

# Multi-Resolution Climate Ensemble Parameter Analysis with Nested Parallel Coordinates Plots

Junpeng Wang, Xiaotong Liu, Han-Wei Shen, *Member, IEEE*, and Guang Lin

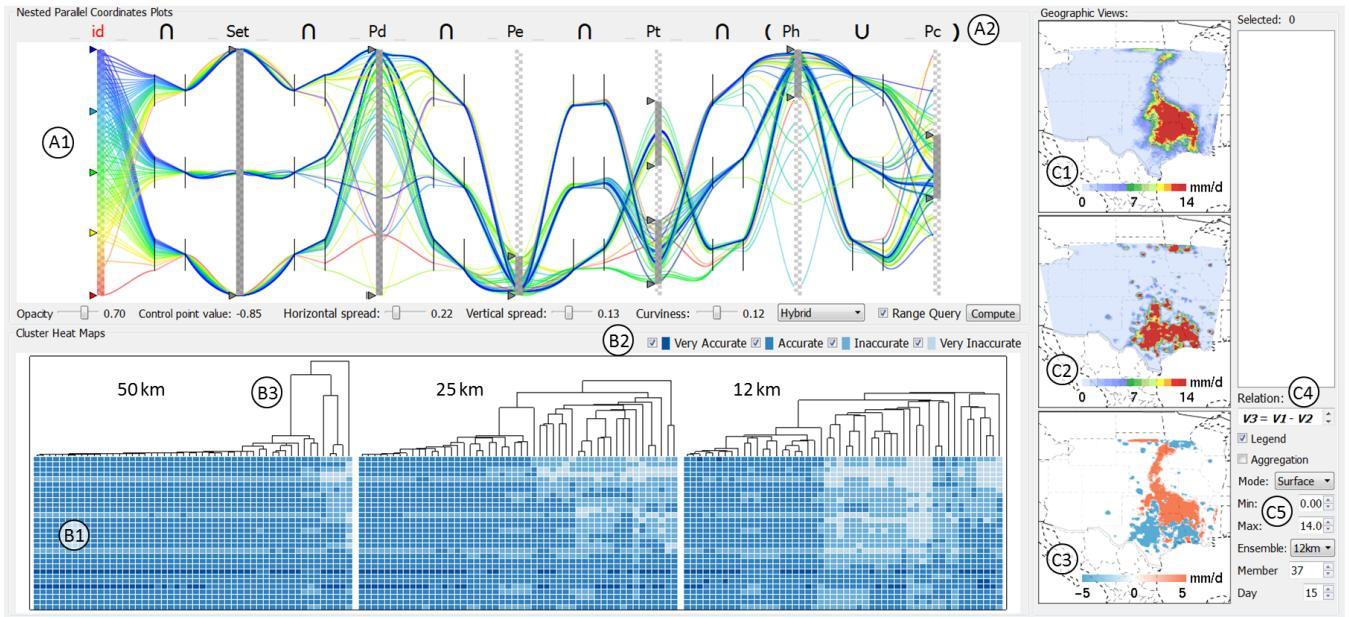


Fig. 1. The proposed visual analytics system: (A1) the nested parallel coordinates plot (NPCP) for analyzing multi-dimensional parameter correlations across resolutions; (A2) high-dimensional range query in NPCP with a set expression; (B1, B2) heat maps for ensemble quality overview and color legends for different quality levels; (B3) dendograms for grouping similar ensemble members; (C1, C2, C3) geographic views for one ensemble item, one observation item and the difference between them; (C4) equation for calculating the difference between an ensemble item and an observation item; (C5) control widgets for ensemble exploration.

**Abstract**— Due to the uncertain nature of weather prediction, climate simulations are usually performed multiple times with different spatial resolutions. The outputs of simulations are multi-resolution spatial temporal ensembles. Each simulation run uses a unique set of values for multiple convective parameters. Distinct parameter settings from different simulation runs in different resolutions constitute a multi-resolution high-dimensional parameter space. Understanding the correlation between the different convective parameters, and establishing a connection between the parameter settings and the ensemble outputs are crucial to domain scientists. The multi-resolution high-dimensional parameter space, however, presents a unique challenge to the existing correlation visualization techniques. We present Nested Parallel Coordinates Plot (NPCP), a new type of parallel coordinates plots that enables visualization of intra-resolution and inter-resolution parameter correlations. With flexible user control, NPCP integrates superimposition, juxtaposition and explicit encodings in a single view for comparative data visualization and analysis. We develop an integrated visual analytics system to help domain scientists understand the connection between multi-resolution convective parameters and the large spatial temporal ensembles. Our system presents intricate climate ensembles with a comprehensive overview and on-demand geographic details. We demonstrate NPCP, along with the climate ensemble visualization system, based on real-world use-cases from our collaborators in computational and predictive science.

**Index Terms**—Parallel coordinates plots, parameter analysis, multi-resolution climate ensembles

## 1 INTRODUCTION

Weather predictions and climate projections are usually conducted with numerical simulations taking into account of geophysical fluid

dynamics and comprehensive physical processes, most notably atmospheric convection [20, 46]. Given a certain state of the atmosphere and the value of atmospheric movement parameters (convective parameters), the simulations execute iteratively based on the known information to derive the state of the next time step. Due to the intricate and uncertain effects of convective parameters on weather prediction models, accuracy is hard to achieve with just a single run of the simulations [46, 47]. To improve the accuracy, different parameter adjusting strategies (parameterization schemes) have been introduced based on the conceptual or empirical relationships between the input convective parameters and the simulation outputs. These parameterization schemes sample the convective parameters space and adjust parameter settings using certain heuristic to strive for more accurate simulations.

• Junpeng Wang, Xiaotong Liu and Han-Wei Shen are with The Ohio State University. E-mail: {wang.7665, liu.1952, shen.94}@osu.edu.

• Guang Lin is with Purdue University. E-mail: guanglin@purdue.edu.

Manuscript received 31 Mar. 2016; accepted 1 Aug. 2016. Date of publication 15 Aug. 2016; date of current version 23 Oct. 2016.

For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TVCG.2016.2598830

The outputs resulted from simulation runs with distinct parameter settings constitute an ensemble. Different spatial resolutions of a simulation capture different scales of the physical features that domain scientists intend to simulate. Consequently, multi-resolution spatial temporal climate ensembles are common in climate modeling.

The knowledge of how different convective parameters interact among themselves, i.e. pairwise correlations, when using a certain simulation resolution and whether the correlations stay the same across different resolutions are crucial for domain scientists to understand the behavior of convective parameters. In addition, understanding how uncertainty in the output of a climate model is attributed to different sources of uncertainty in the model input is significantly important to precisely portray the relationship between the convective parameters and the climate model. Specifically, domain scientists are interested to know: (1) What is the correlation between different parameters and is the correlation transferable across resolutions? (2) What is the connection between the convective parameters and the spatial temporal ensembles? These questions are still open for researchers who study multi-resolution ensemble parameters. In the context of visualization, multi-dimensional visualization techniques such as parallel coordinates plots [16] or scatterplot matrices [7] are not readily applicable to understand the correlations of parameters used for different resolutions. Although parameter sensitivity analyses have been conducted to reveal the scale to which one parameter is affecting the accuracy of a simulation, the sensitivity analysis models [11, 17] that measure parameter behaviors with monotonous numerical values are not intuitive enough to demonstrate the intricate relationship between the multi-resolution parameters and the large spatial temporal ensembles. In addition, domain scientists are more interested in visually quantifying the difference resulted from parameter changing (visual parameter space analysis [34, 40]). Therefore, a visual analytics system that involves domain scientists in the loop of exploration and provides both parameter and ensemble visualization is urgently needed.

In this paper, we present an integrated visualization system to help scientists explore and analyze convective parameters and spatial temporal ensembles, both in multiple resolutions. To tackle the challenge of visualizing multi-resolution high-dimensional parameter space, we propose *Nested Parallel Coordinates Plot (NPCP)*, which combines the superimposition and juxtaposition design to demonstrate both the intra-resolution and inter-resolution correlations of different parameters. It also supports multi-resolution high-dimensional range queries with various set operations (such as union, intersection, and difference) for flexible visual filtering. To effectively analyze the multi-resolution spatial temporal climate ensembles, we apply the *overview+detail* exploration technique [42] to our visualization. Heat maps and dendograms are linked together to help domain scientists gain an overall understanding of large multi-resolution ensembles. Furthermore, multiple side-by-side geographic views are provided to show spatial and temporal details on demand. They allow scientists to intuitively compare ensemble runs with the observed ground truth. In the system, parameter visualization and ensemble visualization are closely linked together to enhance the scientists' ability of visual reasoning. In summary, our contributions in this paper are twofold:

1. We propose a novel design to augment Parallel Coordinates Plots (PCPs). The design enhances the capability of PCPs in demonstrating the relationship of numerous parameters from multiple resolutions in hybrid comparative views, integrating superimposition, juxtaposition and explicit encodings. It also leverages the power of logic set operations in visual filtering.
2. We work closely with climate scientists to study multi-resolution convective parameters and provide an integrated visual analytics system that visualizes both multi-resolution high-dimensional parameter space and large spatial temporal ensembles.

## 2 RELATED WORK

**Parallel Coordinates Plots for High-Dimensional Data Visualization.** One of the most popular and effective high-dimensional correlation visualization approaches is the Parallel Coordinates Plot

(PCP) [18]. When the number of data instances is large, PCP tends to get cluttered because of the massive overplotting. Over the last decade, much research has been conducted to reduce visual clutter, such as ordering the axes of PCPs [2, 31] and the use of frequency and density plots [3]. The recent survey work [16] by Heinrich and Weiskopf has thoroughly discussed these details. In addition, bundling polylines has been considered to be effective in conveying cluster information [8, 25, 48] in PCPs. Palmas et al. [30] bundled polylines with polygonal strips for a faster overview of the cluster information and data trend in high-dimensional data sets. Heinrich et al. [14] replaced the line segment between axes with a pair of cubic Bézier curves, and curves corresponding to the data instances of the same cluster are bundled together for better separation. Parallel coordinates matrices [6] demonstrate correlation patterns between axes of PCPs in multiple small juxtaposed views organized based on certain screen-space metrics. The continuity of polylines is no longer maintained in these juxtaposed views. Effective preservation of users' mental map [26] when shifting from one view to another is a crucial issue. Another limitation of the matrix based representations is the redundancy in the symmetric matrix layout. Heinrich et al. [15] tried to eliminate the redundancy and maintain the polyline continuity using a graph-theoretic approach. Their approach effectively utilizes the space and provides a quick correlation overview, though it takes more time to identify the correlation between two specific dimensions. We try to maintain the continuity of polylines in our new PCP design in this paper. In our domain application, quickly identifying the correlation between two specific dimensions is more important than listing all possible correlation patterns.

**Comparative Visualization.** Visual analysis of multi-resolution data involves visual comparison of data across resolutions. Gleicher et al. [9] divided the design space of comparative visualization into superimposition, juxtaposition and explicit encoding, while Javed and Elmquist [19] detailed explicit encoding by overloading and nesting. Juxtaposition is the most popular visualization design for side-by-side comparison, without overplotting or occlusion that may occur in superimposition [9]. Conventional side-by-side visual comparison can be further enhanced by alternative juxtaposition designs [21, 22, 23] as well as interactions [44]. In this work, we propose a hybrid comparative visualization of parallel coordinates plots by combining the three comparative designs (i.e. superimposition, juxtaposition and explicit encoding), and provide smooth view transition between alternative designs through simple interactions.

**Ensemble Visualization.** An ensemble is a collection of outputs from a sequence of simulation runs. Each simulation run (ensemble member) uses slightly different input parameter settings. This type of data is common in multiple domains, such as climate modeling [36, 39] and high-energy physics [12, 32]. Domain scientists study ensemble data to analyze uncertainties of their simulations [5, 35, 36, 37]. The common features of ensemble data are large, multi-dimensional and usually containing both spatial and temporal information. Due to these complex features, ensemble visualization is a non-trivial task [29]. Numerous visualization approaches have been introduced to demonstrate multi-dimensional spatial temporal ensembles. Glyph-based visualizations [4, 12, 39] encode different facets of data into distinct visual channels. Small multiples and coordinated views show the same facet of different ensemble members in different displays, thus demonstrating effective side-by-side comparisons [33, 37]. Different machine learning and statistical methods have also been adopted for visual analytics on multi-dimensional data [1, 10]. Existing climate ensemble visualization includes SimEnvVis [28], Ensemble-Vis [37] and Noodles [39]. SimEnvVis applies comparative visualizations to demonstrate the multi-dimensional and multi-variate facets of the climate ensemble. Ensemble-Vis is facilitated with numerous statistical visualization tools. It demonstrates uncertainties in climate ensemble with multiple linked displays. Noodles specializes in providing quantitatively evaluations of geospatial uncertainties in meteorological ensemble with glyph-based techniques. Our focus, in this paper, is on exposing the input parameter correlations of climate models running with different resolutions, and establishing the connection between the convective parameter settings and the ensemble outputs.

### 3 MOTIVATION, BACKGROUND AND APPROACH OVERVIEW

In this section, we explain the motivation and background of climate ensemble visualization. Following the specific requirements from our domain science collaborator, we provide an overview of our approach.

#### 3.1 Motivation

In meteorology, *convection* is responsible for the formation of rainfall. It is the transfer of heat from a warmer region to a cooler one by moving warm liquid or gas from the heated area to the unheated area. Weather predictions are conducted through climate simulation models that simulate the atmospheric motions. Given a certain state of the atmosphere, these models discretize the spatial region of interest into small grid cells and compute iteratively on these cells to derive the next state of the atmosphere. The *simulation resolution* indicates the size of the grid cells. In these models, *convective parameters* have been formalized to control the atmospheric convective motions, such as the starting height, downdraft mass and kinetic energy etc. *Convective parameterization* is the process of tuning these parameters. Different parameter adjusting strategies, i.e. *convection parameterization schemes*, have been proposed to accelerate the convergence of the simulations while maintaining the accuracy.

Given the fact that there are many sources of uncertainties in weather predictions, it is necessary to quantify them through multiple simulation runs with different parameter settings. The output of these simulations is a *climate ensemble*. Averaging the result of different ensemble runs can mitigate the effects of outliers and thus improving the prediction accuracy. More sophisticated studies and analyses on the ensemble are usually performed to derive more accurate predictions. Climate models with different simulation resolutions can capture features of different scales that scientists want to simulate. The performance of climate models is sensitive to both physical parameterization and the model resolution. It is not guaranteed that high resolution models always perform better than low resolution models. In addition, models with different simulation resolutions have different computational costs. Hence it is critical to develop an effective data analytics tool to analyze multi-resolution simulation models and identify their advantages and limitations.

#### 3.2 Domain Data Characterization

**Parameter Set.** A *parameter setting* is a combination of values used for different convective parameters in one execution/run of the climate model. Different runs of the simulation model use different parameter settings. A collection of all parameter settings used in an ensemble is a *parameter set*. Considering each parameter as a dimension, multiple parameters with different settings form a high-dimensional space. Multiple parameter sets will be used to generate multi-resolution ensembles.

**Ensemble Set.** The output from one execution of the climate model usually contains prediction results over a period of time. For example, one execution of the simulation can output the precipitation of the future 10 days. For each day, the output contains the precipitation values at different 2D locations (latitude/longitude coordinates). We call the predicted precipitation of a single day as an *ensemble item*, the entire output from one execution of the simulation as an *ensemble member*. All members of one ensemble usually contain the same number of ensemble items. In this example, each ensemble member contains 10 ensemble items. Multiple ensemble members constitute one *ensemble set*. Therefore, a climate ensemble set is a collection of the spatial temporal ensemble members. Ensemble sets with different simulation resolutions form multi-resolution climate ensembles.

**Observation Set.** To study convective parameters and evaluate climate ensembles, satellites are often used to observe and record real-world weather conditions. One *observation set* is the observed weather for a specific region over a period of time. The corresponding observation set of our previous example will be the precipitation of the simulated area over the 10 days. Each day is an *observation item*. This observation set can be considered as the ground truth to calibrate the convective parameter settings. It can also be used to measure the qual-

ity of the ensembles. Based on different settings during the collections, the observation set can also have different resolutions.

#### 3.3 Task Analysis

The main objectives of the domain scientist (from computational and predictive science) that we collaborate with are to: (i) identify the correlation between different parameters and compare the correlation of convective parameters across simulation resolutions; (ii) establish connections between convective parameter settings and the accuracy of climate simulations. Over three months closely working with our collaborator and meeting biweekly to discuss the requirements, we refine the two general objectives into the following detailed requirements.

**R1: Understanding Intra-Set Parameter Correlation.** The intra-set parameter analysis targets at exposing the correlations between different parameters over ensemble members of the same simulation resolution. For example, what is the correlation between two specific convective parameters when the output ensemble members are accurate? How did the correlation change over the ensemble members (correlation in early/later ensemble members) of the same resolution? The answers to these questions will help the domain scientist gain better understanding of the convection parameterization scheme.

**R2: Comparing Inter-Set Parameter Correlations.** Dissimilar behaviors of the convective parameters have been observed in different simulation resolutions. This requirement aims to find out how the convective parameters are different when accurate simulations were achieved in different simulation resolutions and how significant the difference is.

**R3: High-Dimensional Parameter Range Query.** The domain scientist has some empirical expectations on the optimal range of certain convective parameters. For example, probably the accurate simulation results are more frequently observed when the coefficient related to downdraft mass flux rate (one convective parameter, denoted as  $P_d$ ) is between [0.9, 1.0] and the maximum turbulent kinetic energy (another convective parameter, denoted as  $P_t$ ) is in the range of [4.0, 6.0] or [8.0, 10.0]. What are the values of other parameters when both  $P_d$  and  $P_t$  are within their empirical optimal ranges, i.e.  $\{0.9 \leq P_d \leq 1.0\} \cap \{4.0 \leq P_t \leq 6.0\} \cup \{8.0 \leq P_t \leq 10.0\}$ , or at least one of them is in the ranges, i.e.  $\{0.9 \leq P_d \leq 1.0\} \cup \{4.0 \leq P_t \leq 6.0\} \cup \{8.0 \leq P_t \leq 10.0\}$ ? This requires query on the high-dimensional convective parameter space.

**R4: Ensemble Member/Item Quality Evaluation.** Given an output from the climate simulation, the domain scientist is eager to know its accuracy and how it is different from the observed result (collected from satellites). An overall quality distribution of ensemble members/items will be helpful. In addition, it is crucial for the domain scientist to examine the details of individual ensemble member/item.

**R5: Ensemble Member Comparison.** Finding the distribution of similar ensemble members is also important. Comparing the parameter settings of similar ensemble members is one of the most effective ways to study parameter behaviors. Therefore, ensemble member comparison is a critical task.

**R6: Demonstrating Both Spatial and Temporal Facets of Ensembles.** The spatial information in climate ensembles is significantly important for domain scientists. For example, if scientists focus on the precipitation prediction, they not only need to know the amount of precipitation, but also need to know the spatial distribution of the precipitation. Temporal evolution of precipitation is another important component when studying climate ensembles.

#### 3.4 Approach Overview

Based on the requirements from the domain scientist, we formalize two major visualization tasks: parameter visualization (R1, R2 and R3) and ensemble visualization (R4, R5 and R6). To visualize the multi-resolution high-dimensional parameter sets, we propose a novel technique, called *Nested Parallel Coordinates Plot (NPCP)*; whereas an *overview+detail* exploration strategy has been employed for understanding the large spatial temporal ensemble sets.

Figure 2 shows the workflow of our visualization system. The quality of ensemble sets is evaluated through comparing with the corresponding observation sets (R4). The quality values along with the

parameter sets are visualized in the NPCP to demonstrate the intra-resolution (R1) and inter-resolution (R2) correlations of convective parameters. Intuitive high-dimensional range query (R3) and smooth view transition can be performed via simple interactions with the NPCP. The quality values of ensemble members are also used to compare them (R5). A linked visualization of heat maps and dendograms is employed to demonstrate the overall quality and similarity of ensemble members. The interaction with NPCP, heat maps and dendograms leads scientists to examine on the spatial temporal details of an ensemble member/item (R6). Multiple juxtaposed geographic views serve this requirement.

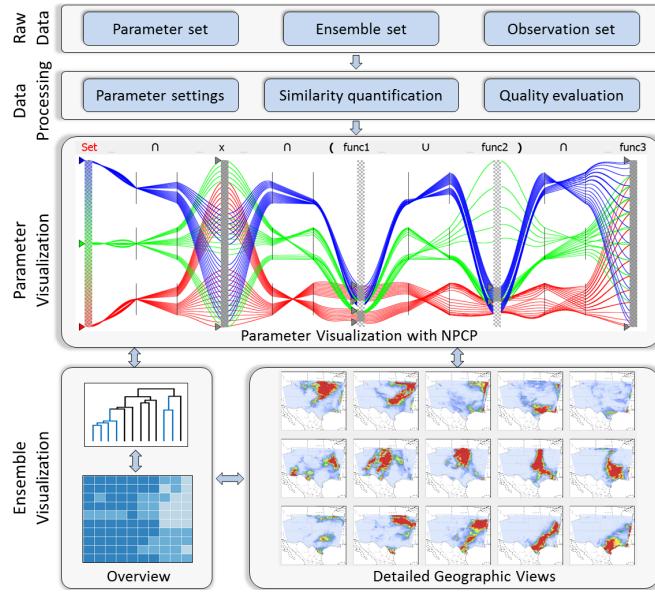


Fig. 2. Data processing pipeline and system overview.

#### 4 MULTI-RESOLUTION CONVECTIVE PARAMETER VISUALIZATION

This section focuses on fulfilling the requirement R1, R2 and R3. The input datasets for visualization are sets of convective parameters from climate simulations. Each set (i.e. *parameter set*) corresponds to ensemble runs of a certain resolution. Multiple parameters with different values in each resolution form a high-dimensional space. Most visualization of such high-dimensional data employs Parallel Coordinates Plot (PCP) [18], which is particularly useful in revealing the relational patterns between multiple dimensions in a single visualization. A conventional PCP visualization encodes high-dimensional data instances as polylines across axes (Figure 3A). We base our parameter visualization on PCP for its popularity and intuitiveness in visualizing multi-dimensional data. However, when visualizing the parameter data from multiple resolutions (sets), the effectiveness of PCPs is limited due to overplotting of polylines, as in this case the number of polylines is proportional to the number of parameter sets (Figure 3A). Consequently, the question is how to improve PCP to facilitate visual comparative analysis of data with multiple sets. To answer this question, we first review the existing design of PCPs, and then present our design.

##### 4.1 Reviewing PCPs from a Comparative Perspective

We review the existing designs for PCPs and highlight the need for a new design of PCPs to visualize the unique features of multi-resolution convective parameters in this section. A synthetic multi-set high-dimensional dataset (3 sets, 5 dimensions) is generated. Table I shows the formulas for the synthetic data. With distinct values of  $x$ , the formulas in each row generate various data instances for each set. The problems we are facing are: (1) visualizing the correlation between different dimensions of a set, such as the correlation between  $x$  and

$func1$  in set 1 ( $x$  and  $e^x$ ), i.e. requirement R1 (*intra-set correlation*); (2) comparing the correlations of specific dimensions in different sets, such as the correlations between the dimension  $func2$  and  $func3$  in set 2 ( $e^x$  and  $\sin(x)$ ) and set 3 ( $\sin(x)$  and  $\cos(x)$ ), i.e. requirement R2 (*inter-set correlation*).

Table 1. Formulas for the synthetic dataset used for experiments.

	set	$x$	$func1$	$func2$	$func3$
1		$x$	$e^x$	$-x$	$\cos(x)$
2		$x$	$\cos(x)$	$e^x$	$\sin(x)$
3		$x$	$-x$	$\sin(x)$	$\cos(x)$

**Superimposed PCP.** Figure 3A shows the result when we draw the synthetic dataset in a typical PCP. This design is ineffective of demonstrating the correlation between different dimensions (due to the overplotting), let alone the comparison between sets. From the comparative visualization perspective, such typical PCP uses the **Superimposition** design [9], which overlays all three sets in a single visualization. Nevertheless, since the three sets share the same axes, we can directly compare the data range and data distribution of the three sets in this design. For example, it is evident that set 1 (blue polylines) has a wider value range on axis  $func1$  than the other two sets.

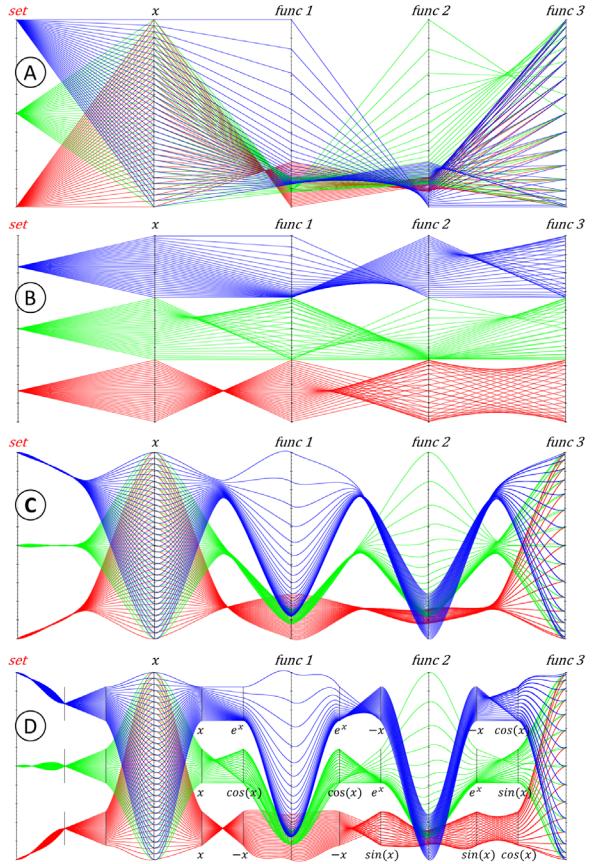


Fig. 3. (A) a conventional PCP with superimposition design; (B) three PCPs with juxtaposition design; (C) superimposed PCP with curve bundling [14] as explicit encoding; (D) nested parallel coordinates plot.

**Juxtaposed PCPs.** A straightforward approach to address the overplotting problem is to separate the three sets and visualize them in three distinct PCPs, as shown in Figure 3B. This design, which encourages side-by-side visual comparison of multiple facets of a complex dataset, is called **Juxtaposition** [9, 24]. Since sets are separated, the correlation between different dimensions in each set (*intra-set correlation*) is clearly revealed. The relational pattern of corresponding dimensions from different sets (*inter-set correlation*) is also obvious.

The disadvantage for this design is the inconvenience of comparing data range and data distribution from different sets. Although the juxtaposition design alleviates the problem of visual clutter, it predominantly relies on the use of the viewer's memory and attention shifts to make connections between different sets [9], which can add burdens onto the viewer's mental effort to interpret the relational patterns between sets regarding one axis.

**PCP with Explicit Encodings.** Figure 3C shows the superimposed PCP, but replacing each line segment between axes with two joint Bézier curves [14]. The joints of curves belonging to the same set are aggregated to a small vertical range to bundle the curves. The distortion of curves helps to maintain the geometric continuity and to trace data instances across axes. Different sets are effectively separated in the superimposed PCP by virtue of the bundling. The above mentioned visual encodings, such as bundling and distorting, are called **Explicit Encodings** in comparative visualization [9]. Although bundling enhances the perception of inter-set correlations in PCPs, the relational patterns between different axes regarding one set (*intra-set correlation*) are hindered due to bundling distortion, as curves from the same sets are aggregated to a single location or a small vertical range.

**Hybrid PCP.** To effectively show both intra-set and inter-set correlations (R1 and R2), none of the aforementioned designs is solely sufficient. Although one can resort to separate views of different PCP designs for different tasks, users may need to switch between different views multiple times during the exploration process, which can be time-consuming and labor-intensive. In addition, since the focus of interest is displayed separately in different views, users may lose their mental map when the view changes. To solve this problem, we propose *Nested Parallel Coordinates Plot (NPCP)* (shown in Figure 3D), which assembles superimposition, juxtaposition and explicit encoding designs for combining their respective advantages. More importantly, through simple interactions with the NPCP, users can control the ratio of superimposition and juxtaposition in the 2D space and smoothly transition from one design to the other, which significantly contributes to preserving users' mental map. The core idea is to nest multiple juxtaposed PCPs into the original superimposed PCP and employ curve bundling as explicit encoding. In this design, each line segment between axes is replaced with two cubic Bézier curves and one shorter line segment. The short line segments from different data instances of the same set constitute the relational pattern between axes (*intra-set correlation*). The pattern is demonstrated in a nested juxtaposed PCP. By virtue of the bundled curves on both sides of the shorter segments, nested juxtaposed PCPs from different sets can be separated and flexibly repositioned for side-by-side comparisons (*inter-set correlation*). The NPCP also demonstrates the data range and distribution on the axes of the original superimposed PCP.

## 4.2 Design of Nested Parallel Coordinates Plot

We now explain more details about the design of NPCP. Each original line segment between axes of the superimposed PCP is replaced with a pair of cubic Bézier curves ( $b_{i,1}$  and  $b_{i,2}$ ) and a shorter line segment ( $l_i$ ), as shown in Figure 4. The two curves enable us to flexibly distribute the nested juxtaposed PCPs, whereas the shorter line segment maintains the original relational pattern. We present the design details of NPCP through the explanation of its three parameters: horizontal spread, vertical spread and curviness.

**Horizontal Spread.** We call the axes of the original superimposed PCP as primary axes ( $X_i$  and  $X_{i+1}$ ) and the axes of the nested juxtaposed PCPs as secondary axes ( $W_i$  and  $W_{i+1}$ ). Horizontally, the nested juxtaposed PCPs will always be in the middle of two primary axes. The distance between a pair of secondary axes is defined as *horizontal spread*, denoted as  $\alpha$  in Figure 4.

**Vertical Spread.** The nested juxtaposed PCPs of different sets are uniformly distributed along the middle line of two primary axes. Their vertical positions depend on the number of sets in the NPCP. For example, if the number of set is three, three nested juxtaposed PCPs will be listed between every two neighboring primary axes. To avoid overlap, the maximum vertical length of each will be one third length of the primary axes. The *vertical spread*, denoted as  $\beta$  in Figure 4, describes

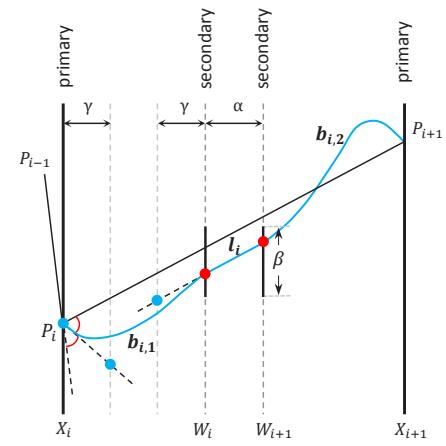


Fig. 4. One straight line segment  $\overrightarrow{P_i P_{i+1}}$  between axes is replaced by two cubic Bézier curves ( $b_{i,1}$  and  $b_{i,2}$ ) and one shorter line segment ( $l_i$ ).  $\alpha$  and  $\gamma$  represents the horizontal spread, vertical spread and curviness.

the vertical length of the secondary axes.

**Curviness.** The *curviness*, denoted as  $\gamma$ , is used to depict the distortion scale of the two cubic Bézier curves. The geometric shape of a cubic Bézier curve is decided by four control points. The start and end control points, that are on the curve, will be decided by the corresponding dimension's value of the data instance. The other two control points in the middle will be determined by the curviness value.

The secondary axes are shorter than the primary axes in NPCP, but the corresponding primary and secondary axis (such as  $X_i$  and  $W_i$ ) are representing the same dimension of the data and using the same value range. With the positions of a data instance on axis  $X_i$  and  $X_{i+1}$ ,  $P_i$  and  $P_{i+1}$ , we can derive its positions on  $W_i$  and  $W_{i+1}$ . The shorter line segment  $l_i$  can then be decided by connecting the two points on  $W_i$  and  $W_{i+1}$ , i.e. the two red points in Figure 4. As we can see,  $l_i$  will be parallel to  $\overrightarrow{P_i P_{i+1}}$ . The value range of the same dimension from different sets may vary a lot. For example, in Figure 3A, the value ranges of three sets in dimension *func2* are very different. When separating three sets from a superimposed primary axis to three juxtaposed secondary axes, if the secondary axes all use the same value range, some of them cannot be fully utilized and the correlation pattern in nested juxtaposed PCPs will be hindered, as shown in Figure 5 (left). To address this problem, we allow NPCP to normalize the values of the same dimension in different sets and map them to the entire range of the corresponding secondary axis, as demonstrated in Figure 5 (right).



Fig. 5. The  $\sin(x) - \cos(x)$  correlation pattern in NPCP. Left: values on the secondary axes are not normalized; right: values on the secondary axes have been normalized. In the right figure,  $l_i$  is not parallel to  $\overrightarrow{P_i P_{i+1}}$ , but the correlation pattern is more obvious.

The data instance's coordinates on  $X_i$  and  $W_i$  decide the positions of the start and end control points of the curve  $b_{i,1}$ . The other two control points are between  $X_i$  and  $W_i$ . Point  $P_{i-1}$  in Figure 4 denotes the intersection of the original polyline with axis  $X_{i-1}$ , i.e. the primary axis on the left of  $X_i$ . The extension of  $\overrightarrow{P_{i-1} P_i}$  beyond  $P_i$  and  $\overrightarrow{P_i P_{i+1}}$  form an angle. The second control point of  $b_{i,1}$  is on the bisector of this angle but  $\gamma$  away from axis  $X_i$ . If  $X_i$  is the leftmost axis, the bisector of the horizontal line and the segment  $\overrightarrow{P_i P_{i+1}}$  will be used instead. The third control point is on the extension of  $l_i$  beyond axis  $W_i$  and its distance to  $W_i$  is  $\gamma$ . With these four control points in order, we derive the cubic

Bézier curve  $b_{i,1}$ . Similarly, one can derive  $b_{i,2}$ . The curve generation method explained here is similar to [14]. The last question is how to decide the vertical order of the nested juxtaposed PCPs. Two methods have been used. One is using the index of sets, in which the vertical position of each set's nested juxtaposed PCPs is consistent across primary axes. For example, if we have three sets, the first set's juxtaposed PCPs will always take the first one third vertical space of the primary axes. Figure 3D is an example of this case. Another way is to sort sets based on the centroid of the original line segments in each set between every pair of neighboring primary axes. This method respects the values of data instances and usually has less curve distortion. However, the vertical positions of sets' nested juxtaposed PCPs may be different across primary axes. Figure 6B and 6C are examples of this case. The nested juxtaposed PCPs of the second set (green curves) are not always in the middle.

With the horizontal, vertical spread ( $\alpha, \beta$ ) and the number of sets, one can derive the ratio of space used for the nested juxtaposed PCPs. For example, if the number of sets is three, the juxtaposed space ratio will be  $3\alpha\beta$ . The continuous value ranges of  $\alpha$  and  $\beta$  guarantee smooth transitions among completely superimposed, completely juxtaposed and hybrid views. The control on curviness helps to preserve the continuity of polylines across axes.

#### 4.3 Interaction with NPCP

Except the basic interactions for PCPs, such as zooming and brushing, we provide several additional interaction techniques in NPCP.

**View Switching.** NPCP demonstrates intra-set and inter-set correlations with superimposed and juxtaposed views. These views are controlled by the horizontal spread and vertical spread of the NPCP. The curviness adjusts the geometric shape of curves and enhances continuity. In NPCP, we offer three default combinations of these parameters. These combinations avoid the tuning of these parameters and help to efficiently switch among completely juxtaposed view (Figure 6A), completely superimposed view (Figure 6D) and hybrid views (Figure 6B, 6C). Users can also adjust the value of parameters in each combination on demand. To simplify the user interface, we hide the widgets for the three control parameters. They will only be displayed when users want to adjust the value of the three parameters.

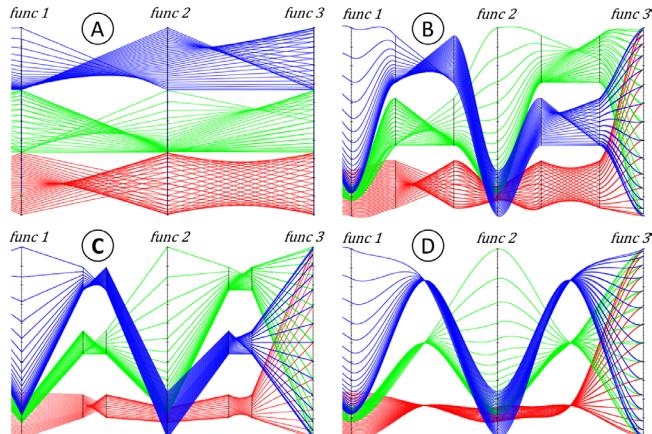


Fig. 6. The effects of three NPCP parameters. (A)  $\alpha = 1.0, \beta = 0.33, \gamma = 0.0$  (juxtaposition ratio is:  $3\alpha\beta=100\%$ ); (B)  $\alpha = 0.4, \beta = 0.25, \gamma = 0.1$ ; (C)  $\alpha = 0.16, \beta = 0.12, \gamma = 0.0$ ; (D)  $\alpha = 0.0, \beta = 0.0, \gamma = 0.15$ .

**Flexible Color Encoding.** The color of a polyline in conventional PCPs is decided by a certain dimension's value of the corresponding data instance. The axis corresponding to this dimension is called the *active axis* in NPCP. The rest axes are *passive axes*. We embed a *color/transparency control widget* (Figure 7) on the active axis to enable flexible color encoding of polylines in NPCP. The control points on the widget control the color/transparency of the color strip overlaid on the axis, as well as the color/transparency of polylines. Linear interpolation between control points smooths the color/transparency trans-

sition. Users can add/edit/remove control points to modify the visual appearance of polylines. Similar control widgets are also embedded in passive axes. However, passive axes have no control on colors; they only affect the transparency of polylines. Control points on these axes have opacity value either 0 or 1. The transparency value of a polyline is derived from the multiplication of the opacity values from all axes (both active and passive). Therefore, as long as one axis has opacity 0, the entire polyline (high-dimensional data instance) will be invisible.

Selections are usually performed via brushing on polylines in PCPs. However, due to the overlap of polylines, preciseness is hardly achieved when data size is large. With the flexible color encoding of NPCP, many polylines can be hidden (set to be transparent) before the brushing. The color/transparency control widgets, together with brushing, provide two-step selections in NPCP. The widgets also help in flexibly adjusting the transparency of polylines, which can effectively fade off uninterested data instances when data size is large.

**High-Dimensional Range Query.** As explained in Section 3.3 (R3), the domain scientist needs to perform range query on the high-dimensional parameter space. The query calculation can be done offline and be separated from the visualization. In this case, the offline calculation results will be taken as the input of the visualization. Conversely, the visualization results, which will guide the domain scientist to refine his query, affect the next offline calculation. Parameter analysis is usually conducted through such a loop between calculation and visualization with the interaction from domain scientists. The offline calculation introduces I/O overhead and the resulted latency discourages the analysis process. NPCP supports real-time high-dimensional range query by enabling set operations, i.e. *unions* and *intersections*, between the primary axes. The *differences* operations can be done through adjusting the control points of the color/transparency control widgets embedded in each axis (Figure 7).

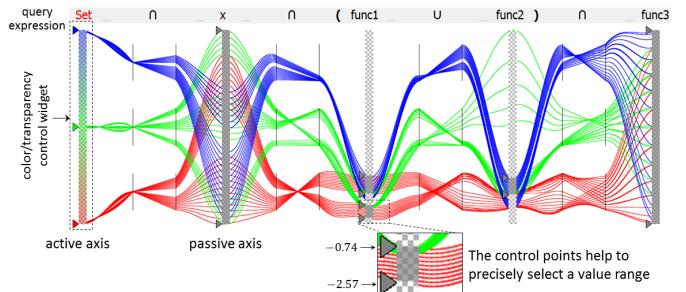


Fig. 7. High-dimensional range query in NPCP with the help of color/transparency control widgets embedded in each axis.

When performing high-dimensional range query with NPCP, the query expression needs to be determined first. Axes of NPCP can then be shuffled based on the appearance order of operands (different dimensions of the data) in the expression. With the help of the control points on the color/transparency control widgets, accurate numerical value range/ranges of each dimension can be selected. In the enlarged view of Figure 7, two control points precisely select the value range  $[-2.57, -0.74]$  on axis  $func1$ . Next, the query expression is visually composed by changing the set operators between axes and adding/removing parentheses on demand. The query expression in Figure 7 is:  $\{1 \leq set \leq 3\} \cap \{3.14 \leq x \leq 3.14\} \cap \{(-2.57 \leq func1 \leq -0.74) \cup \{1.36 \leq func1 \leq 3.33\} \cup \{0.61 \leq func2 \leq 3.29\}\} \cap \{-1 \leq func3 \leq 1\}$ . After the expression is finalized, NPCP evaluates the query expression at background to query out the desired data instances and visualizes them immediately.

#### 5 MULTI-RESOLUTION ENSEMBLE VISUALIZATION

Our visualization system adopts the *overview+detail* exploration technique to demonstrate the multi-resolution spatial temporal ensembles. An overview of the ensembles is demonstrated by heat maps and dendograms, whereas small multiple geographic views are employed for detailed examination on one particular ensemble member/item.

## 5.1 Understanding the Big Picture

The first thing that the domain scientist wants to see from the ensembles is the accuracy of the simulations (R4). The predicted precipitation is the scientist's research focus and its accuracy can be measured by comparing each ensemble item with the corresponding observation item collected from satellites. Mean Square Error (MSE) is the most commonly used comparison metric. However, the metric does not consider the spatial structure of data, which leads to inaccurate comparisons in certain cases. To improve the reliability of accuracy evaluation, we have also calculated the Structure Similarity (SSIM) index [45], in addition to the MSE, for each ensemble item. The SSIM algorithm has been used extensively for image comparison. In our case, the latitude/longitude of a grid point in an ensemble item is similar to the row/column index of a pixel in a 2D image. Therefore, we can easily adapt the SSIM algorithm to assess the quality of ensemble items. Similar to the MSE calculation, the output of SSIM is a numerical value denoting the similarity between an ensemble item and an observation item. Similar items are considered as accurate. Since each ensemble member contains multiple ensemble items, the quality of an ensemble member is described by an accuracy vector whose elements' value represents the corresponding ensemble items' accuracy.

Next, Dynamic Time Warping (DTW) [41] is applied on the accuracy vectors of different ensemble members to measure the similarity between them (R5). The DTW algorithm was originally invented for speech recognition [38]. It measures the similarity of two temporal sequences (feature vectors) by aligning the best matching elements between the two sequences. Since each climate ensemble member has a sequence of time-related accuracy values, DTW can be used here to better align them and provide a more reliable similarity quantification. For example, the precipitation of ensemble member  $B$  might be very similar to ensemble member  $A$ , but only delayed for one day (ensemble item). Simply comparing these two members day by day may lead to the conclusion that they are very dissimilar. DTW can help to handle this case. We compare all pairs of ensemble members using DTW and the output is organized as a symmetric matrix, in which the element of row  $i$  and column  $j$  represents the similarity value between the  $i$ th and  $j$ th ensemble member. We then apply agglomerative clustering [13] on the similarity matrix to group similar ensemble members. The clustering generates a hierarchical tree, in which ensemble members are the leaves and similar members are grouped into the same branch.

Algorithm 1 demonstrates the accuracy calculation, similarity quantification and agglomerative clustering processes. Line 6 computes the SSIM index. For MSE, one can simply replace the SSIM calculation in this line. The result, matSSIM, is a 2D matrix with the first dimension (column) representing the ensemble members and the second dimension (row) representing the ensemble items. Every pair of rows in matSSIM is then taken as the input of DTW (line 12). The output, matSimilarity, is the symmetric similarity matrix. The agglomerative clustering algorithm can then apply on matSimilarity to derive the hierarchical cluster tree (line 16).

---

### Algorithm 1 Quality, similarity quantification and clustering

---

```

1: num_member <- number of ensemble member
2: num_item <- number of ensemble item in each ensemble member
3: // compute SSIM index of each ensemble item
4: for i <= 1; i <= num_member; i++ do
5:   for j <= 1; j <= num_item; j++ do
6:     matSSIM[i][j] <- SSIM(ensemble item, observation item);
7:   end for
8: end for
9: // compute similarity matrix using DTW
10: for i <= 1; i <= num_member; i++ do
11:   for j <= 1; j <= num_member; j++ do
12:     matSimilarity[i][j] <- DTW(matSSIM[i], matSSIM[j]);
13:   end for
14: end for
15: // use similarity matrix for clustering
16: hierarchicalTree <- AgglomerativeClustering(matSimilarity);

```

---

A linked visualization of a heat map and a dendrogram is designed to provide an overview of one ensemble set (Figure 8). Multiple ensemble sets will need multiple such linked visualization views.

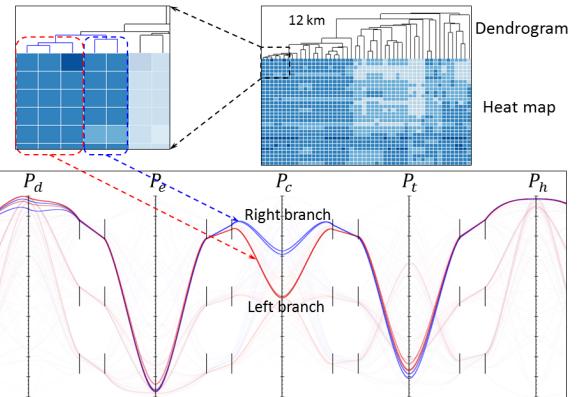


Fig. 8. Ensemble dissimilarity resulted from distinct parameter settings.

**Heat Map.** The horizontal dimension of the heat map is the ensemble members in the same ensemble set. They are ordered according to the agglomerative clustering steps (i.e. the order of leaves in the hierarchical tree). The vertical dimension represents the ensemble items with earlier ones at the bottom. This dimension, therefore, presents the quality of ensemble items over time. We categorize ensemble items into four groups based on their quality: *Very Accurate*, *Accurate*, *Inaccurate* and *Very Inaccurate* (Figure 1 B2), one color for each group in the heat map. Certain groups of ensemble items in the heat map can be filtered out by domain science users. Clicking a certain ensemble item leads the users to its detail in the geographic views.

**Dendrogram.** The leaves of the hierarchical cluster tree represent the ensemble members. They are ordered according to the agglomeration steps. So, the leftmost two leaf nodes (the first two agglomerated members) are the most similar ensemble members. The height (vertical position) of an internal node indicates the dissimilarity scale between its two children. Users can select on the dendrogram to examine certain branches of the tree. The selection will be synchronized in the NPCP view to show the convective parameters used for these ensemble members. Figure 8 demonstrates an example. The domain scientist zoomed into the top left corner of the heat map and selected the first five ensemble members (one branch of the dendrogram). The difference between convective parameters used for the first three members and the last two members is clearly demonstrated in the NPCP.

## 5.2 Drilling Down to Spatial Geographic Views

The geographic views enable scientists to thoroughly examine on a particular ensemble member/item. We present ensembles with geographic views since the spatial information is one of the most important dimensions of ensemble sets (R6).

**Spatial Facet of Ensemble Items.** Three geographic views, one for a climate ensemble item (denoted as  $V_1$ , Figure 1 C1), one for an observation item (denoted as  $V_2$ , Figure 1 C2) and one for their difference (denoted as  $V_3$ , Figure 1 C3), are shown side-by-side. The difference value at each spatial location (grid point) in  $V_3$  is derived by the value at the same location of the ensemble item in  $V_1$  and observation item in  $V_2$ . We allow scientists to customize the calculation of this difference value in the system by providing their own equation to describe the relation between  $V_1$ ,  $V_2$  and  $V_3$  (Figure 1 C4). The most straightforward difference calculation is simply subtracting the observation item from the ensemble item, which is shown in Figure 1 C3.

Our system supports ensemble items aggregation in  $V_1$ . It causes confusion if multiple parameter settings or ensemble members are selected from the NPCP, heat maps or dendograms, but only the item of one member is visualized in  $V_1$ . To address this issue, we provide the option to merge the corresponding items of all selected ensemble members. If selected members are all from the same resolution, the

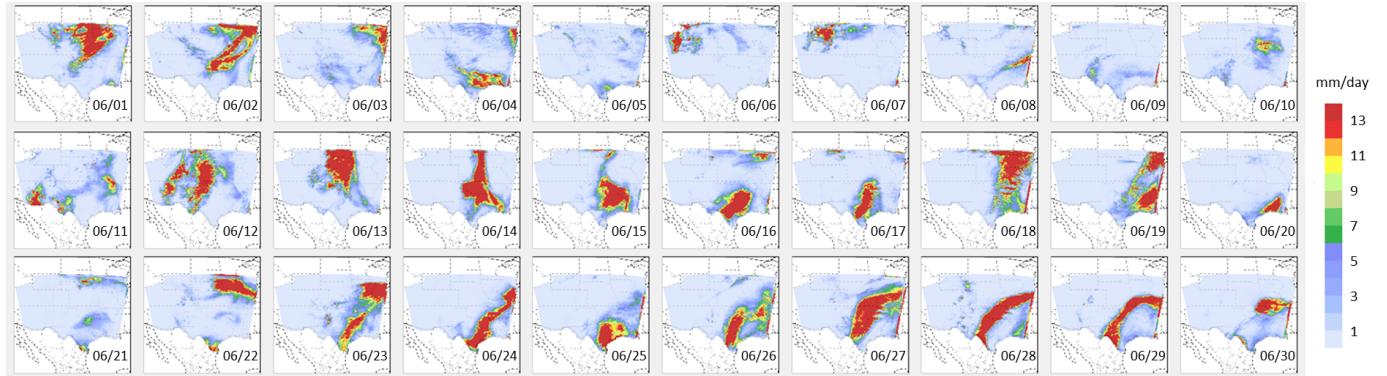


Fig. 9. One ensemble member with 30 items: the predicted precipitation over June 2007, from the 33<sup>rd</sup> member of the 12 km ensemble set.

aggregation of ensemble members on a particular item is conducted through a weighted sum. The weight is decided by the opacity value of the corresponding polyline in the NPCP, which can be adjusted by the color/transparency control widget embedded in each axis. On the other hand, if selected ensemble members are from different resolutions, linear interpolations will be applied first to up-sample the lower resolution ensemble items before the aggregation.

**Spatial Facet of Ensemble Members.** To study spatial temporal climate ensembles, domain scientists need to understand the temporal evolution of ensemble members over geolocations (R6). The ensemble item visualization view (*VI*) can only show the spatial information of one ensemble item. Although we provide scientists with some control widgets (Figure 1 C5) to help them switch between different items, it is difficult to preserve their mental map during view switching. For this reason, we provide the option of showing all items of one ensemble member in multiple juxtaposed geographic views (multiple *VI*). The content of these views can also be multiple observation items (multiple *V2*), or multiple difference views (multiple *V3*). Figure 9 shows one ensemble member with 30 items in 30 small multiple views.

## 6 REAL-WORLD USE-CASES

In this section, we demonstrate the effectiveness of our system with real-world climate ensemble sets from the domain scientist.

### 6.1 Data Description

The domain scientist used the Weather Research and Forecasting (WRF) [43] model to generate climate ensembles with three different spatial resolutions: 50 km, 25 km and 12 km. The number of grid points in each resolution is (latitude×longitude): 43×44, 87×89 and 182×187 respectively. The 50 km ensemble has the lowest grid resolution. These three ensembles were generated to simulate the weather over the Southern Great Plains (SGP) region (latitude: 25.0°~44.0°, longitude: -112.0°~−90°) in June 2007. The Kain-Fritsch [20] convection parameterization scheme is applied in the study and the scientist is interested in the effects of five convective parameters on the precipitation output. These parameters are: the coefficient related to downdraft mass flux rate ( $P_d$ ), the coefficient related to entrainment mass flux rate ( $P_e$ ), the maximum turbulent kinetic energy ( $P_t$ ), the starting height of downdraft above updraft source layer ( $P_h$ ) and the average consumption time of convective available potential energy ( $P_c$ ).

**Parameter Set.** Three sets of parameters have been used to generate the three ensemble sets with the resolution of 50 km, 25 km and 12 km respectively. We use  $P^{50}$ ,  $P^{25}$  and  $P^{12}$  to denote these three parameter sets and they are the input of our NPCP. Each set contains 150 parameter settings; and each setting is a combination of five values representing the values used for the five convective parameters.

**Ensemble Set.** We use  $E^{50}$ ,  $E^{25}$  and  $E^{12}$  to distinguish the three ensemble sets generated using  $P^{50}$ ,  $P^{25}$  and  $P^{12}$  respectively. Each ensemble set has 150 ensemble members. Members of the three ensemble sets all have 30 ensemble items. Each item represents the precipitation of one day. So, the 30 items cover the month of June 2007.

**Observation Set.** Three sets of observed precipitation values over the SGP region in June 2007 were collected from satellites. Their resolutions are: 50 km, 25 km and 12 km; we use  $O^{50}$ ,  $O^{25}$  and  $O^{12}$  to represent them. Each set also contains the precipitation of 30 days. The observed precipitation of one day is one observation item, which can be used to assess the corresponding ensemble item.

### 6.2 Case Studies with Domain Scientists' Feedback

We have verified the effectiveness and usefulness of our system with two domain experts. The first one is the scientist who provided the ensemble data. He was deeply involved in the task abstraction [27] process during the design of our system. The layout of different views (Figure 1) was also suggested by him. The second one is an expert from atmospheric sciences, focusing on dynamical predictability and multiscale convective systems. Both experts have been working with climate data over ten years and are ideal users of our system.

We asked the experts to use the conventional superimposed PCP, juxtaposed PCPs and our NPCP to describe the difference between intra/inter-set correlation patterns. Figure 10 (left) shows part of the NPCP visualization of the convective parameters. Values on all axes increase from bottom to top. The axis  $id$  corresponds to the index of ensemble members, which is their execution order (earlier ensemble members have smaller indices), and the polylines are colored by the  $id$  value of corresponding data instances. Polylines representing early ensemble members are in red; while polylines for later members are in blue. The color strip embedded in axis  $id$  shows the color mapping. It can be seen from the NPCP that the red polylines intersect with the axis  $P_e$ ,  $P_h$  and  $P_t$  in wider ranges; whereas the blue polylines intersect with them in narrower ranges. The narrow ranges where most of the blue polylines intersect with each axis imply the more stable values for each parameter in the later runs of the simulation. The nested juxtaposed PCPs in the bottom, middle and top row of the NPCP demonstrate the correlation patterns between different dimensions in  $P^{50}$ ,  $P^{25}$  and  $P^{12}$  respectively. The differences of the correlation patterns between  $P_h$  and  $P_t$  in  $P^{50}$  and  $P^{25}$  (Figure 10A, 10B), and in  $P^{50}$  and  $P^{12}$  (Figure 10A, 10C) can be seen from the nested juxtaposed PCPs. The experts cannot tell the differences when using the conventional superimposed PCP since patterns from three sets are overlapped, as shown in Figure 10 (middle). The differences can also be identified from the juxtaposed PCPs (Figure 10D and 10E, 10D and 10F). Nevertheless, when the correlation patterns are very similar, the conventional juxtaposed PCPs will not be effective. For example, the experts can hardly differentiate the correlation patterns between  $P_h$  and  $P_t$  in  $P^{25}$  and  $P^{12}$  (Figure 10E, 10F). These two patterns in the nested juxtaposed PCPs of NPCP are also hard to differentiate (Figure 10B, 10C). However, by virtue of the continuous geometric shape of Bézier curves, both experts can trace the blue polylines and find the narrow ranges where they intersect with the primary axis  $P_t$ . The stable values of  $P_t$  in  $P^{25}$  are a little greater than its stable values in  $P^{12}$  (Figure 10G). Similarly, they can find the stable values of  $P_h$  used in  $P^{25}$  are smaller than its stable values in  $P^{12}$  (Figure 10H). The difference of  $P_h$  values in these

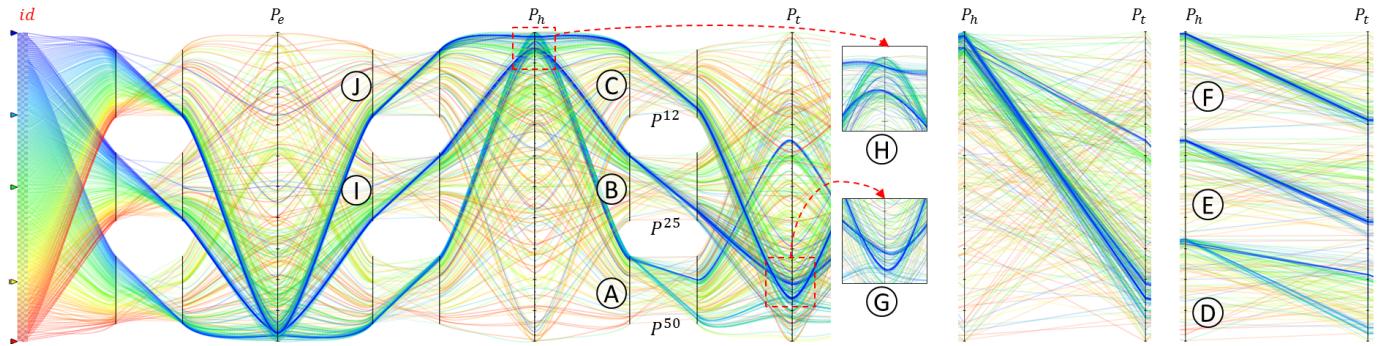


Fig. 10. Left: multi-resolution convective parameter visualization in NPCP; middle: the correlation patterns between  $P_h$  and  $P_t$  in  $P^{50}$ ,  $P^{25}$  and  $P^{12}$  are overlapped in the conventional superimposed PCP; right: the correlation patterns between  $P_h$  and  $P_t$  in  $P^{50}$ ,  $P^{25}$  and  $P^{12}$  in juxtaposed PCPs.

two sets also helped the experts to differentiate the two patterns in Figure 10I and 10J. Both superimposed and juxtaposed comparisons are necessary here. The experts performed high dimensional range query by visually composing set expressions in NPCP. Previously, to find out desired data instances, they have to write different Python/Matlab scripts and use separate visualization packages. A minor change to the query requires repeating the entire data processing pipeline. The intuitive visual query interaction of NPCP significantly reduced the time for data query.

We also asked the experts to describe their findings about the three ensemble sets and describe the predicted precipitation according to the most accurate member they found using our system. Three heat maps in Figure 1 B1, from left to right, demonstrate the overall quality of 50 ensemble members selected from  $E^{50}$ ,  $E^{25}$  and  $E^{12}$  respectively. With the help of the color legends (Figure 1 B2), both experts agreed that the precipitation predictions for early days are more accurate than later days and the lower resolution ensembles have higher overall quality. They commented that simulations with finer grid resolutions are not guaranteed to have higher overall quality. With the ensemble quality and cluster information from heat maps and dendograms, the system directed both experts to the 33<sup>rd</sup> member of  $E^{12}$  (the leftmost leaf in the dendrogram of  $E^{12}$ ). Figure 9 shows the precipitation predicted by this member. The color red specifies the spatial regions where the precipitation is high. In general, the 5<sup>th</sup>, 9<sup>th</sup>, and 21<sup>st</sup> of June have less precipitation than other days and the west coast regions have less precipitation than the east coast regions. Temporal evolutions, like the moving of precipitation from the south east corner to the middle region from the 25<sup>th</sup> to the 27<sup>th</sup> of June, can also be found in the figure.

### 6.3 Discussion

Overall, both domain experts found the results promising. The first expert was satisfied with the effectiveness of the NPCP in demonstrating parameter correlations across resolutions; while the second expert liked the heat maps and dendograms in assessing and clustering ensemble members. Meanwhile, both experts confirmed that the linked views helped them build the connection between parameter settings and ensemble members. The system has the potential to be used for other applications and NPCP can be easily adapted to other multi-set data. In this work, we specifically focused on addressing the domain tasks from the first domain expert, as well as fulfilling the functionality of the system. Generalizing the technique and analyzing other datasets with the system would be good directions for future exploration.

There are several limitations in our current design of NPCP. First, the visual representation of NPCP looks complex in certain cases. We designed NPCP particularly for the objective of revealing the intra/inter-set correlations, which can hardly be accomplished by the conventional PCPs. For more general multi-variate data analysis, we would expect the conventional PCPs to be used more often. Second, short-time training may be required to help users get familiar with the system, especially for users who have no prior experience with the conventional PCPs. Our first expert used PCPs before this work. We

only spent time on explaining the three parameters of NPCP to him and he could adjust NPCP to desired views and explore ensembles with the system independently in less than 30 minutes. The second expert had no prior knowledge about PCPs. It took her around one hour to fully understand the system (including the time to understand the data). Therefore, we would expect the average training time to be less than or around one hour. Third, similar to the conventional PCPs, the superimposed facet (primary axes) of NPCP could result in visual clutter if the number of data instances is large. The nested juxtaposed PCPs separate data instances by sets, thus reducing visual clutter to some extent. This is more obvious when the number of sets is large since more juxtaposed PCPs will be used. Given that the size of the domain data we are dealing with is not extremely large, we put less emphasis on the scalability issue in the current work.

## 7 CONCLUSION AND FUTURE WORK

In this paper, we proposed *Nested Parallel Coordinates Plot (NPCP)* for the visualization and analysis of multi-resolution high-dimensional convective parameter space in climate modeling. NPCP integrates superimposed PCP and juxtaposed PCPs into one view and employs curve bundling as explicit encoding to demonstrate intra-resolution and inter-resolution parameter correlations. It is also equipped with intuitive high-dimensional range query that assists domain scientists to promptly retrieve the desired data instances. An *overview+detail* spatial temporal ensemble exploration and visualization system is also developed. We integrated NPCP into the system to help domain scientists establish the connection between the complex parameter settings and the intricate climate ensembles. Several use-cases with real-world climate data demonstrate the effectiveness of the system. In the future, we would like to work with more domain scientists, explore other climate data using the system, and develop more formal and statistical case studies to further establish the effectiveness of the system in parameter analysis and ensemble visualization. We also plan to apply NPCP to other application areas, in which the domain data can be generalized as multi-set high-dimensional data.

## ACKNOWLEDGMENTS

The authors would like to thank Wen-wen Tung for providing valuable feedback to this work. This work was supported in part by NSF grants IIS-1250752, IIS-1065025, and US Department of Energy grants DE-SC0007444, DE-DC0012495, program manager Lucy Nowell.

## REFERENCES

- [1] N. Andrienko and G. Andrienko. A visual analytics framework for spatio-temporal analysis and modelling. *Data Mining and Knowledge Discovery*, 27(1):55–83, 2013.
- [2] M. Ankerst, S. Berchtold, and D. A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. In *Information Visualization. IEEE Symposium on*, pages 52–60. IEEE, 1998.
- [3] A. O. Artero, M. C. F. de Oliveira, and H. Levkowitz. Uncovering clusters in crowded parallel coordinates visualizations. In *Information Visualization, 2004. IEEE Symposium on*, pages 81–88. IEEE, 2004.

- [4] A. Bock, A. Pembroke, M. L. Mays, L. Rastaetter, A. Ynnerman, and T. Ropinski. Visual Verification of Space Weather Ensemble Simulations. In *Proceedings of the IEEE Vis*, 2015.
- [5] H. Chen, S. Zhang, W. Chen, H. Mei, J. Zhang, A. Mercer, R. Liang, and H. Qu. Uncertainty-aware multidimensional ensemble data visualization and exploration. *Visualization and Computer Graphics, IEEE Transactions on*, 21(9):1072–1086, 2015.
- [6] A. Dasgupta and R. Kosara. Pargnostics: Screen-space metrics for parallel coordinates. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1017–1026, 2010.
- [7] N. Elmquist, P. Dragicevic, and J.-D. Fekete. Rolling the dice: Multi-dimensional visual exploration using scatterplot matrix navigation. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1539–1148, 2008.
- [8] Y.-H. Fu, M. O. Ward, and E. A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *Proceedings of the conference on Visualization'99: celebrating ten years*, pages 43–50. IEEE Computer Society Press, 1999.
- [9] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.
- [10] D. Guo, J. Chen, A. M. MacEachren, and K. Liao. A visualization system for space-time and multivariate patterns (vis-stamp). *Visualization and Computer Graphics, IEEE Transactions on*, 12(6):1461–1474, 2006.
- [11] Z. Guo, M. Wang, Y. Qian, V. E. Larson, S. Ghan, M. Ovchinnikov, P. A. Bogenschutz, C. Zhao, G. Lin, and T. Zhou. A sensitivity analysis of cloud properties to clubb parameters in the single-column community atmosphere model (scam5). *Journal of Advances in Modeling Earth Systems*, 6(3):829–858, 2014.
- [12] L. Hao, C. G. Healey, and S. A. Bass. Effective visualization of temporal ensembles. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):787–796, 2016.
- [13] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction (second edition)*. Springer, 2009.
- [14] J. Heinrich, Y. Luo, A. E. Kirkpatrick, H. Zhang, and D. Weiskopf. Evaluation of a bundling technique for parallel coordinates. *Information Visualization Theory and Applications*, pages 594–602, 2012.
- [15] J. Heinrich, J. Stasko, and D. Weiskopf. The Parallel Coordinates Matrix. *Procdeeing of Eurographics Conference on Visualization—Short Papers*, pages 37–41, 2012.
- [16] J. Heinrich and D. Weiskopf. State of the art of parallel coordinates. *STAR Proceedings of Eurographics*, 2013:95–116, 2013.
- [17] Z. Hou, M. Huang, L. R. Leung, G. Lin, and D. M. Ricciuto. Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the community land model. *Journal of Geophysical Research: Atmospheres*, 117(D15), 2012.
- [18] A. Inselberg. The plane with parallel coordinates. *The visual computer*, 1(2):69–91, 1985.
- [19] W. Javed and N. Elmquist. Exploring the design space of composite visualization. In *Visualization Symposium (PacificVis), 2012 IEEE Pacific*, pages 1–8. IEEE, 2012.
- [20] J. S. Kain. The kain-fritsch convective parameterization: an update. *Journal of Applied Meteorology*, 43(1):170–181, 2004.
- [21] J. Kehrer, H. Piringer, W. Berger, and M. E. Groller. A model for structure-based comparison of many categories in small-multiple displays. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2287–2296, 2013.
- [22] X. Liu, Y. Hu, S. North, and H.-W. Shen. Correlatedmultiples: Spatially coherent small multiples with constrained multi-dimensional scaling. In *Computer Graphics Forum*. Wiley Online Library, 2015.
- [23] X. Liu and H.-W. Shen. The effects of representation and juxtaposition on graphical perception of matrix visualization. In *Proceedings of the 33rd annual ACM conference on Human factors in computing systems*, pages 269–278. ACM, 2015.
- [24] X. Liu and H.-W. Shen. Association analysis for visual exploration of multivariate scientific data sets. *Visualization and Computer Graphics, IEEE Transactions on*, 22(1):955–964, 2016.
- [25] K. T. McDonnell and K. Mueller. Illustrative parallel coordinates. In *Computer Graphics Forum*, volume 27, pages 1031–1038, 2008.
- [26] K. Misue, P. Eades, W. Lai, and K. Sugiyama. Layout adjustment and the mental map. *Journal of visual languages and computing*, 6(2):183–210, 1995.
- [27] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014.
- [28] T. Nocke, M. Flechsig, and U. Böhm. Visual exploration and evaluation of climate-related simulation data. In *Simulation Conference, 2007 Winter*, pages 703–711. IEEE, 2007.
- [29] H. Obermaier and K. I. Joy. Future challenges for ensemble visualization. *Computer Graphics and Applications, IEEE*, 34(3):8–11, 2014.
- [30] G. Palmas, M. Bachynskyi, A. Oulasvirta, H. P. Seidel, and T. Weinkauf. An edge-bundling layout for interactive parallel coordinates. In *Visualization Symposium (PacificVis)*, pages 57–64. IEEE, 2014.
- [31] W. Peng, M. O. Ward, and E. A. Rundensteiner. Clutter reduction in multi-dimensional data visualization using dimension reordering. In *Information Visualization. IEEE Symposium on*, pages 89–96. IEEE, 2004.
- [32] M. N. Phadke, L. Pinto, O. Alabi, J. Harter, R. M. Taylor II, X. Wu, H. Petersen, S. A. Bass, and C. G. Healey. Exploring ensemble visualization. In *IS&T/SPIE Electronic Imaging*, pages 82940B–82940B. International Society for Optics and Photonics, 2012.
- [33] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. Schwalm, R. Cook, E. Bertini, and C. Silva. Similarityexplorer: A visual inter-comparison tool for multifaceted climate data. In *Computer Graphics Forum*, volume 33, pages 341–350. Wiley Online Library, 2014.
- [34] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. R. Schwalm, D. N. Huntzinger, R. Cook, E. Bertini, and C. T. Silva. Visual reconciliation of alternative similarity spaces in climate modeling. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1923–1932, 2014.
- [35] K. Potter, P. Rosen, and C. R. Johnson. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *Uncertainty Quantification in Scientific Computing*, pages 226–249. Springer, 2012.
- [36] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pasucci, and C. Johsson. Visualization of uncertainty and ensemble data: Exploration of climate modeling and weather forecast data with integrated visus-cdar systems. In *Journal of Physics: Conference Series*, volume 180, page 012089. IOP Publishing, 2009.
- [37] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pasucci, and C. R. Johnson. Ensemble-vis: A framework for the statistical visualization of ensemble data. In *Data Mining Workshops. ICDMW'09. IEEE International Conference on*, pages 233–240. IEEE, 2009.
- [38] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 26(1):43–49, 1978.
- [39] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. J. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1421–1430, 2010.
- [40] M. Sedlmair, C. Heinzl, S. Bruckner, H. Piringer, and T. Moller. Visual parameter space analysis: A conceptual framework. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):2161–2170, 2014.
- [41] P. Senin. Dynamic time warping algorithm review. *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA*, pages 1–23, 2008.
- [42] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.
- [43] W. C. Skamarock, J. B. Klemp, J. Dudhia, D. O. Gill, D. M. Barker, M. G. Duda, X.-Y. Huang, W. Wang, and J. G. Powers. A description of the advanced research wrf version 3. *Technical Note NCAR/TN468+STR, National Center for Atmospheric Research, Boulder, CO*, 2008.
- [44] C. Tominski, C. Forsell, and J. Johansson. Interaction support for visual comparison inspired by natural behavior. *Visualization and Computer Graphics, IEEE Transactions on*, 18(12):2719–2728, 2012.
- [45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.
- [46] H. Yan, Y. Qian, G. Lin, L. Leung, B. Yang, and Q. Fu. Parametric sensitivity and calibration for the kain-fritsch convective parameterization scheme in the wrf model. *Climate Res*, 59:135–147, 2014.
- [47] B. Yang, Y. Qian, G. Lin, R. Leung, and Y. Zhang. Some issues in uncertainty quantification and parameter tuning: a case study of convective parameterization scheme in the wrf regional climate model. *Atmos. Chem. Phys*, 12(5):2409–2427, 2012.
- [48] H. Zhou, X. Yuan, H. Qu, W. Cui, and B. Chen. Visual clustering in parallel coordinates. In *Computer Graphics Forum*, volume 27, pages 1047–1054. Wiley Online Library, 2008.