Exploring Variability within Ensembles of Decadal Climate Predictions

Christopher P. Kappe, Michael Böttinger, and Heike Leitte,

Abstract—Ensemble simulations are used in climate research to account for natural variability. For medium-term decadal predictions, each simulation run is initialized with real observations from a different day resulting in a set of possible climatic futures. Understanding the variability and the predictive power in this wealth of data is still a challenging task. In this paper, we introduce a visual analytics system to explore variability within ensembles of decadal climate predictions. We propose a new interactive visualization technique (clustering timeline) based on the Sankey diagram, which conveys a concise summary of data similarity and its changes over time. We augment the system with two additional visualizations, filled contour maps and heatmaps, to provide analysts with additional information relating the new diagram to raw data and automatic clustering results. The usefulness of the technique is demonstrated by case studies and user interviews.

Index Terms—Clustering, ensemble simulations, climate research, visual analysis.

1 Introduction

Climate simulations with coupled atmosphere-ocean models have been carried out for many years in order to estimate possible future climate changes due to increasing atmospheric greenhouse gas concentrations. In contrast to weather forecasting, where models need to be initialized with actual observations, multi-century climate simulations with coupled models are dominated by boundary conditions and do not require precise initialization data. Recent research tries to close the gap between deterministic weather forecasts and probabilistic climate projections by initializing ocean models with ocean reanalysis data and employing ensemble simulation techniques. The goal is a "seamless prediction" on varying temporal scales. For the time scale of up to ten years, so-called decadal climate prediction systems are developed that aim at forecasts of variations in temperature and precipitation. Both factors have a large impact on societal systems (droughts, floods).

Two key quantities in these ensemble simulations are ensemble spread and statistical skill. The ensemble spread is a potential measure for the internal variability of the climate system. The statistical skill of a forecast system is estimated by comparing retrospective forecasts (so called hindcasts) with observations. Conventional analyses mainly focus on spatial representations of temporal means [1] as well as temporal developments of spatial means [2]; analyzing different forecasts together with different spreads for various forecast periods is difficult. There is a lack of tools that allow a compromise between displaying every single ensemble member in full detail and summarizing a whole ensemble just by its mean and maybe the standard deviation.

To address these challenges, we present an interactive visual analytics system that supports climate scientists in better structuring and understanding the large amount of data contained in spatio-temporal ensemble simulations. Our system, which is the result of joint work of computer scientists and climatologists, enables identifying major patterns in their data using clustering. We also provide a variety of tools to verify clustering results. Afterwards analysts can visually explore the temporal evolution of pattern sizes and their variability. To this end, the system offers three linked views: a clustering timeline, a heatmap view and filled contour maps. The clustering timeline is inspired by Sankey diagrams [3], [4] and depicts the temporal evolution of the clustering. The heatmap view helps analysts validate the clustering results as it visualizes all the pair-wise distances between ensemble members, and the filled contour maps link the analysis results back to the simulated/observed data.

Our contributions are described as follows:

- design principles for analyzing variability in decadal climate simulations;
- an interactive visualization system to enable analysts to better understand patterns and variability in ensembles and to compare ensembles to each other;
- a Sankey-inspired diagram that we call *clustering timeline* for communicating clustering results in ensembles.

2 Related Work

In the following we introduce the scientific background of the application data and discuss related work in ensemble visualization, clustering of ensemble members, and flow charts for temporal data.

2.1 Prediction Systems and Climate

Prediction systems in environmental sciences fall into two major categories. Weather forecasts are regularly produced for up to ten days. They are based on most recent observations, data assimilation techniques and ensemble simulations

- *C. Kappe and H. Leitte are with the Department of Computer Science, TU Kaiserslautern, Germany.*
E-mail: {kappe,leitte}@cs.uni-kl.de
- *M. Böttinger is with the Deutsches Klimarechenzentrum GmbH, Hamburg, Germany.*
E-mail: boettinger@dkrz.de

Manuscript received September 06, 2017; revised February 16, 2018.

with atmospheric models. Specifically for short time scales, these prediction systems achieve high forecast skills. Typical uninitialized climate simulations using coupled models of the Earth system span longer time periods of up to hundreds or thousands of years, and due to the internal variability of the climate system, only statistical descriptions of the data for longer time periods such as 20 or 30 years can be used for analyses.

By developing a decadal climate prediction system, the German national research project MiKlip aims to fill this zone of lacking predictability. A decadal prediction system simulates not only the climate response to future natural and anthropogenic forcing but also the future evolution of internal climate variability. The most common visualization techniques for the analyses are color-mapping of scalar measurements on the respective domain and time series [5]. These techniques are also broadly used in communication with the public (like on websites) [6], [7].

Recent research has addressed visualization of the data combining color-mapping of the ensemble-mean temperature anomaly with a visualization of the corresponding per-vertex standard deviation and the predictive skill in the spirit of multi-field visualization [8]. In contrast to the latter method, which for each point in time visualizes statistical quantities describing the ensemble together with the forecast field, we aim here at a more detailed analysis of the full spatio-temporal ensemble data set.

2.2 Ensemble Visualization for Spatial Data

Several techniques have been proposed to visualize (time-dependent) field ensembles. Liu *et al.* present a comparative visualization of vector field ensembles looking at common paths of trajectories [9]. Shu *et al.* analyze the variability in scalar field ensembles. High-variance regions are extracted and tracked over time. Their connectivity is shown in an *EnsembleGraph* [10]. Bensema *et al.* use statistical analysis of ensembles of fields to classify the modality classes and integrate these findings in color-coded representations [11]. Demir *et al.* combine volume visualization and information visualization of derived quantities to compare 3D ensembles [12]. A common approach is also to look into distributions of contours in ensembles as discussed in [13]. These papers, however, lack a visualization of the temporal development of clusters of ensemble members such as the one we propose.

Another noteworthy tool dedicated to the visualization of ensembles or general probabilistic forecasts is Albero [14]; here, clustering is not so much the concern but it is an aid in the visual analysis of the uncertainty in Numerical Weather Prediction Systems and its coordinated multiple views approach partly draws on the same building block as our approach (such as colormap visualizations).

2.3 Clustering of Ensemble Members

In the context of ensemble data, clustering can be used to reduce the data size and complexity by deriving groups of ensemble members with similar features. Here, the relative size of these groups represents the probability of occurrence for these features.

Kothur *et al.* [15] apply clustering to geographically referenced temporal profiles to mine for patterns in time and within the ensemble (while we divide time series into steady slices for our clustering). Using the presented approach users can detect temporal profiles representing geophysical processes and compare two datasets to each other.

Bordoloi *et al.* employ clustering on spatial probability density functions. On the one hand, pixels are clustered with respect to the local uncertainty, on the other hand whole realizations are clustered [16]; the temporal development, however, is no concern. Correa *et al.* perform clustering on data fraught with uncertainty by taking into account the fact that – together with the whole dataset – the distance between individual data points becomes uncertain, which has to be considered in clustering [17]. Bruckner *et al.* propose a tool to cluster ensembles of unsteady visual effects are after merging several time steps into continuous segments [18]. Recently Obermaier *et al.* [19] presented an ensemble exploration technique that combines a flow chart based on the simulation parameter space with spatial observations. In spirit, this system is relatively close to our approach, however, their notion of trends depends on local features like the scalar value at a specific position, while we compute clusters based on overall similarity of the scalar fields.

Ferstl *et al.* propose an approach to clustering and visualization of ensemble data [20] similar to ours, with the focus on meteorological data and under the assumption that clusters do not merge in the course of time (that we do not share). Clustering applied to time-dependent flow data can also be found in [21]; it builds upon *Parallel Sets* [22] which relates to cluster visualization insofar as cluster affiliation can be viewed as a categorical attribute.

Similar to our approach, clustering and visualization of the underlying dissimilarity matrix is used by Jarema *et al.* [23] and Wang *et al.* [24], where the latter does not show pairwise distances but the Mean Squared Error (with respect to observation data) of an ensemble member (column) over several time steps (rows).

2.4 Flow Charts and Stacked Graphs

Flow charts and stacked graphs combine multiple data series in a single chart using a parallel layout with or without intermediate space. Hence, they are closely related to our problem of visualizing temporally changing cluster membership. The following papers were influential works regarding the design of the clustering timeline.

Havre *et al.* introduced *ThemeRiver* [25], employing a river metaphor to stacked graphs to visualize temporal thematic variations in documents. *Streamgraphs* [26] present novel layout strategies that significantly improve the visual appearance and readability. Dork *et al.* visualize a continuously updating information stream with interactively stacked graphs [27]. The TextFlow [28] and ThemeDelta [29] systems for time-varying document collections are able to extract and track topics with a model that allows splits and merges, similar to the time-dependent clustering that we use. In TextFlow the layout is computed based on a directed acyclic graph, an approach that we also pursue. And ThemeDelta uses a visual design that is not so different from our solution with many individual paths still recognizable in the timeline.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2810919, IEEE Transactions on Visualization and Computer Graphics

Waser *et al.* present the world lines technique which uses a tree-like flow chart to visualize a multitude of flooding simulations [30]. Ensemble techniques mentioned before also make use of flow charts to ease communication [10], [19].

3 System Design

In this section, we discuss the visualization challenges and design rationale of our system. We have roughly followed *Munzner's Nested Model* [31] of domain problem characterization, data/operation abstraction design, encoding/interaction technique design and algorithm design.

3.1 Analysis Tasks

The presented system was developed in close collaboration with domain experts. The web-based tool is constantly available to them with regular updates. Based on our common meetings and interviews we compiled the following list of analysis tasks with them. This list formalizes our mutual understanding and expresses problems within the application domain and challenges faced by the respective researchers.

Q.1) *Are there dominant patterns in the data? How distinct and reliable are they?* To structure the large-scale simulations, fundamental patterns have to be found. Many of them are known to climatologists from experience. The question is whether we can extract meaningful patterns automatically. How can we provide visualization tools to communicate those patterns and the variability of matched data points?

Q.2) *How can the structure, magnitude and variability of the discovered patterns be validated and communicated?* Climate simulations contain by nature a lot of natural variance. How can we make sure that we pick up relevant patterns? The climatologists also want to be given means that help them link the new findings to their data in well-known formats.

Q.3) *How persistent are those patterns over time? Do realizations, which appear to be similar at a point in time, stay in the same cluster or do they change frequently?* The analysts want to understand the temporal evolution of climatic patterns. When do they occur and how do they change over time? It is also important to them to understand the contribution of the individual simulation runs to the clusters and their temporal evolution.

Q.4) *Can the prediction skill be improved by looking at particular clusters?* The prediction skill is a measure for the success of predictions. The climatologists are interested in the question whether the system can help them further improve the prediction skill.

Q.5) *How do multiple ensembles compare to each other? What are the effects of different initializations? How do different regimes compare?* Prediction systems are commonly evaluated using many individual simulations (initialization at many time steps). Climatologists want to understand how these initializations compare to each other and how strong the effects are. They also want to compare multiple initialization models to understand the effect of the provided model input.

3.2 Analysis Challenges and Design Rationale

As stated earlier, the state-of-the-art in the visual analysis of decadal climate simulations ranges from 1D and 2D plots to multi-field scalar field videos. These types of visualizations are very intuitive and easy to compute.

However, while the current routine provides an easy measure to estimate the overall quality of a simulation model, it lacks detailed analysis functionality for spatio-temporal processes. The analysis of videos with multi-field visualizations is a very coarse and mentally exhausting task that provides only rough qualitative insights and may easily suffer from missing important pieces of information in the data. In close collaboration with climatologists, we identify a set of design goals that help analysts get a better understanding of decadal climate ensemble simulations.

D.1) **Variability visualization:** We use clustering to identify major patterns in climate simulation ensembles. In our system we will provide means to visualize these patterns and enable the analyst to compare patterns and raw data to estimate variability in the input. A second source of variability are the likelihood of climatic phenomena covered by the distribution of temporal patterns in the ensemble. The system requires means to communicate the likelihood and temporal evolution of patterns.

D.2) **Storytelling metaphor:** As we are working with long time series in ensembles, a storytelling metaphor is desired by the analysts to rapidly see and explain important patterns in the data (Q1-4). The metaphor shall cover global trends and enable the analyst to highlight individual members or subgroups to explain their findings. It shall also allow for extra space to add annotations such as raw data, derived parameters, or visualizations of additional analyses. We chose a Sankey-inspired visualization as the starting point for the analysis as it has a primary axis for time and clearly depicts the temporal clustering results. The horizontal layout allows for easy annotation and can be used readily as centerpiece for communication.

D.3) **User confidence:** Climatologists have a thorough understanding of mathematics and want to understand the decisions made in the automated analysis. Clustering is a critical part in our pipeline and it is important to develop visualizations that allow the analyst to explore clustering decisions (Q1-4). Hence, we make the simulation data accessible to the analyst in a well-known format (filled contour maps) and we include augmented heatmaps to inform about clustering decisions and the confidence of the clustering.

D.4) **Interactive pattern unfolding:** A critical design rationale is interactive data exploration enabling the analyst to interact with the data directly, see results immediately, and have a linked interface between the different visualization modalities, shedding light on varying aspects of the data and derived quantities (Q1-5). The system should provide an overview of how variability changes within the ensemble over time to identify interesting patterns (e.g. cluster structure). Starting from here, the analyst shall be able to gain further insight into patterns and be provided with means to reason about the causes of them.

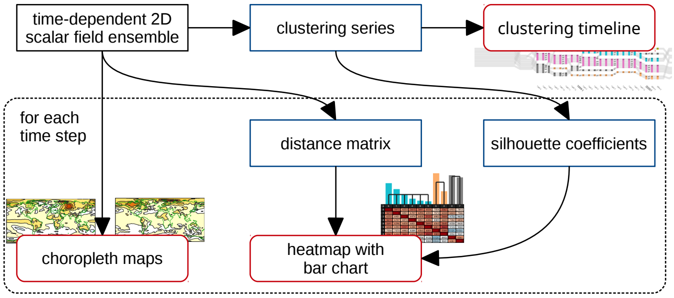


Figure 1. System overview: Our system features a data processing part (blue rectangles) and a visualization part (red rounded rectangles). The clustering timeline provides an overview of temporal similarities and clustering. The augmented heatmap and filled contour maps help analysts interpret the clustering results and temporal patterns.

3.3 System Overview

The whole system is implemented as an HTML5 application that can either be used remotely (client-server mode) or locally. An analysis session starts with a data processing part, where the necessary data is obtained (uploaded by the user or selected from the repository of the server), the distance matrices and clustering are computed, and the results are finally passed to the visualization application. The remaining analysis is interactive. The system consists of three major views (red rounded rectangles in Figure 1): the clustering timeline, an augmented heatmap, and filled contour maps.

The clustering timeline provides an overview of temporal similarities and clustering.

The augmented heatmap and filled contour maps help analysts interpret the clustering results and temporal patterns by displaying detailed information about the raw (scalar field) and computed (scalar field distances, silhouette coefficients) data for single time steps.

4 Cluster Analysis

Data processing is an integral part of the presented system to provide the analyst with a meaningful structuring of the huge amount of data. In discussions with the domain experts we settled on cluster-based analysis to uncover major trends and variability in the simulations. As climate datasets feature cyclic as well as irregular patterns on multiple timescales, we decided on distance-based clustering with each scalar field being an individual data point to uncover recurring patterns. To make this more precise, suppose we have an ensemble simulation with 10 members (simulation runs) and 100 time steps each. This results in 10×100 data points, each representing a 2D scalar field, that are clustered concurrently.

What remains to be discussed is the choice of distance function (Section 4.1), the chosen clustering algorithm (Section 4.2), and methods for the validation of the clustering results (Section 4.3).

4.1 Distance Function

The input data to the clustering are scalar fields, and finding a good distance function (metric) for such high-dimensional data points is a difficult task. The choice commonly heavily depends on the application domain, prior knowledge, and targeted features. On the one hand, choosing a good distance function is critical to obtain meaningful clustering results, on the other, the function can easily be changed to account for novel findings and data-dependent questions. Hence, we decided to start with a standard choice that is well-accepted in the application domain. For the results presented in this paper, we applied a vertex-based distance function for two fields s_1 and s_2 using the $p1$-norm:

$$\mathsf{d}(s_1, s_2) = \sum_{v \in V} \omega(v) \cdot |s_1(v) - s_2(v)|$$

where V is the (common) set of vertices. The weight factor ω accounts for the varying cell sizes when using regular grids based on spherical coordinates (the polar areas are covered more densely than the equatorial area). Notice that we use the same distance function for comparison of scalar fields (e.g. cluster means) with ground truth data (see error charts in Figure 7). But we normalize the result with the sum of the weights to gain more meaningful labels, namely a measure in Kelvin (difference of temperatures in degree Celsius). Similar results were obtained using related measures such as the l_2 metric or the cosine similarity.

4.2 Clustering Algorithm

There are several well-established generic clustering algorithms which can always be taken into consideration.

There are the density-based methods [32] which are intriguing because they can discover clusters of arbitrary (non-spherical) shape; the number of clusters need not (but also cannot) be given by the user. It depends, however, on a parameter defining the minimum number of points a cluster must have (could be set to 1) and – more critically – a minimum distance between points that reflects what “dense” means for the given dataset. Estimating a good value for this parameter works best with a high number of points in a low-dimensional space. Because our application is contrary to this, we dismissed this clustering approach. However, for larger ensembles whose members are previously simplified (e.g. a reduced grid resolution) it may be reconsidered.

There are also hierarchical clustering approaches [33], [34], [35]. But these are less straight forward to work with (automatically) because in order to get a classical, non-hierarchical clustering result as we desire, one first has to find a cut through the resulting dendrogram. However, we think hierarchical clustering may eventually be applicable, too, ideally as a variable, interchangeable step in the analysis workflow.

Eventually, we chose the *k-means* algorithm [36] for the following reasons: (i) the algorithm is easy to understand, (ii) only the number of clusters k has to be provided (the most suitable k can be determined automatically, however), (iii) it directly yields both a clustering and the corresponding cluster centers which reflect the represented climate states.

The required input parameters for k-means are the desired number of clusters k and initial positions for the clusters. Initial data points for the cluster centers are chosen randomly. To avoid locally optimal solutions, we run k-means multiple times (on average 40 times) and choose the clustering with the best silhouette coefficient (see below.) Unless the user explicitly specifies a desired number of clusters, we search for a good partitioning of the data by

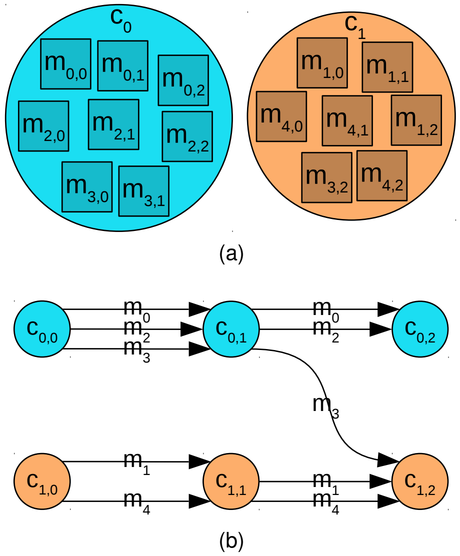


Figure 2. (a) Clustering of ensemble members m_i split into individual time steps $m_{i,t}$. (b) A series of clusterings (constructed from (a)) modeled as graph. Nodes represent a cluster k at a certain time step t (labels: $c_{k,t}$), edges exist for each ensemble member between nodes of consecutive time steps.

trying multiple settings for k and select the one which yields the best silhouette coefficient.

4.3 Cluster Quality Analysis

The criterion used to finally pick a k is the silhouette coefficient [37] of the clustering (or short the silhouette). This number can be computed for individual points, single clusters or a complete clustering (the set of all clusters that have been determined for a dataset). The silhouette is always normalized to be in the range $[-1, 1]$. For an individual point i it states how well that point belongs to its own cluster a and at the same time how distinct it is from its nearest other cluster b. It is defined as:

$$\mathsf{sil}(i) = \frac{\mathsf{d}(i,b) - \mathsf{d}(i,a)}{\max(\mathsf{d}(i,b), \mathsf{d}(i,a))}$$

where $\mathsf{d}(i,x)$ denotes the distance (by means of the previously defined function) between i and the respective cluster centroid. Resulting values can be interpreted as follows. Close to 1: Good, the point lies very central in its own cluster. Close to -1: Bad, the point should actually have been put in another cluster. Close to 0: The point is hard to cluster as it lies just between two clusters.

The silhouette of a cluster or a whole clustering is simply the mean of the silhouette of the respective points. As a very intuitive measure of the quality of a clustering, the silhouette tends to increase with a higher number of clusters – more and more independently of the actual clustering; a clustering where each object is in its own cluster will always get the highest score. To accommodate this, we follow the approach to add a small factor (we use 0.92) that penalizes very small clusters.

5 Visual Encoding Methods

As detailed before, our system consists of three major views: the clustering timeline (Section 5.1), the distances-heatmap/silhouette-plot (Section 5.2), and the filled contour maps (Section 5.3). All views are linked with interactions (Section 5.4).

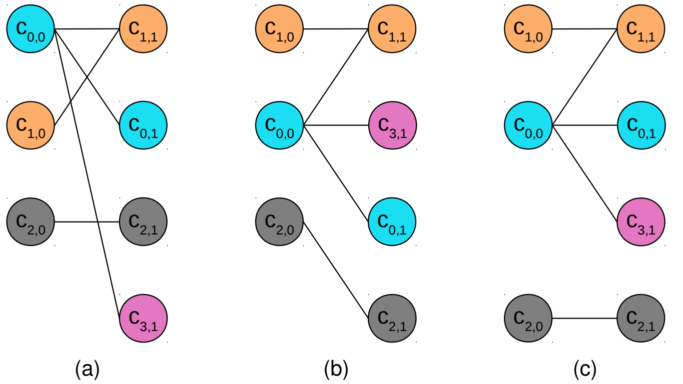


Figure 3. A bipartite graph and its possible embeddings. Nodes are labeled $c_{c,t}$ where c the cluster ID and t is the time step. (a) Arbitrary. (b) Minimized crossings. (c) Node positions optimized for smooth time step transitions.

5.1 Clustering Timeline

The centerpiece of our system is the *clustering timeline* (see e.g. Figure 4 top) which gives a concise overview over the similarity structure in the data for an entire ensemble. Our new design and interaction techniques are inspired by Sankey diagrams [3], [4] and alluvial diagrams [38] both of which represent flow quantities in time-dependent settings.

5.1.1 General Semantics

The clustering timeline is basically a visualization of the directed graph that is defined by the series of clusterings output by the clustering algorithm. We define nodes and edges as follows (see Figure 2 for a small example). A node v represents a cluster $v.c$ at a time step $v.t$. An edge (u, v) represents an ensemble member that belongs to cluster $u.c$ at the time step $u.t$ and to cluster $v.c$ at time step $v.t$. It must hold that $v.t = u.t + 1$ (edges may only exist between nodes of subsequent time steps; no ensemble member may "skip" a time step). There is an edge for each ensemble member for each pair of time steps $(t, t + 1)$. Because, in general, clusters contain more than one member, and members which have been clustered together once often stay in the same cluster for some time, a lot of parallel edges can be expected in this model.

Notice that we distinguish between *local* and *global* clusters (in a temporal sense). When, for one time step, there are three clusters, one may – locally – identify them as clusters 0, 1 and 2. But we also have the notion of clusters which may exist over several time steps. These stem from the concurrent clustering of scalar fields of all time steps. So they exist in different instances at different time steps (ensemble members may come and go) but share a common cluster centroid. These time-independent clusters get a unique global ID, respectively, and also a different color. These colors come from a predefined set of colors chosen with regard to aesthetics and distinguishability based on the proposal in [39]. In case of an extraordinary high number of clusters the set is dynamically extended as proposed in [40].

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2810919, IEEE Transactions on Visualization and Computer Graphics

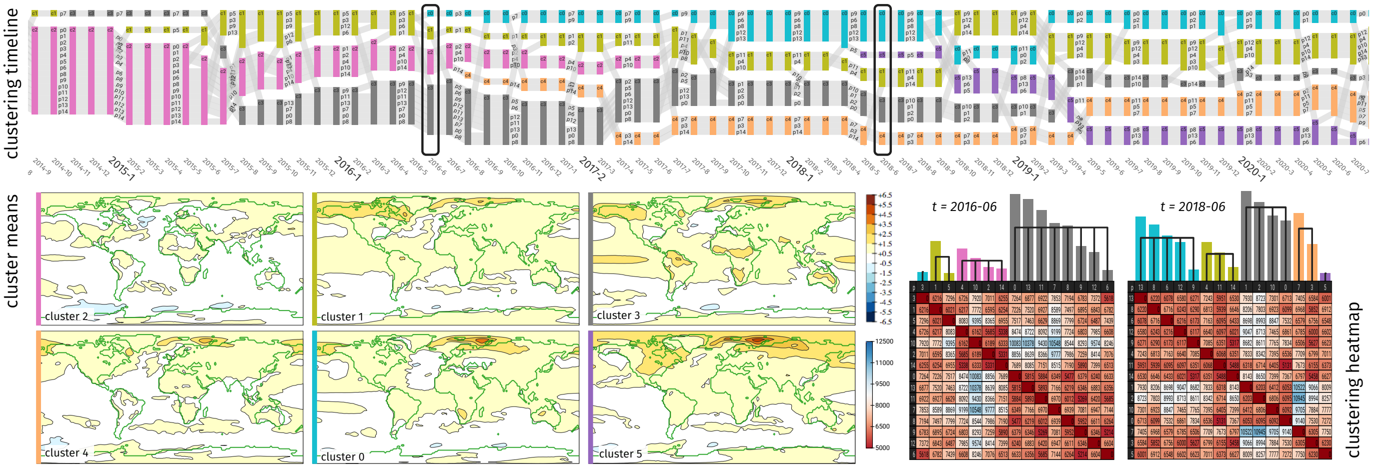


Figure 4. Cluster-based climate ensemble analysis: (top) The clustering timeline gives an overview of temporal occurances of clusters and the transitions of simulation runs between them. (bottom left) The cluster means encode the represented climatic state of each cluster independent of time. (bottom right) Clustering heatmaps support the validation of clustering results (here for the time steps highlighted with rectangles).

5.1.2 *Layout Computation*

Envisioning a good layout for the clustering timeline, we think the following objectives should be met. Clusterings should be arranged along a horizontal timeline. So, the x-coordinate of a node should correspond to the time step it refers to. The y-coordinate should be the same as the one of the node with the same global ID from the previous and next time step, so that a temporal coherence (if existent) is visible at once. The edges should have a constant width (vertical extent) so that the width of a bundle of them corresponds to the number of members exhibiting the same development. Then the nodes must have a height corresponding to number of incoming and outgoing edges so that they cover exactly the "entrance" and "exit" area of the node (we have opted for rectangles as glyphs for the nodes).

To make the path of an ensemble member through the graph a continuous line, the vertical position at which it joins a cluster node must be the same when it leaves it again. So, to compute an embedding one has to split the cluster nodes into several nodes for each member, and the original nodes become a mere constraint to layout the finer nodes next to each other (vertically). Meanwhile the main concern is to reduce crossings of the member paths and to keep them as horizontal as possible (see Figure 3). In general, minimizing the *bipartite crossing number* is an NP-complete problem [41]. With the goal to deliver a responsive application we have implemented an algorithm that optimizes the layout locally and terminates when the number of changes falls below a certain threshold or a maximum number of iterations is reached.

The approach we follow can be seen as a special case of *layered graph drawing*, first proposed by Sugiyama *et al.* [42] where in our case the layers are the sequence of time steps. In the following we outline the implementation for the layout of the clustering timeline. Because the number of ensemble members per time step is constant, one can picture the data for the layout algorithm as an $n_t \times n_m$ grid (where n_t is the number of time steps and n_m is the number of ensemble members). The basic task is to figure out a good *rank* (one of the n_m vertical positions in this grid) for each ensemble member at each time step. This rank will then translate relatively easy to a y-coordinate.

The grid is initialized with a naive listing of the members per cluster for each time step, respectively, and then optimized iteratively. The optimization assesses the current layout with respect to the number of edge crossings. If swapping the position of neighboring nodes reduces the number of crossings, this new layout is kept. But as mentioned above, the fact that nodes must not swap positions with nodes belonging to a different cluster has to be kept in mind.

5.2 Clustering Heatmap

As clustering is a critical part in the analysis process, appropriate means are required to understand and validate the automatically derived results. Therefore, we integrate an augmented heatmap view that includes information about pairwise scalar field distances and about the clustering. Figure 5 (bottom) gives an example of such a visualization.

For a single time step, the heatmap shows the pairwise distances in colored table cells. We chose a blue-white-red colormap for rapid perception of three major classes (low, medium, high similarity) and scaled the range to the minimal/maximal distance within the entire ensemble (over all time steps) to ensure color stability and comparability. Sorting the matrix according to the clustering ensures easy analysis of the intra-cluster variability. Within each cluster, the data points are sorted according to the silhouette coefficient with decreasing order.

The heatmap is augmented with silhouette information – comb-like structures above the heatmap in resemblance to dendrograms for hierarchical clustering. Each comb connects all members of a cluster (vertical lines) and its height encodes the silhouette coefficient for the respective cluster. The higher the comb, the higher the silhouette and thereby the similarity to the respective cluster mean.

An additional bar chart underneath the silhouette comb encodes the silhouette of each cluster member. Large divergence from the baseline of the comb indicates that the data point does not fit too well into the cluster and is close to another one.

5.3 Filled Contour Maps

Working with an abstraction of the input data, such as the clustering timeline, it is important to be able to have a look at the original input data; both for validation of the clustering results and for climatological interpretation. Therefore we have included a geospatial visualization of the time-dependent 2D scalar field ensemble using color-mapping of the scalar values. We have opted for filled contour maps based on a discrete colormap that the climatologists are already familiar with. Recently such visualizations based on binned scalar ranges have been studied with the result that users can read such maps indeed accurately [43], sometimes better than continuous colors. We have further enhanced the visualization with additional isolines and line overlay of the continental outlines. We offer this scalar field visualization not only for all ensemble members for all time steps but also for all the cluster means and the ensemble mean.

5.4 User Interactions

We allow several kinds of user interactions in our tool. The various visualization modules of our software can be combined dynamically by the user to gather all the desired information on demand. At first the focus is on the clustering timeline. Its size can be adjusted while scroll bars allow to navigate through the visualization if a large zoom factor is chosen or the time series is particularly long.

The visualizations of the ensemble field data, the distance matrices and silhoutte bar charts (introduced above) can be added on demand and also be adjusted in size and position in the overall visualization layout. Because these views correspond to single time steps only, we have implemented them to keep in sync with a time step controller similar to those of video players (back/forward buttons, slider); the currently selected time step is also highlighted in the clustering timeline.

Within the clustering timeline, one or more ensemble members can be selected (the state is toggled by left click on edges or nodes) to highlight their paths through the graph (Figure 5 (top)). This way it becomes particularly easy to see when an ensemble member changes from one cluster to another. By default this is not always easy to spot because the path of a single member may drown in the set of parallel edges.

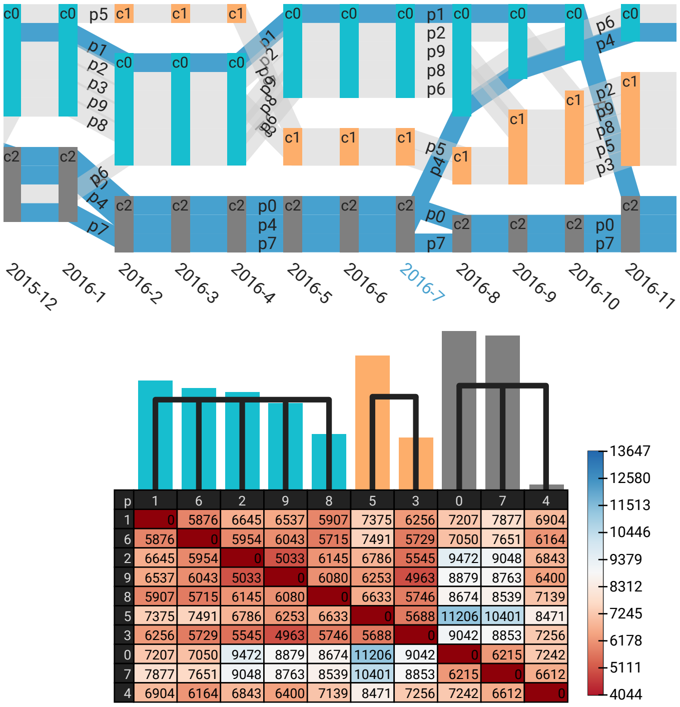
The user can also get a quick view of the scalar fields included in a cluster by hovering over a node with the mouse. The shown images are miniatures of the filled contour maps that can also be looked at in the separate view.

Lastly the user can adapt a *horizontal scaling* parameter which scales the visuals (unproportionally) along the x-axis. Thus the timeline for a long series can be made to fit the window. The distinct colors of the clusters together with the optimized layout make the resulting distorted image still well-readable for the purpose of an overview (see the video in the supplemental material).

6 Results

Driven by the previously defined analysis tasks (Section 3.1), we conduct three case studies together with experts. In the first one (Section 6.2), we analyze the clustering procedure, its parameters, output, and quality (Q.2+3+4). In the second case study (Section 6.3), we look into data variability and temporal cluster structures to gain insight about the simulation skill. A comparative analysis is conducted with respect to ground truth data and two different initializations (Q.1+3+5). In the last case study, we analyze and compare three simulation ensembles that use different numerical models (Q.5).

In the following we first give some details about the ensemble datasets so that the reader can comprehend the discussion of the findings.



p	1	6	2	9	8	5	3	0	7	4
1	0	5876	6645	6537	5907	7375	6256	7207	7877	6904
6	5876	0	5954	6043	5715	7491	5729	7050	7651	6164
2	6645	5954	0	5033	6145	6786	5545	9472	9048	6843
9	6537	6043	5033	0	6080	6253	4963	8879	8763	6400
8	5907	5715	6145	6080	0	6633	5746	8674	8539	7139
5	7375	7491	6786	6253	6633	0	5688	11206	10401	8471
3	6256	5729	5545	4963	5746	5688	0	9042	8853	7256
0	7207	7050	9472	8879	8674	11206	9042	0	6215	7242
7	7877	7651	9048	8763	8539	10401	8853	6215	0	6612
4	6904	6164	6843	6400	7139	8471	7256	7242	6612	0

Figure 5. Top: Clustering timeline (cropped left and right) with ensemble members p1, p0, p4 and p7 selected. Bottom: Distance heatmap and silhouette chart for month 2016-7. Notice how the imminent change of cluster of member p4 is indicated by a silhouette bar that has almost dropped to 0 (to be read as "in between two clusters").

6.1 Datasets

All datasets are simulations of the air temperature at two meters height, expressed as *anomalies* (deviations) of some predefined state (see Section 6.1.1 for details). The data is given on a regular grid (in spherical coordinates) that covers the whole earth. The spatial resolution is 192 × 96 (longitude × latitude). In general the ensembles contain 10 to 15 members. A summary of the datasets is provided in Table 1.

We will consider two simulations starting in 1997 and 1998 using the baseline initialization. Three simulations start in 2013 and use different initializations. The first two datasets are so called hindcasts, i.e. numerical simulations of the past that enable the climatologists to compare their simulations with so called *reanalysis* data. These are consistent time-dependent gridded data sets of the atmospheric state derived on the basis of observations, an atmosphere model and a data assimilation system. Reanalysis data represent a best fit of the numerical model to observational data. For the studies

Table 1
Dataset overview: For the analysis we use five ensemble simulations. The *name* indicates the prediction system and year of initialization. *# runs* gives the number of ensemble members and *# time steps* the number of time steps per simulation with the respective real time (*time range*). The *reanalysis* dataset contains real-world temperature observations [44].

name	# runs	# time steps	time range
baseline1-1996	10	109	[1997, 2006]
baseline1-1997	10	109	[1998, 2007]
baseline1-2013	10	109	[2014, 2023]
prototype-gecco2-2013	15	109	[2014, 2023]
prototype-oras4-2013	15	109	[2014, 2023]
reanalysis	1	229	[1997, 2016]

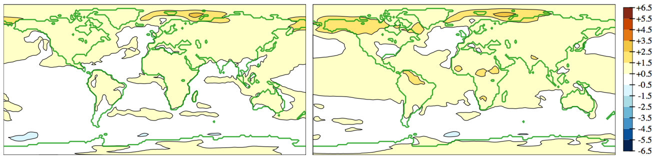


Figure 6. Cluster means of *prototype-oras4-2013* for $k = 2$: (left) cluster 0 continental warming, (right) global warming.

described here we used data from "The NCEP/NCAR 40-year reanalysis project" [44], which we refer to as ground truth or observation data in the following (see also Table 1).

6.1.1 Climate Simulation Settings

The decadal climate predictions that we have worked with were developed within the MiKlip project [2]. The prediction system is based on the coupled earth system model MPI-ESM of the Max Planck Institute for Meteorology (MPI-M). The initialization uses the ORAS4 and GECCO2 [45] ocean reanalyses [46] for the ocean and for the atmosphere ERA40 [47] until 1989 and ERA-Interim [48] thereafter. The global atmospheric model component ECHAM is used with a horizontal resolution of T63/L47 (~200 km, 47 levels vertically) and 1.5 degrees/L40 in the oceanic component MPIOM. We have used the global 2m temperature anomalies relative to the period 1961–2010 for each ensemble member on a monthly basis. The anomalies were low-pass filtered with a one-year running mean. A central question will be how the ensemble members differ and where analysts can see similarities.

6.1.2 El Niño and La Niña

Two specific phenomena that play a role in the sections below are *El Niño* and *La Niña*, which are also known, respectively, as warm and cool phases of the El Niño/Southern Oscillation (ENSO) [49]. ENSO is associated with anomalies in the sea surface temperature (SST) in a region in the equatorial Pacific (between approximately the International Date Line and 120 °W). In this relatively large region the SST is known to rise above (El Niño) or drop below (La Niña) the average temperature at irregular intervals with strong effects on nature and humanity.

6.2 Cluster Analysis

The first step in our analysis procedure is the k-means clustering of an ensemble dataset into k clusters of similar fields; recall that these clusters are computed over the whole time series, so at a select time step not all k clusters need exist. We will walk through this analysis session using dataset *prototype-oras4-2013*. As detailed before, our system provides the opportunity to automatically compute the optimal k parameter based on the silhouette coefficient (result here: $k = 2$). The two resulting cluster means are depicted in Figure 6: cluster 0 represents continental warming, and cluster 1 global warming both by about $1\,°\mathrm{C}$. Comparing the means to the filled contour maps of the 15 simulation runs, we see that the means reflect the major trends in the data well. What they still lack is a more detailed representation of extremal events such as El Niño and La Niña. Hence, the domain experts explore multiple other k-values.

To find a suitable parameter for k we provide several visual aids as depicted in Figure 4. An important cue are the cluster means as we have seen in the previous paragraph. Figure 4 shows the cluster means for $k = 6$, i.e., six representative climate patterns. The $k = 2$ clusters are still present: cluster 4 (orange) and cluster 1 (green). What we also observe is that we get more variations of the two initial patterns. The upper row (pink, green, gray) shows increasing temperature combined with warming in the Pacific Ocean (El Niño event). The lower row (orange, blue, purple) reflects increasing global warming with cooler than average temperatures in the Pacific Ocean (La Niña or neutral event). Similar findings are made for other settings of k.

To avoid overfitting during the clustering, additional information is required. This is provided through the clustering timeline (Figure 4 for $k = 6$). It shows for each time step how many ensemble members belong to a certain cluster and how their affiliation changes over time. For $k = 2$ most simulations feature two large strands that vary in size, often in a quasi cyclic pattern. As k increases, the strands in the clustering timeline get more decomposed. For $k = 6$ (Figure 4) we can still observe major trends (dominant strands). As k is further increased, this structure starts to dissolve and many single-member strands occur. For our data, we found that $k \in [5, 7]$ is commonly a good setting and that the results are stable across the values.

An important design goal was also to allow for high user confidence and trust in the technique. Hence, we also provide more detailed visual and numerical feedback about the clustering using the clustering heatmap. Two sample heatmaps are given in Figure 4 (bottom, right). The respective time steps are highlighted in the clustering timeline in the same figure by black rectangles. Red color in the heatmaps indicates good agreement and we observe that there are well-pronounced cluster squares along the diagonal. The silhouettes bar charts on top indicate which members belong to which cluster and how well the members fit into their cluster. Overall, we see that the clustering works well and groups similar simulation runs together. Furthermore, we can observe natural variability in the data. Often we observe smooth transitions between major states rather than clearly isolated clusters; so certain clusters appear to be relatively similar. For the left heatmap there is high similarity between the (small) blue and green and the (bigger) pink cluster, and for the right heatmap between the (now bigger) blue and green cluster. Pink represents here a rather neutral state

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2810919, IEEE Transactions on Visualization and Computer Graphics

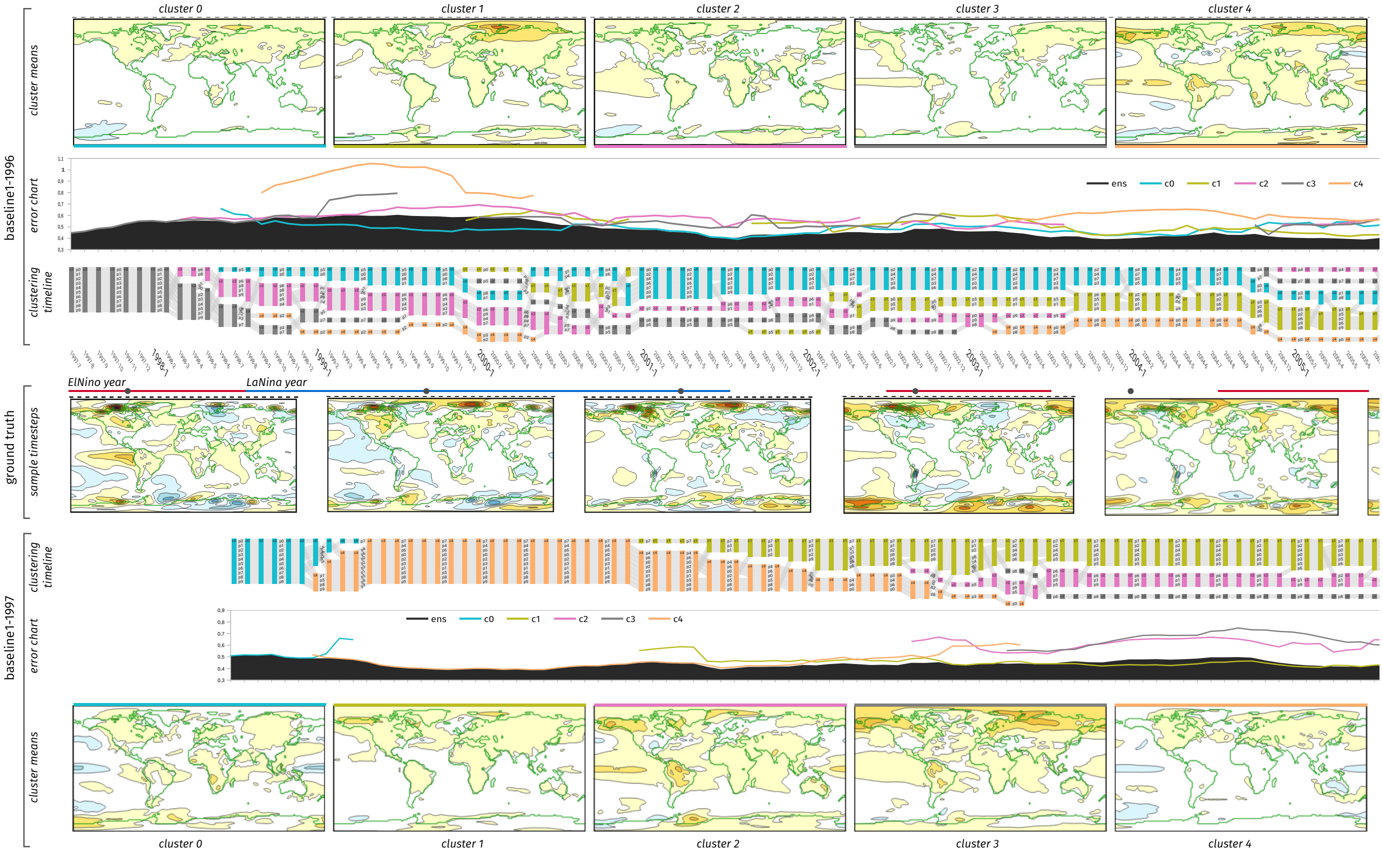


Figure 7. Visual analysis of the ensembles *baseline1-1996* and *baseline1-1997*: The figure is a summary of the relevant charts to analyze and compare the two datasets. For each dataset, we provide the clustering timelines, the error chart, and the cluster means (see labels on the left). The central part contains ground truth information including the time labels, indication of El Niño and La Niña years (red and blue bars), and some sample filled contour maps from the ground truth. Dots on top indicate the respective time step.

with little isolated warming, while blue and green represent stronger global warming combined with La Niña and El Niño characteristics, respectively. As the data transitions slowly between these states, it becomes clear that additional parts in the clustering heatmap will also feature high similarity. This can aid in a detailed analysis to understand major trends in an ensemble analysis.

6.3 Variability Analysis for a Single Ensemble

In the second case study, we walk through two analysis sessions for a single ensemble. We start with dataset *baseline1-1996* and continue with *baseline1-1997*. As both datasets overlap in time, we finish with a discussion of the two analyses.

6.3.1 *baseline1-1996*

Based on previous experience, the domain experts settle in the clustering for $k = 5$. The respective cluster means are depicted in Figure 7 (top): (cluster 0) represents warming in the north pole area, (cluster 1) represents warming over the land masses, (clusters 2, 3, 4) represent different degrees of warming combined with El Niño events. The color code for each cluster is given by the bar above/below each image.

The same colors are used in the clustering timeline (Figure 7 (top part, bottom chart)) to ease identification of clusters over time. The depicted data covers nine years ranging from 1997 to 2005. The clustering timeline features a good agreement of the ensembles in the early phase (until 1998), where all members belong to cluster 4 (El Niño pattern, little warming). This is in good agreement with the ground truth data (Figure 7 (center part)) where we see a strong El Niño event. The color code on top of the ground truth maps indicates El Niño and La Niña years. The x-coordinates are aligned for all temporal charts to allow easy comparison. The timescale is provided in the center of the figure. Looking at the remainder of the clustering timeline, we see that there is commonly no clear trend for one of the clusters. Most of the time many different clusters are present and often feature equal shares in members.

In the years 1998 through 2001, we went through a La Niña period (see ground truth images for a sample). This phenomenon is hardly present in the simulation data and hence no corresponding cluster is identified by k-means. For this time period, most runs belong to either the pink or the blue cluster. We also observe that all other clusters are present with a few members; this indicates a large variability in the ensemble and little agreement.

In the remaining time steps (starting in 2001), the blue cluster (little warming) is dominant, which agrees with observations, followed by the green cluster which represents global warming. Throughout this phase, we also observe a constant presence of clusters indicating an El Niño event (pink, gray). Overall, we can state that variability is very high in this ensemble, and in most time periods many clusters

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2018.2810919, IEEE Transactions on Visualization and Computer Graphics

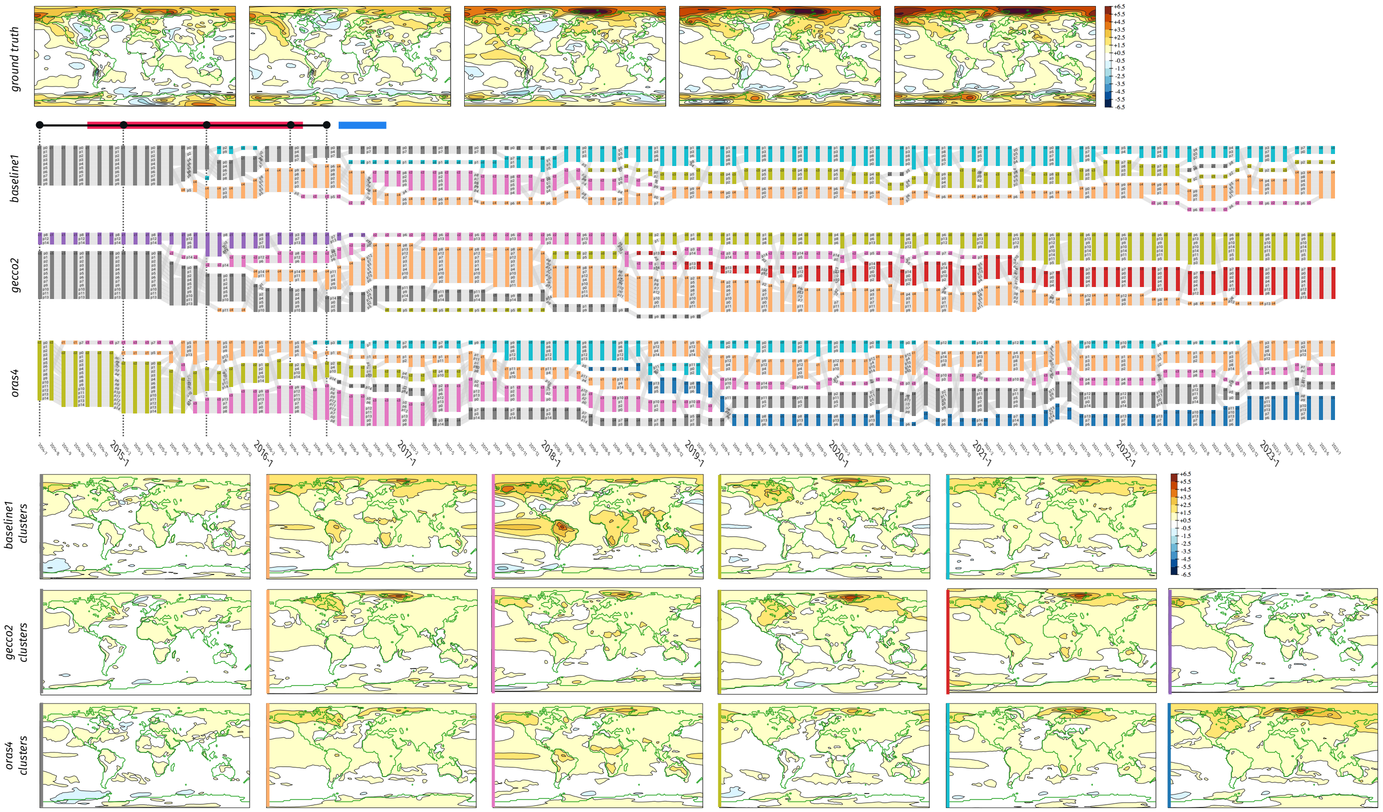


Figure 8. Comparative analysis of multiple simulation models: (top) ground truth filled contour maps for the time range in the past. Dots on the timeline indicate the respective time point. Color encodes observed El Niño (red) and La Niña (blue) events. (center) Clustering timelines for the three models with $k = \{5, 5, 6\}$ for the three models. (bottom) Cluster center of the three ensembles sorted with respect to matching patterns.

with highly varying predictions are present.

We want to analyze this in more detail and look at the prediction skill of the different clusters. Therefore, the analyst is provided with an error chart representing the mean deviation from the ground truth for each cluster (colored lines) and for the ensemble mean (black area). The error chart confirms the initial observations: We start with an average error of $0.45\,°\mathrm{C}$ which increases to $0.6\,°\mathrm{C}$ in the La Niña phase. Best performance during this time is featured by the blue cluster (no equatorial warming). In the second half of the simulation we can observe that the error for the entire ensemble is fairly low ($< 0.5\,°\mathrm{C}$) and commonly smaller than the error for the single clusters.

6.3.2 baseline1-1997

A different picture is given by the clustering timeline of dataset *baseline1-1997* (Figure 7 (bottom part)). We see clear phases of dominance of single clusters, starting with blue, then orange, and finishing with green and pink. The corresponding cluster means are provided below. Blue stands for a La Niña phase with remainders of warming in the equatorial area. Orange features a La Niña pattern with no additional warming. The green cluster represents warming over the land masses, and pink indicates a strong El Niño combined with general warming.

In general, these phases agree well with observed data. In 2001–2002, there is a tie between orange (La Niña) and green (land mass warming), and the ground truth lies actually between those two types of predictions. In 2004–2005, we observe a strong presence of green (warming over land masses) and pink (warming + El Niño), and this was in fact a period when El Niños were observed every other year.

The error chart reveals a generally low error. As ensemble members are often concentrated in one of the clusters, we see that the ensemble error coincides with the error of this cluster. Most of the time, the ensemble error is not higher than any cluster error except for winter 2003/04.

6.3.3 Summary

In the two analysis sessions, we saw two different types of outcomes. Both simulations use the same prediction system with different initializations. In the second case (*baseline1-1997*), the predictions worked really well and little variability is seen in the clustering timeline. In the first case (*baseline1-1996*), we see a much more structured clustering timeline with strong fluctuations in clusters and poorer prediction performance in the first half.

6.4 Comparative Analysis of Multiple Models

In Section 6.3, we analyzed and compared ensembles of same model but for two different time ranges. Now we compare the datasets *baseline1-2013*, *prototype-gecco2-2013*, and *prototype-oras4-2013* (short: *baseline1*, *gecco2*, *oras4*) which cover the same time range (2014–2023) but were created using different models. *Baseline1* contains 10 simulation runs, *gecco2* and *oras4* 15 each. The respective analysis graphic is given in Figure 8.

Using $k = 5$ resulted in good clustering results for the baseline1 dataset. The cluster means are highly descriptive and feature good similarity results in the distances heatmap. Using the same parameters for the datasets *gecco2* and oras4 achieves satisfactory results; but interesting subordinate features (e.g. missing temperature increase in the Pacific Ocean) are often located between clusters and occurred half the time in one cluster and half in the other. Hence, k is increased to 6, which results in coherent climate patterns (cluster means). Figure 8 (bottom) shows the cluster means for the three models sorted in such a way that, where possible, matching patterns are in the same column for all three simulations. The gray cluster represents little warming. Orange, pink, and red represent a $1\,^{\circ}\mathrm{C}$ warming in large parts of the earth combined with El Niño events of varying intensity. The purple cluster represents an El Niño event while general warming is absent. Green, light blue, and dark blue represent a general warming trend which is not present in parts of the Pacific Ocean.

As we have observed in the previous session, we usually obtain high agreement in the initial phase (first year), which is also present in *baseline1* and *oras4*. *Gecco2* contains two clusters in this phase: gray and purple. *Baseline1* features the gray cluster, *oras4* the green one. Looking at the respective cluster means, we see that gray and green are very similar. Green features a more extended warming than gray and warmer temperatures at the north pole. The *baseline1* simulation is more extreme and its gray cluster is located between the gray and green one for the other two simulations. *gecco2* contains several simulation runs that predict an El Niño. Looking at the ground truth data, we see that 2015/16 featured a very strong El Niño event which becomes visible as strong warming in the eastern tropical Pacific Ocean. This feature is most prominent in the third image (date: April 2015) and is well covered by *gecco2* and *oras4*, which feature strong purple, orange and pink clusters standing for the respective climatic pattern.

Ground truth data ends at this time phase and the remainder of the simulations predict the future. Here we see different outcomes. *Baseline1* features strong blue and green clusters (general warming + cool Pacific) starting from 2018 with several temporally limited increases of the orange cluster (warming + moderate El Niño). The gray cluster is no longer present. *gecco2* and *oras4* feature a wide spread of possible futures. In the beginning *gecco2* contains the warmer clusters and *oras4* the more moderate ones.

7 Discussion

We have found that our combination of visualizations works well for the specified tasks. The clustering timeline effectively conveys the temporal development of the ensemble data (begin, continuation, end, division or merge of clusters of similar ensemble members). Other clustering techniques or distance measures to plug into the system remain to be studied in detail. Even with the already convincing qualitative results, there may be computationally more efficient algorithms than the multi-run k-means and the l_1 metric that considers each pair of scalars of two fields to be compared.

In this paper, we particularly focus on climate simulation ensembles, but in its core, the system can deal with any time-dependent (high-dimensional) point cloud as long as the number of points remains constant over all time steps and a meaningful distance between two points can be defined.

7.1 User Feedback

Our project had a clear focus on a good general usability and a maximal utility for our collaboration partners, which has been manifested in regular meetings where we got feedback concerning the current state of the software and discussed the next necessary steps. Hence, we concentrated on *user experience* (UE) as evaluation scenario [50], [51]. The following exemplifies a few of the features of the final implementation that were explicitly asked for based on a prototype of the software.

From early on climatologists stated that they need the permanent possibility to complement the clustering timeline with filled contour maps. They desired means to compare the normalization of a scalar field's error compared to the ground truth so that it reflects an actual temperature difference. Another point of discussion concerned the most helpful coloring of the clustering timeline. First shown a coloring for the cluster nodes which reflected the clusters' silhouette coefficient, the users pointed out that a coloring scheme to distinguish the global clusters easily in an overview view (many time steps visible at once) would be a better aid in the analysis of the long-term behavior of an ensemble. In its early stages, the focus of the application had been a time step by time step analysis approach using a close-up view of the clustering timeline where the text labels were sufficient.

Since we were concerned about the climatologists' wishes during the development, we can ultimately report generally positive user feedback. They were eager to see the clustering results and were very happy with the way they are communicated by the clustering timeline. The filled contour maps, which were deemed indispensable, worked really well for them in the interactive setting. The distance matrices and silhouette plots proved to be the ideal tool to increase the confidence in the clustering results, and to gain a deeper knowledge about how they come to be.

When a software becomes more and more production-ready with a growing set of features, the system configuration and interface can develop into something that is no longer easy and intuitive. Eventually the following graphical user interface was implemented for the application, to everyone's satisfaction. There is a control bar that offers two main entrance points for users: A *Files* menu to load and save data, and a *Tools* menu to work with the data (compute a clustering, create certain visualization, etc.). The actions of these tools can be undone or redone with different parameters. These named parameters can be viewed and adjusted in a dedicated area directly below the control bar. The overall design allows for a relatively smooth workflow (as proposed in this paper) without too much clicking and navigating through menus while we maintained the possibility to extend the software, e.g. with new tools, without major changes.

7.2 Lessons Learned

The presented interface has been developed over the course of 1.5 years with regular feedback rounds. In the following we state lessons-learned from this interaction.

Incorporate established visualization techniques. Expert users, especially in the natural sciences, usually have a strong background in mathematics and data analytics. Most of them have used visualization in their data analyses and communication process throughout their career. Commonly, they employ static ”standard” visualizations and animations such as scalar field visualizations, scatter plots, and line charts which were derived from the raw data through a mathematical procedure. Including (a subset of) these techniques is important to give expert users a convenient start and build trust in the proposed technique.

In our case, we integrated filled contour maps of time-dependent 2D scalar field ensembles in an interactive small multiples approach. This allows the user to check the raw data and understand the new system. We also integrated summary statistics like the error chart (see e.g. Figure 7) to provide an already familiar type of visualization that links our newly proposed clustering timeline to well known statistical summaries.

Allow for export of intermediate results and respective import. We have implemented saving of the filled contour maps, the heatmap and silhouette plot, the abstract clustering results and the clustering timeline. This has given us the following benefits:

- Users can stop halfway through an analysis session and resume it later without having to reproduce the input (data and parameters).
- Less waiting for potentially lengthy computations like the clustering because it can be faster to load results from a file than recompute them each time.
- These intermediate results can easily be shared and for example viewed with other software.
- Data from alternative sources can easily be used instead of the “in-house” results (e.g. other scalar field visualization, other clustering results). (In this sense our web application is really a plug and play toolbox.)

Use visual links in dashboards. In the early stages of our implementation we ensured the detection of matching elements in the dashboard through click interaction. For example, if the user wanted to see the cluster mean for a certain cluster they had to click in the cluster in the clustering timeline and vice versa. The respective complementary element in the other visualization would then have been highlighted. This proved to be rather tedious and we opted for permanent visual links of the elements. In Figure 7, for example, we use the same color in the clustering timeline, the boundary of its respective means in the contour plots, and the lines in the error chart. This proved to be much more convenient than the interactive version.

Use automatic dynamic scaling for different chart types. In the current tool the user can interactively fill the dashboard with the charts they are interested in. In the early implementation we used fixed standard sizes for each chart type that allowed for good visual inspection. Working with varying datasets we realized that this can often lead to extremely large dashboards that require a lot of scrolling when using for example large number of ensembles or very long time-series. In the current versions, we opt for a convenient fit of the charts in the given screen space and allow the user to adjust sizes as necessary.

8 Conclusion and Future Work

In this paper we develop and link a set of visualizations that supports the in-depth analysis of climate simulation ensembles. This development is the result of a close collaboration with domain experts. We have gathered questions that the ensemble datasets pose, and have derived designs for visualizations which address these questions and help solving them. We compute a series of clusterings for the time-dependent 2D scalar fields and visualize it in our so-called clustering timeline, a specialized version of the Sankey diagram. The time-step-wise clustering result itself can be validated with further graphics that we provide in the software. This includes a heatmap showing the pairwise distance of the ensemble members, a bar chart showing the silhouette coefficient for the clusters and filled contour maps of the scalar fields.

We have implemented the software as a web application that the climatologists could easily try out at various stages of its development. In a regular feedback loop we fine-tuned our goals and the respective implementation, so that in the end, the domain experts were very content with the software. They are able to execute analysis sessions, finding the software flexible enough to always adapt to their current need of information. With the help of our clustering approach they can get insights into the data which were not attainable before.

Having a well-working and approved visualization for the clustering series, one research direction we aim at for future work is a further specialization of the clustering approach. For example, we want to explore more sophisticated distance functions, maybe based on a prior feature extraction.

We also see room for improvement concerning the scalability of the clustering timeline; not so much for larger ensembles – up to 100 members should not be a problem. After that maybe just thinner lines might be a solution, even though a completely different drawing of the “flow” might become desirable then. More challenging will be the question of how a larger number of time steps could be handled, as may arise because of higher temporal resolutions or predictions for the further future. Then our visualization, showing every single time step, is probably not the best solution to convey the “big picture”. A hierarchy of timelines with ever increasing temporal summarization (as in [19]) may be a remedy.

Another research direction could be the question if and how it may be possible to tweak the visualization of the “raw data”, like our filled contour maps, so that a manual comparison becomes less tedious and the computed clustering can be comprehended or validated more easily.

Acknowledgments

The authors would like to thank Holger Pohlmann of the Max-Planck-Institute for Meteorology for providing the data from the German Ministry of Education and Research project MiKlip (www.fona-miklip.de) and frequent discussion and suggestions concerning the visualization interface and the results presented in this paper. NCEP/NCAR reanalyses data were received from www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html.

REFERENCES

[1] C. Kadow, S. Illing, O. Kunst, H. W. Rust, H. Pohlmann, W. A. Müller, and U. Cubasch, "Evaluation of forecasts by accuracy and spread in the miklip decadal climate prediction system," *Meteorologische zeitschrift*, vol. 25, no. 6, pp. 631–643, Dec. 2016.

[2] J. Marotzke, W. A. Müller, F. S. E. Vamborg, P. Becker, U. Cubasch, H. Feldmann, F. Kaspar, C. Kottmeier, C. Marini, I. Polkova, K. Prömmel, H. W. Rust, D. Stammer, U. Ulbrich, C. Kadow, A. Köhl, J. Kröger, T. Kruschke, J. G. Pinto, H. Pohlmann, M. Reyers, M. Schröder, F. Sienz, C. Timmreck, and M. Ziese, "Miklip: A national research project on decadal climate prediction," *Bulletin of the american meteorological society*, vol. 97, no. 12, pp. 2379–2394, 2016.

[3] P. Riehmann, M. Hanfler, and B. Froehlich, "Interactive sankey diagrams," in *IEEE Symposium on Information Visualization, 2005*, IEEE, 2005, pp. 233–240.

[4] M. Schmidt, "The sankey diagram in energy and material flow management," *Journal of industrial ecology*, vol. 12, no. 1, pp. 82–94, 2008, ISSN: 1530-9290.

[5] C. Kadow, S. Illing, I. Kröner, U. Ulbrich, and U. Cubasch, "Decadal climate predictions improved by ocean ensemble dispersion filtering," *Journal of advances in modeling earth systems*, 2017, ISSN: 1942-2466.

[6] fona-miklip. (2016). Decadal forecast for 2017–2026, [Online]. Available: www.fona-miklip.de/decadal-forecast-2017-2026/decadal-forecast-for-2017-2026 (visited on 07/12/2017).

[7] DWD, MPI-M, and UHH. (2017). Seasonal forecasts, [Online]. Available: www.dwd.de/EN/ourservices/seasonals_forecasts/time_series.html (visited on 07/12/2017).

[8] M. Böttinger, H. Pohlmann, N. Röber, K. Meier-Fleischer, and D. Spickermann, "Visualization of 2D uncertainty in decadal climate predictions," in *Workshop on Visualization in Environmental Sciences (EnvirVis 2015)*, 2015, pp. 1–5.

[9] R. Liu, H. Guo, J. Zhang, and X. Yuan, "Comparative visualization of vector field ensembles based on longest common subsequence," in *2016 ieee pacific visualization symposium (pacificvis)*, IEEE, 2016, pp. 96–103.

[10] Q. Shu, H. Guo, J. Liang, L. Che, J. Liu, and X. Yuan, "Ensemblegraph: Interactive visual analysis of spatiotemporal behaviors in ensemble simulation data," in *2016 ieee pacific visualization symposium (pacificvis)*, IEEE, 2016, pp. 56–63.

[11] K. Bensema, L. Gosink, H. Obermaier, and K. I. Joy, "Modality-driven classification and visualization of ensemble variance," *Ieee trans. on visualization and computer graphics*, vol. 22, no. 10, pp. 2289–2299, 2016.

[12] I. Demir, C. Dick, and R. Westermann, "Multi-charts for comparative 3d ensemble visualization," *Ieee transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 2694–2703, 2014.

[13] T. Pfaffelmoser and R. Westermann, "Visualizing contour distributions in 2d ensemble data," *Eurovis-short papers*, pp. 55–59, 2013.

[14] A. Diehl, L. Pelorosso, C. Delrieux, K. Matković, J. Ruiz, M. E. Gröller, and S. Bruckner, "Albero: A visual analytics approach for probabilistic weather forecasting," in *Computer graphics forum*, Wiley Online Library, vol. 36, 2017, pp. 135–144.

[15] P. Kothur, M. Sips, H. Dobslaw, and D. Dransch, "Visual analytics for comparison of ocean model output with reference data: Detecting and analyzing geophysical processes using clustering ensembles," *Ieee transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1893–1902, 2014.

[16] U. D. Bordoloi, D. L. Kao, and H.-W. Shen, "Visualization techniques for spatial probability density function data," *Data science journal*, vol. 3, pp. 153–162, 2004.

[17] C. D. Correa, Y.-H. Chan, and K.-L. Ma, "A framework for uncertainty-aware visual analytics," in *Ieee symposium on visual analytics science and technology, 2009*, IEEE, 2009, pp. 51–58.

[18] S. Bruckner and T. Möller, "Result-driven exploration of simulation parameter spaces for visual effects design," *Ieee transactions on visualization and computer graphics*, vol. 16, no. 6, pp. 1468–1476, 2010.

[19] H. Obermaier, K. Bensema, and K. I. Joy, "Visual trends analysis in time-varying ensembles," *Ieee trans. on visualization and computer graphics*, vol. 22, no. 10, pp. 2331–2342, 2016.

[20] F. Ferstl, M. Kanzler, M. Rautenhaus, and R. Westermann, "Time-hierarchical clustering and visualization of weather forecast ensembles," *Ieee transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 831–840, 2017.

[21] M. Jarema, J. Kehrer, and R. Westermann, "Comparative visual analysis of transport variability in flow ensembles," *Wscg*, vol. 24, no. 1, pp. 25–34, 2016.

[22] F. Bendix, R. Kosara, and H. Hauser, "Parallel Sets: Visual Analysis of Categorical Data," in *IEEE Symposium on Information Visualization*, 2005, pp. 133–140.

[23] M. Jarema, I. Demir, J. Kehrer, and R. Westermann, "Comparative visual analysis of vector field ensembles," in *Ieee conference on visual analytics science and technology, 2015*, IEEE, 2015, pp. 81–88.

[24] J. Wang, X. Liu, H.-W. Shen, and G. Lin, "Multi-resolution climate ensemble parameter analysis with nested parallel coordinates plots," *Ieee transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 81–90, 2017.

[25] S. Havre, E. Hetzler, P. Whitney, and L. Nowell, "Themeriver: Visualizing thematic changes in large document collections," *Ieee transactions on visualization and computer graphics*, vol. 8, no. 1, pp. 9–20, 2002.

[26] L. Byron and M. Wattenberg, "Stacked graphs–geometry & aesthetics," *Ieee transactions on visualization and computer graphics*, vol. 14, no. 6, 2008.

[27] M. Dork, D. Gruen, C. Williamson, and S. Carpendale, "A visual backchannel for large-scale events," *Ieee transactions on visualization and computer graphics*, vol. 16, no. 6, pp. 1129–1138, 2010.

[28] W. Cui, S. Liu, L. Tan, C. Shi, Y. Song, Z. Gao, H. Qu, and X. Tong, "Textflow: Towards better understanding of evolving topics in text," *Ieee trans. on vis. and computer graphics*, vol. 17, no. 12, pp. 2412–2421, 2011.

[29] S. Gad, W. Javed, S. Ghani, N. Elmqvist, T. Ewing, K. N. Hampton, and N. Ramakrishnan, "Themedelta:

Dynamic segmentations over temporal topic models," *Ieee trans. on visualization and computer graphics*, vol. 21, no. 5, pp. 672–685, 2015.

[30] J. Waser, R. Fuchs, H. Ribicic, B. Schindler, G. Bloschl, and E. Groller, "World lines," *Ieee transactions on visualization and computer graphics*, vol. 16, no. 6, pp. 1458–1467, 2010.

[31] T. Munzner, "A nested model for visualization design and validation," *Ieee transactions on visualization and computer graphics*, vol. 15, no. 6, 2009.

[32] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Proceedings of second international conference on knowledge discovery and data mining*, vol. 96, 1996, pp. 226–231.

[33] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[34] F. Corpet, "Multiple sequence alignment with hierarchical clustering," *Nucleic acids research*, vol. 16, no. 22, pp. 10 881–10 890, 1988.

[35] J. F. Navarro, C. S. Frenk, and S. D. White, "A universal density profile from hierarchical clustering," *The astrophysical journal*, vol. 490, no. 2, p. 493, 1997.

[36] S. Lloyd, "Least squares quantization in pcm," *Ieee trans. on inform. theory*, vol. 28, no. 2, pp. 129–137, 1982.

[37] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of comp. and appl. math.*, vol. 20, pp. 53–65, 1987.

[38] M. Rosvall and C. T. Bergstrom, "Mapping change in large networks," *Plos one*, vol. 5, no. 1, e8694, 2010.

[39] P. Tol, "Colour schemes," *Sron technical note*, no. 2.2, SRON–EPS, 2012, See also https://personal.sron.nl/ pault/.

[40] C. Glasbey, G. van der Heijden, V. F. Toh, and A. Gray, "Colour displays for categorical images," *Color research & application*, vol. 32, no. 4, pp. 304–309, 2007.

[41] M. R. Garey and D. S. Johnson, "Crossing number is NP-complete," *Siam journal on algebraic discrete methods*, vol. 4, no. 3, pp. 312–316, 1983.

[42] K. Sugiyama, S. Tagawa, and M. Toda, "Methods for visual understanding of hierarchical system structures," *Ieee transactions on systems, man, and cybernetics*, vol. 11, no. 2, pp. 109–125, 1981.

[43] L. Padilla, P. S. Quinan, M. Meyer, and S. H. Creem-Regehr, "Evaluating the impact of binning 2d scalar fields," *Ieee transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 431–440, 2017.

[44] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, *et al.*, "The ncep/ncar 40-year reanalysis project," *Bulletin of the american meteorological society*, vol. 77, no. 3, pp. 437–471, 1996.

[45] A. Köhl, "Evaluation of the gecco2 ocean synthesis: Transports of volume, heat and freshwater in the atlantic," *Quarterly journal of the royal meteorological society*, vol. 141, no. 686, pp. 166–181, 2015.

[46] M. A. Balmaseda, K. Mogensen, and A. T. Weaver, "Evaluation of the ecmwf ocean reanalysis system oras4," *Quarterly journal of the royal meteorological society*, vol. 139, no. 674, pp. 1132–1161, 2013.

[47] S. M. Uppala, P. Kållberg, A. Simmons, U. Andrae, V. Bechtold, M. Fiorino, J. Gibson, J. Haseler, A. Hernandez, G. Kelly, *et al.*, "The era-40 re-analysis," *Quarterly journal of the royal meteorological society*, vol. 131, no. 612, pp. 2961–3012, 2005.

[48] D. Dee, S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, P. Bauer, *et al.*, "The era-interim reanalysis: Configuration and performance of the data assimilation system," *Quarterly journal of the royal meteorological society*, vol. 137, no. 656, pp. 553–597, 2011.

[49] C. Wang and J. Picaut, "Understanding ENSO physics — A review," *Earth's climate*, Geophysical Monograph, vol. 147, C. Wang, S.-P. Xie, and J. A. Carton, Eds., pp. 21–48, 2004.

[50] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical studies in information visualization: Seven scenarios," *Ieee transactions on visualization and computer graphics*, vol. 18, no. 9, pp. 1520–1536, Sep. 2012, ISSN: 1077-2626.

[51] T. Isenberg, P. Isenberg, J. Chen, M. Sedlmair, and T. Möller, "A systematic review on the practice of evaluating visualization," *Ieee transactions on visualization and computer graphics*, vol. 19, no. 12, pp. 2818–2827, Dec. 2013, ISSN: 1077-2626.

Christopher P. Kappe received his B.Sc. and M.Sc. in Applied Computer Science from the University of Heidelberg, Germany, in 2012 and 2015 respectively. He has specialized in visualization, always engaged in and motivated by interdisciplinary collaborations where cutting edge scientific research asked for new visualization solutions. He is currently a PhD student at TU Kaiserslautern and his research interests include visual analytics with a focus on clustering of multivariate data.

Michael Böttinger received his Diploma (equivalent to M.Sc.) in Geophysics from the University of Hamburg, Germany, in 1988. He started as a scientist in the field of climate modeling at the Max Planck Institute for Meteorology. In 1990 he joined the German Climate Computing Center (DKRZ), where he leads the visualization and public relations group. His research is application oriented and focuses on scientific visualization of climate model data.

Heike Leitte received the M.Sc. degree (Diplom) in computer science in 2006 and the PhD degree in 2009 from the University of Leipzig. In 2009/10, she worked as a postdoctoral researcher at Swansea University. From 2010 until 2015, she was assistant professor at Heidelberg University, and became full professor for visual information analysis at TU Kaiserslautern in 2015. Her research interests include visualization of unsteady, multivariate, and high-dimensional data with applications in biology, engineering, and climate research.