Exploration of Climate Data Using Interactive Visualization*

FLORIAN LADSTÄDTER, ANDREA K. STEINER, BETTINA C. LACKNER, BARBARA PIRSCHER, AND GOTTFRIED KIRCHENGAST

Wegener Center for Climate and Global Change (WegCenter), and Institute for Geophysics, Astrophysics and Meteorology/Institute of Physics, University of Graz, Graz, Austria

JOHANNES KEHRER AND HELWIG HAUSER

Department of Informatics, University of Bergen, Bergen, Norway

PHILIPP MUIGG AND HELMUT DOLEISCH

SimVis GmbH, Vienna, Austria

(Manuscript received 6 August 2009, in final form 25 November 2009)

ABSTRACT

In atmospheric and climate research, the increasing amount of data available from climate models and observations provides new challenges for data analysis. The authors present interactive visual exploration as an innovative approach to handle large datasets. Visual exploration does not require any previous knowledge about the data, as is usually the case with classical statistics. It facilitates iterative and interactive browsing of the parameter space to quickly understand the data characteristics, to identify deficiencies, to easily focus on interesting features, and to come up with new hypotheses about the data. These properties extend the common statistical treatment of data, and provide a fundamentally different approach. The authors demonstrate the potential of this technology by exploring atmospheric climate data from different sources including reanalysis datasets, climate models, and radio occultation satellite data. Results are compared to those from classical statistics, revealing the complementary advantages of visual exploration. Combining both the analytical precision of classical statistics and the holistic power of interactive visual exploration, the usual workflow of studying climate data can be enhanced.

1. Introduction

The ever-increasing demand for insight into the earth's climate system has led to an enormous amount of data produced in the last few decades by the climate research community. The increasing complexity and number of global climate models (GCMs), together with a great variety of observational data (Solomon et al. 2007), result in large, multivariate, and time-dependent datasets. Interactive visual exploration helps to quickly gain informative views on the data. It becomes a valuable complement to statistical data analysis methods.

Visualization can be seen in the context of more general data mining concepts. Data mining, vaguely defined in literature, refers to identifying useful patterns in large observed datasets (Fayyad et al. 1996; Friedman 1997; Goebel and Gruenwald 1999). In exploratory data analysis (EDA) as one data mining task (Tukey 1977), interactive visualization plays an important role (Wong 1999; Keim et al. 2002; de Oliveira and Levkowitz 2003). No assumptions of an underlying data model need to be presumed. This is in contrast to confirmative analysis methods, where visualization supports the verification of existing hypotheses about the data (Schumann and Müller 2000). Visual data exploration uses the unique ability of human vision to detect patterns and thus can help the users to identify relevant characteristics of the dataset and to ultimately come up with hypotheses about the data.

* Supplemental information related to this paper is available at the Journals Online Web site: http://dx.doi.org/10.1175/2009JTECHA1374.s1.

Corresponding author address: Florian Ladstädter, Wegener Center for Climate and Global Change (WegCenter), and Institute for Geophysics, Astrophysics, and Meteorology/Institute of Physics, University of Graz, Leechgasse 25, A-8010 Graz, Austria.
E-mail: florian.ladstaedter@uni-graz.at

DOI: 10.1175/2009JTECHA1374.1

In atmospheric and climate sciences, visualization is still most frequently applied in the form of simple unconnected plots (e.g., scatterplots, contour plots) to demonstrate characteristics of the data. Interactive visualization techniques for EDA tasks are hardly used so far. The reasons for this may be both a lack of suitable tools covering the specific needs of the geophysical community and that the available advanced visualization techniques are hardly known in this community (Hibbard et al. 2002; Nocke et al. 2008). Analyses were performed by Macêdo et al. (2000), who explored multivariate ocean–atmosphere datasets using the XGobi software tool (Swayne et al. 1998). In Doleisch et al. (2004), the interactive visual field exploration tool SimVis (Doleisch et al. 2003) was used to visualize a simulated meteorological dataset of the Hurricane Isabel, which struck the U.S. East Coast in 2003. SimVis was then later also applied to climate model data and to reanalysis datasets by Kehrer et al. (2008) and Ladstädter et al. (2009), showing the ability of the tool to explore large climate datasets. Hobbs et al. (2010) showed the benefits of visual data exploration in a case study using geographic and atmospheric data. Cuntz et al. (2007) presented a framework for interactive visualization of flow data output from climate model simulations. Interactive visualization was also used by Sukharev et al. (2009) to analyze correlations in time-varying multivariate climate data. For concise surveys over available tools and related work, see Hibbard et al. (2002), Nocke (2007), Aigner et al. (2008), Nocke et al. (2008), and Fuchs and Hauser (2009).

It is important to point out that classical statistical methods usually need a hypothesis beforehand to work. Features that were not known or at least anticipated beforehand are hard to find. Although providing meaningful quantitative analysis, these methods are less well suitable for the undirected search or for identifying hypotheses without prior knowledge. Classical statistics is therefore much better suited for hypothesis testing, whereas interactive visualization can aid in generating these hypothesis (Kehrer et al. 2008).

With interactive visualization techniques such as linked views and ''brushing'' (selecting) data items, interesting features of huge datasets can be effectively identified and interdependencies between parameters can be discovered. Local deviations in the datasets can be revealed more easily by visual exploration than by classical statistical methods, where the domain of analysis is usually limited to certain areas of interest according to some hypotheses. Exploring the dataset interactively might easily unveil unexpected features, leading to a new view on the data characteristics.

Atmospheric and climate researchers can benefit from these properties of exploration techniques. The potential to explore the whole dataset at once is often convenient when dealing with large atmospheric datasets. Finding interrelations between parameters that were not anticipated can help to discover unexpected features. This is not easily achieved with simple plots as they are commonly used in atmospheric and climate sciences.

In this study, we aim to introduce advanced interactive visualization to atmospheric, climatic, and oceanic scientists. We apply the technology to two climate datasets, investigating trend characteristics of atmospheric parameters. To further illustrate the utility, some of the results are compared to the outcome of classical statistical analysis of trend data. Furthermore, we investigate processing differences in remote sensing satellite data.

This paper is structured as follows: in section 2 the concepts and the tool employed are presented. The climate datasets used for demonstration are introduced in section 3. This is followed in section 4 by a number of example results showing the application of the tool and concepts to climate data, combined with some comparative results from classical statistical methods. The final section presents conclusions.

2. Interactive visual exploration methods

In this section, we present key elements of advanced visualization techniques. As outlined in the introduction, a previously formulated hypothesis is not needed to start working with the data. The aim is to enable the user to iteratively gain knowledge about interesting data characteristics. This knowledge ideally creates new questions or leads to hypotheses, which can then be explored and analyzed in more detail (Kehrer et al. 2008).

In this study, the interactive field exploration tool SimVis is applied (Doleisch et al. 2003). SimVis integrates advanced visualization technology and was originally developed for the exploration and analysis of large, multivariate 4D data resulting from computational fluid dynamics simulations as they are used, for example, in the automotive sector. The framework has only recently been applied to other areas as well, such as meteorological and climatological datasets (Doleisch et al. 2004; Kehrer et al. 2008). Even though it is able to handle large datasets, SimVis can be run on desktop PCs typical for a standard scientists workplace (a large monitor, as well as large RAM storage, is useful for smooth interactive manipulation). Major concepts realized in SimVis are summarized in the following. For a better understanding of the concepts, a video[1] is provided showing an example of interactive exploration.

[1] See online supplement.

a. Interactive feature-based visualization

This concept is founded on reducing the overall data to subsets (features) that exhibit properties of interest [e.g., regions with a high signal-to-noise ratio (SNR)]. To express interest in a certain feature, the user can brush parts of the data using the mouse. Brushing simply means to select data points directly within a visualization by, for example, defining or modifying selection rectangles in a space spanned by two parameters such as temperature and humidity (Becker and Cleveland 1987). A numerical refinement of the brush selection constraints is also possible.

b. Brushing and linking in multiple views

Multiple concurrently shown views present different aspects of the data. These views not only act as passive representation of parts of the dataset, they are also used to interactively explore and select the data. The different views include spatial 3D views and various types of 2D plots such as histogram, scatterplot, parallel coordinates view, and a curve view showing the variation of a parameter over time. Any of the parameters of the multivariate dataset can be assigned to the views. The user then expresses interest in a certain subdomain of the data via interactive brushing. Technically, this process assigns a degree of interest (DOI $\in [0, 1]$) attribute to every data point. For a discrete feature classification, this will be either 0 or 1 for context data and for brushed data, respectively. Through smooth brushing also fractional DOI values can be assigned (Doleisch and Hauser 2002). The information is immediately propagated and highlighted in all other views, providing the linking between the views and the displayed parameter spaces. With this brushing and linking concept, the user can apply constraints to the data and instantly check the resulting distribution in all other representations (Baldonado et al. 2000).

c. Focus and context visualization

Data items with high DOI values are considered to be in focus and are drawn in an emphasized way, whereas low DOI values lead to a reduced style representation (context). This distinction is especially useful in the 3D view to deal with visual clutter. Here, the focus and context concept helps to easily discriminate between the features and their context, with the latter shown in transparent gray. The concept is employed in all available views; for example, the scatterplot view emphasizes data items in focus by drawing them on top of the context items.

d. Derived-data attributes

Depending on the type of data to be explored, the features of interest might not be directly accessible from the dataset. Climate researchers are often interested in temporal trends derived from some existing parameters. This can be achieved through a flexible derived-data concept, where every available data parameter set can be the source of new derived parameters. To transform these parameters, a variety of predefined mathematical operations is available (e.g., algebra, derivatives). Arbitrary operations can be combined to user-defined formulas.

3. Datasets

As examples to demonstrate interactive visual exploration in the field of climate research, representative types of datasets are investigated: (i) reanalysis and model data and (ii) remote sensing data from radio occultation (RO).

a. Reanalysis and model trend data

The 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) seasonal-mean data[2] (Simmons and Gibson 2000) are used for the time period 1961–2002. Because it is based on observational data, incorporating many different meteorological observations, the time series is exposed to irregularities such as changing data sources, leading to varying data quality.

The second dataset consists of output of the fifth-generation atmosphere–ocean general circulation model (AOGCM) ECHAM5[3] (Roeckner et al. 2003), where a simulation of the B1 scenario of the Intergovernmental Panel on Climate Change (IPCC) Fourth Assessment Report is taken for the time period 2001–64. The run is chosen to be reasonably representative for a projected climate trend of the investigated time period, an evaluation that is based on trend calculations for several GCMs (Lackner et al. 2009). The simulation is complemented by the ECHAM5 twentieth-century run before 2001 (back to 1961) in the example, where necessary. The model output data can be expected to be without nonstationarity deficiencies in the majority of the data regime.

For these two datasets, June–August (JJA) seasonal-mean fields of temperature and geopotential height (corresponding to the geometric elevation above mean sea level normalized to standard earth surface gravity) on 18 pressure levels[4] ranging from 1000 to 10 hPa and on a horizontal resolution of $2.5° \times 2.5°$ are explored. For ECHAM5, the topmost level (10 hPa) is left out in

[2] Obtained from the ECMWF data server.

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

[4] The pressure levels are 1000, 925, 850, 775, 700, 600, 500, 400, 300, 250, 200, 150, 100, 70, 50, 30, 20, and 10 hPa.

the exploration because of known limitations (Cordero and de Forster 2006).

To access trend characteristics of both datasets, they are complemented using the derived-data functionality of SimVis (see also Ladstädter et al. 2009). Being an interactive exploration framework, SimVis does not provide statistical functions as needed for regression analysis. To be able to access meaningful features of the dataset, we therefore use the available mathematical operations to derive the following new parameters: We define the signal-to-noise ratio as the ratio of the trend to the detrended standard deviation. The linear trend is calculated as a moving difference on smoothed data, where the time periods used for both trend calculation and smoothing are 25 years for ECHAM5 and 15 years for ERA-40, respectively. All these computations can be done in the SimVis framework during the investigation process. They are applied separately for each of the space–time grid points within the framework.

More specifically, the data y of any parameter are first smoothed using a moving arithmetic averaging, resulting in y^{av}. The linear trend b_i is then computed centered for each time frame as a moving difference between the edge data points: that is, $b_i = N^{-1}(y^{\mathrm{av}}_{i+N/2} - y^{\mathrm{av}}_{i-N/2})$, where i denotes the center point of the time frame and N is the time period length. This gives the slope for the corresponding trend fit curve y^{FIT}_{ij} valid for all data points of the current time frame $i - N/2 \le j \le i + N/2$: that is, $y^{\mathrm{FIT}}_{ij} = y^{\mathrm{av}}_{i-N/2} + [j - (i - N/2)]b_i$. The trend fit curve is then used to remove the trend from the original data to obtain the detrended standard deviation s_i as a measure for the natural variability of the data for the current time frame:

$$s_i = \left[\frac{1}{N-1}\sum_{j=i-N/2}^{i+N/2}(y_j - y^{\mathrm{FIT}}_{ij})^2\right]^{1/2}. \qquad (1)$$

The SNR is then simply defined as the ratio of the trend signal to the variability of the data,

$$\mathrm{SNR}_i = \frac{b_i}{s_i}. \qquad (2)$$

This computation indicates well the utility of SimVis to provide derived-data attributes (here, SNR fields derived from basic fields of temperature and geopotential height) that are subsequently useful as complementary fields in the visual data exploration process.

b. GPS RO data

Global positioning system (GPS) RO satellite data originating from the Challenging Minisatellite Payload (CHAMP) mission (Wickert et al. 2004) are investigated as a further example. The GPS RO method is a remote sensing satellite technique delivering high-quality atmospheric datasets with long-term stability and high vertical resolution (e.g., Kursinski et al. 1997; Foelsche et al. 2009). The raw RO measurements, together with precise orbit data, are first transformed into phase delay data, from which atmospheric profiles are retrieved, such as refractivity (closely equivalent to density), geopotential height, and temperature. Although the RO technique possesses intrinsic self-calibrating properties, the processing of raw RO data leading to derived refractivity can introduce residual differences between the output of different processing centers (structural uncertainty; e.g., Ho et al. 2009).

In this study, differences of monthly-mean climatologies of refractivity retrieved by three different RO processing centers for the time period 2002–06 are explored. The refractivity climatology data used stem from the Wegener Center for Climate and Global Change (WegC; University of Graz, Austria) and further two processing centers: reference centers 1 (RefC1) and 2 (RefC2). The data of two processing versions of the Wegener Center are used: WegC1 as used by Ho et al. (2009) and WegC2 as used by Steiner et al. (2009). WegC1 used phase delay and orbit data from RefC1, and WegC2 used phase delay and orbit data from RefC2.

To compare the data of the four processing runs, monthly refractivity climatologies in 36 5° latitude bands (within 90°S–90°N) and at 200 altitude levels (from 0.2 to 40 km) are used. The difference between the individual datasets and their mean is explored, $\Delta x = x - \overline{x}$, with $\overline{x} = (x_1 + x_2 + x_3 + x_4)/4$. For the refractivity profiles, the relative difference $\Delta x/\overline{x}$ is used to account for the refractivity essentially decreasing exponentially with altitude. Not aiming at a concise study of the topic (for a study on RO structural uncertainties, see Ho et al. 2009), the exploration of these datasets shall serve as another example of fast hypothesis generation using interactive visualization.

4. Application to climate data

In this section, exemplary results for the exploration of atmospheric climate datasets are presented. A prominent topic in today's climate research is the detection of trends in atmospheric parameters caused by climate change. Although most measured data are available from the earth's surface, recent modeling efforts and a growing amount of data available for upper-air parameters through satellite and radiosonde measurements (Solomon et al. 2007; Steiner et al. 2009) show that the upper troposphere–lower stratosphere

(UTLS) region[5] reacts sensitively to climate change. We use interactive visual exploration to quickly get an overview over the various investigated datasets, to identify regions in time and space where robust indicators for climate change emerge, and to compare datasets from various radio occultation data processing centers. By comparing some of the results with classical statistics, we demonstrate the difference of the approaches and present an example where the exploratory method revealed a feature of the dataset not found otherwise. All figures presented (except one from classical statistics) are screenshots directly captured from the work with the SimVis tool (with trivial annotations added). Being an interactive tool, the shown features are inherently better represented on screen; the printed snapshot figures, in particular, show somewhat less color contrast.

a. Exploring reanalysis data

To demonstrate the basic steps of an interactive exploration session, we examine the geopotential height field of the ERA-40 dataset. In Fig. 1, a scatterplot (top center, showing the time channel versus the SNR channel) is used in the SimVis framework to select data points in space and time where the trend signal emerges from the data noise ($|\mathrm{SNR}| \geq 1$). Because we do not want to define a sharp boundary between data with high and low significance, in order not to lose any potential feature close to the boundary, we do take advantage of smooth brushing.[6] The smooth brush effectively assigns fractional DOI values ($\in]0, 1[$) to data points in the ranges $1.0 \leq \mathrm{SNR} \leq 1.5$ and $-1.5 \leq \mathrm{SNR} \leq -1.0$ [because SNR represents a trend signal to noise, as seen in Eq. (2), it can also be negative].

The result of the brushing is immediately shown in all other open views, where the selected data points are highlighted in red according to their DOI value. It is notable in the curve view (Fig. 1, bottom) that, although the bulk of data is located around the zero line (label 1), there are certain time series with high significance (in red, label 2, referred to as outliers in the following) that deviate from the main trends. The scatterplot versus pressure levels (Fig. 1, top right) shows that at the topmost levels the variation becomes much larger, and outliers with high significance and positive trends can be observed at stratospheric levels.

The outliers attract our interest, so we add another brush to the curve view, as shown in Fig. 2a. The brush selects all curves going through the selection box (turquoise rectangle) applied at the year 1974 layer. As a result, features selected in all views are shown in red, features selected only in the current view are shown in blue, and the context information (not selected) is shown in black. The two scatterplots in Fig. 2 show that the feature distribution resulting from the combined selection of outlier trends and high values of SNR is related to the stratosphere region (Fig. 2b) and to southern high latitudes (Fig. 2c). Having found this, literature research reveals that the high variation in southern high latitudes in the ERA-40 dataset is a spurious feature (Santer et al. 2004). The improvement of the data quality with time (as can be observed, e.g., in the curve view in Fig. 1) can be explained with the assimilation of satellite data after 1979 (Uppala et al. 2004). For ERA-40, such explanations are readily available. However, applying interactive visual exploration to new datasets can efficiently and effectively trigger new research on spotted characteristics of interest.

Based on what we learned on the outliers in Fig. 2, we proceed with applying a not-selection to all data south of 60°S and a selection of the post-1979 time period: that is, deselecting domains including data of inferior quality. This results in a distribution of data points with high significance, as shown in Fig. 3. The scatterplot in Fig. 3a shows a positive trend signal throughout the troposphere turning to a negative signal in the stratosphere. To locate the patterns of good trend signals geographically, the scatterplot in Fig. 3b[7] is underlaid with a land–sea mask showing the continents. High significance in the longitude–latitude distribution can be found above the continents in the tropics and at northern midlatitudes. Note that, in Fig. 3a, red (tropospheric) domains emerge visually that have only weakly been seen in the pressure level view of Fig. 1; this occurs, because in Fig. 1 the nonselected data (in blue) had quantitatively exceeded the selected data (in red) as a result of including all pre-1979 information.

b. Exploring climate model data

In this section, the geopotential height and temperature fields for a B1 scenario run of ECHAM5 are explored. To underpin the effectiveness of the approach, the findings are compared to results from statistical analyses for the same B1 run.

A first example of the exploration of the ECHAM5 dataset is shown in Fig. 4. The geopotential height of pressure levels is a good indicator for global warming, because the pressure surfaces will rise when there is thermal expansion of the underlying air masses. In Fig. 4a,

[5] The UTLS extends from about 5 to 35 km (≈500 to ≈10 hPa).

[6] The violet rectangle in Fig. 1, which is in fact a NOT-brush deselects the part then shown in black.

[7] Note that all pressure level data are integrated into this view.

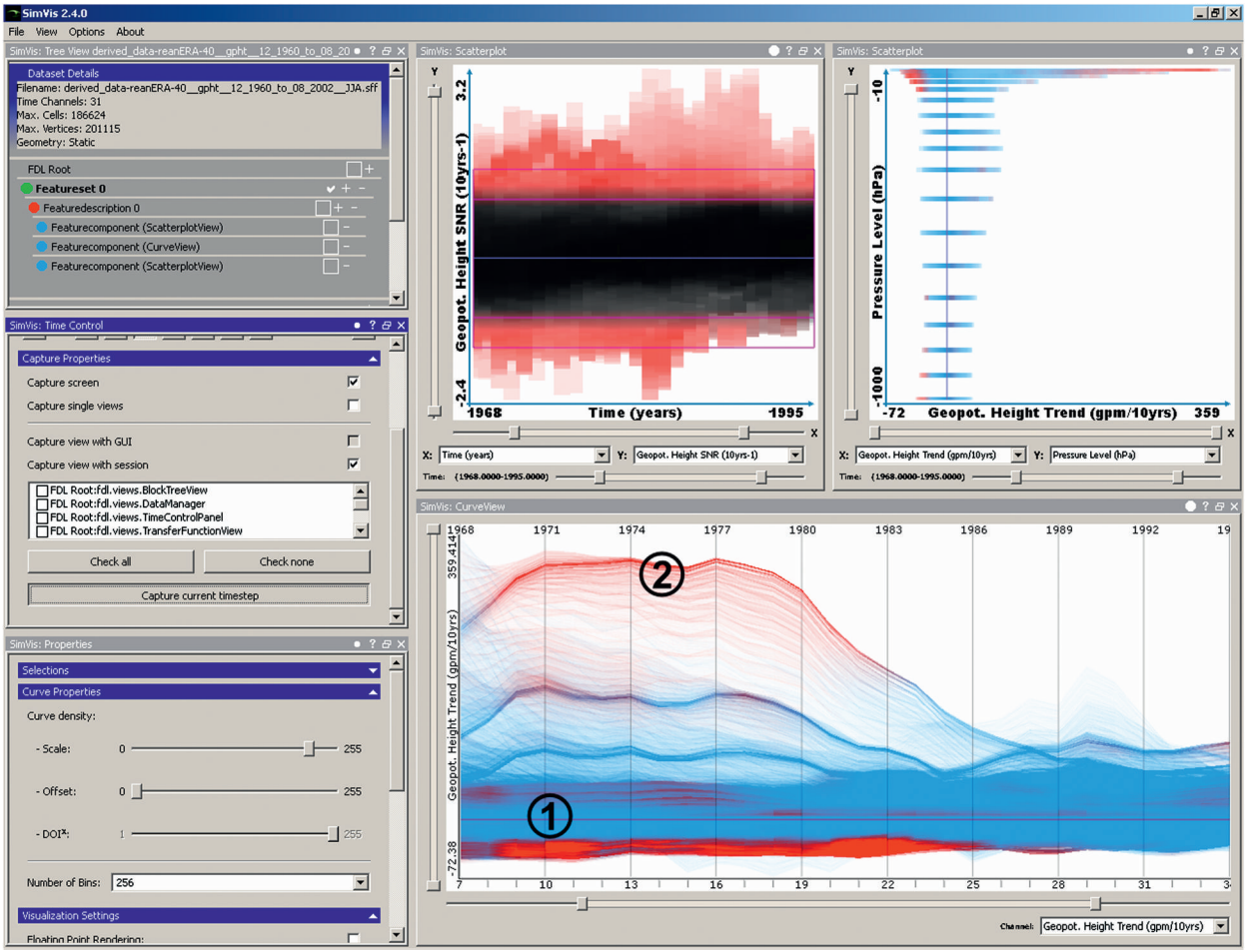

FIG. 1. Exemplary visual exploration session with SimVis, exploring the geopotential height field of the ERA-40 dataset. (top middle) Scatterplot view showing time vs SNR. Smooth brushing is applied to select data points of high significance (violet rectangle). (top right), (bottom) The other views are immediately updated, showing selected features in red, whereas the rest remains blue. (top right) Scatterplot view of the variation of the trend with pressure levels. (bottom) Curve view of the variation of the trend with respect to time with labeling of the zero line (1) and of some deviating trends (2).

the geopotential height field shows regions of high significance at the beginning and at the end of the 1961–2064 time period. The variation of the trend over time in Fig. 4b reveals a clearly visible reversion of the geopotential height trend from negative to positive. To further investigate this behavior, we select these features in subsequent scatterplots (not shown here) by brushing the first time period until 2001 and the second time period from 2001 to 2064 separately. The resulting pressure level distribution of the selected trend values is shown in Figs. 4c,d. Evidently, the significantly negative trends in the time period before 1990 stem from the stratosphere (Fig. 4c), where cooling lowers the pressure levels. Later, past about 2010, this effect becomes outweighed by the raising of the layers in the troposphere because of warming there and significantly positive trends emerge in the upper troposphere (Fig. 4d). Interactive visual exploration was found to be an effective approach for revealing these features.

Figure 5 gives a first overview of the trends in the temperature field, with high SNR values again depicted in red. Figure 5a shows the variation of the temperature trend over time from 2001 to 2064, where negative temperature trends show high significance throughout the whole time period. Positive trends become more significant around 2030 (recall that these are trends at individual grid points with no area or height layer averaging). It is important to emphasize again that the purpose of the interactive visual exploration is not to deliver quantitative results (as is usually the case in statistics) but rather to quickly come up with hypotheses about the data and to explore the data without stating precise quantitative findings. Figure 5b indicates that the most sensitive region for detecting temperature trends

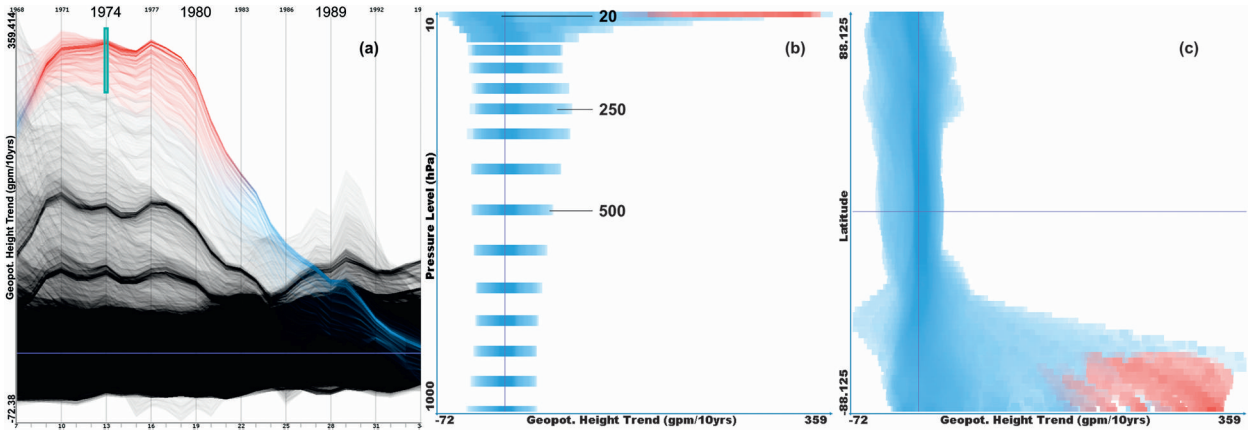

FIG. 2. Inspecting outliers of the geopotential height trend field of ERA-40. Brushing outliers (turquoise rectangle) in (a) the curve view and scatterplot views as a function of (b) pressure levels and (c) latitude, with the selected outliers in red. Some pressure levels are labeled for better orientation.

can be found in the UTLS, with negative trends in the lower stratosphere and positive trends in the upper troposphere. The longitude–latitude distribution of significant trends is shown in Fig. 5c (Northern Hemisphere summer most visible in the JJA trends); in Fig. 5d, the respective latitude ranges for the positive (tropospheric) and negative (stratospheric) trends are highlighted. Interactive visualization proved effective for this fast first overview of the dataset.

Next, we aim to compare our results with classical statistical trend testing (Wilks 2006). Using classical statistical analysis, linear trends were calculated over 2001–50 at each pressure level, and the significance of the trends was determined using Student's t test. Figure 6a illustrates the regions for which this analysis was performed. For regions and pressure levels fulfilling the criteria of statistical significance above 90% and goodness-of-fit coefficient R^2 (the coefficient of determination) above 0.5, Fig. 6b shows the corresponding box colored (for details on the analysis, see Lackner et al. 2009). We compare trends over 2001–50 only, because a linear fit is not an adequate approximation for the changing trend over the whole time period starting in 1961.

For comparison of the features shown in Fig. 5 with statistical results, we choose spatial subdomains according to the regions used in Fig. 6. This is one of the major strengths of interactive visual exploration: While, in statistical analysis, it can be difficult to keep an eye on the whole domain at the same time, generally forcing to preselect areas of interest, the visual exploration can be done on the whole dataset at once, selecting features interactively and iteratively.

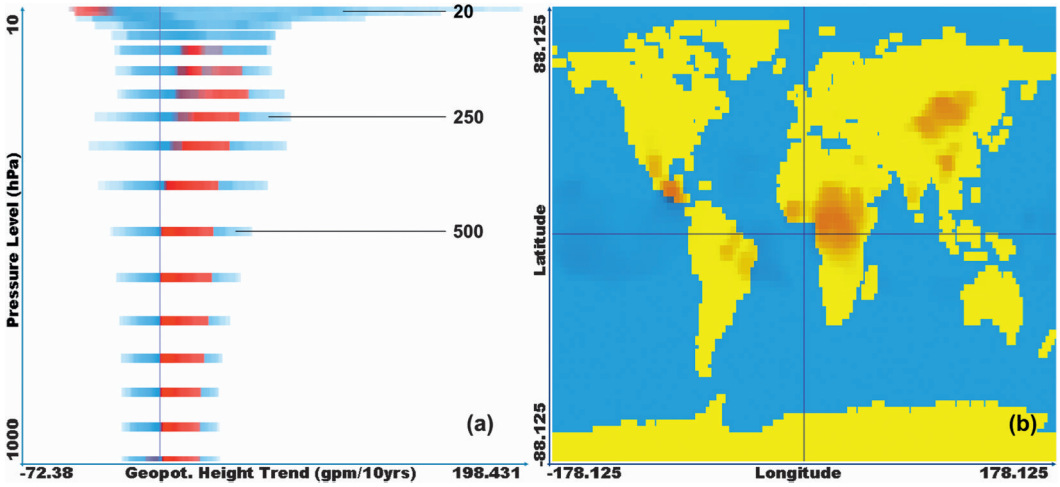

FIG. 3. Removing the southern high latitudes and restricting the analysis to the post-1979 time. Scatterplot view as a function of (a) pressure levels and (b) geographic location, with the selected data in red.

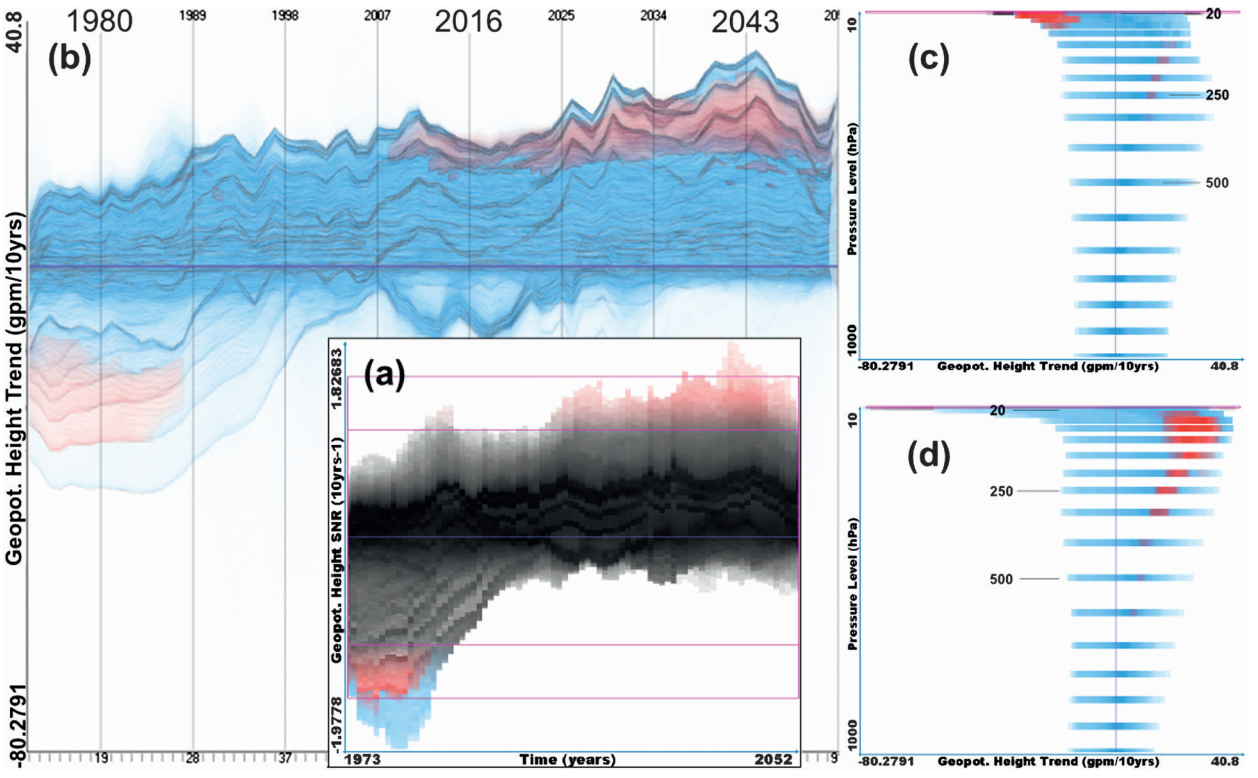

FIG. 4. Exploration of the geopotential height field of an ECHAM5 run. (a) As in Fig. 1, high SNR values are brushed. (b) A curve view shows the time variation of the geopotential height trend (1961–2064), and scatterplot views of trends vs pressure levels show the data for the time periods (c) 1961–2001 and (d) 2001–64 separately, with the selected data in red.

Figure 7 illustrates this. We choose the Canada–Greenland–Iceland (CGI) region as an example, where we brush the region in a longitude–latitude scatterplot (Fig. 7a). The resulting distribution of values showing high SNR as a function of pressure levels (Fig. 7b) can immediately be compared with the statistical result in Fig. 6b, where a significant negative temperature trend signal is shown for the topmost levels 30 hPa upward and

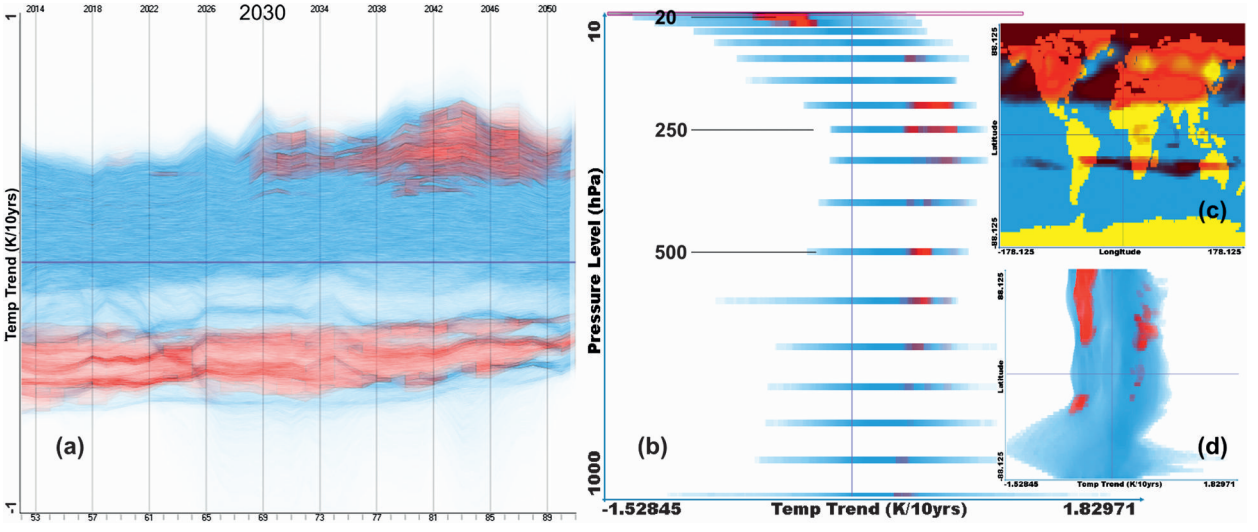

FIG. 5. Exploration of the temperature field of the ECHAM5 run. High SNR values have been brushed as in Fig. 4 (not shown). Shown are (a) a curve view of temperature trends (2001–64) and (b) scatterplots of temperature trends vs pressure levels, as function of (c) geographic location and (d) vs latitude, with the selected data in red.

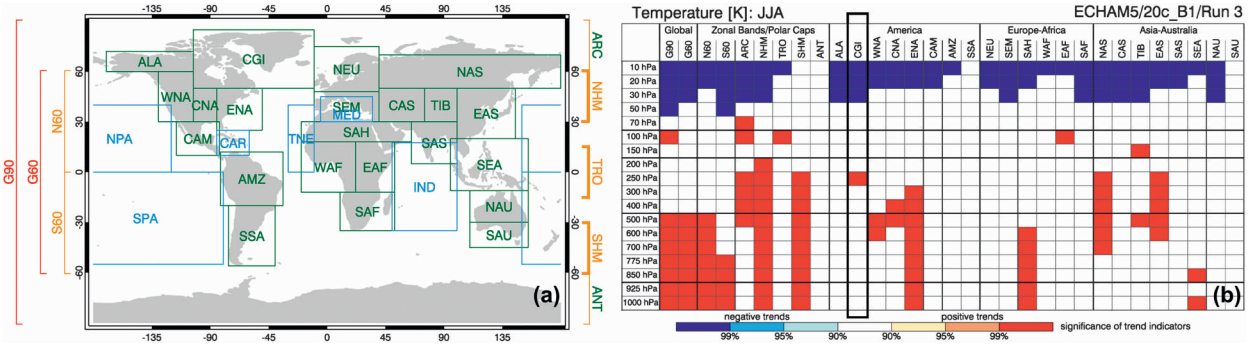

FIG. 6. (a) Regions used in a classical statistical analysis, corresponding mostly to the IPCC regions (Solomon et al. 2007; Lackner et al. 2009) and (b) temperature trend results for the ECHAM5 run for the time period 2001–50. In (b), boxes of regions with statistically significant positive trends are marked in red, and those with significant negative trends are marked in blue.

a positive trend is shown at the 250-hPa level. This indeed agrees very well with Fig. 7b, where the data points in red are clustered at the topmost levels and the 250-hPa level. Some less pronounced features can also be seen in other levels (better visible on screen). Figure 7c shows that the trend significance over time is very stable at stratospheric levels, whereas the upper-tropospheric signal appears most significant for the temporal center of the moving trend estimates around the year 2030.

Other investigated regions (not shown) compare similarly well with statistical results. This gives us confidence that the interactive visual approach, although not meant to give quantitatively precise results, is able to serve as a good exploratory complement to the classical statistical approach.

As a final example for the climate model data exploration illustrated in Fig. 8, we focus on the distinct band of data points with high SNR in the Southern Hemisphere subtropics in Fig. 5c. Brushing the range of about 20°–30°S (Fig. 8a), we see that it stems from the stratosphere with negative trend values (Fig. 8b; less pronounced features on, e.g., the near-surface 1000-hPa layer with positive trend values can be seen on screen). We regard it as a good example for the power of visual, undirected exploration that this band was not part of the zonal bands defined in Fig. 6a [there is a gap between the tropical (TRO) and the Southern Hemisphere mid-latitudes (SHM) band; see legend at the right of Fig. 6a] and that this pronounced feature has therefore been ''overlooked'' in the statistical analysis based on this definition.

c. *Exploring GPS RO data*

To introduce the RO datasets, the climatological temperature profiles of the WegC1 processing are shown in Fig. 9. The typical tropical temperature profile shape emerges from the set of profiles very clearly (Fig. 9a). A higher variation of the profiles can be seen at high latitudes (Figs. 9b,c), with more isothermal shape than expected. Somewhat less variability occurs over the Arctic

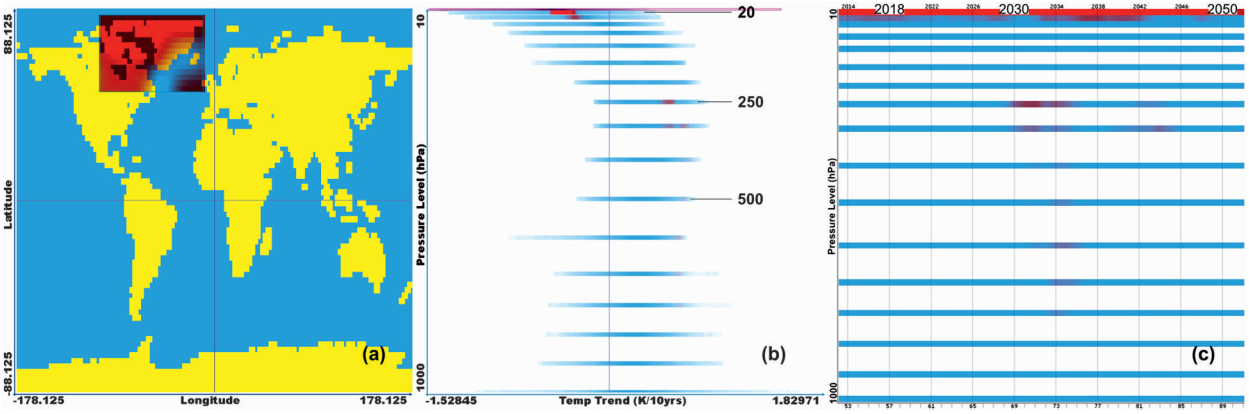

FIG. 7. Inspecting the CGI region of Fig. 6: (a) brushing the region (highlighted rectangle); (b) scatterplot of temperature trends vs pressure levels; and (c) curve view showing the time variation of trend significances, with the selected data in red.

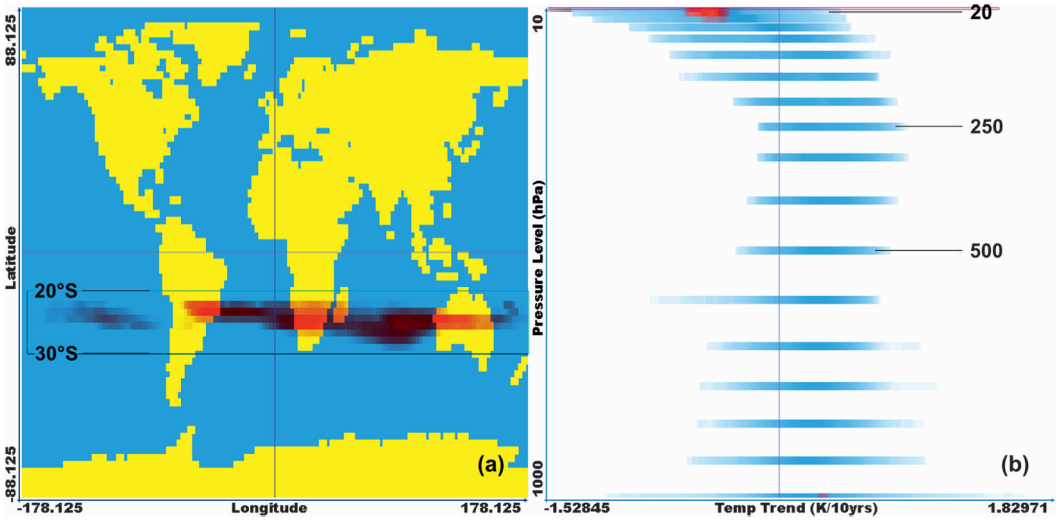

FIG. 8. Inspecting the feature in the Southern Hemisphere subtropics found in Fig. 5. Shown are scatterplots as function of (a) geographic location (with the 20°–30°S band marked) and (b) pressure levels, with the selected data in red.

region (Fig. 9b) than over the Antarctic one (Fig. 9c), where also the absolutely coldest conditions prevail.

Having set this context, we explore relations between refractivity climatologies of the four processing versions (WegC1, WegC2, RefC1, and RefC2) shown in Fig. 10. The number of contributing individual profiles to each monthly-mean climatological profile differs between the four datasets because of different quality criteria used in the processing systems. In Figs. 10a–d, the relation between the relative difference of refractivities to the four-dataset mean and the corresponding difference of the number of profiles to the four-dataset mean number of profiles is shown. While RefC1 and RefC2 numbers (Figs. 10a,b) lie tentatively above average, the WegC processings (Figs. 10c,d) show a number of profiles tentatively below average. The reason is that the WegC processing uses somewhat more rigid quality criteria.

The relative differences show no obvious systematics in this view, so we inspect them further. The parallel coordinates view shows a set of chosen parameters as vertical coordinate axes that are placed next to each other (Inselberg and Dimsdale 1990). Each multivariate data point is then represented as a polyline that connects the corresponding data values on the parallel axes. This makes it a suitable representation for detecting interdependencies between the parameters. For example, if most of the line segments between two adjacent coordinate axes are parallel to each other, there is a strong correlation between the attributes represented by the

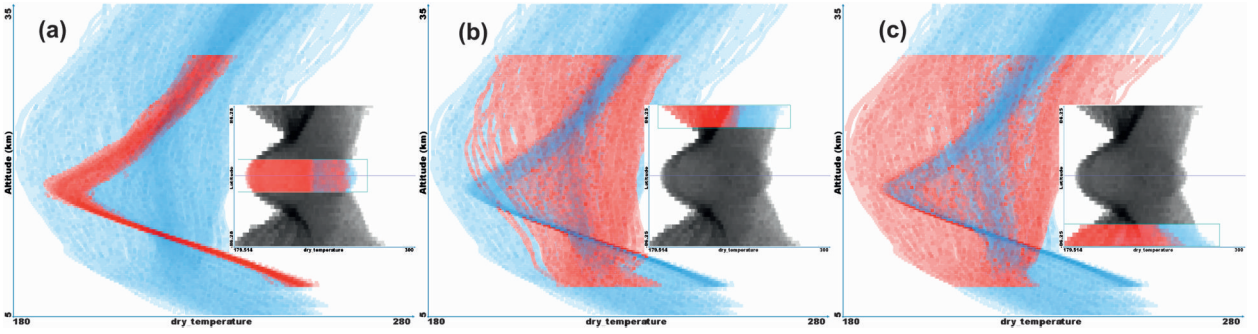

FIG. 9. Zonal mean monthly-mean temperature profiles (5–35 km) from the CHAMP GPS RO data for the time period 2002–06, processed by WegC1 (cf. section 3). The brushes (shown in the inset) selected (a) the tropics and the (b) northern and (c) southern high latitudes, and the selected data are shown in red. Data at altitude levels outside 8–30 km are not in red because the corresponding refractivity data of all centers overlap fully in 8–30 km only: removed by a NOT-brush in another view (not shown).

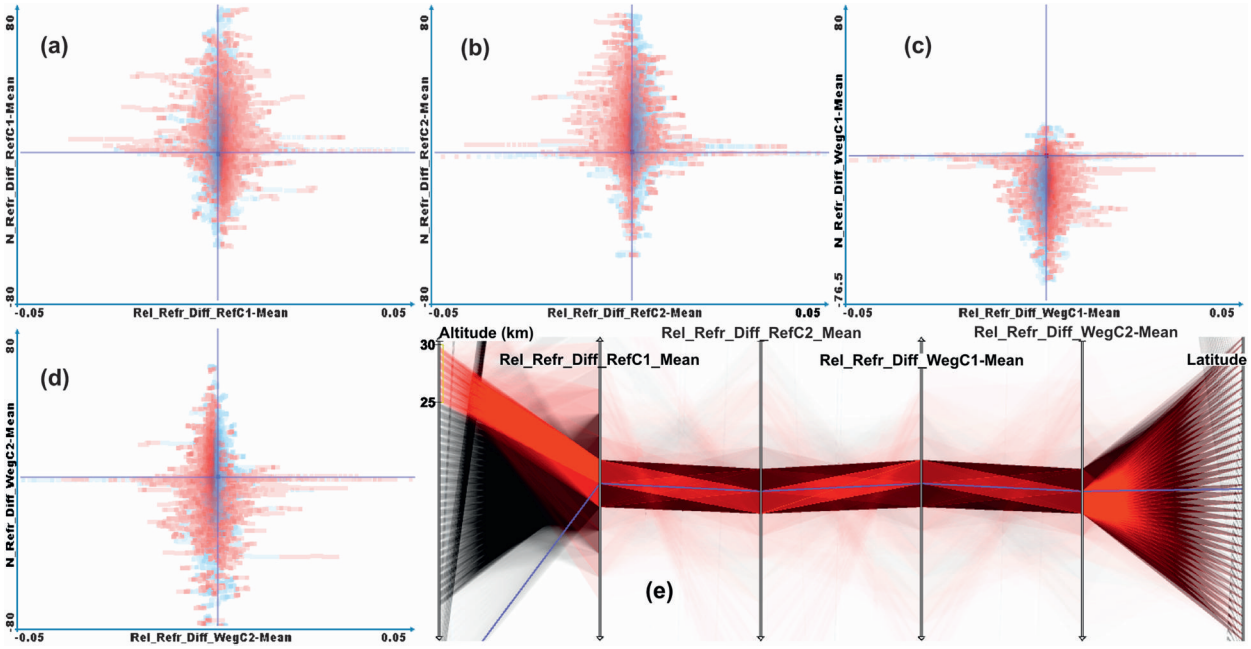

FIG. 10. Relation between the number of profiles and relative refractivity differences against the four-dataset mean for (a) RefC1, (b) RefC2, (c) WegC1, and (d) WegC2. Data outside 8–30 km deselected in a previous step (not shown) appear in blue, and selected data appear in red. (e) A parallel coordinates view (explained in the text) was brushed to select the altitude range from about 25 to 30 km (leftmost coordinate), with the selected data in red. The other coordinates show relative refractivity differences of each dataset (four middle) and latitude (right). The blue line connects the (zero) origins of all coordinates.

respective axis. The view allows brushing for each coordinate separately. Figure 10e shows the result when choosing the altitude as one parameter of interest (left), followed by the relative refractivity differences to the four-dataset mean (from left to right) of RefC1, RefC2, WegC1, and WegC2 and the latitude as last coordinate (right). For the selected 25–30-km altitude range, RefC1 and WegC1 show a tendency to larger relative refractivities than the mean and vice versa for RefC2 and WegC2. This behavior is understood by recalling from section 3 that the main difference between the processing versions of WegC1 and WegC2 lies in the switch from using the raw data processed in a first step by RefC1 to the raw data processed by RefC2. The opposite tendency of the WegC1 and the WegC2 refractivity data compared to the mean suggests that the resulting climatologies from RO data are mainly influenced by this first processing step toward the phase delay and orbit data. A possible dependence on latitude could be easily spotted (rightmost coordinate) but appears not to exist.

Practically, such small residual processing differences, especially above 25 km, are well understood by the RO science community (e.g., Ho et al. 2009; Steiner et al. 2009) and further mitigated by improved processing. Here, it served as an instructive example of how interactive visualization could spot these fast and effectively in a multicenter, multiyear remote sensing dataset for climate.

5. Conclusions

The amount of geophysical data available to scientists grows at a remarkable pace. The instrument to handle such large multivariate datasets is most commonly classical statistics. As an innovative approach, interactive visual exploration is shown to be a valuable technology to explore atmospheric and climate data and to be complementary to statistical analysis. It uses the power of human vision to detect features and relations within multiple parameters of a multidimensional dataset.

Here, we presented some of these visualization concepts to the atmospheric, climate, and oceanic research community. Using the field exploration technology SimVis, the potential of undirected interactive exploration was shown for three representative example datasets, including reanalysis data, climate model data, and GPS radio occultation (RO) satellite data. We performed interactive exploration of atmospheric climate fields of temperature, geopotential height, and refractivity. Our approach efficiently detected deficiencies in the reanalysis dataset and identified parameters and regions reacting most sensitive to climate change (for more details on this topic, see

Kehrer et al. 2008; Ladstädter et al. 2009). Small differences between satellite datasets were spotted quickly and effectively in the context of quality assurance for climate applications of RO data.

We exploited that the visual exploration approach is clearly distinctive to classical statistical methods in that it does not require prior knowledge about the dataset. Features that attract the attention of the user can be interactively selected and iteratively explored in more detail. The resulting hypotheses about the data can then be quantitatively evaluated and confirmed by statistical analysis. We showed that the properties of the approach make it easy to get an overview over general characteristics of the dataset, to focus on certain subdomains of the parameter space, to generate hypotheses about the data, and to even spot relevant features that might have been overlooked in statistical analysis. Comparison with statistical results confirmed the utility of the technology, showing that interactive exploration forms a valuable complement to classical statistics, permitting a fast, efficient, and instructive analysis of the data.

Visualization enabling a direct and interactive access to atmospheric, climatic, and oceanic datasets can support the process of gaining insight into the data characteristics. It is easy to visually analyze and explore features of interest such as trends, differences between datasets, or interdependencies between available parameters. Many more applications can be envisioned, dependent on the specific dataset; we see the potential to apply the presented approach to other climatological and geophysical datasets, beyond the example types of data used in this study. In conclusion, we encourage the combination of the instruments of classical statistical analysis with advanced exploratory data analysis methods in an integrated workflow, helping to handle the vast amount of today's data output with which geoscientific research is dealing.

Acknowledgments. The authors acknowledge GFZ (Potsdam, Germany) and UCAR (Boulder, CO) for CHAMP RO data and ECMWF (Reading, United Kingdom) for the ERA-40 reanalysis data. We furthermore acknowledge the modeling groups, the PCMDI and WRCP Working Group on Coupled Modelling (WGCM), for access to ECHAM5 data as part of CMIP3 multimodel data. SimVis GmbH is thanked for its efforts in SimVis development (available online at http://www.simvis.at), and J. Fritzer, M. Borsche, and U. Foelsche (Wegener Center) are thanked for their efforts in RO processing development and climatology preparations. The work was financed by Austrian Science Fund (FWF) Grants P18733-N10 and P18837-N10; RO processing development was also by ESA and the Austrian Research Promotion Agency (FFG-ALR) via projects ProdexCN2 and EOPSCLIM.

REFERENCES

Aigner, W., S. Miksch, W. Müller, H. Schumann, and C. Tominski, 2008: Visual methods for analyzing time-oriented data. *IEEE Trans. Visualization Comput. Graphics,* **14,** 47–60.

Baldonado, M. Q. W., A. Woodruff, and A. Kuchinsky, 2000: Guidelines for using multiple views in information visualization. *Proc. Workshop on Advanced Visual Interfaces,* Palermo, Italy, ACM, 110–119, doi:10.1145/345513.345271.

Becker, R. A., and W. S. Cleveland, 1987: Brushing scatterplots. *Technometrics,* **29,** 127–142, doi:10.2307/1269768.

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

Cuntz, N., A. Kolb, M. Leidl, C. R. Salama, and M. Böttinger, 2007: GPU-based dynamic flow visualization for climate research applications. *Simulation und Visualisierung 2007 (SimVis 2007),* SCS Publishing House, 371–384.

de Oliveira, M. C. F., and H. Levkowitz, 2003: From visual data exploration to visual data mining: A survey. *IEEE Trans. Visualization Comput. Graphics,* **9,** 378–394.

Doleisch, H., and H. Hauser, 2002: Smooth brushing for focus + context visualization of simulation data in 3D. *Proc. WSCG,* Plzeň, Czech Republic, Eurographics, 147–154.

——, M. Gasser, and H. Hauser, 2003: Interactive feature specification for focus + context visualization of complex simulation data. *Proc. Symp. on Data Visualisation (VISSYM '03),* Aire-la-Ville, Switzerland, Eurographics Association, 239–248.

——, P. Muigg, and H. Hauser, 2004: Interactive visual analysis of Hurricane Isabel—A description in more detail. VRVis Research Center Tech. Rep. TR-VRVis-2004-058, 6 pp.

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

Foelsche, U., B. Pirscher, M. Borsche, G. Kirchengast, and J. Wickert, 2009: Assessing the climate monitoring utility of radio occultation data: From CHAMP to FORMOSAT-3/COSMIC. *Terr. Atmos. Oceanic Sci.,* **20,** 155–170.

Friedman, J. H., 1997: Data mining and statistics: What's the connection? *Proc. 29th Symp. on the Interface: Computing Science and Statistics,* Houston, TX, Interface Foundation of North America, 7 pp.

Fuchs, R., and H. Hauser, 2009: Visualization of multi-variate scientific data. *Comput. Graphics Forum,* **28,** 1670–1690.

Goebel, M., and L. Gruenwald, 1999: A survey of data mining and knowledge discovery software tools. *ACM SIGKDD Explorations Newsletter,* No. 1, Association for Computing Machinery, New York, NY, 20–33.

Hibbard, B., M. Böttinger, M. Schultz, and J. Biercamp, 2002: Visualization in earth system science. *ACM SIGGRAPH Comput. Graphics Quarterly,* Vol. 36, No. 4, Association for Computing Machinery, New York, NY, 5–9, doi:10.1145/637357.637361.

Ho, S.-P., and Coauthors, 2009: Estimating the uncertainty of using GPS radio occultation data for climate monitoring: Intercomparison of CHAMP refractivity climate records from 2002 to 2006 from different data centers. *J. Geophys. Res.,* **114,** D23107, doi:10.1029/2009JD011969.

Hobbs, J., H. Wickham, H. Hofmann, and D. Cook, 2010: Glaciers melt as mountains warm: A graphical case study. *Comput. Stat.,* in press.

Inselberg, A., and B. Dimsdale, 1990: Parallel coordinates: A tool for visualizing multi-dimensional geometry. *Proc. First Conf. on Visualization,* San Francisco, CA, IEEE Computer Society, 361–378.

Kehrer, J., F. Ladstädter, P. Muigg, H. Doleisch, A. Steiner, and H. Hauser, 2008: Hypothesis generation in climate research with interactive visual data exploration. *IEEE Trans. Visualization Comput. Graphics,* **14,** 1579–1586.

Keim, D., W. Mueller, and H. Schumann, 2002: Visual data mining. Eurographics 2002 State of the Art Reports, 49–68.

Kursinski, E. R., G. A. Hajj, J. T. Schofield, R. P. Linfield, and K. R. Hardy, 1997: Observing earth's atmosphere with radio occultation measurements using the global positioning system. *J. Geophys. Res.,* **102,** 23 429–23 465.

Lackner, B. C., A. K. Steiner, F. Ladstädter, and G. Kirchengast, 2009: Trend indicators of atmospheric climate change based on global climate model scenarios. *New Horizons in Occultation Research: Studies in Atmosphere and Climate,* A. K. Steiner, et al., Eds., Springer, 247–259.

Ladstädter, F., A. K. Steiner, B. C. Lackner, G. Kirchengast, P. Muigg, J. Kehrer, and H. Doleisch, 2009: SimVis: An interactive visual field exploration tool applied to climate research. *New Horizons in Occultation Research: Studies in Atmosphere and Climate,* A. K. Steiner et al., Eds., Springer, 235–245.

Macêdo, M., D. Cook, and T. Brown, 2000: Visual data mining in atmospheric science data. *Data Min. Know. Discovery,* **4,** 69–80.

Nocke, T., 2007: Visuelles Data Mining und Visualisierungsdesign für die Klimaforschung (Visual data mining and visualization design for climate research). Ph.D. thesis, University of Rostock, 246 pp.

——, T. Sterzel, M. Böttinger, and M. Wrobel, 2008: Visualization of climate and climate change data: An overview. *Digital Earth Summit on Geoinformatics 2008: Tools for Global Change Research (ISDE'08),* Ehlers et al., Eds., Wichmann, 226–232.

Roeckner, E., and Coauthors, 2003: The atmospheric general circulation model ECHAM5: Part 1. Max-Planck-Institute for Meteorology Rep. 349, 140 pp.

Santer, B. D., and Coauthors, 2004: Identification of anthropogenic climate change using a second-generation reanalysis. *J. Geophys. Res.,* **109,** D21104, doi:10.1029/2004JD005075.

Schumann, H., and W. Müller, 2000: *Visualisierung—Grundlagen und allgemeine Methoden*. Springer, 370 pp.

Simmons, A. J., and J. K. Gibson, 2000: The ERA-40 project plan. European Centre for Medium-Range Weather Forecasts ERA-40 Project Report Series No. 1, 63 pp.

Solomon, S., D. Qin, M. Manning, M. Marquis, K. Averyt, M. M. B. Tignor, H. L. Miller Jr., and Z. Chen, Eds., 2007: *Climate Change 2007: The Physical Science Basis*. Cambridge University Press, 996 pp.

Steiner, A., G. Kirchengast, B. C. Lackner, B. Pirscher, M. Borsche, and U. Foelsche, 2009: Atmospheric temperature change detection with GPS radio occultation 1995 to 2008. *Geophys. Res. Lett.,* **36,** L18702, doi:10.1029/2009GL039777.

Sukharev, J., C. Wang, K.-L. Ma, and A. T. Wittenberg, 2009: Correlation study of time-varying multivariate climate data sets. *Proc. Second Pacific Visualization Symp.,* Beijing, China, IEEE VGTC, 161–168.

Swayne, D. F., D. Cook, and A. Buja, 1998: XGobi: Interactive dynamic data visualization in the X window system. *J. Comput. Graphical Stat.,* **7,** 113–130.

Tukey, J. W., 1977: *Exploratory Data Analysis*. Addison-Wesley, 688 pp.

Uppala, S., and Coauthors, 2004: ERA-40: ECMWF 45-year reanalysis of the global atmosphere and surface conditions 1957–2002. *ECMWF Newsletter,* No. 101, ECMWF, Reading, United Kingdom, 2–21.

Wickert, J., T. Schmidt, G. Beyerle, R. König, C. Reigber, and N. Jakowski, 2004: The radio occultation experiment aboard CHAMP: Operational data analysis and validation of vertical atmospheric profiles. *J. Meteor. Soc. Japan,* **82,** 381–395.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 648 pp.

Wong, P., 1999: Visual data mining. *IEEE Comput. Graph. Appl.,* **19,** 20–21.