Eurographics/ IEEE-VGTC Symposium on Visualization 2010
G. Melançon, T. Munzner, and D. Weiskopf
(Guest Editors)

Computer Graphics Forum, 29(3):813-822, June 2010.
DOI: 10.1111/j.1467-8659.2009.01697.x
The definitive version is available at http://diglib.eg.org/ and http://onlinelibrary.wiley.com/

Brushing Moments in Interactive Visual Analysis

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

[1]Department of Informatics, University of Bergen, Norway, www.ii.uib.no/vis
[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^{th} vs. $(k+1)^{th}$ 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.*

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 [Shn96]. Visual analysis uses proven concepts such as coordinated multiple views, interactive feature specification via brushing, focus+context visualization, and on-demand data derivation [Hau06].

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 [HBSB02, MGKH09]. 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 [LPK05] (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 [Ham04].

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) [MMY06]. These data characteristics can be estimated traditionally or in a robust way [KW04, FMW08]. Additionally, measures of *outlyingness* help to identify extreme observations that substantially deviate from the rest [MMY06]. These interesting opportunities to analyze data distributions, however, also generate 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 data sets and there is a long history of related work [Tuf83]. Interestingly, the slightly younger history of visualization research relates back to early works that were inspired by considerations from statistics [Tuk77, CCKT83, CM88, Cle93]. Even systems for the visual data exploration can be traced back to these [War94, SCB98, TU08]. 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 [Rob07] for an overview). WEAVE [GRW*00] and SimVis [DGH03] 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 [War94]). Logical combinations of brushes in multiple linked views enable the specification of complex features [DGH03, Wea09]. The selection information is used to visually discriminate the specified features from the rest of the data in a focus+context visualization style [Hau05].

The treatment of *multi-run data* is rather new to the visualization community [LPK05]. 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 [CvN00]. Nocke et al. [Noc07, NFB07] 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. [MGKH09] 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. [KLDP02] 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 [KDP01] constructing a 3D volume with the probability density function (PDF) of the data as voxel values. Mathematical and procedural operators [LPK05] are proposed to transform multi-run data into a form where existing visualization techniques are applicable (e.g., pseudocoloring, 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. [PHBG09] 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 [MNS06, ODH*07] 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 [Wea09, Wea10] 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 μ) and the variability or dispersion (variance σ^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 $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, and the second moment by the empirical variance $s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Skewness can be estimated as $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3/s^3$, and kurtosis as $\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4/s^4-3$. 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 data sets 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* [HF96], 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 [Tuk77]: the lower or first quartile $q_1 = q(\frac{1}{4})$, the median or second quartile $q_2 = q(\frac{1}{2}) = \text{med}(x_1,\dots,x_n)$, and the upper or third quartile $q_3 = q(\frac{3}{4})$. Robust estimates for the standard deviation are the *interquartile range* IQR $= 0.741\cdot(q_3-q_1)$ and the *median absolute deviation*

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

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 σ of a normal distribution.

Two robust descriptors of the shape of the distribution are the *octile-based* skewness skew$_{oct}$—this is a special case of a quantile-based skewness coefficient [Hin75]—and an octile-based kurtosis measure kurt$_{oct}$ [Moo88]:

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

where $e_i = q(\frac{i}{8})$ is the i^{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 [FMW08]:

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

As for the classical estimates, the k^{th} moments (skew$_{\text{MAD}}$ and kurt$_{\text{MAD}}$ for $k=3,4$) are made comparable to the normal distribution, and thus $c_3=0$ and $c_4=3$ [FMW08]. 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 (skew$_{\text{MAD}}$, kurt$_{\text{MAD}}$ in Eq. 3) still include such outliers. Therefore, 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 [MMY06]:

$$z = \frac{x_i-\bar{x}}{s} \quad \text{and} \quad z_{\text{MAD}} = \frac{x_i-\text{med}(x_1,\dots,x_n)}{\text{MAD}(x_1,\dots,x_n)}. \qquad (4)$$

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 [RFG05].

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 [TC06]. 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 [WG68] or the spread vs. level plot [Tuk77]—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 [WH06]. 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 [BGM04]. 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 [BGM04].

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 [Ham04]). 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 data set with a total of 100 (10×10) runs. For every run, the data is given for 500 years on 2D sections (latitude × 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 data set is already challenging (5 dimensions, i.e., $3 \times$ 2D sections, time, and two run parameters with 10×10 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 [Cra05, Hel08]. 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 [NCIS02] 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 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 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 [KMSZ06]). 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:

1st moment	median ⇩ ⇧	$\mathcal{T}_{rob}$ ⇄	mean $\mathcal{T}_{ord}$ ⇩ ⇧	⇄ $\mathcal{T}_{rob}$	median ⇩ ⇧
2nd moment	MAD ⇩ ⇧	⇄	std.-dev. ⇩ ⇧	⇄	IQR ⇩ ⇧
3rd moment	skew_{MAD} ⇩ ⇧	⇄	skewness ⇩ ⇧	⇄	skew_{oct} ⇩ ⇧
4th moment	kurt_{MAD}	⇄	kurtosis	⇄	kurt_{oct}

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

- an **order transformation** $\mathcal{T}_{ord}(t_{ord}, m)$ is used to increment or decrement the k^{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 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

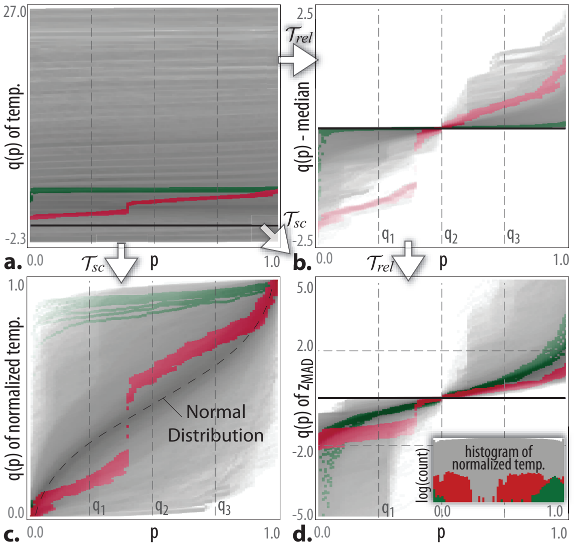

Figure 1: *Different quantile plots show distributions of multi-run data:* ***a*** *shows the original temperature values. The distances to the distribution's median are shown in* ***b****. This view is normalized by the* MAD *in* ***d*** *to identify outliers. The individual distributions in* ***a***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

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 $\mathcal{T}_{sc}(\text{norm}_z, a_y)$ applied to Fig. 1a, see Tab. 2). The plot in Fig. 1d is suitable for investigating outliers located above or below ± 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.

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).
norm_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 Φ 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* $\mathcal{T}_{sc}$*.*

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 indicated 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 [WG68] 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 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.

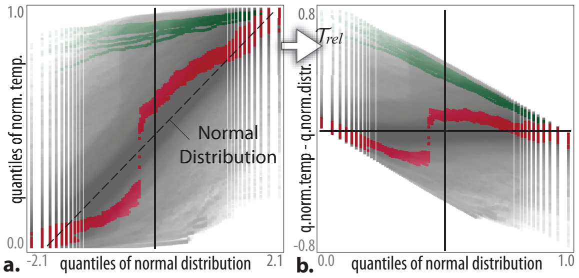

Figure 2: *A Q–Q (quantile-quantile) plot in **a** compares the sample distribution to a standard normal distribution. Applying a view transformation, deviations from the indicated line are investigated in a detrended Q–Q plot in **b**.*

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^{th} vs. $(k+1)^{\text{th}}$ moment (see Sec. 4.3.1). The view transformation $\mathcal{T}_{rob}$ is the next most important one, changing a traditional to a robust measure. Corresponding views of type k^{th} vs. k^{th} moment (traditional and/or robust measures) are discussed in Sec. 4.3.2. 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).

4.3.1. Views depicting the k^{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).

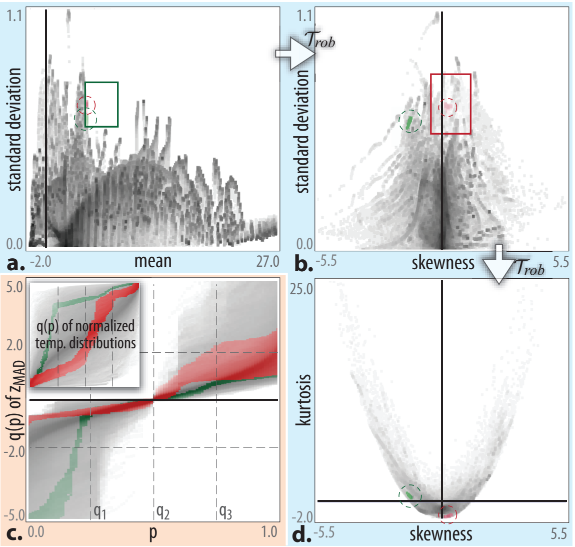

Figure 3: *Basic view setup showing combinations of all four moments in **a**, **b**, **d** (aggregated data part). The quantile plot in **c** is utilized to identify possible outliers. Interesting distributions are brushed and highlighted in color.*

Skewness and kurtosis form a pattern in Fig 3d, known as a Fleishman system [Fle78]. 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 corre-

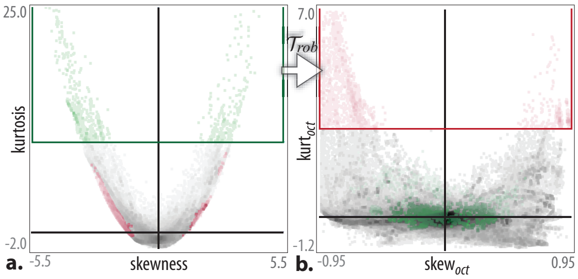

Figure 4: *Traditional vs. octile-based measures for skewness and kurtosis: High skewness values are brushed in **a** and apparently result from outliers since the corresponding robust measures in **b** (green) yield values closer to zero.*

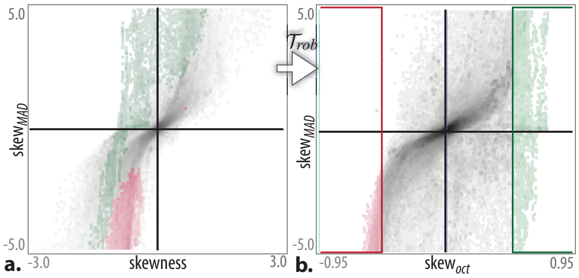

Figure 5: *Comparing traditional vs. median/*MAD*-based vs. octile-based skewness. Some of the green highlighted points with positive* skew_{oct} *selected in **b** even have a negative value for the traditional skewness in **b**.*

sponding 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 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).

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 $\sqrt{24/n}$ [Cra05] ($n = 100$, 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* [Tuk77] 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 [Tuk77] for further details).

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

Views of this category result from robustifying transformations of a k^{th} vs. k^{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 transformation 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., skew_{oct} based on octiles and skew_{MAD} based on the median/MAD). Samples approximately located along the diagonal are normally distributed. High absolute values for skew_{oct} are brushed in Fig. 5b. Some points with a positive 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 ±1 in the inset and are highlighted in red and green, respectively. The related distributions in Fig. 6b form an interesting pattern of

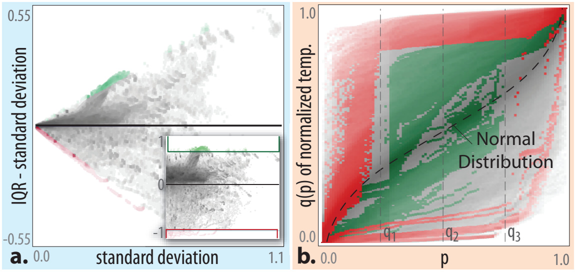

Figure 6: ***a*** *shows the result of a relating transformation applied to a standard deviation vs. IQR plot. Items along the diagonals are selected in a transformed view (inset) and correspond to distributions with a peaked shape in* ***b****.*

peakedness, which can be further investigated looking at the corresponding kurtosis values, for instance.

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 [BGM04]. A global sensitivity analysis (GSA) based on the Morris method [Mor91] 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* [DH02] 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 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.

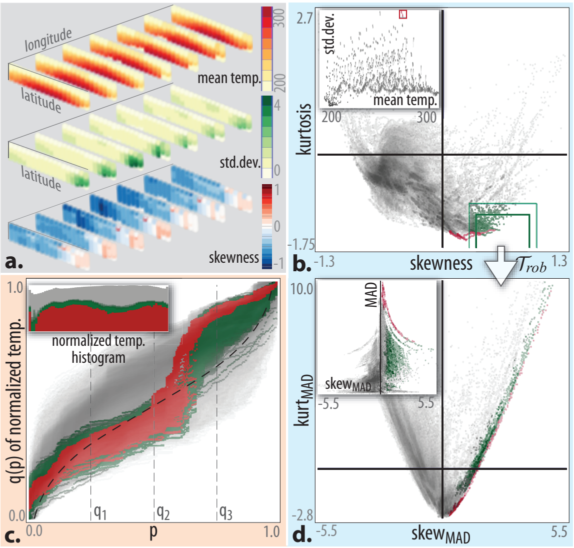

Figure 7: *The 3D atmosphere is shown in* ***a****, encoding mean, standard deviation, and skewness at timestep 80. Interesting data characteristics are brushed in* ***b*** *and refined in the inset, the corresponding distributions are investigated in a quantile plot in* ***c****. A robustified version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

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 skew$_{\text{MAD}}$ 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 [Mor91] of sampling the input parameter space.

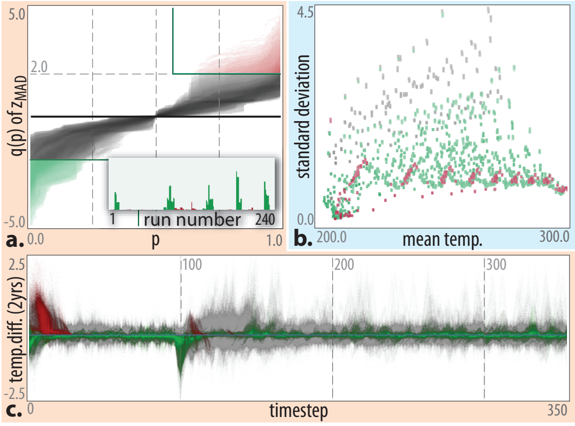

Figure 8: *Negative outliers selected in **a** form a repetitive pattern with respect to the run parameter (highlighted green in the inset). Positive outliers (red) form a repetitive pattern in the mean vs. standard deviation plot in **b**. The temporal evolution can be seen in a function graphs view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

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^{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 [KMSZ06]. 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 instance, 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.

References

[BGM04] Bauer E., Ganopolski A., Montoya M.: Simulation of the cold climate event 8200 years ago by meltwater outburst from Lake Agassiz. *Paleoceanography 19* (2004).

[CCKT83] Chambers J., Cleveland W. S., Kleiner B., Tukey P.: *Graphical Methods for Data Analysis*. Chapman and Hall, 1983.

[Cle93] Cleveland W. S.: *Visualizing Data*. Hobart Press, 1993.

[CM88] CLEVELAND W. C., MCGILL M. E. (Eds.): *Dynamic Graphics for Statistics*. Wadsworth, 1988.

[Cra05] CRAWLEY M.: *Statistics: An introduction using R*. Wiley, 2005.

[CvN00] COOKE R., VAN NOORTWIJK J.: Graphical Methods for Uncertainty and Sensitivity Analysis. In *Sensitivity Analysis*, Saltelli, Chan, Scott, (Eds.). Wiley, 2000, pp. 245–266.

[DGH03] DOLEISCH H., GASSER M., HAUSER H.: Interactive feature specification for focus+context visualization of complex simulation data. In *Proc. VisSym 2003* (2003), pp. 239–248.

[DH02] DOLEISCH H., HAUSER H.: Smooth brushing for focus+context visualization of simulation data in 3D. *Journal of WSCG 10*, 1 (2002), 147–154.

[Fle78] FLEISHMAN A.: A method for simulating non-normal distributions. *Psychometrika 43* (1978), 521–532.

[FMW08] FILZMOSER P., MARONNA R., WERNER M.: Outlier identification in high dimensions. *Computational Statistics & Data Analysis 52*, 3 (2008), 1694–1711.

[GRW*00] GRESH D., ROGOWITZ B., WINSLOW R., SCOLLAN D., YUNG C.: WEAVE: a system for visually linking 3-D and statistical visualizations applied to cardiac simulation and measurement data. In *Proc. IEEE Visualization* (2000), pp. 489–492.

[Ham04] HAMBY D. M.: A review of techniques for parameter sensitivity analysis of environmental models. *J. Environmental Monitoring & Assessment 32*, 2 (2004), 135–154.

[Hau05] HAUSER H.: Generalizing Focus+Context Visualization. In *Scientific Visualization: The Visual Extraction of Knowledge from Data* (2005), Springer, pp. 305–327.

[Hau06] HAUSER H.: Interactive visual analysis – an opportunity for industrial simulation. In *Simulation and Visualization* (2006), pp. 1–6.

[HBSB02] HIBBARD B., BÖTTINGER M., SCHULTZ M., BIERCAMP J.: Visualization in Earth System Science. *SIGGRAPH Comput. Graph. 36*, 4 (2002), 5–9.

[Hel08] HELTON J. C.: Uncertainty and sensitivity analysis for models of complex systems. In *Computational Methods in Transport: Verification and Validation*, vol. 62 of *Lecture Notes in Computational Science and Engineering*. 2008, pp. 207–228.

[HF96] HYNDMAN R. J., FAN Y.: Sample quantiles in statistical packages. *The American Statistician 50*, 4 (1996), 361–365.

[Hin75] HINKLEY D.: On power transformations to symmetry. *Biometrika 63* (1975), 101–111.

[KDP01] KAO D., DUNGAN J. L., PANG A.: Visualizing 2D probability distributions from EOS satellite image-derived data sets: a case study. In *Proc. IEEE Visualization 2001* (2001), pp. 457–460.

[KLDP02] KAO D., LUO A., DUNGAN J., PANG A.: Visualizing spatially varying distribution data. In *Proc. Intl. Conf. Information Visualization (IV '02)* (2002), pp. 219–225.

[KMSZ06] KEIM D., MANSMANN F., SCHNEIDEWIND J., ZIEGLER H.: Challenges in visual data analysis. In *Proc. Intl. Conf. Information Visualisation (IV '06)* (2006), pp. 9–16.

[KW04] KIM T.-H., WHITE H.: On more robust estimation of skewness and kurtosis. *Finance Res. Letters 1*, 1 (2004), 56–73.

[LPK05] LOVE A., PANG A., KAO D.: Visualizing spatial multivalue data. *IEEE Comput. Graph. Appl. 25*, 3 (2005), 69–79.

[MGKH09] MATKOVIĆ K., GRAČANIN D., KLARIN B., HAUSER H.: Interactive visual analysis of complex scientific data as families of data surfaces. *IEEE Trans. Vis. Comp. Graph. 15*, 6 (2009), 1351–1358.

[MMY06] MARONNA R., MARTIN D., YOHAI V.: *Robust Statistics: Theory and Methods*. John Wiley & Sons, 2006.

[MNS06] MÜLLER W., NOCKE T., SCHUMANN H.: Enhancing the visualization process with principal component analysis to support the exploration of trends. In *Proc. Asia Pacific Symp. on Information Visualisation (APVIS '06)* (2006), pp. 121–130.

[Moo88] MOORS J.: A quantile alternative for kurtosis. *The Statistician 37* (1988), 25–32.

[Mor91] MORRIS M.: Factorial plans for preliminary computational experiments. *Technometrics 33*, 2 (1991), 161–174.

[NCIS02] NORTH C., CONKLIN N., INDUKURI K., SAINI V.: Visualization schemas and a web-based architecture for custom multiple-view visualization of multiple-table databases. *Information Visualization 1*, 3-4 (2002), 211–228.

[NFB07] NOCKE T., FLECHSIG M., BÖHM U.: Visual exploration and evaluation of climate-related simulation data. In *Proc. Winter Simulation Conference* (2007), pp. 703–711.

[Noc07] NOCKE T.: *Visual data mining and visualization design for climate research*. PhD thesis, Dept. of Computer Science and Electrical Engineering, Univ. of Rostock, 2007. (in German).

[ODH*07] OELTZE S., DOLEISCH H., HAUSER H., MUIGG P., PREIM B.: Interactive visual analysis of perfusion data. *IEEE Trans. Vis. Comp. Graph. 13*, 6 (2007), 1392–1399.

[PHBG09] PATEL D., HAIDACHER M., BALABANIAN J.-P., GRÖLLER M. E.: Moment curves. In *Proc. IEEE Pacific Vis 2009* (2009), pp. 201–208.

[RFG05] REIMANN C., FILZMOSER P., GARRETT R.: Background and threshold: critical comparison of methods of determination. *Science of the Total Environment 346* (2005), 1–16.

[Rob07] ROBERTS J. C.: State of the art: Coordinated & multiple views in exploratory visualization. In *Proc. Coordinated & Multiple Views in Exploratory Visualization* (2007), pp. 61–71.

[SCB98] SWAYNE D. F., COOK D., BUJA A.: XGobi: Interactive Dynamic Data Visualization in the X Window System. *J. Computational and Graphical Statistics 7*, 1 (1998), 113–130.

[Shn96] SHNEIDERMAN B.: The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Visual Languages* (1996), pp. 336–343.

[TC06] THOMAS J., COOK K.: A visual analytics agenda. *IEEE Comput. Graph. Appl. 26*, 1 (2006), 10–13.

[TU08] THEUS M., URBANEK S.: *Interactive Graphics for Data Analysis: Principles and Examples*. Chapman & Hall, 2008.

[Tuf83] TUFTE E. R.: *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, Connecticut, USA, 1983.

[Tuk77] TUKEY J. W.: *Exploratory Data Analysis*. Addison-Wesley, 1977.

[War94] WARD M.: XmdvTool: Integrating multiple methods for visualizing multivariate data. In *Proc. IEEE Visualization* (1994), pp. 326–336.

[Wea09] WEAVER C.: Conjunctive visual forms. *IEEE Trans. Vis. Comp. Graph. 15*, 6 (2009), 929–936.

[Wea10] WEAVER C.: Cross-filtered views for multidimensional visual analysis. *IEEE Trans. Vis. Comp. Graph. 16*, 2 (2010), 192–204.

[WG68] WILK M. B., GNANADESIKAN R.: Probability plotting methods for the analysis of data. *Biometrika 55*, 1 (1968), 1–17.

[WH06] WALLACE J. M., HOBBS P. V.: *Atmospheric Science—An Introductory Survey*. Elsevier Academic Press, USA, 2006.