eBoF: Interactive Temporal Correlation Analysis for Ensemble Data Based on Bag-of-Features

Zhifei Ding[1], Rongtao Qian[1], Siru Chen[1], Lingxin Yu[1], Jiahao Han[1], Yu Zhu[1], and Richen Liu[1]

[1]Nanjing Normal University

October 31, 2022

Abstract

We propose eBoF, a novel time-varying ensemble data visualization approach based on bag-of-features (BoF). In the eBoF model, we take a simple and monotone interval from all target variables of ensemble scalar data as a local feature patch of BoF model and the duration time of each interval (i.e., feature patch) as its frequency. The feature clusters in ensemble runs are then identified based on the similarity of temporal correlations. eBoF generates the clusters together with their probability distribution across all the feature patches while storing the geo-spatial information, which is often lost in the traditional topic modelling or clustering algorithms. The probability distribution across different clusters can help to generate reasonable clustering results evaluated by the domain knowledge. We conduct several case studies and performance analyses. We also consult the domain experts to evaluate the proposed eBoF model. Evaluation results suggest the proposed eBoF can provide insightful and comprehensive evidence on ensemble simulation data analysis.

Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

eBoF: Interactive Temporal Correlation Analysis for Ensemble Data Based on Bag-of-Features

Zhifei Ding | Rongtao Qian | Siru Chen | Lingxin Yu | Jiahao Han | Yu Zhu | Richen Liu*

School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University, China
Jiangsu Engineering Laboratory of Intelligent Information Processing and Software, Nanjing Normal University, China

Correspondence
*Richen Liu, School of Computer and Electronic Information/School of Artificial Intelligence, Nanjing Normal University, China.
Email: richen@pku.edu.cn

Summary

We propose eBoF, a novel time-varying ensemble data visualization approach based on bag-of-features (BoF). In the eBoF model, we take a simple and monotone interval from all target variables of ensemble scalar data as a local feature patch of BoF model and the duration time of each interval (i.e., feature patch) as its frequency. The feature clusters in ensemble runs are then identified based on the similarity of temporal correlations. eBoF generates the clusters together with their probability distribution across all the feature patches while storing the geo-spatial information, which is often lost in the traditional topic modelling or clustering algorithms. The probability distribution across different clusters can help to generate reasonable clustering results evaluated by the domain knowledge. We conduct several case studies and performance analyses. We also consult the domain experts to evaluate the proposed eBoF model. Evaluation results suggest the proposed eBoF can provide insightful and comprehensive evidence on ensemble simulation data analysis.

KEYWORDS:
feature extraction, computing and processing, ensemble data analysis, data clustering

1 | INTRODUCTION

The improvement of computation power and the development of supercomputers make simulations frequently used in different scientific domains, such as climatology, meteorology, aerodynamics, and oceanography. A simulation ensemble is a set of simulation runs generated from models with different initial values and boundary conditions[1]. Scientists can study the uncertainties and parameter sensitivities of the corresponding model by analyzing simulation members in an ensemble. The certain dynamic processes of ensemble data are simulated to study the accumulated uncertainty over time.

Studying such a process is necessary to understand the simulation model. On the one hand, domain scientists are concerned about the evolution of certain variables in the model. For instance, changes in temperature, pressure, precipitation, and other related attributes in climate research can help predict disastrous or abnormal weather conditions. On the other hand, studying the whole dynamic process is expected to reveal detailed information about the evolution of uncertainty in the model. Such information helps scientists be aware of the uncertainty accumulation along time , understand and control it. Simulation data include the data in simulation parameter space, simulation process, simulation output, and even those in situ visualization. Most scientific simulation data are spatio-temporal, which are often characterized as multi-faceted, including multi-run, multi-variate, multi-dimensional[2].

Time-varying is another intrinsic characteristic of ensemble simulation data. It is difficult to develop visualization and analysis methods for dealing with a single aspect. Handling them simultaneously is a huge challenge. Domain experts in climate

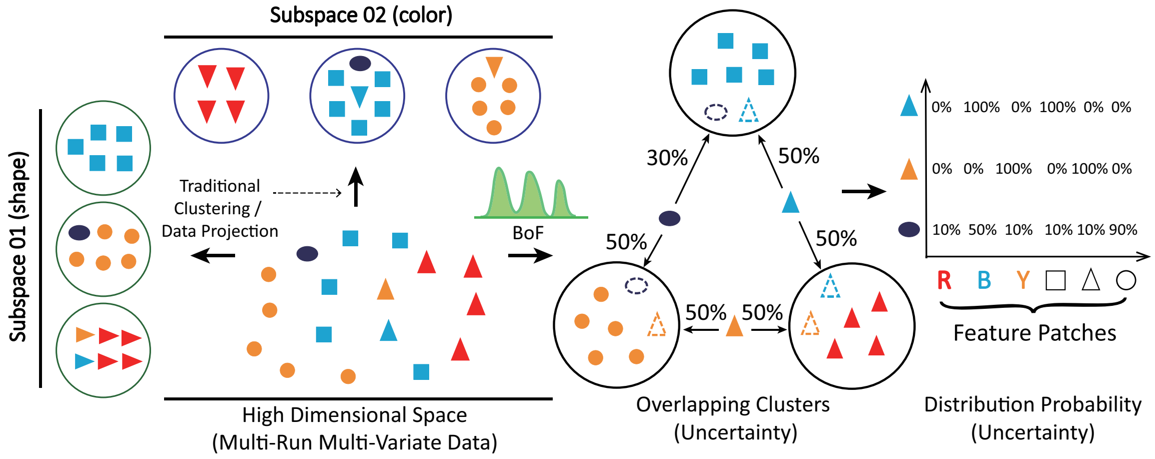


Fig. 1. The illustration of the comparison between bag-of-features and the traditional clustering algorithms. The overlapping set information among clusters and the probability distribution across feature patches can be output by bag-of-features.

study reported that temporal correlation patterns are significant characteristics for analyzing climate simulations. Johnson[3] and Wong et al.[4], summarized that analysis of time-varying features is a top challenge in scientific visualization and visual analysis. Although the time dimension of data is of great concern, direct visualization of spatial–temporal data must still compromise between temporal dimension and spatial domain even now. Thus, in the present study, we couple the multi-faceted aspects of data with time-varying variables. The relationship between time-varying variables is different in individual simulation runs. In some simulation runs, they are intimately correlated with each other. However, in other runs, they may not follow such a pattern. For example, in simulation run A, the change in precipitation is strongly related to that of temperature in a given region, whereas the correlation is relatively weaker in simulation run B. In simulation run C, the change in precipitation is more related to the change in humidity. Thus, one of our goals is to extract temporal correlation patterns from different simulation runs and compare the correlations.

A straightforward method to extract similar temporal correlation patterns from ensemble data is clustering or dimension reduction. Data clustering is essential to compare the differences and similarities of the aggregation of different variables across ensemble runs.This process can be performed using k-series approaches (i.e., k-means[5], k-medoids[6] and k-nearest-neighbor[7]), non-negative matrix factorization (NMF)[8], minimum spanning tree cluster analysis[9], principal component analysis (PCA)[7], density-based spatial clustering of applications with noise[10], and agglomerative hierarchical clustering [11,12,13] to extract ensemble feature. The selection of clustering algorithms is application-specific and is significantly affected by data features[14]. Most existing clustering approaches divide data by classifying data instances with similar features into the same group, as shown in Figure 1 (left).It is an effective way to reveal prominent data patterns. However, several patterns are not salient because of the complexities of the data (e.g., the dark ellipse, yellow triangle and the blue triangle in Figure 1), especially ensemble data characterized as multi-faceted. Moreover, the interrelation between time-varying variables of ensemble data can sometimes reveal the characteristics of the simulation model.In most cases, the relationships between data instances and the clustering groups obey a certain probability distribution instead of a mutual exclusion relationship. Furthermore, all clusters and data instances can be expressed by the probability composition of all feature patches, which includes significant uncertainty information and could be used to compare simulation patterns, as shown in Figure 1. In this paper, we propose a new approach named eBoF based on bag-of-features (BoF), to reveal the temporal change correlations across simulation runs. As an evolution of bag-of-word (BoW)[15] in text analysis, BoF[16] is flourishing in the domains of computer vision and image processing. For example, Li Fei-Fei et al.[17,18] extended BoW to an application-specific BoF that analyzes images of natural scenes. The evolution of data behavior in simulations can be effectively extracted from a group of ensemble runs by eBoF. The extracted feature clusters can further provide insightful and comprehensive evidence on the temporal correlation of data. BoF was chosen over traditional clustering algorithms because of several reasons. **First**, BoF can cluster and classify multiple variables simultaneously by defining all data from multiple target variables as different feature patches. For example, a cluster may correspond to an aggregated region with similar features: “temperature rise,”“humidity rise,” and “precipitation rise”. **Second**, BoF clustering results and all feature patches from different simulation runs obey a certain probability distribution (uncertainty) instead of the mutual exclusion relationship in Figure 1 (right). The set overlapping uncertainties can be further employed to obtain insights into data analysis and comparison because a cluster in BoF is extracted based on almost all feature patches and the weights of all feature patches are associated with their frequencies. **Third**, the flexibility of feature definition allows the feature definition to be changed easily according to the goals and data features. To the best of our knowledge, this study is the first to introduce the BoF model in analyzing ensemble simulation data.

In BoF, an image of a scene can be represented by a collection of local regions. In eBoF, we make a metaphor from high-dimensional image data to ensemble simulation data. The time-varying variables of different runs at a given spatial location are encoded as a BoF. Monotone temporal trends at each location from all target variables,such as, precipitation, temperature, and humidity in climate data, are considered as its feature patch while taking the duration time as the frequency of the patch. The feature clusters in ensemble runs then are achieved and identified based on the similarity in spatial patterns and temporal trends. The extracted clustering results and all feature patches from different simulation runs obey a certain probability distribution, which can be used to derive the data of overlapping clusters (uncertainty). We evaluate the eBoF by performing case studies conducted on two ensemble simulation data, consulting domain experts, and then receiving many positive feedback and suggestions. The evaluation shows that the proposed eBoF can provide insightful and comprehensive evidence on ensemble simulation data.

2 | RELATED WORK

In this section, we discuss related works from two perspectives: the visualization of spatial-temporal simulation data, and the BoF technique.

	Whitaker et al.	Ferstl et al.	Jarema et al.	Höllt et al.	Potter et al.	Bergner et al.	Berger et al.	Matkovic et al.	Matkovic et al.	Doleisch et al.	Kothur et al.	Muigg et al.	Woodring et al.	Hummel et al.	Biswas et al.	Demir et al.	Kumpf et al.	Splechtna et al.	Ferstl et al.	Hazarika et al.	Bruckner et al.	Liu et al.	Bordoloi et al.	Demir et al.	Hao et al.	He et al.	Wang et al.	Harris et al.	Hruschka et al.	Schuller et al.	Csurka et al.	Sivic et al.	Mirzargar et al.	Liu et al.	Fofonov et al.	Jiang and Ngo	Smith et al.	Caicedo et al.	Jegou et al.
SSV	■	■	■	■	■		■	■														■											■						
PSV				■		■	■	■	■	■	■											■												■	■				
DETP				■		■	■	■	■	■		■	■	■																									
K-series					■									■	■	■	■	■																					
PDG	■	■	■							■	■		■	■					■	■	■	■																	
AHC		■	■																■				■	■	■	■	■												
BO"X"																												■	■	■	■	■				■	■	■	■

Table 1 The most related work on multi-space ensemble visualization techniques. The row labels are different techniques on data analysis and visualization, including simulation space (SSV: boxplot-based visual summarization,glyph-based visualization,isocontour and isosurface), parameter space (PSV:trial-and-error exploration,overview-to-detail exploration,focus-and-context exploration, visual steering), and feature space (DET: feature definition, extraction, tracking and exploration, k-series: k-means, k-medoids, k-nearest-neighbor, PDG: PCA, DBSCAN and GMM, AHC: agglomerative hierarchical clustering, BO"X":BoW,BoVW and BoF), where k-series, PDG, AHC could be summarized as data clustering and data reduction.

2.1 | Spatio-Temporal Simulation Data Visualization

Large-scale simulations often generate ensemble data that own time-dependent features (whose analytics is one of the challenges in visualization),and multi-run, multi-variate as well as many other properties.

We categorize the multi-space ensemble data analysis according to the multi-space analysis [2,14] into three types, i.e., simulation space analysis, parameter space analysis, and feature space analysis, as shown in Table 1.

Simulation Space Analysis. Model stability, model fidelity, and spatiotemporal resolution are tightly concerned by domain experts when conducting simulations. To better handle the outliers and missing data during the process of simulation, a lot of techniques including boxplot-based visual summarization, glyph-based visualization, and silhouette-based illustrative rendering emerged for simulation space exploration. The boxplot-based approaches target at getting the trend aggregation results, revealing summary statistics in simulation data, such as uncertainty information. For example, a method named *contour boxplot* [19] *and curve boxplot* [20]is proposed. It mainly presents a generalized method on extending functional data depth to contours and demonstrated methods for displaying the resulting boxplots for 2-D climate simulation data and 2-D computational fluid dynamics.

Glyph-based method is a reasonable choice to visualize the uncertainty, sensitivity, or stainability presented in the whole simulation processes. Regarding 2-D vector field ensembles, a glyph-based technique [21] encodes the variation modes into shapes

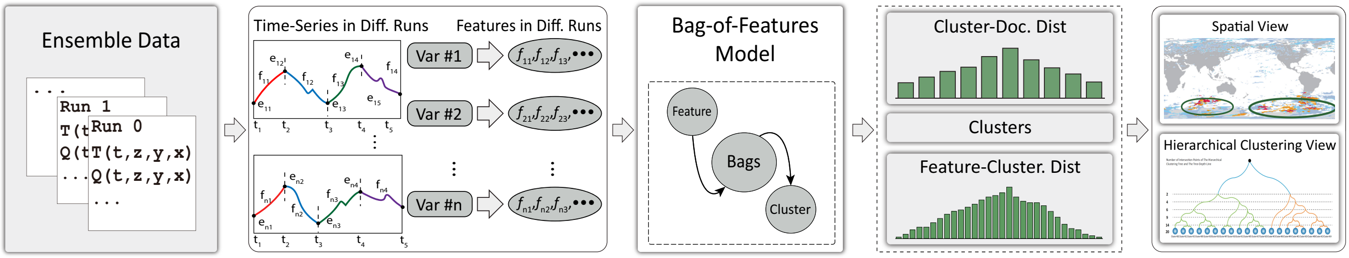


Fig. 2. The pipeline of the eBoF. Monotone intervals are extracted from the target time-varying variables of ensemble data. They are then encoded into feature patches, which are the input of the eBoF model. The uncertain overlapping information of the inter-cluster and the inter-run will be visualized and analyzed in the designed set-based visualization.

and directions of glyphs. As an illustration, a visualization system developed by Hollt et al.[22] that allows users to specify a critical height value in their glyph design. Besides, streamline variability plots [12], was proposed as a more flexible tool, i.e.,for visualizing the statistical properties. What's more,a LCSS-based measurement was proposed by Liu et al.[10] to compute ensemble pathlines similarity field and uncertainty field.

Parameter Space Analysis. Large-scale state parameters are generated by using the same simulation model with various values of control parameters[23]. Considering that the inputs could be multi-variate, high-dimensional parameter space exploration could be tedious and time-consuming even for the experts. Related techniques could be summarized as trial-and-error, focus and context (F+C), overview-to-detail, and visual steering.

Trail-and-error method is seldom used due to its inconvenience and inefficiency. Focus and context (F+C) performs much better than trail-and-error method when handling large data sets with the nature of high dimensionality. Matkovic et al.[24] propose a parallel coordinates plot (PCP)-based parameter space exploration technique that allows users to brush the parameter space in PCP by using an F+C scheme. The overview-to-detail method follows the principle: "Overview first, zoom and filter, then details on demand". For example, ParaGlide[25] is designed for interactive exploration of parameter spaces (overview) of multi-dimensional simulation models, which endeavors to facilitate the process of refining models by guiding data generation using a region-based interface for parameter sampling (details). Besides, visual steering could help domain experts explore the parameter space interactively[26]. Glyph-based approach is used for visual mapping to address the challenges posed by spatio-temporal simulation data visualization. Glyph is often employed to vector field ensemble visualization which was limited to a 2-D scenario. [27,28] Glyph is also often used to make a metaphor in simulation space [22].Fofonov and Linsen [29] generalized the isosurface similarity to a field similarity and further to a multi-field similarity.

Feature Space Analysis. Feature definition, extraction and tracking are essential for ensemble data before exploring feature space in simulation processes. Techniques like feature definition language (FDL)[30,31], hierarchical merge tree and Reeb graph[32] are widely used in this process. While data procession and reduction are significant when comparing the differences and similarities of the aggregation of data sets. By using methods like sampling, clustering, classification and reduction, time-varying behavior characteristics, major trends and outliers in data could be captured[33,34]. For example, principal components analysis (PCA) technique is widely adopted in ensemble feature extractions[7,12,35]. For example, Ferstl et al.[12] use PCA to convert streamlines into a structure preserving Euclidean space. Furthermore, a lot of approaches are implemented by feature extraction algorithms like k-series in the work[5,36,6], and the work based on DBSCAN[10,26], AHC[12,13,37,38,39,40] and topic modeling[41,42,43,44,45,46,47,48,49,50,51,52,53].

2.2 | Bag-of-Features Techniques

BoW[15] becomes a solid foundation of many work in domains like text mining[54], text categorization, information retrieval and speech emotion recognition[55], etc. A base method of BoF is BoW, which is routinely employed in text processing. Furthermore, bag-of-visual-words (BoVW)[56] represent each image by a histogram of the visual words[57]. Some works[58,59] in fields like computer vision use BoW to refer to BoVW, too. As an evolution of BoW, bag-of-features (BoF)[16] is flourishing in computer vision and many other domains. BoF is a resilient method that models image structure robustly[60], where the representative patterns are codified as visual feature, quantize the observations into a set of feature patches, while the frequency vector of such bag is calculated for presentation[61,60].

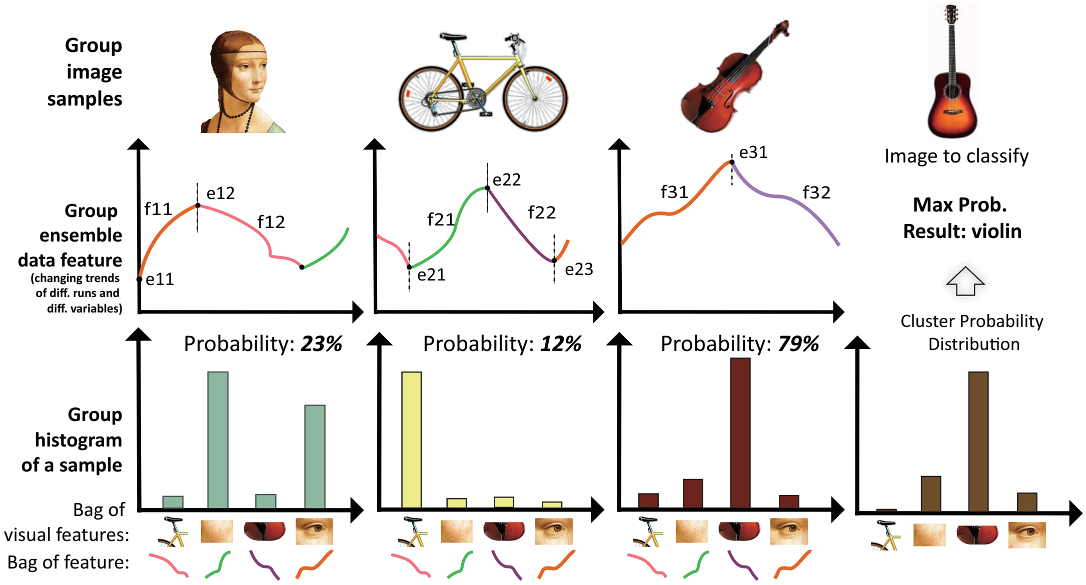


Fig. 3. Image samples can be expressed by local feature patches in BoF, where each local feature with a simple and single semantic can be defined as a feature patch. Similarly, we take a simple and monotone interval from all target variables of ensemble scalar data as a local feature patch. Image courtesy of Svetlana Lazebnik and Fei-Fei Li.

3 | MODEL DESIGN AND FEATURE ENCODING

That three or four variables generally in an ensemble data could be selected as the *target variables*, even though the eBoF itself is not limited by the number of variables. We can choose three or four variables to compare the correct conclusions. Two criteria are met when choosing variables: firstly, it is easy to express numerically, and secondly, it has a great influence on the result. Our hierarchical clustering is mainly about spatial correlation. The value of each voxel in the whole space is subtracted, and then clustering.

In this section, we introduce the proposed eBoF model. The pipeline of the approach is shown in Figure 2. Data features are extracted from the time-varying target variables of the raw ensemble scalar data.

We then encode all the time-varying variables from different runs into the feature patches and take them as the input of our eBoF model. The clustering results and all feature patches from different simulation runs obey a certain probability distribution, which can be used to derive the data of overlapping clusters (uncertainty).

3.1 | Overview and General Idea of the eBoF Model

The BoF model has been widely used in image feature extraction, which comes from BoW in text analysis. An image of a scene in BoF can be represented by a collection of local image patches, as shown in Figure 3. We create a metaphor from high-dimensional image data to ensemble simulation data. The time-varying variables of different runs at a given spatial location are encoded as a BoF. The high-dimensional image samples (or vectors) can be expressed by a collection of local feature patches in BoF (Figure 3), where each local feature with a simple and single semantic can be defined as a feature patch. Similarly, we take a simple and monotone interval (temporal trend) from all target variables of ensemble scalar data as a local feature patch and the duration time of each interval (patch) as its frequency (occurrence probability). The feature clusters in ensemble runs are then identified based on the similarity of temporal correlations, which correspond to a bag of interrelated feature patches in BoF. For example, given two simulation runs run_i and run_j of an ensemble data, if a dependent variable D has strong correlations with an independent variable I in the same period, such as from Feb. to May if the time resolution is month, the similarity value of run_i and run_j would be increased to some extent.

3.2 Feature Definition and Encoding

We define the feature patch as an interval of monotone continuous rising or falling of a variable over time. Suppose we have m ($m > 1$) ensemble runs,

$$R = \left\{ R_1, R_2, ..., R_m \right\}, \tag{1}$$

and in each run R_i, there are n ($n \geq 1$) target variables,

$$V_i = \left\{ V_{i,i}, V_{i,2}, ..., V_{i,n} \right\}, where\ 1 \leq i \leq m. \tag{2}$$

For each variable $V_{i,j}$ at certain location, we can find l ($l > 1$) extrema over time steps t_1, t_2,..., t_l (including the values v_{i,j_1} and v_{i,j_T} in the first and last time steps, respectively), i.e.,

$$E_{i,j} = \left\{ e_{i,j_{t_1}}, e_{i,j_{t_2}}, ..., e_{i,j_{t_l}} \right\}, \tag{3}$$

where $e_{i,j_{t_1}} = v_{i,j_1}, e_{i,j_{t_l}} = v_{i,j_T}, 1 \leq j \leq n$. The feature can be extracted as the interval between two neighboring extrema, formally

$$f_{i,j_{t_k}} = \left[e_{i,j_{t_k}}, e_{i,j_{t_{k+1}}} \right], where\ 1 \leq k \leq l. \tag{4}$$

Then we obtain a group of features from variable $V_{i,j}$:

$$F_{i,j} = \left\{ f_{i,j_{t_1}}, f_{i,j_{t_2}}, ..., f_{i,j_{t_{l-1}}} \right\}. \tag{5}$$

Considering all variables of all runs involves a feature bag, the feature set can be finally defined as

$$F = \bigcup_{i=1}^{m} \bigcup_{j=1}^{n} F_{i,j}, where\ 1 \leq i \leq m, 1 \leq j \leq n. \tag{6}$$

The key to extracting features is to detect local extrema for each variable at each location. We extract the monotone intervals according to the variable values over time. Each interval that contains more than two continuous time steps is considered a valid feature patch. Otherwise it is considered an outlier, as shown in the outlier in Figure 4 (a), which is circled in red. It illustrates the extraction of maximum or minimum points from a time-varying variable. The outlier is important in some application-specific ensemble simulation model analyses. However, the goal of eBoF is to analyze a set of continuous temporal trends and extract temporal correlations presented in multiple variables.

Two factors determine whether or not multiple features can be considered as an identical feature patch: the start time step and the average change rate of features. Features that originate from different runs or locations may be labeled as the same feature patch, thereby allowing run comparisons on temporal correlations. It means that we will encode two feature patches as different values when their start time-steps are not identical. The rational is that features with different start time-steps may have different meanings from a physical point of view. But when the same feature is compared, it starts at the same time, so it can be compared at the same time. Feature patches are also distinguished based on the average change rate to support the analysis of temporal trends for multiple target variables and then obtain their temporal correlations.

Suppose the duration time of a feature patch is [t$_i$, t$_j$], the average changing rate is defined as the slope of the feature curve, i.e.,

$$r = \sum_{x=i}^{j-1} (v_{x+1} - v_x), where\ j > i + 1. \tag{7}$$

The whole range of r is [-90°,90°]. We subdivide this range into several histogram bins. Figure 4 (b-c) illustrate the classification of features.Figure 4 (b) shows all feature classifications. The green arrow corresponds to the feature patch in Figure 4 (c),and it does not show all the feature classification in Figure 4 (b).

We use the adaptive method to divide the angle range. The adaptive method automatically adjusts the boundary or constraint conditions of the angle range according to the data characteristics of the processed data during processing and analysis. This method adapts to the statistical distribution and structural characteristics of the processed data to achieve the best processing effect. Considering that more features appear when the angle is close to 0°, we divide the angular bins with more features. Thus, the number of features belonging to each bin is nearly identical.

The feature patches are encoded into an identical value if they have the same start time-steps and fall into the same bin. For example, Figure 4 (a) shows that the four features are encoded into different values because they have different start time-steps. The feature patch frequency in BoF is further defined as the number of continuous time steps. It can also be considered as the weight of the feature patch in the eBoF model. The rational is that the weight of the temporal trends should be increased in eBoF if the duration time is prolonged.

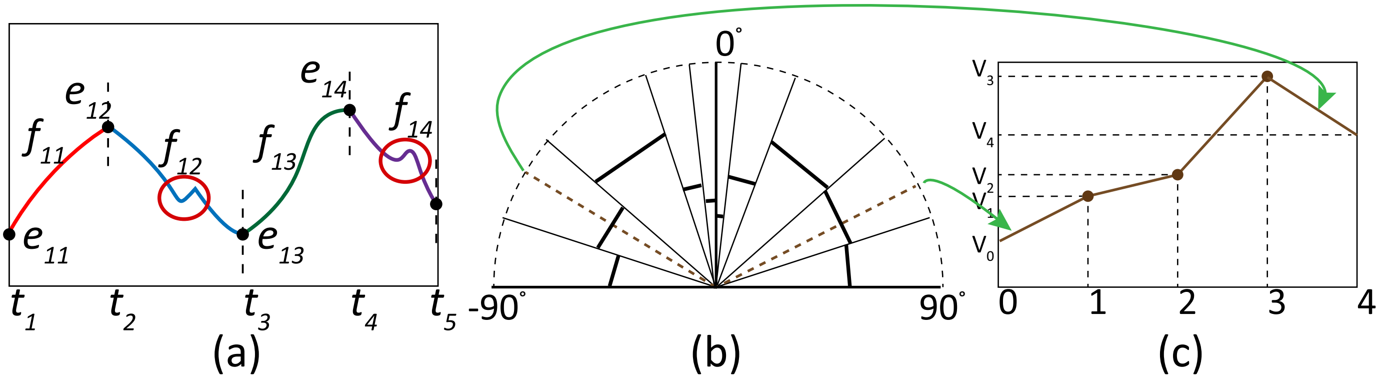


Fig. 4. Feature patch extraction and definition. (a) We extract one monotone interval as a feature patch. An interval that contains more than two continuous time steps would be considered as a feature patch. (b) The range of the average change rate is subdivided into multiple bins adaptively according to the angular histogram of change rate; (c) Valid intervals will be put into one of the bins and then be encoded into different feature patches.

Spatial View. We provide a spatial view to show the regional pattern clustering of a given cluster. The spatial view shows the regional pattern clustering of a specified feature cluster by using the fuzzy probabilities generated from the eBoF model. We extract the probability of the given cluster from all grid points and then visualize the distribution of the cluster over the entire spatial domain in the geographical space. We use a color mapping scheme to visualize regions with different probabilities. Then the clusters of regions with similar probabilities are clearly displayed. From this view, some interesting spatial patterns could be discovered when conducting the run comparison. It represents that multiple runs have similar temporal correlation patterns over some spatial regions.

Hierarchical View. A hierarchical tree is designed to decrease the sensitivity of eBoF to the initial number of clusters because the clusters with close distance can be automatically merged by the hierarchical merging algorithm of eBoF. In specific, we define the distance between the two feature clusters according to the probability values of their key feature patches and the spatial distribution values. We then build a hierarchical clustering tree between the feature clusters based on their distances, providing a reference for users to select similar feature clusters for merging.

In the hierarchical view, we construct a hierarchical tree that can be divided into several levels to explore the hierarchical relationship between clusters. Feature clusters close to each other are recommended to be combined into new feature clusters, and these new clusters will be set on a higher level of the tree depth. Therefore, different depths of the tree indicate the different degrees these clusters can be merged. The distance of the original spatial clusters can be calculated with the following formula.

$$d = \sum_{i=1}^{N} (c(a) - c(b))^2 \tag{8}$$

The earth is divided into more than 10,000 grids, N represents the number of grids, and d represents the distance between the original spatial cluster a and the original spatial cluster b. Function c(a) represents the number of the original spatial cluster a in a certain grid. The function is to perform hierarchical clustering of these spatial clustering results. It is equals the distance between the classes, and the hierarchical clustering is performed again between the classes. Function c(b) represents the number of the original spatial cluster b in a certain grid.

The detailed descriptions about the distance calculation between clusters are shown in Algorithm **??**. Nevertheless, the eBoF can only alleviate the occlusion as much as possible instead of removing them completely. The balloons will move up and down iteratively until the algorithm converges to avoid as much overlapping as possible, when users drag the any one balloon in the Cluster Comparison View or the Run Comparison View. Recalculate and iterate the energy value of the pixel points on the line until the balloon position is stable.

The balloons, e.g., the clusters in the Cluster Comparison View or the ensemble runs in the Run Comparison View, are merged automatically according to the result of hierarchical clustering. The distance of which is defined as the probability differences of two clusters for all the grids. Besides, it is also allowed to merge the clusters in the two views manually as mentioned above. Finally, the glyphs in the merged balloons can be also merged automatically. The balloon merging algorithm is described in Algorithm **??**.

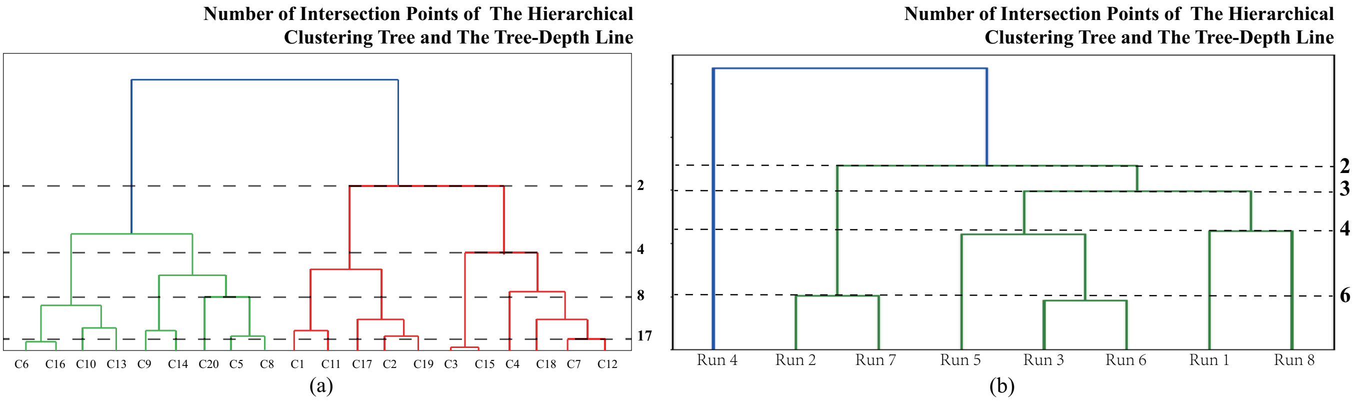


Fig. 5. The hierarchical relationship between clusters (a) and simulation runs (b) of the dataset GEOS-5. Different clusters or runs have different temporal correlations. Clusters or runs close to each other are recommended to be merged into a new node in the tree.

4 | RESULTS AND EVALUATIONS

We evaluate the proposed eBoF on two time-varying ensemble datasets, i.e., GEOS-5 (The Goddard Earth Observing System Model, Version 5)[62] and MOZART-4 (Model of Ozone and Related Tracers, Version 4)[63]. The experiments are performed on the computer with the configuration of Intel Core i7-8550U CPU operating at 1.80 GHz and 8 GB RAM.

Dataset	Dim	#Runs	#IC	HHT	t_p	t_{BoF}
GOES-5	288×182×72×24	8	20	7	7.2	443.6
MOZART-4	192×96×28×24	8	15	10	2.4	49.9

Table 2 Average timing results (in seconds) of the two ensemble datasets in our experiments, where “Dim" is the dimensional size of the dataset, “IC" is the initial number of clusters of eBoF, “HHT" is the height of the hierarchical tree t_p and t_{BoF} are the running time of data preprocessing and BoF model execution, respectively. All the timing results are the averaging results of three tests.

We test the performance of eBoF for the model, as shown in Table 2. Four different step sizes are taken to compute the average timing results. Each result comes from the average performance over three tests. It takes more time for GEOS-5 case than MOZART-4 case to preprocess data and run the BoF model to generate initial clusters because the data size and resolution of the previous dataset are much larger.

4.1 | Case Study of GEOS-5 Dataset

The GEOS-5 simulation data from an atmospheric model provided by NASA Goddard Space Flight Center. This model has a horizontal resolution of 1°×1.25° lat-lon grid and 72 vertical pressure levels transitioning from 1 atm (near the terrain surface) to 0.01 hPa (near 80 km). Previously, scientists performed an eight-run atmospheric CO2 ensemble simulation and generated an output consisting of 24 monthly average data (from January 2000 to December 2001). The total size of the dataset for eight runs is about 76 GB.

In this case, we take three scalar variables, including temperature (T), specific humidity (Q), and total precipitation (TPREC) as target variables, to compare the clusters and simulation runs. They are the variables domain experts are interested in. The initial cluster number of GEOS-5 is set as 20. eBoF is not sensitive to the initial number of clusters because the clusters with close distance can be automatically merged by the hierarchical merging scheme. It also allows users to merge the clusters manually according to their domain knowledge.

A hierarchical clustering and a hierarchical tree merging scheme are used to merge the data clusters. Clusters with close distance can be automatically merged according to the specific tree depth. The distance between clusters C17 and C19 in Figure 5

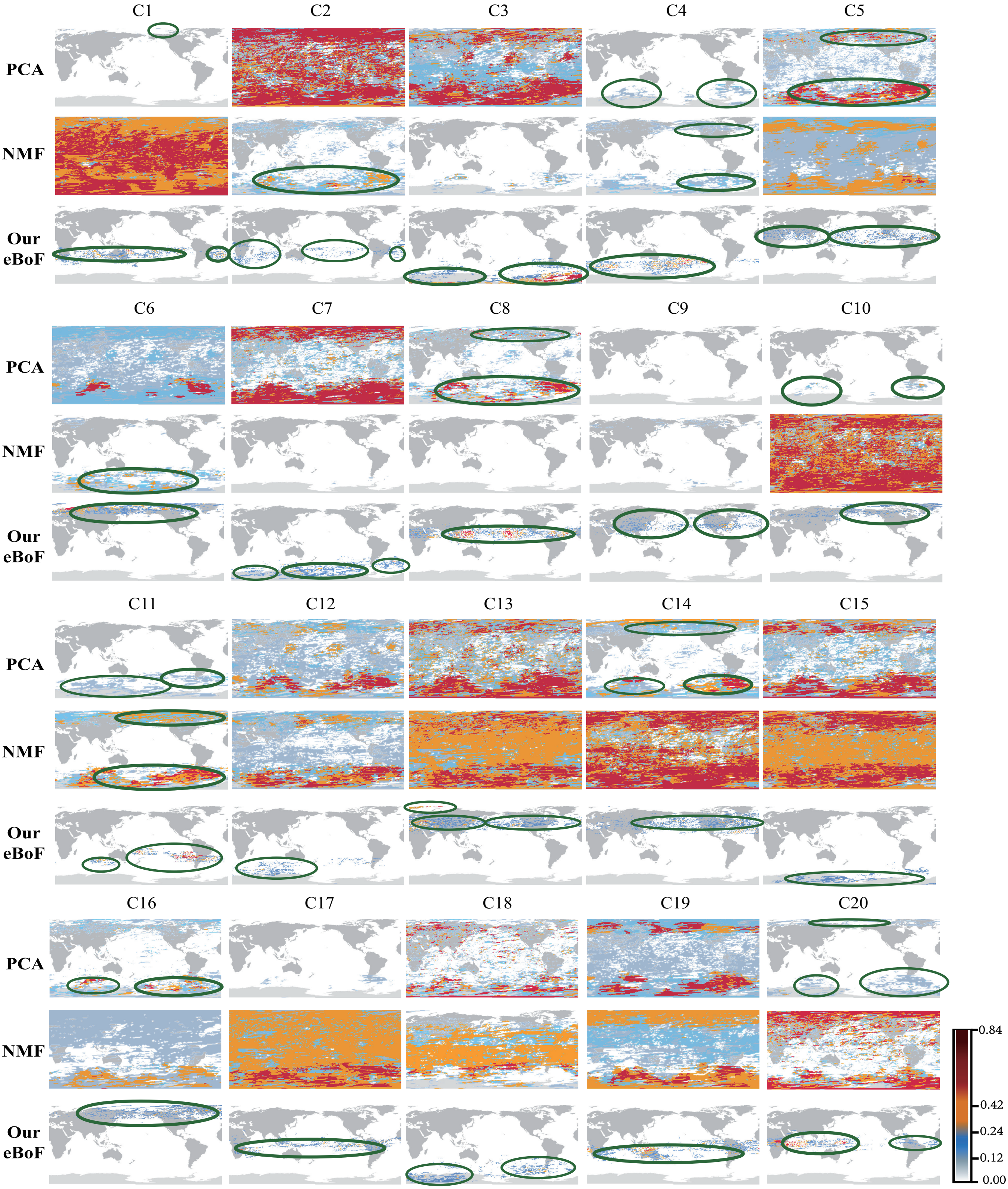
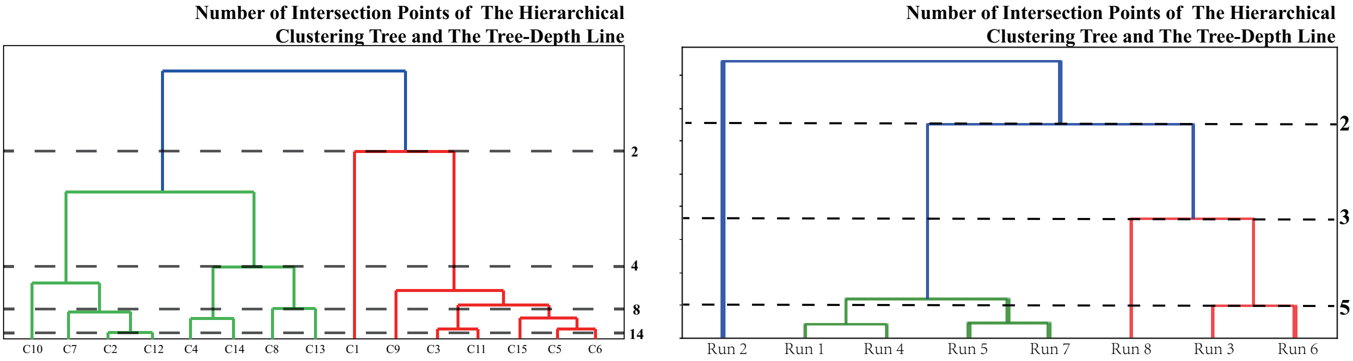


Fig. 6. The spatial view results of temporal correlation clustering for the three target variables (temperature, humidity and precipitation) in the GEOS-5 data. It is more reasonable for the results output by the proposed eBoF according to the feedback of domain experts, because more temporal change correlation patterns in a cluster distributed in some concentrated local regions. We found that similar patterns in C2, C3, C5, C7, C12, C13, C14, C15, C18 and C19 extracted by PCA are distributed in both the Northern Hemisphere and the Southern Hemisphere. Similarly, similar patterns in C1, C4, C10, C11, C12, C13, C14, C15, C16, C17, C18, C19 and C20 extracted by NMF are distributed in both the Northern Hemisphere and the Southern Hemisphere. The value in the legend means the occurrence probability of the cluster run at the corresponding location on the map. According to the legend, the probabilities from 0.0 to 0.12 is in gray, 0.12 to 0.24 is in blue, 0.24 to 0.42 is in orange, and 0.42 to 0.84 is in red.



(a)

(b)

Fig. 7. The hierarchical relationship between clusters (a) and simulation runs (b) of the dataset MOZART-4. Different clusters or runs have different temporal correlations. Clusters or runs close to each other are recommended to be merged into a new node in the tree.

(a) is the closest one. The automatic merging algorithm merges these two clusters first. The tree depth and the cluster number are decreased by one to be 17. Then, C6 and C16, C5 and C8, C3 and C15, and C7 and C12 are also relatively close. They are the candidate merged pairs in the next merging step. Thus, we can select an optimum cluster number according to the tree depth of the hierarchical clustering tree, which indicates that the clustering of eBoF is not sensitive to the initial number of clusters.

The user can select the variable of interest by selecting the cluster level. The number next to the dotted line represents the level of the cluster, and the number of the number represents the level. For example, the number next to a dashed line is 4, indicating that there are 4 clusters under this line. C5 and C8 are in the same green line, indicating that the two can be in one class. C20, C5, and C8 can be placed together indicating that these three can be placed in one class.

In addition, the three target variables in the spatial regions corresponding to the low-level nodes are strongly interrelated to each other in a given time range.

Domain scientists believe that the proposed BoF enables them to obtain an insight into seasonal change patterns in this case. The seasonal change patterns are interrelated for the three target variables in the same time range. Their similarities and differences are determined according to their location on the map.

For example, similar seasonal change patterns can be found in C9 and C14 because they are concentrated in a local region in the northern hemisphere as shown in Figure 6. These two groups of merged clusters are generated according to the hierarchical tree merging scheme. All these 20 clusters show high interrelationships between the three target variables. In simulation run merging, the hierarchical tree (b) is used to merge the simulation pair of Run #2 and Run #7, Run #3 and Run#6, and Run #1 and Run#8. The order of the simulation run merging is largely determined by correlation patterns, which are feature patches in eBoF.

4.2 | Case Study of the MOZART-4 Dataset

The second dataset is a simulation output from the MOZART-4 model. To evaluate ozone formation under different anthropogenic emission scenarios, scientists conduct a series of perturbation experiments. The simulation results comprise eight simulation runs, including two reference runs and six perturbation runs. The six perturbation runs are conducted by alternatively switching off human pollutants from six Eurasian regions. Those source emission regions include Europe (related to EU run), India (IN run), Middle East (ME run), southeast Asia (SA run), eastern Asia (EA run), and mid-Asia (MA run). The remaining two runs are generated for reference. With such an experimental design, scientists would like to investigate the relative importance from different regions to the global ozone concentration. The data have 24 monthly time steps from Jan 2000 to Dec 2001. The simulation has a spatial resolution of 192 × 96, with 28 vertical layers. The total data size is about 13 GB.

In this case, the three target variables we focused on are ozone (O_3), nitric oxide (NO) and oxygen (O_2), which also attracted intensive attention from experts. The initial number of feature clusters is set as 15. As mentioned before, eBoF is not sensitive to the initial number of clusters because of the automatic merging function supported by the hierarchical tree merging scheme and the manual merging function.

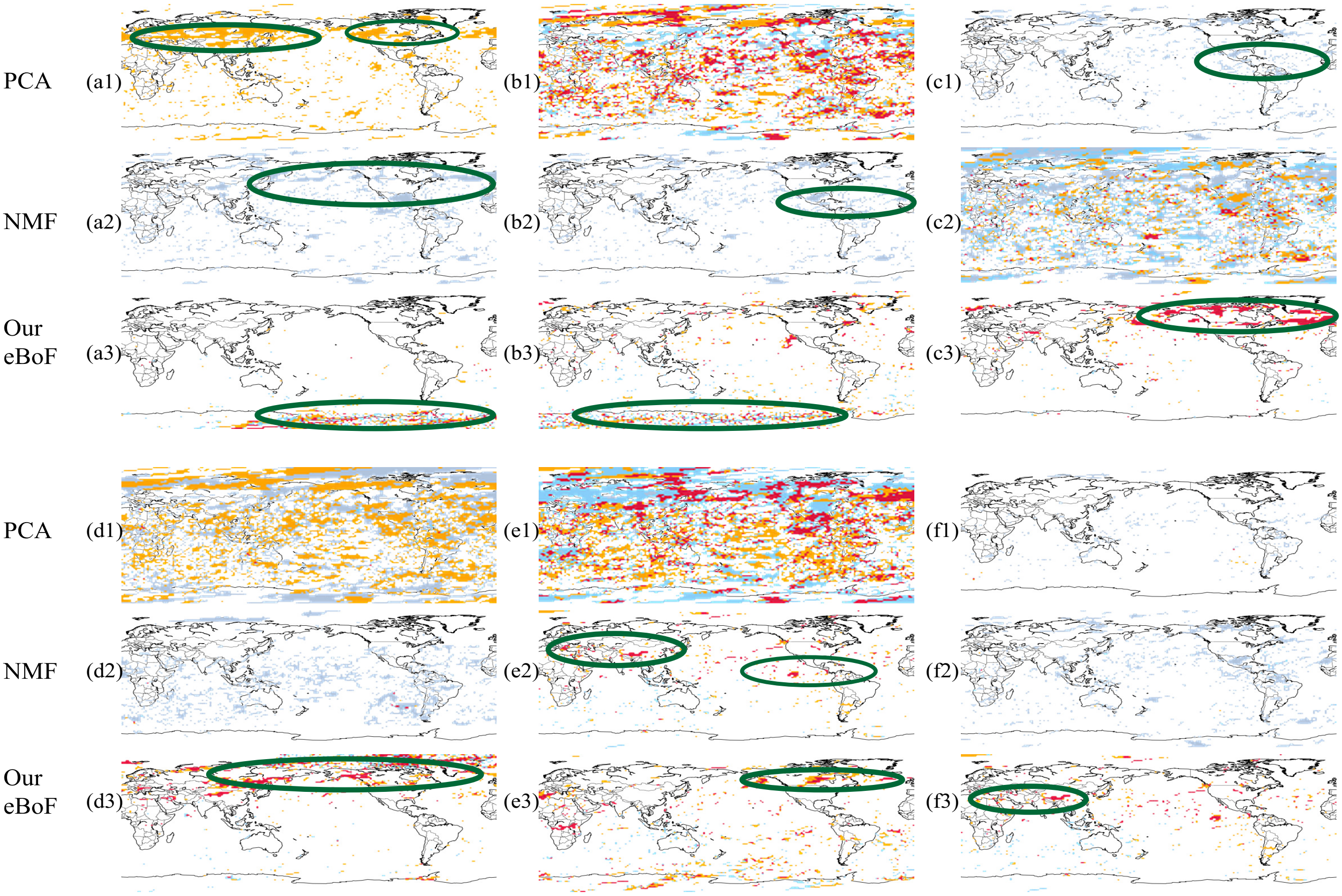


Fig. 8. The six representative clusters in the clustering space for the three target variables (i.e., nitric oxide, oxygen and ozone) in the MOZART-4 data. It is more reasonable for the results output by the proposed eBoF according to the feedback of domain experts, because more temporal change correlation patterns in a cluster distributed in some concentrated local regions.

The hierarchical clustering and the hierarchical tree merging scheme can be used to merge the data clusters, as shown in Figure 7 (a). According to a given tree depth, clusters with close distance can be automatically merged. For example, the distance between clusters C2 and C12 is the smallest one. Thus, these two clusters would be merged first. The tree depth and the cluster number would be decreased by one to be 14. Then, C4 and C14, C8 and C13, and C3 and C11 would also be considered as relatively close pairs, and they would be recommended as the candidate merged pairs of the next merging. Thus the optimum cluster number could be conjectured according to the tree depth of the hierarchical clustering tree, which again indicates that eBoF is not sensitive to the initial number of clusters. With the guidance of the hierarchical clustering tree, the automatical merging result with a recommended optimal merging level can be achieved. Similar to the GEOS-5 data, the clusters could also be manually merged according to the domain knowledge.

Furthermore, in the cluster pairs of C2 and C12, C4 and C14, C3 and C11, and C5 and C6, the three target variables have a very close relationship with each other, as shown in Figure 7 (a). The trend and level of their changes are approximately the same. For example, the levels of all variables are higher or lower in the first year, and their changes are minimal in the second year. This result can be used to help domain experts obtain insights into their change patterns in different countries.

Nevertheless, Run #2 is not recommended to be combined with any other runs in this case because it is too far from any other ones, as illustrated in the hierarchical tree Figure 7 (b). It is considered as a large branch with significant temporal irrelevance, which means that the temporal correlation of the three variables in this run is distinct from the patterns in other runs.

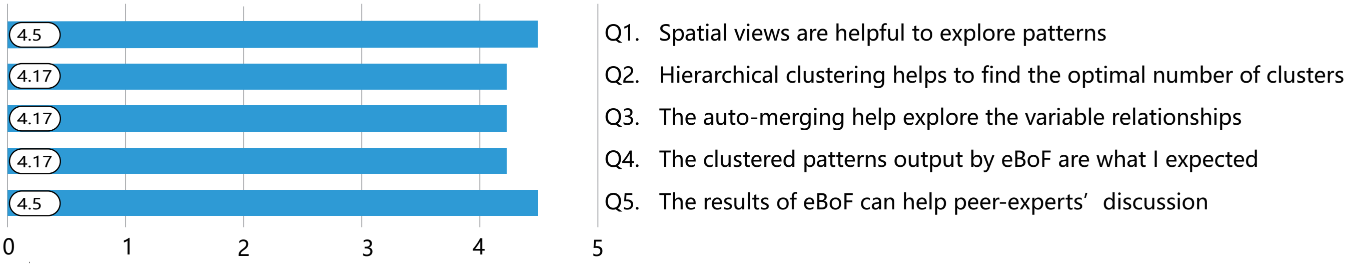


Fig. 9. The questionnaire results from domain experts.

4.3 | Evaluations on Method Comparison

We also apply the traditional dimensionality reduction algorithm in GEOS-5 and MOZART-4 data for comparison. PCA and NMF are commonly used methods of dimensionality reduction. It is usually used to explore and visualize high-dimensional datasets. It can also be used for data compression and preprocessing. PCA can synthesize correlated high-dimensional variables into linearly independent low-dimensional variables, called principal components. The principal component can retain the information of the original data as much as possible.

That the patterns in spatial views generated by NMF and PCA are dispersedly distributed, which means the distribution of one cluster is not concentrated in some regions, as is shown in Figure 6 and Figure 8. Nevertheless, the clusters obtained by the proposed eBoF are more concentrated, which is more reasonable and in accordance with our common sense. As shown in the green circle in the picture, the color is concentrated in a certain place on the map, indicating that the results can be clustered.

Similar patterns in C2, C3, C5, C7, C12, C13, C14, C15, C18 and C19 of GEOS-5 extracted by PCA are distributed in the northern and southern hemisphere, as shown in the first row in Figure 6. Similar patterns in C1, C4, C10, C11, C12, C13, C14, C15, C16, C17, C18, C19, and C20 extracted by NMF are distributed in both the northern and the southern hemispheres, as shown in the second row in Figure 6. However, sharing similar precipitation patterns in the northern and southern hemispheres in a given time range is almost impossible according to the domain knowledge. The reason is that the temporal characteristics of variables will be lost through dimensionality reductions (a set of matrix operations) in NMF and PCA, whereas PCA and NMF achieve topic extraction through dimensionality reduction. Through dimensionality reduction technology, multiple variables are converted into a few principal components, which can reflect most of the information of the original variables. Therefore, the spatial information presented in the clusters extracted by PCA and NMF is unexpected.

The pattern distribution extracted by the three methods for the three target variables (i.e., the dependent variable ozone and the independent variables oxide and oxygen) of the MOZART-4 data are shown in Figure 8. We can find the several clustered patterns extracted by PCA (the first row) and NMF (the second row) are scattered globally, e.g., Figure 8 (b1), Figure 8 (d1) and Figure 8 (e1) for PCA and Figure 8 (c2) for NMF.

4.4 | Domain Expert Feedback

We have consulted domain experts in climate studies and received considerable positive feedback. All experts are familiar with the two data of GEOS-5 and MOZART-4. Two of them provided oral feedback, and three of them used questionnaires.

According to the questionnaire in Figure 9, our results are generally in accordance with the background knowledge of the domain experts. The results of the questionnaire are as follows: According to Q1 (the average score is 4.5), it can be seen that the designed spatial view can be used to explore the extracted patterns. According to Q2 (the average score is 4.17) and Q3 (the average score is also 4.17), most domain experts think that the hierarchical clustering tree helps them achieve the number of clusters. Most domain experts believe that exploring the relationship between multiple sets of data is helpful. According to Q4 (the average score is 4.17), the clustering results output by eBoF are the same as expected. Finally, most domain experts believe the results generated by eBoF can help peer-experts’ discussion according to Q5 (the average score is 4.5).

We also received some oral feedback from several meteorological domain experts who have engaged in ensemble simulation and ensemble data analysis for a long time. The experts agree that the analysis of these aggregation patterns is useful, especially the analysis of trend aggregation over time. They also deem that analyzing the degree of correlation of the aggregation pattern (hierarchical clustering view) is reasonable. This macroscopic relationship is certainly presented, and any aggregation pattern is not necessarily right or wrong.

Regarding the GEOS-5 case, domain experts are interested in the degree of temporal correlation analysis for the three variables in different simulation runs, namely, temperature, humidity, and precipitation. Regarding the Mozart-4 case, the domain experts are also interested in the temporal correlation analysis for the three selected variables. However, the claims of the visualization cannot be determined to be correct or not because the atmospheric cycle is complex. Therefore, various macroscopic relationships exist, and determining which aggregation pattern is necessarily right or wrong is not directly possible, although they believe the overall results are in accordance with their domain knowledge.

The experts also provided some suggestions. They believe the design (i.e., the encoding scheme of eBoF) enables them to see some seasonal changes patterns in the northern and southern hemispheres in the GEOS-5 data. In particular, they would like to understand further the correlation comparisons between the current time-step and all the averaging time-steps. In this way, they can see some distinctive correlations of the climate patterns and can conduct better research in their future work. For example, for a variable (e.g., precipitation), we can encode different values (D-values) of all target temporal variables into the intervals. The D-value is the original values minus the averages of all the years in the dataset. Analyzing the impact of independent variables on dependent variables could be useful in the view of outlier precipitation, which is a follow-up work discussed in Section 5.

5 | DISCUSSION

The proposed eBoF is designed to extract and reveal the temporal correlation patterns in the multi-run, multivariate and spatio-temporal data. However, the work has some limitations, which we plan to address in the follow-up work:

The target variables selected for each dataset could be interesting to domain experts, closely relevant variables are also preferable. For example, domain experts expect to analyze variables relevant to the precipitation patterns in the GEOS-5 data, and the related impact factor on O_3 in the MOZART-4 dataset. Even though we select only three variables to evaluate eBoF, the eBoF itself is not limited by the number of variables. However, finding the independent variables to the dependent variables is inconvenient for domain experts if too many variables are selected.

Domain experts believe the design (i.e., the encoding scheme of eBoF) enables them to see some seasonal changes patterns. In particular, they would like to understand further the correlation comparisons of D-value (see Section 3.2) to identify some distinctive correlations of climate patterns.

The number of clusters selected in the hierarchical tree depends on the domain knowledge. The current selection criterion is a simple trial-and-error method. Users can select different tree depths to a specific cluster number according to the result of the spatial view.

Finally, we plan to design an application-specific overlapping information visualization such as setting visualization[64] to illustrate the intermediate steps of fuzzy clustering, because the evolution of overlapping information is also significant for the domain experts to see the inter-correlations of target variables and ensemble simulation runs.

6 | CONCLUSIONS

We present a new approach eBoF by using BoF, a commonly-used clustering algorithm in computer vision, to reveal the temporal correlations of multi-variate across multiple simulation runs. The time-varying variables of different ensemble runs at a given spatial location are encoded as a BoF. Monotone temporal trends from all target variables at each location are encoded into the feature patch while taking their duration time as the frequency. The feature clusters in ensemble runs then are achieved and identified based on the similarity in spatial patterns and temporal trends.

The probability distribution (overlapping information) across different clusters can help generate reasonable clustering results according to domain knowledge. The BoF algorithm itself can retain the geo-spatial information that is often lost in the traditional clustering algorithms. Furthermore, we design a hierarchical clustering tree to help users find an optimal merging level and then obtain the number of clusters. We demonstrate the results by conducting case studies on two ensemble simulation datasets. Results suggest that the proposed eBoF can further provide insightful and comprehensive evidence on ensemble simulation data according to the feedback of domain experts.

References

1. Guo H, Yuan X, Huang J, Zhu X. Coupled Ensemble Flow Line Advection and Analysis. *IEEE Transactions on Visualization and Computer Graphics* 2013; 19(12): 2733–2742.

2. Chen X, Shen L, Sha Z, et al. A survey of multi-space techniques in spatio-temporal simulation data visualization. *Visual Informatics* 2019; 3(3): 129–139.

3. Johnson CR. Top Scientific Visualization Research Problems. *IEEE Computer Graphics and Applications* 2004; 24(4): 13-17.

4. Wong PC, Shen HW, Johnson CR, Chen C, Ross RB. The Top 10 Challenges in Extreme-Scale Visual Analytics. *IEEE Computer Graphics and Applications* 2014; 32(4): 63-67.

5. Biswas A, Lin G, Liu X, Shen HW. Visualization of Time-Varying Weather Ensembles across Multiple Resolutions. *IEEE Transactions on Visualization and Computer Graphics* 2017; 23(1): 841–850.

6. Potter K, Kniss J, Riesenfeld R, Johnson CR. Visualizing summary statistics and uncertainty. In: . 29. ; 2010: 823–832.

7. Hummel M, Obermaier H, Garth C, Joy KI. Comparative visual analysis of Lagrangian transport in CFD ensembles. *IEEE Transactions on Visualization and Computer Graphics* 2013; 19(12): 2743–2752.

8. Choo J, Lee C, Reddy CK, Park H. UTOPIAN: User-Driven Topic Modeling Based on Interactive Nonnegative Matrix Factorization. *IEEE Transactions on Visualization and Computer Graphics* 2013; 19(12): 1992–2001.

9. Obermaier H, Bensema K, Joy KI. Visual Trends Analysis in Time-Varying Ensembles. *IEEE Transactions on Visualization and Computer Graphics* 2016; 22(10): 2331–2342.

10. Liu R, Guo H, Zhang J, Yuan X. Comparative Visualization of Vector Field Ensembles Based on Longest Common Subsequence. In: ; 2016: 96–103.

11. Wang J, Liu X, Shen HW, Lin G. Multi-Resolution Climate Ensemble Parameter Analysis with Nested Parallel Coordinates Plots. *IEEE Transactions on Visualization and Computer Graphics* 2017; 23(1): 81–90.

12. Ferstl F, Bürger K, Westermann R. Streamline variability plots for characterizing the uncertainty in vector field ensembles. *IEEE Transactions on Visualization and Computer Graphics* 2016; 22(1): 767–776.

13. Ferstl F, Kanzler M, Rautenhaus M, Westermann R. Visual Analysis of Spatial Variability and Global Correlations in Ensembles of Iso-Contours. *Computer Graphics Forum* 2016; 35(3): 221–230.

14. Wang J, Hazarika S, Li C, Shen HW. Visualization and Visual Analysisof Ensemble Data: A Survey. *IEEE Transactions on Visualization and Computer Graphics* 2019; 25(9): 2853–2872.

15. Harris ZS. Distributional Structure. *WORD* 1954; 10(2-3): 146–162.

16. Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos. In: ; 2003: 1470.

17. Fei-Fei L, Perona P. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In: . 2. ; 2005: 524–531.

18. Fei-Fei L, Fergus R, Perona P. Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding* 2007; 106: 59–70.

19. Whitaker RT, Mirzargar M, Kirby RM. Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Transactions on Visualization and Computer Graphics* 2013; 19(12): 2713–2722.

20. Mirzargar M, Whitaker RT, Kirby RM. Curve boxplot: Generalization of boxplot for ensembles of curves. *IEEE Transactions on Visualization and Computer Graphics* 2014; 20(12): 2654–2663.

21. Jarema M, Demir I, Kehrer J, Westermann R. Comparative visual analysis of vector field ensembles. In: ; 2015: 81–88.

22. Höllt T, Magdy A, Chen G, et al. Visual analysis of uncertainties in ocean forecasts for planning and operation of off-shore structures. In: IEEE. ; 2013: 185–192.

23. Matkovic K, Gracanin D, Jelovic M, Ammer A, Lez A, Hauser H. Interactive visual analysis of multiple simulation runs using the simulation model view: Understanding and tuning of an electronic unit injector. *IEEE Transactions on Visualization and Computer Graphics* 2010; 16(6): 1449–1457.

24. Matkovic K, Gracanin D, Klarin B, Hauser H. Interactive visual analysis of complex scientific data as families of data surfaces. *IEEE Transactions on Visualization and Computer Graphics* 2009; 15(6): 1351–1358.

25. Bergner S, Sedlmair M, Moller T, Abdolyousefi SN, Saad A. ParaGlide: Interactive parameter space partitioning for computer simulations. *IEEE Transactions on Visualization and Computer Graphics* 2013; 19(9): 1499–1512.

26. Bruckner S, Moller T. Result-driven exploration of simulation parameter spaces for visual effects design. *IEEE Transactions on Visualization and Computer Graphics* 2010; 16(6): 1468–1476.

27. Liu R, Guo H, Yuan X. A bottom-up scheme for user-defined feature exploration in vector field ensembles. In: IEEE. ; 2015: 155–156.

28. Liu R, Guo H, Yuan X. User-defined feature comparison for vector field ensembles. *Journal of Visualization* 2017; 20(2): 217–229.

29. Fofonov A, Linsen L. Projected Field Similarity for Comparative Visualization of Multi-Run Multi-Field Time-Varying Spatial Data. In: . 38. ; 2019: 286–299.

30. Doleisch H, Gasser M, Hauser H. Interactive feature specification for Focus+Context visualization of complex simulation data. In: . 3. ; 2003: 239–248.

31. Muigg P, Kehrer J, Oeltze S, et al. A four-level Focus+Context approach to interactive visual analysis of temporal features in large scientific data. In: . 27. ; 2008: 775–782.

32. Berger W, Piringer H, Filzmoser P, Gröller E. Uncertainty-aware exploration of continuous parameter spaces using multivariate prediction. In: . 30. ; 2011: 911–920.

33. Woodring J, Ahrens J, Figg J, Wendelberger J, Habib S, Heitmann K. In-situ sampling of a large-scale particle simulation for interactive visualization and analysis. *Computer Graphics Forum* 2011; 30(3): 1151–1160.

34. Köthur P, Sips M, Dobslaw H, Dransch D. 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* 2014; 20(12): 1893–1902.

35. Hazarika S, Dutta S, Shen HW. Visualizing the Variations of Ensemble of Isosurfaces. In: ; 2016: 209—213.

36. Splechtna R, Matković K, Gracanin D, Jelović M, Hauser H. Interactive visual steering of hierarchical simulation ensembles. In: ; 2015.

37. Hao L, Healey CG, Bass SA. Effective Visualization of Temporal Ensembles. *IEEE Transactions on Visualization and Computer Graphics* 2016; 22(1): 787–796.

38. Demir I, Kehrer J, Westermann R. Screen-space silhouettes for visualizing ensembles of 3D isosurfaces. In: ; 2016: 204–208.

39. He W, Liu X, Shen HW, Collis SM, Helmus JJ. Range likelihood tree: A compact and effective representation for visual exploration of uncertain data sets. In: ; 2017: 151–160.

40. He Q, Iyengar A, Nejdl W, Pei J, Rastogi R., eds., *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*; ACM: 2013.

41. Yang Q, Agarwal D, Pei J., eds., *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*; ACM: 2012.

42. Dhillon IS, Koren Y, Ghani R, et al., eds., *The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, August 11-14, 2013*; ACM: 2013.

43. Chen X, Lebanon G, Wang H, Zaki MJ. , eds., *21st ACM International Conference on Information and Knowledge Management, CIKM'12, Maui, HI, USA, October 29 - November 02, 2012*; ACM: 2012.

44. Apté C, Ghosh J, Smyth P., eds., *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*; ACM: 2011.

45. Rao B, Krishnapuram B, Tomkins A, Yang Q., eds., *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, July 25-28, 2010*; ACM: 2010.

46. Macdonald C, Ounis I, Ruthven I. , eds., *Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011*; ACM: 2011.

47. Xiong H, Karypis G, Thuraisingham BM, Cook DJ, Wu X., eds., *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*; IEEE Computer Society: 2013.

48. Zaki MJ, Siebes A, Yu JX, Goethals B, Webb GI, Wu X., eds., *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*; IEEE Computer Society: 2012.

49. Webb GI, Liu B, Zhang C, Gunopulos D, Wu X. , eds., *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*; IEEE Computer Society: 2010.

50. Jones GJF, Sheridan P, Kelly D, Rijke dM, Sakai T., eds., *The 36th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '13, Dublin, Ireland - July 28 - August 01, 2013*; ACM: 2013.

51. Hersh WR, Callan J, Maarek Y, Sanderson M. , eds., *The 35th International ACM SIGIR conference on research and development in Information Retrieval, SIGIR '12, Portland, OR, USA, August 12-16, 2012*; ACM: 2012.

52. Ma W, Nie J, Baeza-Yates R, Chua T, Croft WB., eds., *Proceeding of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July 25-29, 2011*; ACM: 2011.

53. Crestani F, Marchand-Maillet S, Chen H, Efthimiadis EN, Savoy J., eds., *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2010, Geneva, Switzerland, July 19-23, 2010*; ACM: 2010.

54. Da Silva NF, Hruschka ER, Hruschka Jr ER. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems* 2014; 66: 170–179.

55. Schuller B, Müller R, Lang M, Rigoll G. Speaker independent emotion recognition by early fusion of acoustic and linguistic features within ensembles. In: ; 2005.

56. Csurka G, Dance C, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. In: . 1. Prague. ; 2004: 1–2.

57. Zhang Y, Jin R, Zhou ZH. Understanding Bag-of-Words model: a statistical framework. *International Journal of Machine Learning and Cybernetics* 2010; 1(1-4): 43–52.

58. Smith A, Chuang J, Hu Y, Boyd-Graber J, Findlater L. Concurrent visualization of relationships between words and topics in topic models. In: ; 2014: 79–82.

59. Jiang YG, Ngo CW. Bag-of-Visual-Words expansion using visual relatedness for video indexing. In: ; 2008: 769–770.

60. Caicedo JC, Cruz A, Gonzalez FA. Histopathology image classification using Bag-of-Features and kernel functions. In: Springer. ; 2009: 126–135.

61. Jégou H, Douze M, Schmid C. Packing Bag-of-Features. In: ; 2009: 2357–2364.

62. Lucchesi GKKJKR. GEOS-5 Production Operations.. 2012.

63. Emmons , L. K. Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4).. *Geoscientific Model Development* 2010.

64. Collins C, Penn G, Carpendale S. Bubble Sets: Revealing set relations with isocontours over existing visualizations. *IEEE Transactions on Visualization and Computer Graphics* 2009; 15(6): 1009–1016.