J Vis (2021) 24:317–330
https://doi.org/10.1007/s12650-020-00709-z

REGULAR PAPER

**Renpei Huang · Li Chen · Xiaoru Yuan**

# A visual uncertainty analytics approach for weather forecast similarity measurement based on fuzzy clustering

Received: 1 July 2020 / Revised: 26 August 2020 / Accepted: 8 September 2020 / Published online: 3 January 2021


**Abstract** Forecast calibration methods based on historical similar atmospheric state are effective means weather forecast accuracy. Conventional approaches search similar forecasts on the basis of predefined similarity formulas and provide calibration recommendations to forecasters. However, these approaches ignore the uncertainty of similarity measurement, which affects calibration efficacy significantly. This study proposes a similarity weight adaptive algorithm for high-dimensional data on the basis of fuzzy clustering to characterize the uncertainty of similarity measurements. Without prior knowledge, the algorithm computes the uncertainty of the similarity between data in the fuzzy set space iteratively on the basis of membership and then determine weight distribution by maximizing the differentiating ability of each dimension. This study further presents a visual analysis framework on the basis of the weight adaptive algorithm for the exploration of uncertainty in meteorological data and the optimization of similarity measurement method. This framework has coordinated views and intuitive interactions to enable the visualization of the similarity uncertainty distribution and support the iterative visual analysis of similarity weight distribution in each dimension that combines domain knowledge. We illustrate a case study using real-world meteorological data to verify the efficacy of the proposed approach.

**Keywords** Uncertainty visualization · Fuzzy clustering · Weather forecast

## 1 Introduction

Weather forecasting predicts the atmospheric state at a certain position by using modern technology. Professional forecasters release weather reports after collating and calibrating the forecast data. As the scale of the weather forecast results continuously increases, analyzing the raw data and making decisions directly become difficult for experts. Visualizing meteorological data through visual analysis method to improve the ability of forecasters to understand large-scale high-dimensional data is a hot issue in data visualization and meteorology research. This issue is essential to human life and development. Numerical weather prediction (NWP) is the mainstream method in weather forecasting. This method can predict the atmospheric parameter of a specific time and space quantitatively and generate the probability density distribution of the

R. Huang · L. Chen (✉)
School of Software, BNRIST, Tsinghua University, Beijing, People's Republic of China
E-mail: chenlee@tsinghua.edu.cn

R. Huang
E-mail: huangrp2013@gmail.com

X. Yuan
Key Laboratory of Machine Perception (Ministry of Education), and National Engineering Laboratory for Big Data Analysis and Application, Peking University, Beijing, People's Republic of China

future atmospheric state (Leutbecher and Palmer 2008). However, given the chaotic atmospheric system and the inconsistency among the idealized assumptions of numerical simulations, NWP results are uncertain and produce bias inevitably. Forecasters must calibrate the bias to improve weather forecast accuracy. Conventional calibration approaches are based on data similarity. According to the pre-defined weighted similarity measurement, these approaches retrieve the forecast results in similar atmospheric states in historical forecast data and generate calibration recommendations to forecasters.

The key issue that affects weather forecast calibration accuracy is whether the uncertainty in the forecast data is characterized and understood effectively (Du and Kang 2014). For the calibration method based on similarity, the impact is focused on the uncertainty of similarity between data. The influence of the similarity uncertainty is reflected in time and space domains. In the time domain, the forecast bias distribution of extreme weather usually has large variance that reduces the accuracy of the predefined similarity measurement. In the space domain, the inconsistency between the scope of the predefined similarity measurement and the area to be calibrated also reduces similarity retrieval accuracy. Skilled forecasters can accept or reject the calibration recommendations given by the automated method on the basis of their domain knowledge and comprehension of similarity uncertainty. However, forecasters and their domain knowledge cannot participate in defining the similarity measure, thereby reducing calibration accuracy.

The use of visual analysis approach to support the exploration of the similarity uncertainty in large-scale complex forecast data and facilitate automated methods with human knowledge is important for the bias calibration work. However, two major challenges arise. First, defining and computing the uncertainty of similarity between data without prior knowledge is difficult. The representation of uncertainty should be established by modeling the data and statistical induction. Only the exact coordinates of the data distributed in the data space cannot define the uncertainty of the similarity between the data. Designing an uncertainty measurement approach that can expand the data space without prior knowledge is necessary. Second, existing forecast calibration visual analysis approaches are difficult to apply to the optimization of similarity measurement for forecast data. The existing visualization work on forecast calibration is mainly focused on improving calibration efficiency, but few visual analysis methods are designed to incorporate user domain knowledge into the algorithm execution process.

In this study, we present a visual uncertainty analysis approach of weather forecast similarity measurement on the basis of fuzzy clustering to summarize and visualize the uncertainty of similarity in meteorological forecast data. Specifically, we characterize the uncertainty distribution of the similarity in each dimension quantitatively on the basis of the membership of each datum to each fuzzy cluster without prior knowledge. Subsequently, we propose a weight adaptive algorithm to determine the distribution of dimension weight automatically by maximizing the differentiating ability of each dimension. Thus, we summarize the representation of similarity uncertainty for the exploration and optimization of similarity measurement. On the basis of the weight adaptive algorithm, we present a visual analysis framework for the exploration of similarity uncertainty in meteorological data. It has coordinated views and intuitive interactions to enable the visualization of the similarity uncertainty distribution and support the iterative visual analysis of similarity weight distribution in each dimension that combines domain knowledge. Particularly, the framework provides an uncertainty distribution exploration on the basis of a probability density map matrix and glyphs arranged in time series. The visual analysis framework combines visualization and interaction to assist domain experts in analyzing the uncertainty of similarity in the weather forecast data. A case study using real-world meteorological data evaluates the efficacy of our approach. We present our primary contributions as follows.

- We present a similarity weight adaptive algorithm for high-dimensional data on the basis of fuzzy clustering which can calculate the uncertainty of the similarity between data in the fuzzy set space iteratively on the basis of membership without prior knowledge and determine weight distribution by maximizing differentiating ability.
- We propose a visual analysis framework for the exploration of similarity uncertainty and the optimization of similarity measurement to enable the visual analysis of the similarity weight distribution and support the iterative visual analysis of similarity measurement that combines domain knowledge.

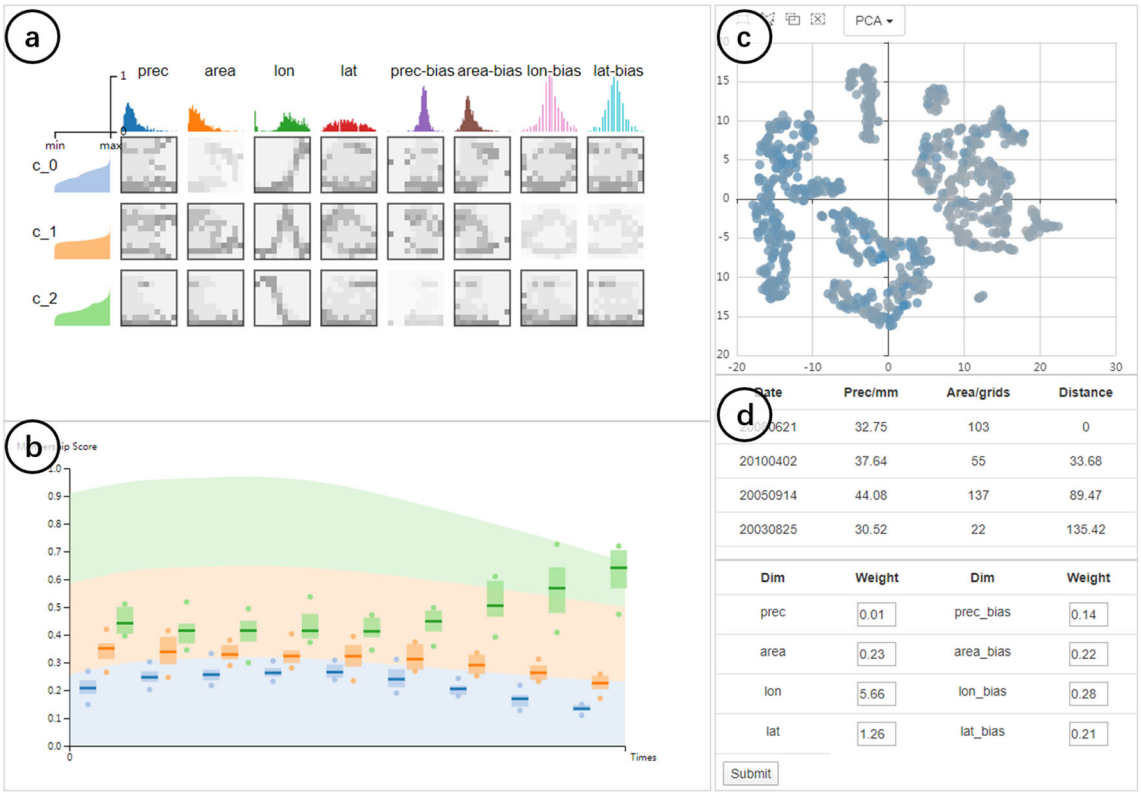


**Fig. 1** Interface of the visual uncertainty analytics approach. **a** Cluster-dimension matrix view displays the uncertainty of similarity in each dimension by probability density map. **b** Clustering process view shows the change in uncertainty of all data by stacked line chart and shows the change in membership in each iteration of selected data by glyphs. **c** Data distribution scatterplot illustrates the similarity distribution on the basis of dimension reduction method. **d** Weight adjustment widget assists in setting initial parameter and data recommendation

## 2 Related work

The literature that overlaps with this work can be categorized into three categories, namely uncertainty visualization, fuzzy clustering visualization, and calibration in the meteorological forecast.

### 2.1 Uncertainty visualization

Uncertainty visualization is an important branch of data visualization and faces many challenges (Bonneau et al. 2014).

Ferstl et al. (2016a, b, 2017) generated variability plots to explore the distribution trend of statistical variable in scalar field, vector field and time series data. Whitaker et al. (2013) and Mirzargar et al. (2014) extended the low-dimensional statistical method to high-dimensional meteorological ensemble data. Their approaches can generate statistical models of curves and contours in the data. Pfaffelmoser et al. (2011, 2013) and Sanyal et al. (2010) visualize the uncertainty at each grid point of weather forecast ensemble data by circular glyphs and special coloring scheme. Potter et al. (2009) and Wang et al. (2017) proposed multiple linked views to assist the data exploration and comparison in high-dimensional data.

The existing uncertainty visualization work can effectively support the exploration and analysis of the uncertainty of multiple types of data (Liao et al. 2016, 2018). However, visual analysis methods that are designed to incorporate user domain knowledge into the algorithm execution process remain limited. In this study, we extract the similarity uncertainty in the intermediate results of the back-end algorithm and visualize the uncertainty to support the iterative visual analysis of similarity weight distribution that combines domain knowledge.

### 2.2 Fuzzy clustering visualization

Compared with other hard clustering approaches, fuzzy clustering is more feasible in many scenarios, because it does not completely divide the data into a specific cluster (Zhao et al. 2018). The visualization of fuzzy clustering is widely used in high-dimensional scalar data. Sharko et al. (2008), Sharko and Grinstein (2009) visualized the fuzzy clusters by Radviz to make a metaphor of data membership distribution to all clusters. Berthold and Hall (2003) employed parallel coordinates to clearly illustrate the membership distribution of fuzzy points. Mao and Jain (1995) reduced the dimension of fuzzy clustering result by principal component analysis (PCA) and displayed the membership distribution in 2D space while maintaining the similarity distance as much as possible. Rueda and Zhang (2006) mapped the fuzzy clustering result into a hyper-tetrahedron to show the geometric correlation between the data. Gasch and Eisen (2002) used heatmap to visualize the sorted matrix of the fuzzy clustering result to assist experts in similar pattern detections.

These studies revealed the similarity between fuzzy clustering results through various methods. However, although they focused on showing the similarity of the membership vectors in the fuzzy concept space, they ignored the similarity of the data value in each dimension. In this study, our approach visualizes the joint distribution of each dimension and membership to characterize the uncertainty of the similarity between data.

### 2.3 Calibration in meteorological forecast

The process that forecasters detect and correct the prediction bias produced by NWP products is called calibration. Many automated methods were proposed to improve calibration efficiency and accuracy. The automated methods are divided into statistical- and similarity-based methods. Statistics-based methods make calibration recommendations by analyzing the statistical characteristics of historical data, such as model output statistics (Glahn and Lowry 1972), bias-corrected relative frequency method (Hamill and Whitaker 2006). The similarity-based method guides calibration on the basis of historical predictions under similar climate conditions in historical data, such as Analog (Hamill et al. 2015). In addition, Raftery et al. (2005) used Bayesian model to analyze multi-source data sets and assigned weights to assist calibration. These methods can improve the calibration results, but the results obtained by experienced forecasters who calibrate independently are often accurate. This implies that incorporating the forecaster's experience into the calibration algorithm helps improve the calibration accuracy.

Besides automated methods, certain works combine visual analysis with the weather forecast calibration process to assist forecasters in making calibration decisions. Liao et al. (2015) presented a visual voting framework to improve calibration efficiency of each grid point. Wang et al. (2015) proposed a rain band extraction method on the basis of Gaussian mixture model, and designed a system to assist rain band validation. Gong et al. (2016) described a salience-based visual analysis system to help forecasters in regional corrections. These studies focused on improving the efficiency of forecast calibration through visual analysis. However, they cannot help forecasters understand the uncertainty in forecast data. In this study, our approach enables forecasters to explore the uncertainty distribution in the forecast data and make further decisions.

## 3 Back-end engine

As shown in Fig. 2, the back-end engine characterizes the uncertainty from forecast data, and determine weight distribution by maximizing differentiating ability of each dimension. The algorithm takes high-dimensional scalar data as input, which is a scalar dataset that includes rainfall from 2002 to 2013 in the USA, with 8 data dimensions and a time step dimension. The back-end algorithm computes on 8 data dimensions of the dataset at all time steps. In this section, we first explain the uncertainty of similarity measurement characterized by membership in the fuzzy concept space. Then, we describe the algorithm that is used to adaptively determine the weight distribution by maximizing the differentiating ability of each dimension.

### 3.1 Uncertainty of similarity measurement

Modeling and analyzing the similarity measurement by assigning dimension weights is necessary to define the similarity among various types of high-dimension weather forecast data. Clustering is an effective

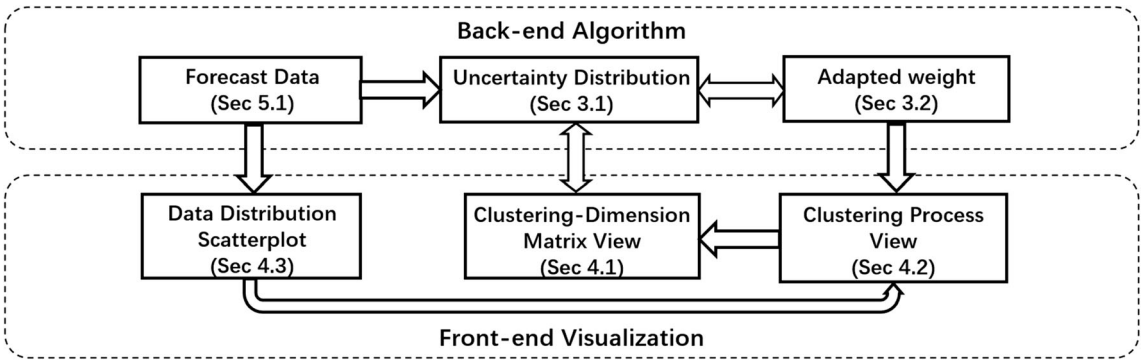


**Fig. 2** Data processing pipeline and an overview. The back-end algorithm characterizes the uncertainty from forecast data and determines weight distribution by maximizing differentiating ability of each dimension. The views in front-end visualization displays the result of each phase of the back-end algorithm and supports the interactive analysis for experts to analyze the data

method to characterize the similarity distribution of data without prior knowledge. The clustering method assigns cluster labels to each datum and adjusts the label on the basis of the distance between the data and cluster center, so that the data in the same cluster is as similar as possible, whereas the data in different clusters are not. Fuzzy clustering (Ruspini 1969; Zadeh et al. 1996) is a type of clustering method, that is also known as soft clustering. In other hard clustering methods, the data either belong to the same cluster simultaneously or to two completely different clusters. Thus, an "either identical or completely different" relationship exists between each datum. This relationship can assist in identifying the features of every cluster, but cannot support the quantitative analysis of the similarity between the data. The fuzzy clustering method does not absolutely divide the ascription of the data but uses the possibility that the data belongs to each cluster, that is, the uncertainty in describing the clustering result of the data. This method weakens the concept of clustering, but can quantitatively characterize the uncertainty of the relationship between the data and clusters, as a basis for judging the similarity between the data.

Most data in the world cannot be classified on the basis of the "either identical or completely different" criteria. The degree to which they belong to a concept is uncertain. Concepts defined in a certain domain that cannot be classified by "either identical or completely different" criteria, such as youth (age), warmth (temperature), and dusk (time), are called fuzzy concepts. Membership describes the possibility that a value in the definition domain belongs to a fuzzy concept. On this basis, a fuzzy set represents a set of values and their membership to a fuzzy concept. Formally, we let the membership function defined on the data set $X = \{x\}$ for the fuzzy concept $C$ be $\mu_C(x)$, whose range is [0, 1]. $\mu_C(x) = 1$ means that the data x completely belongs to the concept $C$, which is equivalent to $x \in C$. $\mu_C(x) = 0$ means that the data x completely different from the concept $C$, which is equivalent to $x \notin C$. Then, according to the membership function, the fuzzy set defined on $X = \{x_i | i = 1, 2, \ldots, N\}$ for the fuzzy concept $C$ can be expressed as:

$$C = \{(\mu_C(x_i), x_i) | x_i \in X\} \tag{1}$$

where $\mu_C$ is the membership function of the fuzzy concept $C$. For a series of fuzzy concepts $C$, we let $\mu_{ij}$ be the value of the membership function of the fuzzy concept $C_j$ at the data $x_i$. We define the membership matrix $U = \mu_{ij}$, which is used to describe the membership of every data point for each fuzzy concept. Then,

$$U = \begin{bmatrix} \mu_{C_1}(x_1) & \ldots & \mu_{C_k}(x_1) \\ \vdots & \ddots & \vdots \\ \mu_{C_1}(x_N) & \ldots & \mu_{C_k}(x_N) \end{bmatrix} \tag{2}$$

where $U$ can be regarded as a matrix composed of k-dimensional vectors $\mathbf{x_i}(\mu_{C_1}(x_i), \ldots, \mu_{C_k}(x_i))$ in the feature space formed by the domains of membership functions of each fuzzy concept. Let vector $\mathbf{x_i}$ be the membership vector of data $x_i$. The membership vector, which corresponds to all data in the data set, constitutes the membership matrix $U$. Therefore, $U$ represents the uncertainty distribution of similarity between each datum in the fuzzy concept space. According to $U$, the similarity between the data can be measured in the original data space.

Fuzzy clustering is used to generate membership matrix $U$ to describe the uncertainty distribution in the fuzzy concept space. In fuzzy clustering, each cluster is a fuzzy set. This method calculates the membership of each datum to each fuzzy set by optimizing the distance between clusters and cluster members. Fuzzy C-means clustering (FCM clustering) algorithm (Bezdek 2013) is a widely used algorithm for fuzzy clustering. The algorithm comes from the generalization of the optimization goal of k-means clustering. The purpose of FCM is to minimize a objective function $J(U, C)$ under the constraint that $\sum_{j=1}^{k} \mu_{ij} = 1, i = 1, 2, \ldots, N$. $J(U, C)$ is expressed as:

$$J(U, C) = \sum_{i=1}^{N} \sum_{j=1}^{k} \mu_{ij}^{m} \times \mathrm{dis}(x_i, c_j) \tag{3}$$

where $N$ is the size of the data set, $k$ is the number of expected fuzzy sets, $C = \{c_j | j = 1, 2, \ldots, k\}$ is the set of central coordinates of all fuzzy sets, $U$ is the membership matrix, *dis* is the similar measurement function, and $m$ is a hyperparameter that is usually set to 2.

The algorithm iteratively maintains the central set $C$ of the fuzzy set and the membership matrix $U$, so that it continuously approaches the optimal solution until the goal. According to the initialized fuzzy set center $C_0$, the FCM algorithm iteratively calculates the membership matrix $U = \{\mu_{ij}\}$ and the fuzzy cluster center $c_j$ to convergence through two steps.

*Step 1*

$$\mu_{ij} = \frac{1}{\sum_{t=1}^{k} \frac{\mathrm{dis}(x_i, c_j)^{\frac{2}{m-1}}}{\mathrm{dis}(x_i, c_j)}} \tag{4}$$

*Step 2*

$$c_j = \frac{\sum_{i=1}^{N} \mu_{ij}^{m} x_i}{\sum_{i=1}^{N} \mu_{ij}^{m}} \tag{5}$$

The algorithm does not require a priori model for classification reference. In addition, the running time can be adjusted by the threshold, thereby achieving flexible control of clustering accuracy and time cost. The membership vector $\mathbf{x_i}$ of each datum $x_i$ for each fuzzy set can be generated after the two-step iteration. The uncertainty of the similarity between each datum and each fuzzy set can be described by the membership vector, and the data distribution in the fuzzy concept space can be characterized as well. Thus, the similarity between each datum can be reflected objectively.

### 3.2 Weight distribution adaptive algorithm

With the given dimension weights, the clustering result can be evaluated according to the data concentration in the same cluster and the data dispersion in different clusters. The more concentrated the data within the same cluster, and the more dispersed they are in different clusters, the more rational the clustering result is. Similarly, with the given clustering result, a rational distribution of dimension weights can be generated by adjusting the dimension weights to make the clustering result consistent with the distribution of "concentrated in the same cluster, dispersed in different clusters."

When projecting the clustering result on each dimension, if the current dimension projection conforms to the criteria of rational clustering results, then this dimension has a strong ability to distinguish data. Otherwise, if the clustering result is mixed, the ability to distinguish the data of this dimension is weak. Therefore, the data differentiation ability of each dimension can be estimated by quantifying the concentration and dispersion of the projection of the clustering result. According to the criteria of rational clustering results, a discriminant coefficient $B_p$ is designed to measure the degree of concentration and dispersion in each dimension. Fuzzy clustering quantifies the uncertainty of the data that belongs to each fuzzy set through membership $\mu_{ij}$. Thus, $\mu_{ij}$ can represent the probability that the data $x_i$ belongs to the fuzzy set $C_j$. According to the conditional probability, $B_p$ is expressed as:

$$B_p = \frac{\frac{1}{k} \sum_{j=1}^{k} \left( c_j^{(p)} - \frac{1}{N} \sum_{i=1}^{N} x_i^{(p)} \right)^m}{\frac{1}{k} \sum_{j=1}^{k} \frac{\sum_{x_i \in C_j} P(x_i | C_j) \left( x_i^{(p)} - c_j^{(p)} \right)^m}{\sum_{x_i \in C_j} P(x_i | C_j)}} \tag{6}$$

where $x_i^{(p)}$ and $c_j^{(p)}$ represent the projection of the data and clustering center on the p-th dimension, $P(x_i|C_j) = \mu_{ij}$, indicating the probability that the data $x_i$ belongs to the fuzzy set $C_j$, $m$ is the same hyperparameter in Eq. 3. The numerator part of $B_p$ calculates the average distance from each cluster to the data center, which quantifies the degree of data dispersion. The denominator part calculates the average distance between the data and its cluster center, which quantifies the degree of data dispersion under the p-th dimension projection.

The similarity weight distribution for each dimension in the original data space can be generated by $B_p$ for all dimensions in the high-dimensional scalar data. The expression of the weight $w_p$ in the p-th dimension is:

$$w_p = \frac{B_p}{\sum_{i=1}^{d} B_i} \tag{7}$$

where $d$ is the number of dimensions in the high-dimensional data.

As the similarity measurement after weight adjustment can distinguish data more effectively than before, the clustering result is more rational based on the adjusted weight. Therefore, the weight distribution adaptive algorithm combines dimension weights generation with the fuzzy clustering method. In the algorithm, the dimension weights are updated after each two-step iteration of FCM, and the membership matrix and fuzzy set center is calculated by the updated dimension weights in the next iteration. By incorporating the dimension weight update step into the FCM iteration, we can iteratively optimize the clustering result and the distribution of dimensional weights, thereby improving the accuracy of the similarity measurement. Concurrently, the algorithm can support the subsequent visual analysis design that combines user domain knowledge and automated algorithms.

## 4 Front-end visualization

In this section, we introduce the visual design and interaction of views in the visual analysis framework for similarity measurement of high-dimensional meteorological forecast data. The visualization consists of four parts: the cluster-dimension matrix view, clustering process view, data distribution scatterplot, and the weight adjustment widget, as shown in Fig. 1. The clustering process view (Fig. 1b) can provide an overview of the algorithm execution process, thereby assisting forecasters and domain experts to understand the changing trend of uncertainty with the algorithm iteration. We use the cluster-dimension matrix view (Fig. 1a) to display the details of the uncertainty distribution of data similarity in each iteration, and supports filtering and screening of extreme data and weakly related dimensions through simple interaction. The data distribution scatterplot (Fig. 1c) can display the distribution of data in the original data space under the current weight. The weight adjustment widget (Fig. 1d) can perform a similar data recommendation and initial weight setting to assist the forecaster in visual analysis and calibration of forecast data. The views can visualize the execution process and results of the back-end algorithm. According to the visualization in each view, users can interact with the cluster-dimension matrix view to adjust the cluster or dimension involved in the algorithm, or directly input a suitable dimension weight through the weight adjustment widget. The back-end algorithm will use the input weight as the initial value, and only use the dimension and weight specified by the user interaction during the iteration process. We visualize the results after the algorithm is re-executed in each view for subsequent iterative visual analysis.

### 4.1 Cluster-dimension matrix view

The cluster-dimension matrix view (Fig. 1a) shows the distribution of data in the data space and fuzzy concept space under the current similarity weight of the back-end algorithm during each iteration based on the density map matrix.

#### *4.1.1 Visual designs of the matrix*

The main body of the view is a density map matrix. Each row of the matrix corresponds to a fuzzy cluster $C_j$. Each column corresponds to a dimension $p$ of the original high-dimensional scalar meteorological forecast data. The vertical axis of the density map corresponds to the range [0, 1] of membership functions of fuzzy clusters from bottom to top. The horizontal axis of the density map corresponds to the value range

of the data in each dimension, with the minimum value on the left and the maximum value on the right. Therefore, each element of the matrix is a density distribution map of the data which represents the joint density distribution of the data in the feature space composed of membership and data values. We divide the density of data distribution into five levels from sparse to dense and encode the corresponding gray values in order from light to dark, as shown in Fig. 3a. The joint density distribution visualized by each element of the density map matrix can intuitively reflect the membership relationship between the projected value of the weather forecast data on the corresponding dimension and corresponding fuzzy cluster, and then reflect the correlation and uncertainty of the data and dimension. Figure 3b shows the distribution of lower and higher probability density. The left density map shows a strong correlation between the data projection and the membership function. The uncertainty is low and the dimension differentiating ability is strong. The data projections in the right density map are chaotically distributed throughout the membership-dimension feature space. The uncertainty is large, thereby indicating that the dimension is weak in distinguishing the corresponding fuzzy clusters and data. Generally, the more concentrated the density distribution of the data (that is, when the dark patches in the density distribution map are concentrated and continuous), the lower is the similarity uncertainty characterized by the density distribution.

#### *4.1.2 Design of each element of the matrix*

We display the name of each dimension of meteorological forecast data and present a histogram of the frequency of data projection distribution above each column of the probability density matrix. As shown in Fig. 3b, the horizontal axis of the histogram is the value range of corresponding dimension, which is the same as that of the probability density map. The vertical axis represents the data frequency. The histogram shows the distribution of data in corresponding dimension. We present the cumulative frequency distribution graph of the membership of meteorological forecast data on each fuzzy cluster on the left side of the matrix. The horizontal axis is the same as the vertical axis of the density map, which is the range of the membership function [0,1]. The vertical axis represents the cumulative frequency, which can reflect the membership distribution of all data in the corresponding fuzzy cluster. The histogram of data projection and cumulative frequency of membership can help user understand the uncertainty of data similarity in the density distribution matrix and combine their own domain knowledge for interaction and subsequent decision-making.

#### *4.1.3 Interactions*

The cluster-dimension matrix view is linked with the clustering process view to show the similarity distribution of data in different iterations, and enables the interaction of clicking to filter the cluster and dimensions that participate in the back-end algorithm. In Eq. 2, the probability distribution $P(x|C_j)$ of the data $x$ belonging to the cluster $C_j$ is computed according to the similarity uncertainty represented by the membership. Then, the ability to distinguish the data in the p-th dimension can be quantified. However, a probability distribution $P_0(x|C_j)$ with large uncertainty reduces the accuracy of the clustering result, and the dimension $p_0$ may affect the similarity measurement accuracy. Conventional methods pose difficulty in avoiding the accuracy decrease caused by high uncertainty. The cluster-dimension matrix view enables users to explore the data cluster and dimensions, and analyze the uncertainty distribution of the data similarity combined with their own domain knowledge. The cluster-dimension matrix view enables users to filter the probability density map with large uncertainty through clicking interaction to avoid reducing the accuracy of the similarity measurement weight. This condition means that the influence of the probability distribution corresponding to the filtered density map in the back-end algorithm is removed. As shown in Fig. 1a, black boxes indicate the probability density maps involved in the back-end algorithm, whereas transparent probability density maps are removed from the algorithm. Users can switch the black box and transparent state by clicking on the probability density map. After interacting with the view and determining the parameters, user can rerun the adaptive weight algorithm according to the new parameters at the current time step.

### 4.2 Clustering process view

The clustering process view (Fig. 1b) shows the change in the overall data uncertainty with iteration during the establishment of the similarity measurement, and the change in the membership distribution of specific data with the iteration.

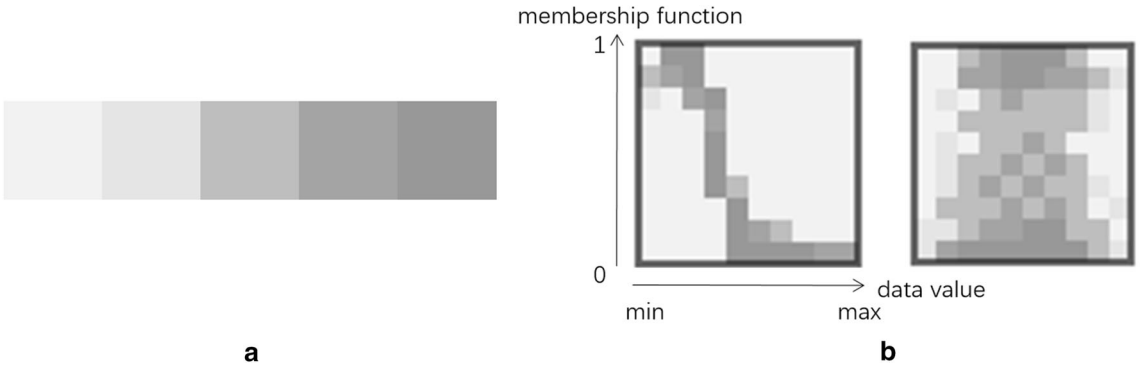


**Fig. 3** **a** Graded representation of grayscale. **b** Probability density map with low uncertainty (left) and high uncertainty (right)

#### *4.2.1 Visualization of overall uncertainty*

The horizontal axis of the clustering process view is the time axis, which indicates the number of iterations. The vertical axis represents the values of the fuzzy entropy and membership. The stacked line graph in the view shows the change in the fuzzy entropy of weather forecast data during clustering. We define the fuzzy entropy as follows:

$$H = \sum_{j=1}^{k} \sum_{i=1}^{N} -\mu_{ij} \ln \mu_{ij} - (1 - \mu_{ij}) ln(1 - \mu_{ij}) \tag{8}$$

where $\mu_{ij}$ is the value of the membership function of fuzzy set $C_j$ at data point $x_i$. The fuzzy entropy describes the overall data uncertainty. The higher the fuzzy entropy, the greater the uncertainty in the data. Particularly, the overall fuzzy entropy of the data is the sum of the fuzzy entropy for each fuzzy cluster. We present the fuzzy entropy of each fuzzy cluster in the clustering process view. A stacked line chart is formed by stacking several line charts that correspond to the color of the same fuzzy cluster in other views. A single color line chart shows the fuzzy entropy of the corresponding fuzzy cluster in each iteration, and the stacked line represents the changes of the fuzzy entropy of all data, that is, the uncertain changes.

#### *4.2.2 Visualization of selected data points*

The clustering process view also supports visualization of the membership of the selected data. The view is linked with the data distribution scatterplot such that, when the user selects clustering or filtering data, the membership of the selected data is automatically visualized by arranging glyphs on the time axis. The glyph for visualization is shown in Fig. 4a. The dots at the top and bottom indicate the corresponding values of the 95th percentile of the data. The upper and lower sides of the rectangle in the glyph indicate the values that correspond to the two quartiles. The horizontal line inside the rectangle corresponds to the median membership value of the selected data. We arranged the membership distribution glyphs of the selected data in different iterations according to the number of iterations and also arranged the glyphs in the same iteration horizontally according to the default order of the fuzzy clusters. Users can analyze the similarity relationship between the selected data and each fuzzy cluster according to the glyph and can select the iteration round of the data displayed in the cluster-dimension matrix view by clicking on different glyphs.

### 4.3 Data distribution scatterplot

The data distribution scatterplot (Fig. 1c) displays the similarity distribution of weather forecast data and enable users to filter the data.

The data distribution scatterplot can reduce the dimension of the high-dimensional meteorological forecast data through dimension reduction methods, such as PCA and t-SNE, to directly visualize the similarity distribution in 2D space. Each point in the view represents a high-dimensional data point, the position of which maps the high-dimensional data to the 2D coordinates after the dimension reduction. We expressed the data membership for each fuzzy cluster by the gray level of the point. If the membership is 0, then the point is completely gray, and if the membership degree is 1, then the color of the point is consistent

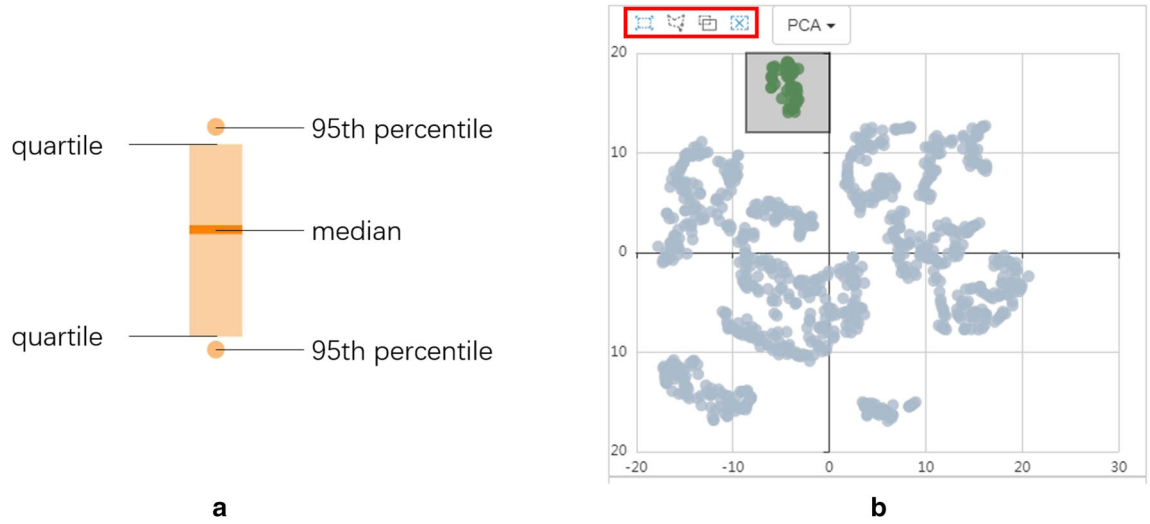


**Fig. 4** **a** Glyph of data membership in the clustering process view. The glyph shows the value of 95th percentile, quartile, and median of the membership of selected data. **b** Data filtering control (red rectangular box marked area) and filtering result in the data distribution scatterplot

with the representative color of the fuzzy cluster. Given that displaying the membership distribution of all clusters on the data through color and gray scale simultaneously can cause severe visual occlusion, the view only shows the distribution of membership of one cluster at a time and enables users to interact with the current display to switch the cluster. As shown in Fig. 4b, the user can select the data of interest by selecting the data filtering control (indicated by a red rectangle) provided. We updated the corresponding visualization results in other views to assist the user in analyzing the similarity of selected data in combination with the visualization of each view. The user can also select various dimension reduction algorithms through a drop-down menu to change the dimension reduction result of the weather forecast data. The view will recalculate and visualize the position of each datum point according to the algorithm selected by the user.

### 4.4 Weight adjustment widget

The weight adjustment widget assists the user in setting algorithm parameters, such as initial weight settings and similarity benchmark data point settings. With no preset parameters, the back-end algorithm assumes that the weights of all dimensions in the original weather forecast data are equal. The user can set the default initial weight of the algorithm through the weight adjustment control, and analyze the algorithm process according to the initial weight adjustment to optimize the predefined similarity measurement and improve its adaptability to the current dataset. The widget also enables the user to select or enter a similarity reference data point in a standard format. Once the reference data point is determined, the widget can automatically recommend several data with a small similarity distance from the reference point according to the current dimension weight. The data can be visualized as a list, thereby providing a reference for users to calibrate the weather forecast data.

## 5 Evaluation and discussion

In this section, we verify the efficacy of our visual uncertainty analysis framework through a case study using real-world meteorological forecast data. The case demonstrates the common workflow of the visual analysis process to assist domain experts in exploring and analyzing the similarity weight distribution in each dimension that combines domain knowledge. A discussion evaluated the system in terms of algorithm design, system usability, scalability, and limitation.

## 5.1 Data description

The data used in the case correspond to the rainfall forecast data extracted from the historical weather reanalysis data. The rainfall forecast data comprise historical forecast and observation data, which are generated by GEFS Hamill et al. (2013) and CCPA Hou et al. (2014), respectively. The rainfall in the reanalysis data is extracted by binarization with a precipitation threshold. Nearly, a thousand of rainfalls over 25 mm in the USA from 2002 to 2013 are encompassed by the rainfall forecast data. The eight scalar dimensions in rainfall forecast data, namely average precipitation (prec), rainfall coverage area (area), longitude (lon) and latitude (lat) of the rainfall geometric center coordinate, average precipitation bias (prec-bias), coverage area bias (area-bias), longitude bias (lon-bias) and latitude bias (lat-bias) between the forecast and observation data, are used for visual analysis.

## 5.2 Case study

The case study aims to assist domain expert in exploring the similarity uncertainty in the historical rainfall forecast data without prior knowledge, analyze the similarity characteristics of the rainfall, and summarize the similarity measurement. We worked with an expert, who is an experienced forecaster. Figure 5a1 shows the cluster-dimension matrix view of the rainfall forecast data in the last iteration of the weight adaptive algorithm under the default initial value. The membership distribution membership of the three fuzzy clusters are shown on the left side of the view. The expert found that all of the membership distributions in each fuzzy cluster in the data are average, and the result of the fuzzy cluster c_1 is relatively better than the others. Figure 5b1 shows the fuzzy entropy during the iterative process of the back-end algorithm and the change of membership of top 50% data in c_1. The stacked line chart shows that the overall uncertainty of all data did not decrease with the change in similarity weight, and the membership of the data to cluster c_1 hovers between 0.4 and 0.5. As the data similarity distribution shown in Fig. 5c1 presents, most data points are orange-gray, indicating that most data in the dimension reduction space may belong to the fuzzy cluster c_1. Thus, the expert considered that the fuzzy clusters and similarity weights cannot distinguish the data well. Through the analysis of the visualization, the current weight distribution cannot characterize the data similarity effectively. Therefore, filtering the probability density map and executing the back-end algorithm again are necessary. According to Fig. 5a1, the expert maintained the initial weights unchanged, and selected the data (framed by solid red rectangles) locate at row “c_0,” ”c_1” and column lon, “lat,” “area-

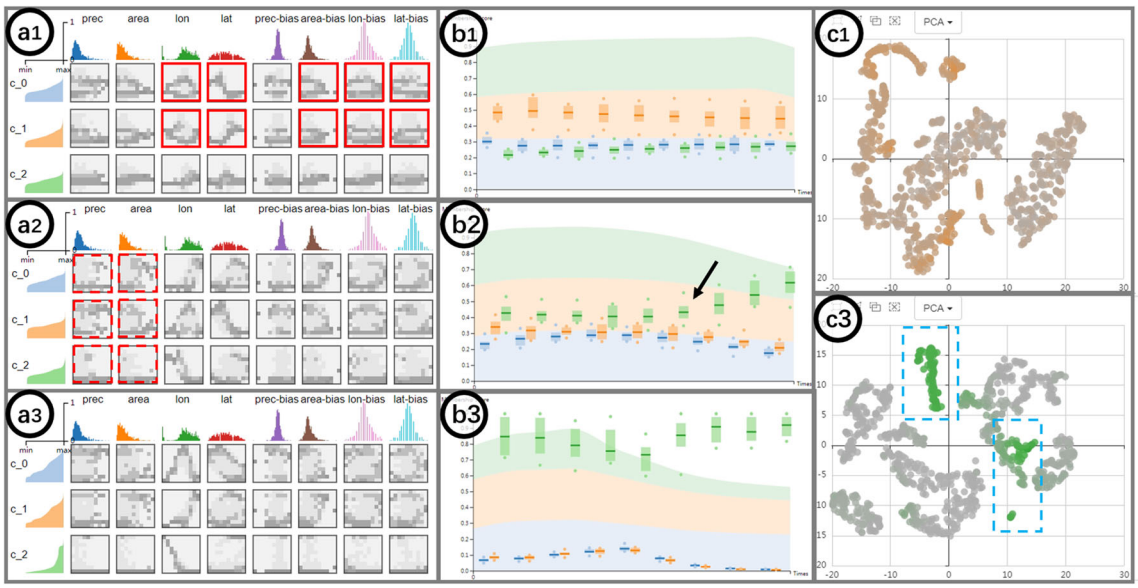


**Fig. 5** Visual analysis views in case study. Subscript *i* represents the visualization of the i-th time back-end algorithm execution. (**a**1–**a**3) Cluster-dimension matrix view in each weight adaptation result. (**b**1–**b**3) Change in fuzzy entropy and membership in the clustering process view in each weight adaptation result. (**c**1, **c**3) Data distribution scatterplot after the first and third weight adaptation results

bias," "lon-bias," "lat-bias" in the dimension matrix view for next execution of the back-end algorithm, because their corresponding density matrix has low uncertainty.

The second weight adaptive algorithm is executed by selecting the corresponding parameters of the aforementioned density maps interactively. The clustering process view of the second weight adaptation result is shown in Fig. 5b2, which also shows the change of the membership of the fuzzy cluster c_2. From the iteration round indicated by the black arrow, the overall data uncertainty begins to decrease, and the data membership starts to approach 0 and 1. The expert selected the 8th iteration in the clustering process view. Then, the density distribution matrix of data in the 8th iteration is shown as Fig. 5a2. On most density maps that correspond to clusters, the data density distribution shows the correlation between the dimension range and the membership function range. However, some density maps, such as the density maps at (c_1, prec) and (c_2, area), have high uncertainty. According to the expert, the similarity of heavy rain is more dependent on geographical characteristics than the atmospheric state. Therefore, the expert kept the initial weights unchanged, and unselected the data (framed by dotted red rectangles) locate at column "prec," "area" in the dimension matrix view, because their corresponding density matrix has high uncertainty.

We present the third weight adaptation result in Fig. 5b3. This iteration reduces the fuzzy entropy of the data significantly, that is, the uncertainty. Figure 5c3 visualizes the data similarity distribution after adapting the dimension weights. The data distribution with high membership for clustering c_2 is relatively concentrated(in the blue dotted rectangles), and the membership of other data for c_2 is almost 0, indicating that the clustering results and dimension weight distribution can characterize the similarity between data effectively. The data density distribution matrix of the last iteration is shown in Fig. 5a3. The overall distribution of the membership of each fuzzy cluster is concentrated at 0 and 1, indicating that the data have less uncertainty about the membership of each fuzzy cluster. The expert found that the dimension "lon," which represents longitude, has a strong correlation with the membership distribution of each cluster. Thus, the expert inferred that the longitude of the rainfall center can measure the similarity more effectively than other dimensions. The final weight distribution result shows that the weight of "lon" is 4.46, the weight of "lat" is 2.48, and the weights of the other dimensions are all less than 1, thereby also explaining the validity of the above-mentioned inference.

## 5.3 Discussion

The case study shows that the visual uncertainty analysis framework can assist domain experts in exploring the uncertainty of similarity and summarizing the similarity measurement of the given meteorological forecast data. Moreover, guiding conclusions are recommended to assist domain experts in similar weather condition retrieval and forecast calibration. We conducted a short semi-structured interview with the expert we worked with to evaluate our system.

### *5.3.1 Algorithm design*

The similarity weight adaptive algorithm based on fuzzy clustering can fit the similarity measurement by maximizing the differentiating ability of each dimension in high-scalar data without prior knowledge. Given the high complexity of the fuzzy clustering algorithm, the number of iterations of the algorithm should be limited as the size of forecast data increases, which may decrease the algorithm accuracy.

### *5.3.2 System usability*

The expert stated that *The system can assist the optimization of similarity measurement and improve the accuracy of forecast calibration*. He agreed that the system can assist user to explore the uncertainty of similarity measurements in meteorological forecast data and optimize the weight distribution of each dimension by iterative visual analysis that combines domain knowledge. The expert was satisfied with the iterative visual analysis process, because the system *combines the experience of local geographical calibration with the weight* and *decrease the error of similarity detection*. However, the new clustering results generated by the back-end algorithm after parameter adjustment through user interaction occupy the visual elements, such as color and transparency, thereby causing visual confusion. Having a certain understanding of each view and the visual analysis process before conducting the iterative visual analysis is necessary.

### *5.3.3 Scalability*

The back-end algorithm is theoretically feasible for any high-dimensional scalar data but not suitable for excessively large-scale data because of the time complexity of the algorithm. The visual analysis system can construct, adapt, and optimize similarity measurement by combining domain knowledge through visual analysis. However, for the task goals where expert knowledge and human decision-making have limited improvement on the results, the accuracy of the visual analysis results might be worse than those of other methods.

### *5.3.4 Limitation*

First, the use of probability density map matrix to characterize the internal uncertainty of the data causes information loss. To enhance efficiency, we used statistical histogram method to enumerate the number of data in each unit to represent the probability distribution. However, discretization leads to information loss, which hinders the user's understanding of data and uncertainty. However, other methods for generalizing the probability density map, such as kernel density estimation, introduces additional uncertainties into the data. Second, the expert emphasized that the system lacks visualization of spatial information in meteorological forecast data. The spatial distribution of the data will also affect the calibration accuracy. Third, this study only verifies the feasibility and effectiveness of the method through case analysis and does not evaluate the ability of the method to assist in the analysis of data and the weight fitting of similarity measurements quantitatively.

## 6 Conclusion and future works

In this study, we present a visual uncertainty analysis approach for the similarity of high-dimensional meteorological data on the basis of fuzzy clustering. We describe a similarity weight adaptive algorithm for high-dimensional data on the basis of fuzzy clustering to characterize the uncertainty of similarity measurements. Without prior knowledge, the algorithm computes the similarity uncertainty iteratively between the data in the fuzzy set space on the basis of membership and then determines weight distribution by maximizing the differentiating ability of each dimension. On the basis of the back-end algorithm, we further present a visual analysis framework for the exploration of uncertainty in meteorological forecast data and the optimization of similarity measurement method. This framework has coordinated views and intuitive interactions to enable the visualization of the similarity uncertainty distribution. The visual analysis framework supports an iterative visual analysis of similarity weight distribution in each dimension that combines domain knowledge. A case study that uses real-world meteorological data demonstrates the efficacy of the approach in exploring similarity uncertainty and optimizing similarity measurement.

We plan to generalize the joint distribution of similarity uncertainty in different dimensions for further analysis. We aim to design an efficient framework to analyze similarity uncertainty across time steps. We also plan to promote the weight adaptive algorithm to other high-dimensional scalar data.

**Acknowledgements** We are grateful for the valuable feedback and comments provided by the anonymous reviewers. This research is partially supported by the National Natural Science Foundation of China (Grant Nos. 61972221, 61572274) and NNW2018-ZT6B12 (National Numerical Windtunnel project).

## References

Berthold MR, Hall LO (2003) Visualizing fuzzy points in parallel coordinates. IEEE Trans Fuzzy Syst 11(3):369–374

Bezdek JC (2013) Pattern recognition with fuzzy objective function algorithms. Springer, Berlin

Bonneau G-P, Hege H-C, Johnson CR, Oliveira MM, Potter K, Rheingans P, Schultz T (2014) Overview and state-of-the-art of uncertainty visualization. In: Shriver B (ed) Scientific visualization. Springer, Berlin, pp 3–27

Du J, Kang Z (2014) A survey on forecasters view about uncertainty in weather forecasts. Adv Meteorol Sci Technol 4(1):60–69

Ferstl F, Bürger K, Westermann R (2016a) Streamline variability plots for characterizing the uncertainty in vector field ensembles. IEEE Trans Vis Comput Gr 22(1):767–776

Ferstl F, Kanzler M, Rautenhaus M, Westermann R (2016b) Visual analysis of spatial variability and global correlations in ensembles of iso-contours. In: Hauser H, Benes B (eds) Computer graphics forum, vol 35. Wiley, New York, pp 221–230

Ferstl F, Kanzler M, Rautenhaus M, Westermann R (2017) Time-hierarchical clustering and visualization of weather forecast ensembles. IEEE Trans Visual Comput Gr 23(1):831–840
Gasch AP, Eisen MB (2002) Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. Genome Biol 3(11):research0059.1
Glahn HR, Lowry DA (1972) The use of model output statistics (MOS) in objective weather forecasting. J Appl Meteorol 11(8):1203–1211
Gong C, Chen L, Zhu Z (2016) A visualization system for calibrating multimodel ensembles in weather forecast. J Vis 19(4):769–782
Hamill TM, Whitaker JS (2006) Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. Mon Weather Rev 134(11):3209–3229
Hamill TM, Bates GT, Whitaker JS, Murray DR, Fiorino M, Galarneau TJ Jr, Zhu Y, Lapenta W (2013) Noaa's second-generation global medium-range ensemble reforecast dataset. Bull Am Meteorol Soc 94(10):1553–1565
Hamill TM, Scheuerer M, Bates GT (2015) Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. Mon Weather Rev 143(8):3300–3309
Hou D, Charles M, Luo Y, Toth Z, Zhu Y, Krzysztofowicz R, Lin Y, Xie P, Seo D-J, Pena M et al (2014) Climatology-calibrated precipitation analysis at fine scales: statistical adjustment of stage IV toward CPC gauge-based analysis. J Hydrometeorol 15(6):2542–2557
Leutbecher M, Palmer TN (2008) Ensemble forecasting. J Comput Phys 227(7):3515–3539
Liao H, Chen L, Song Y, Ming H (2016) Visualization-based active learning for video annotation. IEEE Trans Multimed 18(11):2196–2205
Liao H, Wu Y, Chen L, Chen W (2018) Cluster-based visual abstraction for multivariate scatterplots. IEEE Trans Vis Comput Gr 24(9):2531–2545
Liao H, Wu Y, Chen L, Hamill TM, Wang Y, Dai K, Zhang H, Chen W (2015) A visual voting framework for weather forecast calibration. In: 2015 IEEE scientific visualization conference (SciVis), pp 25–32
Mao J, Jain AK (1995) Artificial neural networks for feature extraction and multivariate data projection. IEEE Trans Neural Netw 6(2):296–317
Mirzargar M, Whitaker RT, Kirby RM (2014) Curve boxplot: generalization of boxplot for ensembles of curves. IEEE Trans Vis Comput Gr 20(12):2654–2663
Pfaffelmoser T, Reitinger M, Westermann R (2011) Visualizing the positional and geometrical variability of isosurfaces in uncertain scalar fields. In: Hauser H, Benes B (eds) Computer graphics forum, vol 30. Wiley, New York, pp 951–960
Pfaffelmoser T, Mihai M, Westermann R (2013) Visualizing the variability of gradients in uncertain 2d scalar fields. IEEE Trans Vis Comput Gr 19(11):1948–1961
Potter K, Wilson A, Bremer P-T, Williams D, Doutriaux C, Pascucci V, Johnson CR (2009) Ensemble-vis: a framework for the statistical visualization of ensemble data. In: 2009 IEEE international conference on data mining workshops, pp 233–240
Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. Mon Weather Rev 133(5):1155–1174
Rueda L, Zhang Y (2006) Geometric visualization of clusters obtained from fuzzy clustering algorithms. Pattern Recognit 39(8):1415–1429
Ruspini EH (1969) A new approach to clustering. Inf Control 15(1):22–32
Sanyal J, Zhang S, Dyer J, Mercer A, Amburn P, Moorhead R (2010) Noodles: a tool for visualization of numerical weather model ensemble uncertainty. IEEE Trans Vis Comput Gr 16(6):1421–1430
Sharko J, Grinstein G, Marx KA (2008) Vectorized radviz and its application to multiple cluster datasets. IEEE Trans Vis Comput Gr 14(6):1427–1444
Sharko J, Grinstein G (2009) Visualizing fuzzy clusters using radviz. In: 2009 IEEE 13th International conference information visualisation, pp 307–316
Wang Y, Fan C, Zhang J, Niu T, Zhang S, Jiang J (2015) Forecast verification and visualization based on Gaussian mixture model co-estimation. In: Hauser H, Benes B (eds) Computer graphics forum, vol 34. Wiley, New York, pp 99–110
Wang J, Liu X, Shen H-W, Lin G (2017) Multi-resolution climate ensemble parameter analysis with nested parallel coordinates plots. IEEE Trans Vis Comput Gr 23(1):81–90
Whitaker RT, Mirzargar M, Kirby RM (2013) Contour boxplots: a method for characterizing uncertainty in feature sets from simulation ensembles. IEEE Trans Vis Comput Gr 19(12):2713–2722
Zadeh LA, Klir GJ, Yuan B (1996) Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers, vol 6. World Scientific, Singapore
Zhao Y, Luo F, Chen M, Wang Y, Xia J, Zhou F, Wang Y, Chen Y, Chen W (2018) Evaluating multi-dimensional visualizations for understanding fuzzy clusters. IEEE Trans Vis Comput Gr 25(1):12–21