A Visual Voting Framework for Weather Forecast Calibration

Hongsen Liao[a,*], Yingcai Wu[b,†], Li Chen[a,‡], Thomas M. Hamill[c,§]
Yunhai Wang[d,¶], Kan Dai[e,‖], Hui Zhang[a,**], Wei Chen, *Member, IEEE*[b,††]
[a]School of Software, Tsinghua National Laboratory for Information Science and Technology, Tsinghua University
[b]State Key Lab of CAD & CG, Zhejiang University
[c]NOAA Earth System Research Lab, Physical Sciences Division
[d]Shenzhen Institutes of Advanced Technology, Chinese Academy of Science
[e]National Meteorological Center of CMA

ABSTRACT

Numerical weather predictions have been widely used for weather forecasting. Many large meteorological centers are producing highly accurate ensemble forecasts routinely to provide effective weather forecast services. However, biases frequently exist in forecast products because of various reasons, such as the imperfection of the weather forecast models. Failure to identify and neutralize the biases would result in unreliable forecast products that might mislead analysts; consequently, unreliable weather predictions are produced. The analog method has been commonly used to overcome the biases. Nevertheless, this method has some serious limitations including the difficulty in finding effective similar past forecasts, the large search space for proper parameters and the lack of support for interactive, real-time analysis. In this study, we develop a visual analytics system based on a novel voting framework to circumvent the problems. The framework adopts the idea of majority voting to combine judiciously the different variants of analog methods towards effective retrieval of the proper analogs for calibration. The system seamlessly integrates the analog methods into an interactive visualization pipeline with a set of coordinated views that characterizes the different methods. Instant visual hints are provided in the views to guide users in finding and refining analogs. We have worked closely with the domain experts in the meteorological research to develop the system. The effectiveness of the system is demonstrated using two case studies. An informal evaluation with the experts proves the usability and usefulness of the system.

Keywords: Weather forecast, analog method, calibration, majority voting, visual analytics.

Index Terms: Human-centered computing [Visualization]: Visualization application domains—Geographic visualization;

1 INTRODUCTION

Numerical weather prediction (NWP) has been practiced operationally at many forecast centers since the middle of the 20th century [17]. Multiple simulations of the future weather are created from slightly perturbed initial states to simulate the uncertainty contributed by imperfections in the forecast model itself. Given the chaotic nature of the atmosphere, even small initial differences increase rapidly with time; thus, *probabilistic forecasts* of future states are more theoretically tenable than deterministic forecasts for domain experts [2]. Despite the rapid progress in the development of NWP, problems still exist. The forecasts may be systematically too warm or too cold, too wet or too dry. Meanwhile, it is a common occurrence to have ensembles of forecasts that have too little spread, and the true state lies outside the ensemble range. *Calibration* is then the process to address the deficiencies to improve the forecasts.

Many meteorological groups are exploring the "post-processing" methods for ensemble forecast. Of particular interest are the analog methods, which have shown promising results [9]. The idea of the method is to find past forecasts (i.e., analogs) in a geographically limited region that resemble the current forecast. A probabilistic estimate of the weather is then formed from the observed data at the time of the past forecasts. However, different similarity measures can be used to find the similar past forecasts and none is regarded perfect [9]. Meanwhile, no effective solution has been proposed to deal with the threshold problem in the analog methods.

Inspired by the majority voting method commonly used in machine learning, we propose a novel user-steered voting framework to refine the analogs obtained using three widely used analog methods. The framework consists of three interactive, coordinated views; each corresponds to one analog method. Through interactions on the coordinated views, users cast their votes for analogs. A two-step solution is proposed to address the threshold problem of the analog methods as well as ensure the reliability of the final forecasts. Based on the framework, we work closely with the domain experts to develop a visual analytics system that empowers the experts to calibrate the forecasts effectively. Two case studies and user evaluations demonstrate the usability of the system.

The main contributions of this paper are as follows:

- A characterization of the calibration problem in the operational weather forecasting.
- A novel voting framework that effectively combines different analog-based methods using three coordinated visualizations.
- A visual analytics system based on the voting framework that assists forecasters in the calibration using analog methods.

2 RELATED WORK

Meteorological Data Visualization Meteorological data visualization has become an important topic since tens of years ago [13, 29]. Many practical visualization tools have been developed to support domain research, such as Vis5D [12], Ferret [10] and GRADS [1]. However, a gap still exists between the advanced visualizations and domain work in climate research [28].

Several visual analytics systems have been developed through close collaborations with meteorologists. Kehrer et al. [16] have proposed a novel visualization pipeline to support hypothesis generation from large scale climate data. To visualize climate variability changes, Janicke et al. [15] have used the wavelet analysis to perform the multi-scale visualization. Lundblad et al. [18] have developed an application to identify significant trends and patterns within

*e-mail:liaohs082@gmail.com
†e-mail:ycwu@cad.zju.edu.cn
‡e-mail:chenlee@tsinghua.edu.cn. Li Chen is the corresponding author.
§e-mail:tom.hamill@noaa.gov
¶e-mail:cloudseawang@gmail.com
‖e-mail:daikan1998@163.com
**e-mail:huizhang@tsinghua.edu.cn
††e-mail:chenwei@cad.zju.edu.cn

weather data using interactive information visualization techniques. Doraiswamy et al. [3] have presented a framework for the identification and tracking of cloud systems. These works assist the domain experts in understanding the atmospheric state and contribute to many practical applications. However, there is little visualization work which supports the weather forecast calibration.

Uncertainty Visualization for Meteorological Ensemble Data Many researchers have contributed to the issue of uncertainty visualization for meteorological data. MacEachren et al. [19] have provided a detailed introduction to the visualization of geo-spatial information uncertainty. Pang et al. [20] have done a lot of research on geo-spatial data visualization. To reveal the probabilistic nature of the data, Potter et al. [24] have described a framework that visualizes the numerical weather ensemble data with linked views. Sanyal et al. [26] have designed an informative ribbon and glyphs to visualize the uncertainty in multiple numerical weather model. Pöthkow et al. [23] measure the positional uncertainty of isocontours with the isocontour density and the level crossing probability field. Pfaffelmoser et al. [21] have provided a color mapping and glyph based visualization solution for visualizing the variability of gradients in 2D scalar fields. Using the Lagrangian-based distance metric, Guo et al. [7] have evaluated and visualized the variation that exists in ensemble runs. Poco et al. [22] have proposed an iterative visual reconciliation solution for similarity spaces in climate model research. Whitaker et al. [30] have introduced the contour boxplots to visualize and explore the contours in ensemble fields.

However, most of the visualization work have not discussed about the consistency between the forecast data and the observed data, which is one of the domain experts' main concerns in the weather forecasting. Moreover, the previous methods cannot be directly applied to visualize the large scale historical data.

Calibration in Meteorological Research The goal of the calibration is to detect and correct the potential errors in the weather forecasts before publications. Glahn et al. [4] have used the linear regression, which is also known as "Model Output Statistics", to calibrate the forecast. Raftery et al. [25] have used Bayesian model averaging to calibrate forecast ensembles. Gneiting et al. [5] have detailedly discussed the probabilistic forecast and the calibration. Although these methods have demonstrated usefulness, they all lack interactive tools which can effectively integrate domain knowledge into the statistical processes.

3 Domain Tasks and Dataset Description

This section briefly discusses the domain tasks and describes the data used in our system.

3.1 Domain Tasks

Three main tasks in forecast calibration have been identified through close collaboration with forecasters, observations on their routine work, and detailed discussions about their working flow.

- T1 **Generating Initial Forecast** An initial forecast is produced using a post-processing method, such as the analog method.
- T2 **Detecting Regions of Interest (ROI)** The initial forecast derived after the post-processing step could have potential biases. Thus, forecasters need to detect ROIs where biases exist.
- T3 **Applying Detailed Calibrations** Calibrations are applied to the detected ROIs statistically or manually according to the professional knowledge of the experts.

3.2 Dataset Description

The reforecast data and the observed data are used in this study.

Reforecast Data The reforecasts are from the US NCEP Global Ensemble Forecast System (GEFS). The GEFS reforecasts are grid data with a resolution of $\sim 0.5^\circ$. They comprise an 11-member ensemble of forecasts which run every day. The reforecasts span from 1985 to present. A variety of forecast variables are produced. Detailed descriptions of the reforecasts can be found in [8].

Observed Data The observed data are the NCEP Climatology-Calibrated Precipitation Analysis (CCPA) data. The CCPA data are grid data with a resolution of 0.125°. The data spans from 2002 to present and covers the continental US. The analysis data are saved every 6 hours. In practical usages, the analysis data are regarded as the ground truth description of the real weather state. Detailed descriptions of the CCPA data can be found in [14].

4 Visual Voting Framework

In this section, we introduce a set of analog methods, and present the visual voting framework used to combine the methods and support the weather forecast calibration, which is the task of **T3**.

4.1 Analog Methods

The analog method has two successive main steps: the step of analog retrieval and the step of probabilistic forecast generation. The analog retrieval step is designed to find the past forecasts with similar data in the ROI. The difference between the current forecast and the past one is defined as the root mean square (RMS) difference of a variable or the weighted sum of the RMS differences (aggregated RMS differences) of several variables. Meanwhile, the mean of the ensemble forecasts is used for the calculation of variable RMS differences. Thereafter, N analogs with the smallest differences or the analogs that satisfy a specified threshold constraint are selected. The probability distribution of the observed data from the corresponding dates of the analogs is used to provide an estimate of the event probability. For example, the probability of the precipitation exceeding 5 mm is 50% if the observed data of 10 out of 20 selected dates exceed 5 mm. More details can be found in [9].

The analog method has many variants with different ROI sizes and difference norms (a blend match of several forecast variables or an independent variable). In our system, three widely used variants are selected, including global to local similarity measurements.

- C1 **RMS Difference of Aggregated Variables in a Large Region:** This measure is adopted to ensure that the selected dates share a similar global atmospheric state of the meteorological event with the current date.
- C2 **RMS Difference of Aggregated Variables in a Small Region:** A small region is the smallest grid cell with four grid points in the corners. This measure is one of the most straightforward measures and is mostly used in the practice.
- C3 **RMS Difference of Separate Variables in a Small Region:** This is a *detailed* measure. A particular variable could be much more significant in some scenarios for the calibration.

The variants retrieve analogs from different meaningful perspectives, but these may lead to rather diverse probabilistic forecasts when employing different thresholds.

4.2 Visual Voting

Majority voting is one of the most fundamental and popular ensemble methods for classification. Inspired by this method, we regard each variant of the analog methods as a classifier for the past dates. A past date can be classified as selected or unselected.

Although majority voting is a well-understood method, it is tedious and error prone to manually set thresholds for each of the variants independently. Therefore, we need to provide suggestions to guide the user in making decision for the thresholds. Wherein, we design a two-step solution: iterative threshold searching for **C1** and **C2**, followed by threshold suggestion and adjustment for **C3**, as shown in Fig. 1. Among the three variants of the analog methods, **C1** and **C2** provide global and local aggregated descriptions for the similarity, respectively, and **C3** is a detailed measure. The goal of this design is to ensure that the selected dates are similar to the current one both globally and locally. Then adjustments can be made

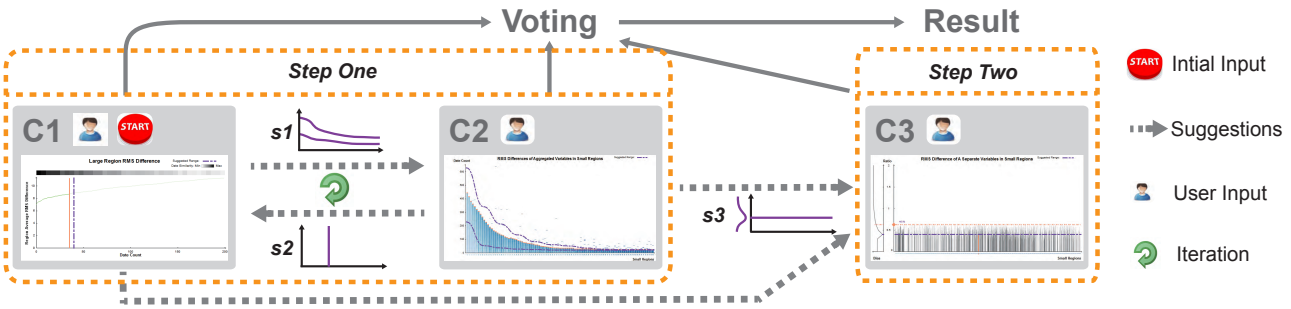

Figure 1: Visual Voting Framework. The framework comprises two main steps. The first step is the iterative threshold searching for **C1** and **C2**. The second step is the threshold suggestion and adjustment for **C3**. After thresholds are set for all the three classifiers, majority voting is applied in order to achieve more reliable forecasts. Visual feedback of the result is presented to assist threshold setting through the whole framework.

according to the detailed measure of **C3**. Meanwhile, suggestions are provided for each of the measures. In the framework, the user is expected to compare the suggestions with their expectations and then determine the acceptable thresholds according to his expertise.

Among the three classifiers, the threshold for **C1** is the most intuitive for the domain experts to set manually. Therefore, our solution begins with the initial input for **C1**, as indicated in Fig. 1.

4.2.1 Iterative Threshold Searching for C1 and C2

This step aims to determine the proper thresholds for **C1** and **C2**. After the initial input is set, a classification result can be obtained using **C1**. According to the suggestion from the domain experts, the selected forecasts should be similar to the current one both globally and locally. Therefore, the dates that selected using **C2** should cover a specific portion (namely, covering rate) of those selected using **C1**. The covering rate for small region r_i is defined as follows:

$$CR_{r_i} = \frac{1}{T} \sum_{t \in Q} \sigma(r_i, t, \beta_i) \tag{1}$$

where Q is the set of selected dates using **C1**, T is the number of dates in Q, β_i is the threshold for the small region, and $\sigma(r_i,t,\beta_i) = 1$ if date t is selected using **C2** for r_i, otherwise $\sigma(r_i,t,\beta_i) = 0$. Given a covering rate (namely, CR_{r_i}, from the user), the RMS differences of the small region are sorted first. The threshold β_i is then the smallest threshold by which the selected dates of **C2** cover the specified rate of the ones from **C1**. In our system, we experimentally use a covering rate range of 0.6 to 0.8 to maintain the similar dates selected using **C1** and leverage the variance between the dates selected using **C1** and **C2**. Moreover, a lower bound of 20 and a upper bound, which is the higher value between the suggested threshold and the one used for the large region, are used to ensure the effectiveness of the result from **C2**, as indicated by $S1$ in Fig. 1.

The user can then estimate the possible proper thresholds for **C2** based on their expertise and the suggestion range. Subsequently, the similarity $S_{R,t}$ of a date t for a large region R can be computed using small regions, which choose t as one of the most similar dates.

$$S_{R,t} = \sum_{r_i \in R} \frac{\sigma(r_i,t,\beta_i)}{n_i} \tag{2}$$

where r_i indicates a small region within R, β_i is the selected threshold for the small region, n_i is the selected analog number for r_i under the threshold of β_i, and $\sigma(r_i,t,\beta_i) = 1$ if date t is selected using **C2** for r_i, otherwise $\sigma(r_i,t,\beta_i) = 0$. Thereafter, the similarity is normalized and used as another meaningful cue for the date selection in the large region. The suggested selected dates for the large region are then those which satisfy the similarity with smallest RMS differences among all the past dates, as indicated by $S2$ in Fig. 1. In our system, we experimentally use the similarity of 0.6 to ensure the similarity of the selected dates for the large region.

New thresholds can be set iteratively until good similarities from **C1**, and satisfactory covering rates from **C2** are obtained.

4.2.2 Suggestion and Adjustment for C3

The next step is to adjust the threshold for **C3**. In our implementation, we use the ratio of the variable RMS difference and the current forecast variable value as the threshold. The ratio is a better measure compared with the absolute RMS difference. For example, the RMS difference of 5 mm has different impacts under the precipitation of 3 mm and 30 mm, but the ratio can handle it well.

The suggested threshold θ of the ratio values for **C3** is then estimated by minimizing the bias between the selected dates by **C3** and those by **C1** and **C2**, as indicated by $S3$ in Fig. 1.

$$min \sum_{t} \sum_{r_i \in R} abs(C^3_{r_i}(t,\theta) - C^1_R(t)) + abs(C^3_{r_i}(t,\theta) - C^2_{r_i}(t)) \tag{3}$$

where $C^3_{r_i}(t,\theta)$ is the class label of **C3** for the small region r_i on date t, $C^1_R(t)$ is the class label of **C1**, and $C^2_{r_i}(t)$ is the class label of **C2**. The labels of the selected and unselected dates are set to 1 and 0, respectively. We sample the ratio value with a small step size to achieve a set of bias values, and the ratio value with the minimum bias (Formula 3) is used as the suggestion. Then the domain user can leverage the suggested threshold and adjust the threshold for **C3** to filter past dates whose ratio values are out of scope. When multiple variables are employed, the threshold suggestion and adjustment are conducted independently for each variable. As a result, the selected dates for **C3** are the dates selected by all the variables.

Through this voting framework, relatively proper thresholds for each of the analog methods can be achieved through the user's supervisions and interactions. Combined with the majority voting, the framework can produce more reliable results for domain usages.

5 System and Visualization Design

In this section, we provide an overview of the system design and then introduce the detailed visualization design in the system.

5.1 System Design

Our system is designed to assist the domain experts' routine work. Thus, a companion system workflow is designed according to their calibration workflow, as illustrated in Fig. 2. According to the utilities of different views and the goal of each task introduced in Section 3.1, we divide our system workflow into five stages:

1.Data Overview: This stage helps domain experts obtain an overview of the ensemble data.

2.Post-processing: This stage generates initial probabilistic forecasts using predefined parameters and completes the task of **T1**.

3.ROI Detection: An RMS difference glyph is designed to assist in detecting ROIs, where the initial probabilistic forecasts might need refinements. Through this process, the task of **T2** is supported.

4.Visual Calibration: Coordinated views are designed to support the visual voting framework. This stage provides support for the task of **T3**, which is our main focus in this study.

5.Comparison and Event Analysis: The adapted forecasts generated after the visual calibration are compared with the initial ones

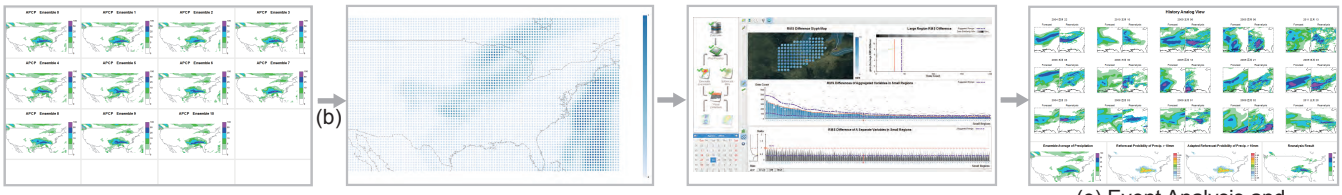

Figure 2: Overview of the system workflow. (a) Overview of the numerical ensemble data. (b) Post processing using the analog method. (c) Detect ROIs where calibrations may be required. (d) Visual calibration. (e) Forecast comparison and similar historical event analysis.

to show how the visual calibration works. Meanwhile, the most similar historical events are presented to verify the forecast.

A calibration task can be completed through all the five stages directed by a user guideline (Fig. 4(e)). We focus on the ROI Detection stage and the Visual Calibration stage in the following. The usage of other stages are demonstrated in the case studies.

5.2 ROI Detection

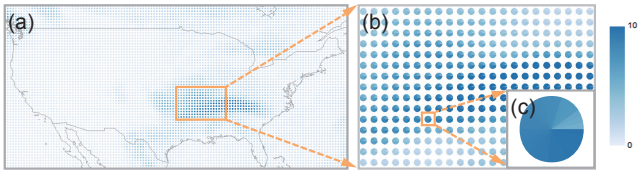

Figure 3: RMS Difference Glyph: (a) Overview of the whole map. (b) Region of interest. (c) Glyph for a small region.

This RMS difference glyph view serves the stage of ROI detection (Fig. 3(a)). The aggregated RMS differences of the selected analogs for a small region are conveyed through a circular glyph as illustrated in Fig. 3(c). The angle is cut into N parts to visualize the sorted RMS differences of the N selected analogs. Color is used to encode the RMS difference. We choose the color sequence from colorbrewer2.org [11] to ensure the linear expression of RMS difference values. The glyph is placed according to the position of the small region on the map. A user can zoom in to obtain a detailed view of the RMS differences or zoom out to achieve an overview of the whole map (Fig. 3(b)). Furthermore, the user can brush a region for the subsequent analysis stages. With this view, a user can efficiently locate ROIs where analogs with larger RMS differences are selected, and further calibrations might be needed. In this design, the glyph is used for its rich applications in domain research. Meanwhile, the view conforms to the information seeking mantra of "Overview first, zoom and filter, then details on demand" [27].

5.3 Visual Calibration

Coordinated views and interactions are designed to implement the visual voting framework.

5.3.1 View Design

The calibration view comprises one geographical view (Fig. 4(a)) and three views that correspond to the three selected analog methods (Fig. 4(b) for **(C1)**, Fig. 4(c) for **C2** and Fig. 4(d) for **C3**).

Geographical View: The geographical view is the RMS difference glyph view laid on a geographical map (Fig. 4(a)). This view presents the glyphs for the selected ROI from the previous stage.

Region RMS Difference View: The region RMS difference view is used to visualize the large region RMS differences and similarities of past dates (Fig. 4(b)) for **C1**. This view comprises a line chart (Fig. 4(b1)) and a color bar (Fig. 4(b2)).

The line chart visualizes the sorted RMS differences of the large region. The x axis of the chart conveys the sorted dates based on the RMS differences, and the y axis presents the difference values. The RMS differences provide visual cues for the threshold setting, and the orange line shows the current selected RMS difference threshold (selected analog number). The color bar is used to visualize the similarity encoded using the selected dates from **C2**. The sorted dates are separated into a series of bins. Each bin is encoded with gray color to represent the average similarity value of the corresponding dates. The suggested threshold is then indicated by a purple line as shown in Fig. 4(b1). In this view, the user is expected to select a threshold smaller than the suggested value according to the increasing trend of the line chart and colors from the color bar.

Small Region RMS Difference View: This view is a pixel bar chart used to convey the aggregated RMS differences for the small regions within the ROI (Fig. 4(c)). Each bar represents a small region and the RMS differences of all the past dates are encoded into the pixel color in the y axis direction (see Figure 7(a)). The view is designed to support the threshold setting for **C2** (Fig. 6(a)) and visualize the final voting results (Fig. 6(b) and 6(c)).

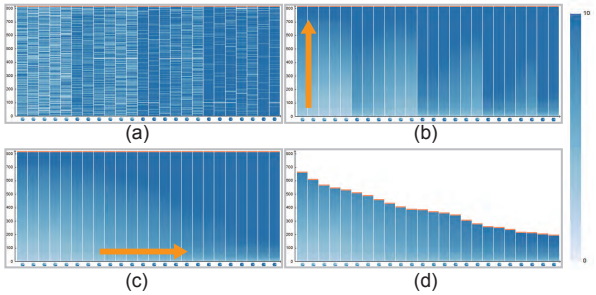

Figure 5: Sorting for the pixel bar chart: (a) Initial data visualization without sorting. (b) Sorting according to the aggregated RMS differences for each small region. (c) Sorting according to the average aggregated RMS differences for all the small regions. (d) A typical threshold setting for the sorted pixel bar chart.

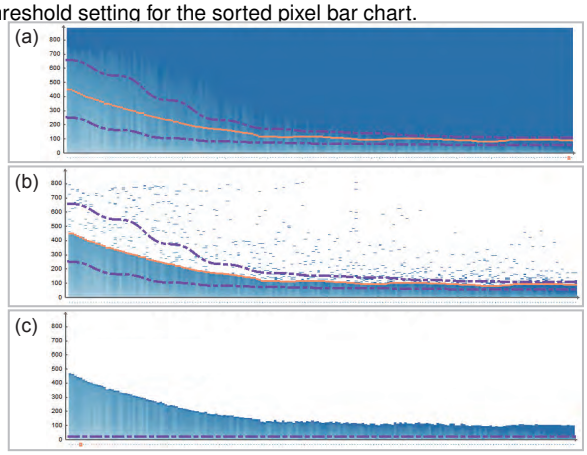

Figure 6: Different utilities of the pixel bar chart: (a) Supporting threshold setting. (b) Visualizing the voting results. (c) Visualizing the voting results by accumulating selected dates.

To support intuitive and efficient threshold setting for each small region, sorting is adopted. First, we sort the aggregated RMS differences for each small region (see Fig. 5(a)). In Fig. 5(b), the date with the smallest RMS difference is placed at the bottom of each bar, and the largest at the top. Then inspired by the common experience that the more unusual the prediction forecast is, the smaller N value should be employed, we sort the bars horizontally according

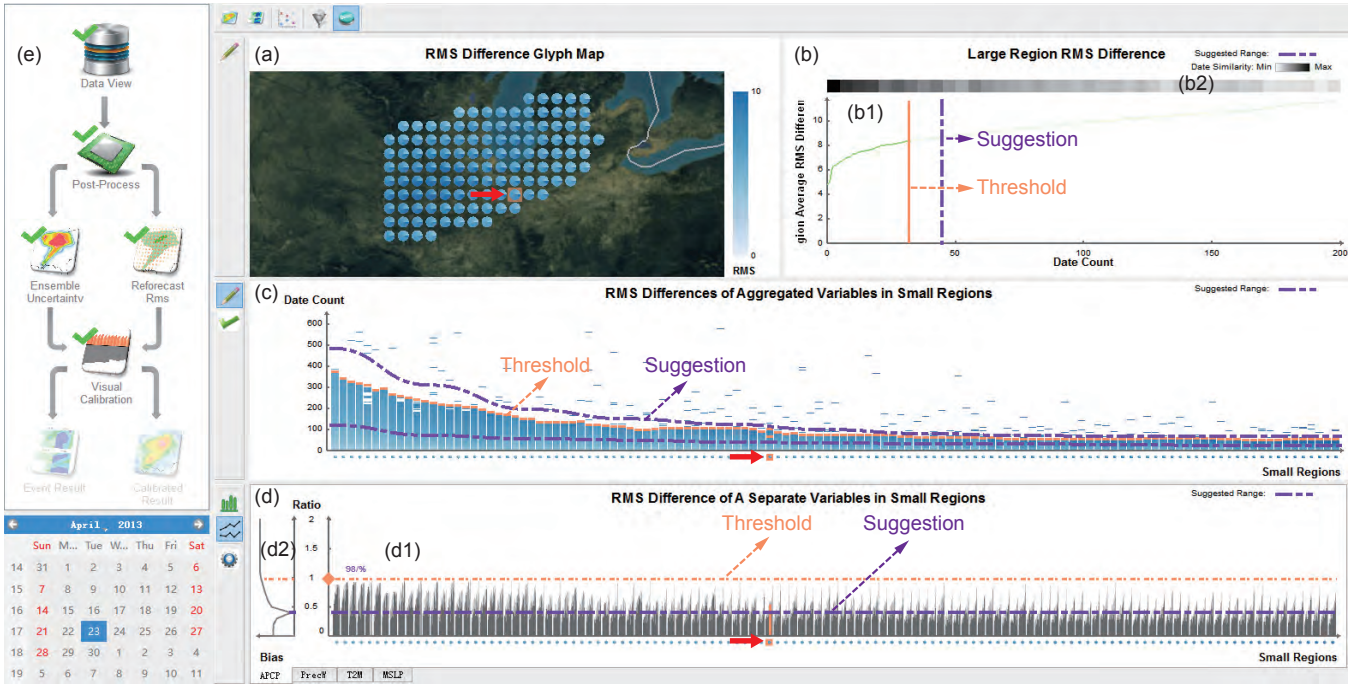

Figure 4: Coordinated views for the calibration stage: (a) The geographical view. (b) The region RMS difference view. (c) The small region RMS difference view (d) The small region variable RMS difference view. (e) The user guide.

to the average aggregated RMS differences of the small regions. In Fig. 5(c), the average RMS differences of the bars increase from the left to the right. A typical threshold selection for the small regions can then be set, as shown in Fig. 5(d). Meanwhile, the suggestions from the classification result of **C1** should be presented to the user, which are highlighted by two purple lines in Fig. 4(c).

To visualize the final voting results, the pixels that represent unselected dates are hidden (Fig. 6(b)), and the remaining pixels can be accumulated to provide a more intuitive impression of the number of selected dates for the small regions (Fig. 6(c)). The suggested lower bound for the threshold is also indicated by a purple line in the view, which is a visual suggestion for the whole framework.

In this view, the pixel bar chart is selected because the chart is intuitive and simple, and can visualize a large amount of data.

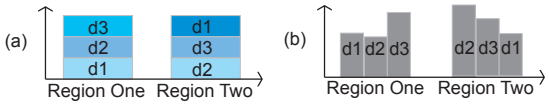

Figure 7: Two illustrative layouts of the bar charts, which contain three selected dates for each of the two regions in (a) the RMS difference view and (b) the variable RMS difference view, respectively.

Small Region Variable RMS Difference View: This view is designed to support the classification using **C3** (Fig. 4(d)). It comprises a bar chart (Fig. 4(d1)) for visualizing the ratio values and a line chart (Fig. 4(d2)) for the bias defined by Formula 3. The bar chart shares the same axis with the pixel bar chart in the Small Region RMS Difference View. For each small region, bars that encode ratio values of past dates are horizontally placed in the x axis with the same order utilized by the y axis of the pixel bar chart, as illustrated in Fig. 7(b). In case of a majority voting of three variants, only those dates that have been selected using **C1** or **C2** are presented. Through the bar chart visualization, outliers can be detected and filtered by adjusting the threshold for the ratio value more easily, compared to the pixel bar chart. The line chart (Fig. 4(d2)) visualizes the biases between the dates selected using **C3** and those selected using **C1** and **C2** under different thresholds for **C3**. The user can achieve a clear view of the bias distribution and estimate the proper threshold while avoiding a huge bias. The suggestion is indicated by a purple line in Fig. 4(d).

5.3.2 Interactions and View Coordination

The four small views are all linked. Double clicking and brushing are supported to enable the setting of the thresholds for **C1** and **C2**, respectively. Suggestions are updated immediately after the user input is completed. The iterative threshold searching for **C1** and **C2** can be efficiently conducted through these interactions. When adjusting the threshold for **C3**, the final results are also updated in real time, as shown in Fig. 6(c). Therefore, the domain user can achieve a clear view of the final voting results and leverage the proper threshold. Highlights are also designed to enhance the linking among views as indicated by the red arrows in Fig. 4, which can effectively enhance the view coordination.

6 Evaluation and Discussion

Two case studies are presented in this section to exhibit the usability of our system. The first one demonstrates the common workflow of our system to help calibrate a forecast. The second one demonstrates the use of our system in detecting an unusual forecast and providing supports for the verification based on historical data.

Experiments have been performed with domain experts to calibrate the probabilistic forecasts of the total accumulated precipitation for 24 hours. Three forecast variables are adopted, namely, total accumulated precipitation (APCP), precipitable water (PREW), and temperature at 2 m (T2M), which are provided by our domain collaborators. Our system uses the data from 2002 to 2013, and only those dates within a window size of 70 that center on the analyzed day are adopted for each year. In addition, the weighted summary of APCP and PREW is used as the aggregated variable with the weights of 0.7 and 0.3, respectively, which is also used by one of our collaborators in his domain research.

6.1 Case One

The data used are from April 27, 2013. After loading the data, our system provides an overview of the data, as is illustrated in Fig. 8(a). A heavy precipitation is recorded in the United States. Post-processing is then performed with the aggregated variables, and initial probabilistic forecasts are generated using the analog method of **C2**. In the experiment, two results with different analog

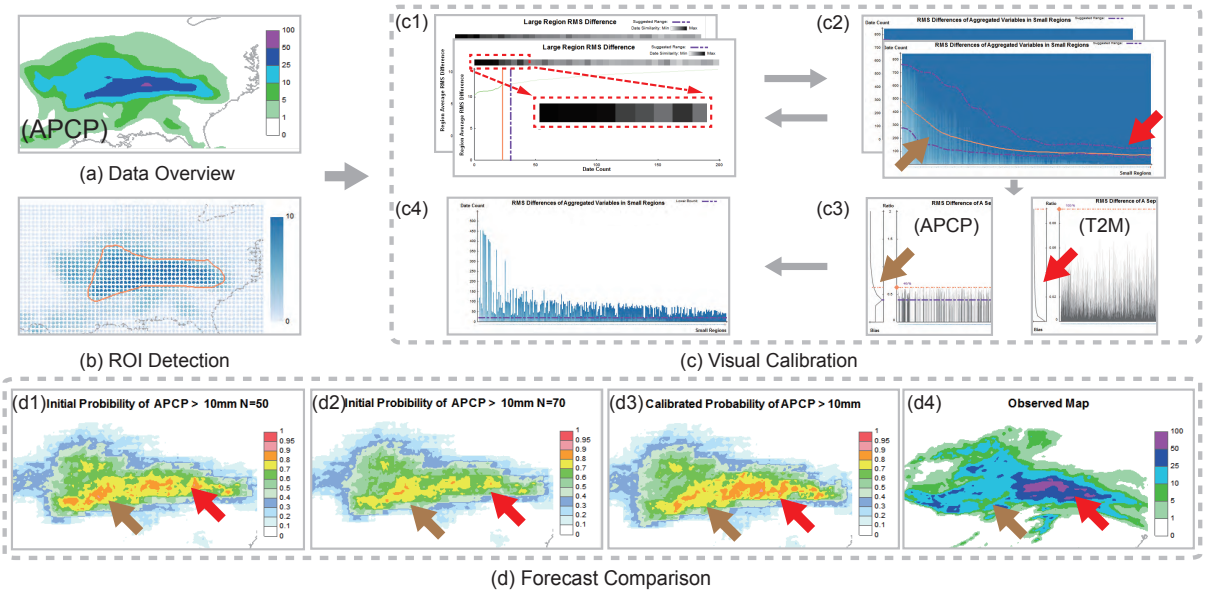

Figure 8: Case One: (a) The data overview stage provides a 2D plot view of the ensemble data. (b) The ROI detection stage helps users detect the ROI, wherein further calibration is required. (c) The visual calibration stage applies the visual voting framework, and users are included to manipulate the analog methods. (d) The forecast comparison stage supports the comparison between the initial forecasts from the post-processing stage and the calibrated ones. The latter can better reflect the observed data in the probability distribution with a high probability in the region with high precipitation, as indicated by the red arrow in (d3) and (d4).

numbers, namely, 50 and 70, are generated, as shown in Fig. 8(d1) and Fig. 8(d2), respectively. The analog number 50 is suggested by one of our domain collaborators based on his previous research. Thereafter, an ROI with high RMS differences is brushed for the further calibration, as illustrated in Fig. 8(b). In case of high RMS differences, biases may occur in the probabilistic forecast of the region. The visual calibration stage begins with an estimation for the threshold of **C1**. Given that the threshold does not necessarily have to be accurate, and that it can be refined through successive iterations, an analog number of 50 is also set as the initial input. Threshold searching iterations are then performed, as shown in Fig. 8(c1) and Fig. 8(c2). For small regions with low RMS differences, high analog numbers are applied, as indicated by the brown arrow in Fig. 8(c2). For the small regions indicated by the red arrow in Fig. 8(c2), RMS differences are mostly over 10 mm, which are high for the current weather. Therefore, the thresholds for these small regions are nearly at the suggested lower bound. In the color bar which encodes the similarity of past dates in the large region, a clear distribution of the similarity is presented, as shown in the red borders in Fig. 8(c1). The threshold for **C1** can then be set to ensure that the selected dates all possess high similarity values. After the refinement iterations, the thresholds are adjusted for the variables. The bias distribution and the suggested threshold for APCP are evident, as indicated by the brown arrow in Fig. 8(c3). However, the minimum bias for T2M is reached with a ratio of nearly 0, as indicated by the red arrow in Fig. 8(c3). This finding indicates that the most similar dates based on APCP and PREW differ significantly from those based on T2M, which confirms the previous research conclusion that precipitation forecast accuracy decreases by including T2M when finding analog dates for the precipitation [9]. The thresholds are then adjusted to filter dates that are clear outliers and decrease the distance between the current threshold for APCP and the suggested threshold, while ensuring that the remaining numbers of selected dates for the small regions are mostly above the lower bound, as shown in Fig. 8(c4). Thereafter, the calibrated probabilistic forecast is generated, as shown in Fig. 8(d3). Meanwhile, the observed map for the same day is shown in Fig. 8(d4). Through these steps, the calibration for this region is completed by the probabilistic forecast generation, and other kinds of forecasts can also be achieved through the calibrated forecasts.

As indicated in the observed map, the precipitation in the region indicated by the red arrow is evidently heavier than that in the region indicated by the brown arrow. Hence, the calibrated probabilistic forecast generated using the visual voting framework can reflect the precipitation distribution better than the initial forecast when the probability is higher in the region indicated by the red arrow than in the region indicated by the brown arrow.

6.2 Case Two

The data used are from August 25, 2013. A heavy precipitation is recorded in Arizona, US (highlighted in the red rectangle in Fig. 9(a)). However, the RMS differences in this region are high. The ROI is then selected from the RMS difference glyph view as illustrated in Fig. 9(b). During the visual calibration stage, the suggested analog number for the large region continues to decrease when we perform threshold searching iterations, as shown by the red arrows in Fig. 9(c1) to 9(c3). The suggested analog number only stops decreasing when the suggested range for the small regions reaches the lower bound set in our system, as indicated by the red arrow in Fig. 9(c4). By this time, the suggested analog number for the large region is nearly below 10. Thus, the current forecast can be regarded as a probable unusual forecast even in the long history. Thereafter, the most similar analogs can be viewed, and two of these analogs are shown in Fig. 9(d1) and Fig. 9(d2). For each analog, the left part of the figure is the forecast of the analog, whereas the right part is the corresponding observed data. Based on the analogs, we can conclude that the forecast is mostly expected to be higher than the observed weather in the region. This conclusion can also be achieved through the probabilistic forecast and confirmed by the observed data for the same day, as shown in the red borders in Fig. 9(d3) to 9(d4). The probability of the precipitation is low in that region, but the forecast precipitation in Fig. 9(a) is high. Moreover, the observed data shows a lower precipitation than the forecast. Through these steps, the calibrated probabilistic forecast can be published for public usages, and the forecasters can make other customized forecasts accordingly.

6.3 Domain Expert Feedback

The proposed system is developed in close collaborations with the meteorologists. A senior forecaster and an experienced meteorol-

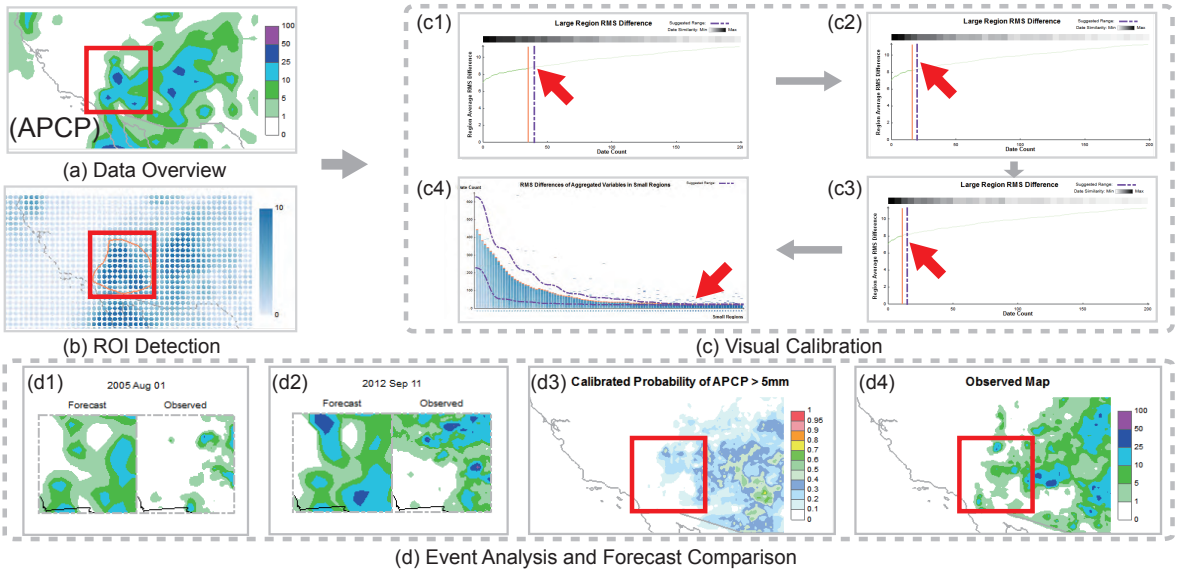

Figure 9: Case Two: This case shows the usability of our system in detecting an unusual forecast. In this case, the suggested analog number for the large region continues to decrease, as indicated by the red arrows in (b). The small suggested analog number shows that the current forecast is unusual even in the long history. Similar previous forecasts are then utilized to assist in further analysis.

ogist who has conducted extensive research on the analog method are involved in the development process. Two meteorologists who specialize in forecast verification then help evaluate our system. We have collected their feedback about the system as well as other important topics, such as the accuracy and speed improvement.

System Evaluation The weather forecaster helps ensure that the system workflow fits the domain experts' routine workflow smoothly through the development. He appreciates the analog methods integrated into the interactive system the most. In terms of visualization, the weather forecaster appreciates the pixel bar chart and the brushing interaction the most. He believes that the view is intuitive and expressive in conveying the RMS differences.

The meteorologist with extensive experiences in analog method research appreciates the interactivity of the tool in supporting the method the most. He said:" *The tool mimics the way a weather forecaster thinks about the weather prediction process. They typically compare today's weather forecasts to past forecasts, think about what actually happened, and construct a mental model for today's forecast. The analog procedure you demonstrate is sort of like an objective way for a forecaster to visualize what's in his brain.*"

The two meteorologists who specialize in verification are extremely interested in our system. Both meteorologists indicate that the tool is highly useful for long-time forecast verification, which covers years of data. The meteorologists appreciate the usability of the method in verifying precipitation level. Meanwhile, they comment that the analog method is ineffective in detecting the shape and position biases of precipitation regions.

Accuracy Improvement The four domain experts all confirm that our system is the first interactive tool they have ever known and used to assist the calibration using the analog method from a long historical perspective. The system enables them to explore the historical data and understand the forecast better, which is difficult to achieve without our system. They agree that this can enhance their justifications for the calibration and assist their routine work.

During the discussions with the domain experts, they point out three typical scenarios under which our system might work well. First, when the numerical prediction model is updated, the potential bias patterns in the forecast might change. Second, novice forecasters usually know little about the potential bias patterns existing in the ensemble forecast. Through our system, they can quickly understand the data and conduct better forecast calibrations. Third, when unusual events occur, forecasters can locate historical similar events quickly with our system and conduct the further analysis. In all these scenarios, forecasters might doubt about how to calibrate the forecast, and our system can provide informative assistances.

Speed Improvement The senior forecaster points out that what our system introduces is a new calibration process for him. Whether our system will help speed up the entire calibration depends on the complexity of the forecast and the forecasters' experiences. For example, an experienced forecaster can complete a forecast calibration very efficiently. Our system might not help speed up, but increase his confidence in the calibration. However, if the bias patterns in the forecast are hard to justify through domain experts' experiences, our system will help speed up the calibration. He also mentions that, there is no way for him to calibrate the forecast from a long historical perspective without our system. Although the analog methods have been proved to be useful, it has still not been widely used in the forecasters' routine work. Among the four domain experts, only one of them is using the method through a console window, which is a text user interface.

6.4 Discussion

Although our system can effectively support the domain experts' routine work, based on the case studies and domain expert feedback, several issues still require further discussions.

First, our framework extends the analog method through the majority voting. Since the analog methods have many variants, the voting is a good choice to deal with the analog selection problem when we use several variants simultaneously. Meanwhile, we have designed a thorough process to assist domain experts in setting thresholds for the three selected variants. This can enhance the step of the analog selection and the generation of the probabilistic forecast. Therefore, the quality of the calibration can be improved.

Second, we have conducted an evaluation of the time that our system might cost through experiments. Currently a typical analysis process for an ROI might cost about 5-8 minutes. The time can be further reduced when the experts become familiar with the system. We can further reduce the time by, for example, enhancing the brushing interaction by selecting patches on a segmented map.

Third, the ROI in our system is a special case and we define it as the region which selects analogs with high RMS differences for generating the probabilistic forecast. It is different from the ROI definitions in some previous research, such as regions with

high data uncertainty [26] and those with predictive error [6]. More specifically, our ROI definition is to detect the region where errors might exist in the generated initial probabilistic forecast.

Fourth, the final results might be different when the same forecast is analyzed by various users. However, the suggestions provided by our system can work as a constraint and guide user interactions to decrease the variances of the results.

7 Conclusion and Future Work

In this study, the calibration problem is characterized and the analog method is utilized to support it. A visual voting framework is proposed to address problems in the existing analog methods. Moreover, a visualization system based on the framework is provided to assist in calibrating weather forecasts. Coordinated views and intuitive interactions are provided to support the involvement of domain experts' professional knowledge in the statistical method. Therefore, the calibration can be better conducted. The system is developed through close collaborations with the domain experts. Case studies and feedback from the domain experts have exhibited the promising usability of our system in supporting the calibration.

In the future, we will integrate our system into the domain system that we are developing with the domain experts. We plan to include additional statistical methods in the system to solve problems that cannot be addressed well by analog methods, such as considerable position and shape biases of precipitation regions and the continuous calibration for meteorological events that last for days.

Acknowledgements

This work was supported in part by the Special Fund for Meteorology Research in the Public Interest of China (GYHY201306002), and in part by the National Natural Science Foundation of China (61272225, 91315302, 51261120376, 61422211, 61232012).

References

[1] F. Berman, A. Chien, K. Cooper, J. Dongarra, I. Foster, D. Gannon, L. Johnsson, K. Kennedy, C. Kesselman, and J. Mellor-Crumme. The grads project: Software support for high-level grid application development. *International Journal of High Performance Computing Applications*, 15(4):327–344, 2001.

[2] R. Buizza, P. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei. A comparison of the ecmwf, msc, and ncep global ensemble prediction systems. *Monthly Weather Review*, 133(5):1076–1097, 2005.

[3] H. Doraiswamy, V. Natarajan, and R. S. Nanjundiah. An exploration framework to identify and track movement of cloud systems. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2896–2905, 2013.

[4] H. R. Glahn and D. A. Lowry. The use of model output statistics (mos) in objective weather forecasting. *Journal of applied meteorology*, 11(8):1203–1211, 1972.

[5] T. Gneiting, F. Balabdaoui, and A. E. Raftery. Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(2):243–268, 2007.

[6] L. Gosink, K. Bensema, T. Pulsipher, H. Obermaier, M. Henry, H. Childs, and K. Joy. Characterizing and visualizing predictive uncertainty in numerical ensembles through bayesian model averaging. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2703–2712, Dec 2013.

[7] H. Guo, X. Yuan, J. Huang, and X. Zhu. Coupled ensemble flow line advection and analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2733–2742, 2013.

[8] T. M. Hamill, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr, Y. Zhu, and W. Lapenta. Noaa's second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society*, 94(10):1553–1565, 2013.

[9] T. M. Hamill and J. S. Whitaker. Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review*, 134(11):3209–3229, 2006.

[10] S. Hankin, D. E. Harrison, J. Osborne, J. Davison, and K. O'Brien. A strategy and a tool, ferret, for closely integrated visualization and analysis. *The Journal of Visualization and Computer Animation*, 7(3):149–157, 1996.

[11] M. Harrower and C. A. Brewer. Colorbrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003.

[12] B. Hibbard and B. Paul. vis5d. *GNU Public Licenced software. University of Wisconsin Space Science and Engineering Center*, 1998.

[13] W. L. Hibbard. Computer-generated imagery for 4-d meteorological data. *Bulletin of the American Meteorological Society*, 67(11):1362–1369, 1986.

[14] D. Hou, M. charles, Y. Luo, Z. Toth, Y. Zhu, R. Krzysztofowicz, Y. Lin, P. Xie, D.-J. Seo, M. Pena, and B. Cui. Climatology-calibrated precipitation analysis at fine scales: Statistical adjustment of stage iv toward cpc gauge-based analysis. *Journal of Hydrometeorology*, 15(6):2542–2557, 2012.

[15] H. Janicke, M. Bottinger, U. Mikolajewicz, and G. Scheuermann. Visual exploration of climate variability changes using wavelet analysis. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1375–1382, 2009.

[16] J. Kehrer, F. Ladstadter, P. Muigg, H. Doleisch, A. Steiner, and H. Hauser. Hypothesis generation in climate research with interactive visual data exploration. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1579–1586, 2008.

[17] C. Leith. Theoretical skill of monte carlo forecasts. *Monthly Weather Review*, 102(6):409–418, 1974.

[18] P. Lundblad, H. Lofving, A. Elovsson, and J. Johansson. Exploratory visualization for weather data verification. In *Information Visualisation (IV), International Conference on*, pages 306–313, 2011.

[19] A. M. MacEachren, A. Robinson, S. Hopper, S. Gardner, R. Murray, M. Gahegan, and E. Hetzler. Visualizing geospatial information uncertainty: What we know and what we need to know. *Cartography and Geographic Information Science*, 32(3):139–160, 2005.

[20] A. T. Pang, C. M. Wittenbrink, and S. K. Lodha. Approaches to uncertainty visualization. *The Visual Computer*, 13(8):370–390, 1997.

[21] T. Pfaffelmoser, M. Mihai, and R. Westermann. Visualizing the variability of gradients in uncertain 2d scalar fields. *Visualization and Computer Graphics, IEEE Transactions on*, 19(11):1948–1961, 2013.

[22] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. R. Schwalm, D. N. Huntzinger, R. Cook, E. Bertini, and C. T. Silva. Visual reconciliation of alternative similarity spaces in climate modeling. *Visualization and Computer Graphics, IEEE Transactions on*, 20(12):1923–1932, 2014.

[23] K. Pöthkow and H.-C. Hege. Positional uncertainty of isocontours: Condition analysis and probabilistic measures. *Visualization and Computer Graphics, IEEE Transactions on*, 17(10):1393–1406, 2011.

[24] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. R. Johnson. Ensemble-vis: A framework for the statistical visualization of ensemble data. In *Data Mining Workshops, IEEE International Conference on*, pages 233–240, 2009.

[25] A. E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski. Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, 133(5):1155–1174, 2005.

[26] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. J. Moorhead. Noodles: A tool for visualization of numerical weather model ensemble uncertainty. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1421–1430, 2010.

[27] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, IEEE Symposium on*, pages 336–343, 1996.

[28] C. Tominski, J. F. Donges, and T. Nocke. Information visualization in climate research. In *Information Visualisation (IV), International Conference on*, pages 298–305, 2011.

[29] L. A. Treinish. Severe rainfall events in northwestern peru (visualization of scattered meteorological data). In *Visualization, IEEE Conference on*, pages 350–354, 1994.

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