Hypothesis Generation in Climate Research with Interactive Visual Data Exploration

Johannes Kehrer, *Student Member, IEEE*, Florian Ladstädter, Philipp Muigg, *Member, IEEE*, Helmut Doleisch, *Member, IEEE*, Andrea Steiner, and Helwig Hauser, *Member, IEEE*

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, as well.

Index Terms— Interactive visual hypothesis generation, interactive visual exploration and analysis, visualization for climate research.

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 multivariate 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 [21], 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 [25]). Comparable to the "discover the unexpected"[TM], as coined by Cook and Thomas [24], 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 [4, 12, 19].

While *computational approaches* (e.g., 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 multivariate 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 [10].

- *Johannes Kehrer and Helwig Hauser are with the Department of Informatics, University of Bergen, Norway, E-mail: {johannes.kehrer, helwig.hauser}@uib.no.*
- *Florian Ladstädter and Andrea Steiner are with the Wegener Center for Climate and Global Change (WegCenter) and the Institute for Geophysics, Astrophysics, and Meteorology (IGAM), University of Graz, Austria, Email: {florian.ladstaedter, andi.steiner}@uni-graz.at.*
- *Philipp Muigg and Helmut Doleisch are with the VRVis Research Center and the SimVis GmbH, Vienna, Austria, Email: {muigg, doleisch}@simvis.at.*

IEEE Trans. Visualization and Computer Graphics, 14(6):1579-1586, 2008.
DOI: 10.1109/TVCG.2008.139

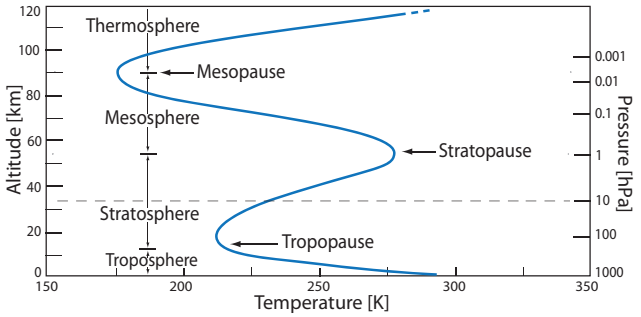

Fig. 1. Illustration of the vertical thermal structure of the atmosphere, reflecting a balance between radiative, convective and dynamical heating and cooling processes of the surface-atmosphere system. Different layers of the standard atmosphere are shown (illustration adapted from Melbourne et al. [13]). Changes in the upper troposphere-lower stratosphere region have strong impact on the Earth's climate system [27].

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 [27]. 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 and satellite-based measurements as well as global climate modeling in the last decades, data for upper air atmospheric variables are also available [23]. 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 [9, 5].

In the context of climate research, large multivariate 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 $1.875^\circ \times 1.875^\circ$ 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, multivariate data analysis, or pattern analysis, to name only three [29]. These methods usually require prior knowledge about the data to narrow down the scope of the analysis (e.g., parameters, boundary conditions).

[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).

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 deserves 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:

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

where g_N is the standard gravity at sea level, ϕ 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 [27]. 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[2] climate model simulation run [18] of the A2 scenario simulations for the Intergovernmental Panel on Climate Change (IPCC) 4th Assessment Report for the time period 1961 to 2064, as well as the geopotential height field of the ERA-40[3] reanalysis dataset [22] 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 20th 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 timeframes 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).

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 [2]. The area of coordinated and multiple views has been steadily developing over the past fifteen years. A good overview is given by Roberts [17]. A comprehensive overview on visual data mining and visualization techniques with respect to climate data is given by Nocke [15].

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

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

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 [2]. 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 [4], 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 [28]). 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 [6]). 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 [7].

In the SimVis system, *smooth brushing* [3] (enabling fractional DOI values) as well as the logical combination of brushes for the specification of *complex features* [2] 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* [2]. 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 [14], 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 γ 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*,

Fig. 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.

where we depict a scalar function over time for each voxel/cell of a volumetric and time-dependent dataset [14]. 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 [8, 16]. We use data aggregation (*frequency binmaps* [16] 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 [10], 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 [10]. 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

[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.

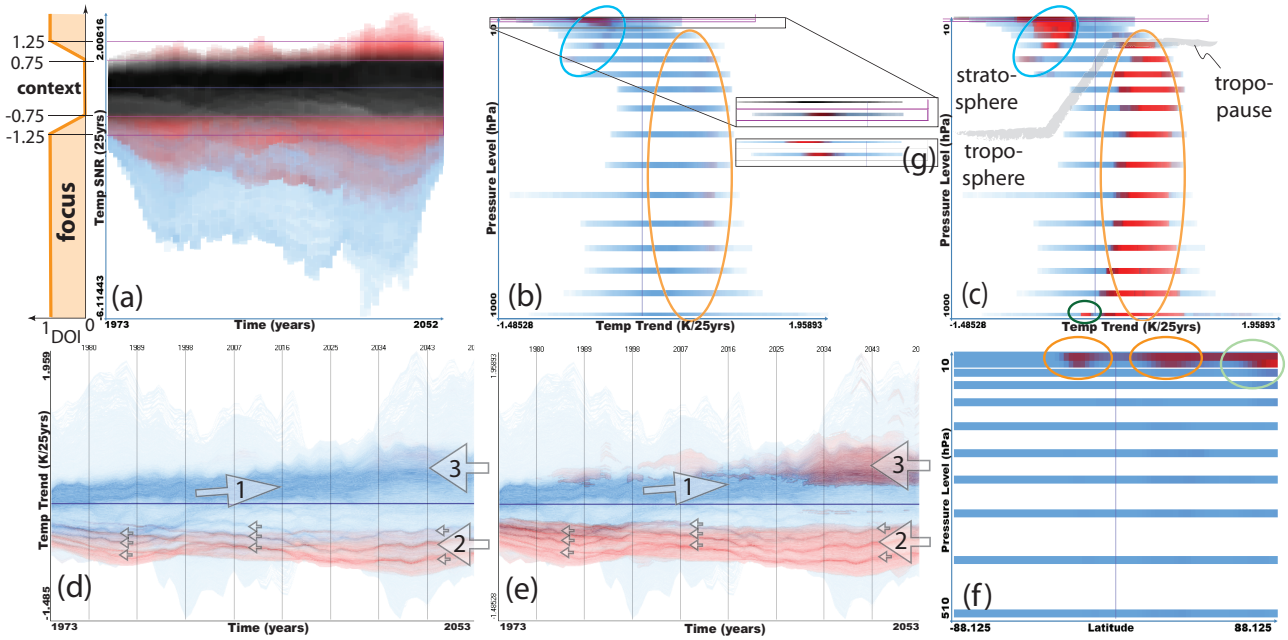

Fig. 3. Hypothesis generation using interactive visual exploration of derived temperature parameters in the ECHAM5 climate model. Features selected in multiple linked view are highlighted in red (focus), features only selected in the current view (2nd level focus) depicted in blue, and context information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

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. 3 (a) 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 separates data of significance (focus) from the context. So we take advantage of the smooth brushing [3] capability of SimVis assigning fuzzy degree-of-interest (DOI) values. Using a smooth NOT-brush (violet rectangle in Fig. 3 (a)) 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. 3 (a)).

As a next step we investigate the corresponding feature with respect to the height. The 2D scatterplot in Fig. 3 (b) 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. 3 (g)). However, according to the literature the ECHAM5 data set has known deficiencies in its highest pressure levels [1]. 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. 3 (b), also in the magnification above Fig. 3 (g)). A negative temperature trend with high significance is still highlighted in the remaining highest pressure levels (indicated by a blue ellipse in Fig. 3 (b) and (c)). 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. 3 (c). 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. 3 (b) 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 3 (c) also the tropopause is visible[6].

Figures 3 (d) and (e) 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. 3 (e) 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. 3 (e) (arrow 3) are highlighted where

[5] As a result, high negative SNR values in the lower part of Fig. 3 (a) 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. 3 (c).

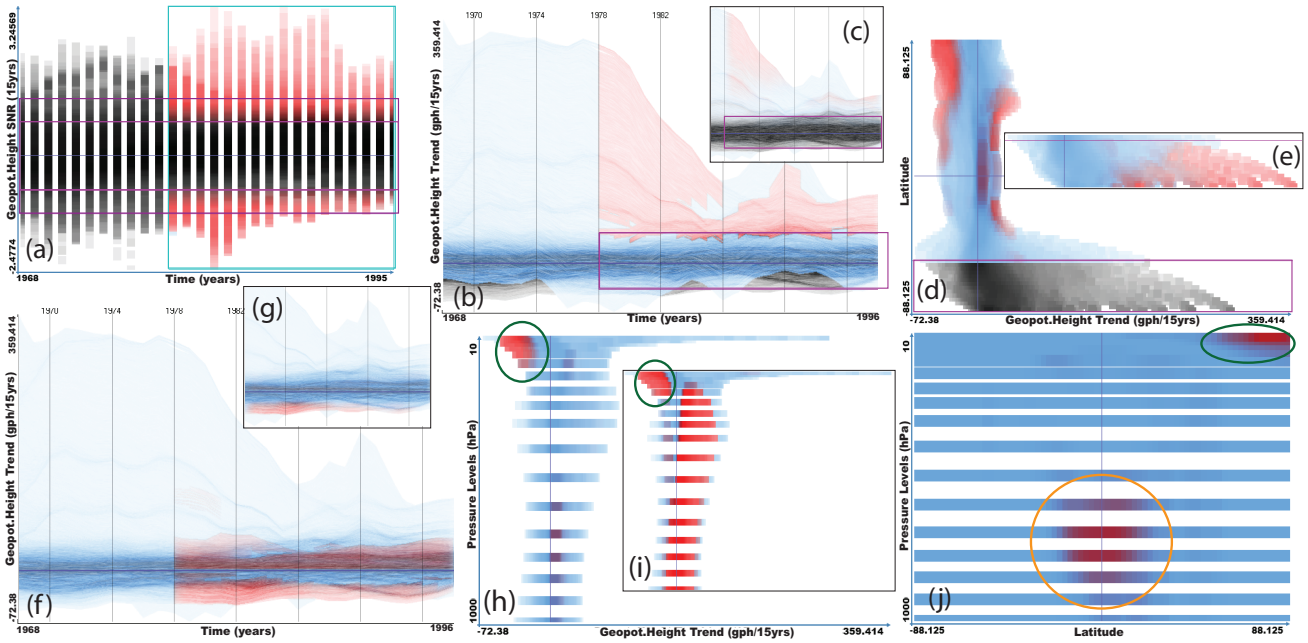

Fig. 4. Hypothesis generation on derived trend in geopotential height fields (ERA-40 reanalysis data set). (a) high SNR values over time (1968–1995) are brushed in a 2D scatterplot. The selection is restricted to the post-1979 era, where satellite measurements were incorporated. (b, c) similarity-based 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).

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 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. 3 (f) 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. 3 (b) and (c), 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).

[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.

ERA-40 Reanalysis Data

In our study, we also examine the geopotential height field of the ERA-40 reanalysis dataset [22] 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. 4 (a) 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. 4 (b)). According to the literature [26], 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. 4 (b) and (c), the main portion of the geopotential height trend is centered around the zeroline. We want to focus 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 Fig. 4 (b) and (c). 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. 4 (b) in order to allow the user to focus on all regions containing features (i.e., low γ value for the DOI enhancement). In order to show the actual significance of the feature it is depicted without enhancement in Fig. 4 (c).

The selection corresponding to the similarity NOT-brush is examined in a 2D 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. 4 (e)). According to Santer et al. [20] 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. 4 (d) 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. 4 (f) 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. 4 (g) 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. 4 (j), 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. 4 (h) and (i). 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 [10] 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 [29]). 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. [10]). 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 Fig. 5 (a) and (b) showing temperature trends (y-axis) vs. latitudes (x-axis), features in (b) 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. 5 (a). 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. 5 (b)). 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 sensitive 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. These examples show how the application of visual exploration techniques—used in an iterative process—contributed to an improved workflow in this application.

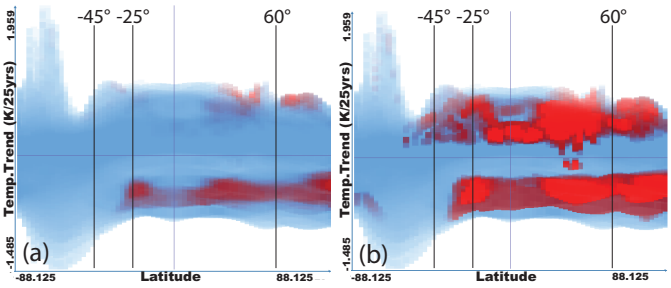

Fig. 5. ECHAM5: Sensitive regions with respect to climate change highlighted in the scatterplot (latitude on x, temperature trend on y-axis) were handed over to statistics for further analysis. In (a) the averaged DOI attribution are depicted in order to visualize the importance of each feature. On the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

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

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

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. [11]).

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. 6 (a) showing SNR values (y-axis) over time (x-axis). Comparing Fig. 6 (b) and Fig. 3 (b) 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.

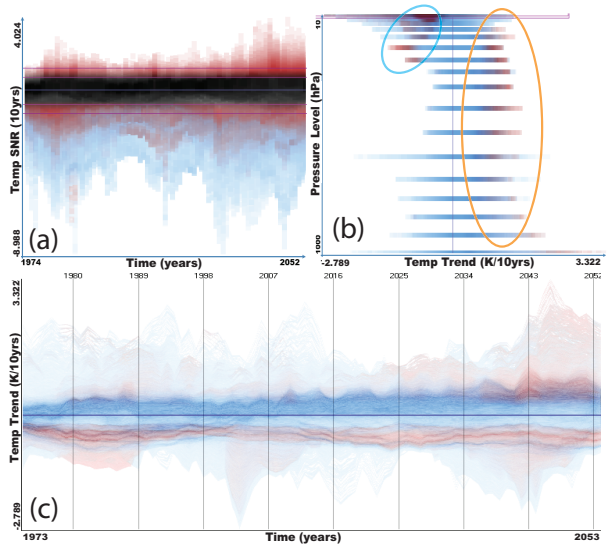

Fig. 6. ECHAM5 temperature: derived parameters computed over 10 years instead of 25 years. The features which were barely visible with 25 years (Fig. 3 (b)) are now highlighted in (b). The function plots of the derived temperature trend seem to contain a lot of noise.

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. 6 (c)). Here, no clear highlighted trends can be identified in the visualization, in contrast to Fig. 3 (d) and 3 (e), 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 Fig. 3 (d) and (e), indicated by small arrows). Using similarity-based brushing we can examine the interrelations between these visible trends and the other dimensions. In Fig. 7 (a) 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 2D scatterplot (SNR over time, see Fig. 7 (b)), 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. 7 (c)), 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. 3 (f)—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

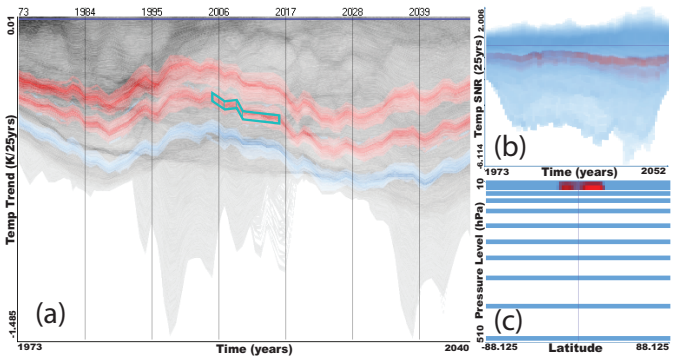

Fig. 7. A prominent visual structure in the function plots view is brushed based on it's similarity to a user defined target function (a). Three families of curves are thereby highlighted. The respective feature contains a relatively high signal-to-noise ratio highlighted in (b), and can be located in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

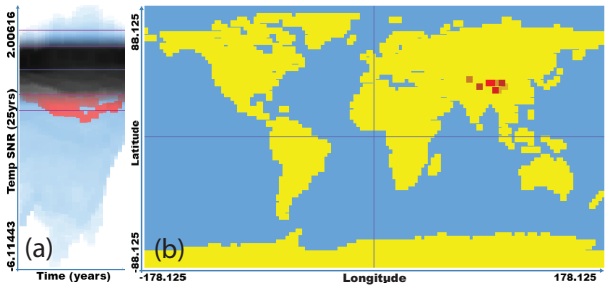

Fig. 8. Cooling trend brushed in the lowest pressure level indicated by a green ellipse in Fig. 3 (c) shows relatively low SNR values (a) and corresponds to a certain geographic area also including the Tibetan Plateau.

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 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. 3 (c)). 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. 8 (a)). 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. 8 (b), 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 $|\text{SNR}| \geq 1$), this feature would have been challenging to detect.

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.

Acknowledgements

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.

References

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

[2] H. Doleisch, M. Gasser, and H. Hauser. Interactive feature specification for focus+context visualization of complex simulation data. In *Proc. VisSym 2003*, pages 239–248, 2003.

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

[4] 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. VisSym 2004*, pages 91–96, 2004.

[5] 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.

[6] G. W. Furnas. Generalized fisheye views. In *Proc. ACM SIGCHI Conf. on Human factors in computing systems (CHI '86)*, pages 16–23, 1986.

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

[8] J. Johansson, P. Ljung, M. Jern, and M. Cooper. Revealing structure within clustered parallel coordinates displays. In *Proc. IEEE Symposium on Information Visualization 2004 (InfoVis 2005)*, pages 125–132, 2005.

[9] T. R. Karl, S. J. Hassol, C. D. Miller, and W. L. M. (Eds.). *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.

[10] 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).

[11] 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 *New Horizons in Occultation Research: Studies in Atmosphere and Climate*. Springer, 2008. (in press).

[12] 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 2005 (Vis '2005)*, pages 623–630, 2005.

[13] 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.

[14] 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. *Comput. Graph. Forum*, 27(3):775–782, 2008.

[15] T. Nocke. *Visuelles Data Mining und Visualisierungsdesign für die Klimaforschung*. PhD thesis, Inst. f. Computer Science, Dept. of Computer Science and Electrical Engineering, Univ. of Rostock, 2007. (in German).

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

[17] J. C. Roberts. Exploratory visualization with multiple linked views. In *Exploring Geovisualization*, pages 159–180. Elseviers, 2004.

[18] E. Roeckner et al. The atmospheric general circulation model ECHAM5. Report no. 349, Max-Planck-Inst. f. Meteorology, Hamburg, 2003.

[19] O. Rübel et al. PointCloudXplore: visual analysis of 3D gene expression data using physical views and parallel coordinates. In *Proc. Eurographics/IEEE-VGTC Symposium on Visualization*, pages 203–210, 2006.

[20] B. D. Santer et al. Identification of anthropogenic climate change using a second-generation reanalysis. *J. Geophys. Res.*, 109:D21104, 2004.

[21] H. Schumann and W. Müller. *Visualisierung – Grundlagen und allgemeine Methoden*. Springer, 2000. (in German).

[22] 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.

[23] S. Solomon et al. Technical summary. In *Climate Change 2007: The Physical Science Basis*. Cambridge University Press, UK, 2007.

[24] J. Thomas and K. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006.

[25] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.

[26] 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.

[27] J. M. Wallace and P. V. Hobbs. *Atmospheric Science—An Introductory Survey*. Elsevier Academic Press, USA, 2006.

[28] M. Ward. XmdvTool: Integrating multiple methods for visualizing multivariate data. In *Proc. IEEE Visualization '94*, pages 326–336, 1994.

[29] D. Wilks. *Statistical Methods in the Atmospheric Sciences*. Academic Press, London, 1995.