

**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

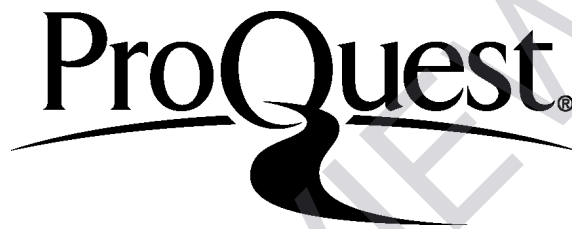
ProQuest Number: 10043944

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10043944

Published by ProQuest LLC (2016). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

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:



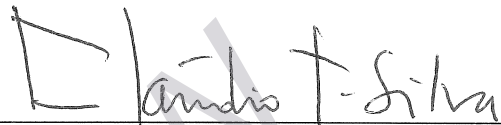
Department Head Signature

7/16/2015

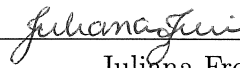
Date

Approved by the Guidance Committee:

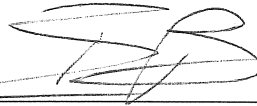
Major: Computer Science



Claudio 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

Microfilm or other copies of this dissertation are obtainable from

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

PREVIEW

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.

PREVIEW

Contents

Vita	iv
Acknowledgements	v
Abstract	vii
List of Figures	xiv
List of Tables	xv
1 Introduction	1
1.1 Motivation	1
1.2 Climate Models	2
1.3 Thesis Statement	3
1.4 Contributions	3
1.5 Outline	5
2 An Exploratory Study of Visualization Use and Design for Climate Model Comparison	6
2.1 Related Work	8
2.2 How Climate Scientists Use Visualization	10
2.3 Methodology	11
2.4 Taxonomy of Design Problems	18
2.5 Matches and mismatches	31
2.6 Solution Redesign	36
2.7 Design Problem Trade-Offs	43
2.8 Guidelines for Avoiding Design Problems	44
2.9 Scope and Impact	47
2.10 Summary	49
3 Perceptual Evaluation of Color Scales for Climate Model Comparison	51
3.1 Related Work	53

3.2	Task Analysis	55
3.3	Choice of Color Scales	57
3.4	Hypotheses Generation	60
3.5	Methodology	61
3.6	Results	67
3.7	Discussion	77
3.8	Summary	79
4	SimilarityExplorer: A Visual Intercomparison Tool for Multifaceted Climate Data	80
4.1	Related Work	81
4.2	Background of Model Intercomparison	83
4.3	Domain Characterization	85
4.4	Visualization Tasks and Design	87
4.5	SimilarityExplorer	90
4.6	Case Studies	95
4.7	Summary	98
5	Visual Reconciliation of Alternative Similarity Spaces in Climate Modeling	99
5.1	Motivation	101
5.2	Related Work	104
5.3	Coordinated Multiple Views	106
5.4	Reconciliation Workflows	109
5.5	Case Studies	117
5.6	Summary	122
6	Using Maximum Topology Matching to Explore Differences In Climate Models	123
6.1	Background	125
6.2	Related Work	127
6.3	Scalar Function Similarity	129
6.4	Implementation	133
6.5	Exploration Framework	136
6.6	Experiments	140
6.7	Case Studies	141
6.8	Discussions	146

6.9 Summary	146
7 Conclusion and Future Work	148
Bibliography	150

PREVIEW

List of Figures

2.1	Illustrating two common visualization use case scenarios and their associated visualization design problems.	7
2.2	One of the few examples of optimal visualization design from our collected sample [24].	11
2.3	Workflow for our qualitative study.	13
2.4	Different levels in the design problem taxonomy.	18
2.5	Design problems found at the encoding stage of the visualization pipeline.	20
2.6	Problems due to clutter: color mixing, visual variable problem: ambiguity, and chart appropriateness: configuration.	22
2.7	Design problems found at the decoding stage of the visualization pipeline.	24
2.8	Problems due to chart appropriateness: mismatch, distortion: scale inconsistency, comparison complexity: superposition overload, communication gap: legend, annotation.	26
2.9	Causes of problems sorted by high percentage of majority disagreement.	33
2.10	Solution redesign for improving the scatter plot shown in Figure 2.1(a).	38
2.11	Solution redesign for the multiple maps in Figure 2.1(b).	41
2.12	Original representation of the spaghetti plot.	42
2.13	Solution redesign for improving the spaghetti plot.	43
3.1	Pairwise comparison of color-coded geographical maps representing climate model outputs encoded with the rainbow color scale.	52
3.2	Three different color scales used in our study and the corresponding luminance plots.	58
3.3	Selection of trials based on data bins.	62
3.4	Examples of map pairs generated based on similar and dissimilar magnitude, and similar and dissimilar spatial distribution.	63
3.5	Data Generation for colormap study.	64
3.6	Overall Relative Error in Task 1.	67

3.7	Effect of Spatial Distribution on Task 1.	68
3.8	Effect of Magnitude on Task 1.	69
3.9	Task 1: Confidence intervals showing estimates of GPP B with respect to the four quadrants.	71
3.10	Confidence vs Error for Task 1.	72
3.11	Task 2 Results.	73
3.12	For Task 3, difference maps showing click spots.	74
3.13	Task 3 Results.	75
3.14	Participant ratings.	76
4.1	Visualizing the complexity of multifaceted climate data.	84
4.2	Similarity computation.	85
4.3	Preserving the mental model and symmetry about spatial and temporal similarity.	89
4.4	SimilarityExplorer is composed of a set of filters (a), similarity views (b, c, d) and data views (e, f).	91
4.5	Data View: Parallel Coordinates.	94
4.6	Comparing multiple output variables.	95
4.7	Comparing model similarity for GPP and analyzing spatiotemporal anomalies for winter and summer (Q1, Q3).	97
5.1	Conceptual model of visual reconciliation.	101
5.2	Matrix view for model structure data.	107
5.3	Iterative visual reconciliation.	110
5.4	Workflow for reconciling model structure with model output.	111
5.5	Synthetic data for validating weighted optimization.	114
5.6	Validation of user feedback based optimization in the MDS plots.	115
5.7	Workflow for reconciling output with structure through feedback.	116
5.8	Reconciling seasonal cycle with model structure similarity.	118
5.9	Iterative exploration of structure-output dependency.	120
6.1	Topology of scalar functions.	124
6.2	f_1 and f_2 are two functions defined on the same domain.	129
6.3	Computing the maximum topology matching.	130
6.4	Computing the topological similarity.	132
6.5	Effect of noise in the neighborhood of a maximum.	135
6.6	We compare three functions f_1 , f_2 , and f_3	137

6.7	The properties view.	137
6.8	The features view.	138
6.9	Parallel coordinates view.	139
6.10	Exploring the features of the MARS model for the Brewers Sparrow data set.	142
6.11	Exploring the GLM model for the Spruce Fir species data.	142
6.12	Comparing MARS with other models for the Brewers Sparrow species.	143
6.13	Locations of a significant minimum-saddle pair in MARS is shown using parallel coordinates.	144
6.14	The response curves corresponding to a minimum-saddle pair.	144
6.15	Comparing MARS and BRT for the Sage Brush species data.	145

PREVIEW

List of Tables

2.1	Connecting design problems to problem consequences sorted by severity	29
3.1	Relative Error Mean and 95% C.I. in Task 1.	69
4.1	Translating tasks into visualization design through a classification scheme.	86
6.1	Time taken to compute the similarity between two models.	140

PREVIEW

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 as 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.

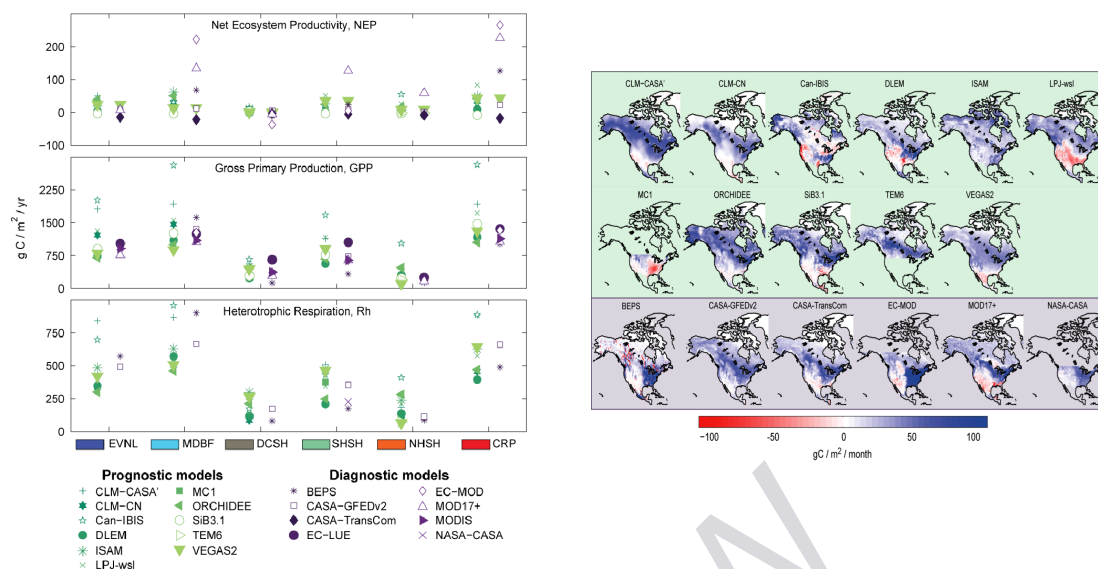
Chapter 2

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.