Environmental Earth Sciences (2019) 78:294
https://doi.org/10.1007/s12665-019-8295-z

THEMATIC ISSUE

An interactive visual analysis tool for investigating teleconnections in climate simulations

Anatoliy Antonov[1] · Gerrit Lohmann[2] · Monica Ionita[2] · Mihai Dima[3] · Lars Linsen[4]

Received: 30 November 2018 / Accepted: 26 April 2019

Abstract
Teleconnections refer to links between regions that are distant to each other, but nevertheless exhibit some relation. The study of such teleconnections is a well-known task in climate research. Climate simulation shall model known teleconnections. Detecting teleconnections in climate simulations is a crucial aspect in judging the quality of the simulation output. It is common practice to run scripts to execute a sequence of analysis steps on the climate simulations to search for teleconnections. Such a scripting approach is not flexible and targeted towards one specific goal. It is desirable to have one tool that allows for a flexible analysis of all teleconnection patterns with a dataset. We present such a tool, where the extracted information is provided in an intuitive visual form to users, who then can interactively explore the data. We developed an analysis workflow that is modeled around four views showing different facets of the data with coordinated interaction. We present a teleconnection study with simulation ensembles and reanalysis data obtained by data assimilation to observe how well the teleconnectivity patterns match and to demonstrate the effectiveness of our tool.

Keywords Interactive visual analysis · Teleconnections · Coordinated views · Spatial data visualization · Multidimensional data projection · Segmentation

Introduction

A common task in climate simulation analysis is to validate the simulation outcome by relating it to measured data often in the form of reanalysis data, i.e., after a data assimilation step. One important phenomenon in climate observations is teleconnectivity. Teleconnections represent relationships between different spatial regions in the climate system such that a certain condition in one region sets an expectation for the condition in the connected regions. For example, increase of pressure at one location may correspond to decrease of pressure at another. Various studies identified a number of prominent patterns, e.g., Walker and Bliss (1932). It is of interest to investigate how well climate simulations reproduce these patterns, and how much the configurations of the patterns change over time.

The analysis of teleconnections is commonly performed by running scripts written for a particular situation, where each change in parameters requires restarting the script. It makes exploration of the data slow and cumbersome. This script-based approach lies in the foundation of web-based tools for data post-processing (Climate Reanalyzer 2011; KNMI Climate Explorer 2013; PSD Web Products and Tools 2019). To the best of our knowledge, there exists no standard software addressing a common workflow for teleconnection research in an interactive and intuitive way, which is also reported in a recent survey on visual analytics of climate networks by Nocke et al. (2015).

In this paper, we present an analysis tool that is based on visual representations of correlation and teleconnectivity information extracted from climate data and interactions that allow for a systematic, user-centric data analysis workflow. The analytical workflow is described in Section "Data analysis workflow". The interaction with coordinated views,

This article is part of a Topical Collection in Environmental Earth Sciences on "Visual Data Exploration", guest edited by Karsten Rink, Roxana Bujack, Stefan Jänicke, and Dirk Zeckzer.

✉ Lars Linsen
linsen@uni-muenster.de

[1] Jacobs University, Bremen, Germany

[2] Alfred Wegener Institute, Bremerhaven, Germany

[3] University of Bucharest, Bucharest, Romania

[4] Westfälische Wilhelms-Universität Münster, Münster, Germany

Published online: 08 May 2019

the visual encodings used for the views, and the information that is extracted for the visual encoding are presented in Section "Data analysis components". For feature extraction, we employ one of the main methods of analyzing patterns in the atmosphere, which is based on the use of the correlation coefficient (Wallace and Gutzler 1981). It defines teleconnectivity for a point as the absolute value of the most negative correlation between the time series of this point and others. Plotting teleconnectivity values for all points together in a map provides a way to see the strongest centers of teleconnectivity and understand their relationships with each other. The one-point correlation maps for these centers reveal the patterns.

We present a real-world scenario by applying our tool to the data from the twentieth century Reanalysis Project and the respective time period extracted from the COSMOS simulations of the last millennium. In Section "Application scenarios", we describe known teleconnectivity patterns, provide details on the reanalysis data and the simulation ensemble, and demonstrate the capabilities of the tool for the comparison of datasets, the extraction of patterns, and the study of the pattern's components. Strengths and limitations of the tool are discussed in Section "Discussion and future work".

Data analysis workflow

The analytical workflow to perform teleconnection pattern analysis using our tool is presented in Fig. 1. The statistical foundation of the presented approach is constructed on the ideas of Wallace and Gutzler (1981): Based on the correlation coefficients between time series, we determine teleconnectivity values.

Thus, the first step in our pipeline is to compute correlation. We can interactively analyze correlations between spatial locations by deploying the interactive correlation map, a map-based visual encoding, where the user can interactively select a reference point and observe correlation patterns with respect to the reference point.

Based on the pairwise correlation, we also compute correlation chains to build groups with strong (positive or negative) correlation. Thus, instead of analyzing a single reference point, we observe regions. Such regions can also be interactively explored in the correlation map.

However, computing correlation chains is a local feature extraction mechanism. To explore the pairwise correlations of all locations in the map we propose a different layout, where the points are not placed due to spatial locations, but due to correlation similarity. Since the layout is computed from the matrix of pairwise similarities using a multidimensional scaling projection, we refer to the layout as projection view.

The second branch of our workflow takes the correlation one step further by computing teleconnections from them. Then, we can also visually encode teleconnections in a map layout called teleconnectivity map.

Based on the teleconnections we can also group spatial locations to form teleconnectivity regions. The regions are defined by applying a thresholding on the teleconnectivity values. The threshold can be interactively modified, allowing for a systematic approach to narrowing down on higher teleconnectivity regions (or widening to lower connectivities respectively). The teleconnectivity regions can be shown in the teleconnectivity map and are also listed in the teleconnectivity links view.

Figure 1 summarizes the analytical workflow using our tool, where the rhombi illustrate interactive input during the feature extraction and visual encoding steps. Once all four views have been created, the full potential of our tool is exploited by coordinated interaction with all four views.

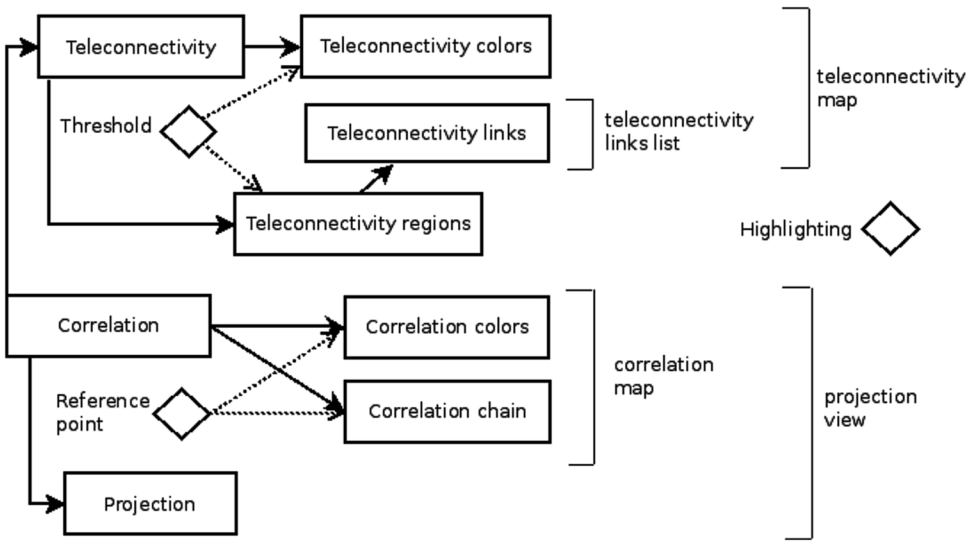

Fig. 1 Data analysis workflow of the tool. Rectangles specify analysis steps and mention the results of these steps. Square brackets show correspondence of analysis results to the visualizations. Rhombi represent user inputs

In the following section, we will provide the details of all mentioned components.

Data analysis components

In this section, we describe how features such as correlations between spatial locations or regions and teleconnections are extracted from the data, how they are visually encoded for the interactive analysis, and which interaction mechanisms support the analytical workflow described above.

Correlation computation

The initial step of the approach is the computation of correlation for all pairs of grid points. For the purposes of this paper, we focus on a single two-dimensional data field. For each pair of grid points p, q, defined by their latitude and longitude, the Pearson's correlation coefficient is applied to the associated time series as

$$r_{pq} = \frac{\sum_t (p_t - \bar{p})(q_t - \bar{q})}{\sqrt{\sum_t (p_t - \bar{p})^2}\sqrt{\sum_t (q_t - \bar{q})^2}},$$

where p_t, q_t are the values of the respective time series at time t, and $\bar{p}$, $\bar{q}$ are the mean values of the time series. Pearson's correlation computes values out of the range $[-1, 1]$, where positive correlation indicates time series with a similar behavior, negative correlation indicates time series with an opposite behavior, and value 0 means no correlation between the time series.

Due to the fact that random time series can exhibit a certain level of correlation, it becomes necessary to estimate statistical significance of computed values. Moreover, a time series of physical simulation output is not purely random and some inertia is observed in the system, i.e., each value influences the subsequent ones. To account for this dependency, autocorrelation for each time series is computed, which is defined as the correlation of a time series with itself shifted by a timestep. The resulting coefficient r_{i0} is used to estimate an effective number of data series items von Storch and Zwiers (2002)

$$n_{ie} = n\frac{1 - r_{i0}}{1 + r_{i0}},$$

where n is the length of the time series. This factor adjusts the length of the time series to its effective length, i.e., the number of entries necessary to describe the time phenomenon. Testing for statistical significance is then performed by executing the Student's t-test using the formula

$$t(r_{ij}) = r_{ij}\sqrt{\frac{n_{ie} - 2}{1 - r_{ij}^2}}$$

where i and j are indices of two points p_i, p_j in the array of all grid points and r_{ij} denotes the Pearson's correlation of the respective time series at the points. If the t-test delivers a value below 0.05, the correlation is said to be statistically significant, below 0.01 even strongly statistically significant.

The correlation matrix with entries r_{ij}, where each grid point is represented by a row and a column, form the data space studied in the visual analysis. By definition, the matrix is symmetric ($r_{ij} = r_{ji}$) and entries on the diagonal are 1 ($r_{ii} = 1$). In the visual encodings, the correlation matrix is exploited as follows:

1. The correlation map displays one selected column at a time, cf. Fig. 2b.
2. The teleconnectivity map displays a column derived from the strongest anti-correlations in each row, cf. Fig. 2a.
3. The projection view shows the selected column overlaid on the inherent structure of the matrix, cf. Fig. 2d.

Correlation chain computation

The teleconnectivity map and links, described below, are indicating the points that expose strong relationships with others, and extrapolate these relationships into regions extracted around local maxima of teleconnectivity. The actual patterns of local and global scale are validated through the correlation maps for these points, corresponding to the respective columns in the correlation matrix.

A correlation chain includes larger groups of points that have strong correlation relationships, both positive and negative, with a chosen reference point. It provides an additional stability assessment for the teleconnections, improving the understanding on whether correlated points cluster around the ends of a link or tend to be located in other regions of the map.

The chain is built iteratively. It starts with the selected reference point. Each subsequent point is the one that has the strongest negative correlation with the current endpoint of the chain and is not in the chain yet. The color alternates between yellow and cyan for even and odd points in the chain, allowing for distinguishing between groups of points that are positively correlated to the reference point and those that are negatively correlated to the reference point, cf. Fig. 2b. For the figures in this paper, the stopping condition is reaching 100 points in the chain.

Correlation map

The correlation map (Fig. 2b) is a geospatial representation depicting correlations with respect to a reference point within physical space. The color represents the correlations of each point with the current reference point. The used

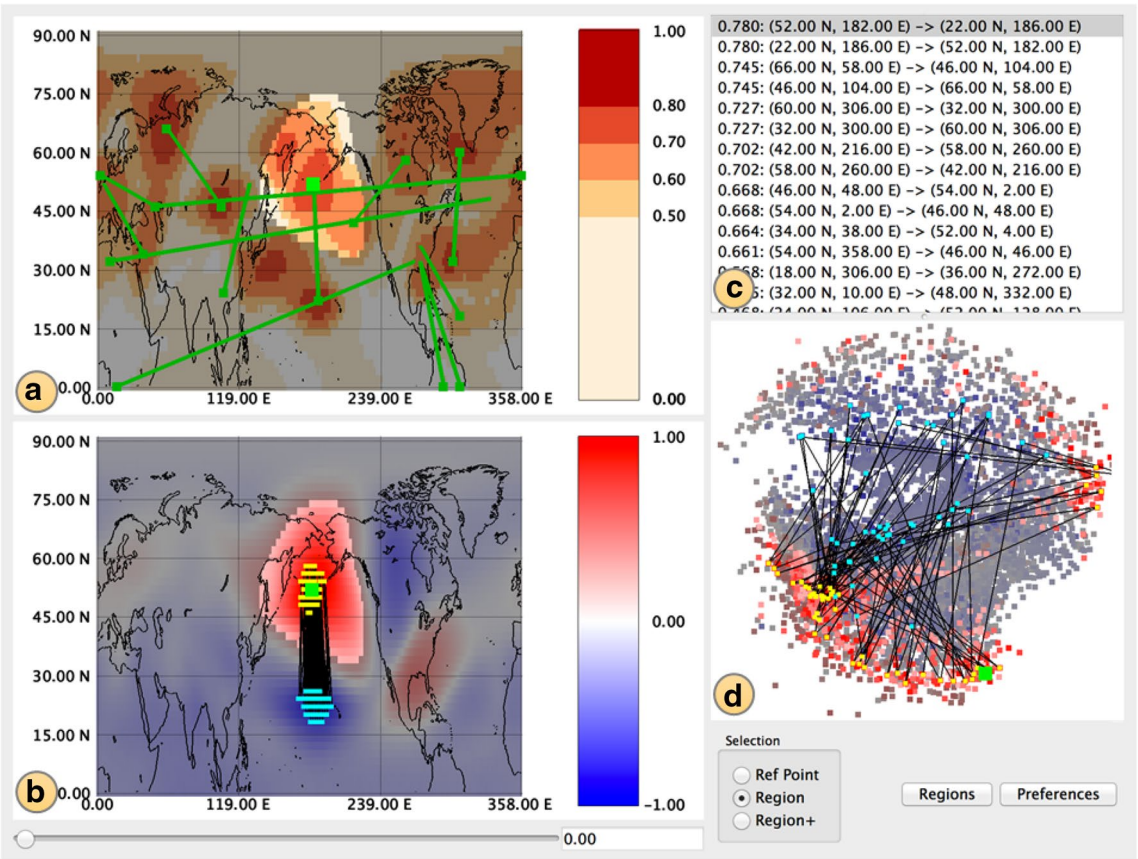

Fig. 2 Interface of the teleconnection analysis tool highlighting the region around the strongest teleconnectivity maximum. Data: NCEP, geopotential height at 500 hPa, northern hemisphere, winter means for 1871–2000. **a** Teleconnectivity map, **b** correlation map, **c** teleconnectivity links list, **d** projection view. The reference point is (52°N, 178°W) with teleconnectivity $T = 0.78$

color scheme is a standard diverging (or bipolar) color map that interpolates the color from dark blue for strong negative correlation via white for non-correlated points to dark red for strong positive correlation, see color legend next to the map in Fig. 2b. The current selection is visualized by keeping the selected points bright and dimming the color of the non-selected points.

The correlation map uses a layout with a rectangular or polar setup, land contours, and the reference grid. It allows for overlaying the color mapping with a geometric representation of the correlation chain. The reference point, which starts the chain, is marked in green, and each consecutive point is marked in cyan for the negatively correlated nodes (an odd number of steps away) and in yellow for the positively correlated ones (an even number of steps away). Nodes are connected with lines.

Correlation projection view

In geoscientific research, it is common to place points on a map where their positions correspond to their latitudes and longitudes in the physical space (as in the correlation map). Here, we suggest projection as an alternative representation of the correlation space of data, where placements of points are defined not by their geographical coordinates, but instead by mutual correlations of the respective time series. The motivation for this step is that we want to see the global structure of correlations, i.e., we try to best capture the correlation between all locations in one visual encoding instead of locally depicting the correlation to a single reference point as in the correlation map. In the projection view, all spatial locations are represented by a point in a 2D visual space. Each point shall be placed close to other points with similar behavior and distant from other points with much different behavior. More precisely, we want to create a view, where the distance between the points matches the dissimilarity of the locations, where similarity is measured using correlation.

To construct this 2D visual space representation, a distance matrix is derived from the correlation matrix, by defining a distance

$$d^*_{ij} = d^*(p_i, p_j) = \frac{1 - r_{p_i p_j}}{2},$$

between two points p_i, p_j, where d^*_{ij} reaches its minimum if the respective time series at the points are perfectly positively correlated, and maximum if they are perfectly negatively correlated.

Then, the task of the particular projection algorithm is to place the points into a Euclidean 2D space to represent the correlation space in terms of the distances between points. For the destination space, the Euclidean distance is applied to the images p'_i and p'_j of the points p_i, p_j:

$$d^{(2)}_{ij} = d^{(2)}(p'_i, p'_j) = \sqrt{(p'_{ix} - p'_{jx})^2 + (p'_{iy} - p'_{jy})^2}.$$

With these definitions, we want to find a 2D embedding, where the 2D Euclidean distances $d^{(2)}_{ij}$ of points in the 2D embedding reflect as much as possible the distances d^*_{ij} derived from the correlation matrix. Thus, we want to compute an embedding that minimizes the differences between $d^{(2)}_{ij}$ and d^*_{ij}. Mathematically, this can be expressed by minimizing the stress or error (Sammon 1969) given by

$$E = \frac{1}{\sum_{i<j}[d^*_{ij}]} \sum_{i<j}^{N} \frac{[d^*_{ij} - d^{(2)}{}_{ij}]^2}{d^*_{ij}}.$$

with d^*_{ij} and $d^{(2)}_{ij}$ denoting the distances between the i-th and the j-th grid point. For this paper, we utilize the technique called Sammon's mapping Sammon (1969) for the projection. It starts with points randomly distributed in the 2D space and then iteratively rearranges them, aimed at minimizing the given error.

The result of this transformation is presented and explored in the projection view. Thus, the projection view (Fig. 2d) is a non-geospatial representation of global correlation relationships within the data. The projection visualization provides a novel perspective on the correlation space and opens possibilities for new findings. The color in this view also represents the correlation of each point with the current reference point. The used color scheme is again a standard diverging (or bipolar) color map that interpolates the color from dark blue for strong negative correlation via white for non-correlated points to dark red for strong positive correlation, see color legend next to the map in Fig. 2b. The current selection is, again, visualized by keeping the selected points bright and dimming the color of the non-selected points. Thus, the projection view also marks the reference point and visualizes the correlation chain in the same way as the correlation map.

The main contribution of the projected view, however, is the global representation of pairwise correlations as distances within a 2D plot. This view spatially restructures the information of the correlation map. Instead of being laid out according to the physical coordinates of points, it expresses the relationships between the grid points in the correlation space. Clusters in the projection view represent points that are strongly (positively) correlated with each other (high similarities of their time series) and much less (or negatively) correlated with points outside the cluster. A dense cluster represents a group of points revealing very high mutual correlation and similar correlations of these points with the rest of the dataset. A slightly less dense group of points stretched along the outside of the projection may indicate that while points exhibit strong positive correlations with each other, their correlations with the rest of the data set are not uniform, or vice versa. Anti-correlated groups of points tend to be placed at the opposite sides of the projection, and this tendency is stronger as the groups get bigger and the strength of their anti-correlation gets higher.

Teleconnectivity computation

Teleconnectivity at a reference point p_i is defined as the absolute value of the strongest anti-correlation observed at p_i with any other point p_j. Given the correlation matrix, we detect the minimum value of row i and take its absolute value:

$$T_i = \left| \min_j(r_{ij}) \right|.$$

In the following, we refer to the triplet (i, j, T_i), for a cell r_{ij} which exposes such a minimum, as a link from the starting point i to the endpoint j with the teleconnectivity value T_i. The function which establishes the correspondence $i \mapsto T_i$ of each point to its teleconnectivity value defines a teleconnectivity map. Representative links for the dataset are selected from the local maxima of the teleconnectivity map, based on the results of the regions extraction algorithm, see below. To test for statistical significance of teleconnectivity values (Section "Correlation computation"), a directional t-test is used, as high values of teleconnectivity are coming from negative values of correlation. Teleconnectivity values and their statistical significance are presented in the teleconnectivity map view, while the representative links are overlaid with the teleconnectivity map and are also enumerated in the teleconnectivity links list (Section "Teleconnectivity map and teleconnectivity links list", cf. Fig. 2c).

Teleconnectivity region extraction

Region extraction is an intermediate step in choosing representative links. The aim is to extract consistent non-overlapping regions and present teleconnectivity relationships between them. A region is defined by the following properties:

1. It has high teleconnectivity values, i.e., the values are statistically significant and above a threshold (in case a threshold is defined and applied).
2. It contains a local maximum of teleconnectivity, i.e., an inner point (or inner subregion in case of a plateau) where teleconnectivity is higher than all teleconnectivities within its neighborhood.
3. It covers a spatially continuous area.
4. It is consistent with respect to correlations between its points, i.e., all points have a statistically significant positive correlation with the region's local maximum.

Our iterative region extraction algorithm starts with excluding the points with teleconnectivity values that are not statistically significant and/or are below the selected teleconnectivity threshold (Condition 1). This initial selection is shown in Fig. 3) as Step 0.

Then, the algorithm iteratively searches for the largest local maximum of teleconnectivity not yet assigned to any region. This maximum becomes the seed for a new region (Condition 2). In Fig. 3), this is shown by the links that are drawn in the form of green lines.

A region-growing approach around the seed point is employed, including neighboring points which have teleconnectivity values above the threshold and which are positively and statistically significantly correlated with the seed point. Continuity of the data in the longitude and latitude dimensions is taken into account (Condition 3). Region growing stops when it reaches the region's limits, i.e., points where the conditions are not met any longer (Conditions 1, 3, 4): At the limits of the region growing all points surrounding the region have teleconnectivity values below the threshold, or are negatively or not statistically significantly correlated with the seed point. The colored areas in Fig. 3) show the resulting regions. The process then enters the next iteration step.

When no unmarked local maxima are left, extraction of the region stops (Condition 2). The final result in Fig. 3) shows, which regions were extracted. The colors here are chosen such that they are well distinguishable and do not exhibit any obvious order, i.e., we chose a categorical color map as indicated by the color legend to the right of Fig. 3). The numbers next to the colors indicate the iteration step. Iteration step 0 shown in grey represents the initial selection after applying thresholding (regions below the threshold), while the number – 1 shown in white shows all regions that have not been selected during the iterative process.

The actual links to be shown in the teleconnectivity views are selected afterwards. From the set of local maxima of extracted regions, the strongest links connecting pairs of maxima between these regions are chosen in each direction.

Teleconnectivity map and teleconnectivity links list

The teleconnectivity map (Fig. 2a) is a geospatial map-based visualization of teleconnectivity and representative links. The map shows a reference grid and land contours in the geographical area of the dataset, and can present the data in a rectangular or polar setup. Hence, the set-up is identical to the correlation map, which eases detecting correspondences.

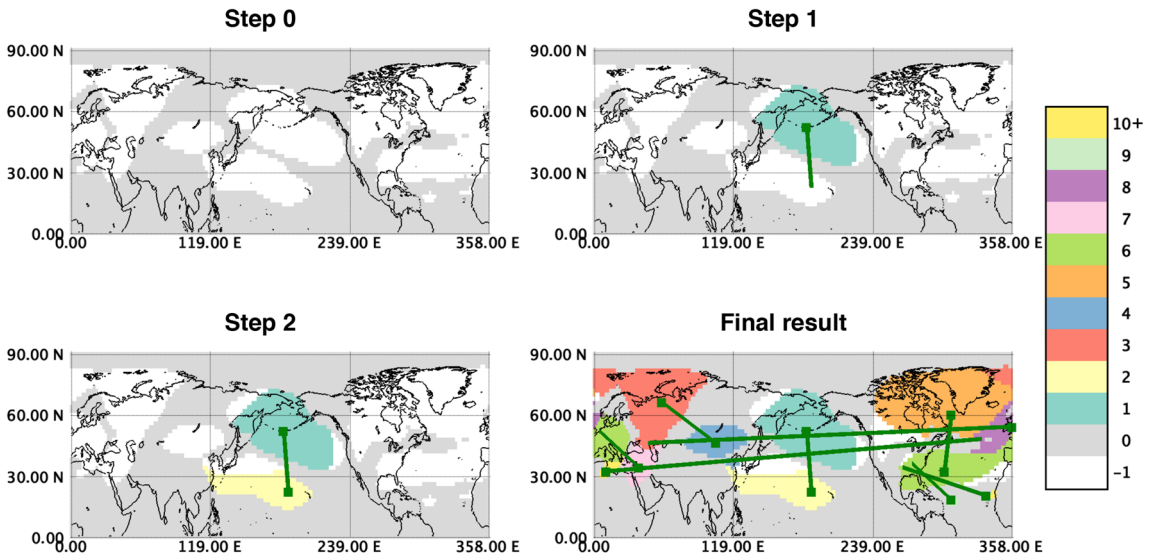

Fig. 3 Region extraction for NCEP dataset with teleconnectivity threshold 0.55. At Step 0, the teleconnectivity threshold is applied, and the excluded area is marked with gray, while regions with values above the threshold stay white. Steps 1–2 show consecutive seed points with respective links and colored regions extracted at first and second iteration. Final result shows the colored extracted regions (color correspond to iteration step) at the state when no local maxima are left in the unmarked areas

Teleconnectivity is visually encoded using color mapping. A user-defined threshold value is used to generate the regions, as described above, as well as for filtering the values in the map. We employ a banded color scheme with decreasing luminance and increasing saturation from the Color Brewer project (Brewer 2003), with higher values receiving darker and more saturated color shades, see color legend next to the map in Fig. 2a. Two extra colors are used for conveying additional information: gray color is used for the points which are not statistically significant, and the values below the user-defined threshold are colored in white. Additionally, if some group of points is selected for highlighting, then these points are shown in full color, while other points are dimmed down in brightness.

Teleconnectivity links are unilateral relations and visualized with a dot at their starting point and a line connecting the starting point to the endpoint. The starting point is the grid point where the representative teleconnectivity value is observed, and the endpoint is the grid point with which strong anti-correlation is observed. The endpoint is not marked if its teleconnectivity value stems from a teleconnectivity with another point, i.e., a point different from the starting point, or if it is in a neighborhood of a stronger maximum. A dot of a slightly bigger size and a brighter color marks the position of the reference point for the correlation views.

The teleconnectivity links list (Fig. 2c) enumerates the links shown on the teleconnectivity map in order of descending teleconnectivity value. For each link, the textual representation of its value, and coordinates of starting and endpoints are shown.

Interaction mechanisms and interactive workflow

Exploration of the results of the analysis is performed in an interactive visual setting with four coordinated views: the teleconnectivity map, the teleconnectivity links list, the correlation map, and the projection view. Two of the views present the teleconnectivity analysis results (Fig. 2a, c), while the other two views present the results of the correlation analysis (Fig. 2b, d).

An important difference of our interactive approach from a traditional script-based approach is that the user is able to interact with the software to examine all aspects of the data at different levels and reconfigure views for multiple analysis tasks and steps after the main resource-consuming operations are completed. Low response times of the system during the interactive analysis enable the user to intuitively explore the data and quickly build knowledge and understanding of the data at multiple scales. In addition, the described approach leverages the concept of coordinated multiple views. This means that several visualizations are shown at the same time, and each action of the user receives fast visual feedback in all views simultaneously, giving a consistent view on multiple aspects of the data and providing assistance and support in making decisions for further actions.

The most basic interaction method is hovering the mouse over points of interest in the views. It triggers small in-place pop-up windows with on-the-spot information, including physical coordinates (latitude and longitude) of the point under the mouse pointer, and its respective value in the view (teleconnectivity for the teleconnectivity map, correlation with the reference point for the correlation map and projection view). Beyond that, the tool accepts three main means of user input: selection of a new reference point, selection of a group of points for highlighting (brushing), and selection of a new teleconnectivity threshold. At the launch of the tool, the default reference point is the top link in the links list (the link with the strongest teleconnectivity), no points are highlighted (no points are dimmed in the views and all points are visualized in full brightness), and the teleconnectivity threshold is zero.

The reference point is used for the color mapping and generating the correlation chain in the correlation map and in the projection view. By selecting a new reference point, the user can validate stability of the teleconnection for the points surrounding the shown link, or simply check the correlation relationships of a point with no link shown in the teleconnectivity map. The reference point is selected by clicking on a point in the point-based views or on a link in the list. The current reference point is reflected in the point-based views as a point of a different color. If it is the starting point of one of the representative links, the respective row is highlighted in the list.

The views also support highlighting, which helps to focus on the most relevant aspects of the data. Depending on the selection mode, a click in the views does not only change the reference point, but also selects its region or connected component in the graph of regions for highlighting. The user can also make a free-form selection in the projection view. The points inside the selection or belonging to the selected regions keep their bright colors while all other points have dimmed colors.

Change of the teleconnectivity threshold modifies the teleconnectivity map, including points above the threshold and excluding points below it. A higher value of the threshold may cause a region to disappear or split into several subregions according to the strengths (i.e., magnitudes) and the number of the local maxima. A lower threshold may cause a new region to appear or two regions to merge. In case of such changes, different teleconnectivity links will be exposed. By interactively modifying the threshold and noticing the changes in the colored area and links, the user can intuitively compare strengths of

the patterns in the dataset, and filter out weaker and less interesting ones.

Implementation and performance

The tool is implemented in C++, using Qt for the user interface, OpenGL for the visualizations, the GNU Scientific Library (GSL) for assessment of statistical significance, and the NetCDF library for reading data provided in NetCDF format. In the current implementation, the tool runs in main memory with longer-term storage of correlation data and projections. Currently, it uses in-core data access, limiting the size of data that can be processed to the size of main memory available to the process.

Computation times for a non-optimized, single-threaded implementation of the feature extractions running on a MacBook Pro with a 2.6 GHz Intel Core i5 processor are shown in Table 1. One can notice that the data loading itself takes seconds to minutes. Other listed steps are the computation of the correlation matrix and the projection. These are the main bottlenecks in the computation and with the current implementation take minutes to hours. However, all these steps are pre-computations. Hence, they only need to be executed once. The pre-computation results can be stored for multiple analysis sessions.

The actual analysis is performed at interactive rates. Thus, all computation steps engaging the user during the visual analysis allow for immediate response. Hence, the users have a smooth user experience during the interactive analysis session.

Application scenarios

In this section, we present an application scenario for our tool. We start by introducing the teleconnection patterns that are known and expected to be observed in the data. The data contain a simulation ensemble including a control simulation and a reanalysis data set, as detailed below. The tasks are to investigate whether the teleconnections in the reanalysis data also occur in the simulated data and how strong the teleconnections are. We present the application of our tool and show how it can help to solve these tasks using the analytical workflow of our interactive visual analysis tool.

Teleconnection patterns

We provide two examples of teleconnection patterns as a reference to the subject of our studies. Usually, a certain name describes a family of patterns with similar configurations, i.e., comparable positions and relative strengths of the centers, while actual configurations may differ due to the methods and data used.

The first pattern we want to introduce is referred to as the North Atlantic Oscillation (NAO) pattern. It dominates the lower atmosphere variability in the North Atlantic sector. It refers to a north–south oscillation in atmospheric mass between the Icelandic low- and the Azores high-pressure centers (Walker and Bliss 1932). The NAO is the dominant pattern of near-surface atmospheric variability over the North Atlantic, accounting for one third of the total variance in monthly sea-level pressure in winter. A standard visualization in the form of a schematic representation of the spatial signature and climate impacts of NAO is given in Fig. 4. The visual representation uses polar coordinates for a map-based representation of the hemispheres.

The positive phase is shown in Fig. 4 to the left. It is associated with higher than normal surface pressure south of 55 °N combined with a broad region with anomalously low pressure throughout the Arctic and subarctic. Consequently, this phase is associated with stronger-than-average winds across the mid-latitudes of the Atlantic onto Europe, with anomalously southerly flow over the eastern United States and anomalously northerly flows across Greenland, the northern part of Canada and the Mediterranean region and enhanced easterly trade winds over the sub-tropical North Atlantic.

During the negative phase, both the Icelandic low and Azores high-pressure systems are weaker than normal, so both middle latitude westerlies and the sub-tropical trade winds are also weak as shown in Fig. 4 to the right. The negative phase brings higher-than-normal pressure over the polar region and lower-than-normal pressure at about 45 °N. The negative phase allows cold air to plunge into the

Table 1 Computation times for pre-computed feature extraction, where loading includes generation of teleconnectivity map, region search, finding teleconnectivity links, and building correlation chain for default reference point

Dataset	Data size	Correlation	Projection	Loading	Total Time
NCEP	91 × 180 × 129	6 m 33 s	4 h 58 m 11 s	2 m 19 s	5 h 7 m 3 s
CTRL	48 × 96 × 129	32 s	21 m 4 s	11 s	21 m 47 s
ENS1	48 × 96 × 129	32 s	20 m 28 s	11 s	21 m 11 s

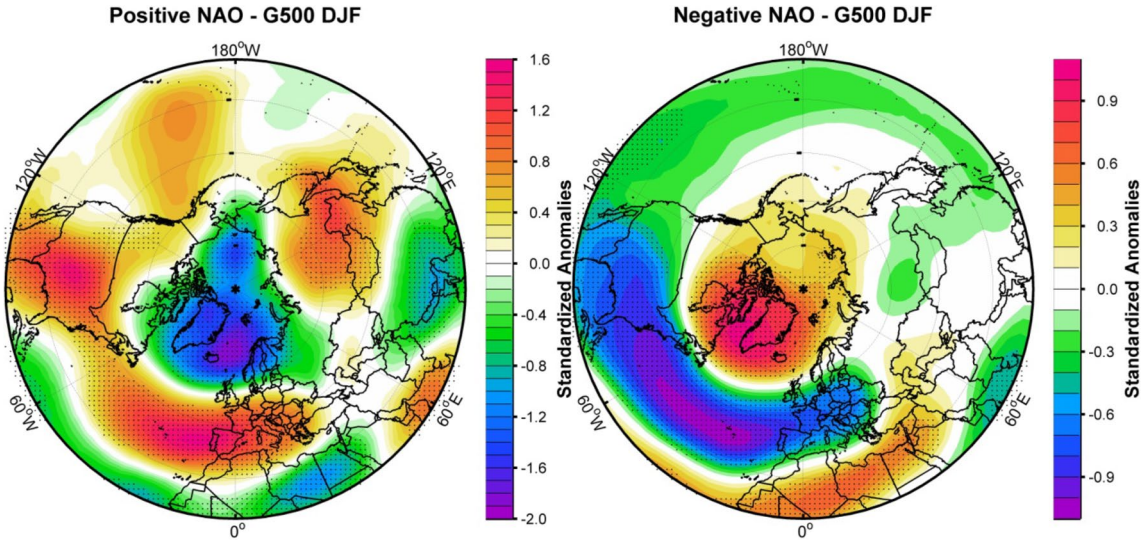

Fig. 4 Composite map between normalized NAO index and normalized geopotential height at 500 mb for positive phase (left) and negative phase (right)

midwestern United States and Western Europe, and storms bring rain to the Mediterranean.

The second pattern that we want to introduce is the Pacific North American Oscillation (PNA). It refers to an alternating pattern between pressures in the central Pacific Ocean and centers of action over western Canada and the southeastern US. It is most pronounced in winter. The PNA is associated with a Rossby wave pattern and refers to the relative amplitudes of the ridge over western North America and the troughs over the central North Pacific and southeastern United States.

The positive phase is shown in Fig. 5 to the left. Teleconnection occurs when deeper than normal troughs occur over the eastern United States and the region of the Aleutians. The positive phase is associated with North–South upper air flow, above-average temperatures over western Canada and the extreme western US, and below-average temperatures across the south-central and southeastern US The associated precipitation anomalies include above-average totals in the Gulf of Alaska extending into the Pacific Northwest of the United States, and below-average totals over the upper midwestern US.

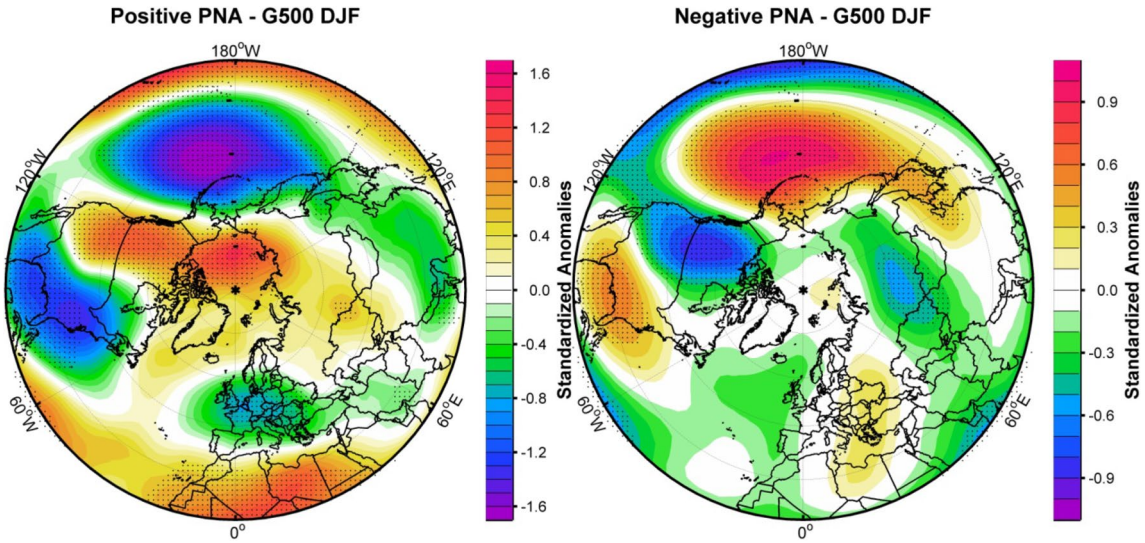

Fig. 5 Composite map between normalized PNA index and normalized geopotential height at 500 mb for positive phase (left) and negative phase (right)

The negative phase is shown in Fig. 5 to the right. Here, the troughs are filled and a ridge over the Rockies is lowered. The negative phase is associated with East–West upper air flow.

Data

In the following, we describe the data that we analyze for teleconnections. The data sets contain northern hemisphere winter means (DJF) of geopotential height at 500 mb level for a period of 130 years (1871–2000).

First, we want to describe the reanalysis data set. It is referred to as "NCEP". It is extracted from the twentieth century Reanalysis Project database, version 2 (NCEP 2011), for the period 1871–2010. This reanalysis data set is derived through a consistent assimilation and modeling procedure that incorporated most available weather and satellite information (Whitaker et al. 2004; Compo et al. 2006, 2011). The spatial resolution of the grid is 2° latitude × 2° longitude.

Second, we want to describe the simulation ensemble that we want to compare to the reanalysis data. Simulations of the last millennium have been conducted using the COSMOS approach (Jungclaus et al. 2010). Here, the Earth system model consists of a general circulation model for the atmosphere ECHAM5 (Roeckner et al. 2003) coupled with the general ocean circulation model MPIOM (Marsland et al. 2003) with a full carbon cycle implementation. The carbon cycle model comprises the ocean biogeochemistry module HAMOCC5 (Wetzel et al. 2006) and the land surface scheme JSBACH (Raddatz et al. 2007). ECHAM5 is run at T31 resolution (3.75°) with 19 vertical levels, resolving the atmosphere up to 10 hPa and MPIOM applies a conformal mapping grid with a horizontal resolution ranging from 22 to 350 km. The model is forced by reconstructions of

1. total solar irradiance,
2. volcanic forcing considering aerosol optical depth at 0.55 μm and effective radius distribution for 10-day time steps and four equi-areal segments,
3. land use change, and
4. anthropogenic greenhouse gases and aerosols.

We use a full-forcing ensemble mean with five members. We refer to it as "ENS1". Moreover, we use a control run, which we refer to as "CTRL". The whole experiments cover the period 800–2005 AD and the different initial conditions for the ensemble members are derived from a 3000 years control integration forced by constant conditions for 1860.

For the presented analysis we use the period 1871–2000, such that we can compare well the reanalysis data to the simulated data well.

Global comparative teleconnectivity analysis

Now, we want to present how our visual analysis workflow is applied to the three data sets NCEP, ENS1, and CTRL described above. We computed correlations and subsequently teleconnection and teleconnection regions. Then, we visually analyze the main teleconnectivity patterns in the three data sets and compare them to each other to observe similarities and differences. During the interactive analysis of teleconnectivity patterns, we interactively adjust the threshold for region extraction.

Figures 6 and 7 provide a comparison of automatically extracted regions and representative links for the three studied data sets showing the teleconnectivity map at two different threshold levels. Figure 6 demonstrates the influence of the threshold on the region extraction. Without thresholding, region growth is only limited by positive correlation with the region seed. This typically leads to the Arctic, the equatorial belt, the Southern Ocean, and the Antarctic being extracted as single regions, possibly under-representing teleconnectivity links between them and other regions.

Figure 7 focuses on the visual analysis of the teleconnectivity. We observe that most of the shown links have points at both ends, i.e., they represent classic teleconnection dipoles, where two points have the most negative correlation on the map with each other. In general, the simulated datasets expose a higher amount of teleconnectivity, particularly in the southern hemisphere between Antarctic and the Southern Ocean. In all cases the link between New Zealand and the Ross Sea is shown, which appears as the globally strongest link for the ENS1 dataset with $T = 0.83$. Another link connects south/southeast of the Indian Ocean with the Antarctic continental border around 120°E.

In comparison to the CTRL simulation, ENS1 exposes lower amount of teleconnectivity over equatorial and tropical regions. In these areas, teleconnectivity above 0.6 can only be seen in the central Pacific and the southeastern United States. There is no clearly visible NAO-like teleconnection observed in the ensemble data. The typical southern counterpart of the teleconnectivity center over Greenland is shifted to the east and is located over Western Europe.

There is a noticeable difference in statistical significance between the datasets (the gray areas in Fig. 7). At the 0.99 level, all grid points of CTRL exposes statistically significant teleconnectivity, ENS1 has a few not statistically significant points in the tropics, while NCEP shows large areas of not statistically significant points at the equator and in the tropics.

Detailed pattern analysis

In a second step, we want to drill down on teleconnectivity patterns that have been extracted during the global analysis

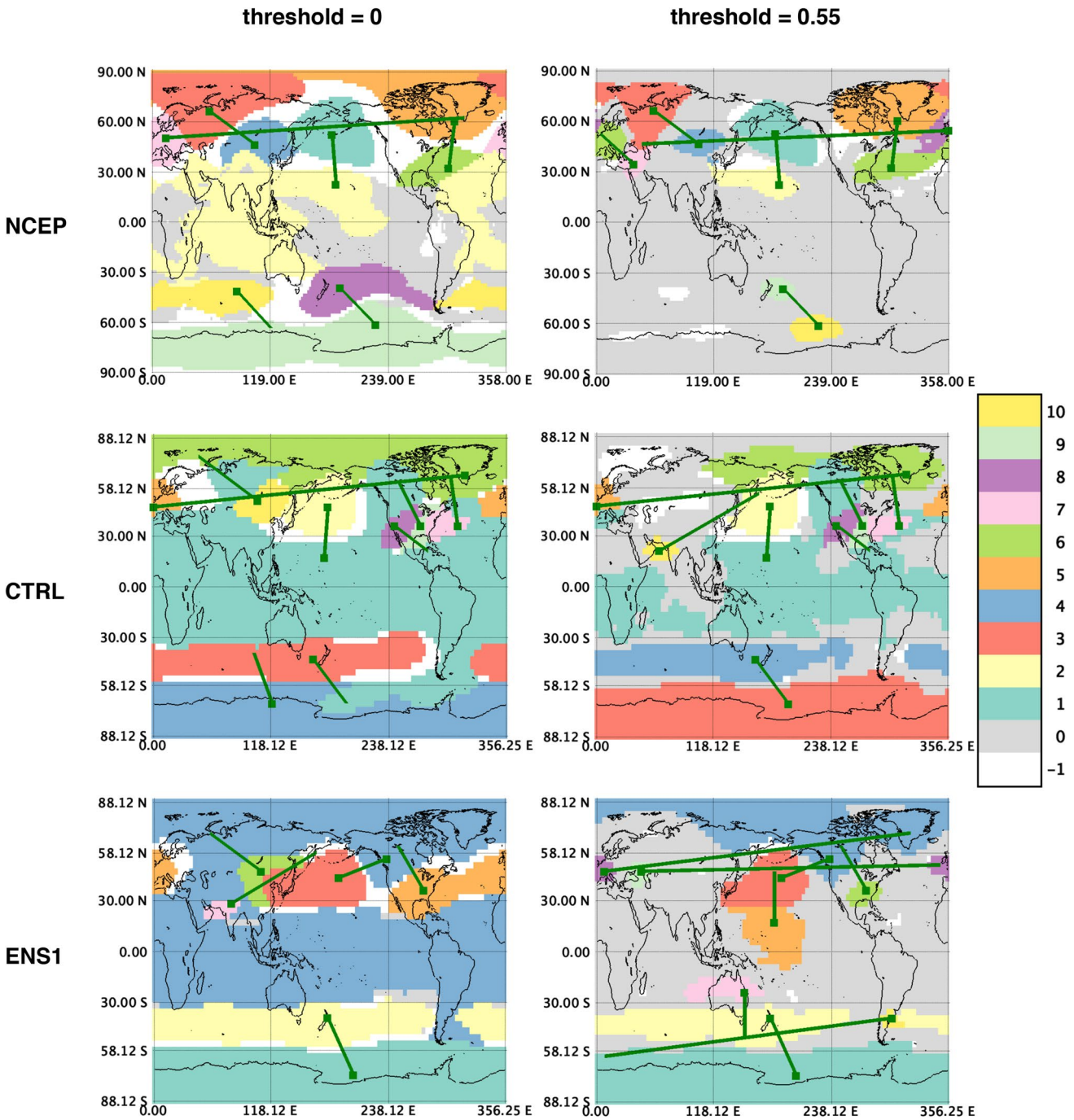

Fig. 6 Comparison of the region extraction results up to the tenth region for different datasets at two threshold levels (NCEP: the reanalysis data; CTRL: the control run data; ENS1: the full-forcing ensemble mean data). Colors correspond to number of iterations executed when the grid point is assigned a value, where zero means no statistically significant teleconnectivity (at 0.99 level) or below the user-defined threshold

described above to investigate individual patterns that attract our attention. The teleconnectivity links list provides a systematic way to look into the individual patterns starting with the most prominent ones.

For a detailed look at the patterns, we restrict the analysis to the northern hemisphere. We start by analyzing the reanalysis data set, which sets our expectations about what to observe in the simulated data sets. Iterating through the list of representative links for the NCEP dataset, we can see that the correlation chain forms a few strong dipoles. Such dipoles occur between the northern and the central Pacific, between Scandinavia and Asia, and between the

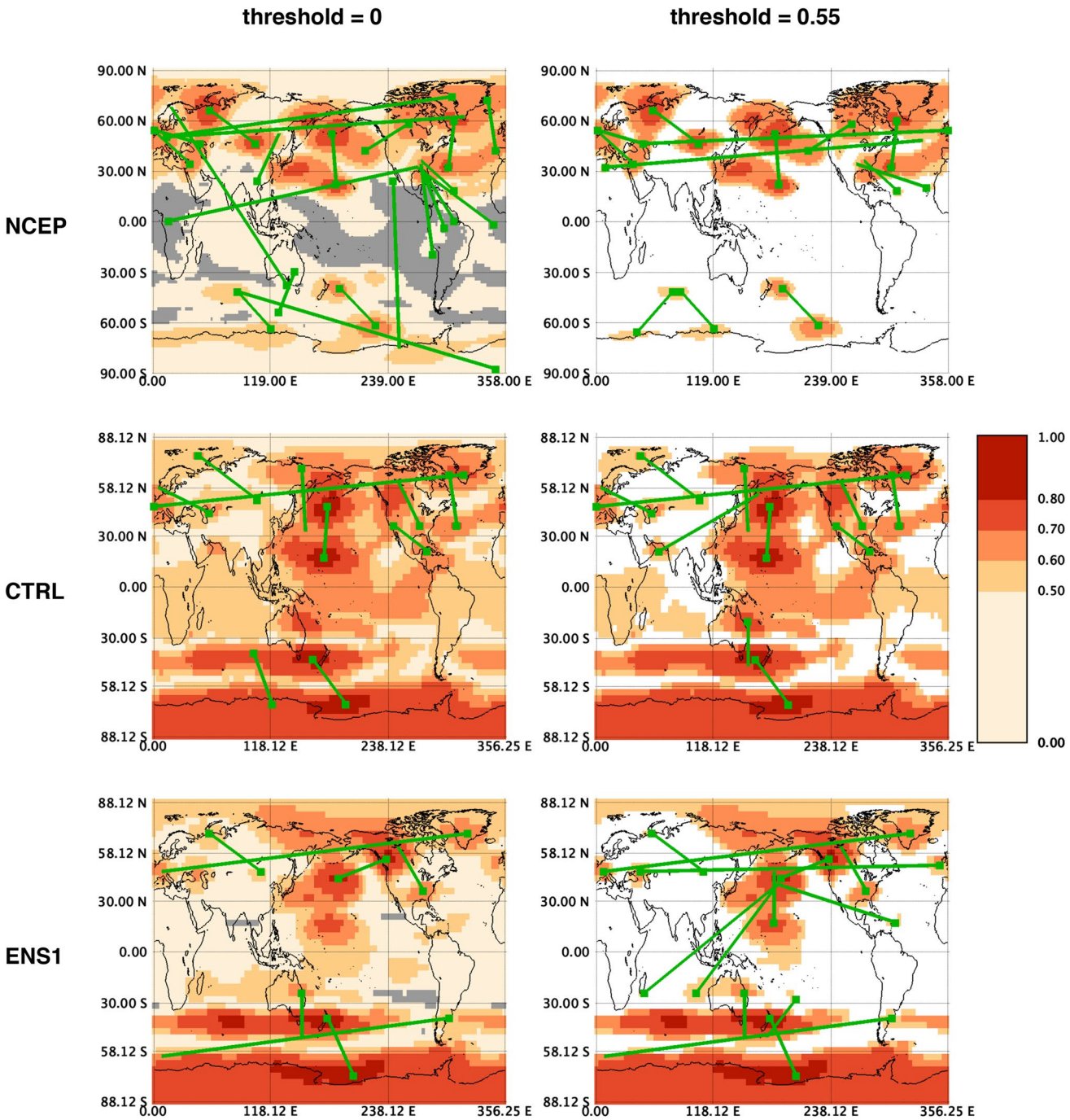

Fig. 7 Comparison of the teleconnectivity map views after completing the region extraction of Fig. 6. Gray color indicates no statistically significant teleconnectivity (at 0.99 level). NCEP: the reanalysis data; CTRL: the control run data; ENS1: the full-forcing ensemble mean data

northern and the western Pacific. We also observe a few patterns with three centers such as Greenland, the northern Atlantic, and Europe as well as the eastern Pacific, Canada, and the Gulf of Mexico. There are also mixtures of these patterns, or less defined patterns with multiple weaker centers.

The strongest teleconnections (Fig. 8) are the dipole in the northern and the central Pacific ($T = 0.78$), the dipole between Scandinavia and Asia ($T = 0.745$), the three-center pattern with Greenland, northern Atlantic, and Europe (the strongest link with $T = 0.727$), and the three-center pattern with the eastern Pacific, Canada, and the Gulf of Mexico (the

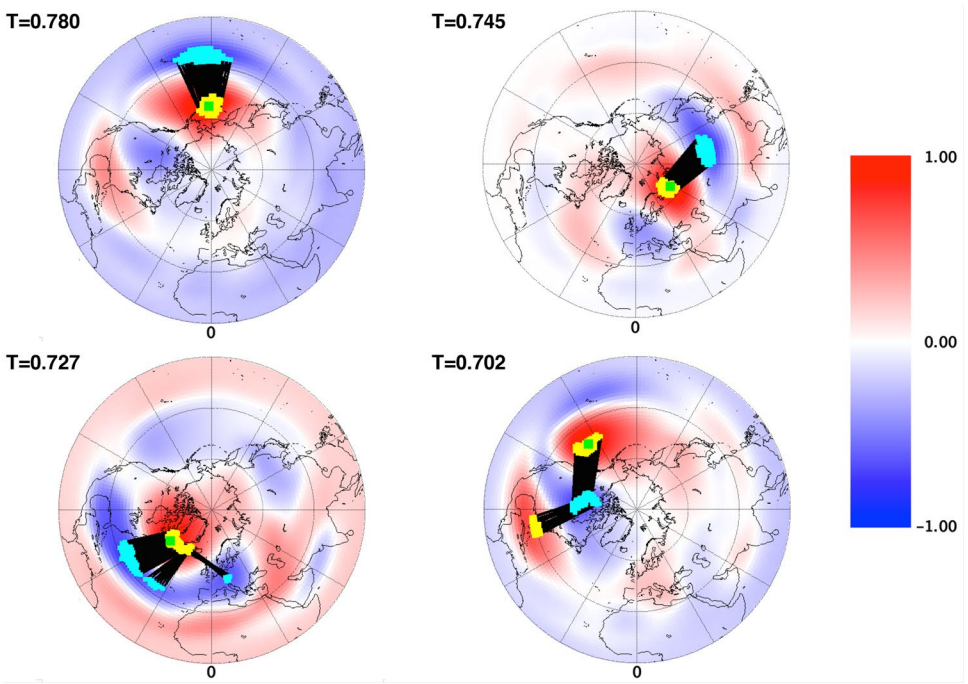

Fig. 8 Strongest teleconnections for the NCEP dataset. (Top-left) $T = 0.78$ at (52°N, 178°W), (top-right) $T = 0.745$ at (66°N, 58°E), (bottom-left) $T = 0.727$ at (60°N, 54°W), (bottom-right) $T = 0.702$ at (42°N, 144°W)

strongest link with $T = 0.702$). Three-center patterns match very well the NAO and PNA patterns described above.

Comparing the correlation maps for $T = 0.78$ and $T = 0.702$ we can see that they represent the same pattern. However, a shift from the center at (52°N, 182°E) to the one at (42°N, 216°E) results in a completely different correlation chain, even if both centers belong to the same extracted region.

In the CTRL simulation data (Fig. 9, top row), we observe the PNA pattern with a center in the northern Pacific and two centers of the opposite sign in the mid-western Pacific and over Canada (the strongest link with

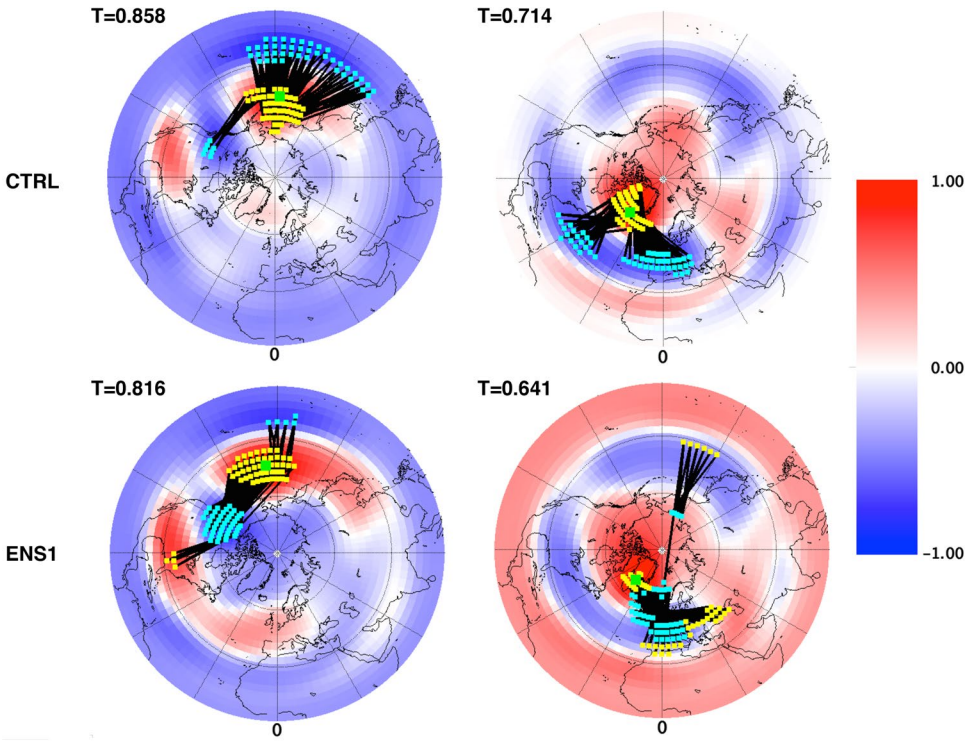

Fig. 9 Teleconnections patterns in the model data. (Left) closest to PNA, (right) closest to NAO. $T = 0.858$ at (46.88°N, 176.25°E), $T = 0.714$ at (65.62°N, 45°W), $T = 0.816$ at (43.12°N, 172.5°W), $T = 0.641$ at (69.38°N, 41.25°W)

$T = 0.858$). Additionally, we see the NAO pattern with a center in the Davis Strait near Greenland and two centers of the opposite sign in the western Atlantic and over Western Europe (the strongest link has $T = 0.714$). A stable dipole exists between the center of the PNA in the mid-western Pacific and an area over the northeastern border of Eurasia, which serves as an attractor for all shown links with $T < 0.7$.

In the ENS1 ensemble data (Fig. 9, bottom row), for the PNA analysis we observe that the strongest link has $T = 0.816$, but the pattern includes four regions and is not as precise as in the reanalysis data. Moreover, the teleconnectivity center suspected for NAO does not produce a clearly separated pattern in the correlation chain.

As the PNA pattern in the ENS1 ensemble data was not clearly pronounced, we would like to analyze it in more detail by investigating the correlations. With the help of the correlation projection view, we can explore the patterns in more depth using a global correlation plot, while the teleconnectivity map shows the respective spatial context in a coordinated view. The top row in Fig. 10 shows highlighted the automatically extracted region, corresponding to the northern and the northwestern Pacific, in the correlation projection view and in the teleconnectivity map. Applying manual selection around the positive end of the correlation chain in the projected view (the middle row in Fig. 10), we can see that these points correspond to the northern Pacific, northeastern United States, and northwestern Europe. Repeating the same interaction for the negative end of the correlation chain (the bottom row in Fig. 10), the central Pacific, Canada, and an area in the central Atlantic receive the highlighting. The areas that are highlighted in the map with these selections are considered to have a similar temporal behavior to the respective parts

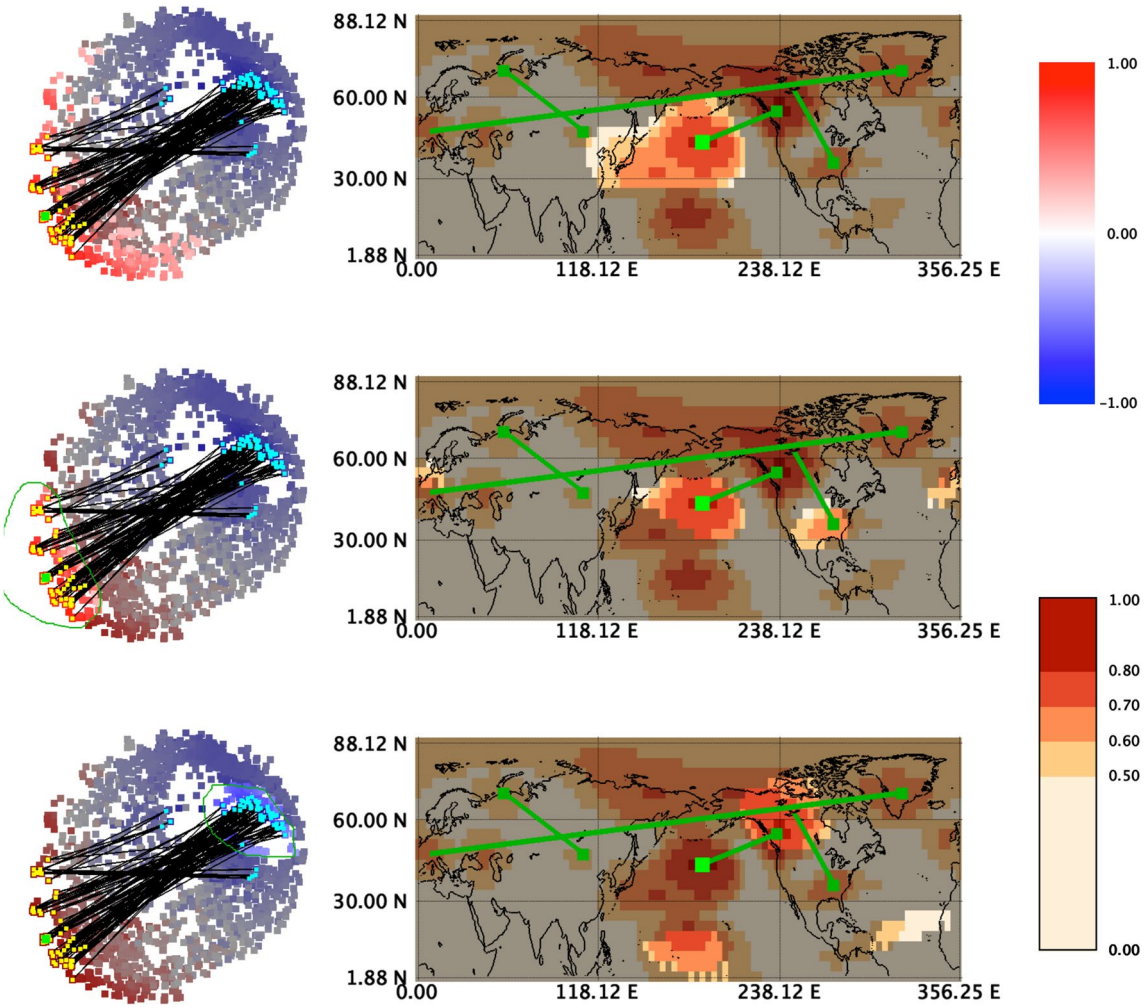

Fig. 10 Exploring the PNA pattern components in the ENS1 data with the help of highlighting in the projection view and the teleconnectivity map. (Top row) region around the start point of the strongest link is highlighted in the views, (middle row) a free-form selection around the positive nodes of the correlation chain in the projection view, (bottom row) a free-form selection around the negative nodes of the correlation chain in the projection view. The reference point is (43°N, 172.5°W), with teleconnectivity $T = 0.816$

of the correlation chain (the ones positively correlated with the reference point, and the ones negatively correlated with the reference point). Consequently, the areas of northwestern Europe and the central Atlantic can be experiencing influence of the PNA pattern.

Discussion and future work

The main features of the software can be summarized as follows:

1. It has a simple graphical interface that allows for visual investigation and interpretation of teleconnections.
2. It provides an additional model validation method based on extracted structures.
3. It allows for easy-to-use comparison of different datasets.
4. With the separation into a pre-computation step and an interactive analysis step, the analysis data can be stored and reused for multiple analysis iterations.

Moreover, this kind of tool development approach emphasizes the importance of bridging the visualization community together with the geoscientific community. This collaboration enables production of tools which are interactive, have a smaller overhead in resource usage, and cover certain workflows without requiring scripting a complex setup of multiple tools.

This project uses temporal correlations and the workflow of Wallace and Gutzler (1981) as the statistical foundation for the tool. Similar tools can be developed using alternative approaches to pattern extraction and analysis, e.g., based on the eigenvector analysis (Barnston and Livezey 1987) or multiple linear regression (van den Dool et al. 2000).

In its current form, the approach focuses on relationships within a single variable at a single vertical level, with time series correlated synchronously. Possibilities for future work include extending the approach to heterogeneous set-ups, e.g., the analysis of correlations between different variables. A promising direction can be seen in development of a similar approach for discovery and exploration of lagged correlations, which can be used for tracing of propagations of natural phenomena in the climate system.

One shall note that teleconnections are also relevant for all types of time scales. The patterns of past climate variability are often related to changes in major phenomena that dominated the climate variability during the last century, like the El Niño Southern Oscillation (ENSO) (Clement et al. 1999; Tudhope et al. 2001; Kitoh and Murakami 2002), the Arctic Oscillation/North Atlantic Oscillation (AO/NAO) (Hurrell 1995; Thompson and Wallace 1998), and the Pacific/North American (PNA) pattern (Wallace and Gutzler 1981). The new technique presented here could be used to explore such teleconnections not only at interannual to decadal time scales, but also to multi-millenial time scales. The latter seems to be relevant for remote effects of glaciation and deglaciation during the last glacial–interglacial cycle (Lohmann 2017).

In addition, we provide an assessment for the quality of data. Typically, the mean of a climate model output is compared to observations. Teleconnectivity and relevant structures can serve as another, more in-depth validation method. From this point of view, the work demonstrates the bridging of the gap between disciplines, namely, how modern visualization methods can be used to respond to a specific geoscientific question and create a more convenient, powerful, and easy-to-use solution. In terms of perspective, it is similar to the EDEN system (Steed et al. 2013). However, the EDEN system focuses on the analysis of multi-variate and multi-field data, while our system has a particular focus on relationships between time series data. This paper presents the initial effort in this direction by creating a tool to explore teleconnections of a single variable in atmospheric simulations.

Conclusion

We presented a tool for identification and analysis of teleconnections. The strongest negative temporal correlations, taken at each point, define a teleconnectivity map that suggests potential pattern centers. Analysis of neighboring points allows for the extraction of regions, relationships between which are taken as representative of the data structure. A correlation map for a pattern center shows degrees of its influence, and the sequence of strongly negatively correlated points starting at the center serves as a stability assessment for the pattern. An additional view provides insights into global correlation relationships within the dataset.

The analysis process is divided into two stages. The first one contains automated computations, and the second one is devoted to interactive exploration. The results of resource-consuming computations are stored on the disk and reused in subsequent sessions with the tool. An analytical workflow is provided.

By comparing three sample datasets with the help of our tool, it is shown that the simulated data expose a higher amount of teleconnectivity when compared with reanalysis data (NCEP). It is also apparent that large variability around the equator in the control run is not present in the ensemble mean data, which may be an artifact of the model simulations or due to the fact that we use an ensemble mean over five members. The patterns extracted from the observation-based NCEP dataset are much better aligned with

well-known climate modes of variability (e.g., NAO and PNA) than for the simulated data.

Acknowledgements This work was funded by Helmholtz Association as a part of Earth System Science Research School, as well as the REKLIM and PACES programmes.

References

Barnston AG, Livezey RE (1987) Classification, seasonality and persistence of low-frequency atmospheric circulation patterns. Mon Weather Rev 115(6):1083–1126

Brewer CA (2003) Colorbrewer. http://colorbrewer2.org. Accessed 2016

Clement AC, Seager R, Cane MA (1999) Orbital controls on the el niño/southern oscillation and the tropical climate. Paleoceanography 14(4):441–456. https://doi.org/10.1029/1999PA900013

Climate Reanalyzer (2011) Climate Change Institute, University of Maine, USA. http://cci-reanalyzer.org/. Accessed 2016

Compo GP, Whitaker JS, Sardeshmukh PD (2006) Feasibility of a 100-year reanalysis using only surface pressure data. Bull Am Meteorol Soc 87(2):175–190

Compo GP, Whitaker JS, Sardeshmukh PD, Matsui N, Allan RJ, Yin X, Gleason BE, Vose RS, Rutledge G, Bessemoulin P, Brönnimann S, Brunet M, Crouthamel RI, Grant AN, Groisman PY, Jones PD, Kruk MC, Kruger AC, Marshall GJ, Maugeri M, Mok HY, Nordli Ø, Ross TF, Trigo RM, Wang XL, Woodruff SD, Worley SJ (2011) The twentieth century reanalysis project. Q J R Meteorol Soc 137(654):1–28

Hurrell JW (1995) Decadal trends in the north atlantic oscillation: regional temperatures and precipitation. Science 269(5224):676–679. https://doi.org/10.1126/science.269.5224.676

Jungclaus JH, Lorenz SJ, Timmreck C, Reick CH, Brovkin V, Six K, Segschneider J, Giorgetta MA, Crowley TJ, Pongratz J, Krivova NA, Vieira LE, Solanki SK, Klocke D, Botzet M, Esch M, Gayler V, Haak H, Raddatz TJ, Roeckner E, Schnur R, Widmann H, Claussen M, Stevens B, Marotzke J (2010) Climate and carbon-cycle variability over the last millennium. Clim Past 6(5):723–737

Kitoh A, Murakami S (2002) Tropical pacific climate at the mid-holocene and the last glacial maximum simulated by a coupled ocean-atmosphere general circulation model. Paleoceanography 17(3):19-1–19-13. https://doi.org/10.1029/2001PA000724 (1047)

KNMI Climate Explorer (2013) Royal Netherlands Meteorological Institute (KNMI). https://climexp.knmi.nl/. Accessed 2016

Lohmann G (2017) Atmospheric bridge on orbital time scales. Theor Appl Climatol 128(3):709–718. https://doi.org/10.1007/s00704-015-1725-2

Marsland SJ, Haak H, Jungclaus JH, Latif M, Röske F (2003) The Max-Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates. Ocean Model 5(2):91–127

NCEP (2011) The 20th century reanalysis project database, version 2, ESRL NOAA. http://www.esrl.noaa.gov/psd/data/20thC_Rean. Accessed 2016

Nocke T, Buschmann S, Donges JF, Marwan N, Schulz HJ, Tominski C (2015) Review: visual analytics of climate networks. Nonlinear Process Geophys 22(5):545–570. https://doi.org/10.5194/npg-22-545-2015. https://www.nonlin-processes-geophys.net/22/545/2015/. Accessed 2016

PSD Web Products and Tools (2019) ESRL NOAA. http://www.esrl.noaa.gov/psd/products. Accessed 2016

Raddatz TJ, Reick CH, Knorr W, Kattge J, Roeckner E, Schnur R, Schnitzler KG, Wetzel P, Jungclaus J (2007) Will the tropical land biosphere dominate the climate-carbon cycle feedback during the twenty-first century. Clim Dyn 29(6):565–574

Roeckner E, Bäuml G, Bonaventura L, Brokopf R, Esch M, Giorgetta M, Hagemann S, Kirchner I, Kornblueh L, Manzini E, Rhodin A, Schlese U, Schulzweida U, Tompkins A (2003) The atmospheric general circulation model ECHAM5, Part I: model description. Max Planck Inst Meteorol Rep 349:1–127

Sammon JW Jr (1969) A nonlinear mapping for data structure analysis. IEEE Trans Comput C–18(5):401–409

Steed CA, Ricciuto DM, Shipman G, Smith B, Thornton PE, Wang D, Shi X, Williams DN (2013) Big data visual analytics for exploratory earth system simulation analysis. Comput Geosci 61:71–82

Thompson DWJ, Wallace JM (1998) The arctic oscillation signature in the wintertime geopotential height and temperature fields. Geophys Res Lett 25(9):1297–1300. https://doi.org/10.1029/98GL00950

Tudhope AW, Chilcott CP, McCulloch MT, Cook ER, Chappell J, Ellam RM, Lea DW, Lough JM, Shimmield GB (2001) Variability in the el nino-southern oscillation through a glacial-interglacial cycle. Science 291(5508):1511–1517. https://doi.org/10.1126/science.1057969

van den Dool HM, Saha S, Johansson A (2000) Empirical orthogonal teleconnections. J Clim 13(8):1421–1435. https://doi.org/10.1175/1520-0442(2000)013<1421:EOT>2.0.CO;2

von Storch H, Zwiers FW (2002) Statistical analysis in climate research. Cambridge University Press, Cambridge

Walker G, Bliss E (1932) World weather v. Mem R Meteorol Soc 4(36):53–84

Wallace JM, Gutzler DS (1981) Teleconnections in the geopotential height field during the northern hemisphere winter. Mon Weather Rev 109:784–812

Wetzel P, Maier-Reimer E, Botzet M, Jungclaus J, Keenlyside N, Latif M (2006) Effects of ocean biology on the penetrative radiation in a coupled climate model. J Clim 19(16):3973–3987

Whitaker JS, Compo GP, Wei X, Hamill TM (2004) Reanalysis without radiosondes using ensemble data assimilation. Mon Weather Rev 132(5):1190–1200