pattern for the `Europe-Eurasia` region.

By using SimilarityExplorer the climate scientists were able to visualize the interdependency between seasonality, region, and model. The fact that the SimilarityExplorer made their analysis more streamlined and efficient was validated from one of their comments:

> "*Without this tool scientists would literally print hundreds of plots and pin them on the wall, this tool solves this problem.*"

They also appreciated the fact that the tool can be easily extended for more models, the benefit is being able to do this with 20 models.

4.7 Summary

In this chapter, we have presented SimilarityExplorer, a visual analysis tool for comparison of multifaceted climate models. Climate scientists are naturally more familiar and comfortable working in one of the two facets of space and time than the other. Most of their exploratory thinking, tools and analyses tend to be biased toward one of them, at the expense of investigations into the other. Because of the relative ease with which users can 'cross the diagonal' from one realm of analysis to the other, the scientists found that "*the SimilarityExplorer offset such natural prejudices and facilitated commensurate symmetry, resulting in more complete exploration and understanding*".

Even with the SimilarityExplorer tool, it requires a lot of time to explore the complete parameter space or perform the analysis from multiple perspectives. Given those limitations, in the next chapters we present two *Visual Analytics Approaches* to detect patterns more efficiently and guide the users in the exploration process.

Chapter 5

Visual Reconciliation of Alternative Similarity Spaces in Climate Modeling

As we mentioned in Chapter 4, the SimilarityExplorer tool makes easier for climate scientists to explore model relationships from multiple perspectives. However, it is cumbersome and often impractical to explore all the options/parameters in order to identify patterns. To deal with this problem, we need to make use of *Visual Analytics Approaches* to automatically identify potential patterns and make use of user's expertise to refine and filter the most interesting patterns. In this chapter we deal with the problem of making sense of alternative ways to create groups based on different descriptors of climate model.

Grouping of data objects based on similarity criteria is a common analysis task. In different application domains, computational methods such as clustering, dimensionality reduction, are used for extracting groupings from data. However, in the real world, with the growing variety of collected and available data, group characterization is no longer restricted to a single set of criteria; it usually involves alternative sets. Exploring the inter-relationship among groups defined by such alternative similarity criteria is a challenging problem. For example, in health care, an emerging area of research is to reconcile patient similarity based on their demographics with that based on their disease history, for targeted drug development [105]. In climate science, an open problem is to analyze how similar outputs from model simulations can be linked with similarity in the model structures, characterized by diverse sets of criteria. Analyzing features of model structures and their impact on model output, can throw light into important global climate change indicators [106].

Redescription mining algorithms have been developed for quantifying and exploring relationships among multiple data descriptors [107]. These techniques have focused on mining algorithms for binary data, where objects are characterized by the presence or absence of certain features. Group extraction based on such computational methods are heavily influenced by parameter settings. Also, it usually takes multiple iterations to find an adequate solution; and in most cases, only approximate solutions can be found. Domain experts need to be involved in this iterative process, utilizing their expertise for controlling the parameters. This necessitates a visual analytics approach towards user-driven group extraction and communication of relationships among the groups, which are characterized by diverse descriptive parameters.

To achieve this goal, we introduce a novel visual analytics paradigm: *visual reconciliation*, which is an iterative, human-in-the-loop refinement strategy for reconciling alternative similarity spaces. The reconciliation technique involves synergy among computational methods, adaptive visual representations, and a flexible interaction model, for communicating the relationships among the similarity spaces. While iterative refinement strategies are not new in visual analytics [108, 109], sense-making of diverse characterization of data spaces is still an emerging area of research [110]. In this context, we introduce the problem of reconciling the characteristics of any data object with respect to alternative similarity spaces, which in this case comprise of boolean and time-varying attributes. The strength of the reconciliation model stems from transparency in presentation and communication of the similarity relationships among diverse data descriptors, with minimal abstraction, and effective visual guidance through visual cues and direct manipulation of the data. The design and interactions are motivated by domain experts' need for visual representations with high fidelity, and a simple yet effective interaction mechanism for browsing through the parameters.

Our concept of visual reconciliation is grounded in our experience of collaborating with climate scientists as part of the MsTMIP [1] project. An open problem in climate science research is to analyze the effect that similarity and differences in climate model structures have on the temporal variance in model outputs. Recent research has shown model structures can have significant impact on variability of outputs [111], and that, some of these findings need to be further investigated in details for exploring different hypotheses. To facilitate the scientific analysis process, we propose an analysis paradigm for reconciling alternative similarity spaces, that leverages the high bandwidth of human perception system and exploits the pattern detection and optimization capabilities of computing models [112, 113].

[1] MsTMIP: Multi-Scale Synthesis and Terrestrial Model Intercomparison Project

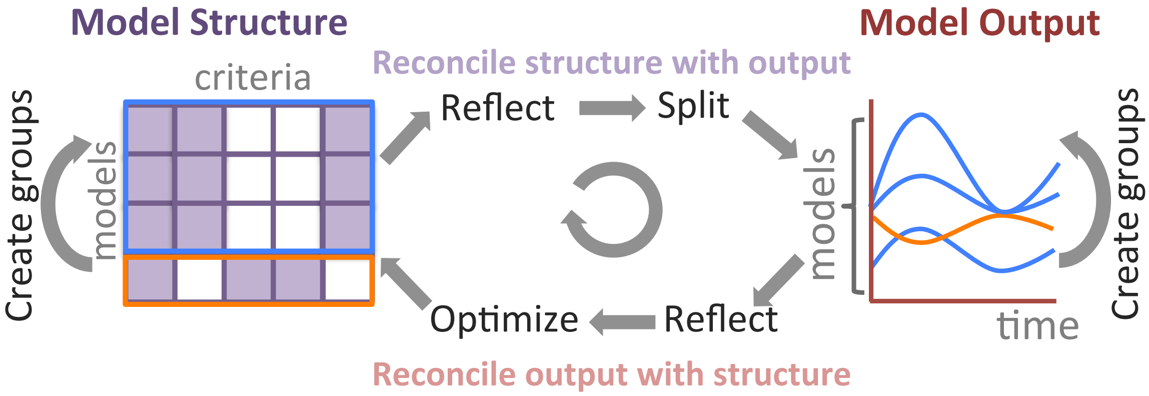


Figure 5.1: **Conceptual model of visual reconciliation** between binary model structure data and time-varying model output data. Iterative creation of groups and derivations of relationship between output similarity and importance of the different model structure criteria. Blue and orange indicate different groups of models.

The key contributions of this work stems from a visual reconciliation technique (Figure 5.1) that i) helps climate scientists understand the dependency between alternative similarity spaces for climate models, ii) facilitates iterative refinement of groups with the help of a feedback loop, and iii) allows flexible multi-way interaction and exploration of the parameter space for reconciling the importance of the model parameters with the model groupings.

5.1 Motivation

Why do we need a new visual analytics technique? Reconciling alternative similarity spaces is challenging on several counts: i) Data descriptors can comprise of different attribute types. From a human cognition point-of-view, reconciling the similarity of climate models across two different visual representations is challenging. There needs to be explicit encoding of similarity [48] that helps in efficient visual comparison and preserve the mental model about similarity. Adaptation of similarity needs to be reflected by dynamic linking between views without causing change blindness; ii) For aligning two different similarity spaces, say computed by two clustering algorithms, we will in most cases get an approximate result. The result will need to be iterated upon with subsequent parameter tuning to achieve higher accuracy. This necessitates iteration, and therefore a human-in-the-loop approach; iii) Domain experts need to *trust* the methodology working at the back-end and interact with parameters for understanding their importance. Fully automated methods do not allow that flexibility. Thereby, a transparent representation with minimal abstraction is necessary where parameters in similarity computation can be influenced by user selections and filters.

As mentioned before, the technique is not restricted to climate models, but for simplifying our discussion in this chapter we specifically discuss the applicability of the visual reconciliation technique in the context of climate modeling.

5.1.1 Problem Characterization

Each TBM is defined by the different input parameters for characterizing ecosystem processes and outputs that quantify the dependency between the carbon cycle and the ecosystem processes. In the context of this chapter, each model has a dual representation of a weighted collection of criteria or descriptive parameters, and time-series for different outputs, for different regions.

Model Structure. Model structure refers to the types of processes considered (e.g., nutrient cycling, lateral transport of carbon), and how these processes are represented through different criteria (e.g., photosynthetic formulation, temperature sensitivity, etc.) in the models. A model simulation algorithm can have different implementations of these processes. These implementations are different from each other due to the presence or absence of the different criteria, that control the specific process. For example, if a model simulates photosynthesis, a group of criteria like `simulating carbon pools`, influence of `soil moisture`, and `stomatal conductance` can be either present or absent. Currently, climate scientists do not have an objective way of choosing one set of criteria over other, that can influence the output. A model structure is a function of these criteria. If there are n criteria, there can be 2^n combinations of this function. In our data, there are 4 different classes of criteria, for energy, carbon, vegetation, and respiration; with each class comprising of criteria, which are about 20 to 30 in number.

Model Output. Model simulation outputs are ecosystem variables that help climate scientists predict the rates of carbon dioxide increases and changes in the atmosphere (*e.g.*, `GPP`, `NPP`, `NEE`).

Relationship Between Model Structure and Output. In the previous chapter, we introduced the SimilarityExplorer (Chapter 4) for analyzing similarity and differences among multifaceted model outputs. Despite the standardized protocol used to derive initial conditions, models show a high degree of variation for `GPP`, which can be attributed to differences in model structural information [111].

Therefore, one of the open research questions in the TBM domain is how similarity or differences in model output can be correlated with that in model structures. The heterogeneity of model structure and model output data makes it complex to derive one-to-one relationships among them. Currently, in absence of an effective analysis technique, scientists manually browse through the theoretically exponential number of model structure combinations, and analyze their output. This process is inefficient and also ineffective due to the large parameter space which can easily cause important patterns to be missed.

In the visual reconciliation technique, we provide a conceptual framework that enable scientists to reconcile model structural similarity with output similarity. We focus on using visual analytics methods for addressing the following high-level analysis questions: i) given all other factors are constant, analyze how different combination of parameters within model structure cause similarity or difference in model output, and ii) by examining time-varying model outputs at different regions, understand which combination of parameters cause the same clusters or groups in model structure.

5.1.2 Visual Reconciliation Goals

As illustrated in Figure 5.1, the visual reconciliation technique enables climate scientists to: i) analyze model structure and use that as feedback for reconciling similarity or differences in model output, and ii) analyze model output and use that as a feedback for comparing similarity or differences in model structure. The reconciliation model focuses on three key goals:

Similarity Encoding and Linking. For providing guidance on choosing the starting points of analysis, the visual representations of both structure and output encode similarity functions. Subsequently, scientists can use those initial seed points for reconciling structure characteristics with output data, or conversely, for reconciling output data with structure characteristics.

Flexible Exploration of Parameters. The visual feedback and interaction model adapts to the analysts' workflow. Scientists can choose different combinations of parameters, customize clusters on model structure and model output side and accordingly the visual representations change, different indicators of similarity are highlighted.

Iterative Refinement of Groups. By incorporating user feedback in conjunction with a computation model, the reconciliation technique allows users to explore different group parameters in both data spaces and iteratively refine the groupings. The key goal here is to understand, which criteria in model structures are most important in determining how the outputs are similar or different over time.

5.2 Related Work

We discuss the related work in the context of the following threads of research: i) automated clustering methods for handling different data descriptors, and visual analytics approaches towards user-driven clustering, ii) integration of user feedback for handling distance functions in the context of high-dimensional data, and iii) visual analytics solutions for similarity analysis of climate models.

5.2.1 Clustering Methods

Different clustering methods have been proposed for dealing with alternative similarity spaces. Pfitzner et al. proposed a theoretical framework for evaluating the quality of clusterings through pairwise estimation of similarity [114]. The area of multi-view clustering [115] analyzes cases when data can be split into two independent subsets. In that case either subset is conditionally independent of each other and can be used for learning. Similarly, authors have proposed approaches towards combining multiple clustering results into one clustered output, using similarity graphs [116]. Although we are also dealing with multiple similarity functions, the goal is to reconcile one with respect to the other.

In this context, the most relevant research in data mining community looks into learning the relationship between different data descriptor sets. The reconciliation idea is similar, in principle, to redescription mining which looks at binary feature spaces and uses automated algorithms for reconciling those spaces [117, 107]. While redescriptions mostly deal with binary data, we handle both binary data and time-varying data in our technique.

Our work is also inspired by the consensus clustering concept, which attempts to find the consensus among multiple clustering algorithms [118] in the context of gene expression data. Consensus clustering has also been applied in other applications in biology and chemistry [119, 120]. In our case, while we are interested in the consensus between similarity of model structure and model output, we also aim at quantifying and communicating the contribution of the different parameters towards that consensus or the lack thereof.

We adopt a human-in-the-loop approach, as automated methods do not provide adequate transparency with respect to the clustering parameters, and also in most cases, iteration is necessary to present reconciliation results. Iterative refinement strategies for user-driven clustering have been proposed for interacting with the intermediate clustering results [108] for tuning parameters of the underlying algorithms [109], and for making sense of dimension space and item space of data [110]. Dealing with diverse similarity functions and at the same time providing a high fidelity visual representation to domain experts which can be interactively refined, are the key differentiators of our work. The reconciliation workflow follows an adaptive process, where the groupings on the model output side are used as an input to the model structure side for: i) providing guidance to the scientists towards finding similar groups with respect to diverse descriptors or criteria, and ii) understanding the importance of criteria, which is handled by an underlying optimization algorithm.

5.2.2 User Feedback for Adaptive Distance Functions

Recently, there has been a lot of interest in the visual analytics community for investigating how computation and tuning of distance functions can be steered by user interaction and feedback. Gleicher proposed a system called Explainers that attempts to alleviate the problem of multidimensional projection, where the axes have no semantics, by providing named axes based on experts' input [121]. Eli et al. presented a system that allows an expert to interact directly with a visual representation of the data to define an appropriate distance function, without having to modify different parameters [122]. In our case, the parameter space is of key interest to the user; therefore we create a visual representation of the parameters, and allow direct user interaction with them. Our user feedback mechanism based weighted optimization method is inspired by the work on manipulating distance functions by Hu et al. [123]. However, the interactivity and conceptual implementation is different, since we are working with two different data spaces, without using multidimensional projections. The modification of distance functions have also been used for spatial clustering, where user selected groups are given as input to the algorithm [124]. Our reconciliation method is similar, in principle to this approach, where the system suggests grouping in one data space, based on the same in other space, by a combination of user selection and computation.

5.2.3 Visual Analytics for Climate Modeling

Similarity analysis of model simulations is an emerging problem in climate science. While visual analysis of simulation models and their spatiotemporal variance have received attention in other domains[125, 126], current visual analytics solutions for climate model analysis [93] mostly focus on xaddressing the problem at the level of a single model and understanding its spatiotemporal characteristics. For example, Steed et al. introduced EDEN [94], a tool based on visualizing correlations in an interactive parallel coordinates plot, focused on multivariate analysis. Recently, UV-CDAT [97] has been developed which is a provenance-enabled framework for climate data analysis. However, like most other tools, UV-CDAT does not support multi-model analysis [127]. To fill this gap, in Chapter 4 we described the SimilarityExplorer tool which was developed to analyze multi-model similarity with respect to model outputs. In this case, we are not only comparing multiple models, but also comparing two different data spaces: model structure and model output. Climate scientists have found that different combinations of model structure criteria can potentially throw light into different simulation output behavior [111]. However, to the best of our knowledge, no visual analytics solution currently exists in climate science to address this problem. For developing a solution, formulating an analysis paradigm precedes tool development because of the complexities involved in handling multiple descriptor spaces. Although there has been some work on hypothesis generation [91] and task characterization [92] for climate science, they are not sufficient for handling the reconciliation problem involving alternative similarity spaces.

5.3 Coordinated Multiple Views

An important component of the visual reconciliation technique is the interaction between multiple views [102] of similarity spaces. In this case we have binary model structure data and time-varying model output data. As we had shown in Figure 5.1, the goal is to let domain scientists create and visualize groups on both sides, and understand the importance of the different criteria in creating those groups. In this section we provide an overview of the different views and describe the basic interactions between those.

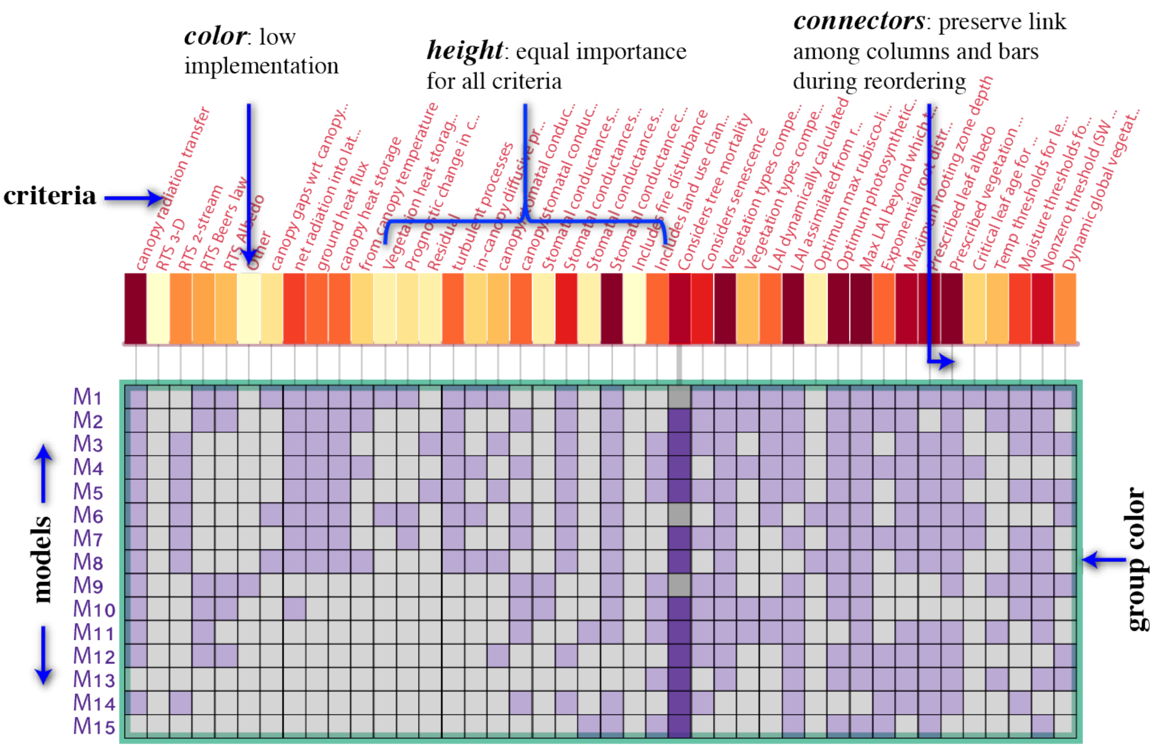

color: low implementation
height: equal importance for all criteria
connectors: preserve link among columns and bars during reordering
criteria
canopy radiation transfer
RTS 3-D
RTS 2-stream
RTS Beer's law
RTS Albedo
Other
canopy gaps wrt canopy...
net radiation into lat...
ground heat flux
canopy heat storage
from canopy temperature
Vegetation heat storag...
Prognostic change in c...
Residual
turbulent processes
in-canopy diffusive pr...
canopy stomatal conduc...
canopy stomatal conduc...
Stomatal conductances...
Stomatal conductances...
Stomatal conductances...
Stomatal conductancec...
Includes fire disturbance
Includes land use chan...
Considers tree mortality
Considers senescence
Vegetation types compe...
Vegetation types compe...
LAI dynamically calculated
LAI assimilated from r...
Optimum max rubisco-li...
Optimum photosynthetic...
Max LAI beyond which t...
Exponential root distr...
Maximum rooting zone depth
Prescribed leaf albedo
Prescribed vegetation ...
Critical leaf age for ...
Temp thresholds for le...
Moisture thresholds fo...
Nonzero threshold (SW ...
Dynamic global vegetat...
models
M1
M2
M3
M4
M5
M6
M7
M8
M9
M10
M11
M12
M13
M14
M15
group color

Figure 5.2: **Matrix view for model structure data**: Rows represent models and columns represent criteria. The variation of average implementation of a criterion for all models is shown by a color gradient from light yellow to red, with red signifying higher implementation. In the default view, all criteria have equal importance or weights, indicated by the heights of the bars. Connectors help visually link the columns and bars when they are reordered independently.

5.3.1 Matrix View

To display the model structure data, which is a two-dimensional matrix of 0's and 1's, we use a color-coded matrix Figure 5.2, which serves as a presence/absence representation of the different criteria for the model structure. This is inspired from Bertin's reorderable matrix [35] and the subsequent interactive versions of the matrix [128].

Since the data is binary, we use two color hues: purple for denoting presence and gray for absence. Visual salience of a matrix depends on the order of the rows and columns and numerous techniques have been developed till data fore reordering [129, 130] and seriation [131]. In this case, the main motivation is to let the scientists visually separate the criteria which have high average non-implementation (indicated by 0's) and those with high average implementation. For providing visual cues on potential groups within the data, we reorder the rows and columns, based on a function that puts the criteria, that are present, to the upper left of the matrix; and pushes those that are absent, to the bottom right.

The colored bars on top of the matrix serve a dual purpose. The heights of the bars indicate the importance or weight of each criteria for creating groups in model structure. The colors of the bars, with a light yellow to red gradient indicate the average implementation of a criterion. For example, as indicated in Figure 5.2, the yellow bar indicates that only three models have implemented that criterion. This gives a quick overview of which criteria are most implemented, and which ones, the least. The grey connectors preserve link among bars and columns during reordering. This is important, especially when criteria bars and the data columns in the matrix are reordered independently.

Groups can be created by selecting the different criteria. For a single criterion, there can be two groups of models: those which do not implement the criteria and have a value 0, and those which implement criteria, and have a value 1. With multiple selections, there can be 2^c combinations, with c being a criterion. In most practical cases, only a subset of these combinations exist in the data.

5.3.2 Time Series View

The model output data, which comprises of a time series for each model, is displayed using a line chart comprising of multiple time series (Figure 5.4(a)). But effective visual comparison of similarity among multiple groups is difficult using this view because of two reasons. First, due to similar trajectory of the series, there is a a lot of overlap, leading to clutter. Second, we are unable to show the degree of clustering using this approach. To resolve these design problems, we use small multiples. Small multiples [9] have been used extensively in visualization, one problem with them is when there are a large number of them, it becomes difficult to group them visually without any additional cues. To prevent this, we create a small multiple for each group. When there are time series for different region, a small multiple can also be created for each region to compare groupings across different regions.

5.3.3 Interaction

An overview of the steps in the interactive workflows between the matrix view and the time series view are shown in Figure 5.1. These actions and operations are described below:

Create Groups. While reconciling model structure with model output, scientists can first observe similarity among the models based on their criteria, and accordingly create groups. This is part of the reconciliation workflow described in Section 5.4.1. In the matrix view, groups can be created on interaction. In the time-series view, groups are either suggested by the system or selected by the user through direct manipulation. This is part of the reconciliation workflow described in Section 5.4.2.

Reflect. Creation of groups triggers reflection of the groups in both views. On the matrix side, this is through grouping of the rows. On the time series side, this is done by color coding the lines.

Split. In the time series view, groups can be reflected by splitting the models into small multiples of model groups.

Optimize. While reconciling model output with structure, to handle the variable importance of the criteria, an *optimization* step is necessary. This workflow starts with the scientist selecting groups in the output, which get *reflected* in the matrix view. Next they can choose to *optimize* the importance or the weights, which leads to subsequent iteration. This reconciliation workflow is described in detail in Section 5.4.2.

5.4 Reconciliation Workflows

In this section we describe how we instantiate the conceptual model of visual reconciliation described in Figure 5.3 by incorporating the coordinated multiple views, user interaction and an underlying computational model. The following workflows provide a step-by-step analysis of how the views and interactions can be leveraged by climate scientists for getting insight into structure similarity and output similarity.

5.4.1 Reconcile Structure Similarity with Output Similarity

In Figure 5.4 we show the different steps in the workflow when the starting point of analysis is the model structure. This workflow relies on visual inspection of structure similarity by using matrix manipulation, and observing the corresponding patterns in output by creation of small multiples. The steps are described as follows:

Create groups. For reconciling model structure with output, it is necessary to first provide visual cues about which models are more similar with respect to the

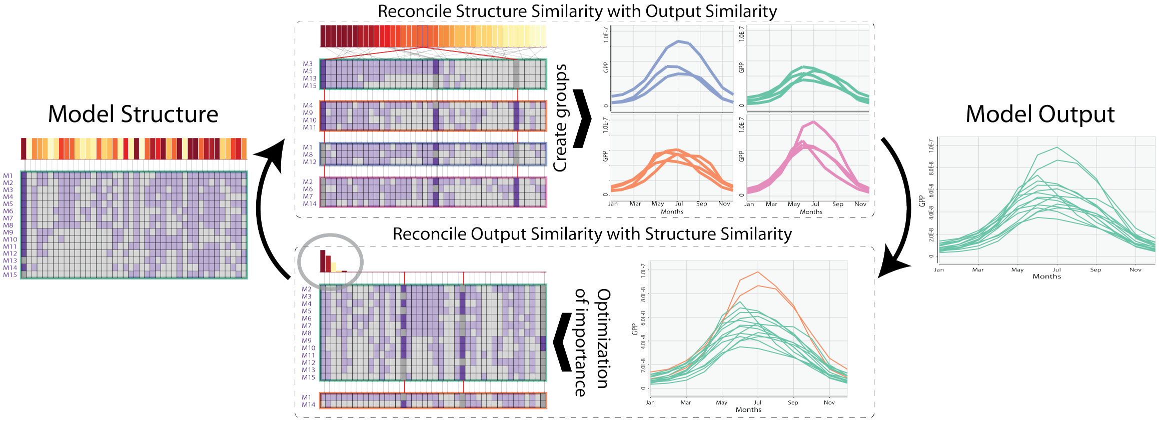

Reconcile Structure Similarity with Output Similarity
Model Structure
Model Output
Create groups
Reconcile Output Similarity with Structure Similarity
Optimization of importance
GPP
Months
Jan
Mar
May
Jul
Sep
Nov

Figure 5.3: **Iterative visual reconciliation** of groupings based on climate model structure and model output. Visual inspection of similarity coupled with an underlying computation model facilitates iterative refinement of the groups and flexible exploration of the importance of the different parameters.

different criteria. For this the default layout of the matrix is sorted from left to right, by high to low average implementation of the different criteria. This is indicated in Figure 5.4(b) by the transition of the importance bars from red to yellow. This gives the scientists an idea of which criteria create more evenly sized groups with 0's and 1's. The criteria which are colored dark red and light yellow will create groups which are skewed: either too many models implement the criteria or they do not. Selecting criteria which are deep yellow and orange, gives more balanced clusters, with around 50 per cent implementation. The highlighted column indicates the criterion with the highest percentage of implementation.

The selected columns are indicated in Figure 5.4(c). These two criteria create four groups. For showing groups of models within the matrix, we introduce vertical gaps between groups, and then draw colored borders around each group. Reordering by columns is also allowed for each group independently as shown in Figure 5.4(c). In that case, the weighted ordering of the bars is kept fixed. For visually indicating the change in ordering we link the criteria by lines. Lines that are parallel indicate that those criteria have not moved due to reordering and share the same position for different groups. Since too many crossing lines can cause clutter, we render the lines with varying opacity. For indicating movement of criteria, we render those lines with higher opacity. To highlight where a certain criterion is within a group, on selection we highlight the line by coloring it red as shown in the figure.

If columns in each group are reordered independently, that shows the average implementation patterns for each group clearly. But it becomes difficult to compare

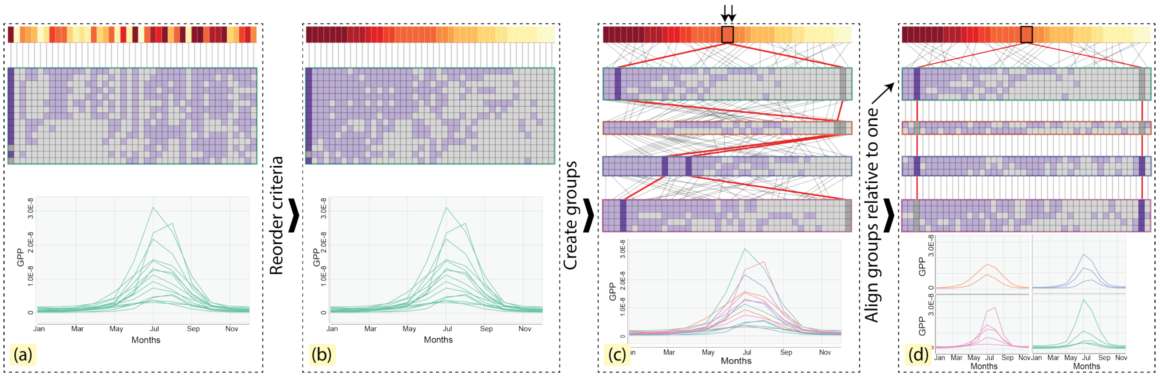


Figure 5.4: **Workflow for reconciling model structure with model output**: This linear workflow relies on matrix manipulation techniques and visual inspection of similarity patterns in the matrix view and the small multiple view.

the implementations of a set of criteria across the different groups. To enable this comparison, user can select a specific group which will be reordered column-wise, and the columns in other groups will be sorted by that order. This is shown in Figure 5.4(d), where the first group from the top is reordered based on the columns, and other groups are aligned relative to that group. As observed, this enables more efficient comparison relative all the implemented and non-implemented criteria in the first group. For example, we can easily find that the rightmost criteria are not implemented by the first group of models, but is implemented by all other groups.

Reflect. The creation of groups in the structure is reflected in the output by the color of the groups. Users can see the names of the models on interaction.

Split. Small multiples can be created for each group (Figure 5.4(d)). The range of variability of models in each small multiple group reflects how similar or different they are. This comparison is difficult to achieve in a time series overloaded with too many lines. This also enables a direct reconciliation of the quality of grouping in model structure with that of the output. For example, as shown in the figure, only the orange group has low variability across models, denoting that the groups based on the criteria in model structure do not create groups where models produce similar output behavior.

5.4.2 Reconcile Output Similarity with Structure Similarity

To reconcile output with structure and complete the loop, we need to account for the fact that different criteria can have different weights or importance in the creation of groups. One of the goals of the reconciliation models is to enable scientists explore different combinations of these criteria that can create groups similar to those in the corresponding model output. However, naive visual inspection is inefficient to analyze all possible combinations without any guidance from the system. For this, we developed a weighted optimization algorithm that complements the human interaction. We describe the algorithm, provide an outline of its validation, and the corresponding workflow, as follows.

5.4.2.1 Weighted Optimization

Using the model structure data and the model output data, we can create two distance matrices. The eventual goal is to learn a similarity function from the output distance matrix and modify the weights of the criteria in the structure distance function for adapting to the output similarity matrix. We describe the problem formulation below.

Let $\hat{M}$ be a matrix representing the model output with size $n \times p$ and $\tilde{M}$ represents the model structure with size $n \times q$. Similarity in model output is computed by the function $\hat{d} : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. This function can be any specialized distance function such as Euclidean, Cosine, etc. For the model structure we use weighted euclidean distance $\tilde{d^w} : \mathbb{R}^q \times \mathbb{R}^q \to \mathbb{R} = \sum_{k=1}^{q} \sqrt{w^k(y_i^k - y_j^k)^2}$, where w^k is a weight assigned to each dimension on $\tilde{M}$.

Using $\hat{d}$ we encode the similarity information of the model output in a distance matrix $\hat{D}$. Our goal would be to find the weights' vector $\mathbf{w} = \{w^1, ..., w^q\}$ which could create a distance matrix for the model structure $\tilde{D}$ containing approximately the same similarity information as the model output. This problem can be formulated as the minimization of the square error of the two distance functions:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} \|\tilde{d^w}(x_i, x_j)^2 - \hat{d}(y_i, y_j)^2\|^2 \\ \text{subject to} \quad & w_k \geq 0, \; k = 1, \ldots, q. \end{aligned} \tag{5.1}$$

where $\|.\|$ is the L_2 norm.

Using this vector $\mathbf{w}$ we can define which criteria are important in the model structure to recreate the same similarity information from the model output. Note that in the previous formulation we have not taken into account the user's feedback. The weights computation step is similar to the one used in weighted metric multidimensional scaling [132] technique.

If we want to incorporate user's feedback into our formulation we can multiply the square errors in Equation 5.1 by a coefficient $r_{i,j}$. This number represents the importance of each pair of elements in the minimization problem. In our approach we allow the user to define groups on the model output, then $r_{i,j}$ will be almost zero or zero for all the elements i, j in a group. Now, we need to minimize:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \sum_{i=1}^{n}\sum_{j=1}^{n} r_{i,j} \| \tilde{d}^{w}(x_i, x_j)^2 - \hat{d}(y_i, y_j)^2 \|^2 \\ \text{subject to} \quad & w_k \geq 0, \; k = 1, \ldots, q. \end{aligned} \tag{5.2}$$

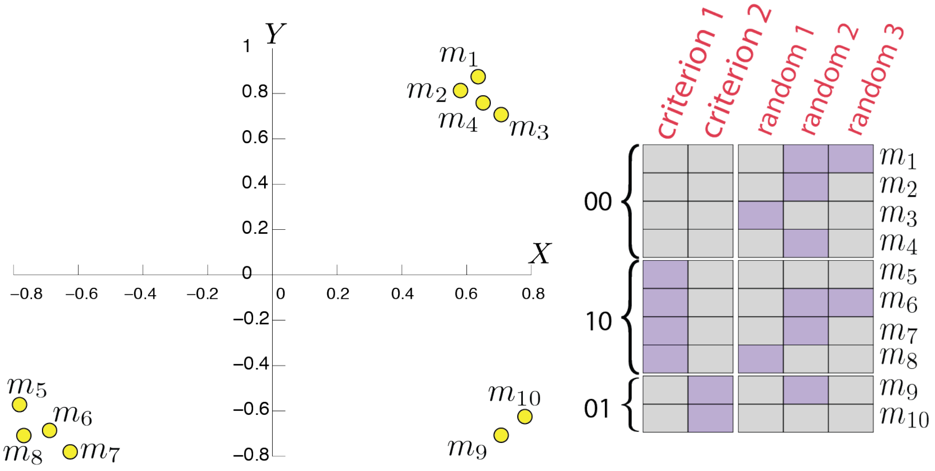
Both equations above can be converted into quadratic problems and solved using any quadratic programming solvers, such as JOptimizer [133] for Java or *quadprog* in MATLAB.

Our approach of incorporating user feedback for computation of the weights is similar to the cognitive feedback model, namely V2PI-MDS [123]. Mathematically the approaches are similar but conceptually they are different on two counts. First, in V2PI-MDS, the high-dimensional data space is represented by the projected data space, and the algorithm attempts to reconcile the two spaces. In our case however, the underlying data spaces are entirely different. We handle this problem by using interactive visualization as a means to preserve the mental model of the scientists about the characteristics of the different data spaces. We could also have used multidimensional projections. But as found in previous work, domain scientists tend not to trust the information loss caused by the dimensionality reduction and prefer transparent visualizations, where the raw data is represented instead [134].

Second, the interaction mechanism for providing feedback to the computation model in the reconciliation model is also different than the V2PI-MDS model. We allow users to define groups within the data, as opposed to direct manipulation and movement of data points in a projection; which is not applicable in our case. Our focus is on the relationship between the weights of the dimensions and the similarity perception they induce. As a result, we let users explore different groupings by using the sorted weights and let them modify the views accordingly. This results in a rich iterative analysis for reconciling the two similarity spaces.

5.4.2.2 Validation

To validate our optimization, we use two synthetic datasets, one for model output and the other one for model structure. The purpose of this validation is to demonstrate



(a) Model Output (b) Model Structure

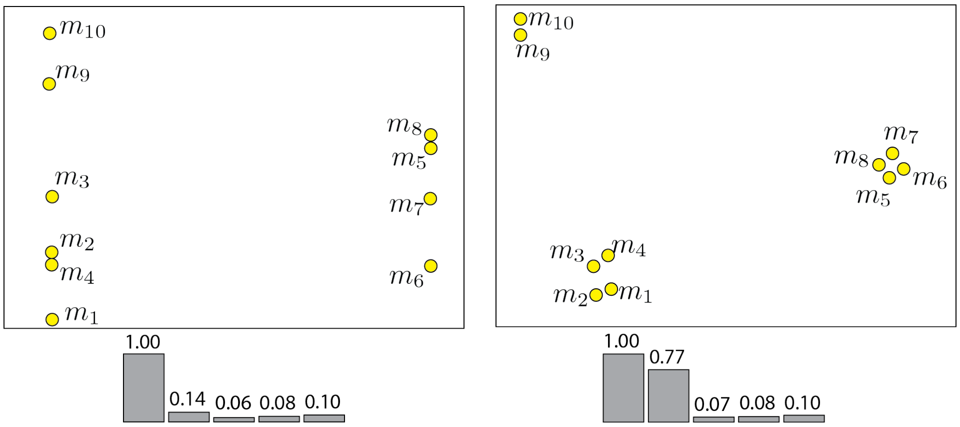
Figure 5.5: **Synthetic data for validating weighted optimization**. Using the model output data in **(a)** and model structure data in **(b)**, we validate the accuracy of the optimization algorithm.

the accuracy of the algorithm in the best case scenario, i.e., when a perfect grouping based on some criteria exists in the data. In most real-world cases, however the optimization will only create an approximation of the input groups.

Our model output is a two-dimensional dataset and we use a scatter plot to visualize it (Figure 5.5(a)). We can notice that we have three well defined groups $\{m_1, m_2, m_3, m_4\}$, $\{m_5, m_6, m_7, m_8\}$ and $\{m_9, m_{10}\}$. Figure 5.5(b) shows our synthetic model structure data which contains boolean values. Each row represents a different model (m_i) and each column a different criterion. The first two criteria were chosen specifically to split the dataset into the same three groups as the model output. For instance when $criterion_1 = 0$ and $criterion_2 = 0$ we can create the group $\{m_1, m_2, m_3, m_4\}$.The next three columns are random values (zero or one).

First, we solve the Equation 5.1 using our synthetic dataset and Euclidean distance for the model output; and we get $\mathbf{w} = \{1.00, 0.14, 0.06, 0.08, 0.10\}$. We use the classical multidimensional scaling algorithm to project the model structure data using the Weighted Euclidean distance. We normalized the weights between zero and one for visualization purpose, but the weighted Euclidean distance uses the unnormalized weights. Figure 5.6(a) shows the two-dimensional data. Our vector $\mathbf{w}$ was able to capture the similarity information from the model output. For example, $\{m_1, m_2, m_3, m_4\}$ is a well defined group. Even though $\{m_5, m_6, m_7, m_8\}$ and $\{m_9, m_{10}\}$ are not mixed, they are not well defined groups.

Next, we incorporate user feedback and set the coefficient $r_{i,j}$ to zero for all



(a) Automatic optimization

(b) Optimization based on user feedback

Figure 5.6: **Validation of user feedback based optimization in the MDS plots.** As we can observe in **(b)**, optimization based on user's feedback gives highest weights to the two criteria which are splitting the models into three groups.

pair combinations in the groups $\{m_1, m_2, m_3, m_4\}$, $\{m_5, m_6, m_7, m_8\}$ and $\{m_9, m_{10}\}$. Solving Equation 5.2 we get the vector $\mathbf{w} = \{1.00, 0.77, 0.07, 0.08, 0.10\}$. Figure 5.6(b) shows the two-dimensional projection of the model structure using the weighted Euclidean distance and $\mathbf{w}$. We notice that now the three groups are well defined. Our algorithm gave the highest weights to the first two criteria ($criterion_1 = 1.0$ and $criterion_2 = 0.7$) which we already knew to have the best combination to split the model structure in the same groups as the model output.

These two experiments show that our formulation accurately gives the highest weights to the most relevant criteria for splitting models into groups, and this can be used to guide the user during the exploration process. In Section 5.5 we will show how this approach works with real data; where in most cases, an approximation of the output group is produced by the algorithm.

5.4.2.3 Workflow

In Figure 5.7 we show how the complete loop starting from output to structure, and back, is executed by user interaction and the optimization algorithm described above. This workflow relies on human inspection of structure similarity through manipulation of the matrix view and observation of the corresponding output in the small multiples of time series. The steps are described as follows:

Create Groups in Output. For suggesting groups of similar outputs, the system uses clustering of time series by Euclidean distance or correlation (Figure 5.7(a)).

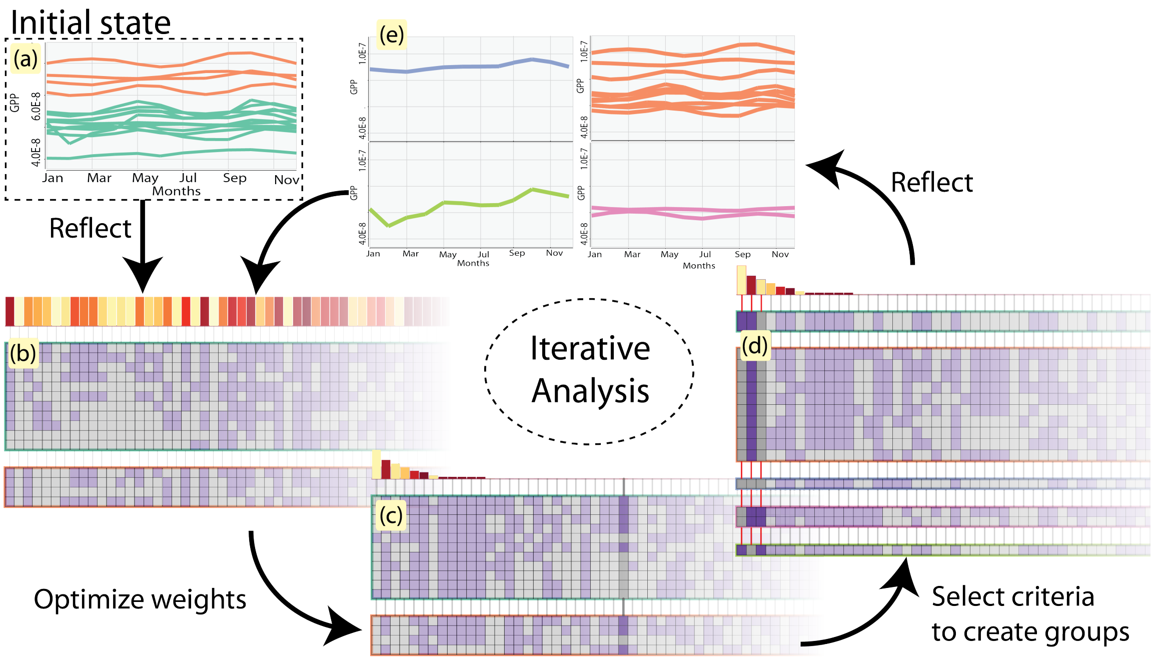


Figure 5.7: **Workflow for reconciling output with structure through feedback**: This iterative workflow relies on weighted optimization based on Equations 5.1 and 5.2, and human initiated parameter tuning and selection for reconciling model structure similarity with model output similarity and vice versa.

While other metrics are available for clustering time series, for this case scientists were only interested in these two. Accordingly, the clusters are updated in the output view.

Reflect in Structure. These clusters are reflected in the model structure side by reordering the matrix based on the groups (Figure 5.7(b)). All the criteria are given equal weights by default, as indicated by the uniform height of the bars. The two views are linked by the color of the groups. Users can also select groups through direct manipulation of the time series in the output view.

Optimize Weights. Next on observing the system-defined clusters, one can choose to optimize the weights for the criteria on the structure side. As shown in Figure 5.7(c), the columns are reordered from left to right based on weights. These weights serve as hints to the user for creating groups on the structure side. The groups are not immediately created to prevent change blindness. The system needs the user to intervene to select the criteria, based on which the groups can be created.

The underlying optimization algorithm as described earlier creates an approximate grouping based on the input. In many cases, as shown in the figure, the highest weight may not give a perfect grouping. By perfect grouping we mean, the optimization algorithm is able to create the exact same groups as the input from the output side. In most cases, the weights for an exact solution might not even exist. By using the optimization, all we get is a group of structure clusters which are as closely aligned with the output as possible.

Create Groups in Structure. Based on the suggested weights, a user can select the two highest weights and create groups, as shown in Figure 5.7(d). There are four possible combinations of these two criteria (with 0's and 1's) and all of them are shown in their own group. In many cases all possible combinations might not exist.

Reflect/Split in Output. The creation of the groups are also reflected on the output side by indicating the group membership of each model by color-coding or by creation of small multiples (Figure 5.7(e)), the output groups created are not perfect, as they do not exactly match with the output groups in the previous step. From this however, the scientists can judge the effect of the two criteria on model output. For example, if for the selected criteria, the presence or absence does not have an impact on the output, that will be reflected in the time series, by their spread or lack of any significant correlation. For inspecting if combining other criteria can give a more perfect grouping on the structure side, that matches with the output, scientists need to continue the iteration and repeat the previous steps.

5.5 Case Studies

We collaborated with 3 climate scientists from the Oak Ridge National Laboratory and from the USDA Forest Service, as part of the Multi-Scale Synthesis and Terrestrial Model Inter-comparison Project (MsTMIP). Each of them have at least ten years of experience in climate modeling and model inter-comparison. MsTMIP is a formal multi-scale 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 [111]. One key goal of MsTMIP is to understand the sources and sinks of the greenhouse gas carbon dioxide, the evolution of those fluxes with time, and their interaction with climate change. To accomplish these goals, inter-annual and seasonal variability of models need to be examined using multiple time-series. Early results from MsTMIP

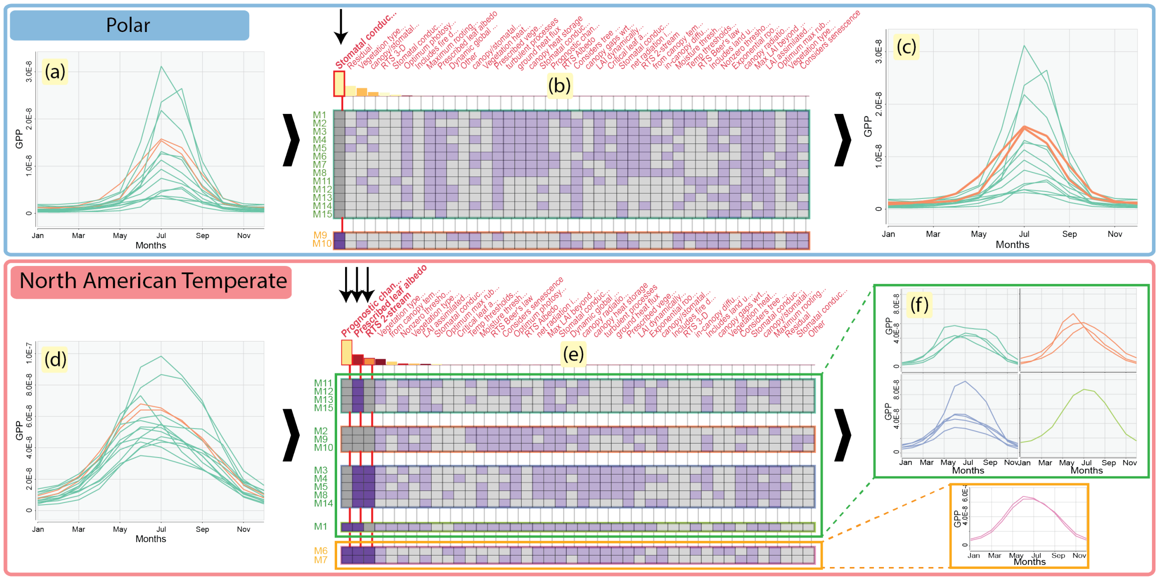


Figure 5.8: **Reconciling seasonal cycle with model structure similarity** using the workflow described in Section 5.4.1. **(a)** Initial user selection in Polar region output. **(b)** Weighted optimization. **(c)** Corresponding output. **(d)** Initial user selection in North American Temperate region. **(e)** Creating groups based on the first three criteria after optimization. **(f)** Small multiple groups of models.

have shown that variation in model outputs could be traced to the same in model structure. Using visual reconciliation, climate scientists wanted to further understand whether similarity or differences in model structure play a role in the inter-annual variability of Gross Primary Productivity (GPP) for different regions. Inclusion of particular combinations of simulated processes may exaggerate GPP or its timing more than any component in isolation. Inclusion of a patently incorrect model structure could dramatically sour model output by itself.

We provided our collaborators with an executable, which they used for a month and reported back to us on their findings, as reported below. Then we conducted face-to-face interviews about the usage of the technique and got positive feedback on how the technique is a first step towards solving the problem of reconciling model structure with output. We describe two cases where our collaborators could find relationships between model structure and model output using a prototype implementation of the visual reconciliation technique. The model structure data is segmented into four classes: energy, carbon, vegetation, and respiration. In this case the scientists wanted to understand the relationship between criteria belonging to energy and vegetation, and their GPP variability in Polar and North American Temperate regions. Each of the model structure datasets consist of about 15 models and about 20 to 30 criteria.

5.5.1 Reconciling Seasonal Cycle Similarity with Structural Similarity

The seasonal cycle of a climate model is given by the trajectory of the time series and the peaks and crests for the different months in a year. Exploring the impact of seasonal cycles for different models with respect to GPP is an important goal in climate science, since the amount and timing of energy fixation provides a baseline for almost all other ecosystem functions. Models must accurately capture this behavior for all regions and conditions before other, more subtle ecosystem processes, can be accurately modeled. The motivation for this scenario was to find if there is any dependency between regional seasonal cycles of models and included model structures with respect to the overarching energy criteria.

The scientists started their analysis in the Polar region by selecting the M9 and M10 models which appeared to be similar with respect to both their GPP values and the timing of their seasonal cycles, as shown in Figure 5.8(a). Their intent was to observe which energy parameter causes M9 and M10 to behave similarly in one group, and the rest in another. They optimized the matrix view to find the most important criterion, which was found to be Stomatal conductance. After this step they chose to select this criterion to split the models into two groups, shown in Figure 5.8(b) and reflected in Figure 5.8(c). The underlying optimization algorithm thus gave a perfect grouping, with the models that implement Stomatal conductance in the orange group, while the rest are in another group. The climate scientists were already able to infer that Stomatal conductance has strong impact on the seasonal cycles of M9 and M10.

Next the scientists selected the M6 and M7 models in the North American Temperate (NAT) region, which appear to be similar with respect to their seasonal cycle and GPP output (Figure 5.8(d)). This grouping is already intuitive and inspires confidence, because of its consistency with the known genealogical relationship of these two models as siblings. With the same goal as the previous case, they optimized the matrix view, and found that Prognostic change was the most important structural criterion to approximately create the two groups. This structural criterion provided a near-perfect segmentation, except for the M1 model, which also implements this parameter, as shown in Figure 5.8(e). In an attempt to get the exact segmentation, they selected the next two most important criteria, which are prescribed leaf index and RTS2-stream. M6 and M7 implement both of these criteria and are in one output group, while the other green output group is split into three sub-groups based on their implementation of these three criteria. The implementation of these three criteria thus has a significant effect on the grouping of these two models with respect to their GPP. The scientists could continue in this way to find more inferences from the implementation or non-implementation of these three structural criteria, by further observing their output in small multiples, as shown in Figure 5.8(f). This shows that the blue group, none of which implement Prognostic change, but all of which implemented the other two, show a greater spread of GPP output values than any other group. In this way, the scientists could reconcile the impact of different energy criteria on the seasonal cycle and regional variability of GPP.

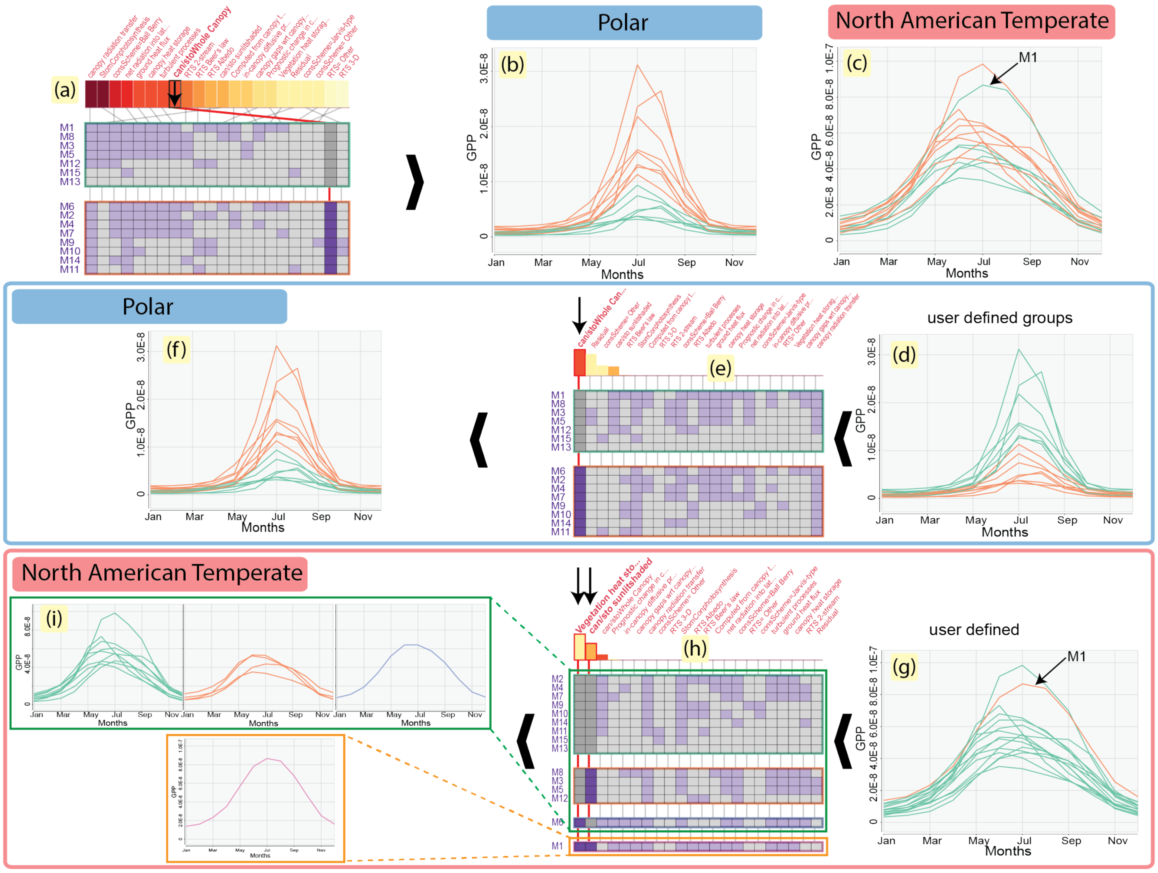


Figure 5.9: **Iterative exploration of structure-output dependency** using a combination of the two workflows for reconciliation. **(a)** Initial user creation of groups; **(b,c)** Corresponding groups in regions; **(d,e,f)** workflow for verifying user-defined groups; **(g,h,i)** workflow for finding the criteria that can potentially cause M1 to be an outlier, and then looking at range of variability in small multiple outputs.

5.5.2 Iterative Exploration of Structure-Output Dependency

In this case, the scientists started by looking at the model structure data for discovering structure criteria that could explain model groups having high and low GPP values across both `Polar` and `NAT` regions. A simple sequential search for criteria is inefficient for reconciliation. To start their analysis, as shown in Figure 5.9(a), the matrix view is first sorted from left to right by the columns having high numbers of implementations. The sorting enabled the scientists to group using a criterion that would cause balanced clusters, i.e., divide the models into equal groups. In this view, these criteria would lie in the center, having orange or deep yellow color. In course of this exploration, they found that the `canopy/stomatal conductance whole canopy` structural criterion splits the group into nearly equal halves. These clusters are represented in the output by green, i.e., not implementing that criterion, and orange, i.e., implementing that criterion. Further, looking at the output, as shown in Figure 5.9(b), scientists found that the orange group has higher `GPP` values and the green group has lower values. In other words, the models that have implemented `stomatal conductance` have higher `GPP` values than the ones that have not implemented this criterion. This grouping is consistent for the `North American Temperate` region, with the exception of the `M1` model, as shown in Figure 5.9(c).

Next, the scientists wanted to verify whether by performing optimization, they can get the same criterion to be the most important for the behavior of `GPP` within the `Polar` region, which represents a different, extreme combination of ecological conditions. They selected the green group, as shown in Figure 5.9(d), and then chose to optimize the matrix view. They found the same criterion (`canopy/stomatal conductance whole canopy`) to have the highest weight, reinforcing the reconciling power of this same group of model structures for explaining differences in `GPP` across two extreme eco-regions. Thus, the criterion that they discovered interactively could be verified algorithmically. Note that, as shown in Figure 5.9(d), only one of the models is classified in a different group than the user-selected group.

For the `NAT` region, the scientists wanted to drill-down to determine what was causing `M1` to behave differently, as was found during the initial exploration. They defined two groups, with one of them only having `M1` as shown in Figure 5.9(g). Once they chose to optimize the matrix, they found that no single criterion could produce the same output groups. However, by combining the two most important criteria, which are `vegetation heat` and `canopy-stomatal sunlit shaded` (Figure 5.9(h)), `M1` was put in a separate group by itself. It was the only model that implemented both of these criteria. Additionally, the scientists also saw that the models in the green group, which did not implement any of these criteria, had a larger range of `GPP` variability than the other model groups (Figure 5.9(i)). They concluded that,by allowing both `more- and less-productive sunlit` and `shaded canopy leaves`, respectively, models which implement these differential processes seem to stabilize the production of GPP, even across extremely different eco-regions, possibly accurately reflecting the actual effect of these processes in nature.

5.6 Summary

In this chapter, we have presented a novel visual reconciliation technique, using which climate scientists can understand the relationships between model structure similarity and model output similarity.

By exploiting visual linking and user-steered optimization, we are able to communicate to the scientists, the effects of different groups of criteria on the variability of model output. Using this technique, scientists could form and explore hypotheses about reconciling the two different similarity spaces, which was not possible before, yet which is crucial for refining climate models; which is reflected in the following comment by one of our collaborators:

> *"Due to imperfect knowledge, understanding, and modeling, correlations in the climate modeling domain may be weakly exhibited at best. This inherent weakness poses the greatest challenge to recognition and reconciliation of such correlations; yet, it is only through the reconciliation of such correlations upon which progress in improving climate models rests."*

Regarding the effectiveness of the reconciliation technique, another collaborator observed that:

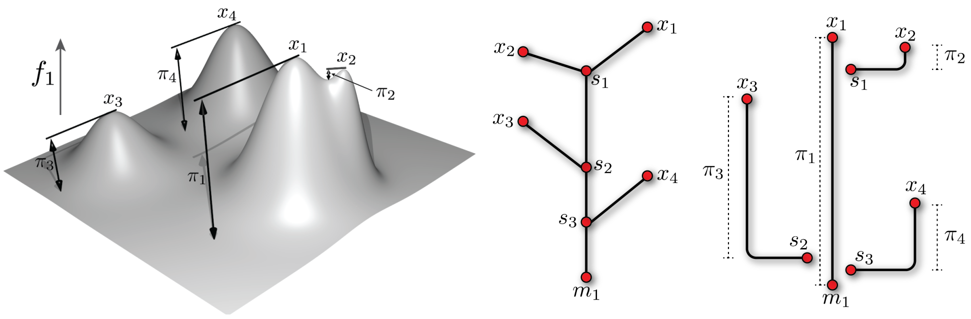
> *"One of the most valuable functions of the technique is to effectively remove from consideration the complications created from model structures, that have little to no effect on outputs, and to effortlessly show and rank the differential effects on output created by seemingly related or unrelated model structures."*

Chapter 6

Using Maximum Topology Matching to Explore Differences In Climate Models

In this chapter, we work with Species distribution models (SDM). SDMs are used to help understand what drives the distribution of various plant and animal species. These models are typically high dimensional scalar functions, where the dimensions correspond to different model algorithms and environment variables, also known as predictors. Ecologists are interested in studying the behavior of these models over its parameter space comprising of its predictors. However, their current approach resorts to visualizing one-dimensional slices of the models. That is, in considering the influence of one specific predictor, the common technique is to select a predictor of interest and fix the values of the other predictors to their mean values, and compare the variation of the models with respect to the selected predictor. This results in a one-dimensional curve known as *response curve* (see Figure 6.11(c)). The main shortcoming of restricting the analysis to considering only one predictor at a time is that it is not possible to obtain an accurate view of the model. This is because, the features resulting from the interactions between the other predictors are lost through such dimensionality reduction. More importantly, even when looking at one-dimensional slices, the response curves are restricted to the fixed value of the other predictors. While there has been some work where ecologists analyze two-dimensional slices of the models [135, 136], the above problems still hold.

The goal of this work is to help ecologists understand the interactions between the predictors in SDMs, and thus have a better understanding of what drives the various species. To this end, we propose a *Visual Analytics Approach* that use computational



(a) (b) (c)

Figure 6.1: **Topology of scalar functions.** **(a)** Height function f_1 defined on a 2-dimensional manifold having 4 maxima. π_i represents the persistence of a maximum x_i. **(b)** The join tree tracks the connectivity of the super-level sets of a scalar function. **(c)** Each branch in the branch decomposition of the join tree corresponds to the path between a creator-destroyer critical point pair.

topology to help explore and compare SDMs directly in the high dimensional space. In particular, we use the extrema of the corresponding scalar functions to *guide* the users towards interesting features of the SDM.

While such exploration of the SDMs will provide more flexibility to the ecologists, manual comparison between the two models is still a time consuming and often impractical process. To overcome this, we propose a novel technique that can be used to compare two scalar functions in a locality-aware manner. We do this by first creating a bipartite graph where the edges correspond to possible correspondence between the extrema of the two functions. The edge weights are defined such that they reflect both the spatial locality of the extrema, as well as the likeness in terms of their function values. The maximum weight matching of the bipartite graph is then computed to obtain the correspondences between the set of extrema. These correspondences are then used to compute a topological similarity measure between the two functions. We also show through experiments the robustness of the matching and the resulting topological similarity measure.

We design a visualization interface which uses the above techniques to help ecologists explore SDMs and analyze the differences between them. Finally, working together with ecologists we demonstrate the effectiveness of our technique and the user interface through several use case scenarios involving SDMs of different species.

6.1 Background

In this section, we provide the necessary background on concepts from computational topology that form the mathematical and algorithmic basis of this work. We refer the reader to the following textbooks [137, 138, 139] for a comprehensive discussions on these concepts.

6.1.1 Morse functions and Species Distribution Models

Let $\mathbb{M}$ denote a d-manifold with or without boundary. Given a smooth, real-valued function $f : \mathbb{M} \rightarrow \mathbb{R}$ defined on $\mathbb{M}$, the *critical points* of f are exactly where the gradient becomes zero. The function f is called a *Morse function* if it satisfies the following conditions [140]:

1. All critical points of f are non-degenerate and lie in the interior of $\mathbb{M}$.
2. All critical points of the restriction of f to the boundary of $\mathbb{M}$ are non-degenerate.
3. All critical values are distinct *i.e.*, $f(p) \neq f(q)$ for all critical points $p \neq q$.

For a Morse function f defined on a d-manifold $\mathbb{M}$, there are $d + 1$ types of critical points indexed from 0 to d. In this work, we are interested in the two most familiar types – *minimum* (with index 0) and *maximum* (with index d), corresponding to a point p whose function value is smaller, or larger, than all other points within a sufficiently small neighborhood of p, respectively. Figure 6.1(a) shows a height function, f_1, defined on a 2-manifold. This function consists of 4 maxima – x_1,x_2,x_3, and x_4.

A species distribution model (SDM) is a d-dimensional function $m : \mathbb{R}^d \rightarrow \mathbb{C}$, where $\mathbb{C} = [0, 1]$ denotes the unit interval. It assigns a probability for the presence of a given species based on the values of its d predicates. In the remaining discussion, we assume that the input SDMs are Morse functions. In case the above conditions do not hold, simulated perturbation of the function [141, Section 1.4] ensures that no two critical values are equal.

6.1.2 Topological Persistence

A sub-level set of a function f, $\mathbb{M}^{(-\infty,a]} := \{x \in \mathbb{M} \mid f(x) \leq a\}$, is the set of all points having function value less than or equal to a. A super-level set is similarly defined as the preimage of the interval $\mathbb{M}^{[a,+\infty)}$.

Consider the sweep of the function f in increasing order of function value. The topology of the sub-level sets changes when this sweep passes a critical point. In particular, at a critical point, either new topology is generated or some topology is destroyed, where topology is quantified by a class of 'cycles'. For example, a 0-dimensional cycle represents a connected component, a 1-dimensional cycle is a loop that represents a tunnel, and a 2-dimensional cycle bounds a void. A critical point is a creator if new topology appears and a destroyer otherwise. It turns out that one can pair up each creator v_1 uniquely with a destroyer v_2 that destroys the topology created at v_1. The persistence value of v_1 and v_2 is defined as $f(v_2) - f(v_1)$, which intuitively indicates the lifetime of the feature created at v_1, and thus the importance of v_1 and v_2.

The function in Figure 6.1(a) consists of three creator-destroyer pairs – (x_2, s_1),(x_3, s_2), and (x_4, s_3). While the global maximum x_1 has a persistence value of ∞, we use a notion of extended persistence where in addition to the above pairs, the global maximum is paired with the global minimum [142]. The persistence values of the set of maxima x_i of the function in Figure 6.1(a) is highlighted as π_i.

Topological persistence of a feature measures the amount of simplification required to smooth the input function in order to remove that feature. This property is later used to define a distance measure between two SDMs.

As mentioned above, in this paper we only consider extreme points of the input function as features. Given an input domain of size n, the persistence of such features can be computed efficiently in $O(n \log n + n\alpha(n))$ time using the union-find data structure[1], versus the usual cubic-time algorithm to compute general topological persistence [143, 144].

6.1.3 Merge Trees

A *join tree* tracks the topology of the super-level sets of the input function, while the *split tree* tracks the topology of the sub-level sets [145]. The join tree and split tree are together known as *merge trees.* Figure 6.1(b) shows the join tree of the function shown in Figure 6.1(a).

The join / split tree is computed using the union-find data structure to keep track of the connected components of the super-level set (or the sub-level set). This procedure also returns the set of creator-destroyer pairs corresponding to the topological features.

[1] The persistence algorithm works for more general topological spaces than manifolds. We only describe the case when it is induced by a function defined on a manifold.

A merge tree can be decomposed into a set of *branches* using the obtained critical point pairs [146]. Each branch corresponds to the path in the merge tree between a creator-destroyer critical point pair. Thus, the height of a branch represents the persistence of the corresponding critical points. Figure 6.1(c) shows the branch decomposition of the join tree in Figure 6.1(b). The smoothing of a function obtained by removing an extremum can be represented abstractly by removing the branch corresponding to that extremum together with all its sub-branches. This observation is key in our algorithm that computes the topological similarity measure between two SDMs.

6.2 Related Work

In this section, we first briefly discuss related work that are used to explore high dimensional functions. Next, we survey topology based techniques that are used for comparing two scalar functions.

6.2.1 Exploring High Dimensional Functions

There are multiple visual analytic techniques to explore the parameter space of high dimensional scalar functions (also referred to as models). Most of these methods are based on sampling the parameter space or using regression algorithms to approximate / predict output from unknown configurations. Matkovic et al. [147] proposed to visualize multirun data as families of data surfaces (with respect to pairs of independent dimensions) in combination with projections and aggregation of the data surfaces at different levels. The same authors [148] also proposed to generate new sample points by interactively narrowing down the control parameters in the visualization via brushing to support visual steering of a simulation. Along the same lines HyperMoVal [149] was designed to visually relate one or more high-dimensional scalar functions with validation data. Later, Berger et al. [150] extended HyperMoVal using regression models for a continuous exploration of the sample parameter space. Similarly, we can find applications of parameter exploration in other domains such as image segmentation [151]. Other approaches partitioned the input space and provided visual analytics strategies for exploration of the input space using one or two parameters at the time [152, 153]. However, all of these approaches require users to manually explore the space in order to identify interesting regions.

Topological abstractions have also been used to create visual representations of high dimensional functions. Topological Landscapes [154, 155] provide a 2D terrain representation having the same contour tree as the input high-dimensional scalar function. When the input are point clouds Oesterling et al. [156] propose to reconstruct the scalar function using density kernels and use topological landscapes to visualize the density of points using a 2D terrain. Geber et al. [157] segment the input domain using an approximate Morse-Scale complex on a cloud of point samples. Then each segment is represented by a curve using a regression. Finally those curves are visualized in 2D space using dimensionality reduction algorithms. While these techniques help users understand the topology of the involved function, it is difficult to use these methods to compare scalar functions as the neighborhood information is lost in the transformation to a 2D representation.

In the ecology domain, while there has been some work on trying to study two-dimensional slices of SDMs [135, 136], ecologists mostly use the SAHM package [158] for VisTrails [96] which supports exploration through one-dimensional response curves.

6.2.2 Comparing Scalar Functions

Early methods of comparing scalar functions directly used the persistence of the critical points of the functions to do so. A distance function, usually bottleneck distance, between the persistence diagrams [159] of two functions are used to compare them. Using an alternate representation, called barcode, Carlsson et al. [160] represented the persistence of the features as intervals on a real line. They then defined a metric to compute the similarity between two barcodes. A disadvantage of using a pure persistence based measure is that they do not capture the neighborhoods of the features.

More recent methods for comparing scalar functions used some form of topological abstraction of the scalar functions to compare them. Morozov et al. [161] defined the interleaving distance between two merge trees as the minimum cost of shifting points in one tree to obtain a mapping of one tree to the other. Beketayev et al. [162] defined a distance between two merge trees by comparing all possible branch decompositions of the two trees. Bauer et al. [163] extended the interleaving distance between two merge trees to Reeb graphs and proposed the functional distortion distance to compare two Reeb graphs, where a Reeb graph is a topological structure which tracks the connectivity of level sets of a scalar function with increasing function value. More recently Narayanan et al. [164] proposed a distance measure between two scalar functions based on the maximum common subgraph between complete extremum

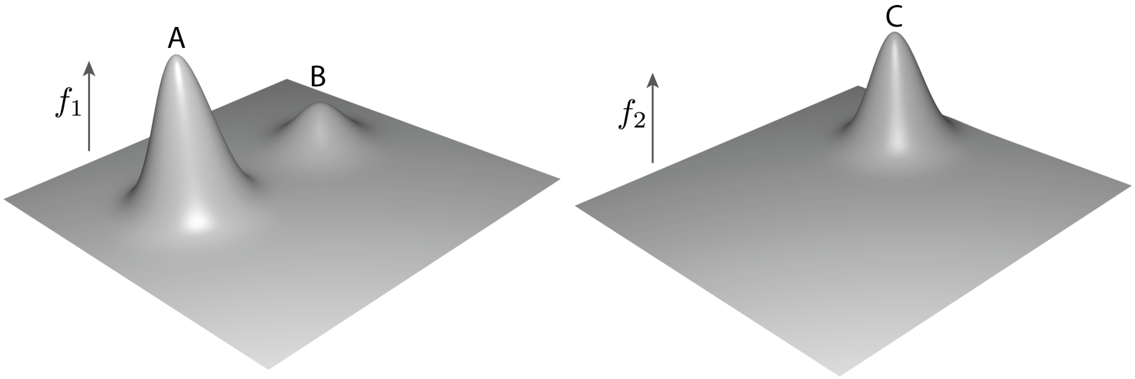


Figure 6.2: f_1 **and** f_2 **are two functions defined on the same domain.** Existing techniques identify peak C to be similar to A instead of B even though B and C are in the same neighborhood of the domain.

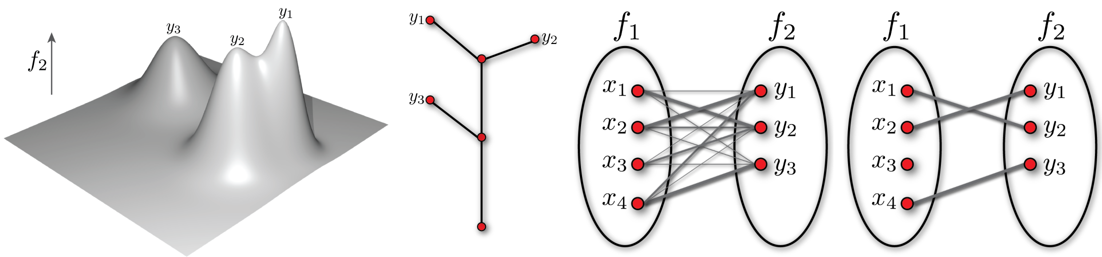
graphs, where an extremum graph is a topological data structure that captures proximity between extreme points in a scalar field [165]. Alternative to computing distance measures, topological structures have also been used to structurally compare two functions. Multi-resolution Reeb graphs have been used for comparing two shapes [166]. Saikia et al. [167] introduced a data structure called extended branch decomposition graphs using which they could compare between all sub-trees of two merge trees. Toplogical abstractions have also been used to identify similar structures within a scalar function [168, 169].

While the above methods capture adjacency based on the connectivity between level sets, they still suffer from two shortcomings. First, it is possible for two adjacent features (adjacent edges) to actually be far from each other. Second and more importantly, the actual locality of the features identified as similar need not be located in the same locality of the domain. For example, consider the two functions shown in Figure 6.2. the above techniques would identify maximum A in f_1 with C in f_2 even though the two maxima are far from each other. However, given that B and C are in a similar locality of the domain, we are interesting in identifying B with C.

Instead of an abstraction, level sets and their properties have also been used for comparing scalar functions [170, 171, 172]. Since these techniques require computing the level sets, extending them to work for high dimensional functions is non-trivial.

6.3 Scalar Function Similarity

We now describe our technique to compare two scalar functions that are defined on the same domain. The main idea is to identify the best match, in terms of the



(a) (b) (c) (d)

Figure 6.3: **Computing the maximum topology matching. (a)** 2-dimensional scalar function f_2 that is compared with the function f_1 in Figure 6.1(a). **(b)** Join tree of f_2. **(c)** The constructed bipartite graph between the maxima of the two functions. **(d)** The computed matching between the maxima.

location and function value, between the set of extrema (of the same type) of the two functions. This matching is then used to compute the similarity measures between the functions.

In this section, we first describe the procedure to identify the correspondence between the set of extrema of two functions. Next we define two similarity measures between the functions and describe how they are computed using the found correspondences. Without loss of generality, the techniques in this section are described with respect to the set of maxima of the functions. The same procedures apply to the set of minima as well.

6.3.1 Maximum Topology Matching

The first stage in identifying the similarity between two scalar functions f_1 and f_2 is to identify the correspondence between the extrema of the functions. Without loss of generality, we assume that the two functions are normalized between 0 and 1.

Let M_1^+ and M_2^+ be the set of maxima of f_1 and f_2 respectively. We first create a complete weighted bi-partite graph $G_T(M_1^+, M_2^+, E^+)$ in which the two partitions corresponds to the maxima of the two functions respectively. Consider a pair of maxima $a \in M_1^+$ and $b \in M_2^+$. Let the difference between their function values be $\delta_{a,b} = |f_1(a) - f_2(b)|$. Let $d_g(a,b)$ denote the distance between the pair of maxima. Since the SDM is defined on $\mathbb{R}^d$, we use the Euclidean distance for this purpose. We assign a weight $w_{a,b}$ to the edge corresponding to the pair of maxima a and b as follows:

$$w_{a,b} = (1 - \delta_{a,b}) \times e^{\frac{d_g(a,b)^2}{r^2}}$$

Here, r is a cut-off radius, which acts as a knob to define the neighborhood sensitivity.

The weight $w_{a,b}$ essentially consists of two parts. A high value of $(1 - \delta_{a,b})$ implies a high similarity between the two maxima in terms of their function value. The weighting term $e^{\frac{d_g(a,b)^2}{r^2}}$ ensures that importance is given to pairs of maxima that are closer to each other, thus preserving the neighborhood locality. Thus a high weight between a pair of maxima implies that they are *similar* not only in terms of their function value, but are also within the same locality of the domain. For example, in order to compare the function f_2 shown in Figure 6.3(a) with the function f_1 from Figure 6.1(a), we create the bipartite graph shown in Figure 6.3(c). The thickness of the edges represents their weights. Note that the edges corresponding to maxima pair that are nearby in the function domain have weight higher than those that are further away.

We next compute the maximum weighted matching [173] on the graph G. A *matching* is defined as a set of pairwise non-adjacent edges. A maximum weighted matching is defined as a matching where the sum of values of the edges in the matching has maximum value. The resultant matching provides the correspondence between the set of maxima of the two functions. For the example of functions f_1 and f_2, the obtained matching is shown in Figure 6.3(d). Note that our technique matches the maxima x_1 to y_2 and x_2 to y_1 due to their proximity. This is unlike existing techniques that do not use the locality information to match features, which would have matched x_1 to y_1 and x_2 to y_2. Also, these techniques would have matched x_3 to y_3 since they use the relative persistence of features when computing similarity.

6.3.2 Topological Similarity

The topological similarity between f_1 and f_2 is defined as the effort required to make the two functions have the same number of maxima in the same neighborhood of the domain. Such similar functions will produce a *perfect matching* in G. A perfect matching is a matching that matches all vertices of the bipartite graph. This quantity is measured as the minimum amount of simplification that is to be performed to attain such a perfect matching.

Consider the function f_1 having the set M_1^+ as its maxima. Let $C \subseteq M_1^+$ be the set of maxima that have a corresponding match in M_2^+. Then $\overline{C} = M_1^+ \setminus C$ is the set of maxima that have to be simplified. The join tree and the appropriate branch decomposition is used to compute, τ_1, the amount of simplification required as follows.

Let r be the root of the join tree T_S. That is, r is the global minimum of the

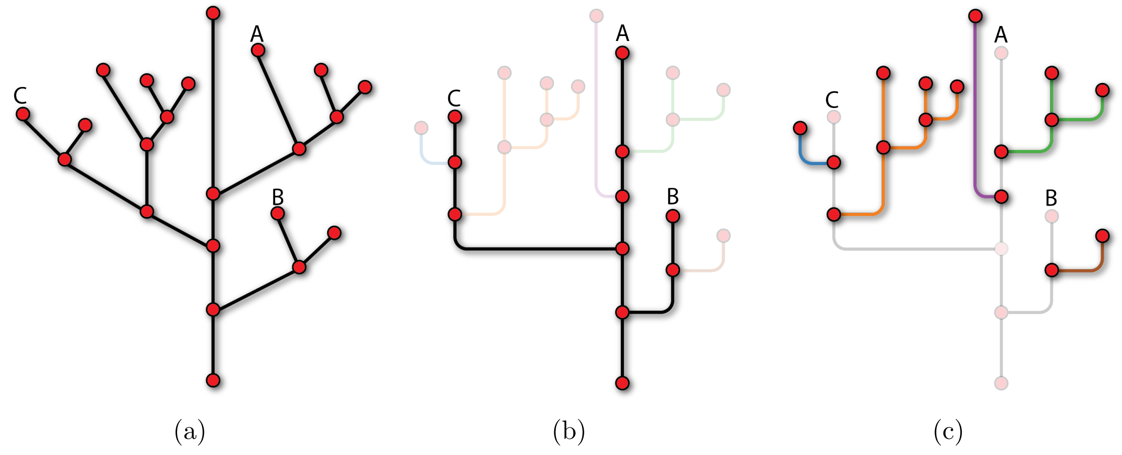


Figure 6.4: **Computing the topological similarity.** **(a)** Example join tree of a function that is being compared. Let maxima A, B and C be matched to maxima of the other function. **(b)** The first step of the algorithm computes the matching tree T'_S, a connected sub-tree of the join tree $T_S$induced by the matched maxima. **(c)** The connected components of $T_S \setminus T'_S$ corresponds to the regions of the function that has to be simplified to obtain a perfect matching. The minimum amount of simplification required to do so measures the topological similarity between the two models.

function f_1. In the first step, we construct the *matching tree* T'_S, the join tree of f'_1 which is the function f_1 in which the set of maxima $\overline{C}$ are removed (simplified). This tree is constructed as follows:

1. For each maximum $m \in C$, construct the path L_m, which is the unique path from the leaf corresponding to m to the root r.
2. the matching tree $T'_S \subset T_S$ is the tree induced by the paths L_m computed above, *i.e.*, $\{T'_S = \bigcup_{m \in C} L_m\}$

Figure 6.4(b) illustrates the matching tree corresponding to the join tree in Figure 6.4(a) when three of its maxima have a matching in the bipartite graph.

Let $\overline{T_S} = T_S \setminus T'_S$. Consider the connected components K of $\overline{T_S}$. Simplifying the set of maxima in $\overline{C}$ is equivalent to removing each of these connected components from T_S. Each of these components corresponds to a connected sub-tree of T_S. The effort τ_k required for simplifying a given component k is equal to the *height* of the largest branch of the corresponding sub-tree. Figure 6.4(c) shows the different components that have to be simplified for the example in Figure 6.4(a).

τ_1 is then computed as the maximum value of τ_k over all components $k \in K$. τ_2, the minimum amount of simplification required for function f_2 is computed in a similar manner. The *topological similarity* $\tau = \max(\tau_1, \tau_2)$ is the minimum simplification required to obtain a perfect matching between the two functions.

6.3.3 Functional Similarity

Given a perfect matching between the maxima of the two topologically similar functions, it is still possible that the matched maxima could differ in their function values. The functional similarity measures this difference. Formally, the functional similarity ϕ is the maximum $\delta_{a,b}$ over all edges (a, b) that are part of the matching. Intuitively, this quantity is used to measure the maximum amount of change required to construct functionally similar functions from topologically similar functions.

6.4 Implementation

In this section, we first provide implementation details describing the adaptation of our similarity technique for the high dimensional SDMs. Next, for completeness we briefly describe the algorithm to compute merge trees [145]. Finally, we discuss the time complexity of our technique and how noise in the input effects the similarity measure.

6.4.1 Discretizing SDMs

In order to efficiently compute the topology of a species distribution model $m : \mathbb{R}^d \rightarrow \mathbb{C}$, the high-dimensional domain of m is approximated as a nearest-neighbor graph, denoted by G, of a set of points sampled uniformly from the domain of $\mathbb{R}^d$. m is then represented by a piece-wise linear (PL) function defined on G. Thus, $m : G \rightarrow \mathbb{C}$. The function is defined on the vertices of the graph and linearly interpolated within each edge.

6.4.2 Computing Merge Trees

Given a PL function m defined on a graph G, the *upper link* of a vertex v is defined as the graph induced by adjacent vertices of v having functional value greater than v. Similarly, the *lower link* of v is defined as the graph induced by adjacent vertices of v having function value less than v. The join tree of m is computed by first sorting the vertices of the G in decreasing order of function value. Next, for each vertex u in this sorted list, the algorithm performs the following operations:

- If u is a maximum (its upper link is empty), create a new component containing u and set u as its *head*.
- If the upper link is not empty, find the components that contain the vertices in the upper link of u. These correspond to the components of the super-level set at $m(u)$. Add an arc between u and the head of each of the components. Merge these components and set u as the head of the merged component. If the number of components is greater than one then u is a *join saddle*.

Similarly, the split tree is computed by traversing the vertices in increasing order of function values.

6.4.3 Analysis

6.4.3.1 Time complexity

Let n be the number of vertices, and m be the number of edges in the graph G that is used to represent the domain of the SDM. Its merge trees can be computed in time $O(n \log n + m\alpha(m))$ [145] using the union-find data structure to maintain the components of the super-level set (sub-level set). Here α is the inverse Ackermann function.

Let the number of extrema in the two functions (leaf nodes of the merge tree), f_1 and f_2, be $t_1 = O(n_1)$ and $t_2 = O(n_2)$ respectively. The created bipartite graph has $n_v = t_1 + t_2$ nodes and $n_e = t_1 \times t_2$ edges. Computing the maximum weight matching can be accomplished in $O(n_v^2 \log n_v + n_v n_e)$ using Dijkstra's algorithm with a fibonacci heap [173].

6.4.3.2 Effect of noise

Noise-based artifacts are common in real world data sets. It is therefore important to consider the effect of noise to the stability of the matching, and the resulting similarity measures. If the original matching remains even with noise, then given the low persistence of the noisy extrema that are added, there is no significant change to the similarity measures. So let us assume that the matching is different from the original. Consider a matched pair of maxima (x_i, y_j) between functions f_1 and f_2. Without loss of generality, assume that there was noise introduced into the function f_2. This would potentially create additional maxima in the neighborhood of y_j. Let the effect on the function value variation due to the noise be bounded by ϵ_f.

Depending on the changes in the weights, there are three scenarios that are possible.

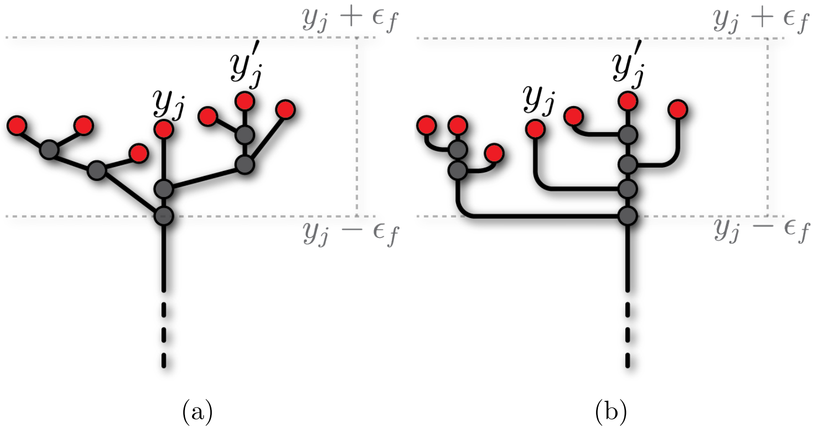


Figure 6.5: **Effect of noise in the neighborhood of a maximum. (a)** Presence of noise could potentially introduce multiple extraneous extrema in the neighborhood of a relevant maximum y_j. **(b)** In case one of the noisy maximum y'_j is matched in the maximum matching, then the amount of simplification needed to perform to remove y_j is bounded by ϵ, while the simplification needed for other extraneous maxima is bounded by 2ϵ.

1. The matching (x_i, y_j) does not change due to noise.

2. The matching algorithm pairs x_i with a maximum y'_j in the neighborhood of y_j. In this case, both the topological similarity and functional similarity change by a maximum of ϵ_f. This is because, y'_j is in the resulting matching tree, and y_j has to be simplified. The persistence of y_j in the new configuration is then bounded by the change in function value (See Figure 6.5), which in the worst case is $2\epsilon_f$.

3. The matching pairs x_i with a maximum y_k not in the neighborhood of y_j. This implies that weight of the edge (x_i, y_k) managed to increase past the weight of edge (x_i, y_j), *i.e.*, the weights $w_{x_i,y_j} \approx w_{x_i,y_k}$. While the weight of the matching in this case would not significantly change, the values of topological similarity and functional similarity could be affected.

We are interested in further exploring Case 3 above when y_j and every other maxima y'_j that was created in the neighborhood of y_j due to noise remains unmatched. If at least one of them is matched to another maximum x'_i, then the change to the topological similarity would be similar to Case 2 above.

Given that the function values are normalized between 0 and 1, the weights of the edges of the bipartite graph is always between 0 and 1. When the weights of the edges under consideration are low, then there are three possibilities:

1. Both y_j and y_k are far away from x_i; or

2. One of the maxima, say y_j is far from x_i but has $\delta \approx 0$, and for y_k, δ is high while it is within the neighborhood of x_i; or

3. Both maxima are in the neighborhood of x_i, but have very high δ (close to 1).

All the above three cases produce an *uneven match*, *i.e.*, the matched pair significantly differ in function values, or are not in the neighborhood of each other. In order to avoid such matches, we perform an additional pruning step to remove such low weight edges from the bipartite graph. Thus this step ensures that there is no significant change in the similarity measures in such cases. Note that we use a value of 10^{-6} in this filtration step, thus ensuring that significant matched pairs are not removed.

On the other hand, let the weights of the edges under consideration be high. Given the exponential decrease in the weights with respect to distance between the maxima, we can safely assume that the two maxima are in the neighborhood of x_i. Assuming that the neighborhood size r is not large (we use 0.1 in our experiments), we can safely infer that the two function values are similar (and high). Thus, there is no effect on the functional similarity. Let s be a saddle that can be reached through a descending path from both y_j and y_k. If there are no other matches in both the sub-trees, from y_j to s and from y_k to s, then there is no change to the topological similarity measure. In case there are other matches, then the persistence of the two maxima in their respective sub-trees decides the maximum change in the topological similarity, which is bounded by $|\pi_{y_k} - \pi_{y_j}|$. As we show in Section 6.6, we found that in practice the changes to topological similarity was indeed small due to noise.

6.5 Exploration Framework

We design a visual interface to help ecologists explore multiple SDMs. We accomplish this through the use of multiple visualizations. The interface consists of 4 views.

6.5.1 Properties View

A matrix is used to represent various properties of different models, as well as the difference pairs of models. The diagonal of this matrix represents the properties of the individual models. The functional distance ϕ between the pairs of models is represented in the upper triangular matrix, while the topological distance τ is

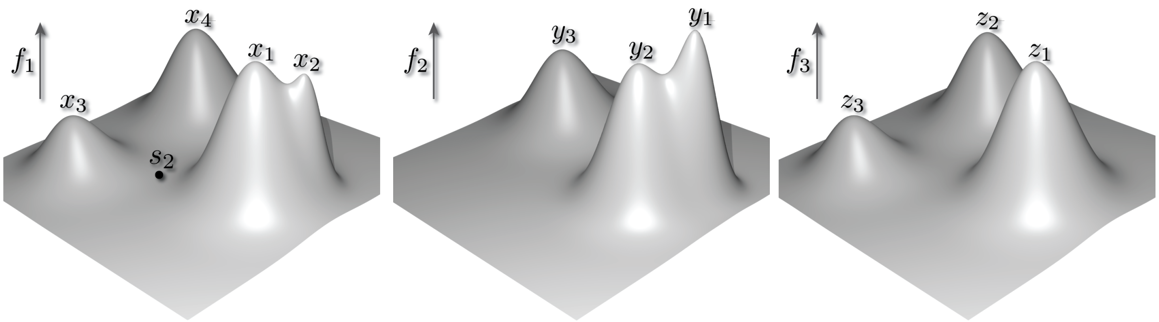
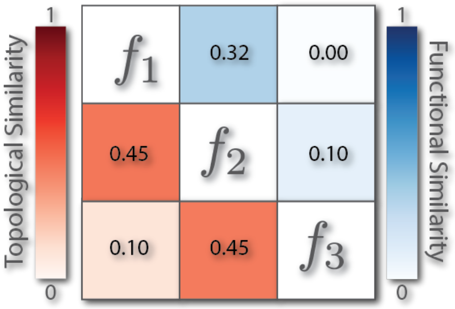


Figure 6.6: **We compare three functions** – f_1, f_2, and f_3, and use this comparison to demonstrate the visualization interface. f_1 and f_2 are the same functions as used in the earlier examples.


Topological Similarity
1
0
Functional Similarity
1
0

f_1	0.32	0.00
0.45	f_2	0.10
0.10	0.45	f_3

Figure 6.7: **The properties view** summarizes the topological similarity and functional similarity between the three sample functions. Note that a lower value is better.

represented in the lower triangular matrix. Figure 6.7 shows the properties view for the three sample functions shown in Figure 6.6. In case of functions f_1 and f_2, the difference is the presence of peak x_3 in f_1, which contributes to the topological similarity. In case of f_2 and f_3, peak equivalent to z_3 is missing in f_2, while a peak equivalent to y_1 is missing in f_3. However, the simplification required to remove z_3 is greater than that required for y_1, which is denoted by their topological similarity.

In order to explore a single SDM, the user can select the diagonal element of this matrix. On the other hand, to compare the given pair of models, the user can select the corresponding cell of the matrix. During this comparison, the user can select to view either the similarities between the models, or the differences.

6.5.2 Features View

This view visualizes the topological features of the selected model(s) as a scatter plot. Each point in the scatter plot corresponds to a topological feature (maximum or

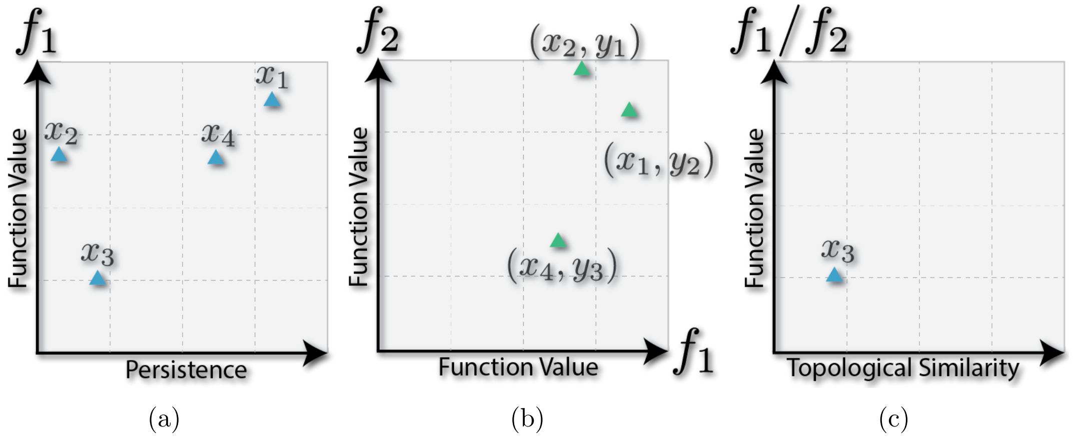


Figure 6.8: **The features view** is a scatter plot denoting the various extrema of the functions. **(a)** When exploring a single function, each point corresponds to the extrema of the function. Here, we show the set of maxima of f_1. **(b)** When exploring the similarities between two functions, each point corresponds to a pair of extrema that are matched. The figure shows the matching between the maxima of f_1 and f_2. **(c)** When exploring differences between two functions, each point corresponds to an extremum that is present in one function, but not in the other. When comparing f_1 and f_2, the maximum x_3 is absent from f_2.

minimum). The axes of the scatter plot are defined based on what the user wants to explore.

Explore single model. In this case, the x-axis of the model corresponds to the persistence (topological significance) of the extrema, while the y-axis corresponds to its function value. This allows the user to choose features during the exploration. For example, in case users are not interested in extrema with a small function value, then they can focus at the appropriate portion of the plot. The extrema of the function f_1 is shown in Figure 6.8(a).

Explore similarities between two models. In this case, each point in the scatter plot corresponds to a pair of extrema that are similar, that is, the pair of extrema that match. The axes corresponds to the function values of the two extrema. This view also provides the intuition for the functional similarity. A functionally similar pair of functions should have all points along the diagonal in this plot. Divergence from the diagonal denotes a disparity in the function values between the two functions in the parameter space in the neighborhood of the extreme point. Figure 6.8(b) illustrates the different matches found between f_1 and f_2 (also see Figure 6.3(d)).

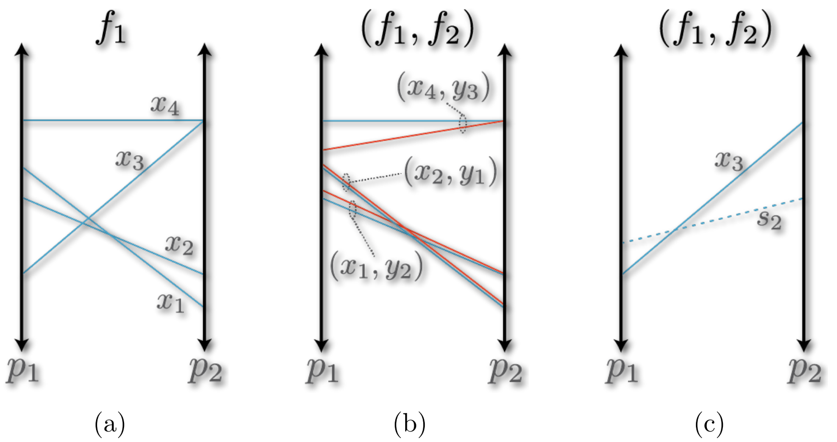


Figure 6.9: **Parallel coordinates view** is used to view the location of extrema of interest in the high dimensional predictor space. **(a)** The set of maxima of f_1. **(b)** The matched set of maxima between f_1 and f_2. **(c)** The maximum that is absent in f_2. We also show the corresponding critical point pair (saddle s_2) when exploring the differences.

Explore differences between two models. In this case, each point in the scatter plot corresponds to an extremum that is present in one function but absent in the other. The point is color coded to denote the function it is part of. The x-axis in this case corresponding to the topological similarity measure, while the y-axis corresponds to the function value. Figure 6.8(c) shows the difference between functions f_1 and f_2.

In this view, the maxima are represented as upward pointing triangles, while the minima are represented as downward pointing triangles.

6.5.3 Parallel Coordinates View

Once features of interest are chosen, the spatial region in the domain corresponding to the selected features are visualized using the parallel coordinates view. This view provides information on the location of the selected extrema in the high dimensional space. Figure 6.9(a) illustrates the locations of all maxima of the function f_1. Figure 6.9(b) and 6.9(c) show the matched maxima and the differing maximum respectively.

6.5.4 Response Curve View

A response curve represents a one dimensional slice of the function. We include this view in our interface since it helps the ecologists understand the different features as they are familiar with this representation. By selecting a feature and predicate

# vertices	# edges	Running time (ms)	
		Merge Trees	Matching
10,000	249,716	206	8
100,000	2,420,083	1,811	13
1,000,000	23,642,201	129,576	44

Table 6.1: Time taken to compute the similarity between two models.

of interest from the parallel coordinate view, the user can view the response curves with respect to the selected predicate. The values of the other predicates are set to those corresponding to the selected extremum. We also show the response curve of the critical point pair corresponding to an extremum. This helps users understand how the function changes. For example, when viewing a minimum-saddle pair, the upward movement of the response curve indicates the approximate shape of the corresponding "valley" in the high dimensional space.

6.6 Experiments

We implemented both the similarity computation and the visualization interface using Java. We used the lemon [174] library for computing the maximum weight bipartite matching. All experiments were run on a MacBook Pro with 2.3 GHz Intel Core i7 processor and 16 GB of memory. In this section, we first report running times for computing the similarity between two function. Next, we demonstrate the robustness of the similarity measure with respect to noise. All our experiments were conducted over three different species data sets, where four models were used for each data set.

6.6.1 Efficiency

In order to test the efficiency of our model comparison technique, we created PL functions by varying the number of sample points. Table 6.1 shows the computation times for each step of our algorithm. Note that computing the initial set of merge trees is a one time operation per model. By pre-computing this, we can accomplish interactive performance even for large graphs having a million sample points and over 23 million edges.

We perform an additional optimization of removing small weighted edges prior to computing the matching instead of a post removal to handle noise (see Section 6.4.3). Thus, given the small number of critical points (compared to the input size) and the above filtering step, the size of the resulting bipartite graph is relatively small. We therefore achieve fast computation of the matching, even though the algorithm has a cubic running time on the number of nodes in the bipartite graph.

6.6.2 Robustness to Noise

In order to test the robustness to noise, we perform two types of experiments. In the first experiment, we fix a function f_1, and artificially induce noise to f_1 to obtain a noisy function f_1^*. The amount of noise induced was bounded by $\epsilon = 10^{-4}$. We then compute the similarity measures between f_1 and f_1^*. Ideally the topological similarity should be *zero*. We performed this experiment for the different models across three data sets. The mean and standard deviation of the topological similarity across these tests were 1.18×10^{-4} and 4.05×10^{-5} respectively. Note that this is less than the 2ϵ bound (see Section 6.4.3).

In the second experiment, we consider pairs of functions, f_1 and f_2. We induce noise into one of the functions, say $f_2^* = f_2 + \texttt{noise}$. We then computed the similarity between f_1 and f_2^*. Ideally, the topological similarity between f_1 and f_2 should be the same as f_1 and f_2^* (*i.e.*, the difference should be 0). In this scenario, we found the mean difference in the topological similarity to be 6.42×10^{-5} and standard deviation to be 6.29×10^{-5}. When looking at individual errors, we found that in several cases, there was no change in the topological similarity demonstrating the robustness of the measure to noise.

6.7 Case Studies

In this section we describe two use case scenarios that are of interest to ecologists. The first case shows how the extrema in the different models can be used to guide ecologists towards interesting features of the model. The second case demonstrates how our similarity comparison technique can be used to identify differences between the models that are otherwise difficult to find.

6.7.1 Exploring an SDM

In this use case, the user is interested in exploring the properties of a single SDM. Using the visual interface, the user first selects the species and model algorithm of interest. Details on the modeling techniques and data are given in [158].

In the first experiment, the user chooses the `Brewers Sparrow` model for the `Brewers Sparrow` species. Figure 6.10(a) shows the set of extrema of this model. An

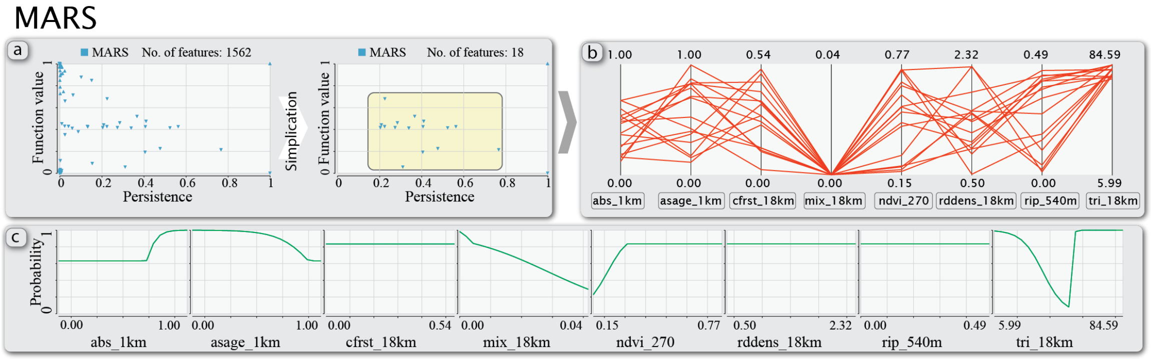


Figure 6.10: **Exploring the features of the `MARS` model for the `Brewers Sparrow` data set. (a)** Given the set of all extrema, the user simplifies to remove all those extrema having persistence less than 0.2. Note that this removes all maxima except the global maximum (at location $(1, 1)$ in the scatter plot). **(b)** The locations of the selected set of minima of the `MARS` model are shown using the parallel coordinates view. **(c)** Note that it is difficult to grasp the presence of high persistent minima (deep valleys) using the default response curves that is common in the analysis of this data.

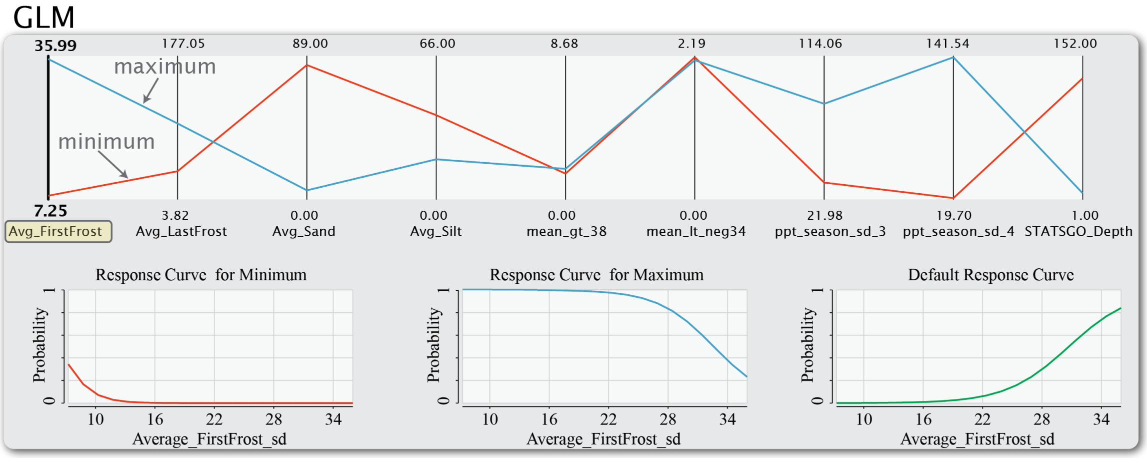


Figure 6.11: **Exploring the `GLM` model for the `Spruce Fir` species data.** Note that the response curves corresponding to the global maximum and minimum of this function is significantly different from the default response curves.

initial simplification is performed to remove noise / less significant extrema. Note that for the `MARS` model, there is a high number of significant minima. Figure 6.10(b) shows the different predictors corresponding to the set of selected minima. It is interesting to note that all of these minima occur when combination of values of `mix_18km` is low and `tri_18km` is high. Such a behavior is clearly not visible using the default response curves [136] shown in Figure 6.10(c).

In the next experiment, the user is interested in exploring the `GLM` model for the Spruce Fir species. The user selects the global maximum to locate the region corresponding to this maximum. The parallel coordinates view, shown in Figure 6.11, shows the coordinates of the maximum together with its critical point pair, which in this case is the global minimum. Note that the response curves at these locations are significantly different from the default response curve.

6.7.2 Exploring differences between given pair of models of a fixed Species

In this experiment, the user first selects the pair of models that are to be compared from the Properties view. The user can now view either the non-matched features or the matched features.

The first experiment considers the differences between MARS and other models for the Brewers Sparrow species. Users can filter features (extrema) having low topological similarity. As shown in the previous use case, the MARS model for Brewers Sparrow contains a large number of significant minima. It can be seen than these minima do not match with any minima of the other models, i.e., there exists no minima in the other models in the corresponding locations. This is illustrated in Figure 6.12 where we look at the different extrema in the features view.

Let us now select a significant difference between GLM and MARS (having high value of τ). Figure 6.13 shows the coordinates of the minimum-saddle pair (intuitively the lowest and highest point of the valley corresponding to the minimum) that is present in MARS, but absent in GLM. The response curves varying the predictor ndvi_270 at the minimum and saddle points shows a significant increase in the shape of the curve indicating a "valley"-like structure in MARS. However, we see a slight decrease in the response curve for GLM indicating the absence of a minimum in that region (and thus

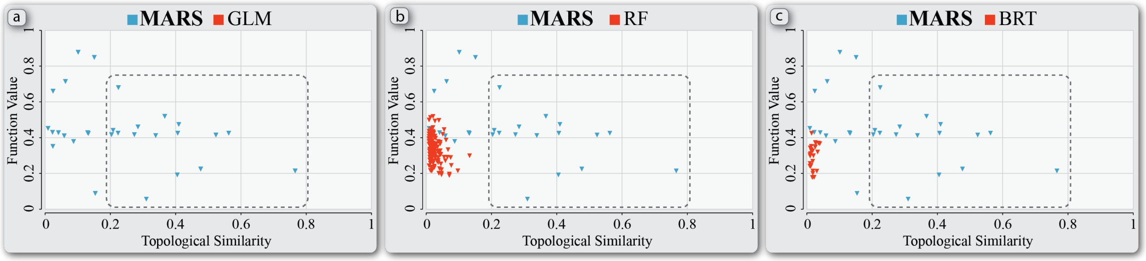


Figure 6.12: **Comparing MARS with other models for the Brewers Sparrow species.** Note that multiple significant minima that are present in MARS are not present in the other models. Also, these constitute the significant differences between these models.

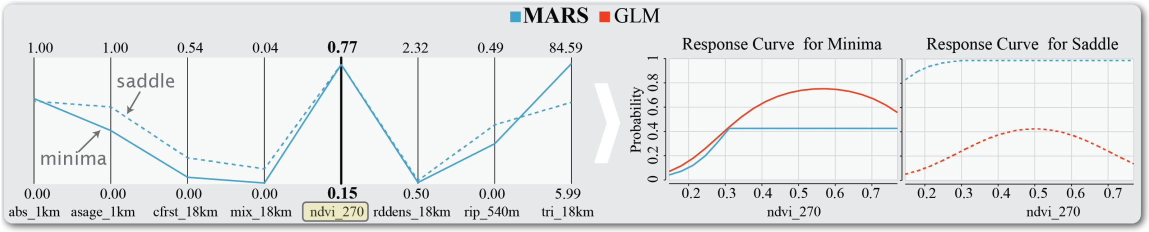


Figure 6.13: **Locations of a significant minimum-saddle pair in MARS is shown using parallel coordinates.** Note the moving up of the response curves of MARS from the minimum to the saddle. At the same location, we see a different behavior for the GLM model.

a difference in the function).

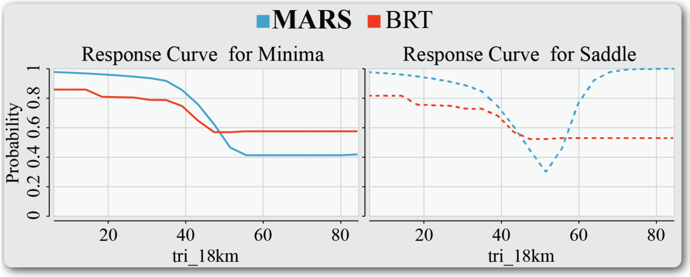


Figure 6.14: **The response curves corresponding to a minimum-saddle pair** of a minimum present in MARS but absent in BRT for the Brewers Sparrow species.

We can notice a similar pattern when comparing MARS with BRT for Brewers Sparrow. Again, one of the significant minimum that is absent in BRT is chosen. Figure 6.14 shows the response curves corresponding to the minimum and saddle for this difference with respect to the predictor tri_18km.

In the next experiment, the user compares the differences between MARS and BRT for Sage Brush. In particular, the user selects the set of significant maxima (having topological similarity > 0.15) in BRT that are not present in MARS (Figure 6.15(a)). Figure 6.15(b) shows the coordinates corresponding to these maxima. Note that all of these differences occur when the value of mean_summer is low. This is counter intuitive when one looks at the default response curves of these two models (Figure 6.15(d)).

6.7.3 Feedback from Ecologists

When we initially provided our tool to the ecologists, they found the results to be a little too abstract and had difficulty in comprehending them. To help them get familiar and better understand the utility of working directly in the high dimensional

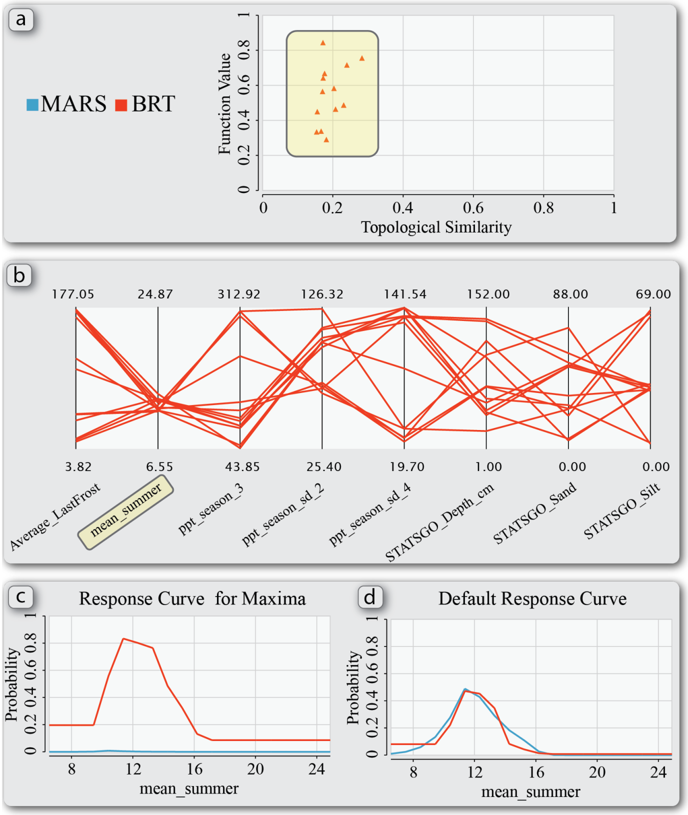


Figure 6.15: **Comparing MARS and BRT for the Sage Brush species data. (a)** Selecting all the significant maxima that are present in BRT but not in MARS. **(b)** Note that such difference mainly occurs at a relatively low value of the mean_summer predictor. **(c)** The response curve at one of the maximum. **(d)** Note that this behavior is counter intuitive to the default response curve, in which we see both MARS and BRT having the same pattern.

space, we used a two dimensional slice of the different models, and setup the software to work with this data. Their familiarity with the features in low dimensions allowed them to relate to the results from our tool. Also, since they could easily visualize the 2D data, the different features were directly apparent.

The examples presented in the previous section highlight the complexity of the response surface when considering an eight dimensional space (that is, using eight predictor variables) and clearly provide new information about the various models used by the ecologists. However, the implications of some of these results was not immediately apparent, which we plan to explore further in the future. As an ecologist collaborator mentioned during one of our interactions:

They also found utility in using our tool as the following response indicates:

> *"The default partial dependence plots show the model prediction for one*

> *variable when all other variables are collapsed to their mean value. However, the models are not applied in this collapsed space but rather in the full n-dimensional space and therefore it is suitable to use tools that allow the user to evaluate the response in the full n-dimensional space."*

In some cases when there is a difference between two models, it is possible that this is due to missing data. In such cases ecologists would have to collect additional data from regions having the differences. So, such a tool can also help in identifying these regions of discrepancies.

6.8 Discussions

6.8.1 Discretization of a High Dimensional Function

Identifying an ideal sample size to represent a high dimensional function is a difficult problem. For all the experiments in this paper, we used a sample size of 100,000 points. We chose this size since we found that the similarity measure computed did not significantly change even on increasing the sample size to above 100,000. This is because increasing the sample size only created noisy extrema which did not affect the similarity measures.

6.8.2 Neighborhood Radius

The neighborhood radius used for weighing the edge weights of the bipartite graph is largely dependent on the application and domain expertise. We used a neighborhood radius $r = 0.1$ for this purpose, and was based on discussions with the ecologists, who did not want the matching features to be far away. It would however be interesting to study the performance of the similarity with varying radius. Patterns from such an experiment could not only help understand the behavior of the functions, but could also help automatically identify the radius.

6.9 Summary

In this chapter, we introduced a topology-based framework that helps ecologist to better understand SDMs and guide them towards interesting features of the model. We also proposed the concept of maximum topology matching that can be used to identify similarities and differences between a given pair of SDMs. Even though the focus was on the ecology domain, our technique is general and can be applied in cases requiring a locality-aware way of comparing scalar functions.

Chapter 7

Conclusion and Future Work

Analyzing multifaceted data presents several challenges for data analysis, cutting across different domains. In this dissertation, by focusing on the specific area of climate model intercomparison, we conducted a thorough investigation of the problem from different perspectives. Through a qualitative study of visualization usage by climate scientists, we were able to reflect on the state-of-the-art in climate data visualization, and identify complex analysis scenarios which require novel visualization solutions [4]. This led to the development of visual analytics techniques for similarity analysis of climate models [6], reconciliation of multiple similarity spaces [7], and maximum topology matching for exploring differences in various climate models directly in the high dimensional domain [8]. Based on the findings and recommendations of our qualitative study, we also conducted a quantitative study for looking at how color maps affect climate model analysis tasks [5]. The concepts, techniques, and studies discussed in this dissertation present new opportunities for research in visualization and visual analytics along different directions, which we discuss below.

The taxonomy of design problems proposed in this dissertation is a first attempt towards understanding how visualization design problems are instantiated in practice. By scaling this approach to multiple domains, we can not only aim to provide improved visualization solutions to domain experts, but also reflect upon the theoretical principles of visualization. An eventual goal is to build a system which extends the best of UV-CDAT [97] and Tableau [175] features: a visualization recommendation system, which adapts to domain experts analytical tasks and provides provenance at the back-end for reproducing the steps which led to the creation of the plots. Especially in scientific disciplines like climate science, biology, etc., where domain experts need to trust what they see on screen, such a provenance-enabled system which helps produce optimal visualizations will fill a gap in the existing data analysis practices.

The results of the study on usage of color maps reveals that appreciation of visualization of best practices is imperative for accuracy of tasks. We believe our study will provide further incentive for conducting more studies on other aspects of effects visualization parameters on real-world data analysis tasks. Further, the results of both the studies also demonstrate the need to rethink the evaluation criteria of visualizations in terms of the high-level goal: whether they are used for analysis or for communication of insights. In visualization, while a lot of research has been addressed for establishing criteria for analysis purposes [9, 33], much less attention has been given to scenarios when domain experts need to convey their insights to a broader audience through visualizations. One of the recent developments in this direction has been the research on storytelling in visualization [57]. An open area of research is to know what levels of abstracts and criteria for effectiveness we should use for visual communication of scientific insights. The findings of our studies can provide a starting point in that respect.

In SimilarityExplorer, the approach of providing multiple perspectives on occurrence and causality of similarity is generalizable to other domains that involve spatiotemporal data, like urban data. We are looking forward to add more features to, and apply SimilarityExplorer for solving problems related to such different domains.

The visual reconciliation technique is not restricted to the climate science domain. One potential application can be in the in the healthcare domain, where the goal is to reconcile patient similarity with drug similarity for personalized medicine development [105]. Another potential application is in the product design domain. For example in the automotive market, car models can be qualified by multitude of features. It will be of interest to automotive companies to reconcile similarity of car models based on their descriptors, with the similarity based on transaction data. In short, we posit that visual reconciliation can potentially serve as an important analytics paradigm for making sense of the ever-growing variety of available data and their diverse similarity criteria. However, we identified multiple challenges that need to be addressed before being a general technique. For instance, make it more scalable, extent the framework to more complex models of time, and increasing diversity of descriptors.

Even though the focus of out maximum topology matching technique was on the climate domain, our technique is general and can be applied in cases requiring a locality-aware way of comparing scalar functions. In addition, because all the climate data sets used in this dissertation are spatiotemporal. Then, techniques and concepts presented in this dissertation can be used to analyze similarities in spatiotemporal data and further problem areas can be addressed through extension of our work.

Bibliography

[1] P. Fox and J. Hendler, "Changing the equation on scientific data visualization," *Science(Washington)*, vol. 331, no. 6018, pp. 705–708, 2011.

[2] D. Huntzinger, W. M. Post, Y. Wei, A. Michalak, *et al.*, "North american carbon program (nacp) regional interim synthesis: Terrestrial biospheric model intercomparison," *Ecological Modelling*, vol. 232, pp. 144–157, 2012.

[3] J. Elith and J. R. Leathwick, "Species distribution models: Ecological explanation and prediction across space and time," *Annual Review of Ecology, Evolution, and Systematics*, vol. 40, no. 1, pp. 677–697, 2009.

[4] A. Dasgupta, J. Poco, Y. Wei, R. Cook, E. Bertini, and C. Silva, "Bridging theory with practice: An exploratory study of visualization use and design for climate model comparison (in press)," *IEEE Transactions on Visualization and Computer Graphics*, 2015.

[5] A. Dasgupta, J. Poco, R. Bernice, E. Bertini, and C. Silva, "Perceptual evaluation of color scales for geospatial analysis of climate data (submitted)," *IEEE Transactions on Visualization and Computer Graphics*, 2015.

[6] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. Schwalm, R. Cook, E. Bertini, and C. Silva, "SimilarityExplorer: A visual inter-comparison tool for multifaceted climate data," *CGF*, vol. 33, no. 3, pp. 341–350, 2014.

[7] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. Schwalm, D. Huntzinger, R. Cook, E. Bertini, and C. Silva, "Visual reconciliation of alternative similarity spaces in climate modeling," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1923–1932, 2014.

[8] J. Poco, H. Doraiswamy, M. Talbert, J. Morisette, and C. Silva, "Using maximum topology matching to explore differences in species distribution models

(submitted)," *IEEE Transactions on Visualization and Computer Graphics*, 2015.

[9] E. Tufte, *The visual display of quantitative information*, vol. 31. Graphics press, 1983.

[10] S. Few, "Show me the numbers," *Designing Tables and Graphs to Enlighten*, 2012.

[11] L. Grammel, M. Tory, and M. Storey, "How information visualization novices construct visualizations," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 16, no. 6, pp. 943–952, 2010.

[12] J. Heer, F. van Ham, S. Carpendale, C. Weaver, and P. Isenberg, "Creation and collaboration: Engaging new audiences for information visualization," in *Information Visualization*, pp. 92–133, Springer, 2008.

[13] N. Kodagoda, B. Wong, C. Rooney, and N. Khan, "Interactive visualization for low literacy users: from lessons learnt to design," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1159–1168, ACM, 2012.

[14] B. chul Kwon, B. Fisher, and J. S. Yi, "Visual analytic roadblocks for novice investigators," in *Visual Analytics Science and Technology (VAST), 2011 IEEE Conference on*, pp. 3–11, IEEE, 2011.

[15] J. Walny, B. Lee, P. Johns, N. Henry Riche, and S. Carpendale, "Understanding pen and touch interaction for data exploration on interactive whiteboards," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 12, pp. 2779–2788, 2012.

[16] P. Isenberg, D. Fisher, S. A. Paul, M. R. Morris, K. Inkpen, and M. Czerwinski, "Co-located collaborative visual analytics around a tabletop display," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 18, no. 5, pp. 689–702, 2012.

[17] S. K. Card, J. D. Mackinlay, and B. Shneiderman, *Readings in information visualization: using vision to think*. Morgan Kaufmann, 1999.

[18] E. H. Chi, "A taxonomy of visualization techniques using the data state reference model," in *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, pp. 69–75, IEEE, 2000.

[19] W. Colin, *Information visualization: perception for design*, vol. 5. 2000.

[20] A. J. Pretorius and J. J. Van Wijk, “What does the user want to see? what do the data want to be?,” *Information Visualization*, vol. 8, no. 3, pp. 153–166, 2009.

[21] C. Weaver, D. Fyfe, A. Robinson, D. Holdsworth, D. Peuquet, and A. M. MacEachren, “Visual analysis of historic hotel visitation patterns,” in *Visual Analytics Science And Technology, 2006 IEEE Symposium On*, pp. 35–42, IEEE, 2006.

[22] A. Dasgupta, M. Chen, and R. Kosara, “Conceptualizing visual uncertainty in parallel coordinates,” in *Computer Graphics Forum*, vol. 31, pp. 1015–1024, Wiley Online Library, 2012.

[23] J. Walny, S. Carpendale, N. Henry Riche, G. Venolia, and P. Fawcett, “Visual thinking in action: Visualizations as used on whiteboards,” *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2508–2517, 2011.

[24] S. Solomon, *Climate change 2007-the physical science basis: Working group I contribution to the fourth assessment report of the IPCC*, vol. 4. Cambridge University Press, 2007.

[25] J. Nielsen and R. Molich, “Heuristic evaluation of user interfaces,” in *In Proceedings, SIGCHI conference on Human factors in computing systems*, pp. 249–256, ACM, 1990.

[26] T. Zuk, L. Schlesier, P. Neumann, M. S. Hancock, and S. Carpendale, “Heuristics for information visualization evaluation,” in *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*, pp. 1–6, ACM, 2006.

[27] C. Forsell and J. Johansson, “An heuristic set for evaluation in information visualization,” in *Proceedings of the International Conference on Advanced Visual Interfaces*, pp. 199–206, ACM, 2010.

[28] K. Charmaz, *Constructing grounded theory: A practical guide through qualitative analysis.* Sage Publications Limited, 2006.

[29] P. Isenberg, T. Zuk, C. Collins, and S. Carpendale, “Grounded evaluation of information visualizations,” in *Proceedings of the 2008 conference on BEyond*

time and errors: novel evaLuation methods for Information Visualization, p. 6, ACM, 2008.

[30] M. Tory and S. Staub-French, “Qualitative analysis of visualization: a building design field study,” in *Proceedings of the 2008 conference on BEyond time and errors: novel evaLuation methods for Information Visualization*, p. 7, ACM, 2008.

[31] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory, “A taxonomy of visual cluster separation factors,” in *Computer Graphics Forum*, vol. 31, pp. 1335–1344, Wiley Online Library, 2012.

[32] J. Hullman, S. Drucker, N. Henry Riche, B. Lee, D. Fisher, and E. Adar, “A deeper understanding of sequence in narrative visualization,” *Visualization and Computer Graphics, IEEE Transactions on*, vol. 19, no. 12, pp. 2406–2415, 2013.

[33] J. Mackinlay, “Automating the design of graphical presentations of relational information,” *ACM Transactions on Graphics (TOG)*, vol. 5, no. 2, pp. 110–141, 1986.

[34] J. W. Tukey, “Graphic comparisons of several linked aspects: Alternatives and suggested principles,” *Journal of Computational and Graphical Statistics*, vol. 2, no. 1, pp. 1–33, 1993.

[35] J. Bertin, “Semiology of graphics: diagrams, networks, maps,” 1983.

[36] S. K. Card and J. Mackinlay, “The structure of the information visualization design space,” in *Information Visualization, 1997. Proceedings., IEEE Symposium on*, pp. 92–99, IEEE, 1997.

[37] A. M. MacEachren, *How maps work: representation, visualization, and design.* Guilford Press, 2004.

[38] W. S. Cleveland and R. McGill, “Graphical perception: The visual decoding of quantitative information on graphical displays of data,” *Journal of the Royal Statistical Society. Series A (General)*, pp. 192–229, 1987.

[39] A. Cairo, *The Functional Art: An introduction to information graphics and visualization.* New Riders, 2012.

[40] C. E. Shannon, “A mathematical theory of communication,” *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[41] C. Jiménez, C. Prigent, B. Mueller, S. Seneviratne, M. McCabe, E. Wood, W. Rossow, G. Balsamo, A. Betts, P. Dirmeyer, *et al.*, "Global intercomparison of 12 land surface heat flux estimates," *Journal of Geophysical Research: Atmospheres (1984–2012)*, vol. 116, no. D2, 2011.

[42] B. E. Rogowitz, L. A. Treinish, and S. Bryson, "How not to lie with visualization," *Computers in Physics*, vol. 10, no. 3, pp. 268–273, 1996.

[43] D. Borland and R. M. Taylor, "Rainbow color map (still) considered harmful," *Computer Graphics and Applications*, vol. 27, no. 2, pp. 14–17, 2007.

[44] M. Borkin, K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, and H. Pfister, "Evaluation of artery visualizations for heart disease diagnosis," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 17, no. 12, pp. 2479–2488, 2011.

[45] M. Harrower and C. A. Brewer, "Colorbrewer. org: An online tool for selecting colour schemes for maps," *The Cartographic Journal*, vol. 40, no. 1, pp. 27–37, 2003.

[46] R. Rosenholtz, Y. Li, J. Mansfield, and Z. Jin, "Feature congestion: a measure of display clutter," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 761–770, ACM, 2005.

[47] T. Keenan, I. Baker, A. Barr, P. Ciais, K. Davis, M. Dietze, D. Dragoni, C. M. Gough, R. Grant, D. Hollinger, *et al.*, "Terrestrial biosphere model performance for inter-annual variability of land-atmosphere co2 exchange," *Global Change Biology*, vol. 18, no. 6, pp. 1971–1987, 2012.

[48] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts, "Visual comparison for information visualization," *Information Visualization*, vol. 10, no. 4, pp. 289–309, 2011.

[49] S. Van Den Elzen and J. J. Van Wijk, "Small multiples, large singles: A new approach for visual data exploration," *Computer Graphics Forum*, vol. 32, no. 3pt2, pp. 191–200, 2013.

[50] D. Kahneman and A. Tversky, "Choices, values, and frames.," *American psychologist*, vol. 39, no. 4, p. 341, 1984.

[51] K. Risden, M. P. Czerwinski, T. Munzner, and D. B. Cook, "An initial examination of ease of use for 2d and 3d information visualizations of web content," *International Journal of Human-Computer Studies*, vol. 53, no. 5, pp. 695–714, 2000.

[52] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical studies in information visualization: Seven scenarios," *TVCG*, vol. 18, no. 9, pp. 1520–1536, 2012.

[53] J. B. Tenenbaum, "Mapping a manifold of perceptual observations," in *Advances in Neural Information Processing Systems 10*, pp. 682–688, MIT Press, 1998.

[54] E. Gomez-Nieto, W. Casaca, L. G. Nonato, and G. Taubin, "Mixed integer optimization for layout arrangement," in *Proceedings,Conference on Graphics, Patterns and Images (SIBGRAPI)*, IEEE, 2013.

[55] T. T. Allen, *Introduction to engineering statistics and six sigma: statistical quality control and design of experiments and systems*. Springer, 2005.

[56] K. Moreland, "Diverging color maps for scientific visualization," in *Advances in Visual Computing*, pp. 92–103, Springer, 2009.

[57] R. Kosara and J. Mackinlay, "Storytelling: The next step for visualization," *Computer*, vol. 46, no. 5, pp. 44–50, 2013.

[58] C. Demiralp, C. Scheidegger, G. Kindlmann, D. Laidlaw, and J. Heer, "Visual embedding: A model for visualization," *Computer Graphics and Applications*, 2014.

[59] B. E. Rogowitz and L. A. Treinish, "Data visualization: the end of the rainbow," *Spectrum, IEEE*, vol. 35, no. 12, pp. 52–59, 1998.

[60] S. Silva, B. Sousa Santos, and J. Madeira, "Using color in visualization: A survey," *Computers & Graphics*, vol. 35, no. 2, pp. 320–333, 2011.

[61] P. K. Robertson, "Visualizing color gamuts: A user interface for the effective use of perceptual color spaces in data displays," *Computer Graphics and Applications, IEEE*, vol. 8, no. 5, pp. 50–64, 1988.

[62] B. E. Rogowitz, D. T. Ling, and W. A. Kellogg, “Task dependence, veridicality, and preattentive vision: taking advantage of perceptually rich computer environments,” in *SPIE/IS&T 1992 Symposium on Electronic Imaging: Science and Technology*, pp. 504–513, International Society for Optics and Photonics, 1992.

[63] S. S. Stevens, *Psychophysics.* Transaction Publishers, 1975.

[64] B. E. Rogowitz, A. D. Kalvin, A. Pelah, and A. Cohen, “Which trajectories through which perceptually uniform color spaces produce appropriate colors scales for interval data?,” in *Color and Imaging Conference*, vol. 1999, pp. 321–326, Society for Imaging Science and Technology, 1999.

[65] B. Rogowitz and A. D. Kalvin, “The ”which blair project”: a quick visual method for evaluating perceptual color maps,” in *Visualization, 2001. VIS’01. Proceedings*, pp. 183–556, IEEE, 2001.

[66] G. Kindlmann, E. Reinhard, and S. Creem, “Face-based luminance matching for perceptual colormap generation,” in *Proceedings of the conference on Visualization’02*, pp. 299–306, IEEE Computer Society, 2002.

[67] C. A. Brewer, A. M. MacEachren, L. W. Pickle, and D. Herrmann, “Mapping mortality: Evaluating color schemes for choropleth maps,” *Annals of the Association of American Geographers*, vol. 87, no. 3, pp. 411–438, 1997.

[68] C. A. Brewer and L. Pickle, “Evaluation of methods for classifying epidemiological data on choropleth maps in series,” *Annals of the Association of American Geographers*, vol. 92, no. 4, pp. 662–681, 2002.

[69] A. M. MacEachren, C. A. Brewer, and L. W. Pickle, “Visualizing georeferenced data: representing reliability of health statistics,” *Environment and Planning A*, vol. 30, no. 9, pp. 1547–1561, 1998.

[70] Kitware, “The Visualization Toolkit (VTK) and Paraview.” `http://www.kitware.com`.

[71] L. D. Bergman, B. E. Rogowitz, and L. A. Treinish, “A rule-based tool for assisting colormap selection,” in *Proceedings Conference on Visualization*, p. 118, IEEE Computer Society, 1995.

[72] C. Tominski, G. Fuchs, and H. Schumann, “Task-driven color coding,” in *Information Visualisation, 2008. IV’08. 12th International Conference*, pp. 373–380, IEEE, 2008.

[73] C. Ware, "Color sequences for univariate maps: Theory, experiments and principles," *Computer Graphics and Applications, IEEE*, vol. 8, no. 5, pp. 41–49, 1988.

[74] J. Tajima, "Uniform color scale applications to computer graphics," *Computer Vision, Graphics, and Image Processing*, vol. 21, no. 3, pp. 305–325, 1983.

[75] H. Levkowitz and G. T. Herman, "Color scales for image data," *IEEE Computer Graphics and Applications*, vol. 12, no. 1, pp. 72–80, 1992.

[76] M. X. Zhou and S. K. Feiner, "Visual task characterization for automated visual discourse synthesis," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 392–399, ACM Press/Addison-Wesley Publishing Co., 1998.

[77] R. Amar, J. Eagan, and J. Stasko, "Low-level components of analytic activity in information visualization," in *IEEE Symposium on Information Visualization*, pp. 111–117, IEEE, 2005.

[78] D. A. Keim, "Designing pixel-oriented visualization techniques: Theory and applications," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 6, no. 1, pp. 59–78, 2000.

[79] J. Kehrer and H. Hauser, "Visualization and visual analysis of multifaceted scientific data: A survey," *IEEE Trans. on Visualization and Computer Graphics*, vol. 19, no. 3, pp. 495–513, 2013.

[80] T. Munzner, "A nested model for visualization design and validation," *IEEE Trans. on Visualization and Computer Graphics*, vol. 15, no. 6, pp. 921–928, 2009.

[81] D. J. Peuquet, "It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems," *Annals of the Association of American Geographers*, vol. 84, no. 3, pp. 441–461, 1994.

[82] G. Andrienko, N. Andrienko, U. Demsar, D. Dransch, *et al.*, "Space, time and visual analytics," *International Journal of Geographical Information Science*, vol. 24, no. 10, pp. 1577–1600, 2010.

[83] N. Andrienko and G. Andrienko, "A visual analytics framework for spatio-temporal analysis and modelling," *Data Min. Knowl. Discov.*, vol. 27, pp. 55–83, July 2013.

[84] A. Malik, R. Maciejewski, N. Elmqvist, Y. Jang, *et al.*, "A correlative analysis process in a visual analytics environment," in *In Proc. IEEE Conference on Visual Analytics Science and Technology*, pp. 33–42, 2012.

[85] R. Maciejewski, S. Rudolph, R. Hafen, A. Abusalah, *et al.*, "A visual analytics approach to understanding spatiotemporal hotspots," *IEEE Trans. on Visualization and Computer Graphics*, vol. 16, no. 2, pp. 205–220, 2010.

[86] R. Maciejewski, R. Hafen, S. Rudolph, S. G. Larew, *et al.*, "Forecasting hotspots a predictive analytics approach," *IEEE Trans. on Visualization and Computer Graphics*, vol. 17, no. 4, pp. 440–453, 2011.

[87] G. Andrienko, N. Andrienko, S. Bremm, T. Schreck, *et al.*, "Space-in-time and time-in-space self-organizing maps for exploring spatiotemporal patterns," vol. 29, pp. 913–922, Wiley Online Library, 2010.

[88] A. Maries, N. Mays, M. Hunt, K. F. Wong, *et al.*, "Grace: A visual comparison framework for integrated spatial and non-spatial geriatric data," *IEEE Trans. on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2916–2925, 2013.

[89] D. Guo, J. Chen, A. M. MacEachren, and K. Liao, "A visualization system for space-time and multivariate patterns (vis-stamp)," *IEEE Trans. on Visualization and Computer Graphics*, vol. 12, no. 6, pp. 1461–1474, 2006.

[90] J. Kehrer, P. Muigg, H. Doleisch, H. Hauser, *et al.*, "Interactive visual analysis of heterogeneous scientific data across an interface," *IEEE Trans. on Visualization and Computer Graphics*, vol. 17, no. 7, pp. 934–946, 2011.

[91] J. Kehrer, F. Ladstadter, P. Muigg, H. Doleisch, A. Steiner, and H. Hauser, "Hypothesis generation in climate research with interactive visual data exploration," *IEEE Trans. on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1579–1586, 2008.

[92] H.-J. Schulz, T. Nocke, M. Heitzler, and H. Schumann, "A design space of visualization tasks," *IEEE Trans.on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2366–2375, 2013.

[93] F. Ladstädter, A. K. Steiner, B. C. Lackner, B. Pirscher, G. Kirchengast, J. Kehrer, H. Hauser, P. Muigg, and H. Doleisch, "Exploration of climate data using interactive visualization*," *Journal of Atmospheric and Oceanic Technology*, vol. 27, no. 4, pp. 667–679, 2010.

[94] C. A. Steed, G. Shipman, P. Thornton, D. Ricciuto, *et al.*, "Practical application of parallel coordinates for climate model analysis," *Procedia Computer Science*, vol. 9, no. 0, pp. 877 – 886, 2012.

[95] Lawrence Livermore National Laboratory, "VisIt: Visualize It in Parallel Visualization Application." `https://wci.llnl.gov/codes/visit`.

[96] J. Freire, D. Koop, E. Santos, C. Scheidegger, C. Silva, and H. T. Vo, *VisTrails*. Lulu Publishing, Inc., 2011.

[97] D. N. Williams, T. Bremer, C. Doutriaux, J. Patchett, S. Williams, G. Shipman, R. Miller, D. R. Pugmire, B. Smith, C. Steed, E. W. Bethel, H. Childs, H. Krishnan, P. Prabhat, M. Wehner, C. T. Silva, E. Santos, D. Koop, T. Ellqvist, J. Poco, B. Geveci, A. Chaudhary, A. Bauer, A. Pletzer, D. Kindig, G. L. Potter, and T. P. Maxwell, "Ultrascale visualization of climate data," *Computer*, vol. 46, no. 9, pp. 68–76, 2013.

[98] J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling, "Climate data challenges in the 21 st century," *Science*, vol. 331, no. 6018, pp. 700–702, 2011.

[99] M. Meyer, T. Munzner, A. DePace, and H. Pfister, "Multeesum: A tool for comparative spatial and temporal gene expression data," *IEEE Trans. on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 908–917, 2010.

[100] M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: reflections from the trenches and the stacks," *IEEE Trans. on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2431–2440, 2012.

[101] M. X. Zhou and S. K. Feiner, "Visual task characterization for automated visual discourse synthesis," in *In Proc. SIGCHI Conf. on Human factors in Computing Systems*, pp. 392–399, ACM, 1998.

[102] J. C. Roberts, "State of the art: Coordinated & multiple views in exploratory visualization," in *In Proc., Coordinated and Multiple Views in Exploratory Visualization*, pp. 61–71, IEEE, 2007.

[103] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *In Proc. IEEE Symp. on Visual Languages*, pp. 336–343, 1996.

[104] A. MacEachren, D. Xiping, F. Hardisty, D. Guo, and G. Lengerich, “Exploring high-d spaces with multiform matrices and small multiples,” in *In Proc. IEEE Symposium on Information Visualization*, pp. 31–38, 2003.

[105] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, “Towards personalized medicine: Leveraging patient similarity and drug similarity analytics,” *target*, vol. 1, no. 1, p. 1.

[106] D. Masson and R. Knutti, “Climate model genealogy,” *Geophysical Research Letters*, vol. 38, no. 8, 2011.

[107] L. Parida and N. Ramakrishnan, “Redescription mining: Structure theory and algorithms,” in *AAAI*, vol. 5, pp. 837–844, 2005.

[108] S. Rinzivillo, D. Pedreschi, M. Nanni, F. Giannotti, N. Andrienko, and G. Andrienko, “Visually driven analysis of movement data by progressive clustering,” *Information Visualization*, vol. 7, no. 3-4, pp. 225–239, 2008.

[109] T. Schreck, J. Bernard, T. Von Landesberger, and J. Kohlhammer, “Visual cluster analysis of trajectory data with interactive kohonen maps,” *Information Visualization*, vol. 8, no. 1, pp. 14–29, 2009.

[110] C. Turkay, A. Lundervold, A. J. Lundervold, and H. Hauser, “Representative factor generation for the interactive visual analysis of high-dimensional data,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2621–2630, 2012.

[111] D. N. Huntzinger, C. Schwalm, A. M. Michalak, K. Schaefer, *et al.*, “The north american carbon program multi-scale synthesis and terrestrial model intercomparison project - part 1: Overview and experimental design,” *Geoscientific Model Development Discussions*, vol. 6, no. 3, pp. 3977–4008, 2013.

[112] E. Bertini and D. Lalanne, “Surveying the complementary role of automatic data analysis and visualization in knowledge discovery,” in *Proceedings of the ACM SIGKDD Workshop on Visual Analytics and Knowledge Discovery*, pp. 12–20, ACM, 2009.

[113] D. A. Keim, F. Mansmann, and J. Thomas, “Visual analytics: how much visualization and how much analytics?,” *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 2, pp. 5–8, 2010.

[114] D. Pfitzner, R. Leibbrandt, and D. Powers, “Characterization and evaluation of similarity measures for pairs of clusterings,” *Knowledge and Information Systems*, vol. 19, no. 3, pp. 361–394, 2009.

[115] S. Bickel and T. Scheffer, “Multi-view clustering,” in *ICDM*, vol. 4, pp. 19–26, 2004.

[116] S. Mimaroglu and E. Erdil, “Combining multiple clusterings using similarity graph,” *Pattern Recognition*, vol. 44, no. 3, pp. 694–703, 2011.

[117] N. Ramakrishnan, D. Kumar, B. Mishra, M. Potts, and R. F. Helm, “Turning cartwheels: an alternating algorithm for mining redescriptions,” in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 266–275, ACM, 2004.

[118] S. Monti, P. Tamayo, J. Mesirov, and T. Golub, “Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data,” *Machine learning*, vol. 52, no. 1-2, pp. 91–118, 2003.

[119] V. Filkov and S. Skiena, “Heterogeneous data integration with the consensus clustering formalism,” in *Data Integration in the Life Sciences*, pp. 110–123, Springer, 2004.

[120] C.-W. Chu, J. D. Holliday, and P. Willett, “Combining multiple classifications of chemical structures using consensus clustering,” *Bioorganic & medicinal chemistry*, vol. 20, no. 18, pp. 5366–5371, 2012.

[121] M. Gleicher, “Explainers: Expert explorations with crafted projections,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2042–2051, 2013.

[122] E. T. Brown, J. Liu, C. E. Brodley, and R. Chang, “Dis-function: Learning distance functions interactively,” in *IEEE Conference on Visual Analytics Science and Technology*, pp. 83–92, 2012.

[123] X. Hu, L. Bradel, D. Maiti, L. House, and C. North, “Semantics of directly manipulating spatializations,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2052–2059, 2013.

[124] E. Packer, P. Bak, M. Nikkila, V. Polishchuk, and H. J. Ship, “Visual analytics for spatial clustering: Using a heuristic approach for guided exploration,” *IEEE*

Transactions on Visualization and Computer Graphics, vol. 19, no. 12, pp. 2179–2188, 2013.

[125] N. Andrienko, G. Andrienko, and P. Gatalsky, “Tools for visual comparison of spatial development scenarios,” in *Information Visualization*, pp. 237–244, IEEE, 2003.

[126] K. Matkovic, M. Jelovic, J. Juric, Z. Konyha, and D. Gracanin, *Interactive visual analysis and exploration of injection systems simulations.* 2005.

[127] E. Santos, J. Poco, Y. Wei, S. Liu, B. Cook, D. Williams, and C. Silva, “UV-CDAT: Analyzing climate datasets from a user’s perspective,” *Computing in Science Engineering*, vol. 15, no. 1, pp. 94–103, 2013.

[128] H. Siirtola, “Interaction with the reorderable matrix,” in *Information Visualization, 1999. Proceedings. 1999 IEEE International Conference on*, pp. 272–277, 1999.

[129] C.-H. Chen, H.-G. Hwu, W.-J. Jang, C.-H. Kao, Y.-J. Tien, S. Tzeng, and H.-M. Wu, “Matrix visualization and information mining,” in *Proceedings in Computational Statistics*, pp. 85–100, Springer, 2004.

[130] H.-M. Wu, Y.-J. Tien, and C.-h. Chen, “Gap: A graphical environment for matrix visualization and cluster analysis,” *Computational Statistics & Data Analysis*, vol. 54, no. 3, pp. 767–778, 2010.

[131] I. Liiv, “Seriation and matrix reordering methods: An historical overview,” *Statistical analysis and data mining*, vol. 3, no. 2, pp. 70–91, 2010.

[132] M. Greenacre, “Weighted metric multidimensional scaling,” in *New Developments in Classification and Data Analysis*, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 141–149, Springer, 2005.

[133] A. Tivellato, “JOptimizer.” `http://www.joptimizer.com/`.

[134] J. Chuang, D. Ramage, C. Manning, and J. Heer, “Interpretation and trust: Designing model-driven visualizations for text analysis,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 443–452, ACM, 2012.

[135] J. Elith, J. R. Leathwick, and T. Hastie, “A working guide to boosted regression trees,” *Journal of Animal Ecology*, vol. 77, no. 4, pp. 802–813, 2008.

[136] D. Zurell, J. Elith, and B. Schröder, "Predicting to new environments: tools for visualizing model behaviour and impacts on mapped distributions," *Diversity and Distributions*, vol. 18, no. 6, pp. 628–634, 2012.

[137] H. Edelsbrunner and J. Harer, *Computational Topology: An Introduction.* Amer. Math. Soc., Providence, Rhode Island, 2009.

[138] A. Hatcher, *Algebraic Topology.* New York: Cambridge U. Press, 2002.

[139] J. Milnor, *Morse Theory.* New Jersey: Princeton Univ. Press, 1963.

[140] K. Cole-McLaughlin, H. Edelsbrunner, J. Harer, V. Natarajan, and V. Pascucci, "Loops in Reeb graphs of 2-manifolds," *Disc. Comput. Geom.*, vol. 32, no. 2, pp. 231–244, 2004.

[141] H. Edelsbrunner, *Geometry and Topology for Mesh Generation.* England: Cambridge Univ. Press, 2001.

[142] P. K. Agarwal, H. Edelsbrunner, J. Harer, and Y. Wang, "Extreme Elevation on a 2-manifold," *Disc. Comput. Geom.*, vol. 36, no. 4, pp. 553–572, 2006.

[143] H. Edelsbrunner, D. Letscher, and A. Zomorodian., "Topological Persistence and Simplification," *Disc. Comput. Geom.*, vol. 28, no. 4, pp. 511–533, 2002.

[144] H. Edelsbrunner and J. Harer, "Persistent Homology — A Survey," in *Surveys on Discrete and Computational Geometry. Twenty Years Later* (J. E. Goodman, J. Pach, and R. Pollack, eds.), pp. 257–282, Amer. Math. Soc., Providence, Rhode Island, 2008. Contemporary Mathematics 453.

[145] H. Carr, J. Snoeyink, and U. Axen, "Computing Contour Trees in All Dimensions," *Comput. Geom. Theory Appl.*, vol. 24, no. 2, pp. 75–94, 2003.

[146] V. Pascucci, K. Cole-McLaughlin, and G. Scorzelli, "The TOPORRERY: computation and presentation of multi-resolution topology," in *Mathematical Foundations of Scientific Visualization, Computer Graphics, and Massive Data Exploration*, Mathematics and Visualization, pp. 19–40, Springer, 2009.

[147] K. Matkovic, D. Gracanin, B. Klarin, and H. Hauser, "Interactive visual analysis of complex scientific data as families of data surfaces," *IEEE TVCG*, vol. 15, no. 6, pp. 1351–1358, 2009.

[148] K. Matkovic, D. Gracanin, M. Jelovic, and H. Hauser, "Interactive visual steering - rapid visual prototyping of a common rail injection system," *IEEE TVCG*, vol. 14, no. 6, pp. 1699–1706, 2008.

[149] H. Piringer, W. Berger, and J. Krasser, "Hypermoval: Interactive visual validation of regression models for real-time simulation," in *Proc. EuroVis*, (Aire-la-Ville, Switzerland, Switzerland), pp. 983–992, 2010.

[150] W. Berger, H. Piringer, P. Filzmoser, and E. Gröller, "Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction," in *Proc. EuroVis*, (Aire-la-Ville, Switzerland, Switzerland), pp. 911–920, 2011.

[151] T. Torsney-Weir, A. Saad, T. Moller, H.-C. Hege, B. Weber, and J.-M. Verbavatz, "Tuner: Principled parameter finding for image segmentation algorithms using visual response surface exploration," *IEEE TVCG*, vol. 17, pp. 1892–1901, Dec. 2011.

[152] T. Muhlbacher and H. Piringer, "A partition-based framework for building and validating regression models," *IEEE TVCG*, vol. 19, no. 12, pp. 1962–1971, 2013.

[153] S. Bergner, M. Sedlmair, T. Moller, S. Nabi Abdolyousefi, and A. Saad, "Paraglide: Interactive parameter space partitioning for computer simulations," *IEEE TVCG*, vol. 19, pp. 1499–1512, Sept. 2013.

[154] G. Weber, P.-T. Bremer, and V. Pascucci, "Topological landscapes: A terrain metaphor for scientific data," *IEEE TVCG*, vol. 13, pp. 1416–1423, Nov. 2007.

[155] W. Harvey and Y. Wang, "Topological landscape ensembles for visualization of scalar-valued functions," *Computer Graphics Forum*, vol. 29, pp. 993–1002, 2010.

[156] P. Oesterling, C. Heine, H. Jänicke, G. Scheuermann, and G. Heyer, "Visualization of high dimensional point clouds using their density distribution's topology," *IEEE TVCG*, vol. 99, no. PrePrints, 2011.

[157] S. Gerber, P. Bremer, V. Pascucci, and R. Whitaker, "Visual exploration of high dimensional scalar functions," *IEEE TVCG*, vol. 16, no. 6, pp. 1271–1280, 2010.

[158] N. Young, "Tutorial for the software for assisted habitat modeling (sahm) package in vistrails," tech. rep., US Geological Survey, Fort Collins Science Center, 2012.

[159] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, “Stability of persistence diagrams,” *Disc. Comput. Geom.*, vol. 37, no. 1, pp. 103–120, 2007.

[160] G. Carlsson, A. Zomorodian, A. Collins, and L. Guibas, “Persistence barcodes for shapes,” in *Proc. SGP*, pp. 124–135, 2004.

[161] D. Morozov, K. Beketayev, and G. Weber, “Interleaving distance between merge trees,” in *Workshop on Topological Methods in Data Analysis and Visualization: Theory, Algorithms and Applications*, TopoInVis’13, 2013.

[162] K. Beketayev, D. Yeliussizov, D. Morozov, G. Weber, and B. Hamann, “Measuring the distance between merge trees,” in *Topological Methods in Data Analysis and Visualization III* (P.-T. Bremer, I. Hotz, V. Pascucci, and R. Peikert, eds.), Mathematics and Visualization, pp. 151–165, Springer International Publishing, 2014.

[163] U. Bauer, X. Ge, and Y. Wang, “Measuring distance between reeb graphs,” in *Proc. SOCG*, pp. 464:464–464:473, ACM, 2014.

[164] V. Narayanan, D. M. Thomas, and V. Natarajan, “Distance between extremum graphs,” in *Proc. PacificVis (to appear)*, 2015.

[165] C. Correa, P. Lindstrom, and P.-T. Bremer, “Topological spines: A structure-preserving visual representation of scalar fields,” *IEEE TVCG*, vol. 17, no. 12, pp. 1842–1851, 2011.

[166] M. Hilaga, Y. Shinagawa, T. Kohmura, and T. L. Kunii, “Topology matching for fully automatic similarity estimation of 3d shapes,” in *Proc. SIGGRAPH*, pp. 203–212, 2001.

[167] H. Saikia, H.-P. Seidel, and T. Weinkauf, “Extended branch decomposition graphs: Structural comparison of scalar data,” *CGF (Proc. EuroVis)*, vol. 33, pp. 41–50, June 2014.

[168] D. Thomas and V. Natarajan, “Symmetry in scalar field topology,” *IEEE TVCG*, vol. 17, pp. 2035–2044, Dec 2011.

[169] D. Thomas and V. Natarajan, “Detecting symmetry in scalar fields using augmented extremum graphs,” *IEEE TVCG*, vol. 19, pp. 2663–2672, Dec 2013.

[170] D. Schneider, A. Wiebel, H. Carr, M. Hlawitschka, and G. Scheuermann, "Interactive comparison of scalar fields based on largest contours with applications to flow visualization," *IEEE TVCG*, vol. 14, no. 6, pp. 1475–1482, 2008.

[171] S. Bruckner and T. Möller, "Isosurface similarity maps," in *Proc. EuroVis*, pp. 773–782, 2010.

[172] D. Thomas and V. Natarajan, "Multiscale symmetry detection in scalar fields by clustering contours," *IEEE TVCG*, vol. 20, no. 12, pp. 2427–2436, 2014.

[173] M. L. Fredman and R. E. Tarjan, "Fibonacci heaps and their uses in improved network optimization algorithms," *J. ACM*, vol. 34, no. 3, pp. 596–615, 1987.

[174] B. Dezs, A. Jüttner, and P. Kovács, "Lemon - an open source c++ graph template library," *Electron. Notes Theor. Comput. Sci.*, vol. 264, pp. 23–45, July 2011.

[175] "Tableau." `http://www.tableau.com`. Accessed: 05-27-2015.