# Interactive Visual Analysis of Multi-faceted Scientific Data

Johannes Kehrer


Dissertation for the degree of
Philosophiae Doctor (PhD)

Supervised by Helwig Hauser
Co-supervised by M. Eduard Gröller

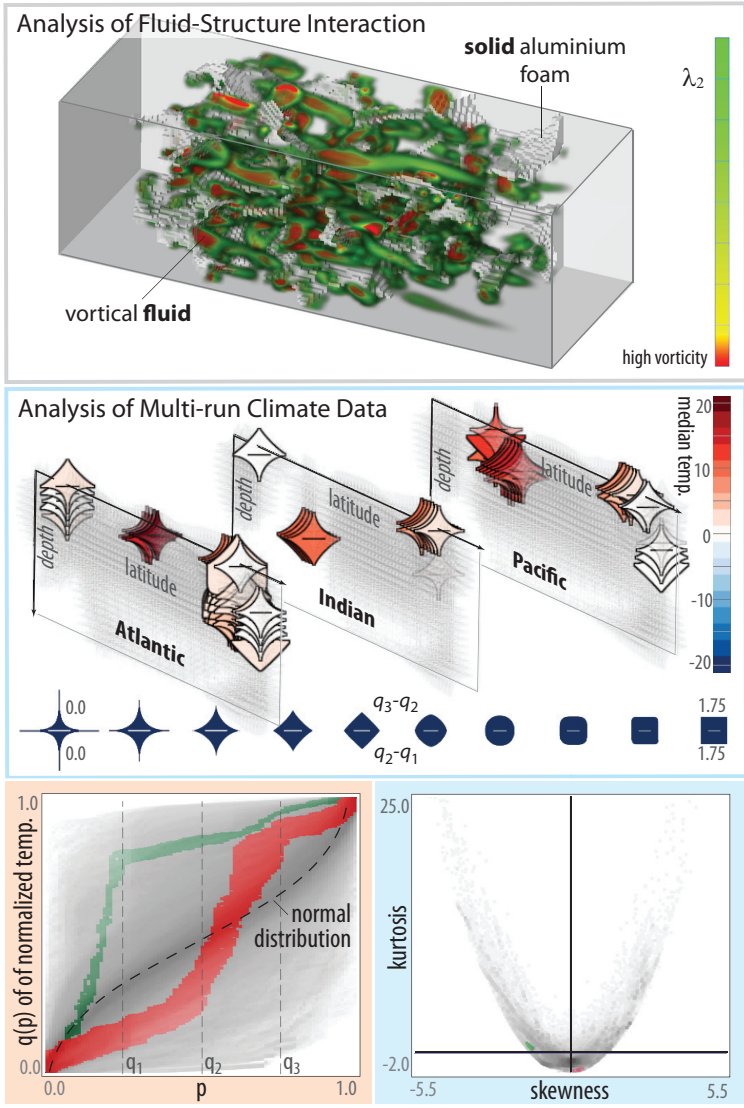
Institute for Informatics
University of Bergen

March 2011

ISBN 978-82-308-1733-9
University of Bergen, Norway
Submitted 2011-03-21 (print version 2011-05-09)

# Interactive Visual Analysis of Multi-faceted Scientific Data

Johannes Kehrer, PhD Thesis




http://www.ii.UiB.no/vis/team/kehrer/thesis/

To my parents,
to my aunt Gretl,
and to Alex and Eva

## Scientific Environment

The research presented in this dissertation has been conducted in the Visualization group at the Department of Informatics, University of Bergen, and while visiting the SimVis GmbH, Vienna, Austria.



# Acknowledgments

First of all, I would like to thank my supervisor Helwig Hauser for his steady support, many fruitful discussions, as well as valuable comments and feedback that contributed to this thesis. He was always available for discussions and taught me how to do research in visualization. My gratitude goes to our cooperation partners at the *SimVis GmbH*, Vienna, Austria, namely Helmut Doleisch, Philipp Muigg, and Wolfgang Freiler. They supported me in many issues related to this work. Helmut and Philipp had already guided my Master thesis and are also coauthors of six joint publications. Every year in December, the SimVis GmbH provided me a great working environment during my research stays in Vienna and gave me shelter from the dark Norwegian winter. I am truly grateful that I could use the SimVis framework as a platform for the research presented here.

I want to thank the present and former members of the *Visualization Group* at the Department of Informatics, University of Bergen, for creating a great and inspiring working environment. Some of you have become friends over the past years, and I am very thankful for that. My special thanks go to Ove Daae Lampe, Daniel Patel, and Ivan Viola for keeping up the good mood, even during stressful periods of paper writing; to Ivan, Veronika Šoltészová, Július Parulek, and Çağatay and Bucu Turkay for nice evenings at Baran Café; to Paolo Angelelli, Åsmund Birkeland, Endre Lidal, Andrea Brambilla, and others for enjoyable lunch breaks; to Jean-Paul Balabanian for introducing me to Smalahove; and to Çağatay and Armin Pobitzer for important help and feedback with respect to data mining and statistics. Further thanks go to our former Master students Stian Eikeland for converting the multi-run data and Andreas Lie for the implementation of the glyph-based renderer that led to a joint publication.

Parts of this work presented here were done in collaboration with the *Wegener Center for Climate and Global Change*, University Graz, Austria. In this context, I want to thank my coauthors Florian Ladstädter, Andrea Steiner, Bettina Lackner, Barbara Pirscher, and Gottfried Kirchengast for many important discussions and valuable input to this work. It was very rewarding to see how positively our technology was integrated in their workflow. Other parts of this work were done in cooperation with the *Potsdam Institute for Climate Impact Research* (PIK), Germany. I want to thank especially Thomas Nocke for fruitful discussions, valuable comments on two of our papers and the related work section, and for supporting a one-week research stay in Potsdam. Further thanks go to Michael Böttinger from the *German Climate Computing Center* (DKRZ) and Laurent Bertino and colleagues from the *Nansen Environmental and Remote Sensing Center* (NERSC), Bergen, Norway. Finally, I liked to thank our coauthor Peter Filzmoser, *Department of Statistics and Probability Theory*, Vienna University of Technology, Austria, for important input and valuable feedback with respect to statistics.

I would like to thank the *Department of Informatics*, University of Bergen, for supporting me and my group during the PhD period, especially Petter Bjørstad, Mark Bezem, Torleiv Kløve, Ida Holen, and Steinar Heldal. My gratitude goes to M. Eduard Gröller and colleagues from the visualization group at the Vienna University of Technology, Austria, as well as my former colleagues from the *VRVis Research Center* in Vienna for many inspiring discussions. I liked to thank Robert Johannessen and Matthew Parker (Univ. of Bergen), Brendan McNulty (Allegro Language Service, Bergen), David Horn, and Andre Alme Rossebø for proofreading parts of this thesis. Also, I am very grateful for the valuable comments of our anonymous reviewers that helped improving our work. Finally, I would like to thank my opponents Min Chen and Heidrun Schumann for the time they spend reviewing this work.

The multi-run climate data are courtesy of the PIK, Germany, and the fluid–structure interaction data are courtesy of Innovative Computational Engineering GmbH (www.ice-sf.at), Leoben, Austria. The Diesel Exhaust System data are courtesy of AVL List GmbH, Graz, Austria, and the ECHAM5 data are courtesy of the Max-Planck-Institute for Meteorology, Hamburg, Germany. Parts of this work were supported by the Austrian Research Funding Agency (FFG) in the scope of the projects "AutARG" (Nr. 819352) and "PolyMulVis" (No. 823855). Other work was done in the scope of the project INDICATE (Nr. P18733-N10) that was funded by the Austrian Science Fund (FWF).

Besides work, I would like to thank my circle of friends for being there for me, especially Alexander Degelsegger and Eva Maria Widmair, who went along with me during the past years. I was always welcome and could stay at their places when I was in Vienna, and we had long Skype calls when I was in Bergen. I also liked to thank Sonia Wu, Hans Gruber, Rebecca and Willi Just, Anton Hahn, Niube Eurídice Márquez, Dave Horn, Daniel Hupfer, Miriam Schneider, Stefan Weninger, Lisa Wawra, Chris and Hannes Felling, Resi Masching, David Palme, Thesi Lackner, Verena Mock, Christoph Neuhauser, and many many more for being great friends over the past years. I feel blessed for knowing all these wonderful people that have enriched my life in so many aspects.

I am grateful for my old group in improvisational theater, *ImproOrange* (or X.Orange as they are called now), which welcomed me whenever I was back in Vienna—also back on stage—, especially Tina Rammlmair, Alexander Fennon, and Therese Garstenauer. Further thanks go to the former and present members of my choir in Bergen, *Studentkoret Blandede Akademikere* (BLAK), especially Kristina Espeseth, Martine Grendahl Sem, Kjersti Juley Nising Sandvold, Raanan Elefant, and Elise Varne. The people in BLAK made the last three years in Bergen special for me, spending countless evenings with singing and celebrating, going on many cottage trips, and having a great time together.

Finally, I want to thank my aunt Gretl Schacherl and my parents Rudolf and Gertrude Kehrer for their love and for everything they have done for me.

# Abstract

Visualization plays an important role in exploring, analyzing and presenting large and heterogeneous scientific data that arise in many disciplines of medicine, research, engineering, and others. We can see that model and data scenarios are becoming increasingly *multi-faceted*: data are often multi-variate and time-dependent, they stem from different data sources (multi-modal data), from multiple simulation runs (multi-run data), or from multi-physics simulations of interacting phenomena that consist of coupled simulation models (multi-model data). The different data characteristics result in special challenges for visualization research and interactive visual analysis. The data are usually large and come on various types of grids with different resolution that need to be fused in the visual analysis.

This thesis deals with different aspects of the interactive visual analysis of multi-faceted scientific data. The main contributions of this thesis are: 1) a number of novel approaches and strategies for the interactive visual analysis of multi-run data; 2) a concept that enables the feature-based visual analysis across an interface between interrelated parts of heterogeneous scientific data (including data from multi-run and multi-physics simulations); 3) a model for visual analysis that is based on the computation of traditional and robust estimates of statistical moments from higher-dimensional multi-run data; 4) procedures for visual exploration of time-dependent climate data that support the rapid generation of promising hypotheses, which are subsequently evaluated with statistics; and 5) structured design guidelines for glyph-based 3D visualization of multi-variate data together with a novel glyph. All these approaches are incorporated in a single framework for interactive visual analysis that uses powerful concepts such as coordinated multiple views, feature specification via brushing, and focus+context visualization. Especially the data derivation mechanism of the framework has proven to be very useful for analyzing different aspects of the data at different stages of the visual analysis. The proposed concepts and methods are demonstrated in a number of case studies that are based on multi-run climate data and data from a multi-physics simulation.

# Related Publications

This thesis is based on the following publications (see part II of the thesis):

**Paper A:** J. Kehrer, F. Ladstädter, P. Muigg, H. Doleisch, A. Steiner, and H. Hauser. **Hypothesis generation in climate research with interactive visual data exploration**. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1579–1586, 2008.

**Paper B:** A. Lie, J. Kehrer, and H. Hauser. **Critical design and realization aspects of glyph-based 3D data visualization**. In *Proc. Spring Conference on Computer Graphics (SCCG 2009)*, pages 27–34, 2009.

**Paper C:** J. Kehrer, P. Muigg, H. Doleisch, and H. Hauser. **Interactive visual analysis of heterogeneous scientific data across an interface**. *IEEE Transactions on Visualization and Computer Graphics*, 17(7):934–946, 2011.

**Paper D:** J. Kehrer, P. Filzmoser, and H. Hauser. **Brushing moments in interactive visual analysis**. *Computer Graphics Forum*, 29(3):813–822, 2010.

The following publications are also related to the thesis:

**Paper 1:** P. Muigg, J. Kehrer, S. Oeltze, H. Piringer, H. Doleisch, B. Preim, and H. Hauser. **A four-level focus+context approach to interactive visual analysis of temporal features in large scientific data**. *Computer Graphics Forum*, 27(3):775–782, 2008.

**Paper 2:** F. Ladstädter, A. Steiner, B. Lackner, G. Kirchengast, P. Muigg, J. Kehrer, and H. Doleisch. **SimVis: an interactive visual field exploration tool applied to climate research**. In A. Steiner, B. Pirscher, U. Foelsche, and G. Kirchengast, editors, *New Horizons in Occultation Research*, pages 235–245. Springer, 2009.

**Paper 3:** F. Ladstädter, A. Steiner, B. 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*, 27(4):667–679, 2010.

**Paper 4:** O. Daae Lampe, J. Kehrer, and H. Hauser. **Visual analysis of multivariate movement data using interactive difference views**. In *Proc. Vision, Modeling, and Visualization (VMV 2010)*, pages 315–322, 2010.

The papers A, 2 and 3 have been done in cooperation with domain scientists from the Wegener Center for Climate and Global Change (WegCenter) and from the Institute for Geophysics, Astrophysics and Meteorology (IGAM), University of Graz, Austria, as well as the with the SimVis GmbH, Vienna, Austria. The latter provided the visual analysis framework upon which this research builds. With respect to **paper A**, I was the principal researcher on the visualization side and F. Ladstädter was the principal researchers on the side of climate research (he is also first author of papers 2 and 3). The work builds also upon a recent extension of the SimVis framework (paper 1), which I did as a Master thesis project [110]. The latter was supervised by P. Muigg and H. Doleisch (at that time at the VRVis Research Center, Vienna) as well as by H. Hauser. Paper 1 also contains a case study that was done in cooperation with S. Oeltze and B. Preim from the Department of Simulation and Graphics, University of Magdeburg, Germany.

**Paper B** is the outcome of a project done together with a former Master student, Andreas Lie, and my supervisor H. Hauser. I was the principal researcher concerning the different aspects of glyph design, while Andreas did mainly the implementation of the proposed glyph renderer. It was a great opportunity for Master student A. Lie to present this paper at the Spring Conference on Computer Graphics (SCCG 2009) in Budmerice, Slovakia.

I was also the principal researcher with respect to the work described in **papers C** and **D**. The former was done in collaboration with my supervisor, H. Hauser, and the SimVis GmbH that provided the platform for this research (same as above for paper A). In this context, the investigation of the related fluid–structure interaction scenario was mainly performed by P. Muigg. The research described in paper D was done in cooperation with my supervisor, H. Hauser, and P. Filzmoser from the Department of Statistics and Probability Theory, Vienna University of Technology, Austria, who helped to secure the soundness of this paper with respect to statistics. The work builds upon the interface concept described in paper C.

Finally, paper 4 represents the research of my colleague, O. Daae Lampe, on the visual analysis of vessel movement data. I contributed to different aspects of the proposed difference views and helped with the write-up of this research.

# Contents

**Scientific Environment** iv

**Acknowledgments** v

**Abstract** vii

**Related Publications** ix

## I Overview

**1 Introduction** 1
- 1.1 Multi-faceted Scientific Data: Characteristics and Challenges . . . 2
- 1.2 Contributions and Thesis Structure . . . 4

**2 State of the Art: Interactive Visual Analysis and Visualization of Multi-faceted Scientific Data** 7
- 2.1 Interactive Visual Analysis . . . 8
  - 2.1.1 Coordinated Multiple Views . . . 9
  - 2.1.2 The Science of Visual Analytics . . . 9
- 2.2 Visual Analysis and Visualization of Time-varying Data . . . 12
  - 2.2.1 Time-dependent Data Visualization . . . 12
  - 2.2.2 Interactive Visual Analysis of Time-dependent Data . . . 14
- 2.3 Multi-variate Data Visualization and Analysis . . . 16
  - 2.3.1 Visualization of Multi-variate Data . . . 17
  - 2.3.2 Visual Analysis of Multi-variate Data . . . 18
- 2.4 Visualization and Visual Analysis of Multi-Modal Data . . . 19
  - 2.4.1 Visual Data Fusion . . . 20
  - 2.4.2 Visual Analysis for Comparison . . . 21
- 2.5 Multi-run Data Visualization and Analysis . . . 22
  - 2.5.1 Visualization of Multi-run Distributions and Derived Data . 23
  - 2.5.2 Interactive Visual Analysis of Multi-run Data . . . 24
- 2.6 Multi-Model Data Visualization and Analysis . . . 26
- 2.7 Chapter Summary and Conclusions . . . 26

**3 Interactive Visual Analysis of Multi-faceted Scientific Data** **29**
3.1 Hypothesis Generation with Interactive Visual Exploration . . . . 29
3.2 Critical Aspects of Glyph-based 3D Visualization . . . . . 32
3.3 Interactive Visual Analysis across Two Parts of Scientific Data . . 34
3.4 A Moment-based Scheme for Interactive Visual Analysis . . . . . 37

**4 Demonstration Cases** **41**
4.1 Exploring Climate Data for Hypotheses Generation . . . . . 41
4.2 Glyph-based Analysis of a Diesel Exhaust System . . . . . 43
4.3 Visual Analysis of a Fluid–Structure Interaction . . . . . 45
4.4 Visual Analysis of Multi-run Climate Data . . . . . 46
4.4.1 Visual Sensitivity Analysis across an Interface . . . . . 46
4.4.2 Moment-based Visual Analysis of Multi-run Climate Data . 49

**5 Conclusions and Future Work** **51**

## II Scientific Results

**A Hypothesis Generation in Climate Research with Interactive Visual Data Exploration** **55**
1 Introduction . . . . . 56
2 Climatological Background . . . . . 57
3 Interactive Visual Data Exploration . . . . . 60
4 Exploring The Two Climate Datasets . . . . . 62
4.1 Hypothesis Generation . . . . . 63
4.2 Parameter Optimization . . . . . 70
4.3 Analyzing Relations Between Selections . . . . . 72
4.4 Further Results . . . . . 73
4.5 Performance Issues . . . . . 74
5 Conclusion and Future Work . . . . . 74

**B Critical Design and Realization Aspects of Glyph-based 3D Data Visualization** **77**
1 Introduction . . . . . 78
2 Related Work . . . . . 79
3 Overview . . . . . 79
4 Selected generic Considerations with respect to Glyph Representation . . . . . 80
4.1 Data Mapping . . . . . 80
4.2 Glyph Instantiation . . . . . 83
4.3 Rendering . . . . . 85

5 Demonstration . . . 86
5.1 Diesel Exhaust System . . . 86
5.2 Hurricane Isabel . . . 89
6 Technical Details . . . 89
7 Summary and Conclusions . . . 92
8 Future Work . . . 92

**C Interactive Visual Analysis of Heterogeneous Scientific Data across an Interface 93**
1 Introduction . . . 94
2 Related Work . . . 96
3 Sample Analysis of a Fluid–Structure Interaction Scenario . . . 98
4 Interactive Visual Analysis across an Interface . . . 101
4.1 The Interface (Structural Relation) . . . 102
4.2 Transfer of Degree-of-Interest Information . . . 105
4.3 Automatic Update of Feature Specification . . . 105
4.4 Strategies for Visual Analysis . . . 107
5 Analysis of Multi-run Climate Data . . . 108
5.1 Basic Setup for the Visual Analysis . . . 109
5.2 Outlier analysis in the aggregated data part . . . 110
5.3 Outlier analysis in the multi-run data part . . . 114
6 Conclusion and Future Work . . . 117

**D Brushing Moments in Interactive Visual Analysis 119**
1 Introduction . . . 120
2 Related Work . . . 121
3 Statistical Background . . . 122
4 A Moment-based Scheme for Visual Analysis . . . 124
4.1 Illustrative Example of Multi-run Climate Data . . . 124
4.2 Generic View Transformations . . . 125
4.3 A Classification Scheme for Moment-based Views . . . 130
5 Demonstration Case . . . 134
6 Conclusions and Future Work . . . 137

**Bibliography 139**

**Errata 157**

# Part I

# Overview

# Chapter 1

# Introduction

Our society is confronted with rapidly growing amounts of scientific data that arise in various areas of medicine, science, engineering, and others. Large-scale measurements of dynamic processes as well as numerical modeling and computational simulation result in multi-variate and time-dependent data that are difficult to analyze. Examples are simulation data from global climate models (GCMs) or computational fluid dynamics (CFD), sensor logs, and medical scans such as magnetic resonance imaging (MRI) or computer tomography (CT). Visualization has proved to be very helpful to explore, gain insight, and explain such data [59, 240]. One takes advantage of the phenomenal capability of the human to process visual information and detect interesting structures and relationships in the data such as patterns, trends and anomalies. However, due to the increasing complexity and heterogeneity of scientific data, an increasing need for sophisticated visualization technology arises [85, 104, 116, 164, 223].

There are three major use cases or application goals for visualization [116, 205]: 1) *visual exploration* is used to investigate unknown data characteristics, to "discover the unexpected" [223], and to come up with promising hypotheses (compare also to Tukey [233]). Starting from concrete hypotheses, 2) *visual analysis* or *confirmative visualization* enables the analyst to examine expected data aspects and to confirm or reject existing hypotheses in a goal-oriented analysis process [116]. Eventually, the 3) *presentation* or dissemination of the findings to different target audiences such as domain experts, decision makers, or the general public (e.g., via web-based services, newspapers or television) is highly important [223, 259]. A lesson that has been learned is that visualization must be tailored with respect to different user goals and tasks [116, 229]. During the visual analysis and exploration process, for example, interaction and flexibility of the application are crucial, using concepts such as multiple linked views and brushing for iterative feature specification [52, 73, 192]. Such a guided human–computer dialog supports a powerful drill-down into different aspects of the data [209]. Hypotheses can be generated and analyzed rapidly, unknown and unexpected features can be discovered, and data trends as well as outliers can be explored interactively.

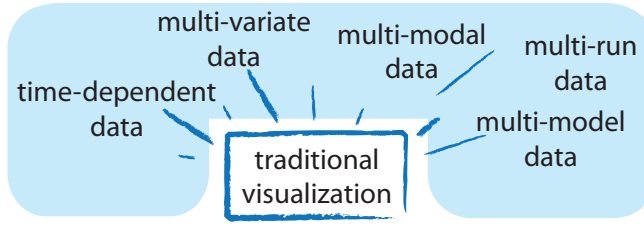



**Figure 1.1:** Multiple, general visualization challenges for multi-faceted scientific data.

## 1.1 Multi-faceted Scientific Data: Characteristics and Challenges

The work in this thesis is motivated by a number of visualization challenges that arise from the heterogeneous nature of *scientific data*. Such data are usually given with a strong inherent reference to space and time and results from a scientific data acquisition method. When talking about *multi-faceted* scientific data, we consider the following (see also Fig. 1.1): 1) *time-dependent* data that represents dynamically changing phenomena; 2) *multi-variate* data consisting of different attributes (data variates) such as temperature or pressure; 3) *multi-modal* data stemming from different acquisition modalities (data sources) that measure or simulate the same phenomenon; 4) *multi-run* data stemming from multiple simulation runs that are computed with varied parameter settings for the simulation model; and 5) *multi-model* data resulting from interrelated simulation models that represent physically interacting phenomena or climate compartments such as ocean and atmosphere. In the following, the characteristics of different kinds of multi-faceted scientific data are discussed together with the related research challenges for visualization and visual analysis.

Advanced computer power allows the simulation of complex dynamic phenomena on high-resolution grids over large timescales (e.g., global climate models [193] or engine simulations [54]). The resulting data often contain multiple data variates per space-time location. The interactive visualization of such data is generally challenging [63, 104, 164]. The important information is commonly hard to identify due to the huge amount of data and their multi-variate characteristics. Important features can often only be extracted when considering multiple data variates and their relations at the same time. Additionally, one has to cope with visual issues such as cluttering and occlusion when representing multiple data variates in the same image. The data is often analyzed using coordinated multiple views that support interactive feature specification via brushing.

The visualization and analysis of time-varying data is challenging too (compare to Aigner et al. [1, 2] and Müller and Schumann [163]). Analysts want to investigate how their data change over time. They want to uncover spatial and temporal patterns (e.g., cyclic behavior or special events), understand major data

trends, and detect anomalies such as outliers. One common goal is to integrate data from multiple time steps in a single image, for instance, by using one spatial axis in the visualization to represent time. Automated analysis methods are often applied in order to abstract time-related data characteristics, for instance, by computing statistical aggregates such as temporal mean values or standard deviations [5]. When designing an analysis framework for time-varying data one also has to consider different data characteristics [1, 2]. Time can, for example, show cyclic behavior such as seasonal trends or changes between day and night, events can happen after each other or at the same time, etc.

Scientific data can often stem from different acquisition modalities that investigate the same physical object/phenomenon. Examples include numerical simulations and measurements such as different types of medical scans (e.g., CT, MRI, or ultrasound data). An analysis task can be to compare data from a climate simulation with observational measurements such as remote sensing in order to find errors and to reduce uncertainties. Here, a challenge is to fuse conflicting multi-modal data in the visualization. The data are possibly given on different data grids (e.g., 2D/3D, unstructured or hybrid) with different temporal or spacial resolutions.

In engineering [152] and climate research [86, 166], so-called multi-run simulations are increasingly often performed to study the variability of a simulation model and to understand the model sensitivity to certain control parameters. According to Hamby [70], the goals of such a *sensitivity analysis* include the identification of model parameters that require additional research, which also reduces the output uncertainty; identifying control parameters that are correlated with the simulation output; or finding insignificant parameters that can be eliminated from the model [70]. The simulation is repeated multiple times with varied settings of the control parameters. In the resulting data, therefore, a collection of values co-exists for the same data attribute at each space/time location [141] (one value for every run). In the analysis, the data is often transformed into an aggregated form, for example, by computing statistical properties with respect to all runs [166]. However, it is a challenge to simultaneously visualize and analyze such large amounts of concurrent data volumes, to extract interesting patterns and trends that occur in different runs, to investigate how many of the runs exhibit a certain pattern, or to study correlations between input and output parameters (compare also to Wilson and Potter [258]).

While dynamic flow is traditionally simulated with respect to a rigid boundary, fluid and solid parts interact during modern *multi-physics* simulations [22]. The solid part can, for instance, be deformed by the surrounding flow. The different data parts are commonly modeled individually on spatially adjoining grids that are connected by a so-called interface. During the simulation, the parts can interact with each other similar to an airplane wing or turbine blades that are deformed by the surrounding flow. In the climate system as well, as another example, components such as atmosphere, ocean, ice, and land interact with each

other. Atmosphere and ocean, for instance, exchange through thermal absorption, precipitation and evaporation. To understand such dynamics, models for the different climate components are coupled in the simulation, commonly with additional coupler modules. Creating a coherent visualization from these multi-model scenarios, which include two or more data parts (e.g., fluid and structure or atmosphere and ocean), is a challenge for visualization research. How can, for instance, feedback and relations between the data parts be investigated?

We consider both multi-run and multi-model data visualization as highly rewarding challenges for visualization research. A large part of this work, therefore, focuses on the interactive visual analysis of these kinds of data. It is important to note that the data can be of different dimensionality such as 2D/3D data, time-dependent data, or higher dimensional data resulting from multi-run simulations with additional independent dimensions for the varied simulation parameters. The data can be given on various kinds of data grids as well, for instance, unstructured or hybrid grids that possibly do not overlap spatially. Designing solutions for interactive visual analysis that address these and other issues of multi-faceted scientific data is a challenging task [104, 116].

As application areas, the main focus of this work is on climate research,[1] and meteorology.[2] These areas are especially interesting—in addition to other application areas such as engineering [128]—since most of the visualization challenges discussed above need to be addressed. Especially climate research has recently gained a lot of public attention concerning the long-term changes in the Earth's climate [214]. A larger part of the work presented in this thesis was done in collaboration with domain researchers. An overview on visualization approaches used in climate research is given by Nocke [165] and Nocke et al. [170]. Lipşa et al. [138], moreover, discusses related work in the context of visualization for the physical sciences.

## 1.2 Contributions and Thesis Structure

The main contributions of this thesis are as follows:

1. Several new approaches are proposed that enable the advanced visual analysis of multi-run data (this is considered an important overall contribution of this thesis as not much research work has been done yet on this topic of increasing relevance).
2. In many scenarios, scientific data consist of multiple interrelated data parts such as the atmosphere and the ocean part of a coupled climate model. A

[1] Climate research is mainly concerned with the analysis and prediction of the overall climate system as well as its variability and long-term behavior [246].

[2] Meteorology is the interdisciplinary science that focuses on the analysis and forecasting of short-term weather phenomena. Especially the investigation and early prediction of extreme weather phenomena such as hurricanes or severe rainfalls are very important.

systematic approach for the visual analysis of such data is proposed that enables the joint investigation of features across related data parts. A so-called *interface* is constructed, which relates individual grid cells between different data parts. During the visual analysis, fractional *degree-of-interest* (DOI) information, resulting from smooth brushing [53], can then be exchanged between the parts. Additional strategies for visual analysis are proposed, where features are iteratively refined, and the analyst works with different data parts simultaneously. The approach is demonstrated on data from a multi-physics simulation as well as multi-run climate simulations.

3. A model for the interactive visual analysis of higher-dimensional data, based on the four statistical moments (mean, variance, skewness, and kurtosis), is proposed. The statistics are computed with respect to selected independent dimensions of the data. Traditional and robust estimates of moments as well as measures of *outlyingness* are integrated in the visual analysis. We propose a set of *view transformations* that support the analyst in navigating the large space of possible views that are based on these statistics. The transformations lead to a classification scheme for informative moment-based views. For depicting the multi-run data distributions, *quantile plots* that are common in statistics are adapted to enable a focus+context style. The proposed model for a moment-based visual analysis is exemplified in different scenarios with multi-run climate data.
4. In the context of climate research, we demonstrate how interactive visual exploration supports the steered generation of promising hypotheses that are subsequently evaluated using classical statistics. In the concrete case, we were looking for atmospheric regions in space and time that represent sensitive and robust *indicators* for climate change. Time-dependent data characteristics such as linear trends and corresponding signal-to-noise ratios are computed using the integrated data derivation mechanism of our visual analysis framework. Strategies are presented, where the derived attributes are interactively explored in order to generate promising hypotheses. Besides identifying such atmospheric regions that react sensitively to climate change, the parameters and boundary conditions for the subsequent computational analysis can be restricted as well. Also, areas with data deficiencies can be identified. The approach is demonstrated in a number of case studies that were done in collaboration with domain experts.
5. Finally, a new glyph is proposed together with structured design guidelines for glyph-based 3D visualization. The task of glyph design is divided into three consecutive steps of data mapping, glyph instantiation, and glyph rendering (compare also to the visualization pipeline [77]). A number of design and realization aspects are discussed according to these steps. For the data mapping stage, we propose strategies to enhance the data such as windowing and/or exponentiation. Important aspects for glyph instan-

tiation are, for example, whether 2D or 3D glyphs should be used and the orthogonality of different graphical properties of the glyph. The latter is related to the possible interference of glyph properties, which hinder the interpretation of the depicted data variates. For the glyph rendering, we suggest to use strategies such as halos in order to address visual cluttering and chroma depth that facilitates depth perception. The proposed design considerations are illustrated on a new glyph that is based on super ellipses. The glyph is designed such that it can be placed in a 3D context and can depict up to six data variates (using color, upper and lower glyph shape, size, rotation, and aspect ratio). The glyph is exemplified in a number of application scenarios including automotive engineering and the visual analysis of multi-run data.

The remainder of this thesis is structured as follows: In chapter 2, we discuss the related state of the art in visualization and the visual analysis of multi-faceted scientific data. The different contributions of this work are then described in more detail in chapter 3. The approaches are exemplified in different application cases in chapter 4. Chapter 5 concludes the first part of this thesis.

Finally, four papers that resulted from this thesis work are given in the second part of this thesis. It should be noted that the papers are not a one-to-one relation to the contributions as listed above—contribution 1 is mainly related to papers C and D, but also uses the glyphs proposed in paper B; contribution 2 is reflected in papers C and D; contribution 3 is detailed in paper D; contribution 4 is related to paper A; and contribution 5 is detailed in paper B, but the glyphs are also used in the context of multi-run data (paper C).

# Chapter 2

# State of the Art: Interactive Visual Analysis and Visualization of Multi-faceted Scientific Data

Multi-faceted scientific data emerge in many areas such as medicine, climate research, physics, or automotive engineering. Visualization and interactive visual analysis have proven to be useful when analyzing such data. Before discussing related approaches from the literature in this chapter, some basic notations should be clarified first.

### Terminology and Structure of the Chapter

In many cases, multi-dimensional *scientific data* can be denoted as $f_d(\mathbf{p})$ where the data variates $f_d$ (e.g., temperature or pressure values) are measured or simulated with respect to points in an $m$-dimensional data domain $\mathbf{p}$ (compare to van Wijk and van Liere [242], for instance). The domain (i.e., the independent data dimensions) can be 2D or 3D space, time, but also independent input parameters to a simulation model. *Multi-run* data, for example, stems from a simulation which is repeated multiple times with varied control parameters, leading to a larger number of concurrent data volumes given for the same space/time [86, 152]. With such data, the word *multi-dimensional* refers to the dimensionality of the independent variables, while *multi-variate* refers to the dependent variables of the data (compare to Wong and Bergeron [261]).

*Multi-modal* data stems from different acquisition modalities such as computer tomography (CT) or magnetic resonance imaging (MRI). While multi-variate data usually results from one modality and describes different physical properties given for the same space/time domain, multi-modal data commonly results from different sources that measure or simulate the same physical phenomenon (e.g., different models of the atmosphere). Accordingly, multi-modal data can be given on different grids and time steps and need to be fused or correlated in the visualization (compare to Fuchs and Hauser [63]). Another important task is the comparison between data stemming from different modalities.

Finally, we refer to *multi-model* data when data stems from different models that simulate related phenomena such as an ocean model and an atmosphere model. During the simulation, these models are often coupled together and can interact and exchange properties (compare also to multi-physics simulations [22]).

The visualization of multi-variate and time-dependent data have been broadly investigated for several years, and a lot of good work has been done. Although these areas belongs to the topics discussed here, we only touch them briefly and refer to other existing state of the art reports. The actual focus of this chapter is on multi-modal, multi-run and multi-model data. Especially multi-run and multi-model data scenarios are relatively new to the visualization community, although these types of data are getting more popular in other domains [22, 86]. For each kind of multi-faceted data, we aim at distinguishing between approaches for visualization and interactive visual analysis, where different tasks are addressed.

The remainder of this chapter is organized as follows: Section 2.1 discusses important concepts in interactive visual analysis such as coordinated multiple views and the combination of computational analysis methods and interactive visualization. Section 2.2 addresses the visualization and visual analysis of time-dependent data, and section 2.3 does this for multi-variate data. The representation, fusion and comparison of multi-modal data are described in section 2.4. Section 2.5 discusses the visual analysis of multi-run data, and section 2.6 addresses challenges for multi-model data. At the start of each section, we attempt to define the related challenges for visualization and visual analysis. At the end of the chapter, an outlook to promising future research and open challenges is given (Sec. 2.7). Note that the list of related works is not meant to be complete, and not all important work could be included in the discussion.

## 2.1 Interactive Visual Analysis

Interactive visualization as well as automated analysis based on statistics or data mining facilitate the understanding of important characteristics in complex data [59, 71, 240]. These areas were developing rather independently from each other for a long time. However, there have also been certain trends on combining automated analysis methods and interactive visualization [118, 210, 233]. While statistical tools commonly utilize static visualization for presentation purpose (confirmatory analysis), Tukey suggests in his seminal work on *exploratory data analysis* [233] to also support direct interaction with the data. Additionally, some of the early works in information visualization were inspired by considerations from statistics [29, 30, 38, 39, 255]. Even certain systems for visual data analysis and exploration can be traced back to these roots [219, 222, 248]. In this context, interaction concepts such as *coordinated multiple views* with linking and brushing are highly relevant and enable a powerful information drill-down process [209]. The history of relations between automated data analysis and interactive visualization eventually led to the recently established initiative on *visual analytics* [116, 223], which is closely related to interactive visual analysis and is discussed in section 2.1.2.

### 2.1.1 Coordinated Multiple Views

The concept of coordinated multiple views has been steadily developing over the last two decades (see Roberts [192] for an overview). Different data variates are simultaneously shown, explored, and analyzed in multiple linked views that are utilized side-by-side. The views can include 2D scatterplots, scatterplot matrices [11], parallel coordinates [91, 102, 172], function graph views [127, 161], or histograms. Interesting data subsets are *brushed* [11] (selected) in the visual display, the related data items are instantly highlighted in the *linked views* (compare to the XmdvTool [248], Polaris/Tableau [217], or ComVis [127, 152], for example). Logical combinations of brushes across multiple views support the specification of complex features, for instance, in a hierarchical feature definition language [52] or in conjunctive visual forms [251]. In cross-filtered views [252], as another example, brushing filters between pairs of views can be enabled/disabled and the data are filtered accordingly. Relationships between multiple variates can thus be explored, also across multiple datasets. Visual analysis frameworks often support the derivation of new data variates from existing ones using computational methods, which facilitates the specification of features [54, 73, 84, 192, 217, 252].

Examples for visual analysis systems for scientific data include WEAVE [66] and SimVis [52]. Such frameworks combine and link *attribute views* such as scatterplots or parallel coordinates [16, 75] with *3D views* of volumetric data [160] (usually given on grids over time). This enables the analyst to investigate multivariate relations of brushed features also in the spatial context (compare to feature localization and local investigation in Oeltze et al. [173], for instance). Instead of a binary selection information, some systems integrate a fractional *degree-of-interest* data attribution $DOI_j \in [0, 1]$ for every data item $j$ (compare to the DOI information in generalized fisheye views [65]). Such an attribution represent the first interpretation level, ranging from data to knowledge [33]. A *smooth brushing* operation [53] results, for example, in a trapezoidal DOI function around the main region of interest in an attribute view. The DOI information is then used in all linked views to visually discriminate interesting features (focus) from the rest of the data (context), leading to a *focus+context visualization* [72, 161]. The focus is thereby visually enhanced, while the rest of the data are depicted in a less prominent style for orientation purpose.

### 2.1.2 The Science of Visual Analytics

Visual analytics is the interdisciplinary science of analytical reasoning facilitated by interactive, visual and analytical methods [114, 115, 223]. Since automated analysis methods only work reliably for well-specified problems, the idea is to combine such approaches with interactive visualization. Visualization can then, for example, support the specification of parameters at different steps of a data

mining algorithm. By interactively and visually exploring the original data as well as derived properties, analysts should be enabled to [223]:

- detect the expected and discover the unexpected;
- find interesting patterns and multi-variate relationships within the data;
- draw conclusions and generate hypotheses based on the visual information;
- reject or verify hypotheses; and
- communicate and present the results of the analytical reasoning process.

Visual analytics aims to combine sophisticated methods from disciplines such as information visualization, data mining,[1] statistics, machine learning, pattern extraction, cognitive and perceptual science, decision science, and human–computer interaction. Such a combination supports the user to effectively and efficiently extract important information from heterogeneous data sources [114]. Shneiderman [210] compares the different philosophies behind exploratory data analysis [233] (used for hypothesis generation) and statistical hypothesis testing. The author suggests, amongst others, to support the user in specifying his/her interest and to keep track of such user decisions using a history mechanism. Shneiderman [210] also recommends that analysis and discovery tools should be easy to use and understand and should support the user in understanding the different steps and outcomes in the statistical analysis process.

Excellent overviews of information visualization and *visual data mining* approaches are given by Keim [113], Keim et al. [118], and de Oliveira and Levkowitz [49]. While visual data mining mainly focuses on the integration of data mining techniques into the visualization, visual analytics aims at integrating other methods of analytical reasoning too (e.g., decision science). Chen [32] discusses visual analytics from the perspective of information theory. Bertini and Lalanne [13, 14] recently survey the integration of visualization and automated analysis in the knowledge discovery. Based on the degree to which such methods are combined, solutions are categorized into computationally enhanced visualizations, visually enhanced mining, and integrated visualization and mining (compare to Keim et al. [118]). The interested reader is also referred to a recent book [115] by the European visual analytics community[2] that discusses aspects such as data management, the analysis of space and time, considerations from cognition and perception as well as evaluation.

Common (semi)automated analysis approaches that are combined with visualization include [13, 14, 118]: *data reduction* via sampling or feature extraction [184]; *clustering* [168, 243] where data items are grouped by similarity; and *dimensionality reduction* that aims to reduce the data dimensionality while

[1] *Data mining* denotes the algorithmic extraction of valuable patterns and models from data. According to Fayyad et al. [57], it is part of a more general process of *knowledge discovery in databases* (KDD), which also includes steps such as data preparation, selection and cleaning.

[2] http://www.vismaster.eu/

maintaining the higher-dimensional data characteristics. Dimensionality reduction approaches include, for instance, principal component analysis [162, 173], which transforms multi-variate data into an orthogonal coordinate system that is aligned with the greatest variance in the data; multi-dimensional scaling [21, 159] (MDS), where higher-dimensional data items are mapped into a lower-dimensional space while preserving the dissimilarities between the items;[3] and self-organizing maps [126, 130] (SOM) which represents an unsupervised learning method that reduces the data dimensionality and also provides a classification of the data. An issue with dimensionality reduction approaches is, however, that it can be hard to mentally relate the derived attributes to the original data. One solution can be to analyze both side-by-side in a coordinated multiple views framework with linking and brushing (see Oeltze et al. [173], for instance).

One should note that an integration of automated and visual techniques is not always desirable. Certain well-defined problems can often be resolved best by automated approaches (compare to Keim et al. [117]). Bertini and Lalanne [13] thus suggest to further study which types of tasks and problems can be addressed best by data mining or by visualization approaches that involve human interaction. Ma [144], moreover, suggests to go a step beyond visual data mining by integrating machine learning methods into the analysis process. Such methods could learn from previous analysis sessions and input data, and abstract away many details of the utilized algorithms, for instance, using case-based reasoning (compare to an infrastructure supporting knowledge-assisted visualization [33]). Only high-level decisions are then left to the user by providing him/her with an "intelligent interface" to the visual analysis [144].

### The Visual Analytics Process

As mentioned earlier, interactive visual analysis enables the user to explore and analyze data in a guided human–computer dialog. The usually employed process follows Shneiderman's *information seeking mantra* [209]: "overview first, zoom and filter, then details-on-demand." If the raw data, however, is too large and complex to be represented in a direct manner, it is necessary to apply automated data abstraction techniques prior to the visualization. Accordingly, Keim proposes an extension to Shneiderman's mantra for visual analytics [116]:

> "Analyze First – Show the Important – Zoom, Filter and Analyze Further – Details on Demand"

Initially, the data are preprocessed in an automatic analysis step, resulting in a condensed representation containing the important aspects of the data. The user gets an overview where he/she can interactively zoom and browse through the data, select data subsets of special interest, or filter uninteresting data. This

[3] Since MDS also maintains the higher-dimensional structure of the data, it is well suitable for subsequent clustering.

helps the analyst to gain knowledge about the data, especially in the case of very large and complex data. This knowledge often leads to new questions and/or hypotheses, which can be explored and analyzed in more detail in an iterative process. One may also want to perform further analysis steps, for instance, by deriving new data attributes from existing ones. The resulting information is again visually analyzed, and so on. Interaction and flexibility of the application are crucial for the analysis process. Yi et al. [266] recently propose a categorization of interaction techniques that are based on the user's intent. Liu and Stasko [139], moreover, investigate how internal representations (mental models) and external visualizations are related to each other. The authors state that such mental models are used during visual reasoning to "simulate" the behavior of the corresponding visualization system [139]. In the visual analysis, the user should be able to query data in many different ways and quickly change what data are represented and how they are visualized [217]. During the analysis process, one takes advantage of human factors such as intuition, creativity, expert knowledge, and the ability to deal with unexpected situations [223].

## 2.2 Visualization and Analysis of Time-varying Data

Time-varying measurements and simulations are ubiquitous in many disciplines such as medicine, climate research, meteorology, or engineering. Being able to understand time-related developments allows one to "learn from the past to predict, plan, and build the future" [1]. When visualizing the data, time can be treated "just" like any other data dimension using, for instance, parallel coordinates, scatterplots, or other information visualization techniques [1]. In many applications, however, time has a very particular meaning and often a central role in the data. Consequently, we see many approaches that support a special treatment of the time dimension. A number of useful reviews of time-dependent data visualization have been published recently [1, 2, 143, 145, 163, 211]. In the following, a general overview on the visualization of time-oriented data is given. Approaches for the visual analysis of time-dependent data are discussed later in section 2.2.2.

### 2.2.1 Time-dependent Data Visualization

Aigner et al. [1] give a systematic view on the visualization of time-oriented data. In their categorization, they consider different characteristics of the time axis such as temporal primitives (discrete time points vs. time intervals) or the structure of time (linear vs. cyclic vs. branching time). These considerations are important when designing a visual analysis system, since they address the data validity and the possible relations among temporal primitives [1]. Moreover, the authors discuss data-related questions (e.g., abstract vs. spatial data, uni-variate

vs. multi-variate data, original vs. derived data) and different visual representations such as static vs. dynamic or 2D vs. 3D representation. Time-dependent data can be visualized, for example, by using animation techniques (e.g., the study of a numerically modeled severe storm [8, 253]), by displaying the data at individual time steps, or by visualizing the evolution of a data variate over time (e.g., by drawing function graphs). According to Müller and Schumann [163], dynamic representations, such as animations, support qualitative statements on the general evolution of the data over time. Static visualizations,[4] on the other hand, are more suitable for making quantitative statements such as comparing different timespans or searching for time-related patterns. The decision whether to use a 2D or 3D graphical representation in information visualization usually depends on the task at hand [1, 205]. However, some kinds of data (e.g., volumetric data, 3D flow data) inherently require a 3D representation.

The ThemeRiver [78] is an example for a static visualization of time-dependent data. Changes in topics in large document collections are visualized with respect to a linear time axis. The number of occurrences of a certain topic is represented as the width of the corresponding river band. Nocke et al. [168] utilize a ThemeRiver approach for the visualization of clustered climate data. Recently, Byron and Wattenberg [26] propose algorithms for stacked graphs where they emphasize considerations of legibility and aesthetics. Another visualization approach for time-dependent data is, for instance, two-tone color mapping [199] which can be used to compactly represent large amounts of time series. In order to support the analysis of cyclic behavior such as the seasons of a year, for instance, helix glyphs [224] placed on a geographic map can be used.

Ma et al. [143, 145] discuss techniques that support the efficient rendering of time-dependent *volumetric data* such as data compression, automated feature extraction, hardware acceleration, or parallel rendering. Jankun-Kelly and Ma [97] study the generation of a single or multiple transfer functions, which capture important structures in time-varying volume data and can be used for batch-mode rendering, for instance. Woodering and Shen [263] propose chronovolumes that represent multiple timesteps in a single image using color composition techniques.

Besides the visualization of time-dependent scalar volumes, also the visualization of time-dependent vector fields is important in many areas. Such approaches for *flow visualization* can be generally classified into [135, 183, 184, 200]: 1) direct flow visualization such as color coding or arrow plots; 2) dense, texture-based approaches using, for instance, spot noise, line integral convolution, or texture advection; 3) geometric flow visualization depicting geometric objects that are extracted/computed from the flow such as streamlines, stream surfaces, streaklines, or pathlines; 4) feature-based techniques that are based on the extraction

[4] It should be noted that the facility of user interaction or parametrization does not influence whether a technique is considered static or dynamic. While the visualization changes automatically in dynamic representations (without the needs of interaction), the visualization is modified manually by user interaction in static representations [205].

of relevant structures such as vortices or shock waves; and 5) partition-based flow visualization that subdivides the domain with respect to certain flow characteristics. While the first three categories depict basic quantities of the flow, the later two provide a more abstracted view on the data.

### 2.2.2 Interactive Visual Analysis of Time-dependent Data

The approaches presented in the previous section usually reach their limits when representing larger amounts of data with several million entries, for instance. Aigner et al. [2] discuss approaches for analyzing time-oriented data where visual and analytical methods are combined. According to Keim's visual analytics mantra [116], (semi)automated data reduction and abstraction techniques are commonly applied, which transform the time-oriented data into a compressed but still representative form. The resulting data can then be visualized instead of the original one. Many approaches for temporal data abstraction come from the field of data mining (see Keogh et al. [119] for an overview). Examples are clustering [243, 168], principal component analysis [162, 173], or wavelet analysis [93, 265]. Moreover, feature specification via interactive brushing or querying methods is often supported in frameworks for time-dependent data analysis.

#### Temporal Data Abstraction

In order to reduce the data complexity or visual cluttering, spatial and/or temporal aggregation is often applied (see López et al. [140] for an overview). With such an approach, data items sharing the same spatiotemporal domain are summarized and depicted instead of the individual data values. According to Andrienko and Andrienko [5], data aggregation can be done either by calculating data characteristics (e.g., the sum, arithmetic mean, variance) or by grouping techniques such as clustering or binning. Aggregation techniques, however, need to be applied with care to preserve important information.

Common analysis approaches for time-dependent *movement data* include the visualization of raw data, computed summaries, or extracted patterns [4]. Andrienko and Andrienko [6], for instance, visualize movement data as flow maps where the spatial domain is subdivided into appropriate areas (based on significant points in the movement) and aggregated trajectories with common start and end points are visualized as arrows. Janoos et al. [100] analyze pedestrian movement trajectories using a wavelet-based feature descriptor in order to detect anomalies. Willems et al. [257] propose a visualization approach based on the convolution of dynamic movement data with a kernel, where the resulting density field is visualized as an illuminated height map. In our own work [47], we propose interactive plots based on kernel density estimates (KDEs) and show differences between different categories (or bins) of aggregated data.

Nocke et al. [168] discuss visualization techniques for clustered climate data such as the ThemeRiver [78], the Rectangular View [169], or the Cluster Calendar View [243]. The latter, for instance, groups time series over a certain period (e.g., month or day) into clusters. These are then visualized using function graphs and also encoded in color in a calendar-like representation. As a result, the frequency of occurrence of each cluster can be seen as well as the daily trends and patterns. Sukharev et al. [218] perform a correlation study of single and pairs of variables using temporal data clustering and segmentation. Aigner et al. [2] discuss the combination of principal component analysis with the visualization of time-dependent climate data (compare also to Müller et al. [162]). Oeltze et al. [173] include correlation analysis and principal component analysis into the visual analysis of perfusion data.

Jänicke et al. [95, 96] apply concepts from information theory in order to automatically extract distinctive structures in time-dependent data. Regions with different temporal behavior than the rest of the field can be identified using local statistical complexity (LSC). The measure assesses the amount of information from the local past that is necessary to predict the local future. While the original approach [96] was limited to 2D data, the authors propose an improved computation method that is also applicable to 3D data [95]. In later work, Jänicke et al. [93] utilize wavelet analysis for the visual exploration of climate variability changes. The authors apply, amongst others, clustering using mutual information in order to identify coherent structures in the data. Chen and Jänicke [34] recently propose a theoretic framework for visualization that is based on information theory. The authors discuss major concepts of information theory and show the broad correlation to phenomena or events in visualization.

### Time-dependent Feature Specification and Analysis

Feature extraction can either be done (semi)automatically or manually [183, 184]. Several applications support the visual analysis of temporal features using interactive brushing or querying techniques. The TimeSearcher [88] is especially designed for the visual analysis of time-dependent data using Time Boxes or angular query widgets. The latter are applied for selecting time series that have a similar slope on a sequence of time steps (compare to angular brushing [75] described in Sec. 2.3.2). Further extensions of the TimeSearcher [23] allow for similarity-based querying of temporal patterns. Konyha et al. [127] introduce line brushes to select function graphs out of a larger number of graphs, which intersect with a simple line segment drawn in the view. Akiba et al. [3] utilize a Time Histogram [129] showing consecutive 1D histograms for every timestep to simplify the specification of transfer functions for time-varying volume data. Wang et al. [247] utilize Time Histograms and clustering for importance-driven visualization of time-dependent data. Data are partitioned into spatial blocks and corresponding importance values are determined using concepts from information theory (compare also to

Jänicke et al. [96]). Woodering and Shen [265] apply wavelet transformation to time-dependent data. The resulting multi-resolution temporal representation is clustered and visualized in a visualization spreadsheet [98] using multiple Time Histograms that also support linking and brushing.

Feature visualization and specification via brushing in coordinated multiple views is also an integral part of the SimVis framework [52, 54]. In previous work [161], we have proposed a four-level focus+context visualization for large amounts of function graphs together with advanced brushing techniques. This extension to the SimVis framework builds the basis for the application study in paper A. Function graphs that are similar to a pattern sketched by the user can be interactively selected. Transfer functions [102] are applied for visual clutter reduction by mapping the number of function graphs per pixel to the pixel's luminance. Aggregation techniques (frequency binmaps [172]) are used in order to maintain the responsiveness of the system, even when interacting with large data. Blaas et al. [16] utilize similar techniques for the visual exploration of large amounts of time-dependent data in parallel coordinates.

## 2.3 Multi-variate Data Visualization and Analysis

The multi-variate characteristics of scientific data are often of special interest, typically in combination with their spatial and/or temporal reference. When investigating, for instance, the fronts of a storm [125] or environmental phenomena such as the El Niño [93, 227] multiple variates and their relation to each other need to be considered. Riley et al. [189], for example, propose a realistic-looking, physics-based multi-field weather visualization that supports the evaluation and prediction of clouds and storms.

Johnson [104] identifies the visualization of multi-variate scientific data (also referred to as multi-field data) as one of the top challenges in visualization research. Wong and Bergeron [261] as well as Fuchs and Hauser [63] provide comprehensive surveys on the topic. Multiple variates can be visualized jointly in a single image, for instance, by using different textures, colors or glyphs, where one usually has to cope with visual cluttering and occlusion. Alternatively, relations between different variates can be visualized by plotting the data in attribute space (e.g., scatterplot or parallel coordinates) or by specifying features across multiple linked views via brushing. Keim [112] classifies information visualization techniques for multi-variate data by data type (e.g., number of variates, hierarchical data), visualization technique (e.g., 2D/3D visualizations, geometrically transformed displays, glyphs), and interaction and distortion technique (e.g., projection, filtering, zoom, distortion, and linking&brushing).

In the next section, we mainly discuss examples for the visualization of multi-variate data such as preattentive graphical features, glyphs, and layering techniques. Approaches for the interactive visual analysis of multi-variate data such

as feature-based visualization are described in section 2.3.2. Visual data fusion and comparative visualization are important tasks as well and will be presented in the context of multi-modal data (Sec. 2.4).

### 2.3.1 Visualization of Multi-variate Data

Multiple data values can be simultaneously represented at a spatial location using *preattentive visual stimuli* such as width, size, orientation, curvature, color (hue), or intensity [40, 79]. These features are rapidly processed by our low-level visual system and can thus be used for the effective visualization of millions of data items [58]. Special care is required, however, if several such stimuli are combined (the result may not be preattentive any more). Healey and Enns [80] propose simple texture patterns and color to visualize multi-variate data. Different data variates are encoded in the individual elements of a perceptual texture using equally distinguishable colors and texture dimensions such as element density, regularity, and height. In later work [81], the authors utilize simulated brush strokes that vary color and perceptual texture to visualize multi-variate weather conditions.

A powerful way of visualizing multi-variate data are *glyphs* (also referred to as icons, see Ward [250] for an overview). It is important to note that some graphical attributes or their relationships can be easier perceived than others. Since glyphs are usually not placed in a dense way, the free space between them can be used for additional information [124]. Max et al. [44, 153], for example, use splatting to render small colored vector glyphs depicting wind velocity combined with contour surfaces representing cloudiness. Treinish [229] visualizes weather data using color contouring on vertical slices and isosurfaces that represent cloud boundaries. At user-defined locations (vertical profiles), the wind velocity and direction are represented by a set of arrow glyphs. Streamlines that follow the wind direction are seeded at each arrow [229]. Nocke et al. [167] use a metaphor-based iconic visualization for maize harvest predictions, which can represent six different data values. Stier et al. [216] use iconified bar and circle representations displaying four and two different aerosols, respectively.

In the context of information visualization, Ward [249] discusses glyph placement strategies such as data or structure-driven placement. Ropinski and Preim [194] propose a taxonomy for glyph-based medical visualization. The authors categorize glyphs according to 1) preattentive visual stimuli such as glyph shape, color and placement, and 2) attentive stimuli that are mainly related to the interactive exploration phase. Additional usage guidelines are proposed, for instance, that glyph shapes should be perceivable unambiguously from different viewing directions. Kindlmann [122] as well as Jankun-Kelly and Mehta [99], for example, use superquadric glyph shapes that fulfill the latter criterion. Ropinsky and Preim [194], moreover, state that parameter mappings should focus the user's attention and emphasize important variates. Our guidelines for glyph-based 3D

visualization (paper B) is inspired by their work. We extend, for example, the preprocessing step (parameter mapping) prior to the creation of individual glyphs. Also, the task of glyph-based 3D visualization is divided into several steps. Related critical design aspects are then discussed in a structured form, for instance, glyph normalization and orthogonality, depth perception, visual cluttering, and the usage of redundancies in the visualization to ease interpretation.

Further approaches for multi-variate data visualization utilize 2D/3D layering techniques, for example. Kirby et al. [124] use concepts from painting when visualizing 2D flow by combining different image layers with glyphs, elongated ellipses, and color. Wong et al. [262] visualize multi-variate climate data by overlaying multiple see-through layers using opacity modulation, filigree graphics, or 2D height maps. They also propose enhanced color maps that highlight flow features such as critical points or vortices. Shenas and Interrante [208] discuss approaches for effectively combining texture with color to visualize multi-variate data. Two-level volume rendering [69, 76] considers segmentation information when visualizing volumetric data. Based on the segmentation, different rendering techniques such as maximum intensity projection (MIP), direct volume rendering, or non-photorealistic techniques are combined in a single visualization. In recent work, Rautek et al. [187] propose semantic layers for illustrative volume rendering, where the mapping of data properties to visual styles can be specified using natural domain language. Further approaches for visual data fusion are discussed in the context of multi-modal data in section 2.4.1.

### 2.3.2 Visual Analysis of Multi-variate Data

Multi-variate data are commonly analyzed using multiple linked views, which were already discussed in section 2.1.1. Attribute views such as scatterplots, function graphs [127, 161], or parallel coordinates [16] allow the user to investigate the relations between data variates and brush interesting subsets of the data. Hauser et al. [75], for instance, introduce angular brushing for parallel coordinates, where data correlations between adjoining axes can be studied by selecting line-segments with a similar slope. Some systems [16, 52, 66] utilize attribute views side-by-side with linked 3D representations of volumetric data, where the specified features can be explored in their spatial relation.

Kniss et al. [125] propose multi-dimensional transfer functions to specify features in meteorological data. Interesting data subsets can be selected both in volume and transfer function space using a set of direct manipulation widgets. To support the higher dimensional classification of volume data, Tzeng et al. [235] propose an intelligent painting interface. Regions can be marked directly on sample slices in the volume space and the data are then classified automatically using a supervised machine learning approach. The training data can then as well be used for classifying other data with similar characteristics. Ma [144] discusses further applications of machine learning and visualization such as flow feature

extraction and feature tracking. Recently, Fuchs et al. [64] combine interactive visual analysis with machine learning. The user specifies an initial hypothesis via linking and brushing, and a heuristic search algorithms then finds alternative or related hypotheses that explain the same feature. The most suitable hypotheses can thus be identified out of a large search space using different fitness criteria.

Johansson et al. [102] utilize clustering and high-precision textures in order to identify structures within parallel coordinates. To overcome the limitations of the output device, they apply user-defined transfer functions that map the number of primitives per pixel to the pixels' luminance. Novotný and Hauser [172] utilize 2D binmaps for aggregating the data between each pair of adjacent axes in parallel coordinates. Outlier detection as well as clustering are applied on the binmaps in order to visualize data trends while preserving outliers in a focus+context style. Further approaches [16, 161] combine both binmaps [172] and transfer functions [102] in order to efficiently visualize large data.

As discussed in the context of visual analytics (Sec. 2.1.2), dimensionality reduction techniques are often applied when analyzing multi-variate data. Data are projected to a low-dimensional space while preserving their meaningful structures and relationships. The grand tour method [7], for example, automatically generates a sequence of orthogonal projections onto a 2D subspace, which can be used in an animation. Seo and Shneiderman [206] introduce the rank-by-feature framework, where low-dimensional projections of multi-variate data such as histograms or scatterplots are ranked based on a user-selected criterion (e.g., correlation or entropy in scatterplots). A triangular matrix represents the possible combinations of data variates in a scatterplot and encodes the corresponding ranking score in color. This supports the user to select interesting views on the data. Tatu et al. [220] propose further quality measures for scatterplots and parallel coordinates that are utilized for ranking these views. Johansson and Johansson [103] propose a system for dimensionality reduction that uses a combination of user-specified quality metrics to preserve important structures in the data. Jänicke et al. [94] transforms data onto a 2D point cloud, where data items with similar multi-variate characteristics are located close to each other. The authors compute a tree where the Euclidean distance between multi-variate data items is minimal. This structure is then utilized when transforming the data to 2D. Additional information is encoded using color and point size, and interesting structures can be selected interactively via brushing [94].

## 2.4 Visualization and Analysis of Multi-Modal Data

Data stemming from different acquisition modalities are common in many physical sciences including climate research, meteorology, physics, and astronomy [138, 238]. A simulation model can, for instance, be validated by comparing it to the output of another model or measurement data from weather stations or

satellite observations. The data in such scenarios can be given on various types of grids (e.g., 2D/3D, unstructured, hybrid) with different time steps or resolutions. Accordingly, multi-modal data often need to be fused in the visualization, for instance, by resampling them to a common grid. Also in the medical domain, data increasingly often stem from different measurement techniques such as CT, MRI, or positron emission tomography (PET). Combining such modalities in a visualization can account for the strengths and weaknesses of the individual ones. Bones, for instance, are best captured by CT, soft tissue such as the brain is better represented with MRI, and PET data can be used to study functional processes in the body. In the following, approaches for visual data fusion (Sec. 2.4.1) and comparative visualization (Sec. 2.4.2) are discussed.

Treinish describes four perspectives for visual data fusion and comparison depending on the visualization task [231]: 1) at the image level using adjacent windows or mosaiced visualizations for qualitative comparison; 2) combining different visualization techniques in the same coordinated system (common view), including interactions such as numerical querying; 3) numerical comparison on the data level using one of the two previous approaches; and 4) coordinated multiple views supporting numerical and visual comparison. When visualizing multi-modal data, the different data sources first need to be registered and normalized to each other in order to make them comparable. The visualization also has to be designed carefully to avoid the introduction of artifacts that can be erroneously interpreted as features [231].

### 2.4.1 Visual Data Fusion

Multi-modal scientific data can be fused at different levels of the visualization pipeline [27, 63]: 1) during data filtering and visualization mapping, for instance, by reducing the data to relevant features or by resampling to a common grid; 2) during accumulation in the rendering stage [15, 68]; and 3) in image stage, for example, using layering techniques [89, 124, 221, 262]. Multi-block flow visualization is an example where simulations are performed on multiple grid types with different resolutions [56]. When visualizing the data, these blocks are commonly fused at the data level by constructing a common grid (compare to cross-mesh field evaluation [37]). As a result, multi-variate visualization techniques as described in section 2.3 can be applied. We find, for instance, rendering approaches for non-uniform grids [46, 190] or hybrid and non-structured grids [160] that are also integrated in frameworks for visual analysis and exploration [52, 215].

Treinish [227] discusses treatments for scattered meteorological data such as constructing a grid using Delauney triangulation, resampling to a regular grid using the nearest neighbor, or weighted average gridding. In later work, the same author proposes a function-based data model [228] that provides uniform data access by adjusting to the data structure and the way data are processed. Consequently, the same operations can be applied to data from various sources without

resampling to a common mesh or unnecessary interpolation. Treinish [230] also proposes a multi-resolution approach for nested meteorological data.

Cai and Sakas [27] propose an approach that uses the different modalities as parameters to a multi-volume illumination model (in the visualization mapping stage). As an alternative, the same authors suggest methods to combine color and opacity from different volumes during accumulation, where each volume has it's own transfer function [27]. Chen and Tucker [35] propose constructive volume geometry, where multiple volumes are combined using algebraic operations. Woodering and Shen [264] build on this approach and propose volume shaders to combine and compare multiple time-dependent volumes using consecutive algebraic set operators as well as numerical operators. For interaction and visualization of the resulting volume tree they utilize image spreadsheets (compare to Jankun-Kelly and Ma [98]). Beyer et al. [15] present a system for preoperative planning of neurosurgical interventions. Similar to two-level volume rendering [69], the authors render segmented multi-modal data directly on the GPU. Burns et al. [25] combine tracked ultrasound data with direct volume rendering using flexible cutaways and importance-driven shading. Context information occluding the ultrasound image can thus be removed and features can be enhanced (compare also to Viola et al. [245]). Ropinski et al. [196] introduce interactive closeups for multi-modal medical reporting.

Grimm et al. [68] performed fusion of multiple intersecting volumes during the rendering stage. They propose V-objects that represent visual properties of the individual volumes such as illumination, transfer function, region of interest, and transformation. The data are visualized efficiently using a brick-wise ray traversal scheme and mono-volume rendering for non-intersecting areas. Similar to this approach, Plate et al. [182] present a framework to render large, arbitrarily oriented volumes using slice-based rendering on the graphics hardware. Their approach also supports out-of-core techniques and volumes given at multiple resolutions. Recently, Lindholm et al. [137] introduce a region-based scene description for GPU-based direct volume rendering. Using binary space partitioning, the depth information of the intersecting geometry is stored in a view-independent way and expensive depth sorting can be avoided.

### 2.4.2 Visual Analysis for Comparison

The objective of comparative visualization is to show similarities and differences between data stemming from two or more sources [175] (e.g., measurements, computational simulations, or different simulation models). According to Verma and Pang [244], comparison can be done at the image level, the data level, or the feature level. Pagendarm and Post [175] illustrate a number of examples where comparison is performed at the data and image level.

*Image level comparison* is the most frequent one. It does not directly operate on the data but on 2D images that result, for example, from a visualization method,

from experiments [267], or Schlieren photography [175]. Examples include side-by-side visualizations where the user has to visually compare images. Other approaches directly represent per-pixel differences by subtracting two registered images from another [47]. For such approaches, the selection of an appropriate color map is highly important, for instance, using a diverging map to discriminate positive and negative differences [19]. Zhou et al. [267] present a study of different comparison metrics that numerically quantify image differences between experiments and visualizations. It should be noted that image level comparison operates on 2D representations where the intermediate information about how the images were created is usually lost. Deviations can, therefore, also result from different visualization settings (e.g., transfer function, point-of-view, lighting conditions) and need not necessarily represent data differences [244].

*Data level comparison* utilizes the raw data as a starting point and often incorporates intermediate information from the rendering process [226]. Sahasrabudhe et al. [198] propose methods for measuring the differences between scalar datasets including spatial and perceptual metrics. Kim et al. [120] propose metrics for data level comparison of direct volume rendering and also incorporate intermediate rendering information in their comparison approach. Sauber et al. [203] introduce multified-graphs that represent correlations between different data variates. Malik et al. [147] recently propose the multi-image view that supports the comparison of series of scans from the same specimen.[5] Such approaches are usually superior to pure image level comparison since they include more information and flexibility. Finally, *feature level comparison* is an extension of data level comparison and is based on features that are extracted from the data. For flow data, such features can be shock waves, vortices, streamlines, or isosurfaces (see the work of Pagendarm and Post [176] or Verma and Pang [244], for instance).

## 2.5 Multi-run Data Visualization and Analysis

The previous section discusses visualization and analysis approaches for a relatively small number of data volumes. For comparing such data, for instance, side-by-side visualizations or isosurfaces can be used [166]. However, the visualization and visual analysis of a larger number of concurrent data volumes requires more sophisticated methods. Such data commonly results from *multi-run* (or ensemble) simulations, which are performed increasingly often in climate research [86, 166] or automotive engineering [152, 180].

Multi-run simulations are an important step in the development of simulation models, where one aims to identify model parameters that have the most influence on the simulation output. In such a *sensitivity analysis* [70, 82, 83], the values of certain model parameters are changed systematically and multiple simulation

[5] To a certain degree, such dataset series resulting from multiple scans can be considered as multi-run data.

runs are computed, accordingly. In the resulting data, a distribution of values is given for the same data variate at each position in space and time (one value for each run). The representation of such multi-run data is rather new to the visualization community [107, 108, 141]. It is especially challenging since the data are often time-dependent, higher-dimensional, multi-variate, and large at the same time [258]. A direct visualization of such time-varying volumes of data distributions is often not feasible. Accordingly, the individual distributions of multi-run values need to be analyzed first, and then derived statistical properties can be visualized (compare to Keim's mantra [116]).

The visualization of multi-run data is especially interesting since it is an alternative approach for representing uncertainty [186, 258]. General approaches for uncertainty visualization are discussed by Pang et al. [177], Johnson and Sanderson [105], and Griethe and Schumann [67]. MacEachren et al. [146], moreover, review approaches for geospatial uncertainty visualization and Skeels et al. [213] survey related approaches for information visualization.

In this section we describe applicable visualization techniques such as coordinated multiple views, visualization of statistical parameters, shape descriptors, and operators (compare also to Love et al. [141]).

### 2.5.1 Visualization of Multi-run Distributions and Derived Data

A standard in statistics for representing data distributions are *box plots* [154], which encode minimum and maximum values, mean, median, and other quartile or percentile information. Kao et al. [108, 107] extend this approach to 2D multi-run data. In some cases, the distribution can be represented adequately by *statistical parameters* such as mean, standard deviation, interquartile range, skewness or kurtosis. The computed statistics are visualized on 2D surfaces using color-coding and bar glyphs. Where this is not the case, the same authors propose a *shape descriptor* approach constructing a 3D volume, where the data range is handled as a third dimension and the probability density function (PDF) of the multi-run data is used as voxel values. Furthermore, the peaks in the PDF are described by a set of shape descriptors (e.g., number of peaks, height, width, and location), which are displayed on orthogonal 2D slices [108].

Another extension of box plots is presented by Potter et al. [186]. The proposed summary plot includes additional statistics of the multi-run data such as skewness, kurtosis and tailing information. These plots, however, cannot be placed in a dense 2D/3D context. Spaghetti plots [50] are commonly utilized by meteorologists to investigate multi-run data, where a contour line is visualized for each run at a selected time step (resembling a pile of spaghetti noodles). Sanyal et al. [202] combine such plots with a ribbon and glyph-based uncertainty visualization. The uncertainty glyphs consist of a number of concentric colored circles that represent the standard deviation, interquartile range, and the width of the

95% confidence interval. In paper C, we use carefully designed glyphs (paper B) to visualize aggregated data properties in a 3D context.

Mathematical and procedural *operators* [141, 142] can be utilized to transform multi-run data into a form where existing visualization techniques are applicable (e.g., pseudo-coloring, streamlines or isosurfaces). The multi-run distributions can, for instance, be compared against a single threshold value or against a reference distribution when drawing contour lines or isosurfaces. This approach is very promising due to its flexibility. However, the usage of the operators and the interpretation of the resulting visualizations require additional training and care from the user.

### 2.5.2 Interactive Visual Analysis of Multi-run Data

In the visual analysis of multi-run data, information visualization techniques such as parallel coordinates or scatterplot matrices can be combined with statistics [41]. Potter et al. [185] propose a framework for analyzing multi-run data, which consists of overview and statistical visualizations such as trend charts or spaghetti plots [50]. Nocke et al. [165, 166] present a coordinated multiple views framework for analyzing a large number of tested model parameters and simulation runs. Statistical aggregations of multi-run simulations can be visualized, for instance, using linked scatterplots, graphical tables, or parallel coordinates. The sensitivity of the model to certain input parameters can be explored via brushing, and the related model runs can be compared in detail (compare also to the work of Matković et al. [152] on injection systems simulations).

Matković et al. [151] visualize multi-run data as families of data surfaces with respect to pairs of independent data dimensions. Using multiple linked views and brushing, the authors analyze projections and aggregations of the data surfaces at different levels (e.g., a 1D profile or single aggregated value per surface). In paper C, we propose a more general interface concept that relates data items between different parts of scientific data and supports the transfer of fractional DOI information. This approach can also be used for multi-run data.

Matković et al. [150] also propose a visual steering approach where new simulation runs are triggered by interactively narrowing down the control parameters in the visualization. This approach realizes a tight integration of visualization and computational simulation. In recent work, Matković et al. [149] propose the simulation model view which is integrated in their coordinated multiple views framework. The view represents the building blocks of the utilized simulation process and model at three different levels of detail (using a histogram, scatterplot or curve view). The approach aims at bridging the gap between the simulation model and resulting multi-run data.

Van Wijk and van Liere [242] propose HyperSlice, which represents higher dimensional data as a matrix of orthogonal 2D slices around an $m$-dimensional focal point. The Prosection Matrix [234] extends this concept by projecting

higher dimensional data points that are in proximity to the 2D slices to scatterplots. The approach supports also filtering via brushing. Piringer et al. [180] build upon these concepts and propose an interactive system for visual validation of regression models. The authors utilize 2D and 3D projections of multi-run data around a user-controlled focal point. Known results can then be compared to model predictions (represented as families of function graphs), which supports the identification of regions with bad fit. Also, derivations from the expected values can be computed and visualized together with other data variates in scatterplots or parallel coordinates.

Bordoloi et al. [17] apply hierarchical clustering techniques on multi-run data. Data can either be clustered along the spatial dimensions by grouping locations with similar statistical properties and probability density functions of multi-run values—this approach helps to identifying spatial structures and patterns, which may result from the simulated phenomenon. Alternatively, the runs can be clustered base on their similarity. Such an approach supports the comparison of different groups of simulation outcomes, where each group can be represented [17]. In recent work, Bruckner and Möller [20] present a result-driven exploration approach for physically-based multi-run simulations. Each volumetric time sequence is first split into similar sequences and thereafter grouped across different runs using a density-based clustering algorithm. This approach supports the user in identifying similar behavior in different simulation runs. In an overview visualization, each cluster is depicted with respect to a common time line. Paths drawn between the clusters show the progression of the individual sequences.

Finally, Correa et al. [43] propose a framework for uncertainty-aware visual analysis. Statistical methods are incorporated such as uncertainty modeling as well as uncertainty propagation and aggregation during data transformations. The authors adopt approaches for data transformation such as regression, principal component analysis, and $k$-means clustering in order to account for uncertainty. A number of views are presented that combine summarized and detailed uncertainty visualizations. Dependent on the task of the analysis, uncertain data can be enhanced or de-emphasized. In later work [31], the same authors augment traditional scatterplots by visualizing sensitivity information, which they considered similar to velocities in a flow field. Sensitivities can thus be represented as tangent lines on the individual points in the flow-based scatterplot. Moreover, the assumed flow field can be visualized using streamlines, and data points can be clustered by proximity to these lines. The proposed approach allows the analyst, for instance, to correlate changes in one variate with respect to another.

We clearly see a lot of potential for future research along with this kind of data. Due to the technological developments in climate research, engineering, and other fields, we see that this kind of data gains increasing importance. Visualization must deal with data that are multi-variate, time-dependent, and multi-run data. It is not at all straightforward to visualize an overview of several hundred runs of

time-dependent 3D data. Advanced data abstraction and aggregation techniques are required that are aware of data trends and outliers.

## 2.6 Multi-Model Data Visualization and Analysis

Multi-model simulations are gaining importance in areas like climate research [86] or multi-physics research [22]. In the climate system, for instance, different compartments such as ocean, ice, surface, and atmosphere are interacting with each other. Ocean and atmosphere exchange through thermal absorption, precipitation, and evaporation, also ice and air are interacting. Accordingly, ocean and atmosphere models are often coupled in the simulation (e.g., ECHAM5/MPI-OM [106]). The models are often not computed on the same types of grid, or for the same time steps. When analyzing feedback between these models, statistical aggregates are usually investigated.

Fluid–structure interactions (FSI), to address another example, are interactions of a deformable or movable structure with an internal or surrounding flow. They are among the most important and—with respect to both modeling and computational issues—the most challenging multi-physics problems, and therefore currently a hot topic in simulation research itself. The variety of FSI occurrences is abundant and ranges from bridges, flexible roofs, or off-shore platforms to micropumps and injection systems, from parachutes via airbags to blood flow in arteries or artificial heart valves, to name just a few [22].

For visualization research it is very challenging to generate a coherent visualization of these datasets, for instance, when one model is simulated on a 2D grid and the other one on a 3D grid. How can different attributes given in the different models be compared to each other? How can selections and features be communicated between different models, when these are given on different grids and time steps? How can data be represented, where there are values missing (e.g., an attribute is simulated in one model but not in the other, or the data are not uniformly available for a spatial dimension). One can image a visualization approach, where different representations are coupled together in the visualization (similarly to the simulation). Our work presented in paper C goes a first step in this direction. However, we definitely see a great potential for future visualization research here.

## 2.7 Chapter Summary and Conclusions

We see that the visualization and interactive visual analysis of multi-faceted scientific data are gaining increased importance in areas such as climate research, engineering and medicine. This is due to the fact that computational power increases rapidly, and measurements are getting more accurate and detailed. Ac-

cordingly, also model and data scenarios are getting more complex. Visualization has been well established to explore and analyze such data and to communicate results from data analysis. With respect to multi-faceted data, we see a variety of interesting challenges that require advanced visualization technology. In this chapter, the related state of the art has been discussed.

As one interesting observation, we see a gap between the techniques used by domain scientists and the approaches available from visualization research. Recent advances in visualization are rarely used in application domains such as climate research (compare to Nocke et al. [170]). A major challenge for future developments is thus to further bridge this gap by including sophisticated visualization technology in the application domain as well as by including knowledge from domain experts when designing visualization solutions [104, 241]. Visualizations should follow guidelines from perception research and human–computer interaction, providing simple graphical user interfaces and advanced visualization methods. Examples are feature-based approaches that (semi)automatically extract unknown and interesting patterns from the data [54, 94]. Especially the combination of automated analysis approaches and interactive visualization methodology—as proposed in the visual analytics agenda [223]—is a promising direction, and we expect to see a lot of more interesting work in this area. A further step is the integration of machine learning methods that can learn from previous data and user input, and configure the control parameters of the visualization based on the acquired knowledge [33, 144].

We identify the visualization and analysis of data stemming from multi-run simulations and interacting simulation models (e.g., coupled climate models or multi-physics simulations) as promising directions for future research, as well as multi-modal visualization. A challenge is to jointly integrate larger amounts of concurrent data volumes in the visualization/analysis, possibly given on different grids and/or with different data dimensionality. Moreover, how to investigate feedback between interacting compartments in the simulation. For multi-variate and time-dependent data visualization, we can find a lot of related work that brings up good solutions. The visualization and analysis of these kinds of data belong to the top challenges in current visualization research [104].

# Chapter 3

# Interactive Visual Analysis of Multi-faceted Scientific Data

This chapter discusses the different contributions that are made by the papers in the second part of this thesis. In section 3.1, an application study of time-dependent climate data is described. We demonstrate how interactive visual exploration can be used to rapidly generate promising hypothesis that are subsequently evaluated using classical statistics. In addition to generating hypothesis, the parameter ranges for the computational analysis can be narrowed down efficiently. The study was done in collaboration with climate researchers and lead to two further publications (Ladstädter et al. [132, 133]), which are discussed here as well. For glyph-based 3D visualization, structured guidelines are proposed that are based on critical design aspects (see Sec. 3.2). A new glyph is presented and used to illustrate the different design considerations.

Section 3.3 describes an approach for interactive visual analysis of heterogeneous scientific data. The data consist of multiple parts such as multi-run data and aggregated statistics as well as data from a multi-physics simulation. By constructing a so-called interface that relates data items across the different parts, the joint investigation of features is supported. The proposed interface builds also the basis for a study of multi-run data that is based on statistical properties (see Sec. 3.4). Traditional and robust estimates for the four moments are computed from the higher dimensional multi-run data. Additional measures of *outlyingness* are incorporated in our framework of visual analysis. A model for interactive visual analysis is proposed, which represents a structured guide to the multitude of opportunities of moment-based visual analysis.

While this chapter focuses primarily on the contributions of my work, a number of related demonstration cases are presented in chapter 4.

## 3.1 Hypothesis Generation with Visual Exploration

The generation of hypotheses via interactive visual exploration is one of the major application goals for visualization [116, 233] (besides confirmative analysis and the presentation of results). While computational analysis such as statistics commonly requires a hypothesis beforehand in order to work, it is occasionally

rather hard to come up with such concrete application questions. It is challenging as well to identify features that are not anticipated prior to the analysis (compare to Shneiderman [210], for example). Statistical methods such as linear trend regression, multi-variate data analysis, or pattern analysis are thus more suitable to quantitatively and accurately check specific hypotheses [71]. Interactive visual exploration, on the other hand, supports the efficient generation of hypotheses in an undirected search process [209, 233]. The whole data can be explored interactively, for instance, by looking at relations between different variates using multiple linked views and feature specification via brushing [73, 192]. Additionally, it is often necessary to derive new data attributes from existing ones using automated analysis methods [113, 116]. The hypotheses resulting from such an exploration process can then be evaluated, for example, using statistical methods or interactive visual analysis (confirmative visualization). Interactive visualization can be utilized in order to understand the output of different stages of the computational analysis and to narrow down the scope of the analysis [260] (e.g., finding appropriate parameter settings and boundary conditions). Interaction and flexibility of the application are crucial for the entire exploration and analysis process, supporting the user to query the data in many different ways [217].

## Hypothesis Generation in Climate Research

In our application case, we were collaborating with climate researchers in order to generate hypotheses related to climate change. Specifically, we were interested in identifying particular atmospheric regions (in space and time) that represent potentially sensitive and robust *indicators* for atmospheric change. Important climate parameters such as temperature or geopotential height, which are among the candidates for such sensitive indicators [62, 109], are investigated. The study is carried out in the SimVis framework for visual analysis [52] that has been extended in order to deal with large and time-dependent data (Muigg et al. [161]). We have integrated new attribute views that are capable of visualizing a larger number of function graphs using a four-level focus+context style. Brushing techniques for time-dependent data were developed as well, where function graphs can be selected based on their similarity to a user-defined pattern that is sketched in the view [161]. Such advanced brushing methods in combination with computation analysis enable a powerful visual analysis approach (compare to the levels of visual analysis by Hauser [74]).

In order to generate the hypotheses, *temporal trends* as well as the corresponding *signal-to-noise ratio* (SNR) values are computed from selected climate parameters using the integrated data derivation mechanism of our analysis framework. The derived data variates are then interactively explored via brushing in order to located the sensitive indicator regions in space and time. An useful analysis pattern thereby is to investigate if implications between data variates such as $a \rightarrow b$ exist in the opposite direction as well ($a \leftarrow b$). The respective statement

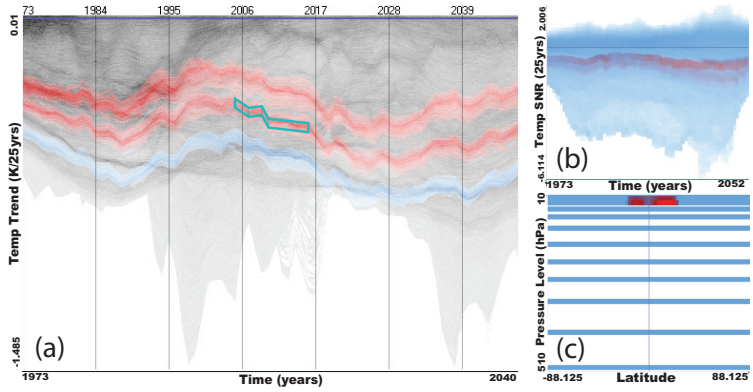



**Figure 3.1:** A prominent visual structure in the function plots view is brushed based on it's similarity to a user-defined pattern (a). The related feature exhibits a relatively high signal-to-noise ratio highlighted in (b), and can be located in the upper pressure levels, centered around the tropical region (c).

can be considered stronger if such an interrelation ($a \leftrightarrow b$) can be confirmed, which can help to direct the analysis. For example, certain temporal trends in the ECHAM5 temperature field [193] can be identified in a function graph view when selecting high absolute SNR values in order to locate sensitive regions. Using a similarity-based brush [161] in Fig. 3.1a, the previously highlighted temporal trends can be selected and checked whether a similar feature emerges in the SNR data (Fig. 3.1b) as well as in the spatial context (Fig. 3.1c).

In addition to generating hypotheses in the first place, visual exploration is utilized to efficiently narrow down certain parameters that are required for the subsequent statistical analysis. The utilized least-squares fitting method [131], for example, requires the timespan over which curves are fitted together with latitude ranges. Using visual exploration, we could figure out that the utilized statistical method reacts more sensitive with respect to the chosen timespan than expected. In contrast to classical statistics, no assumptions of an underlying data model are required for the visual exploration process. The whole field can be investigated at once without the need to preselect certain geographical regions (as required by the utilized statistical approach [131]). Possible data deficiencies can thus be efficiently detected and taken into account.

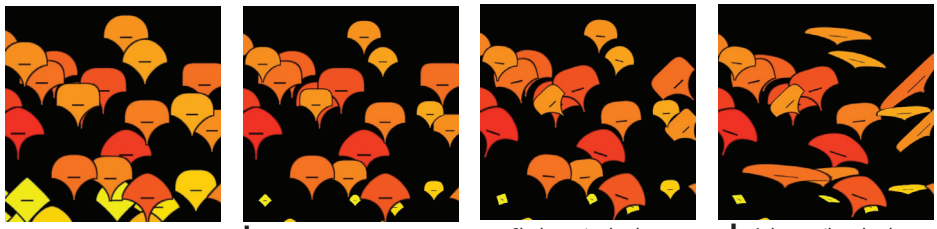
Our approach is demonstrated on a number of representative datasets including measurement and simulation data. The results from the computational analysis of the generated hypotheses give us confidence that visual exploration, although not meant to provide quantitative results, serves as an excellent complement to statistics in an integrated workflow (compare also to Ladstädter et al. [133]). Selected parts of our study can be found in section 4.1. Further details are given in paper A and the work of Ladstädter et al. [132, 133].

## 3.2 Critical Aspects of Glyph-based 3D Visualization

While the previous section focused on time-dependent data, this section addresses the visualization of multi-variate 3D data. Glyphs are often used to simultaneously represent multiple data variates in the same image [250]. The different variates are thereby represented by a set of visual properties of the glyph including shape, size, color, orientation, etc. It is important to note that certain of these properties are more prominent and thus can be easier perceived and related than others (compare to preattentive visual stimuli [40, 79]). An appropriate glyph design is thus crucial for an effective visualization, where different graphical properties are carefully chosen and combined. In this section, we discuss such critical design aspects for a glyph-based 3D visualization and propose related guidelines. This work is inspired by the work of Ropinski and Preim [194]. We divide the task of creating a glyph-based 3D visualization into three stages (compare to the visualization pipeline [77]): 1) during *data mapping*, the data variates are possibly enhanced and mapped to the different glyph parameters; 2) *glyph instantiation* creates the individual glyphs; and 3) during *rendering*, the glyphs are placed in the visualization, where one has to cope with issues such as visual cluttering or occlusion. In the following, we discuss critical design considerations during each of these steps. The different aspects are illustrated with a new glyph-based visualization of 3D data.

We consider it useful that the glyphs expect normalized input such as $[0, 1]$ from the depicted data variates (compare also to Ward [249]). During data mapping, we consider three important steps where the depicted variates are enhanced. First, the data values within a user-selected range $[w_{left}, w_{right}]$ are linearly mapped to the unit interval in order to enhance the contrast (windowing). Values outside the range are clamped to the boundaries. After the windowing, an optional exponential mapping $e(x) = x^c$ can be applied in order to further enhance the data. Finally, a third mapping step enables the user to restrict the output range that should be depicted by a glyph property. Here, also semantics of the data variates can be considered (compare to Ropinski and Preim [194]). Using a reverse mapping, for instance, smaller data values that are possibly more important can be represented in an enhanced style while larger values are deemphasized. When rotation is used to represent a data variate, as another example, the user may want to restrict the rotation angle in this final mapping step ($-45°$ to $45°$).

Several consideration are important for the instantiation of the individual glyphs. When using a 3D glyph shape, one has to account for possible distortions introduced when viewing the glyph from a different point of view [122]. In order to avoid this problem, we strongly suggest to use 2D billboard glyphs instead. In certain scenarios, however, it makes sense to use 3D glyphs, for example, when depicting a flow field via arrow glyphs. Another challenge in glyph design is the *orthogonality* of the different glyph components, meaning that it should be



**a.** 2 data attributes represented as upper/lower glyph shape **b.** Added data attribute to overall glyph size **c.** Glyph rotation has been assigned a data attribute as well **d.** A data attribute has been assigned to glyph aspect ratio

**Figure 3.2:** Adding more attributes to the glyph, while preserving the glyph's orthogonality.


**Figure 3.3:** The upper and lower glyph shape are based on super ellipses and can each represent a data variate. The overall glyph size is normalized in order to account for implicit size changes introduced by the glyph shape.

possible to perceive each property individually (or to mentally reconstruct them as suggested by Preim and Ropinski [194]). In this context, the number of data variates that can be depicted must be seen in relation to the available screen resolution. Large and complex glyphs can be used when only a few data points need to be visualized (compare to the local probe [48], for example). If many glyphs should be displayed in a dense manner, however, a more simple glyph may be desirable [123].

In Fig. 3.2a–d, an increasing number of variates is represented by our proposed glyphs. The use of glyph size and aspect ratio should be handled with care, since these glyph properties may distort the interpretation of others. Size can be used, for instance, to focus on important aspects of the data (similar to a focus+context style). Fig. 3.3 shows how the upper/lower glyph shape represent a data variate by changing from a star (small value), to a diamond, to a circle, and a box representing a large value. Since the changes in shape affects the area (size) of the glyph, we suggest to *normalize* these effects against each other. Accordingly, the overall glyph size is altered in order to compensate for these implicit changes. Another design guideline is the usage of *redundancies*. Our glyph is horizontally symmetric which should make it easier to mentally reconstruct the glyph shape when parts of it are occluded. Important properties can, moreover, be mapped to multiple glyph characteristics in order to reduce the risk of information loss. When designing glyphs, it is especially important to consider how different glyph properties interact with each other and thereby possibly distort the interpretation (compare to glyph size and aspect ratio).

Important aspect when rendering many glyphs in a dense 3D context are depth perception, occlusion, and visual cluttering. *Halos* can help in cases where many glyphs overlap in order to enhance the depth perception and to distinguish individual glyphs (compare to Piringer et al. [181]). For improving the depth perception for non-overlapping glyphs a special color map (*chroma depth* [225]) can be used to represent depth. Finally, appropriate glyph placement [194, 249], interactive slicing, or filtering via brushing are strategies for dealing with occlusion and cluttering issues.

The proposed glyphs are demonstrated in the study of a Diesel particulate filter (Sec. 4.2) and in the visual analysis of multi-run data, where aggregated data properties are represented by the same glyphs (Sec. 4.4.1). Further details with respect to our glyph design can be found in paper B.

## 3.3 Visual Analysis across Two Parts of Scientific Data

Scientific data in classical application scenarios are usually given in a coherent form, similar to a table with rows and columns that is given in relation to space and time. In practice, however, we increasingly often find data and model scenarios that are more heterogeneous. The data consist of *multiple parts* stemming, for instance, from numerical models that simulate different interacting phenomena. Examples are multi-physics simulations such as fluid–structure interactions [22] (FSIs) as well as coupled climate models [86]. While these scenarios are getting increasingly popular in different application fields, they are hardly addressed in visualization research (compare to multi-model scenarios described in Sec. 2.6). Other examples include scenarios where data are given with different dimensionality, for instance, 2D/3D data, time-dependent data, or higher dimensional data stemming from multi-run climate simulations (with additional independent dimensions representing different simulation parameters [86, 151, 152]). In the analysis, one often aims at reducing the data dimensionality, for instance, by computing statistical aggregates with respect to an independent data dimension [5] (e.g., calculating temporal or spatial mean values). Often, only the aggregated data part is further analyzed, accepting that the details from the original data are lost. In our work, however, we integrate both data parts, the original multi-run data and the aggregated data, into the visual analysis.

The challenge with scenarios as described above is to integrate multiple data parts into the visual analysis and to support the investigation of relations and feedback between the parts. One is, for example, interested in the areas of an ocean model that are influenced by adjacent atmospheric regions that exhibit certain characteristics such as high temperatures. How can such a feature from the atmosphere be propagated to the ocean part? It should also be possible to direct the analysis in the opposite direction, for instance, specifying an ocean feature and further examine it in the atmospheric part. Our idea is to use the fractional

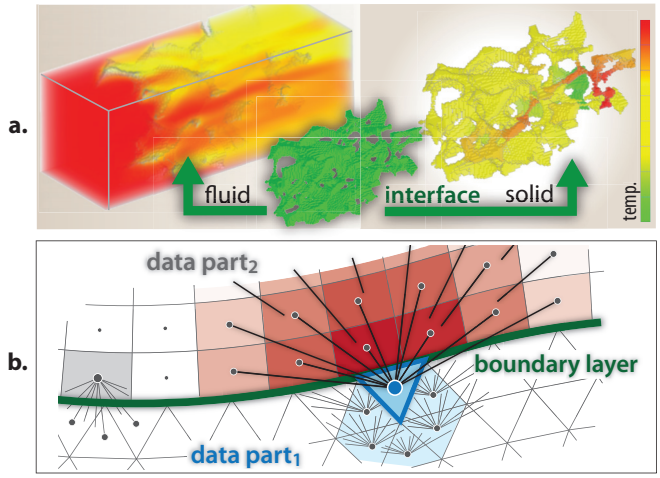



**Figure 3.4:** In a fluid–structure interaction simulation (a), fluid and solid parts are connected via an interface that relates cells sharing a common boundary. A similar interface is constructed for the visual analysis (b). The influence (weights) of the grid cells related to a certain cell (blue) are encoded in red.

*degree-of-interest* (DOI) attribution, resulting from smooth brushing [53], as a common level of data abstraction between the related data parts. Such markups represent the first interpretation level, ranging from data to knowledge (compare to Chen et al. [33]).

We propose a concept that enables the bidirectional transfer of user-specified features between two related data parts. Similar to a fluid–structure interaction scenario (see Fig. 3.4a), we create a so-called *interface*[1] that connects individual grid cells between the two data parts and enables the transfer of DOI information. Our interface is inspired by the data state reference model [36] and consists of: 1) a *structural relation* that specifies which grid cells at which time steps are related between both data parts; 2) a *feature transfer*, i.e., different ways of how the DOI information—resulting from smooth brushing—is exchanged across the data parts between the related grid cells; and 3) an *automatic update mechanism* that keeps the feature specification in both data parts consistent during the visual analysis. A similar coordination space is implicitly given in the model-view-controller pattern [18] or cross-filtered views [252]. With our approach, however, we account for the heterogeneity of independent data dimensions (compare to multi-run data). Features can be transferred as well between spatially adjoining

[1] Many disciplines including physics, biology and computer science utilize this term. According to the Oxford English dictionary, an interface signifies "a point where two things meet and interact."

data parts similar to fluid–structure interactions, leading to a joint focus–context discrimination that accounts for the fractional DOI information.

Analogous to relational data bases, different data items can form an one-to-one, one-to-many, or many-to-many relation (compare to North et al. [171]). This relationship is specified when creating the structural part of the interface, for instance, in a preprocessing step. Additional *weight values* are assigned to each connection between two cells. During the visual analysis, these values are then considered when transferring the fractional DOI information between the data parts and determine the influence a related data item has on the item in question. The relationship between fluid and solid parts in an FSI scenario, for example, can be translated into a many-to-many relation. Grid cells sharing a common boundary are then connected as illustrated for an example cell (blue) in Fig. 3.4b. The weight values are encoded in red, representing that grid cells that are located close-by have more influence than cells located farther apart. A similar relation can be established between (partially) overlapping data parts from multiple sources, which are possibly given at different grids and/or resolutions (i.e., multi-modal data as described in Sec. 2.4). Instead of resampling the data to a common grid, a many-to-many relation can be established, where the weight values represent the spatially overlapping volume of the related grid cells (similar to an interpolation during resampling).

Two kinds of relationships can be specified between multi-run and aggregated data: an aggregated grid cell can be related to the multi-run cells that share the same space and time, and vice versa. Such a one-to-many relationship is utilized in the demonstration cases with multi-run climate data in section 4.4. Alternatively, each run can be considered as an individual data part in addition to the aggregated data. A many-to-many relation can then be established where each multi-run grid cell is related to the other multi-run cells (each located in another data part/run) given for the same space/time as well as the corresponding grid cell in the aggregated data. With the latter setup, a feature specified in one run can then be compared to the related data in the other runs as well as the aggregated data. However, the investigation of such a many-to-many relation for scenarios that incorporate more than two data parts is subject of current work.

The feature transfer previously mentioned provides different options of how the DOI values of the related grid cells can be exchanged. An example is the transfer of the weighted sum of related DOI values, taking the associated weight values into account. Alternatively, only the maximum (weighted) DOI value can be transferred in order to preserve isolated DOI peaks. These alternatives are useful for different stage of the analysis, for instance, starting with a maximum DOI transfer in order to not "lose" features in cells with small weight values due to averaging. At a later stage, the user may then want to switch to a weighted DOI transfer in order to study the degree to which features coexist. In paper C, we propose a set of similar strategies for a visual analysis across two data parts and provide further details with respect to our interface. An example analysis of

fluid–structure interactions is described in section 4.3 and selected parts of our study of multi-run climate data are discussed in sections 4.4.1 and 4.4.2.

## 3.4 A Moment-based Scheme for Visual Analysis

The interface described above can also be employed to transfer data properties between parts of the data (compare to data transformations in the reference model [36]). Using the structural relation specified in the interface, such statistical properties can be computed on-demand via the integrated data derivation module of our framework. In this work, we study the integration of different statistics, aggregated along independent dimensions of higher dimensional multi-run data, into the visual analysis process. A scheme based on statistical moments is proposed, which provides guidelines to the multitude of opportunities during such an analysis. Multi-run data and aggregated properties are thereby related across the interface (one-to-many relation), enabling the analyst to work with both data representations simultaneously. Interesting multi-run distributions can then be selected, for instance, by brushing certain aggregated statistics.

In the analysis, one is commonly interested in data trends and outliers. The four *statistical moments* describe important characteristics of data distributions, that is central tendency (mean), variance, asymmetry (skewness) and peakedness (kurtosis) of the distribution [148]. Furthermore, extreme observations that substantially deviate from the rest of the data can be identified using measures of *outlyingness* [148]. Since such outliers can influence the traditional measures, the moments can be estimated in robust ways as well [60, 121]. This multitude of available statistics, however, also generates a "management challenge" for the analyst: Which statistical moments should be opposed in a scatterplot? Should a traditional or robust estimate be chosen? Should a data transformation such as normalization or scaling be applied to emphasize certain data characteristics?

In our work, we propose a set of *view transformations* as a structured approach to construct a multitude of informative views, based on statistical moments. These transformations can be seen as an extension to classical data transformations and support the analyst in maintaining a mental model of the currently utilized views. We propose: 1) transformations of *moment order* that increment the $k^{\text{th}}$ moment shown in a view to the $(k+1)^{\text{th}}$ moment; 2) transformations of *robustness* replace a classical estimate in a view by a more robust one (e.g., median instead of mean); 3) *relating transformations* that relate the axis in a view, for instance, by subtracting them or computing the ratio; and 4) *scale transformations* that change the scale or unit of a view axis. Relating and scale transformations are closely related to classical data transformations. They support the comparison of view attributes to each other and can be utilized to enhance the depiction of statistical properties, for instance, by using a logarithmic scale.

Transformations of order and robustness, on the other hand, support the classi-

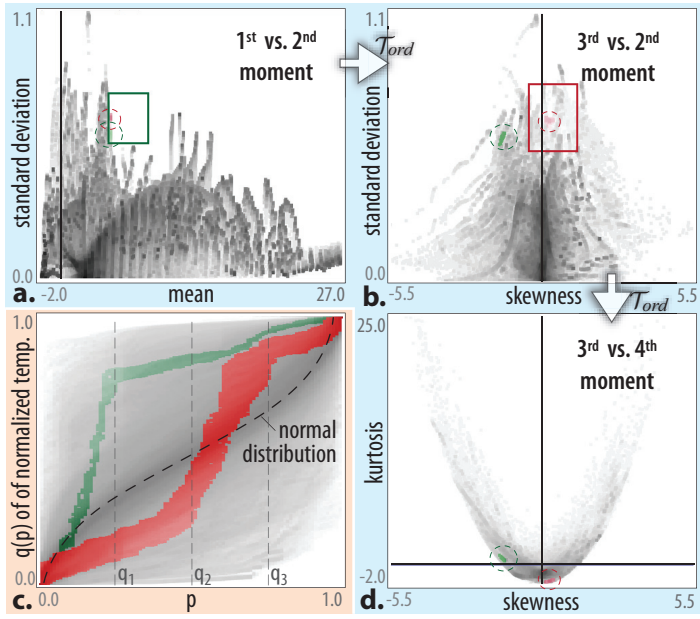



**Figure 3.5:** Basic view setup showing combinations of all four moments in the aggregated data part in (a), (b) and (d). Interesting distributions are brushed and highlighted in the quantile plot in (c).

fication of useful attribute combinations in 2D scatterplots. We present a scheme consisting of views showing: 1) the $k^{\text{th}}$ vs. $(k+1)^{\text{th}}$ moment estimated in a traditional or robust way; 2) traditional vs. robust estimates of the same moment, and 3) different robust estimates of the same moment. The different views can support different tasks such as exploration of relations between different moments or assessment of the influence of outliers by opposing traditional and/or robust estimates. Figs. 3.5a, 3.5b and 3.5d show combinations of all four moments, computed from multi-run climate data. The views result from consecutive transformations of moment order $\mathcal{T}_{ord}$. Such a setup of views supports the investigation of basic characteristics of the related distributions, where the views are arranged such that each of them have an axis in common. By replacing the traditional estimates by their robust alternatives one can, moreover, study the effect of outliers on the measures.

For depicting the individual data distributions of the multi-run data, so-called *quantile plots* and *Q–Q plots* [254] (quantile–quantile plots) are utilized. While these views are common in statistics, they are hardly known in visualization research. A traditional quantile plot depicts the sample quantiles[2] of only a few

[2] A sample quantile $q(p)$ [90] splits a distribution of values $\{x_1, \ldots, x_n\}$ such that at least $np$ of the samples are $\leq q(p)$ and at least $n(1-p)$ values are $\geq q(p)$ where $p \in [0,1]$.

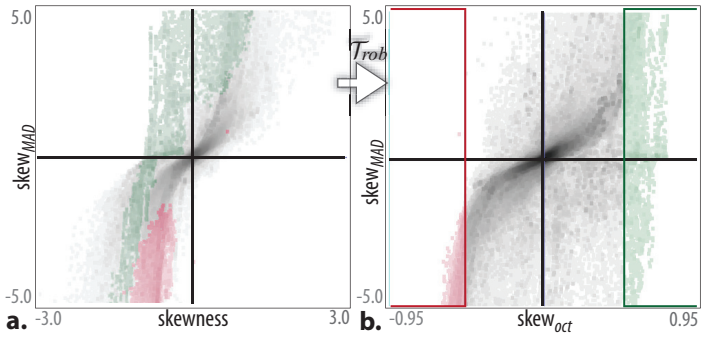



**Figure 3.6:** Comparing traditional vs. median/MAD-based skewness in (a) and two robust estimates for skewness in (b). Some of the points with positive octile-based skewness ($\mathrm{skew}_{oct}$) are selected in (b) and emphasized in green. Some of the corresponding values are even negative when using the traditional estimate (a).

distributions. In our work, however, we visualize all data distributions in the multi-run data using a focus+context style. In Fig. 3.5c, the multi-run temperature values of each location in space/time are normalized to the unit interval using a scale transformation. The resulting values are shown as a sequence of points that monotonically extends from the left to the right. Different distributions can be compared with each other (the shape of a standard normal distribution is depicted as a dashed curve). An interesting combination of mean and standard deviations has been selected in Fig. 3.5a and refined in Fig. 3.5b. Related distributions are thus emphasized in color in Fig. 3.5c. Distributions with negative skewness are highlighted in green. While most of the values of these distributions are located on the top of the figure, certain values that strongly deviate can be seen on the left in Fig. 3.5c. Applying a scale transformation on the y-axis, a measure of outlyingness can be used instead, which supports the further investigation of this feature (compare to Fig. 3c on page 131, paper D).

In our scheme, we provide two robust alternatives for each statistical moment: 1) measures that are based on octiles [87, 157] (special cases of quantiles) that aim at reducing the influence of outliers and 2) measures that utilize the median and *median absolute deviation* from the median (MAD) as robust estimates for mean and standard deviation, respectively [60]. Classical and robust estimates for the same moment can be compared in views that result from robustifying view transformations (see Fig. 3.6). These views of type $k^{\text{th}}$ vs. $k^{\text{th}}$ moment can be utilized in order to assess the influence of outliers on the measures. For normally distributed data, the points should be located along a diagonal. Deviations

The three *quartiles* that are commonly depicted in box plots [233] are examples for such quantiles: lower quartile $q_1 = q(\frac{1}{4})$, median $q_2 = q(\frac{1}{2})$, and upper quartile $q_3 = q(\frac{3}{4})$. The *median* is, moreover, a robust estimate for the mean of a distribution and the *interquartile range* (IQR $= q_3 - q_1$) is a robust measure for the standard deviation.

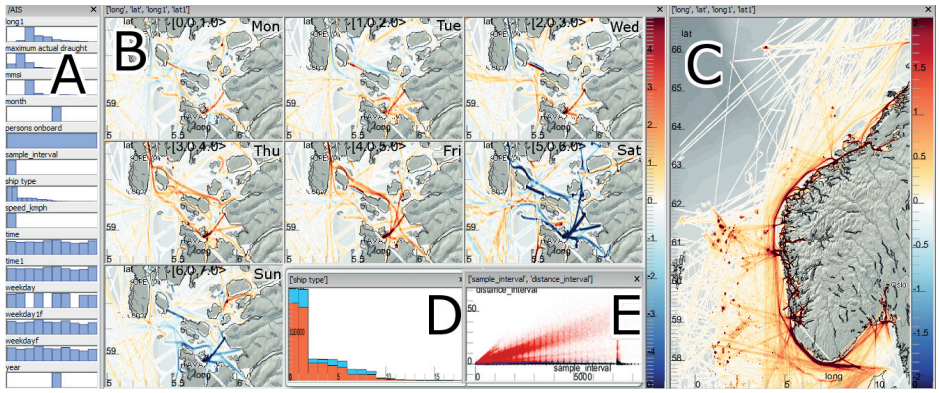



**Figure 3.7:** Visual analysis of ship movement data using interactive difference views. (a) depicts the available data variates as histograms, (b) compares the difference in ship traffic around Stavanger per weekday to the average traffic (c), a histogram of different ship types is shown in (d), and selections can be made in scatterplots (e). Using a context menu, data in a view can be further compared by depicting deviations from the average for different categories or bins (e.g., weekdays, ship type, ship speed). Image adapted from Daae Lampe et al. [47].

from this pattern can then be investigated, for instance, by applying a relating transformation that subtracts one view attribute from the other.

A number of views from the statistics literature can be constructed by the proposed set of view transformations. Examples are the standard and detrended Q–Q plot [254], plots showing skewness and kurtosis that form a so-called Fleishman system [61], as well as the spread vs. level plot [233] that opposes the logarithmic versions of median and interquartile range. We consider the fact that these views agree well with the proposed scheme as evidence that the classification is useful and appropriate. The proposed view transformations, moreover, match the iterative nature of a visual analysis (compare to the Keim's mantra [116]); views are modified step-by-step along with a mental model of the views (compare to Liu and Stasko [139]). Selected results from an application of our scheme in an analysis of multi-run climate data are given in Sec. 4.4. Further details can be found in paper D.

In recent work, we explicitly implement the concept of view transformations in an user interface that rapidly enables the construction of a series of difference views. These views show aggregated differences between movement data using a visualization that is based on kernel density estimates (KDEs). The approach supports the visual analysis of a set of hypotheses, for instance, by comparing ship traffic at different workdays (see Fig. 3.7b). The aggregated movement data for different days are then related to the overall average by showing the particular differences (compare to relating transformations). For further details see Daae Lampe et al. [47].

# Chapter 4

# Demonstration Cases

In this chapter, we demonstrate the different approaches that are described in the previous chapter. Selected results from an exploration of time-dependent climate data are presented in Sec. 4.1. Here, the main goal is the generation of promising hypotheses that are subsequently evaluated with statistical methods. The glyphs that are described in Sec. 3.2 are utilized in the analysis of an Diesel particulate filter (Sec. 4.2) as well as in the visual analysis of multi-run data (Sec. 4.4). In the latter example, multi-run data and aggregated statistics are related via our proposed interface (see Sec. 3.3), which supports the investigation of features across multiple parts of data. The same concept is used in the analysis of a fluid–structure interaction in section 4.3. Finally, our scheme for visual analysis based on statistical moments is demonstrated in section 4.4.2.

## 4.1 Exploring Climate Data for Hypotheses Generation

The goal of this study is the generation of hypotheses related to climate change, which can be subsequently evaluated using classical statistics. Our cooperation partners from climate research were interested in finding specific regions in the atmosphere that represent potential sensitive indicators for climate change. An example for such a hypothesis could be that a cooling trend over several years, located at certain pressure levels in a specific geographic region, can be considered a robust and sensitive indicator for atmospheric change. In order to generate such hypotheses, we utilize the data derivation mechanism of our analysis framework [52, 54] for temporal data abstraction.

First, certain climate variables such as temperature or geopotential height are temporally smoothed, and then *linear trends* are computed as moving differences over a certain timespan. In order to determine the significance of the derived climate signal, the corresponding *signal-to-noise ratio* (SNR) is computed as well. The temporal trends in the aspired indicator regions should exceed a certain SNR, which means that the derived climate signals substantially deviate from the natural climate variability (compare also to Ladstädter et al. [132, 133]). The hypotheses that emerge during the visual exploration process are then evaluated using statistical trend testing [131, 256].

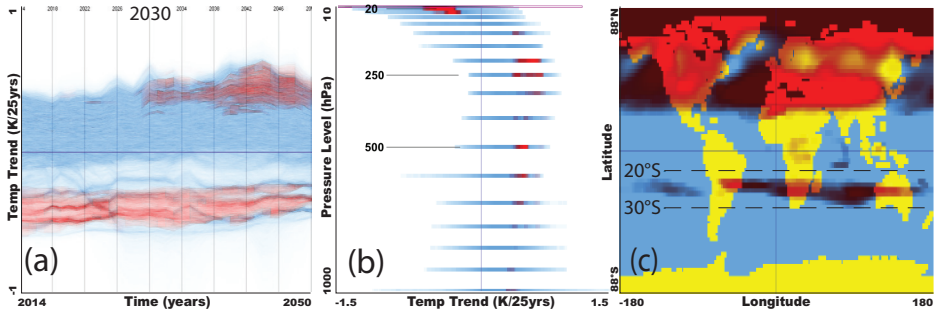



**Figure 4.1:** Exploration of the temperature field of ECHAM5. Areas with high SNR values have been selected in another view (not shown), the corresponding derived temperature trends are highlighted in a function graph view (a). The same features are emphasized in a plot showing trend vs. pressure levels (b) as well as in the geographical context (c). Figure adapted from Ladstädter et al. [133].

In the following, the temperature field of one ECHAM5 climate model simulation run[1] [193] of the B1 scenario of the Intergovernmental Panel on Climate Change (IPCC) 4$^{\text{th}}$ Assessment Report is investigated. The ECHAM5 IPCC 20$^{\text{th}}$ century run is utilized for the period before 2001. In order to reduce the influence of seasonal cycles, seasonal-mean fields of temperature (June-July-August) are used. Linear trends are computed over 25 years, together with the corresponding SNR. In the visual exploration, several linked scatterplots and a function graph view are brought up next to each other. In order to localize the sensitive indicator regions, low absolute signal-to-noise ratios are excluded from the selection (not shown here). Since we do not want a sharp discrimination between regions with high and low significance, smooth brushing is utilized [53] in our study. The uppermost pressure level (10 hPa) is, moreover, excluded since the ECHAM5 data has known deficiencies in these areas [42].

Figure 4.1a depicts the variation of the derived temperature trend over time (2014–2050) in the new function graphs view. The feature highlighting (red color) is enhanced in order to make the sensitive indicator regions more visible. Negative temperature trends show a high significance over the whole investigated time period. Positive trends become more significant around 2030. In Fig. 4.1b, sensitive regions with positive temperature trends are mainly located in the upper troposphere, while indicator regions with negative trends are located in the lower stratosphere. The indicators can be investigated in their geographic context as well (see Fig. 4.1c), where the Northern Hemisphere summer is visible in the June–August trends.

Specific hypotheses generated from these views (i.e., the identified indicator regions) are subsequently evaluated using statistical least-squares-fitting (com-

[1] Max-Planck-Institute for Meteorology (MPI-M) Hamburg, Germany.

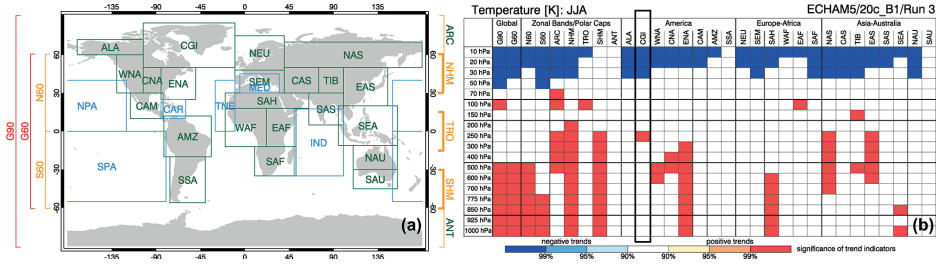



**Figure 4.2:** (a) Statistical trends are computed mainly with respect to the IPCC regions [131, 214]. Temperature trends are computed for the time period 2001–50. Regions with statistically significant positive trends are depicted in red, and significant negative trends are shown in blue. Figure taken from Ladstädter et al. [133]

pare to Lackner et al. [131]). For the Canada–Greenland–Iceland (CGI) region, for instance, the significant negative temperature trend in the topmost pressure levels (10–30 hPa) can be confirmed as well as the positive trend at the 250 hPa level (see Fig. 4.2b). Also other generated hypotheses compare well with the results from statistical analysis (compare to Ladstädter et al. [133]). In Fig. 4.1c, a band of high significance is visible in the Southern Hemisphere subtropics (20°–30°S). This apparently sensitive indicator is located between the zonal bands of the IPCC regions that are commonly utilized in the computational analysis [131, 214]—see the gap between the tropical (TRO) and Southern Hemisphere midlatitudes (SHM) band in Fig. 4.2a. In a traditional statistical analysis, this indicator region would have been "overlooked."

Further results and generated hypotheses can be found in section 4 of paper A, where a different scenario (A2) of the ECHAM5 climate simulation and ERA-40 reanalysis data[2] [212] is used. As one conclusion, we see an excellent potential of generalizing this tight integration of visual exploration and statistical analysis to other scenarios with similar characteristics.

## 4.2 Glyph-based Analysis of a Diesel Exhaust System

We use the glyphs proposed in Sec. 3.2 in a visual analysis of a Diesel particular filter that is part of a Diesel exhaust system studied by Doleisch et al. [54]. The filter traps Diesel particulates (*soot*) and burns them at high temperatures during regeneration cycles. According to Doleisch et al. [54], the domain experts want to understand whether the soot is burned completely and at which locations and how fast it is oxidized. In Fig. 4.3, different data variates are encoded in the glyph-based visualization. Flow temperatures are represented by color and the overall glyph size for redundancy purpose. The amount of soot is depicted

[2]European Centre for Medium-Range Weather Forecasts, Reading, U.K.

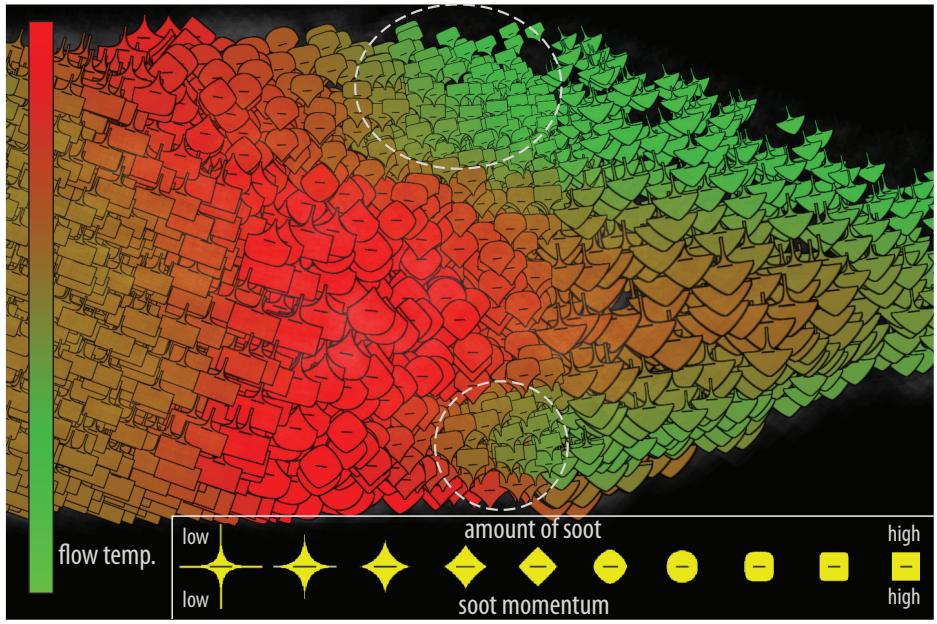



**Figure 4.3:** Soot particles are burned at high temperatures in a Diesel particular filter. Two areas are indicated by circles with a high amount of soot (upper shape) and a decelerated soot reduction (lower glyph shape), which is seemingly related to low temperatures (color) and lower $O_2$ levels (rotation).

by the upper glyph shape, and the lower shape shows the second derivative of the amount of soot (soot momentum[3]). Additionally, the amount of $O_2$, being important for the oxidation process, is depicted by the glyph rotation ($-45°$ to $45°$). Using the glyph-based visualization, the depicted data variates can be related to each other.

In figure 4.3, the oxidation progresses from left to right over time. We can see that an uneven soot oxidation is apparently related to differences in temperature. Two areas with relatively low temperature (green color, see the circles) are located to the right of the peak temperatures of the oxidation process (red). The $O_2$ fraction in these areas is relatively low (counter clockwise rotation), which affects the soot oxidation in addition to low temperatures. The amount of soot in these areas is thus relatively high (upper glyph shape). In contrast, we can see that the amount of soot in areas left to the peak temperatures is rather low (here the soot has been burned already). Using the glyph-based visualization in addition to linking and brushing supports the investigation of important data characteristics.

[3]This information is important, since we are interested in the rate of soot reduction (negative $1^{st}$ derivative) and whether the soot reduction is accelerated (negative $2^{nd}$ derivative) or slowed down (positive $2^{nd}$ derivative), compare to Hauser [74].

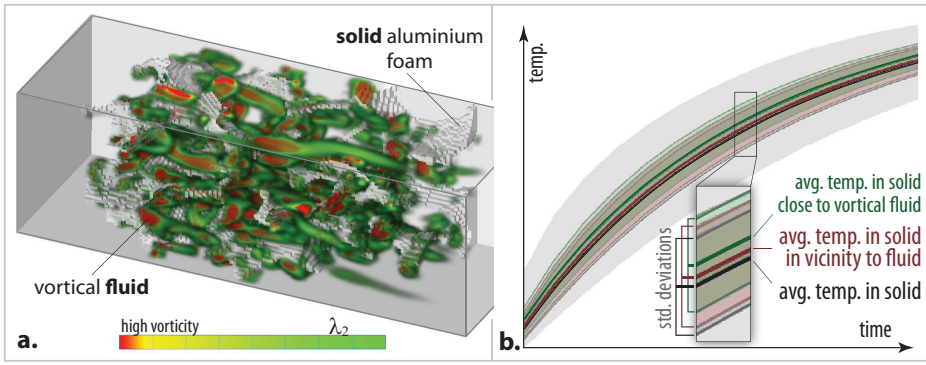



**Figure 4.4:** Visual analysis of heat transfer in an FSI scenario: (a) vortical regions within the flow volume are selected via the $\lambda_2$ criterion [101]. The feature is transferred to the solid part, where statistical properties for different selected regions are shown over time (b).

Such an approach is demonstrated in section 4.4.1, where aggregated properties of multi-run data are depicted by glyphs.

## 4.3 Visual Analysis of a Fluid–Structure Interaction

In the following, data from a multi-physics simulation of warm water flow though a cooler aluminum foam is investigated. Both domains are modeled individually in the simulation as spatially adjoining 3D volumes. Fluid and structure are thereby connected by an interface representing the physical boundary region (see Fig. 3.4a, page 35). The interface enables the exchange of properties such as heat. In the analysis, we investigate how the thermal behavior of the foam is influenced by its micro structure. We utilize the interface concept described in section 3.3. Both fluid and solid data resulting from the simulation are integrated in our visual analysis framework. A many-to-many relation is established between the grid cells located in the boundary region between the data parts (illustrated in Fig. 3.4b). Weight values are specified such that grid cells located farther apart have less influence on each other than cells located close-by.

Vortices are highly important for understanding flow characteristics such as the investigated heat exchange. Vortical flow regions are, therefore, selected using the $\lambda_2$ criterion [101]. Areas with strong vortical properties are depicted in red in Fig. 4.4a. The selection from the fluid part is instantly transferred to the foam part via the interface. In order to derive quantitative properties from the transferred feature, the solid temperature in the selected vicinity of vortical fluid has been averaged and is plotted as a green curve over time in Fig. 4.4b. Additionally, the temperature in the solid part close to the fluid is depicted as

a brown curve, and the overall average temperature in the solid is shown as a black curve (corresponding standard deviations are shown as filled areas in the background; see Unger et al. [236] for more details on this visualization of statistics computed on smoothly brushed features). As shown in Fig. 4.4b, the solid temperature close to vortical flow (green curve) is warmer than the solid in vicinity to the fluid (brown curve) as well as the average solid temperature (black curve). This is a strong indication of a direct relation between turbulent flow around the foam structure and the corresponding heating process.

## 4.4 Visual Analysis of Multi-run Climate Data

In this section, the concept of relating two parts of scientific data by an interface is exemplified in a study of multi-run climate data. Additionally, we demonstrate how a model of visual analysis that is based on statistical moments enables the structured study of different characteristics of multi-run data (Sec. 4.4.2). Data from a multi-run simulation of a palaeoclimatic cold event is investigated [10]. The anomaly was caused by a meltwater outflow from Lake Agassiz, an enormous glacial lake located in the center of North America. Due to climate warming and melting of the Laurentide Ice Sheet, the lake drained approximately 8,200 years ago. The investigated data results from a multi-run simulation of the CLIMBER-2 coupled atmosphere–ocean–biosphere model, which simulates a cooling of about 3.6 K over the North Atlantic [10].

An important goal for the climate modelers is to better understand how sensitively their simulation model reacts to variations in certain control parameters. In such a sensitivity analysis one aims at identifying the parameters with most influence, which can help to validate the model and also guide future research efforts [70]. Multiple simulation runs are, therefore, computed with varied initial parameters. In the following, we perform a visual sensitivity analysis of the ocean part of the CLIMBER-2 model based on the input parameters $diff_h$ and $diff_v$. In section 4.4.2, our moment-based scheme for visual analysis is exemplified on data from the atmosphere-part of the same CLIMBER-2 model [10].

### 4.4.1 Visual Sensitivity Analysis across an Interface

In the multi-run simulation of the ocean model, two ocean diffusivity parameters are altered, one horizontal ($diff_h$) and one vertical ($diff_v$), with ten variations each. The resulting data consist of a total of 100 runs ($10 \times 10$), where the temperature values for each run are given for 500 years on 2D sections (latitude $\times$ depth) through the Atlantic, Indian and Pacific ocean. Since the number of independent dimensions of the multi-run data is challenging, aggregated statistics with respect to the run dimensions of the data are computed. The resulting aggregated data are stored in one data part, where a one-to-many relation is

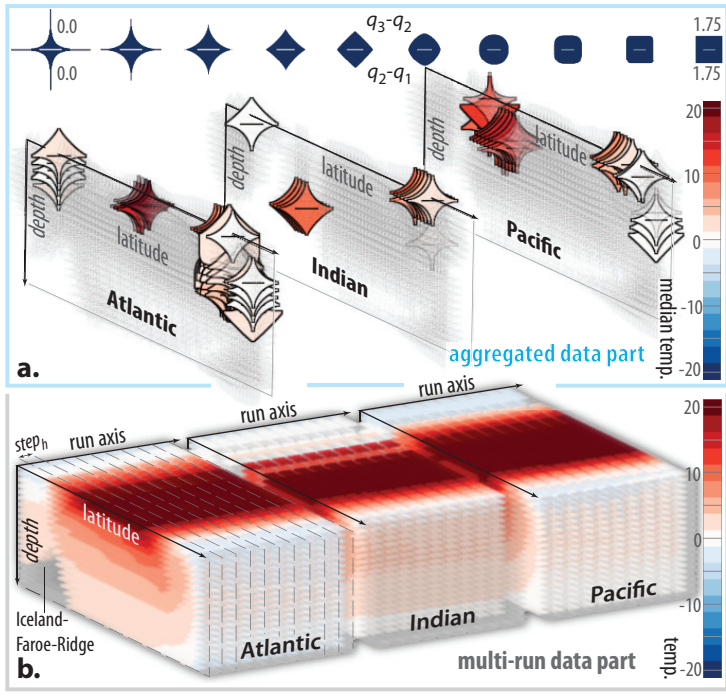



**Figure 4.5:** Aggregated and multi-run ocean data form a one-to-many relation in the interface: (a) glyph-based visualization of four aggregated properties (color, overal glyph size, upper/lower shape). The original multi-run data on 2D cross sections is shown in (b). The run parameters are encoded in one of the spatial dimensions.

established between each aggregated grid cell and the related cells in the multi-run data part (given for the same space and time). In the analysis, one can thus go back and forth between aggregated and multi-run data, using the interface to transfer features in both directions.

The glyphs proposed in section 3.2 are utilized to represent aggregated statistics in a 3D visualization. The quartile information commonly shown in box plots [154] is computed for each multi-run distribution. In Fig. 4.5a, the lower and upper glyph shape represent the distance between distribution's median $q_2$ and the lower and upper quartile ($q_1$ and $q_3$), respectively. The overall glyph size, moreover, represents the interquartile range (IQR $= q_3 - q_1$), which is a robust estimate for the standard deviation. The median temperature is depicted using a diverging color map [19]. The corresponding statistical properties can be explored in the aggregated data using the setup similar to figure 3.5 (page 38). Large interquartile ranges are brushed and opacity represents the corresponding DOI values in Fig. 4.5a. The glyphs give a qualitative overview of the related multi-run distributions. A few locations with large interquartile ranges (larger

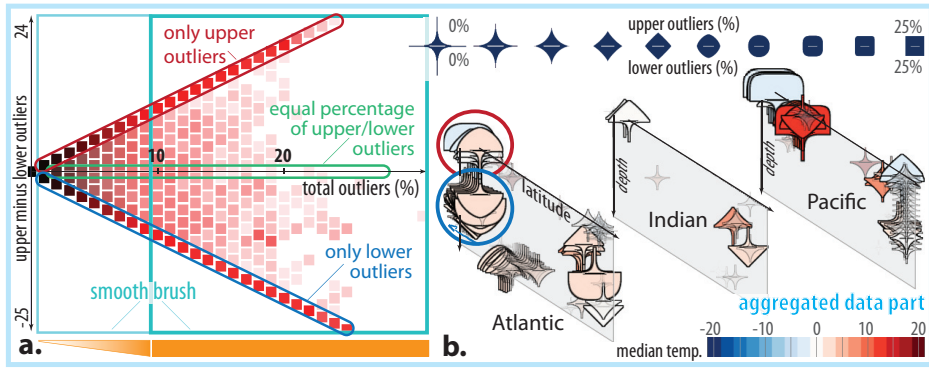



**Figure 4.6:** Analyzing distributions that contain at least 10% outliers: (a) scatterplot showing the percentage of total outliers (x-axis), and a measure to determine how the outliers are distributed (y-axis), i.e., are more located above $q_3$ (upper outliers) or below $q_1$ (lower outliers). The same outlier properties are depicted using glyphs (b).

glyphs) can be seen. The upper and lower glyph shape, moreover, provide information about the skewness of the distribution. The multi-run data for the same time step is shown in Fig. 4.5b. For each run, the cross section (latitude × depth) given for the Atlantic, Indian and Pacific Ocean are shown. The corresponding run settings are encoded in one of the spatial axes of the visualization (run axis). The camera settings in both views are synchronized during interaction.

**Visual Outlier Analysis:** As a next step, we investigate certain multi-run values that deviate significantly from the rest of the corresponding distribution in the multi-run data. Finding such outliers, together with the corresponding run settings, can help to find possible errors in the model as well as unsuitable parameter settings for the simulation. For each distribution in space/time, the total percentage of "mild" outliers is computed [233]: that is the percentage of values above $q_3 + 1.5 \times \text{IQR}$ (upper outliers) plus the percentage of values below $q_1 - 1.5 \times \text{IQR}$ (lower outliers). The scatterplot in Fig. 4.6a depicts the total percentage of outliers, computed for each multi-run distribution, on the x-axis. The y-axis shows the percentage of upper minus lower outliers and, therefore, provides information on how the outliers are distributed. This means, for example, are more outliers located above the upper quartile or below the lower quartile, are they equally distributed, etc. The number of data items per rectangle is encoded in luminance and the corresponding DOI values are shown in color (red represents the focus). Using a smooth brush [53], we focus on distributions that contain more than 10% of outliers.

The corresponding outlier characteristics can be investigated in the spatial context using our glyph-based visualization (see Fig. 4.6b). The percentage of total

outliers is encoded in the overall glyph size, and the percentage of upper and lower outliers is represented by the upper and lower glyph shape, respectively. The median temperature is encoded in color. A group of grid cells with mainly upper outliers is located in the north of the Atlantic (red ellipse). Another group of lower outliers is propagating northwards over time, and downward near the seabed (blue ellipse). This feature can be investigated over time by interactively changing the depicted time step. Here, it extends over the North pole to the other parts of the arctic sea (not shown here). This feature is further analyzed by deriving additional statistics in both aggregated and multi-run data. Computing a measures of outlyingness in the multi-run data (as outlined in Sec. 3.4) eventually supports the identification of concrete settings for $diff_h$ and $diff_v$ that produce certain outlier characteristics that are specified in the aggregated data part. The investigation of these features across the interface was beneficial for this application study. Further details and results are provided in paper C.

### 4.4.2 Moment-based Visual Analysis of Multi-run Climate Data

We demonstrate our moment-based scheme for visual analysis (described in Sec. 3.4) in another study of multi-run climate data. The investigated data stems from the atmosphere-part of the CLIMBER-2 model, which simulates a cooling over the North Atlantic [10]. A global sensitivity analysis [158] is performed in the simulation. The space of model parameter, consisting of seven parameters, is thereby sampled iteratively in order to identify the most influential paramters. The resulting multi-run data consists of 240 runs that are each given for a 3D atmosphere over 500 years. Multi-run and aggregated data are again related via an interface. The four standard moments are initially computed for each distributions of multi-run values. The resulting mean temperature, standard deviation, and skewness are represented in color in Fig. 4.7a (timestep 80). Distributions with higher standard deviations can be seen in southern latitudes together with positive skewness values. A view setup is created that shows combinations of all four moments (aggregated data) and a quantile plot.

Relations between different moments are investigated via brushing. Since there is no clear distinction between focus and context, a smooth brush [53] is utilized. Certain distributions with positive skewness and negative kurtosis are selected in Fig. 4.7b and highlighted in green in the quantile plot in Fig. 4.7c. The distributions in this plots have been normalized (scale transformation) in order to be comparable among each other. The majority of the selected distributions are bimodal, i.e., they have two local maxima (modes) as also shown in the histogram. The runs for these grid cells represent different states of the climate model. The distributions can be compared to distributions depicted in red, which are selected in a mean vs. standard deviation plot (not shown here). As a next step, the effect of outliers on the classical moments is investigated. Several views are replaced by their robust alternatives using a transformation of robustness $\mathcal{T}_{rob}$. Some of

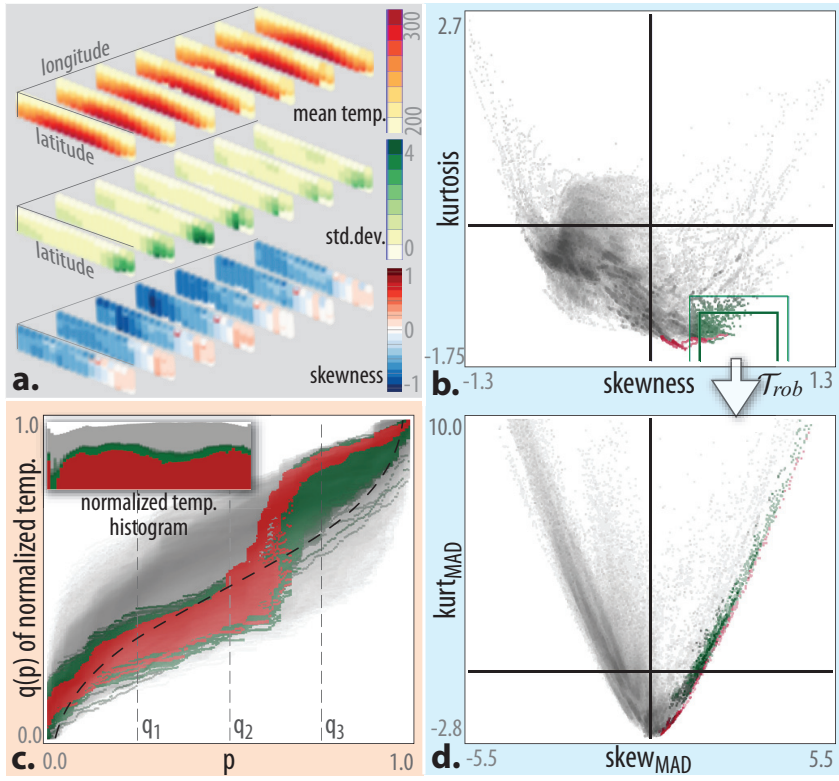



**Figure 4.7:** (a) Aggregated mean temperature, standard deviation, and skewness are shown for the 3D atmosphere of a multi-run climate simulation (timestep 80). Interesting data characteristics are brushed in (b) and the corresponding distributions are investigated in a quantile plot (c). A robust version of view (b) is shown in (d).

the highlighted points (red, green) with negative kurtosis values in Fig. 4.7b are positive when estimated in a robust way (Fig. 4.7d). Further results from this study can be found in Sec. 5 in paper D.

# Chapter 5

# Conclusions and Future Work

Multi-faceted scientific data are becoming a standard in many areas including climate research and engineering. Data are often multi-variate, time-dependent and stem from multi-modal, multi-run, and/or multi-model scenarios. Addressing the different data characteristics described above in an integrated approach is very challenging. In this work, we propose multiple extensions for a visual analysis framework that is based on feature specification via brushing in multiple linked views, focus+context visualization, and on-demand data derivation [52, 73].

The integration of *derived data* resulting from computational analysis such as statistics into the visual analysis process has shown to be beneficial in many scenarios [113, 116, 223]. When exploring time-dependent climate data, for instance, the computation of temporal trends was essential for generating promising hypotheses. It was very rewarding to see how positively our visual analysis technology was adopted in a challenging application domain such as climate research. Another example was the computation of descriptive statistics for analyzing data trends in multi-run data. This kind of data reduction enabled an effective visual analysis where the analyst could change between traditional and robust estimates of the four moments. Additional measures of outlyingness were essential for identifying outliers that substantially deviated from the results of the other runs. This increase of opportunities, however, also generated a challenge for the analyst to maintain an overview of the currently used statistics. Selected view transformations helped to categorize this multitude of informative views based on statistics and aligned well with the iterative nature of a visual analysis (compare to Keim's mantra [116]). Relating transformation can be used, for instance, to investigate deviations from expected patterns or trends stemming, for instance, from a linear regression model. We use similar view transformations for creating difference views in the visual analysis of multi-variate movement data [47].

We identify the visual analysis of data from multi-run simulations and interacting simulation models (e.g., coupled climate models or multi-physics simulations) as promising directions for future research, as well as multi-modal visualization. The proposed interface concept supports a visual analysis that incorporates multiple parts of scientific data. Features that are specified via smooth brushing [53] can be transferred across the interface in several ways, for instance,

transferring the maximum or a weighted sum of related DOI values. For the investigated cases with multi-run data and aggregated statistics, the analysis usually starts at the aggregated level (overview first) where certain data characteristics can be specified via brushing. The feature can then be refined and investigated in detail in the related multi-run data, for instance, using a quantile plot. The analysis can then go back and forth between the data parts, where features are iteratively refined. Similar patterns could be observed when analyzing data from a fluid–structure interaction simulation. Moreover, the proposed glyphs for 3D data visualization were not only useful for representing different variates, but also for analyzing aggregated data properties from multi-run data. Here, it was important to maintain the orthogonality of the different glyph properties such that the different variates can be interpreted separately.

In future work, we aim to integrate further methods from computational analysis such as additional measures of outlyingness, clustering or principal component analysis. Using visualization to understand the stages of the analysis, for instance, finding appropriate parameters or understanding the result of clustering can support a powerful visual analysis process. Especially the identification of common analysis patterns (similar to the visual analytics mantra [116] or the information seeking mantra [209]) can give guidelines for similar scenarios. Especially methods from machine learning such as neuronal networks can extract patterns and knowledge from the data, which can be used for a knowledge-assisted visualization [33].

# Part II

# Scientific Results

# Paper A

# Hypothesis Generation in Climate Research with Interactive Visual Data Exploration

Johannes Kehrer,[1] Florian Ladstädter,[2] Philipp Muigg,[3]
Helmut Doleisch,[3] Andrea Steiner,[2] and Helwig Hauser[1]

[1]Department of Informatics, University of Bergen, Norway
[2]Wegener Center for Climate and Global Change (WegCenter) and Institute for Geophysics, Astrophysics, and Meteorology (IGAM), University of Graz, Austria
[2]VRVis Research Center and SimVis GmbH, Vienna, Austria

## Abstract

One of the most prominent topics in climate research is the investigation, detection, and allocation of climate change. In this paper, we aim at identifying regions in the atmosphere (e.g., certain height layers) which can act as sensitive and robust indicators for climate change. We demonstrate how interactive visual data exploration of large amounts of multi-variate and time-dependent climate data enables the steered generation of promising hypotheses for subsequent statistical evaluation. The use of new visualization and interaction technology—in the context of a coordinated multiple views framework—allows not only to identify these promising hypotheses, but also to efficiently narrow down parameters that are required in the process of computational data analysis. Two datasets, namely an ECHAM5 climate model run and the ERA-40 reanalysis incorporating observational data, are investigated. Higher-order information such as linear trends or signal-to-noise ratio is derived and interactively explored in order to detect and explore those regions which react most sensitively to climate change. As one conclusion from this study, we identify an excellent potential for usefully generalizing our approach to other, similar application cases.

This article was published in *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1579–1586, Nov/Dec 2008. Digital Object Identifier no. 10.1109/TVCG.2008.139. The work was also presented by the main author at VisWeek 2008, Oct. 19–24, Columbus Ohio, US. An error in equation 2 has been corrected here.

# 1 Introduction

We can see that climate change has become a broadly discussed topic—politics, business, and also the general public engage with climate issues in parallel to the work of scientists. Of course, it is prediction which is the most important related aspect—but similar to weather research it is difficult to come up with deterministic results. In this study, we investigate whether we can identify particular subsets in climate data—both in time and space—that potentially represent sensitive and robust *indicators* of atmospheric climate change which possibly have strong predictive power with respect to the long-term development of our Earth's climate. We work with two representative datasets to draw our conclusions.

Improved measurement records (e.g., satellite observations) as well as extensive simulations commonly result in large, time-dependent, and multi-variate datasets which are difficult to manage. Visualization has proved to be very useful for gaining insight into such large and complex data. Three main classes of use cases or application goals can be identified [205], namely (1) visual exploration; (2) interactive visual analysis or confirmative visualization; and (3) presentation (or dissemination).

In our case, we utilize interactive visualization primarily for the early, more explorative steps (compare also to Tukey [233]). Comparable to the "discover the unexpected"$^{\text{TM}}$, as coined by Cook and Thomas [223], we aim at rapidly identifying *promising hypotheses* that afterwards are checked in an analytical, confirmative process (in our cases mostly handled by statistics). Generally, we think that it is easier for visualization to unfold its maximal utility in the context of undirected exploration (as compared to the analysis of clearly specified application questions)—and that, even though we have seen a number of cases where visualization facilitated interactive analysis very effectively [54, 134, 197].

While *computational approaches* such as statistics conveniently provide good means to accurately—and also quantitatively(!)—check specifically formulated hypotheses, it is generally quite challenging to actually derive these specific application questions. Intuition of experts—based on experiences and knowledge gained from many years—leads to promising hypotheses as well as scientific trial-and-error approaches. The emerged availability of powerful visualization technology now turns into substantial support for this important step in scientific work. Instead of cumbersomely searching within many dimensions and extensive content, we effectively shed light onto complex relations within multi-variate data by interactive visual exploration. By looking at the data (and the implicit relations within the data) and by integrating domain knowledge, the user is able to efficiently narrow down on interesting aspects of the data, which is usually achieved in an *iterative process* of repeated visualization and interaction steps. Subsequent analysis is thereby fed with well-informed hypotheses, thus resulting in a streamlined overall process with fewer large-cycle iterations.

In addition to the important step of identifying hypotheses in the first place,

it also turns out to be important to identify the right *parameter settings* and/or *boundary conditions* for the statistical analysis, especially if there are multiple parameters that influence the process. It is one characteristic of modern scientific methodology that it is now possible to vary many more parameters than ever before. While this is useful for a more varied and more detailed analysis, it also generates the significant challenge of managing all this variability. Since parameters also often influence each other, meaning that we usually cannot utilize separability to efficiently identify optimal parameters (one by one), we again welcome support as offered by interactive visualization to act in a more informed, direct way.

In this paper, we demonstrate how interactive visual exploration is used to identify certain regions in space and time which are sensitive to climate change. Even though we successfully used the here employed visualization technology in conjunction with all three types of application questions (confirmation, exploration, presentation), we focus on hypothesis generation in this paper. For analysis, the identified regions are then statistically evaluated. Visual exploration is also used to narrow down the parameter ranges that affect the computational analysis. The entire datasets can be explored at once without the need to preselect certain subsets, as this is done, e.g., in classical trend testing [131].

The remainder of this paper is organized as follows: section 2 gives a brief introduction to the here investigated questions of climate research. In section 3 the employed visualization technology is described. Several concrete details of this application are presented and discussed in Sec. 4. Finally, the paper is concluded in section 5.

## 2 Climatological Background

Climate research is concerned with the analysis of the climate system—composed of the atmosphere (compare to Fig. 1), the hydrosphere, cryosphere, lithosphere, and the biosphere—, its variability and its long-term behavior [246]. The currently most prominent topic in climate research is the investigation of *climate change*, its detection and attribution, whether naturally or anthropogenically induced.

For this purpose, we are interested in determining characteristic spatial and temporal *climate signals* which can be attributed to some cause such as, for example, anthropogenic forcing. These signals are compared with the climate noise to assess the *significance* of the findings. The signal should deviate substantially from the noise to be of use for detecting climate change.

It is not yet completely clear, which physical variable describing the state of the atmosphere is best suited as a sensible parameter for detecting climate change. Previous work mostly concentrates on the surface temperature, not at the least because of the availability of long-term records. With the advent of radiosonde

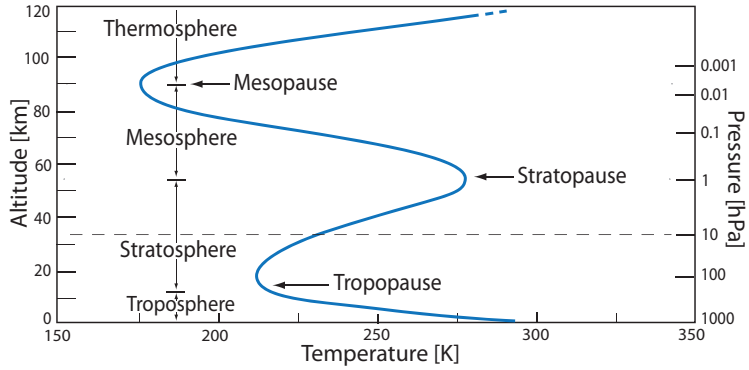



**Figure 1:** Illustration of the vertical thermal structure of the atmosphere, reflecting a balance between radiative, convective and dynamical heating and cooling processes of the surface-atmosphere system. Different layers of the standard atmosphere are shown (illustration adapted from Melbourne et al. [155]). Changes in the upper troposphere-lower stratosphere region have strong impact on the Earth's climate system [246].

and satellite-based measurements as well as global climate modeling in the last decades, data for upper air atmospheric variables are also available [214]. Key *climate parameters* such as temperature, pressure, humidity, or geopotential height can be accessed and are among the candidates to provide a sensitive indicator for atmospheric climate change [62, 109].

In the context of climate research, large multi-variate data fields are commonly investigated. Usually these fields describe the physical state of the atmosphere and can stem from various sources, such as global climate models, reanalysis data (meteorological observations assimilated into a numerical weather prediction model), or measurement records from a single instrument (e.g., satellite data). For climate models, these gridded data can easily constitute a resolution of $1.875° \times 1.875°$ in latitude and longitude, on 16 pressure levels (leading to a grid with about 300K cells), e.g., repeated on 100 time steps.[1]

When it comes to analyzing the data, it is challenging for scientists and practitioners to get a grip on these large time-dependent three-dimensional fields. The common way to gain information is to use classical *statistical methods* such as linear trend regression, multi-variate data analysis, or pattern analysis, to name only three [256]. These methods usually require prior knowledge about the data to narrow down the scope of the analysis (e.g., parameters, boundary conditions).

In this study we focus on the temperature and the geopotential height as interesting key atmospheric variables in climate research. While the temperature is easily comprehensible out of every-day experience, the geopotential height de-

[1]Note, however, that the datasets used in this study consist of 180K cells given at 108 and 42 time steps, respectively, corresponding to a horizontal resolution of 2.5°x2.5° and 18 pressure levels up to 10hPa (as indicated in Fig. 1).

serves a short elaboration: In meteorology and climatology the common measure of height is not the geometric but the geopotential height $z$, which can be seen as the geometric elevation above sea level corrected by Earth's gravitation:

$$z := 1/g_N \int_0^h g(\phi, h')dh' \tag{1}$$

where $g_N$ is the standard gravity at sea level, $\phi$ is the latitude, and $h$ is the geometric elevation. The correction is quite small (less than 1% for $h = 50$km), but using $z$ instead of $h$ is the more natural measure in the application: Using in-situ or remote-sensing measurements of the atmosphere, for example, commonly provides the temperature, pressure and humidity, but not the geometric height. Using the barometric formula (relating the pressure with the height), the geopotential height can be derived directly out of these parameters [246]. Measuring geopotential heights of constant pressure surfaces has therefore become a common approach in climate science, also because the thermal expansion raises the height of the constant pressure surfaces, providing a key parameter to detect climate change.

We consider the temperature field of one ECHAM5 climate model simulation run[2] [193] of the A2 scenario simulations for the Intergovernmental Panel on Climate Change (IPCC) $4^{\text{th}}$ Assessment Report for the time period 1961 to 2064, as well as the geopotential height field of the ERA-40 reanalysis dataset[3] [212] for the time period 1961 to 2002, respectively. Since the ECHAM5 A2 scenario simulation starts in the year 2001, it is complemented using the ECHAM5 IPCC $20^{\text{th}}$ century run before 2001. Using seasonal (northern) summer means (June-July-August) in this example provides us with data without the influence of the seasonal cycle, yielding clear climate signals.

Given this background, we investigate the following *application questions* in this study. We use visual exploration to:

- rapidly generate promising hypothesis, i.e., identify certain regions in space and time which potentially are sensitive to climate change. Thereby we can efficiently narrow down the parameters and/or boundary conditions for subsequent statistical analysis;
- assess the influence of smoothing parameters and trend time-frames on the findings;
- analyze the relations between certain interesting subsets of data in multiple dimensions.

The here employed modern visualization approach provides us with the unique ability to achieve these tasks faster, and also without the usually needed a priori knowledge about the datasets (i.e., to get support in data exploration).

[2]Max-Planck-Institute for Meteorology (MPI-M) Hamburg, Germany.
[3]European Centre for Medium-Range Weather Forecasts, Reading, U.K.

## 3 Interactive Visual Data Exploration

The interactive exploration of the climate data in this application has been carried out in a framework employing a coordinated multiple views setup [52]. The area of coordinated and multiple views has been steadily developing over the past fifteen years. A good overview is given by Roberts [191]. A comprehensive overview on visual data mining and visualization techniques with respect to climate data is given by Nocke [165].

Interactive visual analysis enables users to get into a *visual dialog* with the climate data. The procedure that is usually employed is the following: first an interactive visualization according to user input is generated. This helps the user to gain knowledge about the data, especially in the case of very large and complex datasets. This knowledge often leads to new questions and/or hypotheses, which can be explored and analyzed in more detail in an iterative process. Through interaction the previous visualization results are modified step by step to gain more knowledge and insight into the data. For this process it is crucial, that the tools supporting this knowledge gaining process must be fully interactive and flexible, allowing to query the data in many different ways, even for large datasets.

In this application study we have used and extended the SimVis framework [52]. In contrast to many of the previously published coordinated multiple views prototypes, SimVis is targeted at interactive PC-based handling of large datasets. The previous development of this technology was targeted at the analysis of 3D time-dependent flow simulation data especially in the automotive field [54], but has recently been extended to also cope with various other data types, e.g., measured 3D weather radar data.

In SimVis, multiple linked views are used to concurrently show, explore, and analyze different aspects of multi-field data. The different views that are used next to each other include 3D views of volumetric data (grids, also over time), but also several types of attribute views, e.g., 2D scatterplots and histograms. Interactive feature specification is usually performed in these attribute views. The user chooses to visually represent selected data attributes in such a view, thereby gaining insight into the selected relations within the data. Then, the interesting subsets of the data are interactively brushed directly on the screen (compare also to the XmdvTool [248]). The result of such a brushing operation is reintegrated within the data in the form of a synthetic data attribute $DOI_j \in [0, 1]$ (*degree-of-interest* (DOI), compare to Furnas [65]). This DOI attribution is used in the 3D views of the analysis setup to visually discriminate the interactively specified features from the rest of the data in a focus+context visualization style which is consistent in all (linked) views [72].

In the SimVis system, *smooth brushing* [53] (enabling fractional DOI values) as well as the logical combination of brushes for the specification of *complex features* [52] are supported. A smooth brush results in a trapezoidal DOI function

around the main region of interest in the attribute views. Brush attributes and their composition are explicitly represented in the system and can be interactively adjusted through the integration of a fully flexible derived data concept, a data calculator module with a respective graphical user interface—in this study we will benefit from this feature to derive meaningful parameters with respect to climate change. These new attributes can be derived from existing ones and thereafter are available for full investigation in all linked views. Due to the explicit representation of brush attributes as well as all view settings, analysis sessions can be saved and reapplied to other datasets through the use of a *feature definition language* [52]. This enables an easier and faster comparison of different climate simulation runs, for example.

### New Extensions to the SimVis Framework

In this study we extended the SimVis technology to also work with large climate simulation results, where especially the time-dependent behavior of different attributes is of interest.

To deal with overdrawing and visual cluttering when depicting large amounts of data we developed a *four-level focus+context* visualization [161], with the context information for orientation and also three different levels of focus in every attribute view. The different focus levels result from logical combinations of features, which are specified by the user in a hierarchical scheme based on individual selections. When several colors representing different focus levels are blended together (based on their respective smooth DOI values), it is crucial to have as little color mixing as possible (i.e., avoid the introduction of additional tints). This enables a more straightforward interpretation of the colors and the understanding of corresponding semantics and interrelations of the data. Moreover, the user is enabled to enhance the contrast of the DOI attribution in a view to place emphasis on regions with only a few important data items that otherwise are occluded by large amounts of context data. Therefore, the DOI values used in our color compositing scheme can be enhanced, i.e., $\overline{DOI}_j = DOI_j^\gamma$, where $\gamma$ can be altered by the user within $[0, 1]$. Alternatively, the maximum DOI value per screen pixel can be displayed opaquely on top, allowing to focus only on the features regardless of the relative importance with respect to the overall data.

For the improved visual analysis of the time-dependent climate data, we extended the existing framework with a *function graphs view*, where we depict a scalar function over time for each voxel/cell of a volumetric and time-dependent dataset [161]. In our scenario, this can lead to a dense visualization consisting of hundreds of thousands or even millions of function graphs, which are given at a relatively low number of time steps (e.g., 100). Using customizable transfer functions, the number of function graphs passing through each pixel is mapped to the pixel's luminance, which allows a straightforward interpretation of data trends, prominent (visual) structures within the data, and outliers [102, 172]. We use


**Figure 2:** Interactive visual exploration of climate data: Meaningful climate parameters are derived from the original data which are explored interactively in order to form hypotheses. Statistical analysis confirms or rejects the hypotheses. The results from analysis are generally visualized for illustration. In this pipeline each step can also reflect back on previous steps for efficient information drill down.

data aggregation (*frequency binmaps* [172] which have been extended to incorporate also DOI information) and image space methods to retain the responsiveness even when interacting with such large datasets.

Enhanced brushing techniques were integrated in order to cope with the temporal nature of the data. Time series are classified according to their *similarity* to a user-defined pattern, which can be directly sketched as a polyline by specifying an arbitrary number of control points. Several measurements were incorporated to quantify similarity, including the sum of absolute differences between the gradients (first derivative estimated as forward or central differences) of the function graphs and the target function. The aggregation of differences per time series is then compared to one threshold (for binary classification) or alternatively two thresholds (again with a smooth transition area between focus and context) to obtain fuzzy DOI values.

## 4 Exploring The Two Climate Datasets

In this section, we demonstrate the interactive visual data exploration in the field of climate research. We use the extended SimVis framework to deal with the application questions as introduced in Sec. 2. Our main goal is to rapidly identify promising hypotheses, i.e., certain regions in the atmosphere which are potentially robust indicators for climate change. The emerged hypotheses are then further investigated using statistical analysis [131], and we are able to present some preliminary results already here.

The respective process is illustrated in Fig. 2. Since it is rather difficult to identify the regions sensitive to climate change within the original data, we first derive meaningful parameters. In our case *linear trends* are calculated on smoothed data as moving differences over $N$ years, and the corresponding *signal-to-noise ratios* (SNR) are derived to determine the significance of the respective trends. The computation of these parameters is detailed in Sec. 4.2, and can be performed

and altered directly within SimVis.[4] The sensitive areas in space and time for which the anticipated signal emerges out of the climate noise background can be selected and visualized in all available attributes and views.

In an interactive visual exploration process the promising hypotheses can then be rapidly identified (e.g., certain height/pressure layers given at certain latitudes over a certain timespan). The hypotheses can then be confirmed or rejected using classical *least-squares-fitting* of a linear trend over a fixed timespan and pre-defined geographical region [131]. The results from statistics can be further explored and illustrated using confirmative visualization. The parameters affecting each step in our scenario (e.g., the timespan over which the linear trend is computed, the parameters affecting the visualization, or the boundary conditions for the statistical analysis) can be altered and narrowed down efficiently in this process. This leads to more insight and deep information drill-down.

## 4.1 Hypothesis Generation

In order to quickly come up with new hypotheses, which are otherwise difficult to generate, we first have to consider the features which characterize those atmospheric regions in space and time, which are supposed to be sensitive to climate change. These can be determined by a high absolute SNR, where the derived climate signal (i.e., linear trend) exceeds the natural climate variability. In the following, the temperature field of an ECHAM5 climate model run (A2 scenario), and the ERA-40 geopotential height field will be explored.

The ability to browse the whole field without prior knowledge of its characteristics (as usually required when using computational analysis) is advantageous here. By exploring the data as well as derived attributes with interactive visualization, possible field deficiencies (for example common in certain latitude regions for some reanalysis data) can be efficiently detected and consequently taken into account. Without knowing in advance what the expectations in the data are, interesting features or patterns can be found by browsing interactively through the field. The findings narrow down the scope for a later, more specialized treatment using statistical tools, which then are applied to gain quantitative results.

### ECHAM5 climate model run

We examine the temperature field in an ECHAM5 climate model run, where the derived parameters are computed based on a 25 year moving timeframe ($N = 25$). In Fig. 3a the SNR values of the derived linear temperature trends (y-axis) over the time domain from 1973 to 2052 (x-axis) are shown in a scatterplot. We are interested in regions where the derived climate signal has a high significance (i.e., high absolute SNR values), however, there is no sharp boundary which

[4] The derived data, for instance, for the ECHAM5 climate model results in a 2.38 GB dataset, which can be interactively explored and also saved to and loaded from the hard disk.

separates data of significance (focus) from the context. So we take advantage of the smooth brushing [53] capability of SimVis assigning fuzzy degree-of-interest (DOI) values. Using a smooth NOT-brush (violet rectangle in Fig. 3a) we exclude the data elements with a relatively low SNR from our selection, i.e., a DOI of 0 (context) is assigned to SNR values within $[-0.75, 0.75]$, a DOI value of 1 (focus) where $|SNR| \geq 1.25$, and a DOI from $]0, 1[$ to SNR values from the transition between focus and context (see the illustration on the left of Fig. 3a).

As a next step we investigate the corresponding feature with respect to the height. The 2D scatterplot in Fig. 3b shows derived temperature trend values (x-axis) with respect to pressure levels (y-axis). In the visualization, the averaged DOI values (with respect to the number of data points) are accumulated and highlighted in red according to the DOI. We can see a high significance (represented as pure red) in the topmost layers of the simulation, which may be an indicator region (see inset Fig. 3g). However, according to the literature the ECHAM5 data has known deficiencies in its highest pressure levels [42]. Therefore, we completely exclude the highest 10 hPa level and partly exclude the 20 hPa layer using a smooth NOT-brush[5] (shown in Fig. 3b, also in the magnification above Fig. 3g). A negative temperature trend with high significance is still highlighted in the remaining highest pressure levels (indicated by a blue ellipse in Figs. 3b and 3c). This cooling trend located in the lower stratosphere is supposed to be of high significance with respect to climate change (and thus part of one here generated hypothesis).

We also investigate regions with only few important data points (i.e., possibly weaker indicators). Therefore, the maximum instead of the average of the DOI values are shown in Fig. 3c. Here, a positive (warming) temperature trend is highlighted in most pressure levels of the troposphere (orange ellipse). Since this feature is barely visible in Fig. 3b it is supposed to be a less robust indicator for climate change compared to the prominent cooling trend in the lower stratosphere (blue ellipse). In figure 3c also the tropopause is visible.[6]

Figures 3d and 3e show the variation of the derived temperature trend over time (1973–2052) in the new function graphs view. The DOI values are enhanced in Fig. 3e in order to make the features more visible. The main part of the positive trend curves rises slightly (see the large amount of blue curves close to the zero line, indicated by arrow 1) and is mainly located in the troposphere. Note that only those parts of the curves in Fig. 3e (arrow 3) are highlighted where the respective SNR at the corresponding time step is relatively high. The emphasized warming trend is supposed to be a less robust climate change indicator since it is only visible when the feature representation is enhanced. On the other hand, one

[5] As a result, high negative SNR values in the lower part of Fig. 3a no longer belong to the overall feature and are therefore depicted in blue.

[6] The tropopause is the boundary between the troposphere and the stratosphere. It is higher in the tropics (up to about 17 km) and lower at the poles (up to about 8 km), which is also visible in Fig. 3c.

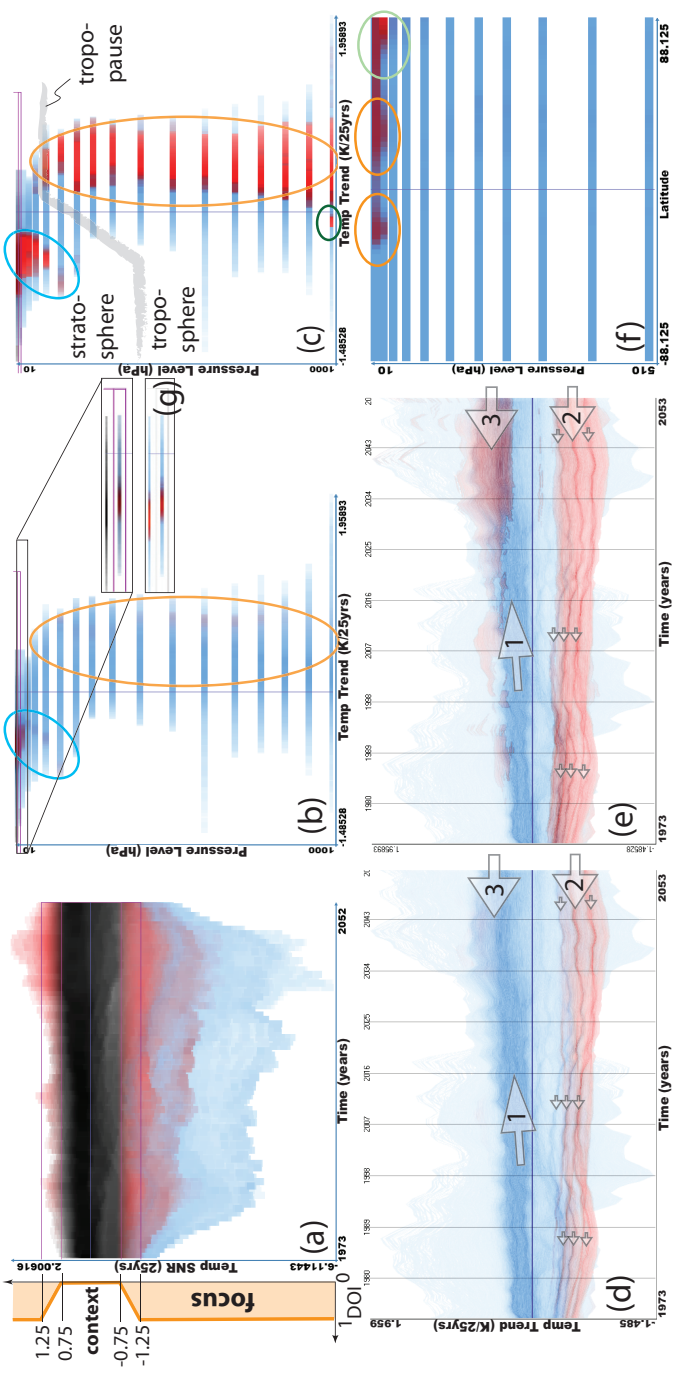



**Figure 3:** Hypothesis generation using interactive visual exploration of derived temperature parameters in the ECHAM5 climate model. Features selected in multiple linked view are highlighted in red (focus), features only selected in the current view (2nd level focus) depicted in blue, and context information in black (more details in the text).

can see that the *negative temperature trend* is very prominent and robust over the whole visible time period (arrow 2)—three traces of curves emerge visually[7] (indicated also by the small arrows). We come back to this later in Sec. 4.3. Therefore the cooling trend stemming from the lower stratosphere is supposed to be a more robust indicator for climate change considering the whole investigated timespan.

An overview of the spatial location of the sensitive regions with high absolute SNR values is given in Fig. 3f showing a latitude (x-axis) versus pressure (y-axis) scatterplot. Two highlighted areas (indicated by orange ellipses) are centered horizontally around the tropical region in the remaining high pressure levels—this feature is discussed in more detail in Sec. 4.3. Another sensitive region is visible in the northern high latitudes in the lower stratosphere (green ellipse). Brushing this region, one can identify the corresponding feature belonging mainly to the negative (cooling) temperature trend (indicated by a blue ellipse) in Figs. 3b and 3c, respectively.

**Generated hypothesis:** The above described visual exploration process lead to the following hypothesis: A promising and robust indicator region with respect to climate change is seemingly located in the lower stratosphere (upper pressure levels in the ECHAM5 temperature field), geographically located in the northern latitudes as well as in the tropics. The corresponding cooling trend is considered to be a robust indicator over the whole investigated timespan. On the other hand, the observed positive trend in the troposphere can be considered less prominent according to visual exploration (some preliminary results from the statistical evaluation are given at the end of this section).

### ERA-40 Reanalysis Data

In our study, we also examine the geopotential height field of the ERA-40 reanalysis dataset [212] for the time period 1961 to 2002 where the derived parameters are based on a 15 year moving timeframe ($N = 15$). As done with ECHAM5, low absolute SNR values are excluded in the 2D scatterplot in Fig. 4a using a smooth NOT-brush (violet color). When examining the evolution of the derived geopotential height trend over time in a function graphs view, high variations in the early years can be observed (see Fig. 4b). According to the literature [237], this is supposed to be a spurious feature. Thus, we restrict our selection to the post-1979 era, where also satellite data were assimilated.

As shown in the function graphs views in Figs. 4b and 4c, the main portion of the geopotential height trend is centered around the zeroline. We want to focus

[7] Brushing one of these traces reveals that each trace corresponds to one specific pressure level in the stratosphere (the lower one to the 10 hPa, the middle one to the 20 hPa, and the upper one to the 30 hPa pressure layer). This feature is an artifact resulting from the resolution of the simulation grid, since the ECHAM5 dataset is computed on discrete pressure levels.

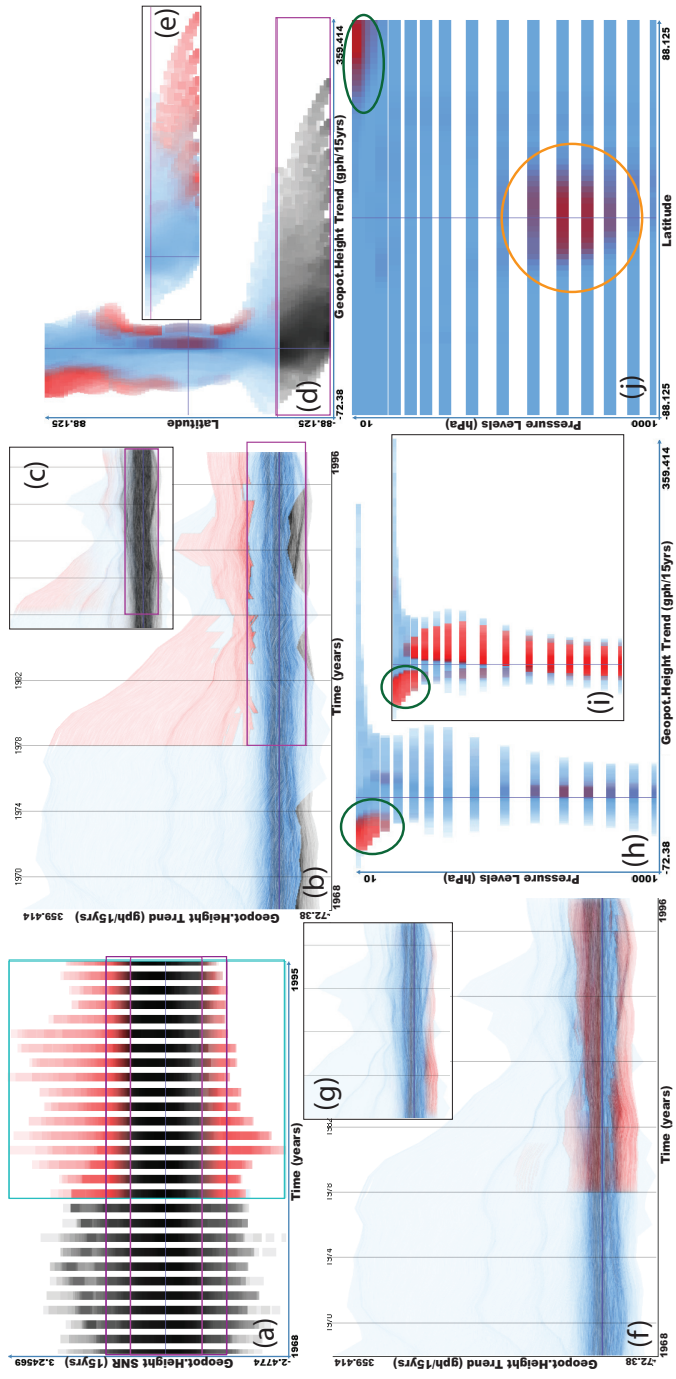



**Figure 4:** Hypothesis generation on derived trend in geopotential height fields (ERA-40 reanalysis dataset). (a) high SNR values over time (1968–1995) are brushed in a 2D scatterplot. The selection is restricted to the post-1979 era, where satellite measurements were incorporated. (b, c) similarity-based brushing of function graphs, which have a high variation, features are enhanced in (b). The resulting feature appears only in southern latitudes (e), which might be a spurious feature. These regions are therefore excluded from the selection in (d). (f, g) function graphs after 1979 having a high SNR are highlighted in red; features are additionally enhanced in (f). (h, i) geopotential height trends (x-axis) vs. pressure levels. A prominent feature is indicated by the green ellipse. features enhanced in (i). (j) sensitive regions with respect to climate change are highlighted in the scatterplot showing latitudes (x-axis) vs. pressure levels (y-axis). Here, two separable areas can be investigated (indicated by ellipses).

on the outliers, which diverge from the observable main data trend. Thus, we use a similarity-based NOT-brush (the violet brush located around the zeroline) in order to select curves with high variations—the resulting feature is highlighted in blue and red in Figs. 4b and 4c. Here, the red curves belong also to the high absolute SNR and post-1979 feature specified in the 2D scatterplot, while the blue curves ($2^{nd}$ level focus) are only selected in the function graphs view by the similarity-based NOT-brush. The visual prominence of the features is moreover enhanced in Fig. 4b in order to allow the user to focus on all regions containing features (i.e., low $\gamma$ value for the DOI enhancement). In order to show the actual significance of the feature it is depicted without enhancement in Fig. 4c.

The selection corresponding to the similarity NOT-brush is examined in a scatterplot showing derived geopotential height trends (x-axis) vs. latitude (y-axis). The highlighted feature shows that the high trend variations brushed in the function graphs view is only prominent in southern latitudes, which seems to be a spurious feature (see Fig. 4e). According to Santer et al. [201] the ERA-40 dataset contains deficiencies in these regions. Therefore, we exclude the latitudes 60°S–90°S from the selection. The result is shown in Fig. 4d highlighting high absolute SNR selections in the post-1979 era.

The variation of the geopotential height trend over time is visually examined in the function graphs view, highlighting the same features in red (post-1979 era, high absolute SNR selection, excluding southern latitudes). In Fig. 4f the features are visually enhanced in order to examine all areas containing brushed data items. One can see that the highlighted regions are vertically centered around the zeroline. On the other hand, the features are depicted without enhancement in Fig. 4g in order to focus on the prominence of the features. Since only the negative trend curves are enhanced, these are supposed to be more significant with respect to climate change than the positive trends.

**Generated hypothesis:** The features (high SNR, post-1979 era, excluding southern latitudes) are highlighted in red in the scatterplot in Fig. 4j, showing latitudes (x-axis) vs. pressure levels (y-axis). Here two structures are very prominent (indicated by two ellipses) and are supposed to be the promising indicators for climate change (and thus part of the here generated hypothesis). The one sensitive region is located in the upper pressure levels and is prominent in northern latitudes (see green ellipse). This feature corresponds to the negative geopotential height trend indicated by a green ellipse in Figs. 4h and 4i. The other sensitive region can be examined in the tropical region in medium pressure levels centered around the 700 hPa level (see orange ellipse). Since the geopotential height has different properties as the temperature also the sensitive regions are differently located. While the promising indicators are mainly located in the uppermost pressure levels of the ECHAM5 temperature field, for the ERA-40 geopotential height field they appear also in the lower to middle troposphere.

### Preliminary Results from Statistical Analysis

The hypotheses which were generated during interactive visual exploration are subject to statistical analysis. The employed *least-squares-fitting* method [131] expects the timespan over which the curves are fitted, and the corresponding latitude range as prerequisites. Linear trends are calculated over the investigated timespan and region. The statistical significance of a trend is determined by the *Students t-test* and the *goodness-of-fit measure*, which is given by the coefficient of determination $R^2$ (compare to Wilks [256]). We define the trend significance and the goodness-of-fit as the quantitative criteria for assessing the sensitivity and robustness of the explored parameter (for further details on the method see Lackner et al. [131]). Since this paper focuses on hypothesis generation, we only give some preliminary results from this analysis. A detailed computational analysis is, however, subject of future work.

For the ECHAM5 dataset, for instance, the high significance for the highlighted features in the lower stratosphere could be confirmed applying the statistical analysis to the higher northern latitude region of 60°N–90°N at the 20hPa–30hPa pressure levels (see the prominent features in the scatterplots in Figs. 5a and 5b showing temperature trends (y-axis) vs. latitudes (x-axis), features in Fig.5b are enhanced). When evaluating the hypothesis generated for the geopotential height field the ERA-40 reanalysis dataset we also got similar results.

On the other hand, the southern latitudes 25°S–90°S over the timespan 2025–2050 were also evaluated. According to the explorative visualization, these areas had a relatively low significance—see the less prominent features in Fig. 5a. However, according to the statistics the same areas returned a strong significance for the chosen timespan stemming mainly from 25°S–45°S. Therefore, the features in this latitude region were again examined using SimVis, but now displaying the maximum DOI values in order to focus on all areas containing features (see Fig. 5b). Still, only small areas with low prominence could be found, even though we already get a slightly improved agreement. Getting back to statistics, we varied the timespans for the least-squares-fit method, i.e., 2020–2045 and 2015–2040, respectively. With these modified parameters also the statistical analysis returned a noticeable lower significance for the respective latitude range, which shows that the least-squares-fit reacts very sensitively to the chosen timerange (the coupling of visualization and statistical analysis was crucial to identify this relation).

Using this iterative approach between visual exploration and computational analysis, we could benefit from the strengths of both domains: Finding the right parameters for statistics is usually cumbersome, however, using interactive visual explorations these parameter ranges could be efficiently narrowed down in an iterative process. Moreover, we could investigate that the applied statistical method reacts more sensitive with respect to the chosen timespan than expected.

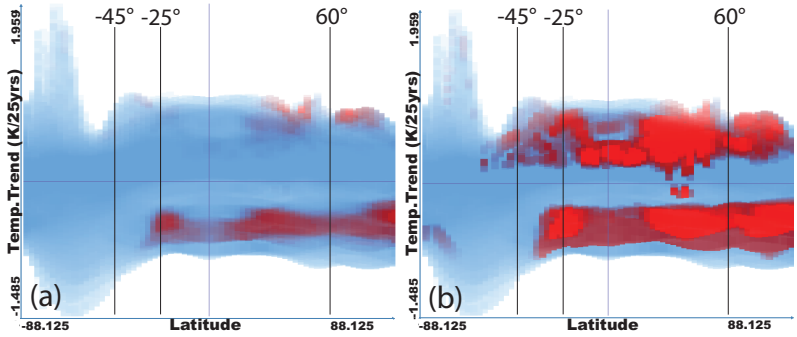



**Figure 5:** ECHAM5: Sensitive regions with respect to climate change highlighted in the scatterplot (latitude on x, temperature trend on y-axis) were handed over to statistics for further analysis. In (a) the averaged DOI attribution are depicted in order to visualize the importance of each feature. On the other hand, the visual representation of the features is enhanced in (b), showing the maximum DOI values.

These examples show how the application of visual exploration techniques—used in an iterative process—contributed to an improved workflow in this application.

## 4.2 Parameter Optimization

As illustrated in Fig. 2 there are several parameters involved in the exploration scenario in this study. It is often challenging to come up with the optimal settings, affecting the respective exploration steps in the pipeline. For example, we derive climate parameters (linear trend, SNR) from the original data in order to form our hypotheses. Thereby, the timeframe over which these calculations are performed significantly affects the derived data, and therefore also influences the following steps in the pipeline. Using interactive visual exploration we can assess the sensitivity of our results to the timeframe. To this end, we have derived the parameters over 10 and 25 years for ECHAM5 and over 10 and 15 years for ERA-40. On the example of ECHAM5, we briefly show how SimVis was used to come up with parameters that then were suitable for our analysis.

In order to be able to calculate meaningful linear trends, the original data is smoothed first using a moving average over a timespan of $N$ years. Then, the *linear trend* of a year $i$ is calculated as a moving difference between the smoothed data $\tilde{y}$, i.e., $trend_i = \frac{1}{N}(\tilde{y}_{i+N/2} - \tilde{y}_{i-N/2})$. The *linear trend fit curve* for each time frame over $N+1$ years is calculated using the derived trend values as a slope, i.e., $fit_{ij} = \tilde{y}_{i-N/2} + [j - (i - N/2)]trend_i$, where $j$ runs from $i - N/2$ to $i + N/2$. As a next step, the fitted trend curve is removed from the original data $y$ to obtain

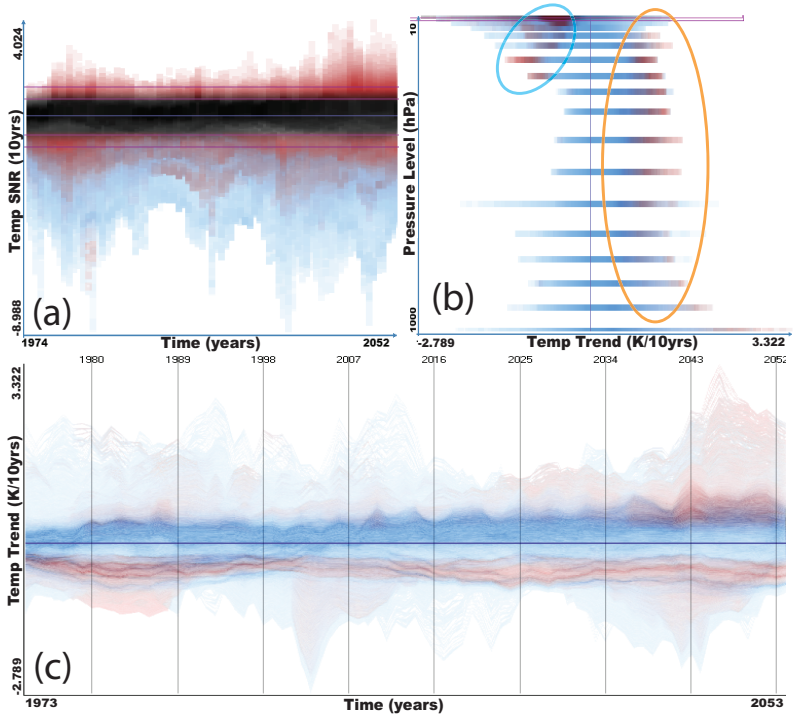



**Figure 6:** ECHAM5 temperature: derived parameters computed over 10 years instead of 25 years. The features which were barely visible with 25 years (Fig. 3b) are now highlighted in (b). The function plots of the derived temperature trend seem to contain a lot of noise.

the detrended standard deviation $s$ for the current timeframe, determining the natural variability of the climate data:

$$s_i = \left[\frac{1}{N-1}\sum_{j=i-N/2}^{i+N/2}(y_j - fit_{ij})^2\right]^{\frac{1}{2}} \tag{2}$$

Finally, the *signal-to-noise ratio* is computed as the ratio of the trend to the standard deviation, i.e., $SNR_i = \frac{trend_i}{s_i}$ (compare to Ladstädter et al. [132]).

The resulting parameters are explored using SimVis, in a similar setting as described in Sec. 4.1. When the ECHAM5 data is smoothed over a shorter time frame (10 instead of 25 years) there are obviously more high-frequency features present in the data, which can also be observed in Fig. 6a showing SNR values (y-axis) over time (x-axis). Comparing Fig. 6b and Fig. 3b shows that averaging over less data points leads to less pronounced formation of features. For the long-term trend in which we are interested, a longer timeframe is clearly favorable,

since the high-frequency characteristics are effectively flattened out and do not show up in the visual exploration.

When examining the linear temperature trends using a function graphs view one gets a high response in the upper and lower trend values (10-years), which also seem to contain a lot of noise (see Fig. 6c). Here, no clear highlighted trends can be identified in the visualization, in contrast to Figs. 3d and 3e, arrow 2. Using 25 years we obtain clearer signals and thus better-defined features. Accordingly, we used 25 years instead of 10 years in the ECHAM5 dataset, and 15 instead of 10 years in the ERA-40 dataset, respectively.

## 4.3 Analyzing Relations Between Selections

Up to now we were performing our investigation mainly in one direction, e.g., brushing high absolute SNR values and examining the resulting feature in other dimensions. In science, this principle is known as implication ($a \rightarrow b$). In the following, we want to check whether this interrelation also exists in the opposite direction, i.e., whether we we get a similar feature in one dimension when specifying a feature in another dimension ($a \leftarrow b$). If this interrelation can be confirmed the respective statement is stronger ($a \leftrightarrow b$).

When examining the derived temperature trends in the function graphs view (ECHAM5, 25 years, see Sec. 4.1), one can visually identify three streams of curves, which were very prominent in the visualization and also seemed to belong to the high absolute SNR feature (highlighted in red in Figs. 3d and 3e, indicated by small arrows). Using similarity-based brushing we can examine the interrelations between these visible trends and the other dimensions. In Fig. 7a such a brush is specified, aiming to approximate the visible structure of the respective curves. Here, similarity is evaluated based on the gradients of the function graphs and the target function. Three families of curves are emphasized in red and blue within the function graphs view (context data depicted in black). The bottom family of enhanced curves stems from the uppermost pressure level, which has been excluded, and is therefore colored in blue (second level feature).

Examining the resulting feature in a scatterplot (SNR over time, see Fig. 7b), one can see that the highlighted curves have a relatively high (negative) signal-to-noise ratio—note, that the high SNR feature is disabled in the scatterplot. The similarity feature is highlighted in another 2D scatterplot (see Fig. 7c), where it is approximately horizontally centered around the zeroline (the tropical region), and located in the uppermost pressure levels. A similar feature can be examined in Fig. 3f—indicated by orange ellipses—when going into the opposite direction (i.e., selecting high absolute SNR values in a scatterplot). However, in the previous examination these two highlighed spots were not very dominant—they were occluded by other highlighted areas in the upper pressure levels, where the most prominent feature was in the high northern latitudes. Due to the use of similarity based brushing, the areas in the tropics containing these families

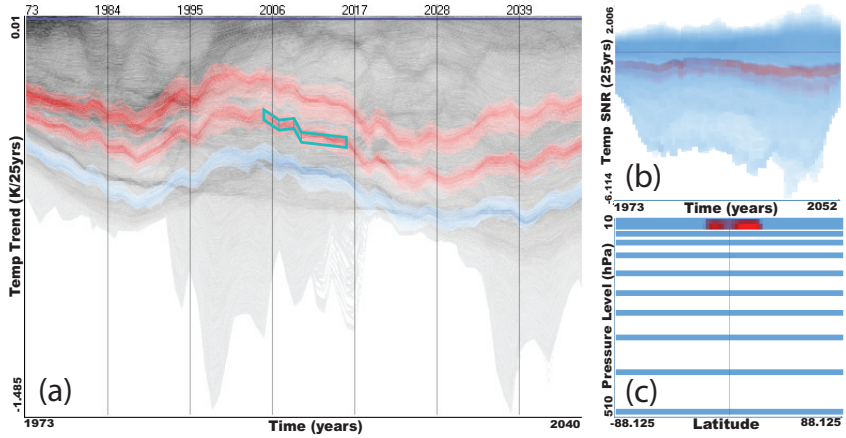



**Figure 7:** A prominent visual structure in the function plots view is brushed based on it's similarity to a user defined target function (a). Three families of curves are thereby highlighted. The respective feature contains a relatively high signal-to-noise ratio highlighted in (b), and can be located in the upper pressure levels, centered around the tropical region (c).

of similar curves could be located. Since this relation seemingly exists in both directions ($a \leftrightarrow b$) the corresponding statement is supposed to be stronger and can be considered for further investigation (e.g., using statistics).

## 4.4 Further Results

When analyzing the ECHAM5 dataset (25 years) in the 2D scatterplot, a negative (cold) temperature trend feature (considering high absolute SNR values) visually emerged in the pressure level closest to the surface (indicated by a green ellipse in Fig. 3c). This feature varies from the more prominent warming trend features with high SNR also located in this pressure level. Brushing the area (green ellipse) with a rectangular brush reveals that this feature corresponds to relatively low SNR values in the timespan 2022 to 2052 (see Fig. 8a). When looking at the geographic location, one can identify that the brushed feature corresponds to a certain area which is mainly located at the Tibetan Plateau (see Fig. 8b, where also a land-sea coloring is incorporated). According to the process illustrated in Fig. 2, the next step would be to use statistical analysis in order to evaluate whether this geographical region has a special characteristic—this is subject of future work. However, using classical statistical analysis, it would have been very challenging to identify this region in the spatial context. Also when using a binary classification scheme instead of smooth brushing (e.g., with a hard selection of $|SNR| \geq 1$), this feature would have been challenging to detect.

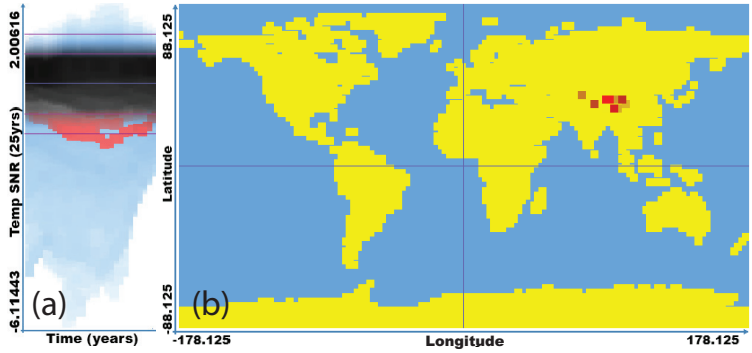



**Figure 8:** Cooling trend brushed in the lowest pressure level indicated by a green ellipse in Fig. 3c shows relatively low SNR values (a) and corresponds to a certain geographic area also including the Tibetan Plateau.

### 4.5 Performance Issues

The presented study was carried out on a system consisting of the following components: The hardware used was a modern PC-based system (Intel Core2 Quad CPU, 4 GB RAM, 320 GB harddisk, 64bit Windows) with a NVIDIA GeForce 8800 graphics card. The SimVis software is written in C++, using OpenGL and Cg shader language.

The two datasets investigated during this case study consist of 180K cells, defined at 42 time steps (ERA-40) and 108 time steps (ECHAM5), respectively. The derived data of ECHAM5 resulted in approximately 2.3 GB of data, for example. Due to algorithmic optimizations and an effective data handling framework, we are able to handle analysis sessions with multiple linked views at interactive framerates. By the use of binning techniques, large amounts of function plots can be depicted and analyzed, while still providing full interactivity. To the best of our knowledge no other comparable system can handle such large amounts of function graphs interactively on a PC.

## 5 Conclusion and Future Work

The generation of hypotheses in climate research is a crucial task. In this paper, we demonstrate the useful integration of state-of-the-art interactive visual exploration technology into the hypothesis generation process in climate research. The goal was to investigate atmospheric regions in space and time that are sensitive with respect to climate change. In order to rapidly come up with promising hypotheses, we explored derived parameter spaces using interactive visual exploration of complex features specified in multiple, linked attribute views. For analysis, the emerged hypotheses were handed over to statistical analysis. Up

to now, the results from visual exploration could already be confirmed in some exemplary cases. We also applied visual exploration in individual cases where the correlation could not be established. Here, our visual exploration framework showed to be especially useful to further investigate these cases, and to improve the understanding of the influence of different parameters on computational analysis. The power of this approach is that no prior knowledge about the data is needed to rapidly formulate hypotheses. Therefore, parameter ranges affecting for instance the computational analysis can be narrowed down efficiently.

Lessons learned from this case study are that interactive visual exploration with the opportunity to interactively drill down into certain aspects of the data (through brushing) substantially supports the exploration and analysis process of climate researchers in many ways. Using interactive visual exploration allowed us to examine the whole field without knowing its characteristics in advance, which showed to be very useful. Interesting features or patterns can be found by browsing interactively through the field. The findings narrow down the scope for a later, more specialized treatment using statistical tools, which then are applied to gain quantitative results. For visualization research it is very rewarding to see how positively new technology is adopted in a challenging application domain. Generally, we see great potential for visualization when performing undirected exploration since it efficiently complements computational analysis (e.g., statistics). We think that the approach presented here of using visual exploration to come up with promising hypotheses and then quantitatively evaluating the results can be generalized to several other scenarios.

In future work we will focus on further fusing statistical methods yielding quantitative results in our visual exploration framework. We also want to perform a detailed quantitative evaluation of the results gained from this study using computational analysis. Here again, we want to show how visual exploration and statistics can interact in a feedback loop to gain in depth insight into the data.

## Acknowledgments

The authors want to thank G. Kirchengast for important discussions, contributions and the supervision of F. Ladstädter, moreover, we thank B.C. Lackner for her help and results from the statistical analysis. The datasets are courtesy of the Max-Planck-Institute for Meteorology, Hamburg, Germany and the European Centre for Medium-Range Weather Forecasts, Reading, UK. This work was supported in part by the Austrian Science Fund (FWF) Project INDICATE P18733-N10.

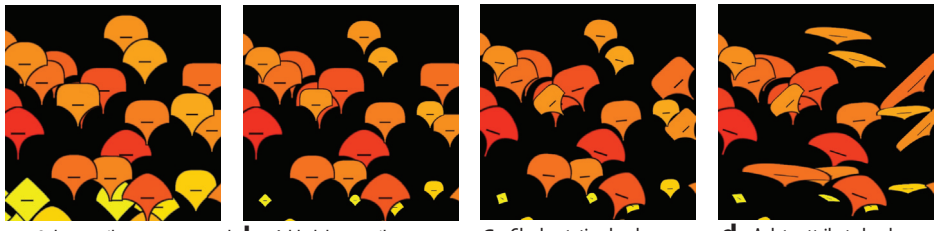
# Paper B

# Critical Design and Realization Aspects of Glyph-based 3D Data Visualization

Andreas E. Lie, Johannes Kehrer, and Helwig Hauser

Department of Informatics, University of Bergen, Norway





**a.** 2 data attributes represented as upper/lower glyph shape **b.** Added data attribute to overall glyph size **c.** Glyph rotation has been assigned a data attribute as well **d.** A data attribute has been assigned to glyph aspect ratio

**Figure 1:** Adding more attributes to the glyph, while preserving the glyph's orthogonality.

## Abstract

Glyphs are useful for the effective visualization of multi-variate data. They allow for easily relating multiple data attributes to each other in a coherent visualization approach. While the basic principle of glyph-based visualization has been known for a long time, scientific interest has recently increased focus on the question of how to achieve a clever and successful glyph design. Along this newer trend, we present a structured discussion of several critical design aspects of glyph-based visualization with a special focus on 3D data. For three consecutive steps of data mapping, glyph instantiation, and rendering, we identify a number of design considerations. We illustrate our discussion with a new glyph-based visualization of time-dependent 3D simulation data and demonstrate how effective results are achieved.

This paper was published in *Proc. Spring Conf. on Computer Graphics (SCCG 2009)*, pages 27–34, 2009. The work was also presented by Andreas E. Lie at the SCCG 2009, April 23–25, Budmerice, Slovakia.

# 1 Introduction

In scientific projects as well as in commercial applications we see an increased utilization of computational simulation for the investigation of natural phenomena. Compared to earlier years, current resulting datasets are 3D instead of 2D, time-dependent instead of single time step, only, and multi-variate with many values per space-time location, to name just three of more recent properties (which soon will be standard in many cases). This means that not only the large size of simulation datasets is challenging, but also its complexity. With this, it is getting more important and more difficult to enable users to "read between the lines", i.e., to better understand the relations between different data dimensions. A variety of useful visualization approaches have been proposed to reveal the information that is contained in high-dimensional data, and we refer to Ward [250] and Bürger and Hauser [24] for a review of some of these approaches.

An interesting approach is to use glyphs to represent multiple data variates per space-time location. A generic (and usually also relatively simple) shape is defined with a set of variable appearance properties, including shape characteristics, color, opacity, etc., that—when instantiated—is parameterized by a subset of the data variates per data item. Glyph-based visualization has been known for many years. It is more than 15 years ago, for example, that de Leeuw and van Wijk [48] proposed the so-called "local flow probe" as an interesting example for glyph-based visualization of 3D flow data. Recently, there is new interest in glyph-based approaches. See, for example, Ropinski and Preim [194] for a recent survey. Several interesting examples of glyph-based visualization that have recently been published, see Oeltze et al. [174], Kindlmann and Westin [123], Ropinski et al. [195], and Meyer-Spradow et al. [156].

An important lesson learned from these recent works is that an appropriate glyph design is crucial for the success of a glyph-based visualization. It was a wide-spread opinion in the related research community for a long time, that "just" knowing the well-published basic principle of glyph-based visualization would well suffice to also utilize this approach successfully. More recently, however, it has been understood that only very well designed glyphs are actually useful. In this paper we therefore discuss critical design aspects of glyph-based visualization with the special focus on 3D data. We exemplify our discussion with a new, glyph-based visualization of time-dependent 3D simulation data and show how effective results can be achieved.

In section 2 we describe similar and related work. In sections 3 and 4 we will present considerations with respect to glyph representation. Section 5 will exemplify different datasets visualized by glyphs created according to these considerations.

## 2 Related Work

Ward [249] discusses different glyph types and placement strategies for glyphs. This work is considered highly relevant for glyph-based visualization. Ropinksi et al. [195] propose glyph placement strategies, and use glyphs on surfaces in 3D for visualizing multi-variate data. Our focus is not placing glyphs on surfaces, but in a truly 3D environment. They also provide a thorough taxonomy for glyph-based visualizations in the medical domain [194]. We aim to build upon this by extending the preprocessing step prior to the creation of the glyphs. Sawant and Healey [204] successfully map several attributes to their glyphs used in flow visualization. Our aim is to use more complex shapes to enable more variates to be mapped to glyph properties. Bertin [12] proposed six retinal variables: shape, size, orientation, color (value and hue) and texture, which stresses the importance of careful and well thought glyph design. Shaw et al. [207] show that it is hard to distinguish similar shapes of super ellipsoids. This is a problem related to the orthogonality of the shapes. Wong et al. [262] fuse several layers of variate visualizations, but suffers slightly from cluttering and occlusion. De Leeuw and van Wijk [48] designed the Probe glyph which could properly visualize twelve different parameters simultaneously. Van Walsum et al. [239] describe features and attribute sets which are extracted from the regions of interest in the data, allowing local minima and maxima to be mapped to icons (or glyphs). We build upon the idea of features and attribute sets, and allow changing of mapping inside the extracted data ranges as well. Densely packed icons have been used to form visual textures [179], representing multi-variate data. Stolte et al. [217] proposed a system named Polaris which does interactive visual analysis and allows data transformations to unveil hidden relations and data. Kindlmann [122] uses different kinds of super quadrics as glyphs to visualize data. Similar work has been done by T.J. Jankun-Kelly and Mehta [99] which use super ellipsoid glyphs to visualize variates in nematic liquid crystals. Piringer et al. [181] show the importance of halos to enhance the depth perception and separability of points. Toutin [225] uses color to assist the viewer in interpreting spatial relation.

## 3 Overview

E. Tufte discusses in his work [232] that developing design techniques for enhancing graphical clarity is crucial. According to the model of the visualization pipeline [77], we suggest to substructure the task of glyph-based 3D visualization into three separate steps, namely Data Mapping, Glyph Instantiation and Rendering. Figure 2 shows the flow from the data mapping step, to glyph instantiation and the final rendering step. It is generally very useful for glyphs to have normalized input from the variables the glyphs represent. While the interval $[-1, 1]$ arguably can be used for such a normalization purpose, we choose to

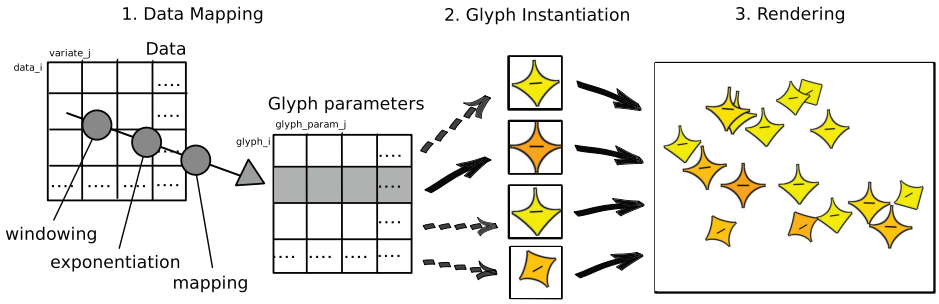



**Figure 2:** Data variates undergo data mapping stages; windowing, exponentiation and mapping. These values are then used to instantiate the corresponding glyphs, (e.g, determining shape, size) and finally the glyphs are rendered into the context.

use the range $[0, 1]$ for the design of our very comprehensible model. Accordingly, the data mapping step comes first in our pipeline, and the glyph instantiation and rendering steps successively. We think its generally useful to consider to data mapping as three elementary stages, namely, windowing, exponentiation, and mapping. Aiming for a glyph-based visualization of 3D data, there is always a question on how to cope with the problem of occlusion and cluttering which results in information loss. We propose three generic options in order to deal with these challenges, namely, halos, chromadepth, and interactive slicing. The glyphs depicted in figure 1 are created by drawing two super ellipses (one for the upper, the other for the lower half) and combining them. These shapes are considered simple, meaning that they are easy to understand and allow for mental completion if they were to overlap or partially occlude each other. The glyphs can have data mapped to them, controlling the upper and lower half, color, rotation, size and aspect ratio. In section 6, we will describe the creation of these glyphs in more detail.

# 4 Selected generic Considerations with respect to Glyph Representation

## 4.1 Data Mapping

In this section we discuss three steps in the data mapping stage where the user has the ability to increase the value of the visualization by adjusting the data variates according to his/her needs. We also cover several aspects of glyph design to provide some guidelines for making the most useful glyphs. Finally, we will cover how to cope with occlusion and cluttering in the visualizations.

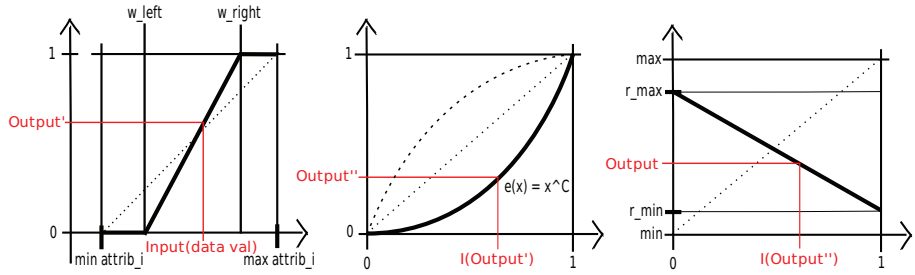



**Figure 3:** (a) User selects $w_{left}$ and $w_{right}$ which adjusts the data mapping ramp for input datavalues. The dotted line represents the default value for the ramp. (b) User can control the data mapping curve by adjusting the C term in $x^C$. The middle dotted line represents the default curve, the curve can be bent upwards or downwards. (c) The output range of the datamapping can be adjusted to fit the datavariates. User selects $r_{min}$ and $r_{max}$ to clamp the output range.

**Windowing:** The process of windowing serves the effect of enhancing differences in data values. This is achieved by clamping the range of selected values to be linearly distributed to the output range. Figure 3a displays such a mapping function for data values. Every value outside the clamped window results in either the mapping functions minimum or maximum output depending on the data value in question. The windowing allows the user to select the $w_{left}$ and $w_{right}$ giving the user complete control over how and where the mapping slope will reside in the data domain. Windowing is a method widely used in medical visualization where it is more commonly known as contrast enhancement, and is often applied to visualize 2D or 3D images. In figure 4 several examples of such windowing are illustrated. The default setting equals a simple linear mapping of the data values (the dotted lines in figure 3a).

**Exponentiation:** After the windowing all datavalues are transformed to the unit range, $[0, 1]$, the next step in our pipeline is to adjust the function that further processes the data. This process is known as exponentiation. Here we allow the user to control the exponent $c$ in the simple, yet powerful function $x^c$. Such functionality makes selected data values easier to distinguish based on which curve was chosen. Figure 3b displays such exponential mapping. This curve exists only inside the window and is by default $x^1$ which is linear mapping. Figure 4 is an example where both windowing and exponentiation have been applied to the data.

**Mapping:** Data may have characteristics that may result in unwanted glyph behaviour. We therefore allow the glyph mapping to be altered according to the users wish for solving such problems. It would, for instance, make little sense to

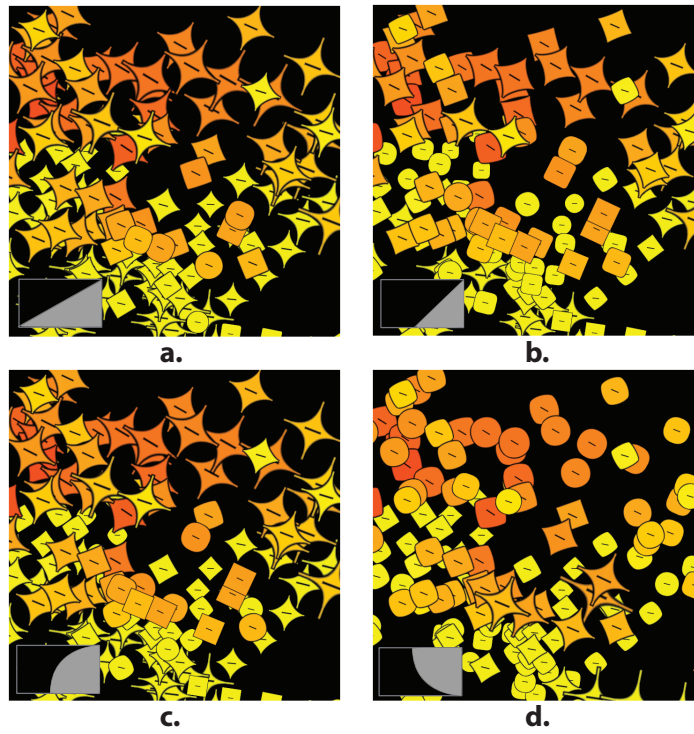
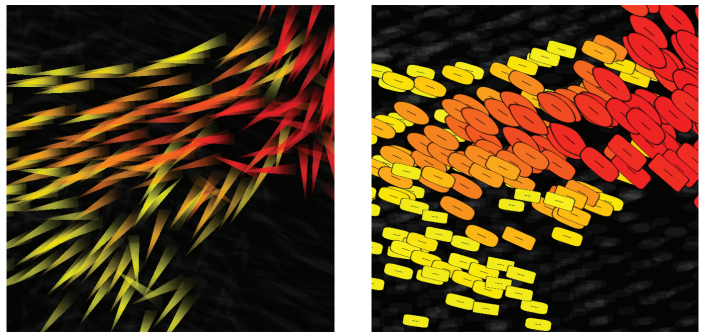



**Figure 4:** Different changes to the datamapping stage done successively: (a) depict default datamapping. In (b) window adjustments have been done to achieve better contrast with respect to higher values. Exponentiation curve has been fine tuned in (c) to reveal more differences among the lower values for the selected range. In (d) the mapping output range has been inverted.

map time to a rotation attribute, which could rotate the glyph $\pm 45°$. Mapping allows us to change the output range of the data mapping pipeline, making this attribute more fitting by restricting the output range to (in this case) $[0.5, 1]$ instead of the regular $[0, 1]$ range. Now time $t = 0$ equals zero rotation, and $t = max$ equals the equivalent max glyph rotation. An example where mapping applies well, is in the case where the user wants to focus on low values for a certain data variate. This variate can now be mapped to glyph size, resulting in more prominent glyphs for low values by reversing the mapping. Figure 3c shows an example of change in the output function. Figure 4d displays the effect of reversing the mapping.

Data mapping consists of three simple steps, which are intuitive and straight forward to specify. These steps are an important aspect for glyph-based visualization. Users of glyph-based visualizations will start by selecting a mapping



**a.** 3D arrow glyphs **b.** 2D glyph billboards

**Figure 5:** (a) 3D glyphs are fine, if data with an inherent relation to the 3D space is shown (such as flow direction); (b) otherwise 2D (billboarded) glyphs are preferred.

(possibly choosing to invert data or clamp output ranges). After specifying the correct output ranges for the data mapping pipeline, they then proceed to change and fine tune the windowing and exponentiation data mapping stages.

## 4.2 Glyph Instantiation

**2D vs. 3D:** Using glyphs for visualizing multi-variate variables in 3D is challenging for the user. To mentally reconstruct the particular values represented by glyphs is non-trivial. Size, orientation and geometric properties are often mapped to the data (because the properties represent strong visual cues). Effects from 3D projection complicate the interpretation of the glyph shapes. Therefore, we suggest to only use 3D glyphs if they are geometric properties and inherently related to the 3D position where they are placed. For example, it makes sense to show arrow based glyphs mapped with the three velocity components in 3D simulation domains. For other visualization we strongly propose to use billboarded 2D glyphs since they avoid distortion of other glyph properties. Figure 5a is such an example of arrow based glyph visualization.

**Orthogonality:** A big challenge in glyph design is the orthogonality of the glyph components. If glyph parts are not visually separable, the interpretation is non-trivial. An example of such a mapping, is to map data values to individual RGB color components, since interpreting the individual color components from a color is very hard. Furthermore, large numbers of variates is hard to accommodate if the glyph shape is simple. The glyph size and complexity must be seen in a direct relation to the resolution of the visualization. If there are few datapoints large glyphs can be used. If there are many data points, simple shaped glyphs that can be displayed in a densely manner are required. Eventually it must be

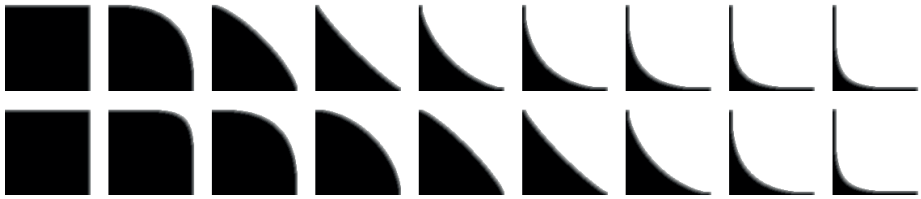


**Figure 6:** Upper row: the attribute is directly mapped to the shape exponent. The lower row: perceptual shape normalization along the diagonal.

assumed that there exist some maximum number of glyph shapes and properties that the user can distinguish and discriminate. In our examples we found it challenging to have 5 or 6 different variates mapped to the glyph. In the figure 1 we can see how well the glyph avoids distortion when having more and more attributes mapped to it. In the first image three variates are mapped. Two to (upper and lower) shape and one to color. The second image has included size as a parameter for the glyph. In the third rotation is introduced, and in the last visualization aspect ratio is mapped for a total of 6 variates. Mapping both size and aspect ratio should be handled with care, since they combined impose a perceptual challenge on the viewer. They can however be utilized, for instance, by using size as a selection parameter. Thus large glyphs have important characteristics (small glyphs do not), and having aspect ratio possibly depicting another parameter of interest. An example mapping can be seen in the Hurricane Isabel visualization, see figure 10, where size depicts amount of clouds, and aspect ratio depicts air velocity.

**Normalization:** A variation of the glyph shape, for example, has the implicit effect of changing the size (i.e., area) of the glyph. We therefore suggest to normalize these effects against each other, e.g., to adapt the overall glyph size in order to compensate the otherwise implicit change of the size due to the shape variation. In our glyph design we also take into consideration that the shapes used for representing the glyphs are visually not equally spaced. Figure 6 shows how the shapes originally where, and how (the lower row) they became after shape equalization. For the shape equalization, we calculated the distance from the center of the shape, to its curve along the diagonal, and used these calculations to select shapes which would result in visually equally spaced shapes. In figure 7 you can see how measurements were made to normalize the shapes.

**Redundancy:** As mentioned above, it is challenging to read glyph-based visualizations, even if designed with care. Using redundancy to depict especially relevant data characteristics is an useful way to emphasize important attributes,

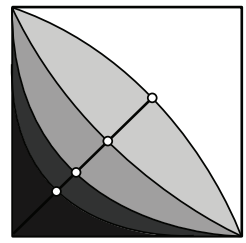


**Figure 7:** We calculate the difference between the innermost and outermost curve, and use this number to equally distribute the shape variations linearly along the diagonal.

and decrease the chances for information loss. The glyphs design inherently displays the same shapes on the right and left side of the glyphs, which allows users to correctly understand how the glyphs are shaped, despite being able to see only partially occluded glyphs in visualizations. The densely packed glyphs in figure 9 relies strongly on this. Figure 9 has color and size mapped to temperature, emphasizing the importance of temperature in the visualization. We also use a vertical bar inside the glyphs, to assist users interpreting the rotation of a given glyph. Figures 1, 9 and 10 all show these bars that assist user interpretation of rotation.

Glyph-based visualization is just one opportunity to visualize multi-variate data. Glyphs are helpful to understand multiple variates simultaneously (e.g. "reading between the lines"). Therefore it is of much more importance to carefully think about inter property aspects of glyph design, i.e. how the different glyph characteristics are dependent on each other, than the individual glyph expressions of data variables. The size and aspect ratio characteristics is an excellent example of this. Eventually it is how all the different variations harmonize that will result in whether or not users are able to achieve their goals.

## 4.3 Rendering

A general problem in 3D visualization is occlusion, depth perception and visual cluttering. Glyph-based visualizations are most comprehensible when selecting (or placing) glyphs in such a manner that only a small number of glyphs are shown simultaneously and that they do not overlap or occlude one another. Often it is not trivial to achieve either small numbers of glyphs, or view angles where glyphs do not overlap. In these circumstances we suggest three approaches; halos, chromadepth, and interactive slicing, that by themselves have been proposed in different situations and scenarios.

**Halos:** A simple but effective way to improve depth perception of discrete primitives in datasets is to include halos around the primitives. This will make the

primitives stand out from other objects, and allow users to mentally complete the partially occluded shapes since the individual glyphs can be identified. This is a technique very common among illustrators for drawing attention toward objects. Piringer et al. [181] and Interrante and Grosch [92] use halos to emphasize discontinuity in depth and to draw the users attention towards objects.

**Chromadepth:** Relative depth perception is hard to cope with by using only halos. The relation of two non-overlapping glyphs are diffucult to determine (which is in front of which). By allowing the use of color to represent depth (instead of a data attribute), similar to the work of Toutin [225], and by using a color scale that either is complementary or has clear continuous change, depth perception can be achieved successfully. The figure 8c is an example of chromadepth and its effect.

**Interactive Slicing:** Occlusion is a major problem when reading glyphs. Since halos and chromadepth does not cope with occlusion, we suggest to employ interactive slicing, a technique that allows for view dependent slice-based visualizations. The user specifies a plane in the 3D visualization, which determines whether the renderer should omit the data or not. This way we can avoid the occlusion problem by suppressing the occluding glyphs in the front of our specified plane. Figure 10 is an example of such slicing, where only the lowest layer of the hurricane Isabel is visualized. Interactive slicing is commonly used in volume visualization.

These three solutions enable the user to cope with a large amount of problems caused by occlusion and visual cluttering. They also enhance and attract the focus of the user to the glyphs with halos and may assist revealing hidden nuggets (valuable information) in the data by interactive slicing.

# 5 Demonstration

In this section we will demonstrate that glyphs can be used in conjunction with very different datasets and successfully depict several different characteristics simultaneously. The datasets are the Diesel Exhaust System-dataset and data from the hurricane Isabel. Both these datasets are studied thoroughly by Doleisch et al. [54, 55].

## 5.1 Diesel Exhaust System

The diesel exhaust system includes a diesel particle filter which traps soot, and burns the soot at over 1000 degrees to oxidize it at different intervals. The dataset contains over 260.000 vertices, and is given at ten timesteps. In Figure 9 timestep

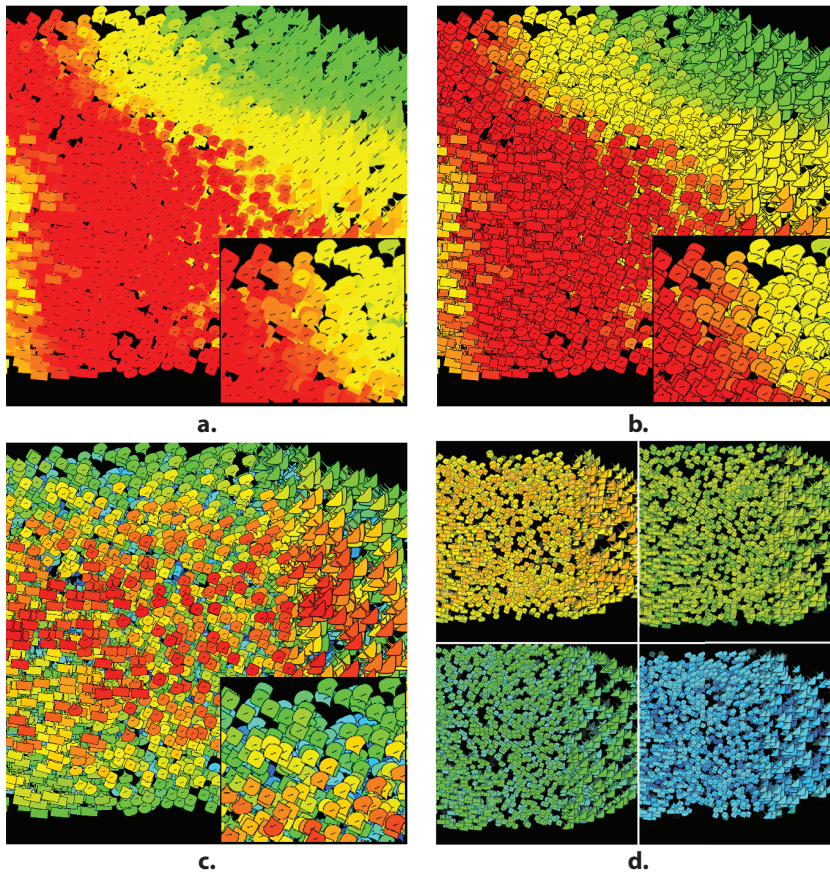



**Figure 8:** These visualizations are from the Diesel particle filter dataset. (a) represents just simple colored glyphs. (b) emphasizes glyph differentiation by adding halos to the glyphs. In (c) color is used to assist the user in interpreting the spatial relation between the glyphs (depth perception). Various orthogonal slices can be seen in (d) where color indicates the depth of the slice taken, similar to (c).

five is visualized. We map data attributes to the glyph properties as specified in table 1.

Exhaust and soot particles from the diesel engine is guided into a diesel particle filter where soot is trapped and oxidized at around temperatures from 600 up to over 1000 degrees celsius. Table 1 shows which data values are mapped to the glyph properties. The oxidation process moves from left to right.

From the visualization in figure 9, we can see that the temperature differences cause an uneven oxidation of the soot. Two areas with non-optimal temperature levels (green) can be located on the right side of the oxidation peak temperatures

**Figure 9:** Diesel particle filter oxidizes soot at temperatures above 1000 degrees Celsius. The color represents the temperature of the process. Upper glyph shape visualizes the amount of soot at that given point, and the lower shape the rate of rate of change. Rotation is the amount of $O_2$ which is needed for the oxidation process. Glyph size is mapped to temperature to achieve redundancy for the visualization. One can see that there exist two green areas right of the peak temperatures where there are high amounts of soot, and the oxidation is non-optimal because of lower temperatures.

| | |
|---|---|
| Color | Flow Temperature |
| Glyph Upper | Soot amount |
| Glyph Lower | Soot amount second derivative |
| Glyph Size | Flow Temperature |
| Glyph Rotation | $O_2$ fraction |

**Table 1:** Glyph property mapping for Diesel Exhaust System dataset.

| | |
|---|---|
| Color | Temperature |
| Glyph Upper | Pressure |
| Glyph Lower | Precipitation |
| Glyph Size | Clouds |

**Table 2:** Glyph property mapping for Hurrican Isabel dataset.

(red). In these areas one can see that the $O_2$ levels are low (rotation), a critical part (in addition to high temperature) to burn as much soot as possible. The visualization also can verify that soot amount left of the peak temperatures is very low.

### 5.2 Hurricane Isabel

The hurricane simulation contains meteorological data from the class 5 hurricane Isabel which ravaged in 2003. The dataset has 24 variables and contains 24 timesteps, each of 100.000 vertices. In Table 2 the glyph property mapping for the visualization in Figure 10. We choose to focus on fast moving air flows that exist close to the surface. Through the use of slicing, only the lowest layer of data (closest to the surface) is selected. Semi-transparent glyphs can be identified in this visualization, which is a direct result of brushing flow velocity vs clouds with smooth degree-of-interest in the SimVis framework.

We can identify the eye of the hurricane in the lower right corner of the visualization in figure 10. The eye is almost surrounded by a wall of precipitation and relatively colder airflows. From this visualization one can see that there is low pressure inside the hurricane, and that cold winds from north mix in with warm air from the south. A cold front can be identified in the higher left parts of the visualization. There exists also a very interesting area directly below this cold front, where a small subset of glyphs (four) identifies a region where there is a high amount of pressure and precipitation.

## 6 Technical Details

Our glyph visualization is integrated into the SimVis framework [51] for assisting the user in visual datamining and analysis. The SimVis framework allows inter-

**Figure 10:** The hurricane Isabel dataset, timestep 12. Color represents temperature, and the amount of clouds is mapped to glyph size. The upper shape represents pressure, and the lower shape precipitation. The visualization depicts the fast moving clouds (specified via brushing), and the eye of the hurricane is visible in the lower right, surrounded by high amounts of precipitation and cold airflow.

active visual analysis of large multi-variate datasets. The renderer allows user changes to the data pipeline to adjust for more optimal glyphs. See section 4.1 on windowing and exponentiation for a thorough explanation. The framework and the plugin was developed in programming languages C++, OpenGL and CG-shader.

We employ a glyph texture atlas containing all possible glyph variations. This texture allows us to externalize the glyph itself, making the glyphs fully independent of the framework. By using such atlases, new glyphs and variations are very likely to appear since the framework is unaware of a glyphs visual characteristics. In the atlas we only represent one quadrant of each glyph, to save valuable space by omitting redundant information. The glyph shapes can easily be reconstructed inside a shader by simple mirror and rotation operations.

Our glyph texture atlas was created by drawing various super ellipses, and saving them in the atlas. We also employed antialiasing to smooth the borders of the glyphs giving them a more visually pleasing look. Halos were saved in a separate channel of the atlas, enabling the shader to include the halo if the user specified so.

The shader would ultimately load the glyph atlas as a lookup texture, picking the correct quadrants (quarters of the glyph) from that texture and mirror and rotate these quadrants to completely draw the glyph itself. Since the texture coordinates can easily be modified inside the shader, we allow for different halves to be drawn in the glyph thus enabling more attributes to be mapped to the glyph. Size, rotation and aspect ratio is also performed by adjusting the texture coordinates to achieve the desired effect.

We choose Super Ellipses as a basis for our glyph shapes. These shapes are simple to understand, and easy to get to parameterized form. The super ellipses can be varied by changing their controlling exponent, from a square (low exponent) through circle and diamond shapes to star shape (with high exponents). This exponent is continuous and therefore well suited for having mapped data to it. These shapes are easy to distinguish from each other, and work very well to convey information of the data values they depict. By having two separate ellipses, one for top and one for bottom, we can map two different datavalues to these parameters. The glyph shapes were perceptually normalized to allow them to represent an even amount of change in the corresponding data. See figure 7 and figure 6. The area of cover of every glyph is also calculated, to allow for size normalization during the glyph instantiation step.

An advantage of simple shapes, is that the viewer of the visualization still can mentally complete the glyphs if they were to overlap and occlude each other. This quality, in addition to visual redundancy, makes simple glyphs very efficient in conveying their information.

We are able to map data to color, size, the two super ellipse halves, rotation and aspect ratio of the glyph. The glyphs can properly visualize six different parameters in addition to the DOI controlled opacity provided by the SimVis

framework. These attributes are closely coupled with retinal variables described by Bertin [12]: shape, size, orientation and color (hue and value).

## 7 Summary and Conclusions

We present an effective way to allow user adjusting of data values that is both straightforward and comprehensible. All data variates may undergo the data mapping steps: windowing, exponentiation, and mapping. These are considered as easy to understand, but powerful tools that allow fine tuning the resulting glyph shapes. The data mapping stage inherently increases the value of the resulting glyphs.

The design of glyphs to be used in visualizations is both complex and crucial. We point out improvements of glyph design by discussing the glyphs inter property aspects (orthogonality and redundancy). We moreover propose to normalize glyph shapes both perceptually and in size to avoid loss of orthogonality while maintaining clarity. 2D shapes are ultimately easier to interpret than their 3D counterparts, and we propose to only use the 3D glyphs when spatial relation is inherent.

We stress to use already existing techniques to help avoid problems with occlusion and cluttering. The use of halos help emphasize discontinuity. Chromadepth and interactive slicing help the user interpret depth.

## 8 Future Work

A user study would help emphasize the strengths and weaknesses of glyph-based visualizations, as well as feedback on the guidelines provided to create such glyphs. Another interesting angle would be to apply MPEG-7 shape descriptors to have perceptually focused metrics for glyphs.

## Acknowledgments

We want to thank Helmut Doleisch and Phillip Muigg (SimVis GmbH, Austria) and Jean-Paul Balabanian (Univ. of Bergen, Norway) for guidance and help regarding SimVis and the implementation of our glyph-based renderer. Thanks to Robert A. Johannessen (Univ. of Bergen) for proof-reading. The Diesel Exhaust System dataset is courtesy of AVL List GmbH, Graz, Austria. The Isabel hurricane dataset is courtesy of the National Center for Atmospheric Research (NCAR), USA.

# Paper C

# Interactive Visual Analysis of Heterogeneous Scientific Data across an Interface

Johannes Kehrer,[1] Philipp Muigg,[2,3]
Helmut Doleisch,[2] and Helwig Hauser[1]

[1]Department of Informatics, University of Bergen, Norway
[2]SimVis GmbH, Vienna, Austria
[3]Institute of Computer Graphics and Algorithms, Vienna University of Technology, Austria

## Abstract

We present a systematic approach to the interactive visual analysis of heterogeneous scientific data. The data consists of two interrelated parts given on spatial grids over time (e.g., atmosphere and ocean part from a coupled climate model). By integrating both data parts in a framework of coordinated multiple views (with linking and brushing), the joint investigation of features across the data parts is enabled. An interface is constructed between the data parts that specifies 1) which grid cells in one part are related to grid cells in the other part, and vice versa, 2) how selections (in terms of feature extraction via brushing) are transferred between the two parts, and 3) how an update mechanism keeps the feature specification in both data parts consistent during the analysis. We also propose strategies for visual analysis that result in an iterative refinement of features specified across both data parts. Our approach is demonstrated in the context of a complex simulation of fluid–structure interaction and a multi-run climate simulation.

This article is accepted for publication in *IEEE Transactions on Visualization and Computer Graphics*, 17(7):934–946, July 2011. Digital Object Identifier no. 10.1109/TVCG.2010.111. Manuscript submitted 24 July 2009; revised 2 Feb. 2010; accepted 20 July 2010; published online 20 Aug. 2010.

# 1 Introduction

Computational simulation is used in science and engineering to investigate dynamic processes and complex phenomena. Interactive visual analysis enables the user to explore and analyze data in a guided human–computer dialog. Using proven interaction schemes such as linking and brushing, a powerful information drill-down process is supported [209]. Visual analysis is based on concepts such as coordinated multiple views, interactive feature specification via brushing, focus+context visualization, and on-demand data derivation [73].

Scientific data in a traditional application scenario is usually given in a coherent form. It can be considered, to a certain degree, as a table with rows and columns that contains multiple data attributes (given in relation to space and time). We call this a *single-part* scenario. In practice, however, we increasingly often find model and data scenarios that are more heterogeneous. They consist of two or more individual data parts that are related to each other. The data parts are, for example, computed with different simulation models, given on various data grids, with different dimensionality (e.g., 2D/3D data). Such *multi-part* scenarios present us with the challenge of integrating multiple data parts in the analysis.

Dynamic flow, for instance, is traditionally simulated with a rigid boundary. In modern *fluid–structure interactions* (FSIs), however, a movable or deformable structure interacts with an internal or surrounding fluid flow. These simulations are becoming more popular and belong, with respect to both modeling and computational issues, to the most challenging of multiphysics problems [22]. Fluid and solid parts are usually modeled individually on spatially adjoining grids that are connected by a so-called *interface*.[1] The latter represents the physical boundary between the two parts and enables them to influence each other during the simulation (compare to airplane wings or turbine blades that are deformed by the surrounding flow). Also in the climate system, as another multi-part scenario, atmosphere, ocean, ice, and land interact with each other. Ocean and atmosphere, for example, interact by means of thermal absorption, precipitation, and evaporation [86]. To understand such dynamic processes, the climate components are usually modeled individually and then coupled in the simulation, often with additional coupler modules.

Creating a coherent visualization from heterogeneous data that consists of two parts (e.g., atmosphere and ocean, or fluid and structure) is a challenge for visual analysis. How can we investigate feedback between the two data parts? The analyst is, for example, interested in areas of an ocean model that are influenced by adjacent hot areas in the atmosphere. The corresponding regions are first selected in the atmosphere via brushing. This feature then needs to be transferred

[1] The term interface is used in many disciplines such as chemistry, physics, biology, or computer science. According to the Oxford English dictionary, it denotes "a point where two things meet and interact", e.g., the surface that connects two physical materials, a biological cell and another material, or a human and a computer (user interface).

to the ocean part where it can be related to ocean features and further analyzed. In our analysis framework, we realize this feature transfer by an interface that connects the two data parts similar to a fluid–structure interaction. Our interface is designed such that the data parts can be given on different grids (e.g., 2D/3D, unstructured, and hybrid), with different resolutions or time-scales.

Another example that can be considered a multi-part scenario is hierarchically organized scientific data. A data part with higher data dimensionality can be related to a part with lower dimensionality, and vice versa. Multi-dimensional scientific data signifies that different attributes (e.g., temperature, pressure) are measured or simulated with respect to an $m$-dimensional data domain. The domain (i.e., the independent data dimensions) can be 2D or 3D space and time, but also input parameters to a simulation model. In climate research or engineering, for instance, so-called *multi-run* simulations have become an important approach to assess simulation models [86, 151]. They are used to evaluate the variability of a model and to better understand how sensitively the model reacts to its input parameters (sensitivity analysis [70]). The values of certain input parameters are varied. Simulation outputs (runs) are then computed for many combinations of the parameters. This leads to multi-run data where a collection of values exists per space/time location [141] (one value for each run).

The analysis of such higher-dimensional scientific data is generally challenging. A natural attempt in such a situation is to reduce the data dimensionality, for instance, by computing statistical aggregations along selected independent dimensions (e.g., averaging with respect to a spatial axis, the time axis, or the input parameters of the simulation). In practice, often only the aggregated data is further analyzed.

In this paper, we demonstrate that it is useful to integrate both the original multi-run data and the aggregated data part (with lower dimensionality) into the visual analysis. Similar to the simulation of a fluid–structure interaction, we construct an interface as a bridge between the two data parts. During the visual analysis, the interface is used to transfer selections (features specified via brushing) between the parts. Thus, complex relations can be investigated within and across the two data parts.

Corresponding to the multi-part scenarios described above, we have researched this problem and present the following contributions with this paper:

- We propose the construction of an interface that enables the joint visual analysis of heterogeneous scientific data that consists of two data parts.
- We propose strategies for visual analysis where the analyst works with both data parts simultaneously.
- We demonstrate the usefulness of our approach in the context of a fluid–structure interaction and a multi-run climate simulation.

## 2 Related Work

The integration of abstract data from multiple sources is common in *information visualization* (e.g., in relational databases [171], or web data [28]). North et al. [171] propose flexible visualization schemas built upon the snap-together visualization model, which enable the user to create multiple-view visualizations analogous to relational data schemas. Polaris/Tableau [217] supports the exploration of data cubes, where data is given at different hierarchical levels. These approaches deal with heterogeneous abstract data. In this paper, we present a visual analysis approach for heterogeneous scientific data usually given on grids over time. Cross-filtered views [252] allow interactive drill-down into relationships between multiple data attributes, also across multiple datasets. Brushing filters between pairs of views can be enabled/disabled. Cross-filtered views are neutral with respect to the data dimensionality and also support the derivation of new data attributes. With our approach, we account for the heterogeneity of the independent dimensions of space and time, similar to scenarios with multi-run data. Features can also be transferred between non-overlapping data parts such as spatially adjoining physical materials or interacting climate components. While the data is filtered with cross-filtered views, our approach leads to a joint focus–context discrimination that is related across heterogeneous data parts.

The area of *coordinated multiple views* has been steadily developing over the past fifteen years (see Roberts [192] for an overview). XmdvTool [248] allows the analysis of complex relations in multi-variate data using combinations of brushes in multiple views. SimVis [52] and WEAVE [66] are just two examples that realize the concept of a visual analysis framework for scientific data. Multiple linked views are used to simultaneously show, explore, and analyze different aspects of multi-variate data. The views are used next to each other and include 3D views of volumetric data (grids, also over time), but also attribute views such as 2D scatterplots, function graph views, or histograms. Interesting subsets of the data are interactively selected (brushed) directly on the screen, the relations are investigated in other linked views (compare also to the XmdvTool [248]).

In some systems, the result of a smooth brushing operation [53] is reintegrated within the data in the form of a synthetic *degree-of-interest* data attribute $DOI_j \in [0, 1]$ for every data item $j$ (compare to the DOI attribution for generalized fisheye views by Furnas [65]). This data attribution represent the first interpretation level, ranging from data to knowledge [33]. Logical combinations of brushes in multiple linked views enable the specification of complex features in a hierarchical feature definition language [52]. The DOI attribution is used in all linked views to visually discriminate interesting features from the rest of the data in a focus+context visualization style [161]. Our framework is based on these concepts, extending the analysis capabilities to scenarios with heterogeneous scientific data. We connect the two data parts by an interface

that transfers fractional DOI information between the parts. Complex features can be specified via (smooth) brushing within and across the data parts.

According to Fuchs and Hauser [63], scientific data stemming from different modalities (e.g., different simulation models, or measurements) can be fused at different levels in the visualization pipeline. In multi-block flow visualization, for instance, simulations are performed on multiple grid types with different resolutions [56]. Since the blocks do not represent different physical materials, a feature transfer across the blocks would not make sense. In the visualization, the blocks are usually fused at the data level (e.g., by constructing one hybrid or unstructured grid). In VisIt, for instance, data from different meshes are evaluated onto a common mesh (cross-mesh field evaluation [37]). Since the data is fused at the data level, it can be considered as a single-part scenario according to our terminology. Treinish [228] proposes a uniform data model that adjusts to the data structure and how the data is processed. Using such a data-/model-centric approach, data from different sources can be fused (or correlated), thus avoiding unnecessary interpolation or resampling to a common mesh. With our approach, fusion is performed at the feature/interpretation level [33] instead of the data level.

The treatment of *multi-run data* is rather new to the visualization community [141]. Information visualization techniques such as parallel coordinates or scatterplot matrices are used in combination with statistics to improve the understanding of the model output from multi-run simulations [41]. Nocke et al. [166] propose a system of coordinated multiple views to analyze a large number of tested model parameters and simulation runs. Statistical aggregations of the multi-run data are visualized, e.g., using linked scatterplots, graphical tables, or parallel coordinates. In their approach, however, the data is given in a coherent data part. Potter et al. [185] propose a framework that consists of overview and statistical visualizations for analyzing multi-run data. Matković et al. [151] visualize multi-run data as families of data surfaces with respect to pairs of independent data dimensions. Projections and aggregations of the data surfaces are analyzed at different levels (e.g., a 1D profile or single value per surface). In our work, we propose a more general interface concept that connects data items between two parts of scientific data and supports the transfer of fractional DOI information. This approach can also be used for multi-run data. In recent work [111], we have integrated traditional and robust estimates of statistical moments in the visual analysis of such data, where we also utilize the interface described here.

Kao et al. [108] visualize distributions over 2D multi-run data, where the distribution can apparently be represented by statistical parameters. For other cases, they propose a shape descriptor approach [107] constructing a 3D volume with the probability density function (PDF) of the data as voxel values. Mathematical and procedural operators [141] are proposed to transform the distribution data into a form where existing visualization techniques can be applied (e.g., pseudo-

coloring, streamlines, or isosurfaces). This operator approach is very promising due to its flexibility. However, it is not integrated in a visual analysis framework that would enable the analyst to interactively specify features within the transformed data. Recently, Potter et al. [186] extend the box plot [154] to include additional statistics. The resulting summary plot depicts different characteristics of multi-run data, however, it cannot be placed in a dense manner. In our multi-run example, we use carefully designed glyphs [136] to visualize aggregated data properties in a 3D context.

## 3 Sample Analysis of an FSI Scenario

Fluid–structure interactions (FSIs) are complex multiphysics problems and currently an important topic in simulation research. In such scenarios, a solid structure interacts with a surrounding fluid flow, for example, by exchanging heat and/or being deformed. The variety of FSI occurrences is abundant and ranges from bridges, flexible roofs, or offshore platforms to micropumps and injection systems, from parachutes to airbags, to blood flow in arteries or artificial heart valves [22]. In the following, the study of heat transfer in an FSI scenario is used to illustrate our proposed methodology. Motivated by this example, we later come up with a more general approach that can also be applied in other scenarios with heterogeneous data such as multi-run data.

In our example, data from a multiphysics simulation of warm water flow through a cooler aluminium foam is investigated. The main goal of the domain experts is to understand how the micro structure of the simulated foam influences its thermal behavior. This knowledge can then be used to derive approximated models of the foam which can be applied within larger scale simulations. A more in-depth understanding of the flow characteristics through the simulated domain can help the application experts to experiment with different foam structures. This eventually leads to more desirable thermal properties of the foam structure.

The modeled domain contains two types of physically different materials, i.e., water and aluminium. The underlying multiphysics simulation, therefore, generates two spatially disjoint result volumes (see Figs. 1a and 1b). Both 3D volumes are connected by an interface which identifies common faces between fluid and solid grid cells (illustrated in Fig. 1). During the simulation, the fluid and solid part can interact with each other via the interface, and exchange properties such as heat.

In the visual analysis, we integrate both data parts, fluid and structure. Vortices are very important in understanding flow characteristics such as heat exchange, which is the primary focus for this example. We are interested in the thermal behavior in the structure part in the vicinity of vortical flow. Since the two data parts do not spatially overlap, a selection of vortex regions in the fluid (specified via brushing) needs to be transferred to the neighboring areas in the

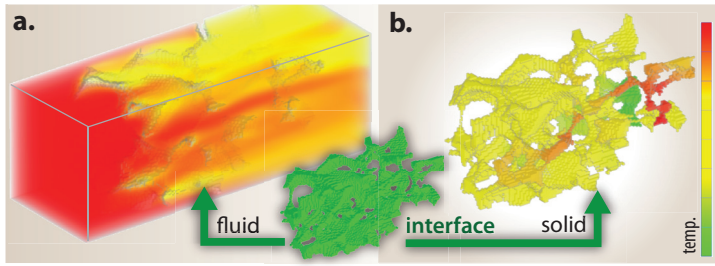



**Figure 1:** The basic structure of the fluid–structure interaction: (a) simulated fluid volume with temperature mapped to color and (b) temperature distribution in the solid part of the data. Both data parts are connected via an interface that relates cells sharing a common face between fluid and solid.

solid part. For this purpose, we construct an interface between the data parts that is similar to the one used in the simulation. The interface is created in advance to the visual analysis, and can be saved and loaded together with the data. Grid cells that are located in the boundary region between fluid and solid are automatically correlated (the technological details are given in Sec. 4). During the visual analysis, user-specified features within these regions are instantly exchanged between the data parts via the interface. The interface can, for instance, be employed to investigate relations between flow phenomena and the resulting temperature changes within the nearby solid.

In Fig. 2a, vortex regions within the fluid part have been selected using the $\lambda_2$ criterion [101]. Color is mapped to the value of $\lambda_2$ with lower values, indicating stronger vortical properties, mapped as red. In Fig. 2b, the $\lambda_2$ selection has been restricted to fluid cells in the vicinity of the aluminium foam using the interface.[2] In order to derive quantitative properties from this selected region, the fluid temperature within it has been averaged and plotted as a green curve over time [236] (see Fig. 2c). Some context is provided by plotting the overall average temperature within the fluid as a black curve and the averaged temperature in the vicinity of the solid as a brown curve (standard deviations are encoded as filled areas in the background). Since the aluminium foam is being warmed by the fluid, the averaged fluid temperature in the vicinity of the foam (brown curve) is lower than the averaged overall fluid temperature (black curve). As indicated by the green and brown curves, it is notable that the fluid temperature close to the solid is warmer when measured in regions of vortical flow.

The next step of the analysis deals with the solid portion of the simulation data. The feature transfer mechanism over the interface works bidirectionally. Thus, it is possible to project the previously defined selection of vortical flow ($\lambda_2$ criterion) onto solid regions in their vicinity. These regions are selected in

[2] The fully selected solid region has been transferred onto the neighboring fluid part where it is combined with the vortex feature.

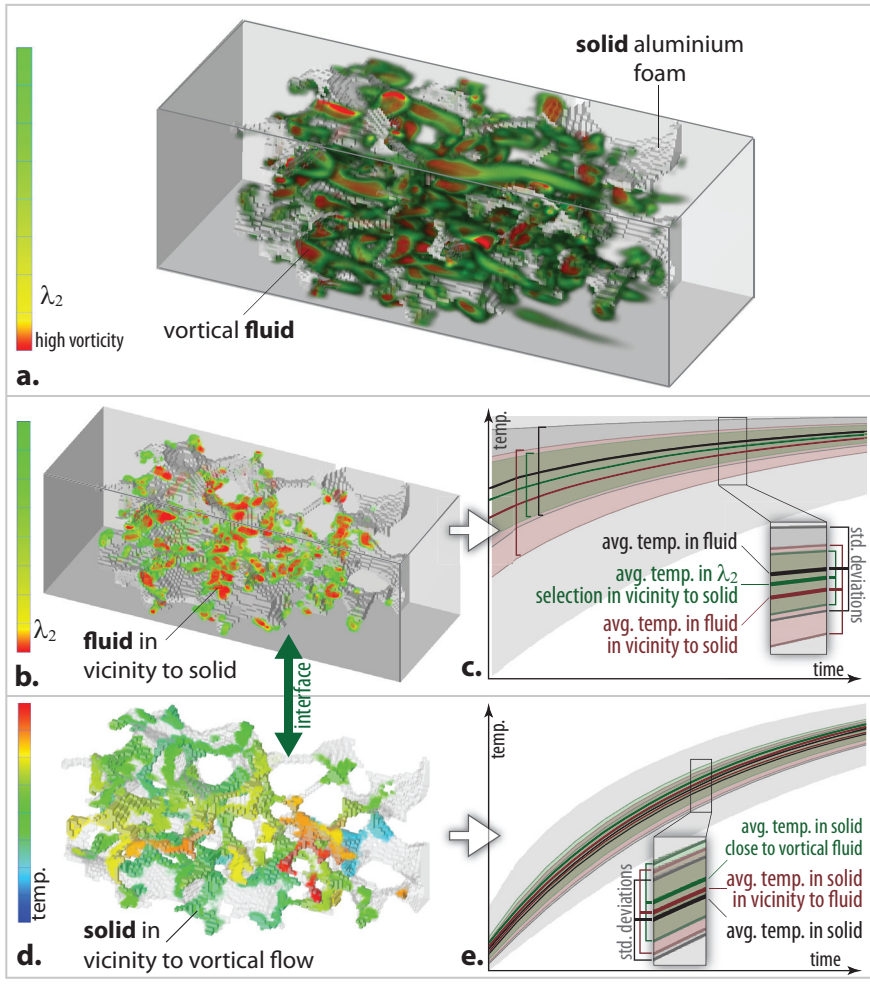



**Figure 2:** Visual analysis of heat transfer using the bidirectional transfer of user-specified features: (a) vortical regions within the flow volume are selected via the $\lambda_2$ criterion [101]. Only fluid regions (b) in the vicinity of the solid and solid regions (d) in the vicinity of vortical flow are visible. In (c, e) statistical properties of selected regions are shown over time.

Fig. 2d, where temperature is encoded in color. The solid portions in the vicinity of vortical fluid (green curve in Fig. 2e) are warmer than the average temperature in the solid (black curve) and also warmer than the remaining solid part in the vicinity of the fluid (brown curve). This is a strong indicator for a direct relation between turbulent flow around the simulated foam structure and the heating process within the structure.

# 4 Interactive Visual Analysis across an Interface

Motivated by the previous example of a fluid–structure interaction, our goal is to enable the joint interactive visual analysis of heterogeneous scientific data. The data consists of two parts (e.g., multi-run and aggregated data or data from a coupled climate model) that are both integrated into the visual analysis. Visual analysis is often based on the concept of user-specified interest per data item (resulting from feature specification via brushing). Such markups represent the first level of semantic abstraction, ranging from data to knowledge [33]. Our idea is to use a synthetic degree-of-interest (DOI) attribution [53] as a common level of data abstraction between two related parts of scientific data. In order to exchange the fractional DOI information, we construct an interface that connects individual grid cells between the data parts (similar to the fluid–structure interaction scenario). Such an abstract coordination space is also implicit in the model-view-controller pattern (see Boukhelifa and Rodgers [18], for instance).

Based on the data state reference model [36], our interface consists of the following four components illustrated in Fig. 3 and described in the following sections:

- the interface describes the *structural relation* between the two data parts (see Sec. 4.1). That is, it specifies which of the grid cells in the one data part are related to certain other cells in the other part, and vice versa. The structural relation can be generated automatically (e.g., in a preprocessing step), and is saved and loaded together with the data parts.
- during the visual analysis, the *transfer of DOI information* represents the functional aspect of the interface. It specifies how the fractional DOI information is exchanged between the data parts (see Fig. 3b and Sec. 4.2). In our fluid–structure scenario, for example, a vortex feature specified in the fluid part is automatically transferred to the solid part where it can be further refined. The feature transfer works in both directions between the data parts.
- the *automatic update of feature specification* represents the dynamic aspect of the interface, which ensures consistency of the features and interactive frame rates during visual analysis. That is, the order in which the DOI information is transferred and updated between the data parts where multiple processes run in parallel (see the arrows illustrating the update process in Fig. 3c).
- we also propose *strategies for visual analysis* across an interface, i.e., the interactive and iterative refinement of features that are specified within and between the two data parts (see Fig. 3d and Sec. 4.4).

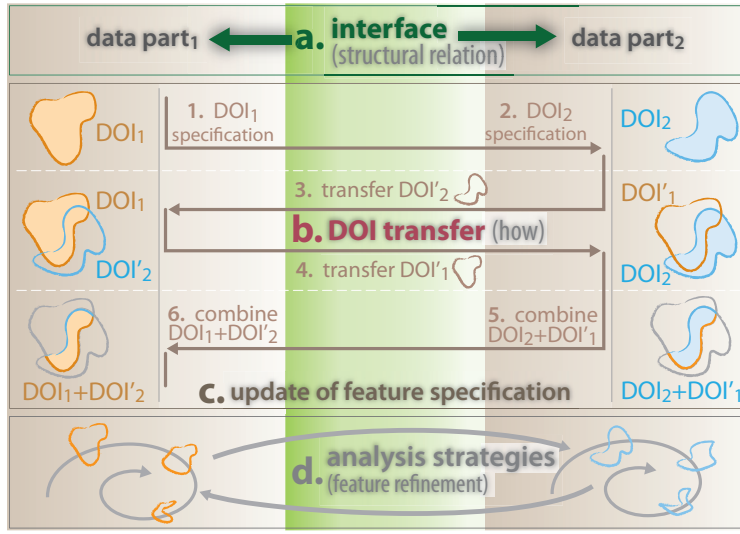



**Figure 3:** In our visual analysis scenario, two parts of the scientific data are connected through an interface: the interface (a) specifies which cells in the two data parts are related to each other, (b) it specifies how the user-specified degree-of-interest (DOI) information is transferred between the data parts. Moreover, it (c) considers dynamic aspects between multiple processes to enable interaction during visual exploration, and (d) enables novel analysis strategies for iterative feature refinement.

## 4.1 The Interface (Structural Relation)

As stated above, the interface specifies the structural relation between the individual grid cells of two parts of the scientific data (see Fig. 3). This relation needs to be generated once for a particular scenario (e.g., in an automatic preprocessing step), and can be saved and loaded. During the visual analysis, the structural relation is then used when transferring features between the data parts. In order to make the interface suitable for different scenarios with heterogeneous data, we need to consider that the two data parts can be given on various kinds of grid, with different data dimensionality, and for possibly different time steps. For all cells in one of the data parts (at a given time step), the interface stores a collection of references to all related cells (and the corresponding time step) in the other part. This allows, for instance, grid cells at a given timestep to be connected to grid cells at multiple time steps, and vice versa (e.g., when the data parts are given for different time intervals). Furthermore, a weight value is assigned to each relation between two cells. This weight determines the amount of influence a related data item has on the item in question. In the FSI scenario, for instance, it may be desirable that fluid and structure cells that are located farther apart have less influence than cells that are relatively close to each other.

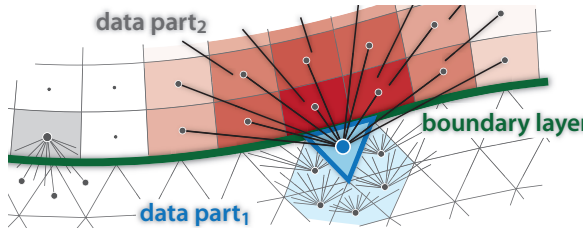



**Figure 4:** Many-to-many relation between two spatially adjoining data parts: a grid cell in one of the data parts can be related to multiple grid cells in the other data part, and vice versa. The weights of the grid cells related to a certain cell (blue) are encoded in red. The different data parts can represent fluid and structure, atmosphere and ocean, or fluid and fluid.

To make the interface as flexible as possible, the relations are separately specified in both directions. In a symmetric scenario, this can also be simplified.

There are three possible ways that data items can be related across different parts of the data [171]: *one-to-one*, *one-to-many*, and *many-to-many*. A one-to-one relation exists also in a traditional multi-variate dataset (single-part scenario) or when different data parts are given for the same grids/time steps. This relation is, therefore, not discussed in further detail here. In the following, we describe the many-to-many relation that exists, for instance, in an FSI simulation. The one-to-many relation is then described in the example of a multi-run scenario.

### Many-to-many relation between two data parts

This kind of relation emerges, for instance, between spatially neighboring data parts such as an FSI simulation. Also in a coupled atmosphere–ocean model simulation, the two models spatially adjoin at the ocean surface and exchange properties through a coupler module (e.g., temperature, precipitation, evaporation). Since the two data parts do not spatially overlap, our approach is to consider the DOI transfer similar to a diffusion process of the features at the boundary between the data parts. This is in agreement, for instance, with the oceanographers' concept of the upper ocean layer that is influenced by the atmosphere (influence is decreasing with depth).

As shown in Fig. 4, the relationship between grid cells sharing a common boundary between the data parts can be translated into a many-to-many interface. The $N$ data items that are close to the boundary layer are connected to $M$ data items which lie in their vicinity in the second data part, and vice versa. As illustrated for the blue grid cell in Fig. 4, the influence of the related grid cells (i.e., the weight values encoded in red) decreases with the spatial distance between the cells.

An interface such as the one used in the fluid–structure interaction example can be automatically constructed as follows (see Fig 4): For every $cell_i$ in data part$_1$

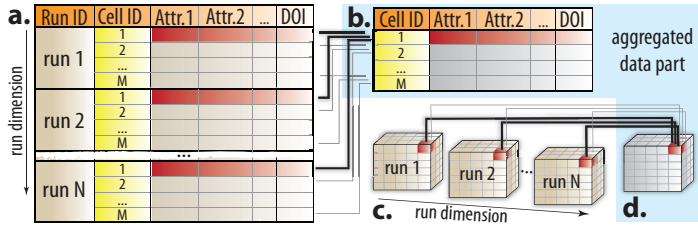



**Figure 5:** One-to-many relation between two data parts with different dimensionality: every $N$ cells in a multi-run data (a, c) are connected to one cell in an aggregated data part (b, d), which share the same space/time (indicated in red).

that is within a certain distance $dist_{\max}$ to the boundary surface, all grid cells in data part$_2$ that are within a distance $dist_{\max}$ to $cell_i$ are added to the collection of related cells. The individual weights for the related cells are, for example, specified as a function of the distance $dist_{i,j}$ between the cells and an importance value of the cell $CI_j$, i.e.,

$$w_j = CI_j \frac{dist_{\max} - dist_{i,j}}{dist_{\max}},$$

where $CI_j$ is usually proportional to the actual volume of the grid cell, giving larger cells a higher influence than smaller ones. In some cases, however, the opposite may be desirable. In simulation, for instance, smaller cells are often used in regions of special interest. In such a case, smaller cells can then receive a higher importance value $CI_j$ than larger cells.

### One-to-many relation between two data parts

This kind of relation exists, for example, between data parts that are specified at two different hierarchical levels. Examples are scale space representations of scientific data where data is given at different resolutions [9] or multi-run and aggregated data that are given with different dimensionality. In the latter case, the higher dimensional data part represents the original multi-run data (with additional independent dimensions for the input parameters to the simulation). In Figs. 5a and 5c, a collection of $N$ values exists for the same data attribute for every grid cell (e.g., 100 temperature values per cell for a simulation with 100 runs). To analyze the distribution of values, statistical properties such as mean or standard deviation can be computed with respect to the run dimension (or another independent data dimension). The result of this aggregation represents the second data part given at a lower dimensionality. In Figs. 5b and 5d, every single cell in the aggregated data part is, therefore, related to the $N$ cells in the multi-run data that share the same space and time, and vice versa.

## 4.2 Transfer of Degree-of-Interest Information

The DOI transfer represents the functional aspect of the interface. It is based on the structural relation between the two data parts (see Sec. 4.1). For every data item $i$ in one data part, the transferred $DOI'_i$ is computed from the related data items in the other part, and vice versa. This transferred DOI information is then combined with the local one in the data part (e.g., logical AND/OR). Since the DOI transfer works bidirectionally, we need to ensure that the transferred feature is not transferred back, which would lead to inconsistencies in the feature specification (see also Sec. 4.3). We propose three different ways of transferring the DOI information: 1) weighted sum, 2) maximum (or minimum) weighted DOI value, and 3) maximum (or minimum) DOI value without weighting. Depending on the user's needs, one can switch between these options during the visual analysis. This opens up interesting opportunities for analytic procedures (see also Sec. 4.4).

With the first approach, the weighted sum of the DOI values of related cells is computed for every data item $i$: $DOI'_i = \frac{1}{\sum_j w_j}\left(\sum_j w_j \cdot DOI_j\right)$. For the one-to-many relation (e.g., when working with hierarchically organized data parts) this represents the transfer of the average of the related DOI values. For the many-to-many relation this kind of DOI transfer can be seen as a diffusion of the DOI information across the interface. For cases in which data is given in a continuous form, the process is similar to integrating over the weighted DOI values of the related cells. This weighted transfer is well suited for examining the degree to which the related cells are part of the focus in the other data part (e.g., 20 out of 100 related cells are selected). However, it has the drawback that isolated DOI features are de-emphasized due to averaging of the DOI values, e.g., when only one of the related cells has a maximum DOI value and all other cells are part of the context.

In order to preserve such DOI peaks, we suggest also to allow the maximum of the weighted DOI values of related cells to be transferred, i.e., $DOI'_i = \max_j(w_j \cdot DOI_j)$. As a third alternative, the user can choose to neglect the weight values, and transfer the maximum value of the related DOI information only. This can be useful, for instance, in order to preserve features even though only a few related cells have large DOI values, or relatively low weights. Examples are grid cells in an FSI scenario or multi-model simulation without considering the actual intercell distance or cell importance. The three methods for DOI transfer are suitable for different stages of a visual analysis, which is discussed in section 4.4.

## 4.3 Automatic Update of Feature Specification

In this section, we describe dynamic aspects of interlinking two data parts. During the visual analysis, the feature specification is automatically updated by multiple threads to ensure consistency and responsiveness of the application.

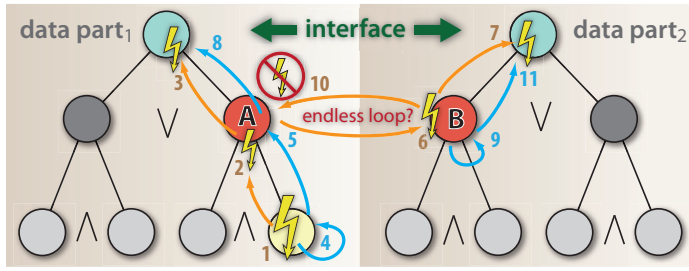



**Figure 6:** Relating complex features that are specified in a hierarchical manner. Yellow arrows represent node invalidations and blue arrows represent updates.

Features can be specified by logical combinations of brushes within and across views. In our framework, the resulting DOI information within an attribute view (e.g., scatterplot, histogram) is represented as a leaf node in a hierarchical feature definition language [52]. The nodes are combined by logical AND/OR-operations in order to specify three levels of focus (see Fig. 6). The different focus levels and the context are encoded in color in every attribute view [161]. The DOI information is thus defined at every node in the feature tree and a flag indicates whether the information is currently up-to-date. As soon as the DOI information at a certain tree node becomes outdated (e.g., when altering a brush in a view), all update processes are suspended. The out-of-date event is propagated up to the tree root (see the flash symbols 1–3 in Fig. 6). Update threads are then restarted, and the feature specification is updated in a depth-first manner, starting with the deepest node in the tree that is out-of-date (steps 4, 5, and 8 in Fig. 6).

The feature trees in two data parts can be related by exchanging the DOI information of two nodes given at the same hierarchy level (e.g., nodes A and B in Fig. 6). The naïve approach is to set the related node B out-of-date after the DOI information in node A is updated (i.e., after step 5)—this is then propagated up to the tree root in data part$_2$ and starts the corresponding update threads. This approach works well as long as the data parts are related only in one direction. If the relation is established in both directions, node A would also be set out-of-date after node B is updated (illustrated in step 10). This would cause an endless loop of updates. To avoid this problem, we do not set node B out-of-date in step 6. Instead a synchronized update of nodes A and B is performed in step 5. Subsequently, only the parent node of B is set out-of-date (step 7).

The sequence of events as to how the related nodes A and B exchange their DOI information is illustrated by arrows in Fig 3c. First, the feature specification in node A ($DOI_1$) and node B ($DOI_2$) is updated, combining the DOI information of the respective child nodes. When exchanging the features via the interface, we need to ensure that the transferred DOI information is not transferred back. This would lead to inconsistencies in the feature specification. In steps 3 and 4 in

Fig. 3c, therefore, the DOI information is first transferred between the data parts (see $DOI'_1$ and $DOI'_2$) and stored temporarily. After that, the transferred DOI can be combined with the local one (steps 5 and 6 in Fig. 3c). During this process, all operations are performed by the threads of only one data part (potential updates of the feature specification in the other data part are suspended).

After nodes A and B have exchanged their DOI information as described above, only the parent node of B is set out-of-date (step 7 in Fig. 6). This restarts the update threads in the feature tree in data $part_2$. Since node B itself has not been set out-of-date, steps 9 and 10—leading to an endless loop—are not performed.

## 4.4 Strategies for Visual Analysis

Interactive visual analysis enables the user to enter a *visual dialog* with the data. The employed procedure usually follows Ben Shneiderman's information seeking mantra [209] (overview first, zoom and filter, details-on-demand) or Keim's recent modification for visual analysis [116] (analyze first, show the important, zoom, filter and analyze further, details-on-demand). The analysis process usually takes place in a single-part scenario. When this is extended to two data parts, the pattern has to be adapted accordingly. Additional iteration loops are introduced between the data parts as illustrated in Fig. 3d. With spatially adjoining data parts, for example, features are iteratively specified in one data part by brushing. The relations of the features—transferred by the interface—are also inspected in the other data part, e.g., in the spatial context using a 3D view or in attribute views (compare to the FSI scenario in Sec. 3). At a certain point, the analysis moves over to the other data part, possibly also with certain iterations, before it can go back to the first data part, and so on.

We have worked through several analysis scenarios with two hierarchically related data parts (Sec. 5 describes one such analysis of multi-run climate data). From these scenarios, we see that it is useful to have views that show the data at the aggregated and detail level next to each other. The analysis usually starts at the aggregated level (overview first). Statistical properties—computed from the data part given with more detail (e.g., the multi-run data)—are investigated at this level. Interesting data characteristics can be selected such as distributions that have a high variability or contain irregularities such as outliers. While interactively brushing the aggregated properties, the collections of related data values are instantly highlighted in another view at the detail level. After several iterations at the aggregated level, the analysis continues in the data part that is given with more detail. The features can be further refined here (e.g., selecting/excluding individual data values that are outliers). The relations are again checked in both data parts, and so on.

An analysis pattern with respect to the DOI transfer is to begin with a maximum transfer first. This is independent of the quantitative influence which the related data items have on each other (e.g., the distance between cells in an

FSI scenario). That is useful, for instance, not to "lose" features in cells with small weight values due to averaging. Such a maximum DOI transfer enables the analyst to look up where features coexist in both data parts. At a certain stage of the analysis, the analyst decides to change to a weighted DOI transfer. This results in a more quantitative analysis of the relations between the data parts, i.e., the degree to which the features coexist. With two hierarchically related data parts (one-to-many), one can investigate how many of the related cells (e.g., in the multi-run data part) are part of the focus. For spatially neighboring parts, the weighted DOI transfer also gives an indication of how close or distant the related cells are. For scenarios with FSI or coupled climate models, this transfer corresponds to the physical properties of a diffusion process.

Another important aspect of related analysis procedures is that data attributes can be transferred across the interface as well (compare to data transformations in the data state reference model [36]). Using an integrated *data calculator* module with a respective graphical user interface, additional data attributes can be derived from existing ones that are possibly located in the other data part. To do so, the structural relation between the data items in the data parts is used (see Sec. 4.1). The new attributes are thereafter available for full investigation in all linked views. We will benefit from this mechanism in the demonstration (Sec. 5), where statistical attributes are derived from multi-run data during the visual analysis.

# 5 Analysis of Multi-run Climate Data

The visual analysis of heterogeneous scientific data is exemplified in the context of a climate data analysis. We investigate data from a multi-run simulation of a palaeoclimatic cold event that was caused by a meltwater outburst from Lake Agassiz, an immense glacial lake located in the center of North America. About 8,200 years ago, the lake drained due to climate warming and melting of the Laurentide Ice Sheet. The investigated data stems from the CLIMBER-2 coupled atmosphere–ocean–biosphere model that simulates a cooling of about 3.6 K over the North Atlantic [10].

With a sensitivity analysis, an important goal for the climate modelers is to better understand the variability of a simulation model with respect to certain model parameters. Identifying those parameters that have the most influence can help to validate the model and also guide future research efforts [70]. Multiple simulation runs are computed with varied initial parameters. In our case, two diffusivity parameters of the ocean model are altered, one horizontal ($diff_h$) and one vertical ($diff_v$), with ten variations each. This leads to a dataset with a total of 100 ($10 \times 10$) runs. For each run, the data is given for 500 years on 2D sections (latitude $\times$ depth) through the Atlantic, Indian, and Pacific ocean. In the following, we present a selection of results from a visual sensitivity analysis

of the ocean part of the CLIMBER-2 model based on the input parameters $diff_h$ and $diff_v$.

## 5.1 Basic Setup for the Visual Analysis

Since the number of independent dimensions in the multi-run ocean data is already challenging (five dimensions, i.e., a 2D section for each ocean, time, and two run parameters with $10 \times 10$ runs), a traditional visual analysis is difficult. Reducing the data dimensionality can help, for instance, by computing statistical aggregates along an independent data dimension. Such an example is to consider averages over time instead of all the individual data values. For the ocean data, we compute statistics with respect to the two run-dimensions. The aggregated data properties are reintegrated in our framework through an attribute derivation mechanism. The result is stored in a separate data part with lower dimensionality than the original data (i.e., a 2D section per ocean over time).

For the visual analysis, we connect the data part that contains the multiple runs and the aggregated data part by an interface. The interface is created automatically during the data conversion and is loaded together with the data parts at the beginning of the analysis session. As discussed in section 4.1, a one-to-many relation is established between each aggregated cell and the collection of multi-run values given for the same space and time (see Figs. 5c and 5d). Brushing, for instance, an aggregated cell also selects the related distribution of values in the multi-run data (at the same timestep). Since the two data parts are connected by the interface, we can go back and forth between the original data and aggregated statistics during the visual analysis.

In the following analysis, we first familiarize ourselves with the data by means of an overview visualization (in the aggregated data part). This is based on glyphs showing derived statistical properties computed from the multi-run data. In the aggregated part, we are able to identify certain cells which contain interesting outliers (with respect to a sensitivity analysis). The selection is automatically transfered via the interface to the multi-run data part. The feature is further investigated and refined, which is also reflected back to the aggregated data part. In the analysis, the parameter settings that lead to the selected outliers can be identified.

First, we want to obtain an overview of the multi-run ocean data. At every timestep, statistical properties are computed from each distribution of multi-run values per grid cell. We are, for instance, interested in distributions where the outputs from different runs have a high variation. For this purpose, we compute the quartile information that is commonly represented in box plots [154]. The three quartiles divide the collection of 100 values per grid cell—one value per run—into four equally populated parts: 25 percent of values are smaller than the *lower quartile* $q_1$, 50 percent are smaller/larger than the *median* $q_2$, and 25 percent of the values are larger than the *upper quartile* $q_3$. The median is a

robust estimate of the center of a distribution (as compared to the mean) and the *interquartile range* ($IQR = q_3 - q_1$) is a more robust estimate for the standard deviation [148]. Carefully designed glyphs, placed as billboards in 3D, can be used to represent multiple properties per grid cell [136].

The glyphs provide qualitative information about the data distribution with respect to the multiple runs. In Fig. 7a, four statistical properties are represented per aggregated cell at timestep 100: the median temperature is encoded in color,[3] the interquartile range is mapped to the overall glyph size, the upper glyph shape represents the distance $q_3 - q_2$, and the lower shape shows $q_2 - q_1$. Large interquartile ranges have been brushed, and opacity represents the respective DOI values. The upper and lower shape of the glyphs are based on super ellipses [136]. Each shape represents an attribute by changing from a star (small value), to a diamond, to a circle, and a box representing a large value (see the glyph legend in Fig. 7a). Even though the figure may contain some visual cluttering, it gives a qualitative overview about the data distribution over all runs (at the given timestep). We see a couple of interesting locations (larger glyphs) where the corresponding distribution of multi-run values have a high variation. The upper/lower glyph shapes also provide information about the skewness of the distribution. Due to its horizontal symmetry, the glyph shape can usually be mentally reconstructed when the glyph is partially occluded. The user can also zoom and rotate the visualization.

Fig. 7b depicts the multi-run data part at the same timestep. For each run, temperature is shown on a cross section through the Atlantic, Indian, and Pacific ocean. The 2D sections (latitude $\times$ depth) are hierarchically arranged next to each other. The two run dimensions of the data are embedded by (re)using one of the spatial dimensions of the visualization (denoted as run axis). The location $r$ along the run axis is determined by the input parameters to the simulations, i.e., $r = \mathit{diff}_h \cdot \mathit{step}_h + \mathit{diff}_v \cdot \mathit{step}_v$, where $\mathit{step}_h$ is chosen slightly larger than $10 \cdot \mathit{step}_v$. This leaves some space between cross sections resulting from different settings for $\mathit{diff}_h$ (illustrated in Fig. 7b). Both step sizes can be specified by the user. During interaction, the camera settings for the aggregated and multi-run view are synchronized.

## 5.2 Outlier analysis in the aggregated data part

As a next step, the influence of the ocean diffusivity parameters on the simulation output is investigated. We focus on grid cells that contain interesting multi-run outliers. These are values resulting from individual runs that strongly diverge from the output of other runs (for the same grid cell and timestep). Identifying such outliers can be useful for finding possible errors in the model or unsuitable

[3] The color maps are based on the work of Brewer [19]. Discrete maps are chosen to allow more quantitative statements about the data.

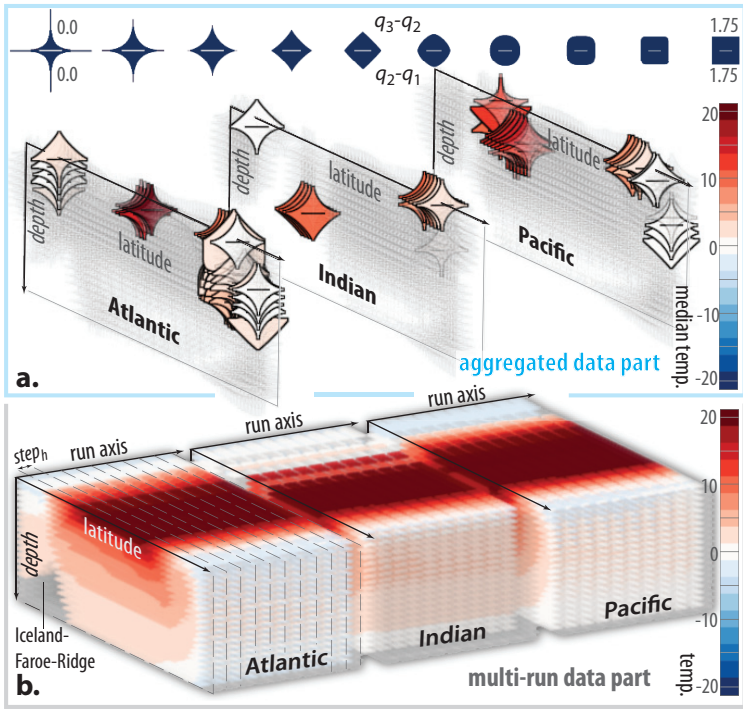



**Figure 7:** Multi-run climate data at timestep 100 given for two hierarchical levels: (a) glyph-based visualization of four aggregated properties from the multi-run data (color, overal size, upper/lower glyph shape). (b) the original multi-run data on 2D cross sections through the Atlantic, Indian, and Pacific ocean. The run parameters are encoded in one of the spatial dimensions (run axis). Camera settings in both views are synchronized.

settings for the model parameters. We compute additional data properties from the multi-run data using the integrated data derivation mechanism of our framework. The resulting properties are stored in the aggregated data part. We create a 2D scatterplot that can answer two questions per multi-run distribution:

- what percentage of the multi-run values given for a grid cell/distribution represent outliers (x-axis), and
- how are the outliers distributed (y-axis). That is, are more outliers located above $q_3$ or below $q_1$, are they equally distributed, etc.

Univariate measures of *outlyingness* often consider the distance of the samples to the data center, normalized by the standard deviation. Such measures can be estimated in a classical or a robust way [148]. Data values that lie more than $1.5 \times IQR$ away from the upper or lower quartile are often considered as "mild" outliers, and values that differ by more than $3 \times IQR$ are considered as "extreme"

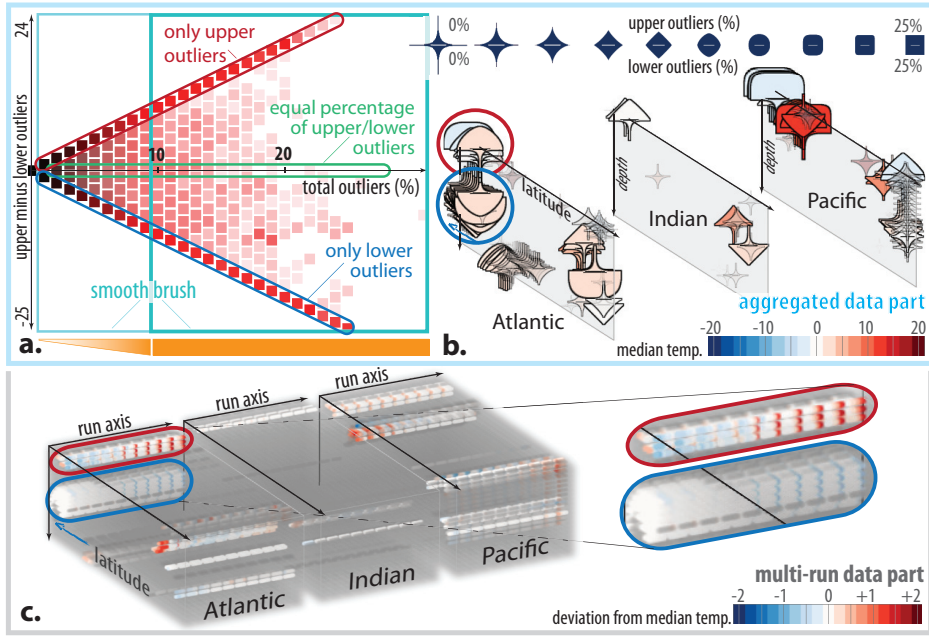



**Figure 8:** Analyzing cells that contain at least 10% of outliers: (a) scatterplot showing the percentage of total outliers (x-axis), and a measure to determine how the outliers are distributed (y-axis), i.e., are more located above $q_3$ (upper outliers) or below $q_1$ (lower outliers). Aggregated outlier properties are depicted using glyphs (b), the selected cells are also shown for the multi-run data (c).

outliers [233]. At this stage of the analysis, we consider mild and extreme outliers as equally important. In Sec. 5.3, however, we treat them differently.

For each distribution of multi-run values at a timestep, we derive the percentage of *upper outliers* (percent data values $\geq q_3 + 1.5 \times IQR$) and *lower outliers* (percent data values $\leq q_1 - 1.5 \times IQR$). The scatterplot in Fig 8a shows aggregated properties for all grid cells and timesteps.[4] The percentage of total outliers per grid cell (at a timestep) is mapped to the x-axis. A measure that expresses whether there are more upper or lower outliers is represented on the y-axis (i.e., upper minus lower outliers). In the view, the number of data items per rectangle is encoded by its luminance and the DOI values are represented by color (pure red represents a maximal DOI value). Grid cells with certain outlier characteristics can be investigated via brushing: Data items at $(0, 0)$ contain no outliers according to the chosen measure. Items along the diagonals contain either only

[4]Since the point size in this plot has been increased, it is similar to a 2D histogram using colored rectangles to represent the bar height.

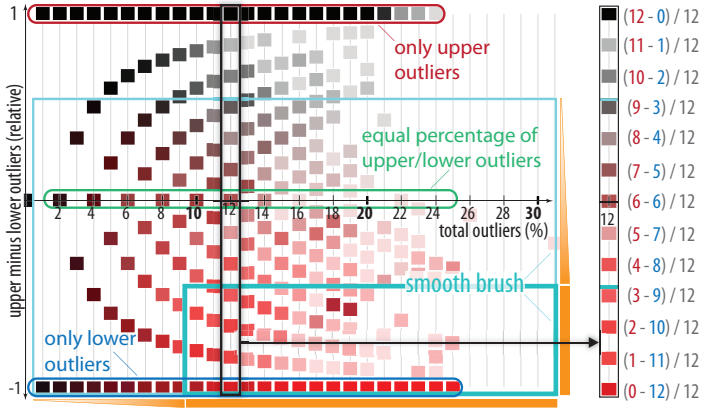



**Figure 9:** Distributing the data items from Fig. 8a uniformly on the vertical axis supports brushing of certain outlier characteristics. Grid cells are selected that contain at least 10 percent of outliers of which at least 75 percent are lower outliers. For the example of 12 percent total outliers, the possible distributions of upper minus lower outliers (red and blue number) is shown.

upper or lower outliers. Items located on the x-axis ($y = 0$) contain the same number of upper and lower outliers. Using a smooth brush [53], we focus on grid cells where more than 10 percent of the multi-run values diverge strongly from the rest (with a transition to cells containing no outliers, illustrated as an orange gradient below Fig. 8a). While brushing these aggregated characteristics, the selection is instantly transferred to the multi-run data part via the interface. The spatial relation of the feature can be investigated in Figs. 8b and 8c.

The glyphs in Fig. 8b depict the derived outlier characteristics at timestep 60. Color represents the median temperature and the overall glyph size represents the percentage of total outliers per cell (at the timestep). The upper and lower glyph shape shows the percentage of upper and lower outliers, respectively. In Fig. 8c, the corresponding deviation of multi-run values from the median temperature is visualized. A group of cells with mainly upper outliers (round upper glyph shape) is visible in the north of the Atlantic (see red ellipses in Figs. 8b and 8c). Another group of cells with many lower outliers is located north of the Iceland-Faroe-Ridge in the Atlantic (see the blue ellipses). By changing the depicted timestep, one can observe that the feature with lower outliers propagates northward and downward near the seabed over time. The feature also extends over the north pole to the other parts of the arctic sea (not visible at this timestep). At a later stage of the simulation, an increasing number of runs results in such lower (cooler) outliers compared to the rest (blue ellipses in Figs. 8b and 8c). We further investigate this feature.

We focus on cells that contain more lower than upper outliers. To allow such a relative selection, the data mapped to the y-axis in Fig. 8a is normalized. The respective data attribute (upper minus lower outliers) is, therefore, divided by the corresponding percentage of total outliers (x-axis). The resulting scatterplot is shown in Fig. 9. For each column of total outliers (x-axis), the combinations of upper and lower outliers are now equally distributed on the vertical axis (this is illustrated for the example of 12 percent total outliers in Fig. 9). Accordingly, it is now possible to brush the ratio between upper and lower outliers. Data items that 1) contain at least 10 percent of outliers at a timestep (x-axis) and 2) have at least 75 percent lower outliers—compared to the percentage of upper outliers—are in full focus (see also the smooth extension of the brush where the DOI linearly decreases, illustrated as orange gradients in Fig. 9). The respective feature is further analyzed in the following section.

## 5.3 Outlier analysis in the multi-run data part

Up to now, our analysis was mainly based on aggregated properties. Since both data parts are connected through an interface, we can go back to the original multi-run data and further refine our selection of lower outliers. In the following, the model sensitivity with respect to the input parameters $diff_h$ and $diff_v$ is investigated for the specified feature. Our goal is to identify 1) the grid cells with the specified outlier characteristics and 2) the parameter settings that result in such outliers. A measure of outlyingness is thus derived in the multi-run data part, which also allows us to differentiate between mild and extreme outliers. For each multi-run value $x_j$, the deviation from the center of the corresponding distribution is normalized by the interquartile range, i.e., $\frac{x_j-(q_1+q_3)/2}{IQR}$. Values inside $[q_1, q_3]$ are thereby mapped to the interval $[-0.5, 0.5]$. Note that the median does not have to be zero on this scale.

The scatterplot in Fig. 10 shows the described measure of outlyingness for the multi-run data (y-axis), and the corresponding deviation from the median temperature per distribution (x-axis). We brush multi-run values that represent extreme outliers with a smooth transition to mild outliers (see the illustration on the right of Fig. 10). In the scatterplot, such extreme outliers are vertically located above or below $\pm 3.5$ and deviate by more than $3 \times IQR$ from the upper or lower quartile, respectively. Mild outliers are located above or below $\pm 2.0$. In Fig. 10, different levels of focus+context [161] are discriminated in color: the context is shown in black, data items only selected in the local view are encoded in blue, and items selected in both data parts are highlighted in red. Since the interface works bidirectionally, the maximum DOI value per multi-run distribution is also transferred to the related grid cell in the aggregated data. Aggregated cells where the related distribution contains only mild outliers accordingly receive a low DOI value.

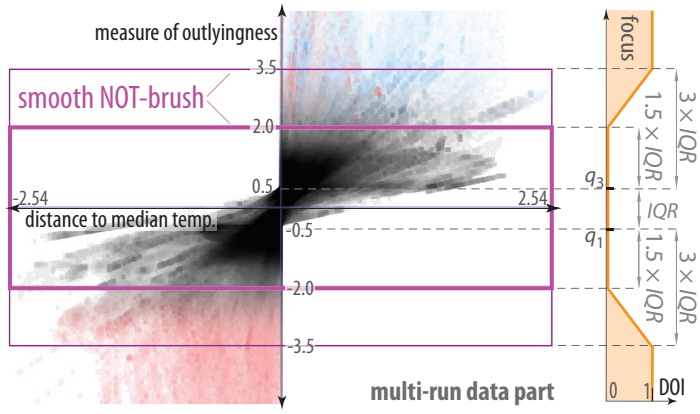



**Figure 10:** Outliers are brushed using derived attributes in the multi-run data part: mild outliers are vertically located above or below $\pm 2.0$ and extreme outliers are located above or below $\pm 3.5$. Features selected in multiple views are highlighted in red (focus), features only selected in the current view are depicted in blue, and context information in black.

As a next step, multi-run values that are relatively similar to the median temperature of the corresponding distribution are excluded from the selection (with a smooth brush on the x-axis in Fig. 10, not shown here). This is to account for distributions with a very small interquartile range, where the chosen measure of outlyingness becomes less significant (as compared to larger interquartile ranges).

In the following, we investigate the temporal evolution of the previously specified feature of lower outliers (aggregated data part) that has been refined in the multi-run data part to also identify the parameter settings causing these outliers. Fig. 11 shows the aggregated and multi-run data at three different timesteps, represented as columns. The aggregated outlier properties are visualized in the top row. A diverging color map is used for the multi-run data (middle row) to encode the deviation from the median temperature per grid cell. A view from above (bottom row) is used to identify the input parameter settings resulting in the selected outliers ($diff_v$ settings are also color-coded).

At timestep 60, the selected multi-run values strongly deviating from the rest of the distribution are visible north of the Iceland-Faroe-Ridge (see the blue ellipses in Figs. 11b and 11c). Since the corresponding $diff_v$ and $diff_h$ settings are spatially encoded in the visualization, we can see that these outliers mainly result from larger $diff_v$ settings[5] (see the inset in Fig 11c). At this stage of the simulation, multi-run values that are also simulated with larger values for the horizontal ocean diffusivity ($diff_h$) deviate earlier from the other runs. For

[5] These runs are located on the right side each, because $diff_v$ input values are encoded with a smaller step size than $diff_h$ values.

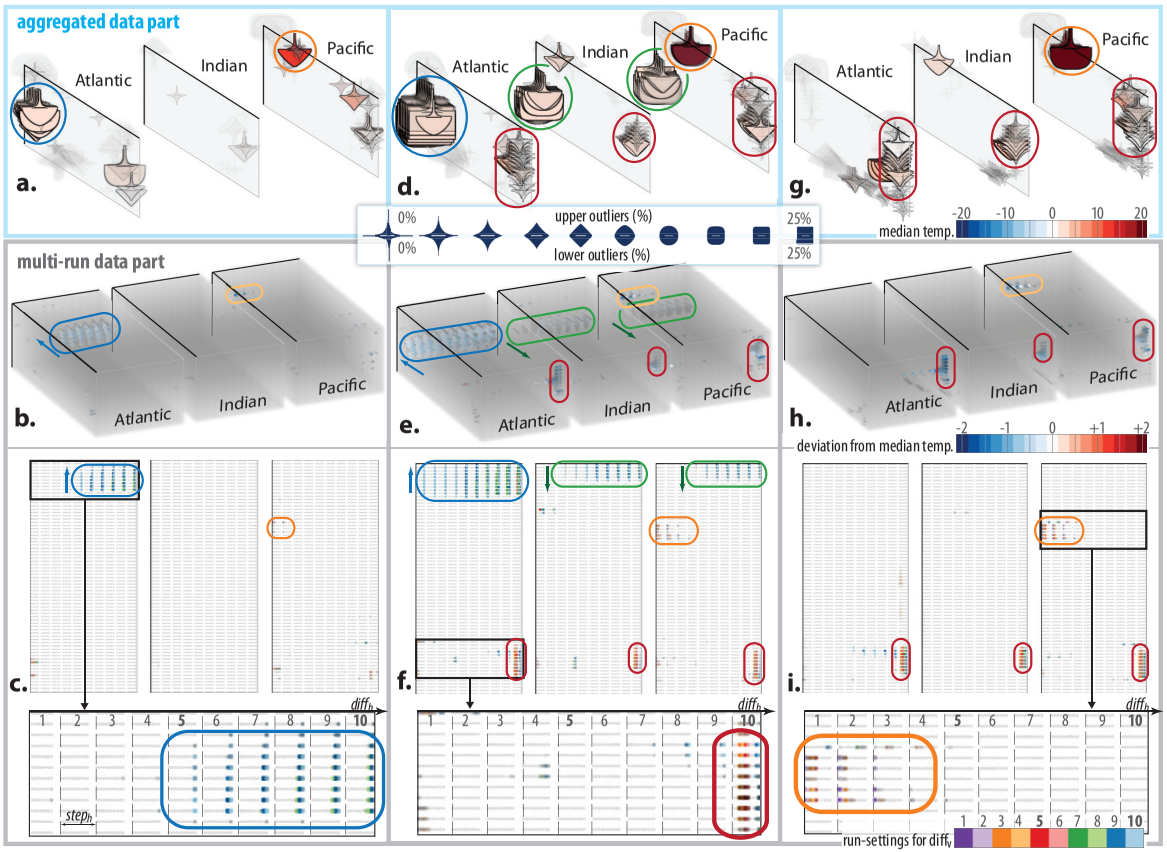



**Figure 11:** Investigation of the specified feature (multi-run and aggregated data) at timestep 60 (left), timestep 120 (middle), and timestep 250 (right). In the top row, four derived properties are visualized (median temp., percentage of upper/lower outliers, and percentage of total outliers). The individual runs that result in outliers are visible in the 3D context (second row), and in a view from above (last row). Here, the two run parameters are embedded by (re-)using one of the spatial dimensions in the visualization ($diff_v$ is also color coded).

these grid cells and run settings, the model changes from its standard behavior to another climate condition.

At year 120 the feature of lower outliers (blue ellipse in the north of the Atlantic) has also propagated to the other parts of the arctic sea (green ellipses in the Indian and Pacific ocean). After extending over the north pole, it is propagating southwards (indicated with arrows). It is still only larger settings for $diff_v$ that produce these outliers (see upper part of Fig 11f). On the other hand, a condition has established in the southern region in all three oceans where a few runs constantly result in different output values to the rest (see red ellipses). These outliers result from large $diff_h$ settings and are, therefore, represented to the right of each ocean. A similar behavior is also visible at year 250 (see the red ellipses in Figs 11h and 11i). At this stage of the simulation, the outliers previously visible in the north have already disappeared. A condition has been established in the north where the runs mainly result in similar outputs.

Over the investigated timespan, certain runs in the northern Pacific also produce a larger number of lower outliers (see the orange ellipses in Fig. 11). These outliers result mainly from smaller settings for $diff_h$ and $diff_v$ (e.g., see the inset in Fig. 11i). As a next step, we change our selection in the aggregated data part to select grid cells that contain more upper than lower outliers. A similar analysis is performed, where the parameter settings producing these upper outliers are investigated. Due to space limitations, this is not shown here.

In summary, we performed a visual sensitivity analysis of a multi-run climate simulation. In our analysis framework, multi-run and aggregated data were integrated and related by an interface, which supports the investigation of features across both data parts. Statistical properties were computed from the distributions of multi-run values. Based on these properties, interesting outlier characteristics could be brushed in the aggregated data part. The feature was automatically transferred to the multi-run data via the interface where it was further investigated. Individual runs that substantially deviate from the other values of the distribution could be identified together with the corresponding input parameter settings. By connecting both data parts via the interface, the analyst can go back and forth between multi-run and aggregated data, which enables a powerful analysis.

# 6 Conclusion and Future Work

The joint visual analysis and exploration of heterogeneous scientific data is a crucial and challenging task. In this paper, we propose a systematic approach to the interactive visual analysis of two heterogeneous parts of scientific data. Analogous to the related simulation scenarios, we construct an interface between the data parts which connects data items in the one part to data items in the other, and vice versa. We propose different ways of how a user-specified degree-of-interest

attribution can be transferred between the data parts. Instead of performing fusion between the parts at the data level—this is often not practical in scenarios including multi-run simulation data or fluid–structure interactions—we perform the fusion on the first semantic/interpretation level explicitly represented as user-specified features [33]. Our approach is demonstrated in two visual analysis scenarios with heterogeneous scientific data, which were conducted in collaboration with domain researchers.

For data parts specified at hierarchically different levels, the integration of derived statistical attributes in the analysis process has shown great potential. It enables the analyst to work simultaneously in both—the data part containing the actual data, and the aggregated data part representing summary information. The analyst can go back and forth in an iterative manner, analyzing the data at different hierarchical levels. Relations between these data parts can thereby be identified through the visualization and iteratively refined. Such a tight integration of a computational and interactive analysis methodology agrees well with the requirements for prototypic visual analytics solutions [116].

In future work we will focus on extending our approach to scenarios with multiple data parts (e.g., given at multiple aggregated levels). We also aim at further integrating statistical properties, yielding quantitative results into our visual analysis framework. Here again, we want to show how visual analysis and statistics can interact in a feedback loop to gain in-depth insight into the data. We also want to identify further analytical patterns that involve our interface.

## Acknowledgments

The authors thank Thomas Nocke, Michael Flechsig, and colleagues from the Potsdam Institute for Climate Impact Research, Germany, for fruitful discussions, valuable comments on this paper, and for providing the climate simulation data. We thank Andreas Lie for implementing the glyph renderer and Stian Eikeland for the data conversion (both from the Univ. of Bergen). We also thank Matthew Parker (Univ. of Bergen), Brendan McNulty (Allegro Language Service, Bergen), David Horn, and our anonymous reviewers for their valuable comments that helped to improve this paper. The CFD data is courtesy of Innovative Computational Engineering GmbH (www.ice-sf.at), Leoben, Austria. This work was supported in part by the Austrian Research Funding Agency (FFG) in the scope of the projects “AutARG” (No. 819352) and “PolyMulVis” (No. 823855).

# Paper D

# Brushing Moments in Interactive Visual Analysis

Johannes Kehrer,[1] Peter Filzmoser,[2] and Helwig Hauser[1]

[1]Department of Informatics, University of Bergen, Norway

[2]Department of Statistics and Probability Theory, Vienna University of Technology, Austria

## Abstract

We present a systematic study of opportunities for the interactive visual analysis of multi-dimensional scientific data that is based on the integration of statistical aggregations along selected independent data dimensions in a framework of coordinated multiple views (with linking and brushing). Traditional and robust estimates of the four statistical moments (mean, variance, skewness, and kurtosis) as well as measures of *outlyingness* are integrated in an iterative visual analysis process. Brushing particular statistics, the analyst can investigate data characteristics such as trends and outliers. We present a categorization of beneficial combinations of attributes in 2D scatterplots: (a) $k^{\text{th}}$ vs. $(k+1)^{\text{th}}$ statistical moment of a traditional or robust estimate, (b) traditional vs. robust version of the same moment, (c) two different robust estimates of the same moment. We propose selected view transformations to iteratively construct this multitude of informative views as well as to enhance the depiction of the statistical properties in scatterplots and quantile plots. In the framework, we interrelate the original distributional data and the aggregated statistics, which allows the analyst to work with both data representations simultaneously. We demonstrate our approach in the context of two visual analysis scenarios of multi-run climate simulations.

This article was published in *Computer Graphics Forum*, 29(3):813–822, 2010. Digital Object Identifier no. 10.1111/j.1467-8659.2009.01697.x. The work was also presented by the main author at EuroVis 2010, June 9–11, Bordeaux, France. In Fig. 3, the annotation of view transformation has been changed ($\mathcal{T}_{ord}$ instead of $\mathcal{T}_{rob}$). An error in the caption of Fig. 5 has been corrected.

# 1 Introduction

The increasing complexity of modern scientific data (from measurements and computational simulations) presents us with new challenges for data analysis. Traditional approaches are often based on the *a posteriori* discussion of expressive statistical properties of the data. Interactive visual analysis, as addressed in this paper, allows the iterative exploration and analysis of data in a guided human–computer dialog. Simple but effective visualization techniques are used in combination with proven interaction schemes such as linking and brushing. This enables a powerful information drill-down process [209]. Visual analysis uses proven concepts such as coordinated multiple views, interactive feature specification via brushing, focus+context visualization, and on-demand data derivation [73].

In many cases, multi-dimensional scientific data can be denoted as $f_d(\mathbf{p})$ where data values $f_d$ (e.g., temperature, pressure values) are measured or simulated with respect to an $m$-dimensional data domain $\mathbf{p}$. The domain (i.e., the independent data dimensions) can be 2D or 3D space, time, but also independent input parameters to a simulation model. In climate research or engineering, for instance, so-called *multi-run* simulations have recently become an important approach to assess simulation models [86, 151]. The input parameters of the simulation are varied and a simulation output is computed for each variation of the parameters (or at least many of them). This leads to a collection of values that exists at every space/time location [141] (one value for every run). Multi-run data is analyzed to assess the variability of the simulation model and to better understand how sensitive the model reacts to a variation of its input parameters (*sensitivity analysis*). Identifying those parameters that have the most influence can help to validate the model and also guide future research efforts [70].

The analysis of high-dimensional data is generally quite challenging, especially if the number of independent dimension is larger than two/three. Reducing the data dimensionality is a natural attempt in such a situation, e.g., by computing statistics along selected independent data dimensions. Such an example is to consider averages over time instead of all the individual data values. In this paper, we demonstrate that it is useful to integrate statistical properties in an interactive visual analysis process. Such an integration opens up the possibility of new informative views on the data as well as opportunities for advanced visual data analysis.

When analyzing data distributions, trends and outliers are often of special interest. The four statistical moments are suitable for describing data trends (with respect to centrality and variance) as well as the shape of the distribution (skewness and kurtosis) [148]. These data characteristics can be estimated traditionally or in a robust way [60, 121]. Additionally, measures of *outlyingness* help to identify extreme observations that substantially deviate from the rest [148]. These interesting opportunities to analyze data distributions, however, also gen-

erate a "management challenge" for the analyst: what perspective is best for a particular analysis task?

In this paper, the integration of traditional and robust statistical moments in the visual analysis is discussed in a structured form. We propose a set of generic *view transformations* that allow the iterative construction of a multitude of informative views, based on these statistics. The transformations lead to a classification scheme for possible attribute/axis configurations in 2D scatterplots. In the analysis framework, we relate the original data—the individual data items from which the statistics are computed—and the derived statistics to each other. Thus, the analyst can work with both data representations simultaneously. Data trends and outliers can be investigated by brushing statistical properties in multiple views, by iteratively altering the depicted view attributes, and by deriving new data attributes on demand.

## 2 Related Work

Visualization and statistics facilitate the understanding of relevant characteristics of complex datasets and there is a long history of related work [232]. Interestingly, the slightly younger history of visualization research relates back to early works that were inspired by considerations from statistics [233, 30, 39, 38]. Even systems for the visual data exploration can be traced back to these [248, 219, 222]. So there is a long history of relations between statistics and visualization.

The area of *coordinated multiple views* has been steadily developing over the past fifteen years (see Roberts [192] for an overview). WEAVE [66] and SimVis [52] are just two examples for according visual analysis frameworks for scientific data. Multiple linked views are used next to each other to concurrently show, explore, and analyze multi-variate data. This includes 3D views of volumetric data (grids, also over time), but also attribute views such as 2D scatterplots, histograms, function graph views, or parallel coordinates. Interesting subsets of the data are interactively selected (brushed) directly on the screen, the relations are investigated in other linked views (compare also to the XmdvTool [248]). Logical combinations of brushes in multiple linked views enable the specification of complex features [52, 251]. The selection information is used to visually discriminate the specified features from the rest of the data in a focus+context visualization style [72].

The treatment of *multi-run data* is rather new to the visualization community [141]. Information visualization techniques (e.g., parallel coordinates, scatterplot matices) are used in combination with statistics, to improve the understanding of the model output from multi-run simulations [41]. Nocke et al. [165, 166] propose a coordinated multiple views system to analyze a large number of tested model parameters and simulation runs. Statistical aggregations of the multi-run data are visualized, e.g., using linked scatterplots, graphical tables, or parallel

coordinates. In recent work, Matković et al. [151] visualize multi-run data as families of data surfaces (with respect to pairs of independent dimensions) in combination with projections and aggregation of the data surfaces.

Kao et al. [108] visualize data distributions over 2D multi-run data, where the distributions can apparently be represented by statistical parameters. For other cases, they propose a shape descriptor approach [107] constructing a 3D volume with the probability density function (PDF) of the data as voxel values. Mathematical and procedural operators [141] are proposed to transform multi-run data into a form where existing visualization techniques are applicable (e.g., pseudo-coloring, streamlines, or isosurfaces). This approach is very promising due to its flexibility. However, it is not integrated in a visual analysis framework that would enable to interactively specify and investigate features within the transformed data attributes.

Recently, Patel et al. [178] visualize moments that describe the distribution of values in a growing neighborhood around a voxel. The resulting curves enable the specification of a transfer function with improved discriminative properties in volume rendering. Others [162, 173] exemplify that the integration of selected data analysis mechanisms (such as principal component analysis, PCA) can support the visual analysis of scientific data.

Finally, the interesting work by Weaver [251, 252] demonstrates the value of a structured discussion of selected aspects of visual analysis approaches. Different opportunities for visual data analysis are analyzed, providing an ordered guide to a multitude of opportunities. With our paper, we provide such a guide to the rich space of opportunities of moments-based interactive visual analysis of scientific data.

## 3 Statistical Background

Statistical moments describe important characteristics of data distributions. The first two moments refer to the central tendency (mean $\mu$) and the variability or dispersion (variance $\sigma^2$). The third and fourth standardized moment characterize the asymmetry (*skewness*) and the peakedness (*kurtosis*) of a distribution, respectively. For a distribution of samples $\{x_1, \ldots, x_n\}$, the first moment can be estimated by the arithmetic mean $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and the second moment by the empirical variance $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$. Skewness can be estimated as $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3/s^3$, and kurtosis as $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4/s^4 - 3$. The subtraction of the constant 3 results in a kurtosis value of zero in case of normally distributed data. Although these classical estimators are very useful in practice, they have to be applied with care. For datasets including outliers the results can be misleading because outliers can have an arbitrarily large influence on these estimators. In the following, we recapitulate more robust estimators for the moments as well as measures of outlyingness that are integrated in our approach.

**Robust estimates of statistical moments:** The *median* is a robust estimate of the center of a distribution. It is a special case of a sample *quantile* [90], which is a value $q(p)$ such that at least $np$ of the observations are $\leq q(p)$ and at least $n(1-p)$ observations are $\geq q(p)$ where $p \in [0,1]$. This gives the three *quartiles* that are, for example, used in box plots [233]: the lower or first quartile $q_1 = q(\frac{1}{4})$, the median or second quartile $q_2 = q(\frac{1}{2}) = \text{med}(x_1, \ldots, x_n)$, and the upper or third quartile $q_3 = q(\frac{3}{4})$. Robust estimates for the standard deviation are the *interquartile range* $\text{IQR} = 0.741 \cdot (q_3 - q_1)$ and the *median absolute deviation*

$$\text{MAD}(x_1, \ldots, x_n) = 1.483 \cdot \text{med}_{1 \leq i \leq n}( \, |x_i - q_2| \, ) \tag{1}$$

from the distribution's median $q_2$. Using the constants 0.741 and 1.483, respectively, allows for a consistent estimation of the standard deviation $\sigma$ of a normal distribution.

Two robust descriptors of the shape of the distribution are the *octile-based* skewness $\text{skew}_{oct}$—this is a special case of a quantile-based skewness coefficient [87]—and an octile-based kurtosis measure $\text{kurt}_{oct}$ [157]:

$$\frac{e_7 + e_1 - 2e_4}{e_7 - e_1} \quad \text{and} \quad \frac{(e_7 - e_5) + (e_3 - e_1)}{e_6 - e_2} - 1.23 \tag{2}$$

where $e_i = q(\frac{i}{8})$ is the $i^{\text{th}}$ octile. Alternative robust measures for skewness ($k = 3$) and kurtosis ($k = 4$) can be obtained by replacing the classical estimates of mean and standard deviation by the robust versions median/MAD [60]:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{(x_i - \text{med}(x_1, \ldots, x_n))^k}{\text{MAD}(x_1, \ldots, x_n)^k} - c_k. \tag{3}$$

As for the classical estimates, the $k^{\text{th}}$ moments ($\text{skew}_{\text{MAD}}$ and $\text{kurt}_{\text{MAD}}$ for $k = 3, 4$) are made comparable to the normal distribution, and thus $c_3 = 0$ and $c_4 = 3$ [60]. While the octile-based skewness and kurtosis coefficient (see Eq. 2) aim to minimize the influence of outliers on the measure, the median/MAD-based moments ($\text{skew}_{\text{MAD}}$, $\text{kurt}_{\text{MAD}}$ in Eq. 3) still include such outliers. Therefore, $\text{kurt}_{\text{MAD}}$ can also be used to identify distributions that contain outliers. If the samples are approximately normally distributed, the median/MAD-based measures yield values close to zero.

**Measures of outlyingness:** Outliers and their identification are of special interest in many practical applications. Univariate measures of outlyingness often consider the distance of the samples $x_i$ to the data center, normalized by the standard deviation. Both center and standard deviation can be estimated in a classical or a robust way. This leads to the classical and the median/MAD-based z-score [148]:

$$\text{z} = \frac{x_i - \overline{x}}{s} \quad \text{and} \quad \text{z}_{\text{MAD}} = \frac{x_i - \text{med}(x_1, \ldots, x_n)}{\text{MAD}(x_1, \ldots, x_n)}. \tag{4}$$

For normally distributed samples, both the classical and the robust z-scores yield values in the interval $[-2, 2]$ for about 95% of the data points. Accordingly, approximately 5% of the samples are identified as potential outliers. For distributions including outliers, only the robust version lead to a reliable tool for outlier identification [188].

# 4 A Moment-based Scheme for Visual Analysis

Descriptive statistics characterize the main features of a distribution of values. The integration of such statistical properties into a visual analysis provides interesting opportunities [223]. However, there is a multitude of alternatives when mapping, for instance, two statistical properties (such as moments) to a scatterplot. Which of the four moments should be plotted against each other? Should a traditional or robust estimate be used? Should some kind of data transformation (such as normalization or scaling) be applied?

In this section, we present a classification scheme for possible combinations of moment-based statistical properties in views. This scheme is constructed by a set of view transformations that are applied consecutively to the attributes mapped to scatterplots. We show how a large set of informative views—including known statistical plots such as the Q–Q (quantile–quantile) plot [254] or the spread vs. level plot [233]—can be constructed iteratively by such view transformations. For illustrative purposes, we start with an example analysis of multi-run climate data. In Sec 4.2, four types of view transformations are described. The resulting view classification scheme is presented in Sec. 4.3.

## 4.1 Illustrative Example of Multi-run Climate Data

Climate research is concerned with the analysis of the climate system, its variability, and long-term behavior [246]. To allow better predictions of future events, it is important to understand the past. The CLIMBER-2 coupled atmosphere–ocean–biosphere model simulates a palaeoclimatic cold event [10]. The anomaly was caused by a meltwater outburst from Lake Agassiz, an immense glacial lake located in the center of North America. About 8,200 years ago, the lake drained due to climate warming and melting of the Laurentide Ice Sheet. The CLIMBER-2 model simulates a cooling of about 3.6 K over the North Atlantic induced by a meltwater outflow into the Hudson strait [10].

We analyze a multi-run simulation of the ocean part of the CLIMBER-2 model. With such an analysis, an important goal for climate modelers is to better understand the variability of a model with respect to certain model parameters (sensitivity analysis [70]). Multiple simulation runs are computed with varied initial parameters. In our case, two diffusivity parameters of the ocean model are altered, one horizontal ($diff_h$) and one vertical ($diff_v$), with ten variations each.

The simulation leads to a dataset with a total of 100 ($10 \times 10$) runs. For every run, the data is given for 500 years on 2D sections (latitude $\times$ depth) through the Atlantic, Indian, and Pacific ocean.

**Basic Setup for the Visual Analysis:** Since the number of independent dimensions in the multi-run ocean dataset is already challenging (5 dimensions, i.e., $3 \times$ 2D sections, time, and two run parameters with $10 \times 10$ runs), a traditional visual analysis is difficult. Reducing the dimensionality can help, for instance, by computing statistical aggregates along independent data dimensions such as time or a spatial axis. For the ocean data, we compute statistics with respect to the run-dimensions. The aggregated properties are reintegrated in our framework through an attribute derivation mechanism. The result is stored in a separate data part with fewer independent dimensions (i.e., $3 \times$ 2D sections over time).

In practice, often only the aggregated data is further analyzed using statistical tools and static visualizations [45, 82]. However, we integrate both the multi-run and aggregated data part in an interactive visual analysis process where they are related to each other. A one-to-many relation [171] is established between an aggregated cell $ac_j$ and the distribution of multi-run values $\mathbf{x}_j = \{x_{1,j}, \ldots, x_{100,j}\}$ given for the same space/time. Both data parts can exchange selection information, i.e., brushing an aggregated cell $ac_j$ selects also the related distribution $\mathbf{x}_j$ in the multi-run data. Fig. 1 on page 128 shows such distributions (highlighted in color) that were selected in the aggregated data part (not shown here).

A so-called *quantile plot* is shown in Fig. 1a for the multi-run data. The sample quantiles $q_j(p)$ of each distribution of temperature values $\mathbf{x}_j$ are plotted on the y-axis with respect to a parameter $p \in [0, 1]$. Traditionally, only a small number of distributions are depicted in such a plot. Using a focus+context style, however, we are able to look at all distributions in the multi-run data. For each location in space/time, the multi-run values of the corresponding distribution are represented as a sequence of points monotonically extending from the left to the right. Brushing statistical properties in the aggregated data facilitate the identification of interesting distributions in the quantile plot. Distributions with a substantially negative kurtosis measure are highlighted in green, and distributions with a high standard deviation are shown in red. Two brushes were used for selection in the aggregated data part. To make the individual distributions in Fig. 1a comparable to each other, we can apply selected transformations on the view.

## 4.2 Generic View Transformations

View transformations can be seen as an extension to classical data transformations. They facilitate the interaction with views during visual analysis and help the analyst to maintain a mental model of the utilized views and their depicted attributes. Starting from a generic view $v$, its appearance is consecutively altered

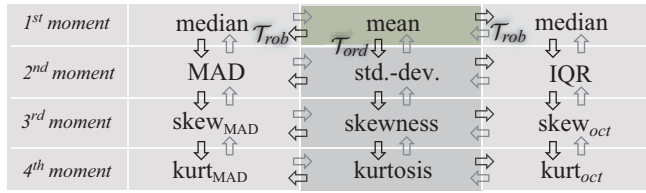



**Table 1:** Traditional and robust estimates of moments: the table is constructed starting from the mean, applying order increasing and robustifying view transformations.

by applying a view transformation $\mathcal{T}$, i.e., $v' = \mathcal{T} \circ v$. Consequently, a large set of informative views can be constructed. The progressive refinement of views using transformations complies with the iterative nature of a visual analysis (compare to the visual analytics mantra [116]). The transformed version of a view can either be used additionally, or it can replace the original view. We propose four types of view transformations to construct our classification of moment-based views (presented in Sec. 4.3). The two main types allow us to switch between the four moments, and their robust and traditional estimates:

- an **order transformation** $\mathcal{T}_{ord}(t_{ord}, m)$ is used to increment or decrement the $k^{\text{th}}$ statistical moment $m$ shown in a view (dependent on the type $t_{ord} : k \to (k \pm 1)$);
- a **"robustifying" transformation** $\mathcal{T}_{rob}(t_{rob}, b)$ chooses a traditional or robust estimate of a moment $m$, depending on the type $t_{rob}$; we provide two robust alternatives per moment, estimates based on quartiles/octiles and others based on the median/MAD.

Order and robustifying view transformations represent the most important construction elements for our view classification scheme. They are used to create the entries in table 1. For practical situations, we provide "shortcuts" to all twelve measures in addition to the respective transformations.

We propose two additional types of view transformations for the analysis, which are closer related to classical data transformations (e.g., normalization, z-standardization):

- a **relating transformation** $\mathcal{T}_{rel}(t_{rel}, a, b)$ that sets a view axis $a$ in relation to a data attribute $b$; dependent on the type $t_{rel}$, for example, the difference ($\ominus$) or ratio ($\div$) of the attributes $a$ and $b$ is computed;
- a **scale transformation** $\mathcal{T}_{sc}(t_{sc}, a)$ changes the scale/unit of an view axis $a$; Example types $t_{sc}$ utilized in our scheme are given in table 2 and discussed in the following.

Scale and relating view transformations both facilitate the comparison of view attributes to each other. Also characteristics in the data/views can be enhanced

| Type $t_{sc}$ | Description |
|---|---|
| $\text{norm}_{[0,1]}$ | Normalizing the samples $x_{i,j}$ of a distribution $\mathbf{x}_j$ to $[0,1]$, i.e., $\frac{x_{i,j}-x_{\min,j}}{x_{\max,j}-x_{\min,j}}$ (with corresponding min-/max-values). |
| $\text{norm}_{\text{z}}$ | Computing the z-score for each distribution (see Eq. 4). |
| $\text{norm}_{\mathcal{N}}$ | Normalization of the samples $x_{i,j}$ with respect to a standard normal distribution $\mathcal{N}$ by computing $\Phi(x_{i,j})$ where $\Phi$ denotes the cumulative distribution function of $\mathcal{N}$. |
| log | Computing the logarithm of the samples, i.e., $\log x_{i,j}$. |

**Table 2:** Different types of scale transformations $\mathcal{T}_{sc}$.

such as deviations from the norm. In the following, we discuss scale and relating view transformation on several example views.

We continue with our illustrative example of multi-run climate data. Since the individual distributions in Fig. 1a stem from different spatial positions (e.g., from hot and also cold regions) the corresponding temperature ranges are quite different. One option to better relate the distributions to each other is a relating transformation $\mathcal{T}_{rel}(\ominus, a_y, \text{med}(a_y))$ applied to the y-axis $a_y$ of the quantile plot. Accordingly, the median is subtracted from the values $x_{i,j}$ of each distribution $\mathbf{x}_j$, i.e., $\tilde{x}_{i,j} = x_{i,j} - \text{med}(x_{1,j}, \ldots, x_{100,j})$. By using the median instead of the mean, also an implicit robustifying transformation is applied. The resulting plot in Fig. 1b shows the quantiles $\tilde{q}_j(p)$ of the differences to the median $\tilde{x}_{i,j}$. It is advantageous that vertical distances in the view still represent temperature differences, however, it is not obvious whether deviations from the median also represent outliers.

To address this issue, another relating transformation $\mathcal{T}_{rel}(\div, a_y, \text{MAD}(\mathbf{x}_j))$ is applied to the view in Fig. 1b. The temperature differences $\tilde{x}_{i,j}$ are thus divided by the corresponding MAD. The resulting plot in Fig 1d depicts the quantiles of the median/MAD-based z-score that represents a robust measure of outlyingness (this view can also be obtained by $\mathcal{T}_{sc}(\text{norm}_{\text{z}}, a_y)$ applied to Fig. 1a, see Tab. 2). The plot in Fig. 1d is suitable for investigating outliers located above or below $\pm 2$ (in contrast to Fig. 1b). Several of the left-skewed distributions highlighted in green, for instance, contain strongly deviating outliers according to the robust z-score measure. On the other hand, selected distributions with high standard deviation (red) apparently belong to distributions with two different modes (local maxima). This can also be seen in a histogram where the values of each distribution are normalized to the unit interval by a scale transformation.

Another option facilitating the comparison of distributions in Fig. 1a is a scale transformation $\mathcal{T}_{sc}(\text{norm}_{[0,1]}, a_y)$ applied to the y-axis $a_y$. The multi-run values of each distribution are thus normalized to the unit interval (see Tab. 2), the resulting quantile plot is shown in Fig. 1c. No presumptions about the individual distributions are required when constructing this plot (in contrast to a Q–Q plot described below). The typical pattern of a standard normal distribution is indi-

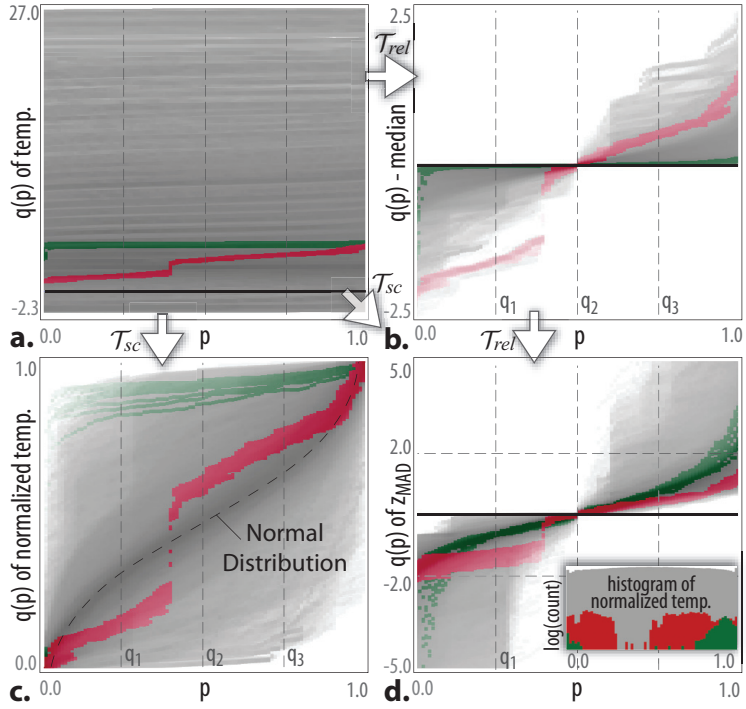



**Figure 1:** Different quantile plots show distributions of multi-run data: (a) shows the original temperature values. The distances to the distribution's median are shown in (b). This view is normalized by the MAD in (d) to identify outliers. The individual distributions in (a) are normalized to $[0, 1]$ in (c). Views in (b, c, d) result from view transformations $\mathcal{T}$ of view (a).

cated as a dashed curve. Interesting distributions that, for instance, deviate from this curve can be observed. Moreover, relations between the quantiles of a distribution can be seen (e.g., comparing the three quartiles with $p = 0.25, 0.5, 0.75$). Contrary to Fig. 1a and b, it becomes clearer that the samples emphasized in green belong to left-skewed distributions where the mass of the distributions is concentrated on the top of Fig 1c. Vertical distances, however, can no longer be interpreted as temperature differences since a relative scale is depicted on the y-axis (compared with Fig 1b).

**Q–Q (quantile–quantile) plots:** A Q–Q plot [254] is commonly used in statistics to compare a distribution of data samples to a theoretical distribution such as a normal distribution. The quantiles of both distributions are, thereby, plotted against each other. We can generate a Q–Q plot by applying a scale transformation $\mathcal{T}_{sc}(\text{norm}_{\mathcal{N}}, a_x)$ on the view in Fig. 1c. The attribute mapped to the

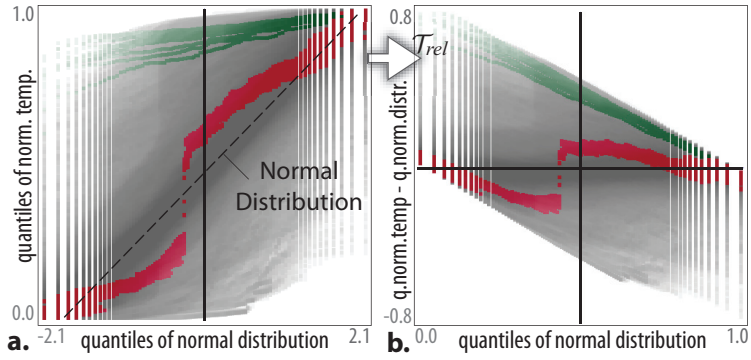



**Figure 2:** A Q–Q (quantile-quantile) plot in (a) compares the sample distribution to a standard normal distribution. Applying a view transformation, deviations from the indicated line are investigated in a detrended Q–Q plot in (b).

x-axis $a_x$ is then normalized with respect to a standard normal distribution $\mathcal{N}$. The resulting view is shown in Fig. 2a where the quantiles of the normalized multi-run data $\hat{q}_j(p)$ are plotted against the quantiles $\Phi^{-1}(p)$ of the standard normal distribution (x-axis). Multi-run values that are normally distributed are (approximately) located along the indicated line. This would be a 45° diagonal in the case of a standard normal distribution and a quadratic plot. Deviations from the line can have different reasons. The distribution may contain outliers that would be located in the upper or lower area of the plot, or the samples may be distributed with a different skewness and/or kurtosis such as a heavy-tailed distribution.

One is often interested in the deviations from the reference distribution (i.e., the diagonal in the Q–Q plot). A *detrended Q–Q plot* (see Fig. 2b) can be used for this purpose. The standard Q–Q plot in Fig. 2a has been vertically sheared by subtracting the attribute mapped on the x-axis from the y-axis—both data attributes, thereby, need to be normalized to approximately the same data range. The detrended Q–Q plot in Fig. 2b is constructed accordingly by two view transformations of Fig. 2a, i.e., $\mathcal{T}_{rel}(\ominus, a_y, a_x) \circ \mathcal{T}_{sc}(\text{norm}_{[0,1]}, a_x)$. Data samples stemming from the same as the reference distribution are located approximately on the x-axis ($y = 0$). Deviations from a normal distribution are represented more explicitly in Fig. 2b and can be investigated, for instance, by brushing (the original Q–Q plot is then used as a reference).

The presented view transformations represent the basic construction elements for our view classification. In future work, we will investigate the inclusion of further view transformations such as relating transformations depicting the principal components of two view attributes or scale transformations performing a contrast enhancement on an axis (e.g., windowing).

## 4.3 A Classification Scheme for Moment-based Views

The four types of view transformations previously discussed are the building elements for our classification of moment-based views. The order transformation $\mathcal{T}_{ord}$ is the most important one, constructing views of type $k^{\text{th}}$ vs. $(k+1)^{\text{th}}$ moment (see Sec. 4.3). The view transformation $\mathcal{T}_{rob}$ is the next most important one, changing a traditional to a robust measure. Corresponding views of type $k^{\text{th}}$ vs. $k^{\text{th}}$ moment (traditional and/or robust measures) are discussed in Sec. 4.3. The views in each category can be further refined, for instance, applying some kind of normalization to the attributes (scale transformation $\mathcal{T}_{sc}$). In cases where one is interested in deviations from the norm (e.g., the diagonal in a view), a view transformation $\mathcal{T}_{rel}$ can relate both view attributes (e.g., by subtraction or division).

#### Views depicting the $k^{\text{th}}$ vs. $(k+1)^{\text{th}}$ moment

This category of views is beneficial for investigating relations between moments. An initial setup of views is created that shows combinations of all four moments simultaneously. This allows the investigation of the basic characteristics of data distributions. We start from a scatterplot showing mean vs. standard deviation in the aggregated data part (see Fig. 3a). The view is altered by applying consecutive transformations of moment order $\mathcal{T}_{ord}$, leading to Fig. 3b and 3d (indicated with arrows). The views are arranged such that each of them have an axis in common. For practical reasons, such a view setup can be provided as a default configuration. In the multi-run data part (see Fig 3c), moreover, a quantile plot shows the median/MAD-based z-score as a robust measure of outlyingness (for alternative plots see Sec. 4.2).

Skewness and kurtosis form a pattern in Fig 3d, known as a Fleishman system [61]. Positive kurtosis values correspond to *leptokurtic* distributions with a more peaked shape and also fatter tails than a normal distribution. In other words, values are more concentrated near the data center, and a higher probability for extreme values exists (thus the kurtosis is also useful to identify distributions with outliers). *Platykurtic* distributions (kurtosis $< 0$), in contrast, have a lower wider peak around the center and thinner tails (i.e., a lower probability of extreme values compared with a normal distribution). Skewness gives additionally an indication whether the data center is shifted within the distribution.

While brushing particular attributes in a view, the relations between moments and distributions can be investigated in the other views. Using two brushes, for instance, an interesting combination of mean and standard deviation is first selected in Fig. 3a and then refined in Fig. 3b. The corresponding distributions with negative and positive skewness are highlighted in green and red, respectively. In the left part of Fig 3c, certain outliers with negative skewness (green) can be seen that strongly deviate from the rest (see also the inset showing a quantile

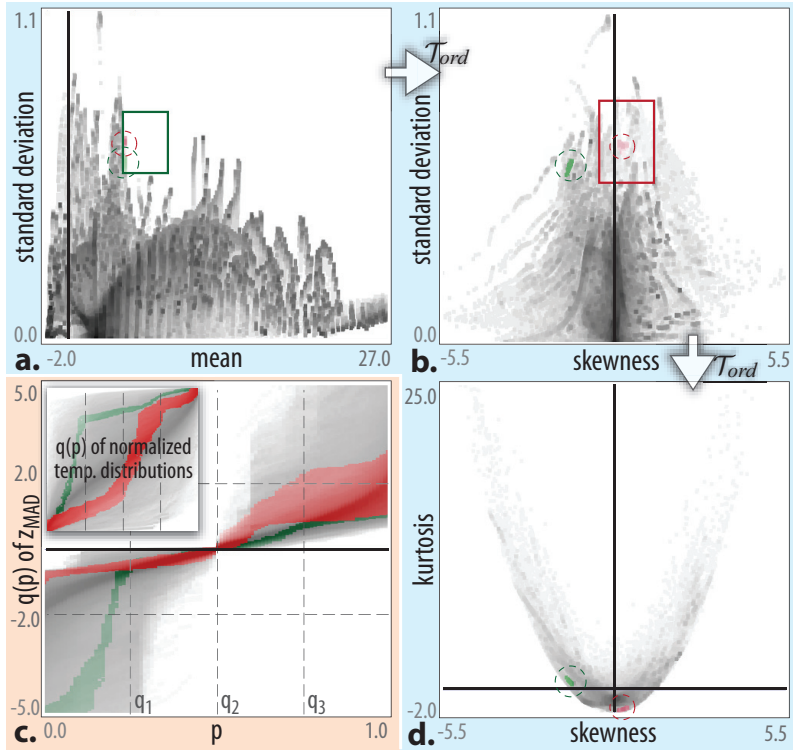



**Figure 3:** Basic view setup showing combinations of all four moments in (a), (b), and (d) (aggregated data part). The quantile plot in (c) is utilized to identify possible outliers. Interesting distributions are brushed and highlighted in color.

plot of normalized temperature distributions, compare to Fig. 1c). During the analysis, a 3D view is used in addition that encodes selected statistical properties in color and gives spatial reference of the selected features using a focus+context style (not shown here).

**Robustifying transformations:** Since the traditional moments can be influenced by outliers, we use robust alternatives for certain plots. In Fig. 4a, the classical skewness and kurtosis measures are opposed to each other. The view transformation $\mathcal{T}_{rob}(\text{rob}_{oct}, \{a_x, a_y\})$ leads to the octile-based measures in Fig. 4b. High skewness/kurtosis values are brushed in Fig. 4a, the corresponding robust measures yield smaller values (emphasized in green) in relation to others. The selected values in Fig. 4a, therefore, apparently result from outliers in the distributions. High octile-based kurtosis values are, moreover, selected in Fig. 4b (colored red).

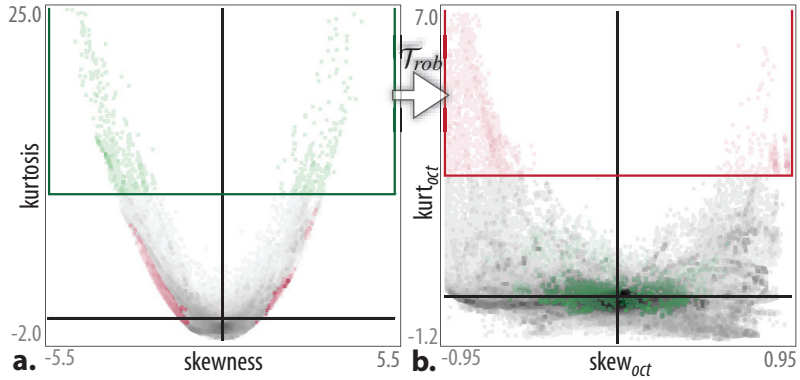



**Figure 4:** Traditional vs. octile-based measures for skewness and kurtosis: High skewness values are brushed in (a) and apparently result from outliers since the corresponding robust measures in (b) (green) yield values closer to zero.

**Scale transformations:** To make the measures in Fig. 4 more comparable to a normal distribution, a scale transformation can be applied. Skewness measures are, therefore, multiplied with a factor $\sqrt{6/n}$ and kurtosis measures with a factor $\sqrt{24/n}$ [45] ($n = 100$, i.e., the number of samples per distribution). For normally distributed values, both the classical and the robust measures then yield values in $[-2, 2]$ for about 95% of the samples.

A *spread vs. level plot* [233] can be obtained by applying $\mathcal{T}_{sc}(\log, \{a_x, a_y\}) \circ \mathcal{T}_{rob}(\text{rob}_{oct}, \{a_x, a_y\})$ to the axes in Fig. 3a. The logarithm of the median (x-axis) is then plotted against the logarithm of the IQR (y-axis). Such a plot is commonly used in statistics to estimate an appropriate transformation for a variance stabilization (e.g., when comparing groups with different variances). The necessary parameters for the transformation can be estimated using the plot (see Tukey [233] for further details).

### Views depicting the $k^{\text{th}}$ vs. $k^{\text{th}}$ moment estimated in a robust and/or traditional way

Views of this category result from robustifying transformations of a $k^{\text{th}}$ vs. $k^{\text{th}}$ moment plot and are useful to assess the influence of outliers on different moment estimates. Examples are mean vs. median, standard deviation vs. IQR (or MAD), skewness vs. octile-based (or median/MAD-based) skewness, etc. Also robust measures can be compared against each other, for instance, IQR vs. MAD, or octile-based vs. median/MAD-based skewness. For a normal distribution, the points in such plots are expected to be located along the diagonal. Therefore, we are especially interested in deviations from the diagonal. A relating transfor-

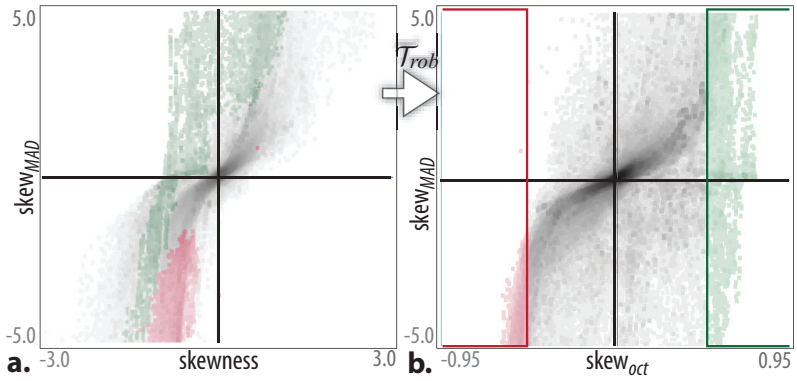



**Figure 5:** Comparing traditional vs. median/MAD-based vs. octile-based skewness. Some of the green highlighted points with positive $\mathrm{skew}_{oct}$ selected in (b) even have a negative value for the traditional skewness in (a).

mation that, for instance, subtracts the x-axis from the y-axis can be beneficial here (compare to the detrended Q–Q plot, Sec. 4.2).

**Comparing estimates of the same moment:** Fig. 5 opposes the traditional skewness to two robust estimates (i.e., $\mathrm{skew}_{oct}$ based on octiles and $\mathrm{skew}_{\mathrm{MAD}}$ based on the median/MAD). Samples approximately located along the diagonal are normally distributed. High absolute values for $\mathrm{skew}_{oct}$ are brushed in Fig. 5b. Some points with a positive $\mathrm{skew}_{oct}$ value (green) even have a negative value for the classical estimate in Fig. 5a. For such distributions with outliers, the traditional measures can be very misleading.

**Relating transformations:** As discussed above, the deviation from the norm is often especially interesting (e.g., the diagonal in some of our plots). Fig. 6a results from a relating transformation of a standard deviation (x-axis) vs. IQR plot where the difference (IQR − standard deviation) is mapped to the y-axis. Several interesting points are located along the diagonals. To enhance the "contrast" of the attribute on the y-axis, another relating transformation $\mathcal{T}_{rel}(\div, a_y, a_x)$ is performed where the y-axis is divided by the x-axis. In the resulting view (see inset) we can brush the diagonals of Fig. 6a. The according points are located close to $\pm 1$ in the inset and are highlighted in red and green, respectively. The related distributions in Fig. 6b form an interesting pattern of peakedness, which can be further investigated looking at the corresponding kurtosis values, for instance.

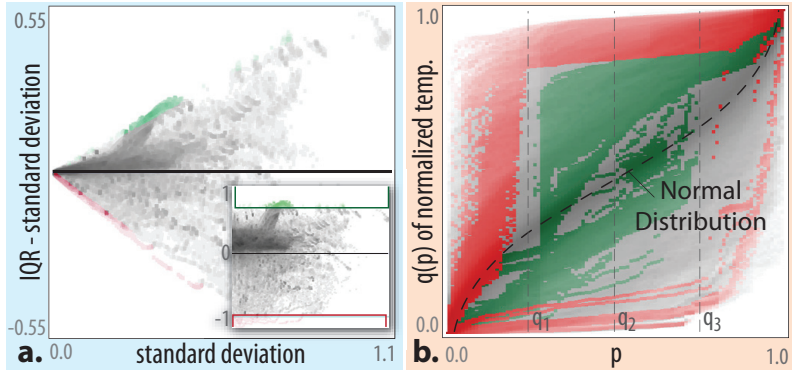



**Figure 6:** (a) shows the result of a relating transformation applied to a standard deviation vs. IQR plot. Items along the diagonals are selected in a transformed view (inset) and correspond to distributions with a peaked shape in (b).

## 5 Demonstration Case

We exemplify our approach in another visual analysis of multi-run climate data. The investigated data stems from the atmosphere-part of the same CLIMBER-2 model where a cooling over the North Atlantic is simulated [10]. A global sensitivity analysis (GSA) based on the Morris method [158] is performed in the simulation. The model parameter space with seven parameters is sampled iteratively to determine the most influential parameters on the model state. The resulting multi-run data represents a 3D atmosphere over 500 years given for 240 runs. As a first step, the four standard moments are computed for the distributions over multiple runs. In Fig. 7a, the resulting mean temperature, standard deviation, and skewness are encoded in color and give a first overview (timestep 80 is shown, which can be changed interactively). Higher standard deviations can be seen in southern latitudes together with positive skewness values. To analyze the data distributions in more detail, a view setup is created (similar to Fig. 3) that shows all four standard moments (aggregated data) and a quantile plot.

Relations between different moments and distributions are explored via brushing. In the scatterplot in Fig. 7b, distributions with positive skewness and negative kurtosis are selected. Since there is no clear boundary separating focus and context, a *smooth brush* [53] is utilized, which results in a trapezoidal degree-of-interest function ($DOI \in [0,1]$) around the main region of interest. The corresponding distributions are emphasized in green in the other views according to the DOI information. In Fig. 7c, a quantile plot depicts normalized temperature values resulting from $\mathcal{T}_{sc}(\text{norm}_{[0,1]}, a_y)$. The majority of the selected distributions are bimodal, i.e., they have two modes (local maxima as shown in the histogram). For these cells, the runs represent two different climate states of the model. In

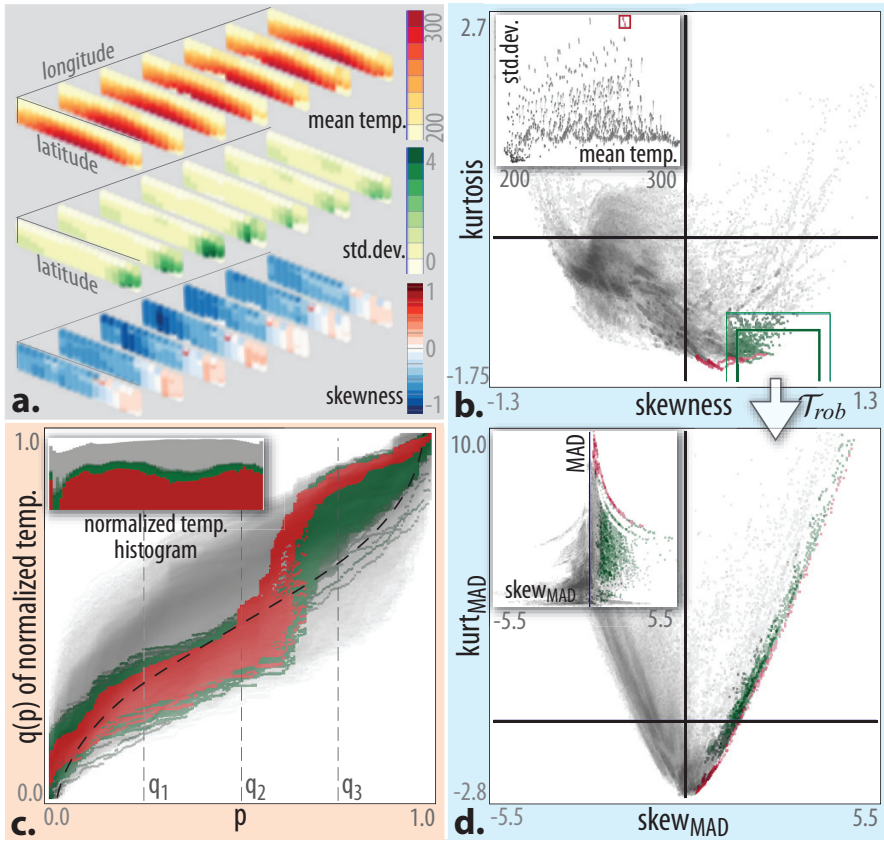



**Figure 7:** The 3D atmosphere is shown in (a), encoding mean, standard deviation, and skewness at timestep 80. Interesting data characteristics are brushed in (b) and refined in the inset, the corresponding distributions are investigated in a quantile plot in (c). A robustified version of (b) is shown in (d).

a scatterplot showing mean vs. standard deviation, the highlighted points form certain clusters. One of them is brushed for further investigation (see inset), the corresponding distributions are highlighted in red. The main characteristics of the two selections can be compared with each other, for instance, in the quantile plot or the skewness vs. kurtosis plot (see Fig. 7b). In the spatial context, these distributions are located in the south in the early timesteps of the simulation.

As a next step, we analyze the influence of outliers on the utilized classical moments. A robustifying transformation $\mathcal{T}_{rob}$ is applied to several of our views. Fig. 7d plots median/MAD-based skewness vs. kurtosis values. Due to outliers, some of the highlighted points (red, green) with positive robust kurtosis values are negative when estimated traditionally (see Fig. 7b). Moreover, certain $\text{skew}_{\text{MAD}}$

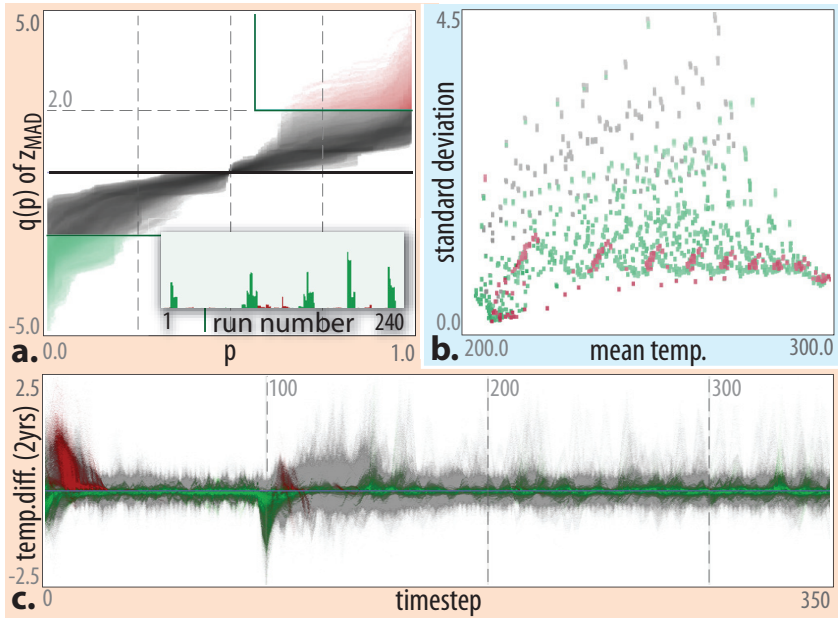



**Figure 8:** Negative outliers selected in (a) form a repetitive pattern with respect to the run parameter (highlighted green in the inset). Positive outliers (red) form a repetitive pattern in the mean vs. standard deviation plot in (b). The temporal evolution can be seen in a function graphs view in (c).

vs. MAD combinations can be seen in the inset that are inversely proportional (red, green). This correlation is not expected and is apparently a characteristic of the investigated data.

A transformed quantile plot showing the robust z-score in Fig. 8a that allows the selection of outliers above +2 (red) and below −2 (green). Positive outliers (red) correspond to a repetitive pattern in the mean vs. standard deviation plot (see Fig 8b showing timesteps 300–500), and stem mainly from different height levels in the atmosphere. To study the relation to the input parameters of the simulation, a histogram (inset) highlights the number of outliers with respect to the run number. A repetitive pattern corresponds to the negative outliers (green) that apparently results from the Morris method [158] of sampling the input parameter space. The runs with the corresponding input parameters result in values that deviate from the rest, which is relatively stable over the investigated timespan. This can be seen in the function graphs view (Fig. 8c) showing bi-annual temperature differences for each simulation cell. The temperature drop at timestep 100 results from the induced meltwater impulse, moreover, positive outliers (red) in the early timesteps of the simulation can be seen.

# 6 Conclusions and Future Work

Statistics are well known for describing important characteristics of data distributions. High-dimensional data can be reduced by considering statistics computed along selected independent data dimensions (instead of the individual values). We have demonstrated that it is rewarding to integrate such a dimension reduction mechanism in the interactive visual analysis of multi-dimensional scientific data. Estimates of the four statistical moments in their traditional or robust form (based on quartiles/octiles or median/MAD), in their original or transformed (scaled) data unit (e.g., normalization to $[0, 1]$, z-standardization), can be combined in a multitude of informative views on the data. We have presented a structured discussion of this rich space of possible moment-based views that can be constructed by consecutive view transformations ($\mathcal{T}_{ord}$, $\mathcal{T}_{rob}$, $\mathcal{T}_{sc}$, $\mathcal{T}_{rel}$). Beneficial configurations of such views have been discussed, including views that oppose the $k^{\text{th}}$ and $(k+1)^{\text{th}}$ statistical moment, views showing a traditional and robust estimate or two robust estimates of the same moment, and views that make relations between data attributes visible by an explicit representation (e.g., division, subtraction).

We experienced a substantial increase of opportunities in the interactive visual analysis as compared to traditional approaches. The tight integration of a computational and interactive analysis methodology is well aligned with Keim's requirements for prototypic visual analytics solutions [116]. We consider the fact that we came across a number of known views from statistics literature (e.g., spread vs. level plot, standard and detrended Q–Q plot), a confirmation that our views scheme is appropriate and useful. Parts of our view classification can even be regarded more general than discussed here, for example, the difference between looking at values in the original data unit, and relative values to better assess deviations from the trend. We also consider describing our classification scheme by means of generic view transformations useful as it tightly matches the iterative nature of a visual analysis: Views are developed step-by-step along with a mental model that is necessary to understand the views and the depicted data properties. An according user interface solution could be developed, where a hierarchical context menu can be used to change between views by applying view transformations.

Interesting opportunities for future work include the extension of the conceptual framework presented here (e.g., including other robust estimates and measures of outlyingness). While we have focused on the use of scatterplots in this paper, we aim at also including other views in our classification. In parallel coordinates, for example, one can bring up all four moments next to each other in their traditional and/or robust form. Moreover, we aim at including further view transformations, for instance, a relating transformation that shows the deviation from a linear/non-linear regression measure between the attributes. Other view transformations could enhance the "contrast" of the depicted attributes, for in-

stance, by applying a windowing or clustering algorithm that also preserves the continuous nature of scientific data.

## Acknowledgments

The authors thank Thomas Nocke, Michael Flechsig, and colleagues from the Potsdam Institute for Climate Impact Research (www.pik-potsdam.de) for fruitful discussions, valuable comments, and for providing the climate simulation data. We thank Helmut Doleisch and Philipp Muigg from the SimVis GmbH (www.simvis.at) for their support, also Armin Pobitzer and Stian Eikeland (both from the Univ. of Bergen) for helpful discussions and data conversion. Finally, we thank our reviewers for their valuable comments.

# Bibliography


[1] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visualizing time-oriented data: A systematic view. *Computers & Graphics*, 31(3):401–409, 2007.

[2] W. Aigner, S. Miksch, W. Müller, H. Schumann, and C. Tominski. Visual methods for analyzing time-oriented data. *IEEE Trans. Visualization and Computer Graphics*, 14(1):47–60, 2008.

[3] H. Akiba, N. Fout, and K.-L. Ma. Simultaneous classification of time-varying volume data based on the time histogram. In *Proc. Eurographics/IEEE-VGTC Symp. on Visualization (EuroVis 2006)*, pages 171–178, 2006.

[4] G. Andrienko, N. Andrienko, J. Dykes, S. Fabrikant, and M. Wachowicz. Geovisualization of dynamics, movement and change: key issues and developing approaches in visualization research. *Information Visualization*, 7(3):173–180, 2008.

[5] N. Andrienko and G. Andrienko. *Exploratory Analysis of Spatial and Temporal Data – A Systematic Approach*. Springer, 2006.

[6] N. Andrienko and G. Andrienko. Spatial generalization and aggregation of massive movement data. *IEEE Trans. Visualization and Computer Graphics*, 17(2):205–219, 2011.

[7] D. Asimov. The grand tour: A tool for viewing multidimensional data. *SIAM J. Scientific and Statistical Computing*, 6:128–143, 1985.

[8] M. Baker and C. Bushell. After the storm: considerations for information visualization. *IEEE Computer Graphics and Applications*, 15(3):12–15, 1995.

[9] D. Bauer and R. Peikert. Vortex tracking in scale-space. In *Proc. Eurographics/IEEE-TCVG Symp. on Visualization (VisSym 2002)*, pages 233–240, 2002.

[10] E. Bauer, A. Ganopolski, and M. Montoya. Simulation of the cold climate event 8200 years ago by meltwater outburst from Lake Agassiz. *Paleoceanography*, 19, 2004.

[11] R. Becker and W. Cleveland. Brushing scatterplots. *Technometrics*, 29(2):127–142, 1987.

[12] J. Bertin. *Semiology of graphics*. Univ. of Wisconsin Press, 1983.

[13] E. Bertini and D. Lalanne. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. *ACM SIGKDD Explor. Newsl.*, 11:9–18, 2009.

[14] E. Bertini and D. Lalanne. Surveying the complementary role of automatic data analysis and visualization in knowledge discovery. In *Proc. ACM SIGKDD Workshop Visual Analytics and Knowledge Discovery (VAKD '09)*, pages 12–20, 2009.

[15] J. Beyer, M. Hadwiger, S. Wolfsberger, and K. Bühler. High-quality multimodal volume rendering for preoperative planning of neurosurgical interventions. *IEEE Trans. Visualization and Computer Graphics*, 13(6):1696–1703, 2007.

[16] J. Blaas, C. Botha, and F. Post. Extensions of parallel coordinates for interactive exploration of large multi-timepoint data sets. *IEEE Trans. Visualization and Computer Graphics*, 14(6):1436–1451, 2008.

[17] U. D. Bordoloi, D. L. Kao, and H.-W. Shen. Visualization techniques for spatial probability density function data. *Data Science J.*, 3:153–162, 2004.

[18] N. Boukhelifa and P. J. Rodgers. A model and software system for coordinated and multiple views in exploratory visualization. *Information Visualization*, 2(4):258–269, 2003.

[19] C. Brewer. Color use guidelines for data representation. In *Proc. Section on Statistical Graphics*, pages 55–60, 1999.

[20] S. Bruckner and T. Möller. Result-driven exploration of simulation parameter spaces for visual effects design. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1468–1476, 2010.

[21] A. Buja, D. Swayne, M. Littman, N. Dean, and H. Hofmann. XGvis: interactive data visualization with multidimensional scaling. *J. Computational and Graphical Statistics*, 2004.

[22] H.-J. Bungartz and M. Schäfer, editors. *Fluid-Structure Interaction: Modelling, Simulation, Optimisation*, volume 53 of *Lecture Notes in Computational Science and Engineering*. Springer, 2006.

[23] P. Buono, A. Aris, C. Plaisant, A. Khella, and B. Shneiderman. Interactive pattern search in time series. In *Proc. IST/SPIE's 17th Ann. Int'l. Symp. Electronic Imaging (VDA '05)*, volume 5669, pages 175–186, 2005.

[24] R. Bürger and H. Hauser. Visualization of multi-variate scientific data. In *Eurographics 2007 State of the Art Reports*, pages 117–134, 2007.

[25] M. Burns, M. Haidacher, W. Wein, I. Viola, and M. E. Gröller. Feature emphasis and contextual cutaways for multimodal medical visualization. In *Proc. Eurographics/IEEE-VGTC Symp. on Visualization (EuroVis 2007)*, pages 275–282, 2007.

[26] L. Byron and M. Wattenberg. Stacked graphs – geometry & aesthetics. *IEEE Trans. Visualization and Computer Graphics*, 14(6):1245–1252, 2008.

[27] W. Cai and G. Sakas. Data intermixing and multivolume rendering. *Computer Graphics Forum*, 18(3):359–368, 1999.

[28] M. Cammarano et al. Visualization of heterogeneous data. *IEEE Trans. Visualization and Computer Graphics*, 13(6):1200–1207, 2007.

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

[30] J. Chambers, W. Cleveland, B. Kleiner, and P. Tukey. *Graphical Methods for Data Analysis*. Chapman and Hall, 1983.

[31] Y.-H. Chan, C. Correa, and K.-L. Ma. Flow-based scatterplots for sensitivity analysis. In *Proc. IEEE Conf. Visual Analytics Science and Technology (VAST 2010)*, pages 43 –50, 2010.

[32] C. Chen. An information-theoretic view of visual analytics. *IEEE Computer Graphics and Applications*, 28(1):18–23, 2008.

[33] M. Chen et al. Data, information, and knowledge in visualization. *IEEE Computer Graphics and Applications*, 29:12–19, 2009.

[34] M. Chen and H. Jänicke. An information-theoretic framework for visualization. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1206–1215, 2010.

[35] M. Chen and J. V. Tucker. Constructive volume geometry. *Computer Graphics Forum*, 19(4):181–193, 2000.

[36] E. Chi. A taxonomy of visualization techniques using the data state reference model. In *Proc. IEEE Symp. Information Visualization (InfoVis 2000)*, pages 69–75, 2000.

[37] H. Childs. *An Analysis Framework Addressing the Scale and Legibility of Large Scientific Data Sets*. PhD thesis, Computer Science Department, University of California, Davis, 2006.

[38] W. Cleveland. *Visualizing Data*. Hobart Press, 1993.

[39] W. Cleveland and M. McGill, editors. *Dynamic Graphics for Statistics*. Wadsworth, 1988.

[40] W. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. American Statistical Association*, 79(387):531–554, 1984.

[41] R. Cooke and J. van Noortwijk. Graphical Methods for Uncertainty and Sensitivity Analysis. In Saltelli, Chan, and Scott, editors, *Sensitivity Analysis*, pages 245–266. Wiley, 2000.

[42] E. Cordero and P. M. de Forster. Stratospheric variability and trends in models used for the IPCC AR4. *Atmos. Chem. Phys.*, 6:5369–5380, 2006.

[43] C. Correa, Y.-H. Chan, and K.-L. Ma. A framework for uncertainty-aware visual analytics. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST 2009)*, pages 51–58, 2009.

[44] R. Crawfis and N. Max. Texture splats for 3D scalar and vector field visualization. In *Proc. IEEE Visualization Conf. (Vis '93)*, pages 261–266, 1993.

[45] M. Crawley. *Statistics: An introduction using R*. Wiley, 2005.

[46] N. Cuntz, A. Kolb, M. Leidl, C. Rezk-Salama, and M. Böttinger. GPU-based dynamic flow visualization for climate research applications. In *Proc. Simulation and Visualization (SimVis)*, pages 371–384, 2007.

[47] O. Daae Lampe, J. Kehrer, and H. Hauser. Visual analysis of multivariate movement data using interactive difference views. In *Proc. Vision, Modeling, and Visualization (VMV 2010)*, pages 315–322, 2010.

[48] W. C. de Leeuw and J. J. van Wijk. A probe for local flow field visualization. In *Proc. IEEE Visualization Conf. (Vis '93)*, pages 39–45, 1993.

[49] M. de Oliveira and H. Levkowitz. From visual data exploration to visual data mining: a survey. *IEEE Trans. Visualization and Computer Graphics*, 9(3):378–394, 2003.

[50] P. Diggle, P. Heagerty, K.-Y. Liang, and S. Zeger. *Analysis of longitudinal data*. Oxford University Press, 2002.

[51] H. Doleisch. SimVis: interactive visual analysis of large and time-dependent 3D simulation data. In *Proc. Winter Simulation Conf.*, pages 712–720, 2007.

[52] H. Doleisch, M. Gasser, and H. Hauser. Interactive feature specification for focus+context visualization of complex simulation data. In *Proc. Eurographics/IEEE-TCVG Symp. on Visualization (VisSym 2003)*, pages 239–248, 2003.

[53] H. Doleisch and H. Hauser. Smooth brushing for focus+context visualization of simulation data in 3D. *Journal of WSCG*, 10(1):147–154, 2002.

[54] H. Doleisch, M. Mayer, M. Gasser, R. Wanker, and H. Hauser. Case study: Visual analysis of complex, time-dependent simulation results of a diesel exhaust system. In *Proc. Eurographics/IEEE-TCVG Symp. on Visualization (VisSym 2004)*, pages 91–96, 2004.

[55] H. Doleisch, P. Muigg, and H. Hauser. Interactive visual analysis of Hurricane Isabel with SimVis. Technical Report TR-VRVis-2004-058, VRVis Research Center, Vienna Austria, 2004. http://www.vrvis.at/simvis/Isabel/.

[56] J. M. Favre. Towards efficient visualization support for single-block and multi-block datasets. In *Proc. IEEE Visualization Conf. (Vis '97)*, pages 425–428, 1997.

[57] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54, 1996.

[58] J.-D. Fekete and C. Plaisant. Interactive information visualization of a million items. In *Proc. IEEE Symp. Information Visualization (InfoVis 2002)*, pages 117–124, 2002.

[59] J.-D. Fekete, J. van Wijk, J. Stasko, and C. North. The value of information visualization. In *Information Visualization*, volume 4950 of *Lecture Notes in Computer Science*, pages 1–18. 2008.

[60] P. Filzmoser, R. Maronna, and M. Werner. Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711, 2008.

[61] A. Fleishman. A method for simulating non-normal distributions. *Psychometrika*, 43:521–532, 1978.

[62] U. Foelsche, G. Kirchengast, and A. K. Steiner. An observing system simulation experiment for climate monitoring with GNSS radio occultation data: setup and testbed study. *J. Geophys. Res.*, 113, 2008.

[63] R. Fuchs and H. Hauser. Visualization of multi-variate scientific data. *Computer Graphics Forum*, 28(6):1670–1690, 2009.

[64] R. Fuchs, J. Waser, and M. E. Gröller. Visual human+machine learning. *IEEE Trans. Visualization and Computer Graphics*, 15(6):1327–1334, 2009.

[65] G. W. Furnas. Generalized fisheye views. In *Proc. Human Factors in Computing Systems (CHI '86)*, pages 16–23, 1986.

[66] D. Gresh et al. WEAVE: a system for visually linking 3-D and statistical visualizations applied to cardiac simulation and measurement data. In *Proc. IEEE Visualization Conf. (Vis 2000)*, pages 489–492, 2000.

[67] H. Griethe and H. Schumann. The visualization of uncertain data: methods and problems. In *Proc. Simulation and Visualization (SimVis 2006)*, pages 143–156, 2006.

[68] S. Grimm, S. Bruckner, A. Kanitsar, and M. E. Gröller. Flexible direct multi-volume rendering in interactive scenes. In *Proc. Vision, Modeling, and Visualization (VMV 2004)*, pages 386–379, 2004.

[69] M. Hadwiger, C. Berger, and H. Hauser. High-quality two-level volume rendering of segmented data sets on consumer graphics hardware. In *Proc. IEEE Visualization Conf. (Vis 2003)*, pages 301–308, 2003.

[70] D. M. Hamby. A review of techniques for parameter sensitivity analysis of environmental models. *J. Environmental Monitoring & Assessment*, 32(2):135–154, 2004.

[71] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[72] H. Hauser. Generalizing Focus+Context Visualization. In *Scientific Visualization: The Visual Extraction of Knowledge from Data*, pages 305–327. Springer, 2005.

[73] H. Hauser. Interactive visual analysis – an opportunity for industrial simulation. In *Proc. Simulation and Visualization (SimVis)*, pages 1–6, 2006.

[74] H. Hauser. Levels of interactive visual analysis. Presentation at Dagstuhl Seminar on Scientific Visualization, June 14–19, Schloss Dagstuhl, Germany, 2009.

[75] H. Hauser, F. Ledermann, and H. Doleisch. Angular brushing of extended parallel coordinates. In *Proc. IEEE Symp. Information Visualization (InfoVis 2002)*, pages 127–130, 2002.

[76] H. Hauser, L. Mroz, G.-I. Bischi, and M. E. Gröller. Two-level volume rendering – fusing MIP and DVR. In *Proc. IEEE Visualization Conf. (Vis 2000)*, pages 211–218, 2000.

[77] H. Hauser and H. Schumann. Visualization pipeline. In Liu and Özsu, editors, *Encyclopedia of Database Systems*, pages 3414–3416. Springer, 2009.

[78] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. ThemeRiver: visualizing thematic changes in large document collections. *IEEE Trans. Visualization and Computer Graphics*, 8(1):9–20, 2002.

[79] C. Healey, K. Booth, and J. Enns. High-speed visual estimation using preattentive processing. *ACM Trans. Computer-Human Interaction*, 3:107–135, 1996.

[80] C. Healey and J. Enns. Large datasets at a glance: Combining textures and colors in scientific visualization. *IEEE Trans. Visualization and Computer Graphics*, 5(2):145–167, 1999.

[81] C. Healey and J. Enns. Perception and painting: a search for effective, engaging visualizations. *IEEE Computer Graphics and Applications*, 22(2):10–15, 2002.

[82] J. Helton. Uncertainty and sensitivity analysis for models of complex systems. In *Computational Methods in Transport: Verification and Validation*, volume 62 of *Lecture Notes in Computational Science and Engineering*, pages 207–228. 2008.

[83] J. Helton, J. Johnson, C. Sallaberry, and C. Storlie. Survey of sampling-based methods for uncertainty and sensitivity analysis. *Reliability Engineering and System Safety*, 91(10–11):1175–1209, 2006.

[84] C. Henze. Feature detection in linked derived spaces. In *Proc. IEEE Visualization Conf. (Vis '98)*, pages 87–94, 1998.

[85] B. Hibbard. The top five problems that motivated my work. *IEEE Computer Graphics and Applications*, 24(6):9–13, 2004.

[86] B. Hibbard, M. Böttinger, M. Schultz, and J. Biercamp. Visualization in Earth System Science. *SIGGRAPH Comput. Graph.*, 36(4):5–9, 2002.

[87] D. Hinkley. On power transformations to symmetry. *Biometrika*, 63:101–111, 1975.

[88] H. Hochheiser and B. Shneiderman. Dynamic query tools for time series data sets: timebox widgets for interactive exploration. *Information Visualization*, 3(1):1–18, 2004.

[89] D. House, A. Bair, and C. Ware. An approach to the perceptual optimization of complex visualizations. *IEEE Trans. Visualization and Computer Graphics*, 12(4):509–521, 2006.

[90] R. J. Hyndman and Y. Fan. Sample quantiles in statistical packages. *The American Statistician*, 50(4):361–365, 1996.

[91] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1(2):69–91, 1985.

[92] V. Interrante and C. Grosch. Visualizing 3D flow. *IEEE Computer Graphics and Applications*, 18(4):49–53, 1998.

[93] H. Jänicke, M. Böttinger, U. Mikolajewicz, and G. Scheuermann. Visual exploration of climate variability changes using wavelet analysis. *IEEE Trans. Visualization and Computer Graphics*, 15(6):1375–1382, 2009.

[94] H. Jänicke, M. Böttinger, and G. Scheuermann. Brushing of attribute clouds for the visualization of multivariate data. *IEEE Trans. Visualization and Computer Graphics*, 14(6):1459–1466, 2008.

[95] H. Jänicke, M. Böttinger, X. Tricoche, and G. Scheuermann. Automatic detection and visualization of distinctive structures in 3D unsteady multi-fields. *Computer Graphics Forum*, 27(3):767–774, 2008.

[96] H. Jänicke, A. Wiebel, G. Scheuermann, and W. Kollmann. Multifield visualization using local statistical complexity. *IEEE Trans. Visualization and Computer Graphics*, 13(6):1384–1391, 2007.

[97] T. Jankun-Kelly and K.-L. Ma. A study of transfer functions generation for time-varying volume data. In *Proc. Eurographics/IEEE-TCVG Workshop Volume Graphics*, pages 51–68, 2001.

[98] T. Jankun-Kelly and K.-L. Ma. Visualization exploration and encapsulation via a spreadsheet-like interface. *IEEE Trans. Visualization and Computer Graphics*, 7(3):275–287, 2001.

[99] T. Jankun-Kelly and K. Mehta. Superellipsoid-based, real symmetric traceless tensor glyphs motivated by nematic liquid crystal alignment visualization. *IEEE Trans. Visualization and Computer Graphics*, 12(5):1197–1204, 2006.

[100] F. Janoos, S. Singh, O. Irfanoglu, R. Machiraju, and R. Parent. Activity analysis using spatio-temporal trajectory volumes in surveillance applications. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST 2007)*, pages 3–10, 2007.

[101] J. Jeong and F. Hussain. On the identification of a vortex. *J. Fluid Mechanics Digital Archive*, 285:69–94, 1995.

[102] J. Johansson, P. Ljung, M. Jern, and M. Cooper. Revealing structure within clustered parallel coordinates displays. In *Proc. IEEE Symp. Information Visualization (InfoVis 2005)*, pages 125–132, 2005.

[103] S. Johansson and J. Johansson. Interactive dimensionality reduction through user-defined combinations of quality metrics. *IEEE Trans. Visualization and Computer Graphics*, 15:993–1000, 2009.

[104] C. Johnson. Top scientific visualization research problems. *IEEE Computer Graphics and Applications*, 24(4):13–17, 2004.

[105] C. Johnson and A. Sanderson. A next step: Visualizing errors and uncertainty. *IEEE Computer Graphics and Applications*, 23(5):6–10, 2003.

[106] J. Jungclaus et al. Ocean circulation and tropical variability in the coupled model ECHAM5/MPI-OM. *J. Climate*, 19(16):3952–3972, 2006.

[107] D. Kao, J. Dungan, and A. Pang. Visualizing 2D probability distributions from EOS satellite image-derived data sets: a case study. In *Proc. IEEE Visualization Conf. (Vis 2001)*, pages 457–460, 2001.

[108] D. Kao, A. Luo, J. Dungan, and A. Pang. Visualizing spatially varying distribution data. In *Proc. Int'l. Conf. Information Visualization (IV '02)*, pages 219–225, 2002.

[109] T. Karl, S. Hassol, C. Miller, and W. Murray, editors. *Temperature Trends in the Lower Atmosphere: Steps for Understanding and Reconciling Differences*. A Report by the Climate Change Science Program and the Subcommittee on Global Change Research, Washington, DC, 2006.

[110] J. Kehrer. Integrating interactive visual analysis of large time series into the SimVis system. Master's thesis, Vienna University of Technology, Austria, 2007.

[111] J. Kehrer, P. Filzmoser, and H. Hauser. Brushing moments in interactive visual analysis. *Computer Graphics Forum*, 29(3):813–822, 2010.

[112] D. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Trans. Visualization and Computer Graphics*, 6(1):59–78, 2000.

[113] D. Keim. Information visualization and visual data mining. *IEEE Trans. Visualization and Computer Graphics*, 8(1):1–8, 2002.

[114] D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. Visual analytics: Definition, process, and challenges. In *Information Visualization*, volume 4950 of *Lecture Notes in Computer Science*, pages 154–175. 2008.

[115] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering the Information Age: Solving Problems with Visual Analytics*. Eurographics Assoc., 2010.

[116] D. Keim, F. Mansmann, J. Schneidewind, and H. Ziegler. Challenges in visual data analysis. In *Proc. Int'l. Conf. Information Visualization (IV '06)*, pages 9–16, 2006.

[117] D. Keim, F. Mansmann, and J. Thomas. Visual analytics: how much visualization and how much analytics? *SIGKDD Explor. Newsl.*, 11:5–8, 2010.

[118] D. Keim, W. Müller, and H. Schumann. Visual data mining. In *Eurographics 2002 State of the Art Reports*, pages 49–68, 2002.

[119] E. Keogh, S. Lonardi, and C. A. Ratanamahatana. Towards parameter-free data mining. In *Proc. ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '04)*, pages 206–215, New York, NY, USA, 2004.

[120] K. Kim, C. Wittenbrink, and A. Pang. Extended specifications and test data sets for data level comparisons of direct volume rendering algorithms. *IEEE Trans. Visualization and Computer Graphics*, 7(4):299–317, 2001.

[121] T.-H. Kim and H. White. On more robust estimation of skewness and kurtosis. *Finance Res. Letters*, 1(1):56–73, 2004.

[122] G. Kindlmann. Superquadric tensor glyphs. In *Proc. Eurographics/IEEE-TCVG Symp. on Visualization (VisSym 2004)*, pages 147–154, 2004.

[123] G. Kindlmann and C.-F. Westin. Diffusion tensor visualization with glyph packing. *IEEE Trans. Visualization and Computer Graphics*, 12(5):1329–1336, 2006.

[124] R. Kirby, H. Marmanis, and D. Laidlaw. Visualizing multivalued data from 2D incompressible flows using concepts from painting. In *Proc. IEEE Visualization Conf. (Vis '99)*, pages 333–340, 1999.

[125] J. Kniss, C. Hansen, M. Grenier, and T. Robinson. Volume rendering multivariate data to visualize meteorological simulations: a case study. In *Proc. Eurographics/IEEE-TCVG Symp. on Visualization (VisSym 2002)*, pages 189–194, 2002.

[126] T. Kohonen. *Self-organizing maps*. Springer, second edition, 1997.

[127] Z. Konyha, K. Matković, D. Gračanin, M. Jelović, and H. Hauser. Interactive visual analysis of families of function graphs. *IEEE Trans. Visualization and Computer Graphics*, 12(6):1373–1385, 2006.

[128] Z. Konyha, K. Matković, and H. Hauser. Interactive visual analysis in engineering: A survey. In *Proc. Spring Conference on Computer Graphics (SCCG 2009)*, pages 31–38, 2009.

[129] R. Kosara, F. Bendix, and H. Hauser. Timehistograms for Large, Time-dependent Data. In *Proc. Eurographics/IEEE-TCVG Symp. on Visualization (VisSym 2004)*, pages 45–54, 2004.

[130] M. Kreuseler and H. Schumann. A flexible approach for visual data mining. *IEEE Trans. Visualization and Computer Graphics*, 8:39–51, 2002.

[131] B. C. Lackner, A. K. Steiner, F. Ladstädter, and G. Kirchengast. Trend Indicators of Atmospheric Climate Change Based on Global Climate Model Scenarios. In *New Horizons in Occultation Research: Studies in Atmosphere and Climate*. Springer, 2008. (in press).

[132] F. Ladstädter, A. K. Steiner, B. C. Lackner, G. Kirchengast, P. Muigg, J. Kehrer, and H. Doleisch. SimVis: an interactive visual field exploration tool applied to climate research. In A. Steiner, B. Pirscher, U. Foelsche, and G. Kirchengast, editors, *New Horizons in Occultation Research*, pages 235–245. Springer, 2009.

[133] 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*, 27(4):667–679, 2010.

[134] R. S. Laramee, C. Garth, H. Doleisch, J. Schneider, H. Hauser, and H. Hagen. Visual Analysis and Exploration of Fluid Flow in a Cooling Jacket. In *Proc. IEEE Visualization Conf. (Vis 2005)*, pages 623–630, 2005.

[135] R. S. Laramee, H. Hauser, H. Doleisch, B. Vrolijk, F. H. Post, and D. Weiskopf. The state of the art in flow visualization: Dense and texture-based techniques. *Computer Graphics Forum*, 23(2):203–221, 2004.

[136] A. E. Lie, J. Kehrer, and H. Hauser. Critical design and realization aspects of glyph-based 3D data visualization. In *Proc. Spring Conference on Computer Graphics (SCCG 2009)*, pages 27–34, 2009.

[137] S. Lindholm, P. Ljung, M. Hadwiger, and A. Ynnerman. Fused multi-volume dvr using binary space partitioning. *Computer Graphics Forum*, 28(3):847–854, 2009.

[138] D. Lipşa, R. Laramee, S. Cox, J. Roberts, and R. Walker. Visualization for the physical sciences. In *Eurographics 2011 State of the Art Reports*, 2011. (to appear).

[139] Z. Liu and J. Stasko. Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *IEEE Trans. Visualization and Computer Graphics*, 16(6):999–1008, 2010.

[140] I. López, R. Snodgrass, and B. Moon. Spatiotemporal aggregate computation: A survey. *IEEE Trans. Knowl. Data Eng.*, 17(2):271–286, 2005.

[141] A. Love, A. Pang, and D. Kao. Visualizing spatial multivalue data. *IEEE Computer Graphics and Applications*, 25(3):69–79, 2005.

[142] A. Luo, D. Kao, and A. Pang. Visualizing spatial distribution data sets. In *Proc. Eurographics/IEEE-TCVG Symp. on Visualization (VisSym 2003)*, pages 29–38, 2003.

[143] K.-L. Ma. Visualizing time-varying volume data. *Computing in Science & Engineering*, 5(2):34–42, 2003.

[144] K.-L. Ma. Machine learning to boost the next generation of visualization technology. *IEEE Computer Graphics and Applications*, 27(5):6–9, 2007.

[145] K.-L. Ma and E. Lum. Techniques for visualizing time-varying volume data. In C. Hansen and C. Johnson, editors, *Visualization Handbook*, pages 511–531. Academic Press, 2004.

[146] A. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, 2005.

[147] M. Malik, C. Heinzl, and M. E. Gröller. Comparative visualization for parameter studies of dataset series. *IEEE Trans. Visualization and Computer Graphics*, 16(5):829–840, 2010.

[148] R. Maronna, D. Martin, and V. Yohai. *Robust Statistics: Theory and Methods*. John Wiley & Sons, 2006.

[149] K. Matković, D. Gračanin, M. Jelović, A. Ammer, A. Lež, and H. Hauser. Interactive visual analysis of multiple simulation runs using the simulation model view: Understanding and tuning of an electronic unit injector. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1449–1457, 2010.

[150] K. Matković, D. Gračanin, M. Jelović, and H. Hauser. Interactive visual steering – rapid visual prototyping of a common rail injection system. *IEEE Trans. Visualization and Computer Graphics*, 14(6):1699–1706, 2008.

[151] K. Matković, D. Gračanin, B. Klarin, and H. Hauser. Interactive visual analysis of complex scientific data as families of data surfaces. *IEEE Trans. Visualization and Computer Graphics*, 15(6):1351–1358, 2009.

[152] K. Matković, M. Jelović, J. Jurić, Z. Konyha, and D. Gračanin. Interactive visual analysis and exploration of injection systems simulations. In *Proc. IEEE Visualization Conf. (Vis 2005)*, pages 391–398, 2005.

[153] N. Max, R. Crawfis, and D. Williams. Visualization for climate modeling. *IEEE Computer Graphics and Applications*, 13(4):34–40, 1993.

[154] R. McGill, J. Tukey, and W. Larsen. Variations of box plots. *The American Statistician*, 32:12–16, 1978.

[155] W. G. Melbourne et al. The application of spaceborne GPS to atmospheric limb sounding and global change monitoring. JPL publication 94-18, Jet Propulsion Lab, Pasadena, CA, 1994. 147 pp.

[156] J. Meyer-Spradow, L. Stegger, C. Döring, T. Ropinski, and K. Hinrichs. Glyph-based SPECT visualization for the diagnosis of coronary artery disease. *IEEE Trans. Visualization and Computer Graphics*, 14(6):1499–1506, 2008.

[157] J. Moors. A quantile alternative for kurtosis. *The Statistician*, 37:25–32, 1988.

[158] M. Morris. Factorial plans for preliminary computational experiments. *Technometrics*, 33(2):161–174, 1991.

[159] A. Morrison, G. Ross, and M. Chalmers. Fast multidimensional scaling through sampling, springs and interpolation. *Information Visualization*, 2(1):68–77, 2003.

[160] P. Muigg, M. Hadwiger, H. Doleisch, and H. Hauser. Scalable hybrid unstructured and structured grid raycasting. *IEEE Trans. Visualization and Computer Graphics*, 13(6):1592–1599, 2007.

[161] P. Muigg, J. Kehrer, S. Oeltze, H. Piringer, H. Doleisch, B. Preim, and H. Hauser. A four-level focus+context approach to interactive visual analysis of temporal features in large scientific data. *Computer Graphics Forum*, 27(3):775–782, 2008.

[162] W. Müller, T. Nocke, and H. Schumann. Enhancing the visualization process with principal component analysis to support the exploration of trends. In *Proc. Asia Pacific Symp. on Information Visualisation (APVIS '06)*, pages 121–130, 2006.

[163] W. Müller and H. Schumann. Visualization methods for time-dependent data – an overview. In *Proc. Winter Simulation Conf.*, pages 737–745, 2003.

[164] T. Munzner, C. Johnson, R. Moorhead, H. Pfister, P. Rheingans, and T. S. Yoo. NIH-NSF visualization research challenges report summary. *IEEE Computer Graphics and Applications*, 26(2):20–24, 2006.

[165] T. Nocke. *Visual data mining and visualization design for climate research*. PhD thesis, Dept. of Computer Science and Electrical Engineering, Univ. of Rostock, 2007. (in German).

[166] T. Nocke, M. Flechsig, and U. Böhm. Visual exploration and evaluation of climate-related simulation data. In *Proc. Winter Simulation Conf.*, pages 703–711, 2007.

[167] T. Nocke, S. Schlechtweg, and H. Schumann. Icon-based visualization using mosaic metaphors. In *Proc. Int'l. Conf. Information Visualization (IV '05)*, pages 103–109, 2005.

[168] T. Nocke, H. Schumann, and U. Böhm. Methods for the visualization of clustered climate data. *Computational Statistics*, 19(1):75–94, 2004.

[169] T. Nocke, H. Schumann, U. Böhm, and M. Flechsig. Information visualization supporting modelling and evaluation tasks for climate models. In *Proc. Winter Simulation Conf.*, pages 763–771, 2003.

[170] T. Nocke, T. Sterzel, M. Böttinger, and M. Wrobel. Visualization of climate and climate change data: An overview. In Ehlers et al., editors, *Digital Earth Summit on Geoinformatics 2008: Tools for Global Change Research (ISDE'08)*, pages 226–232, 2008.

[171] C. North, N. Conklin, K. Indukuri, and V. Saini. Visualization schemas and a web-based architecture for custom multiple-view visualization of multiple-table databases. *Information Visualization*, 1(3–4):211–228, 2002.

[172] M. Novotný and H. Hauser. Outlier-preserving focus+context visualization in parallel coordinates. *IEEE Trans. Visualization and Computer Graphics*, 12(5):893–900, 2006.

[173] S. Oeltze, H. Doleisch, H. Hauser, P. Muigg, and B. Preim. Interactive visual analysis of perfusion data. *IEEE Trans. Visualization and Computer Graphics*, 13(6):1392–1399, 2007.

[174] S. Oeltze, A. Malyszczyk, and B. Preim. Intuitive mapping of perfusion parameters to glyph shape. In *Bildverarbeitung für die Medizin (BVM2008)*, pages 262–266, 2008.

[175] H.-G. Pagendarm and F. H. Post. Comparative visualization – approaches and examples. In *Visualization in Scientific Computing*, pages 95–108, 1995.

[176] H.-G. Pagendarm and F. H. Post. Studies in comparative visualization of flow features. In *Scientific Visualization: Overviews, Methodologies, and Techniques*, pages 211–227, 1997.

[177] A. Pang, C. Wittenbrink, and S. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997.

[178] D. Patel, M. Haidacher, J.-P. Balabanian, and M. E. Gröller. Moment curves. In *Proc. IEEE Pacific Visualization Symp. (PacificVis 2009)*, pages 201–208, 2009.

[179] R. Pickett and G. Grinstein. Iconographic displays for visualizing multidimensional data. In *Proc. IEEE Int'l. Conf. Systems, Man, and Cybernetics (SMC)*, pages 514–519. 1988.

[180] H. Piringer, W. Berger, and J. Krasser. HyperMoVal: interactive visual validation of regression models for real-time simulation. *Computer Graphics Forum*, 29(3):983–992, 2010.

[181] H. Piringer, R. Kosara, and H. Hauser. Interactive focus+context visualization with linked 2D/3D scatterplots. In *Proc. Coordinated & Multiple Views in Exploratory Visualization (CMV 2004)*, pages 49–60, 2004.

[182] J. Plate, T. Holtkämper, and B. Fröhlich. A flexible multi-volume shader framework for arbitrarily intersecting multi-resolution datasets. *IEEE Trans. Visualization and Computer Graphics*, 13(6):1584–1591, 2007.

[183] F. H. Post, B. Vrolijk, H. Hauser, R. S. Laramee, and H. Doleisch. Feature extraction and visualization of flow fields. In *Eurographics 2002 State of the Art Reports*, pages 69–100, 2002.

[184] F. H. Post, B. Vrolijk, H. Hauser, R. S. Laramee, and H. Doleisch. The state of the art in flow visualisation: Feature extraction and tracking. *Computer Graphics Forum*, 22(4):775–792, 2003.

[185] K. Potter et al. Ensemble-Vis: a framework for the statistical visualization of ensemble data. In *Proc. IEEE Int'l. Conf. on Data Mining Workshops*, pages 233–240, 2009.

[186] K. Potter, J. Kniss, R. Riesenfeld, and C. Johnson. Visualizing summary statistics and uncertainty. *Computer Graphics Forum*, 29(3):823–832, 2010.

[187] P. Rautek, S. Bruckner, and M. E. Gröller. Semantic layers for illustrative volume rendering. *IEEE Trans. Visualization and Computer Graphics*, 13:1336–1343, 2007.

[188] C. Reimann, P. Filzmoser, and R. Garrett. Background and threshold: critical comparison of methods of determination. *Science of the Total Environment*, 346:1–16, 2005.

[189] K. Riley, D. Ebert, C. Hansen, and J. Levit. Visually accurate multi-field weather visualization. In *Proc. IEEE Visualization Conf. (Vis 2003)*, pages 279–286, 2003.

[190] K. Riley et al. Visualization of structured nonuniform grids. *IEEE Computer Graphics and Applications*, 26(1):46–55, 2006.

[191] J. C. Roberts. Exploratory visualization with multiple linked views. In A. MacEachren, M.-J. Kraak, and J. Dykes, editors, *Exploring Geovisualization*, pages 159–180. Elseviers, 2004.

[192] J. C. Roberts. State of the art: Coordinated & multiple views in exploratory visualization. In *Proc. Coordinated & Multiple Views in Exploratory Visualization (CMV 2007)*, pages 61–71, 2007.

[193] E. Roeckner et al. The atmospheric general circulation model ECHAM5. Report no. 349, Max-Planck-Inst. f. Meteorology, Hamburg, 2003.

[194] T. Ropinski and B. Preim. Taxonomy and usage guidelines for glyph-based medical visualization. In *Proc. Simulation and Visualization (SimVis 2008)*, pages 121–138, 2008.

[195] T. Ropinski, M. Specht, J. Meyer-Spradow, K. Hinrichs, and B. Preim. Surface glyphs for visualizing multimodal volume data. In *Proc. Vision, Modeling, and Visualization (VMV 2007)*, pages 3–12, 2007.

[196] T. Ropinski, I. Viola, M. Biermann, H. Hauser, and K. Hinrichs. Multimodal visualization with interactive closeups. In *Proc. EGUK Theory and Practice of Computer Graphics (TPCG)*, pages 17–24, 2009.

[197] O. Rübel et al. PointCloudXplore: visual analysis of 3D gene expression data using physical views and parallel coordinates. In *Proc. Eurographics/IEEE-VGTC Symp. on Visualization (EuroVis 2006)*, pages 203–210, 2006.

[198] N. Sahasrabudhe, J. West, R. Machiraju, and M. Janus. Structured spatial domain image and data comparison metrics. In *Proc. IEEE Visualization Conf. (Vis '99)*, pages 97–515, 1999.

[199] T. Saito, H. N. Miyamura, M. Yamamoto, H. Saito, Y. Hoshiya, and T. Kaseda. Two-tone pseudo coloring: Compact visualization for one-dimensional data. In *Proc. IEEE Symp. Information Visualization (InfoVis 2005)*, page 23, 2005.

[200] T. Salzbrunn, H. Jänicke, T. Wischgoll, and G. Scheuermann. The state of the art in flow visualization: Partition-based techniques. In *Proc. Simulation and Visualization (SimVis 2008)*, pages 75–92, 2008.

[201] B. D. Santer et al. Identification of anthropogenic climate change using a second-generation reanalysis. *J. Geophys. Res.*, 109:D21104, 2004.

[202] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *IEEE Trans. Visualization and Computer Graphics*, 16(6):1421–1430, 2010.

[203] N. Sauber, H. Theisel, and H.-P. Seidel. Multifield-Graphs: an approach to visualizing correlations in multifield scalar data. *IEEE Trans. Visualization and Computer Graphics*, 12(5):917 –924, 2006.

[204] A. Sawant and C. Healey. Visualizing flow data using assorted glyphs. *Crossroads*, 14(2), 2007.

[205] H. Schumann and W. Müller. *Visualisierung – Grundlagen und allgemeine Methoden*. Springer, 2000. (in German).

[206] J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *Proc. IEEE Symp. Information Visualization (InfoVis 2004)*, pages 65–72, 2004.

[207] C. D. Shaw, J. A. Hall, C. Blahut, D. S. Ebert, and D. A. Roberts. Using shape to visualize multivariate data. In *Workshop on New Paradigms in Information Visualization and Manipulation*, pages 17–20, 1999.

[208] H. Shenas and V. Interrante. Compositing color with texture for multi-variate visualization. In *Proc. Int'l. Conf. Computer Graphics and Interactive Techniques in Australasia and South East Asia (GRAPHITE)*, pages 443–446, 2005.

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

[210] B. Shneiderman. Inventing discovery tools: combining information visualization with data mining. *Information Visualization*, 1(1):5–12, 2002.

[211] S. Silva and T. Catarci. Visualization of linear time-oriented data: A survey. In *Proc. International Conference on Web Information Systems Engineering (WISE'00)*, volume 1, pages 310–319, 2000.

[212] A. J. Simmons and J. K. Gibson. The ERA-40 Project Plan ERA-40. Project Report Series, no. 1, ECMWF, Reading, UK, 2000. 62 pp.

[213] M. Skeels, B. Lee, G. Smith, and G. Robertson. Revealing uncertainty for information visualization. *Information Visualization*, 9(1):70–81, 2010.

[214] S. Solomon et al. Technical summary. In *Climate Change 2007: The Physical Science Basis*. Cambridge University Press, UK, 2007.

[215] Y. Song, J. Ye, N. Svakhine, S. Lasher-Trapp, M. Baldwin, and D. Ebert. An atmospheric visual analysis and exploration system. *IEEE Trans. Visualization and Computer Graphics*, 12(5):1157–1164, 2006.

[216] P. Stier et al. The aerosol-climate model ECHAM5-HAM. *Atmosheric Chemistry and Physics*, 5:1125–1156, 2005.

[217] C. Stolte, D. Tang, and P. Hanrahan. Polaris: A system for query, analysis, and visualization of multidimensional relational databases. *IEEE Trans. Visualization and Computer Graphics*, 8(1):52–65, 2002.

[218] J. Sukharev, C. Wang, K.-L. Ma, and A. Wittenberg. Correlation study of time-varying multivariate climate data sets. In *Proc. IEEE Pacific Visualization Symp. (PacificVis 2009)*, pages 161–168, 2009.

[219] D. F. Swayne, D. Cook, and A. Buja. XGobi: interactive dynamic data visualization in the X window system. *J. Computational and Graphical Statistics*, 7(1):113–130, 1998.

[220] Tatu et al. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. In *Proc. IEEE Symp. Visual Analytics Science and Technology (VAST 2009)*, pages 59–66, 2009.

[221] R. Taylor. Visualizing multiple fields on the same surface. *IEEE Computer Graphics and Applications*, 22(3):6–10, 2002.

[222] M. Theus and S. Urbanek. *Interactive Graphics for Data Analysis: Principles and Examples*. Chapman & Hall, 2008.

[223] J. Thomas and K. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[224] C. Tominski, P. Schulze-Wollgast, and H. Schumann. 3D information visualization for time-dependent data on maps. In *Proc. Int'l. Conf. Information Visualization (IV '05)*, pages 175–181, 2005.

[225] T. Toutin. Qualitative aspects of chromo-stereoscopy for depth-perception. *Photogrammetric Engineering and Remote Sensing*, 63(2):193–203, 1997.

[226] J. Trapp and H.-G. Pagendarm. Data level comparative visualization in aircraft design. In *Proc. IEEE Visualization Conf. (Vis '96)*, pages 393–396, 1996.

[227] L. Treinish. Case study: Severe rainfall events in northwestern peru (visualization of scattered meteorological data). In *Proc. IEEE Visualization Conf. (Vis '94)*, pages 350–354, 1994.

[228] L. Treinish. A function-based data model for visualization. In *Proc. IEEE Visualization Conf. (Vis '98)*, pages 73–76, 1998.

[229] L. Treinish. Task-specific visualization design. *IEEE Computer Graphics and Applications*, 19(5):72–77, 1999.

[230] L. Treinish. Multi-resolution visualization techniques for nested weather models. In *Proc. IEEE Visualization Conf. (Vis 2000)*, pages 513–516, 2000.

[231] L. Treinish. Visual data fusion for applications of high-resolution numerical weather prediction. In *Proc. IEEE Visualization Conf. (Vis 2000)*, pages 477–480, 2000.

[232] E. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

[233] J. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[234] L. Tweedie, R. Spence, H. Dawkes, and H. Su. Externalising abstract mathematical models. In *Proc. Human Factors in Computing Systems (CHI 96)*, pages 406–412, 1996.

[235] F.-Y. Tzeng, E. Lum, and K.-L. Ma. An intelligent system approach to higher-dimensional classification of volume data. *IEEE Trans. Visualization and Computer Graphics*, 11(3):273–284, 2005.

[236] A. Unger, P. Muigg, H. Doleisch, and H. Schumann. Visualizing statistical properties of smoothly brushed data subsets. In *Proc. Int'l. Conf. Information Visualization (IV '08)*, pages 233–239, 2008.

[237] S. Uppala et al. ERA-40: ECMWF 45-year reanalysis of the global atmosphere and surface conditions 1957–2002. ECMWF Newsletter No. 101, 2004. European Centre for Medium-Range Weather Forecasts, UK.

[238] S. Uselton, J. Ahrens, W. Bethel, L. Treinish, and A. State. Multi-source data analysis challenges. In *Proc. IEEE Visualization Conf. (Vis '98)*, pages 501–504, 1998.

[239] T. van Walsum, F. H. Post, D. Silver, and F. J. Post. Feature extraction and iconic visualization. *IEEE Trans. Visualization and Computer Graphics*, 2(2):111–119, 1996.

[240] J. van Wijk. The value of visualization. In *Proc. IEEE Visualization Conf. (Vis 2005)*, pages 79–86, 2005.

[241] J. van Wijk. Bridging the gaps. *IEEE Computer Graphics and Applications*, 26(6):6–9, 2006.

[242] J. van Wijk and R. van Liere. HyperSlice: visualization of scalar functions of many variables. In *Proc. IEEE Visualization Conf. (Vis '93)*, pages 119–125, 1993.

[243] J. van Wijk and E. van Selow. Cluster and calendar based visualization of time series data. In *Proc. IEEE Symp. Information Visualization (InfoVis '99)*, pages 4–9, 1999.

[244] V. Verma and A. Pang. Comparative flow visualization. *IEEE Trans. Visualization and Computer Graphics*, 10(6):609–624, 2004.

[245] I. Viola, A. Kanitsar, and M. E. Gröller. Importance-driven feature enhancement in volume visualization. *IEEE Trans. Visualization and Computer Graphics*, 11(4):408–418, 2005.

[246] J. M. Wallace and P. V. Hobbs. *Atmospheric Science—An Introductory Survey.* Elsevier Academic Press, USA, 2006.

[247] C. Wang, H. Yu, and K.-L. Ma. Importance-driven time-varying data visualization. *IEEE Trans. Visualization and Computer Graphics*, 14(6):1547–1554, 2008.

[248] M. Ward. XmdvTool: Integrating multiple methods for visualizing multivariate data. In *Proc. IEEE Visualization Conf. (Vis '94)*, pages 326–336, 1994.

[249] M. Ward. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3/4):194–210, 2002.

[250] M. Ward. Multivariate data glyphs: Principles and practice. In C.-H. Chen, W. Härdle, and A. Unwin, editors, *Handbook of Data Visualization*, Springer Handbooks Comp. Statistics, pages 179–198. Springer, 2008.

[251] C. Weaver. Conjunctive visual forms. *IEEE Trans. Visualization and Computer Graphics*, 15(6):929–936, 2009.

[252] C. Weaver. Cross-filtered views for multidimensional visual analysis. *IEEE Trans. Visualization and Computer Graphics*, 16(2):192–204, 2010.

[253] R. Wilhelmson et al. A study of the evolution of a numerically modeled severe storm. *Int. J. Supercomput. Appl. High Perform. Eng.*, 4(2):20–36, 1990.

[254] M. B. Wilk and R. Gnanadesikan. Probability plotting methods for the analysis of data. *Biometrika*, 55(1):1–17, 1968.

[255] L. Wilkinson. *The Grammar of Graphics*. Statistics and Computing. Springer New York, Inc., 2005.

[256] D. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press, London, 1995.

[257] N. Willems, H. van de Wetering, and J. van Wijk. Visualization of vessel movements. *Computer Graphics Forum*, pages 959–966, 2009.

[258] A. Wilson and K. Potter. Toward visual analysis of ensemble data sets. In *Proc. Ultrascale Visualization Workshop*, pages 48–53, 2009.

[259] M. Wohlfart and H. Hauser. Story telling for presentation in volume visualization. In *Proc. Eurographics/IEEE-VGTC Symp. on Visualization (EuroVis 2007)*, pages 91–98, 2007.

[260] P. C. Wong. Visual data mining. *IEEE Computer Graphics and Applications*, 19(5):20–21, 1999.

[261] P. C. Wong and R. D. Bergeron. 30 years of multidimensional multivariate visualization. In *Scientific Visualization: Overviews, Methodologies, and Techniques*, pages 3–33. IEEE Computer Society, 1997.

[262] P. C. Wong, H. Foote, D. L. Kao, R. Leung, and J. Thomas. Multivariate visualization with data fusion. *Information Visualization*, 1(3/4):182–193, 2002.

[263] J. Woodring and H.-W. Shen. Chronovolumes: a direct rendering technique for visualizing time-varying data. In *Proc. Eurographics/IEEE-VGTC Workshop Volume Graphics*, pages 27–34, 2003.

[264] J. Woodring and H.-W. Shen. Multi-variate, time-varying, and comparative visualization with contextual cues. *IEEE Trans. Visualization and Computer Graphics*, 12:909–916, 2006.

[265] J. Woodring and H.-W. Shen. Multiscale time activity data exploration via temporal clustering visualization spreadsheet. *IEEE Trans. Visualization and Computer Graphics*, 15(1):123–137, 2009.

[266] J. S. Yi, Y. a. Kang, J. Stasko, and J. Jacko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE Trans. Visualization and Computer Graphics*, 13(6):1224–1231, 2007.

[267] H. Zhou, M. Chen, and M. Webster. Comparative evaluation of visualization and experimental results using image comparison metrics. In *Proc. IEEE Visualization Conf. (Vis 2002)*, pages 315–322, 2002.

# Errata

Page 24: “(paper refp:glyphs)” changed to “(paper B)”
Pages ix and 93: The citation of paper C was updated with the final information from the publisher