Empirical Analysis of the Subjective Impressions and Objective Measures of Domain Scientists' Visual Analytic Judgments

Aritra Dasgupta, Susannah Burrows, Kyungsik Han, Philip J. Rasch
Pacific Northwest National Laboratory
Richland, USA
{aritra.dasgupta, susannah.burrows, kyungsik.han, philip.rasch}@pnnl.gov

ABSTRACT
Scientists often use specific data analysis and presentation methods familiar within their domain. But does high familiarity drive better analytical judgment? This question is especially relevant when familiar methods themselves can have shortcomings: many visualizations used conventionally for scientific data analysis and presentation do not follow established best practices. This necessitates new methods that might be unfamiliar yet prove to be more effective. But there is little empirical understanding of the relationships between scientists' subjective impressions about familiar and unfamiliar visualizations and objective measures of their visual analytic judgments. To address this gap and to study these factors, we focus on visualizations used for comparison of climate model performance. We report on a comprehensive survey-based user study with 47 climate scientists and present an analysis of: i) relationships among scientists' familiarity, their perceived levels of comfort, confidence, accuracy, and objective measures of accuracy, and ii) relationships among domain experience, visualization familiarity, and post-study preference.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI)

Author Keywords
Information visualization; visual comparison; climate; familiarity; trust; preference; Taylor plot; slope plot

INTRODUCTION
Domain experts' analytical workflow often comprises a set of conventional methods for data analysis and presentation. Since experts are highly familiar with these methods, they tend to inherently have greater confidence in their outputs than in those of new, unfamiliar methods. At the same time, conventional data visualization methods used by practitioners might not be in sync with visualization best practices [9]. However, there is little empirical analysis of the relative effectiveness of familiar and unfamiliar visualization techniques, compared with domain scientists' subjective impressions (e.g., confidence, preference, etc.) or with objective performance measures.

To fill this gap, we contribute a user study with climate scientists, focusing on the common task of comparing and assessing similarities and differences in model fidelity across multiple simulations. We report on three related contributions in this paper. *First*, through a problem characterization phase, we developed a shared understanding of the important visualization tasks for model fidelity comparison, and identified visualization techniques that are currently widely used for these tasks. By applying visualization design principles, we collaboratively developed and selected two sets of visualizations: familiar visualizations modified for more effective visual comparison, and unfamiliar visualizations that had the potential to be more effective than the familiar ones. *Second*, using these two sets we conducted a user study with 47 climate scientists, where we recorded their objective task responses, their subjective impressions [4], such as their perceived levels of comfort, confidence, accuracy when carrying out analysis tasks using a particular visualization, and their preferred visualization as indicated at the close of the study. *Third*, we analyzed study results for understanding the relationships among scientists' domain experience, familiarity, and objective measures of their analytical judgments, the discrepancies between their familiarity and preference levels for each visualization, and also the discrepancies between their subjective impressions and objective performance measures.

PROBLEM AND TASK CHARACTERIZATION
The work reported here results from a *six*-month long collaboration between two climate scientists (co-authors of this paper) and visualization researchers. Following the nested model approach [21], there were four distinct stages in our collaboration: i) characterization of the model fidelity analysis problem, ii) a shared understanding of the visualization tasks, iii) analysis of the state-of-the-art in visualizations used by climate scientists, and iv) participatory design of prototypes. We had frequent face-to-face discussions for facilitating all of these stages, and the outcome of this collaborative design process helped us select appropriate tasks and visualizations, calibrated by experts' degree of familiarity, for the user study. In this section we describe the model fidelity analysis problem and characterize the relevant visualization tasks.

Model Fidelity Analysis: Climate model fidelity is measured by the degree of consistency between models and observations for specific model output variables or features. Scientists

CHI 2017, May 06 - 11, 2017, Denver, CO, USA

ACM 978-1-4503-4655-9/17/05 $15.00
DOI http://dx.doi.org/10.1145/3025453.3025882

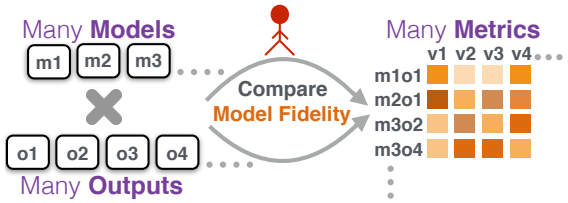

Figure 1. Illustrating the multi-model, multi-output comparison problem where scientists have to visually reconcile similarities and differences with respect to many models, outputs, and metrics. The visualization challenge is twofold: i) scale up to the different *levels* of the categories, e.g., models $m1$, $m2$, $m3$, etc., and outputs $o1$, $o2$, $o3$, $o4$, etc. and ii) at the same time provide visual cues for similarity and dissimilarity based on the values for many metrics ($v1$, $v2$, $v3$, $v4$, etc.).

frequently summarize aspects of the model's *fidelity*, using statistical metrics such as (but not limited to) the root-mean-square error, correlation, and relative variance of the model output variables compared to observations of the same variable [12]. Because a credible climate simulation requires reasonable simulations of many different physical processes and state variables, it is typically not sufficient to evaluate models on a single metric, instead, models must be compared across a suite of such metrics, leading to a more complex analysis situation. In addition, identifying models that exhibit similar or dissimilar patterns of performance across a suite of metrics may lead to important insights into model behaviors. Frequently, trade-offs occur between different aspects of model fidelity, with models that perform better in one area performing worse in another, so scientists must perform nuanced analysis to understand subtle similarities and differences in fidelity with respect to multiple metrics.

Multi-Way Visual Comparison Tasks: Model fidelity analysis is typically supported by visual comparison of multiple metrics across simulations derived from multiple models. We term these tasks *multi-way visual comparison* tasks (Figure 1), since they involve comparison across different dimensions: among models and output variables, their combinations, and across different metrics. The comparison tasks are simpler when evaluating a single model using a single type of statistical metric, e.g. comparing model-observation correlation for temperature and precipitation fields; or when comparing the performance of multiple models for a single variable and type of metric, e.g. identifying the model that produces the lowest root-mean-square-error for cloud extent. However, in reality, when evaluating overall model fidelity, scientists often need to compare across many models (10–15 or more), many output variables (10–15), and more than one metric (e.g. root-mean-square error, global mean bias, correlation, and variance).

As illustrated in Figure 1, a major visualization design challenge in multi-way visual comparison tasks stems from the need to simultaneously express similarity and dissimilarity based patterns across multiple categorical variables (e.g., types of models, output variables, etc.) and numerical variables (e.g., fidelity metrics). Visualizations of model fidelity metrics should enable an expert user to efficiently and accurately perform tasks such as: i) assess and compare the overall fidelity of models across a range of metrics, and ii) identify groups of models that exhibit similar or different patterns of behavior across multiple metrics. In this work, our main motivation was to understand how scientists' analytical judgment for comparison of fidelity across different combinations of models, model outputs and metrics is impacted by different visualization techniques.

RELATED WORK

Our contributions combine two areas of visualization research: i) visual comparison approaches, and ii) evaluation of the user experience of domain experts using visualizations.

Alternative Approaches for Visual Comparison

The design space of visual comparison using data sets that contain a mix of numerical and categorical attributes can be described based on three different approaches. In the first approach, multiple visual variables [2], like shape, color, or symbols, can be used for representing different categorical variables. Such visualizations are popular in the climate science domain [24]. But they do not scale well to the number of levels in a categorical variable, as the number of discriminatory steps using color or shape is limited. The second approach is to use small multiples [29], where depending on the task, a scatter plot or bar chart can be instantiated for each categorical variable. The number of small multiples can be a problem when there are many categories or many levels in a category, and might not scale well for visual comparison. In a third approach, several researchers have proposed alternative representations of categorical data [10]. Researchers have also looked at different interestingness measures for encoding multivariate relationships in mixed data sets [1] and using analytical methods like multiple correspondence analysis [5]. In this work, we focus on model fidelity visualization techniques used in the climate science community. Two popular choices are the Taylor plots [28] and heat maps [24]. We analyzed their merits and demerits, proposed alternatives by modifying these familiar visualizations to suit the visual comparison tasks, and introduced unfamiliar yet potentially optimal visualizations that could lead to better analytical judgment.

Evaluating Domain Experts' Visualization Experience

Previous research has demonstrated that visualization researchers and domain experts often disagree about the efficacy of visualization designs [3] . This can be for two reasons: domain experts tend to trust and prefer familiar analysis methods over new ones [27], and there is a general lack of awareness of visualization best practices and available alternatives [9]. In human-machine relationships, as in human-human relationships, familiarity breeds trust [11], and analysts tend to have more trust in familiar means and mediums. However, some conventional methods that are widely used for model performance visualization may scale poorly to more complex data. Novel approaches may offer better support for complex comparisons of model fidelity across multiple models and variables. To investigate how much domain scientists trust and prefer different visualization techniques, we followed the strategy in McAllister's survey [20], which was subsequently adapted for evaluating domain experts' trust in cyber security interfaces [27] and in complex dynamic systems [30]. We adopt a relevant subset of these questions for our evaluation.

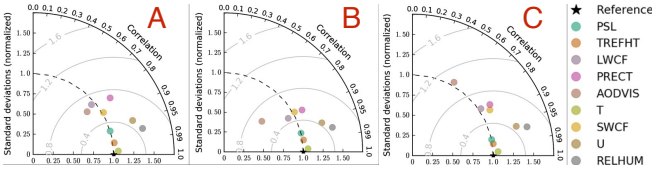

Figure 2. Small multiples of Taylor plots where each plot represents a model and each data point is an output variable. Taylor plots are popularly used for model fidelity comparison. Since the default layout with many symbols is less effective for comparison across many (≥ 3) models and many (≥ 3) variables, we designed this small multiple version, which our collaborators rated positively and was eventually used in the study.

Domain experts' trust in and preference for visualization systems is an emerging area of research [26]. In our work, we focus exclusively on the efficacy of alternative visualization designs and how they can inspire comfort, confidence, and preference in climate scientists.

VISUALIZATION DESIGN ALTERNATIVES

The multi-way visual comparison tasks abstracted during the problem characterization phase were utilized for analyzing the merits and demerits of different model fidelity visualization techniques. This was done through a review of the literature and discussions with two climate scientists (co-authors on this paper), one of whom has more than 30 years of experience in climate modeling. In the course of our discussions we narrowed down on a subset of such tasks that climate scientists most frequently perform as part of their analysis routine. By applying visualization design principles in the context of model comparison [9], we found that familiar visualizations such as the Taylor plots, heat maps, and bar charts, that are used for such tasks, might not be the optimal choice. These led us to propose design alternatives by modifying familiar visualizations and introducing new ones. In this section, we describe their design rationale.

Modification of Familiar Visualizations

We collaboratively critiqued the familiar visualizations by analyzing if they effectively communicate similarity in model fidelity and if that information can be recovered by scientists accurately and efficiently. Based on our analysis, we proposed the use of improved versions of the familiar visualizations.

Small Multiples of Taylor Plots: A Taylor plot [28] graphically summarizes how closely a pattern (or a set of patterns) matches observations. The similarity between two patterns is quantified in terms of their correlation, their centered, normalized root-mean-square difference and their variability (represented by their standard deviations). These plots are especially useful in evaluating multiple (< 3) output variables of complex models or in gauging the relative skill of many different models for a small number of output variables (< 3). As shown in the Taylor plot in Figure 2, each circle represents a model output. The light gray contours show the centered normalized RMSE, which is proportional to the distance from the reference point (1, 1). The angular axis indicates the correlation with respect to the observational data. The radial axis indicates the variance ratio, i.e. the ratio of the model's standard deviation to the standard deviation of the observations, which was not used in this study.

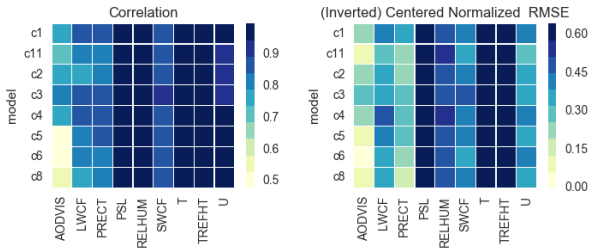

Figure 3. Heatmaps representing model performance based on two different metrics provide a quick overview of overall differences and variability. To help in distinguishing among small differences, we use a segmented color scheme. Using of color to express quantitative differences can lead to less accurate judgment.

When a single Taylor plot [28] is used to summarize many (> 3) variables across many (> 3) models, information can be obscured by overlap and clutter. In such plots, multiple visual variables are used to represent the different levels within categorical variables, which can reduce the accuracy and efficiency of visual analysis. We addressed this problem by using small multiples of Taylor plots [29]. In the small multiples used in this study, each Taylor plot (Figure 2) represents a different model, while color is used to encode different output variables. This encompasses both superposition (for variables) and juxtaposition (for models) categories for visual comparison [13]. Similar to identifying shapes in scatter plots, one can compare the shapes of point clouds in these Taylor plots for a multi-way comparison among models and variables. This small multiple approach can suffer from two limitations: visual scalability when the number of models is greater than four or five, and difficulty in comparing across different output variables using their relative positions on the small multiple plots. Other strategies for using Taylor plots are possible, but each strategy we considered suffers from similar difficulties in comparing many models. We feel Taylor plots are most useful when comparing a few (< 3) models for multiple output variables, or multiple models for a few (< 3) output variables.

Variants of Heat Maps and Bar Charts: A heat map is one of the most widely-used visualizations for comparing model outputs or model performance [6]. We designed a heat map with a continuous color scheme [14] where yellow indicated bad performance while blue indicated good performance. We used two heat maps for representing two different performance metrics. In the heat maps, all models and variables are represented within a *large single* plot [31], in contrast to the *small multiples* used for the Taylor plots. By using color to represent variables, heat maps can support an efficient comparison of model fidelity and aid identification of similarities and dissimilarities by visual pattern recognition, but the use of color might limit the accuracy of quantitative comparisons [33].While similarity metrics could be used to sort variables in heat maps in a more meaningful fashion, we chose to use alphabetical sorting because of its common use in the climate science literature, and to ease comparison of the same variables across multiple heat maps, where different orderings for each heat map can lead to high comparison complexity.

We found bar charts to be a very familiar visual representation for climate scientists, which are often used to examine relative distributions among variables, but not necessarily for multi-

way comparison tasks. We also found variants of bar chart representations in the climate science literature, including stacked bar charts, with different stacks representing different categories [24]. By using spatial position along a common axis to encode information, bar charts can support accurate estimation and comparison of small differences in quantitative values [33]. To support visual comparison among multiple models, we used a small multiple approach where each small multiple represents a model, and in order to encode multiple metrics, two sets of bar charts are provided, where each set represents a metric. While bar charts enable quantitatively accurate evaluations, comparison across multiple bar charts for many models and variables may be difficult and inefficient.

Design of Unfamiliar Visualizations: Slope Plots

We introduced the slope plots (Figure 4), a variant of the slopegraph proposed by Tufte [29], to our collaborators as an alternative, yet potentially unfamiliar, solution for multi-way visual comparison. Our aim was to overcome the following shortcomings of the existing visualizations. i) The original Taylor plots do not scale for many models and variables ii) The small multiples of Taylor plots and bar charts may not support optimal comparison amongst multiple (10) models and multiple (10) variables. iii) The heat map, with a large single approach, supports efficient comparison, but because it uses color to represent quantities, may reduce accuracy. With the slope plot, we also followed the large single approach, but with a position-based encoding and explicit cues of similarity/dissimilarity, we aimed at making the comparison tasks. more accurate and efficient. We describe the design rationale and implementation in detail below.

Slope plots encode one numerical variable (representing a metric) in a single plot, while multiple such plots can be juxtaposed for multi-way visual comparison. The main rationale behind the slope plot was to allow comparison by both superposition and juxtaposition [13], so that scientists can assess variability among both output variables and models quickly and accurately.

The layout of the slope plot is inspired by the parallel coordinates visualization [17]. Slope plots have an intentional horizontal layout to avoid confusion with line plots with time on the horizontal axis. In the slope plot (Figure 4), each horizontal axis represents the same numerical variable, correlation on the left plot and centered RMSE on the right. The axis can be discretized based on the number of levels in a category, which in this case are the different models, represented by color. The number of axes is equal to the number of levels in another categorical variable, in this case, the model outputs.

Polylines connecting the axes represent different models and their slopes are not indicative of functional relationships among the outputs. Relative slopes indicate the similarities and differences among the models. By linking different points, we added Gestalt effects of connectivity [18] and similarity, where parallelism or non-parallelism of lines (representing models) will give visual cues of overall similarity [7] or dissimilarity with respect to both models and variables. The spread of the points on the horizontal axis indicates variability across models. The advantage of slope plots over the other

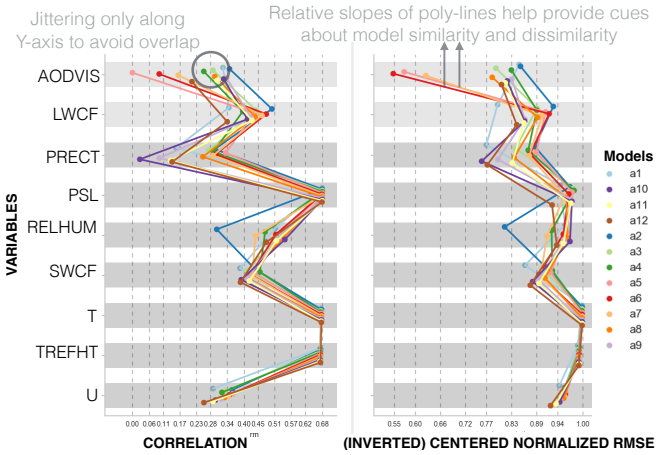

Figure 4. The slope plot was designed as an alternative to the existing model fidelity visualization techniques. Each polyline represents a model and their relative slopes indicate similarities and dissimilarities with respect to different output variables, which are represented on the horizontal axes. This design offers an efficient way to look at variability across both models (polylines) and variables (rows). Jitter is added along the Y-axis to avoid over-plotting.

visualizations is that it scales well to the association among multiple levels of two categorical variables (i.e., models and outputs) while at the same time preserving the perceptual cues about similarity and dissimilarity.

One of the problems with spatial encoding is the over-plotting of multiple data points. To minimize such over-plotting, we introduced jitter along the vertical axis. This does not distort the values as the vertical axis is categorical, and due to the jitter lines do not overlap and can be precisely identified.

STUDY DESIGN

We conducted a within-subjects study to evaluate climate scientists' subjective and objective measures of analytical judgment using the familiar and potentially unfamiliar visualization techniques. The study was divided into three sections.

In the first section, we asked participants questions regarding their demographic characteristics and years of domain experience, and provided training on the tasks and visualizations that would be used for the study. For this study, it was important to isolate any domain-specific bias from the assessment of the visualizations. Climate scientists, when comparing models across multiple metrics, may at times assign different weights to different output variables, depending on the importance of those variables to the analytical task. Our instructions made it clear that for the purposes of the study, participants should treat all variables and metrics as equally important. We also emphasized that the goal of this study was not to *identify the best model*, but to assess the *effectiveness of the analysis methods*, namely, the different visualization techniques.

In the second section, participants were asked to perform three tasks for each visualization they were shown. These tasks were designed to be relevant for scientists in developing an understanding of the information contained in multiple fidelity metrics across multiple model simulations.

In the third section, we selected a set of questions that assessed scientists' subjective experience of the visualizations with respect to different dimensions of trust (e.g., perceived accuracy, efficiency, comfort, confidence) and to their preference for a particular visualization technique. We describe the different aspects of our study design in detail below.

Data and Visualization Generation

We discuss the data and visualization generation based on the performance metric outputs.

Models and Outputs: We used simulation output from a number of variants of version 5 of the Community Atmosphere Model (CAM5, [23]), the atmosphere component of the Community Earth System Modeling (CESM) project [16]; and descendants of that model developed for a variety of projects supported with funding from the U.S. Department of Energy (DOE) Earth System Model Program, which we call generically "CAM5" variants hereafter. We selected output variables that participate in many important ways in the climate system. These variables play important roles in the Earth's radiation budget and cloud features (LWCF, SWCF, RELHUM, AODVIS), the hydrologic cycle (PRECT), and general circulation features (T, TREFHT, and U).

Fidelity Metrics: The model performance metrics selected here are amongst the metrics that are commonly and widely used in the climate modeling community [22]. These metrics, the correlation and the normalized standard variance, are straightforward to calculate and interpret. The two metrics relate to two different aspects of model performance: the correlation is a measure of *pattern matching*, while the normalized standard variance is a measure of how well the magnitude of the variability of the model agrees with the variability of observations. The model-observation agreement is often evaluated in a climatological sense, i.e., an average January from the model simulation is compared with an average observed January, or an average is computed for a three-month period identified as a season. In keeping with that practice, the visualizations presented in this study represent fidelity metrics computed for each of those seasons. We used two data sets from two seasons as a repeated measure.

Normalization and Variability: A higher correlation score indicates *good* model performance, as the output is highly correlated with the observation. A higher nRMSE score on the other hand signifies *poor* model performance. Therefore for semantic equivalence between a plot for correlation and a plot for nRMSE, where high values represent *good* performance, we invert the nRMSE scores by subtracting each value from the maximum value in the data set. Within each pair of plots in the study, one indicates the correlation and the other indicates the inverted nRMSE score. In order to ensure a reasonable amount of variability among the fidelity scores for both the metrics, we selected data set pairs where the average standard deviation across the models for a pair was the highest.

Visualizations: We selected four visualizations for the study, based on our understanding of scientists' familiarity. These were small multiples of Taylor plots, heatmaps, small multiples of bar charts, and slope plots, as discussed earlier. We expected the slope plot to be the least familiar visualization.

For ordering the rows and columns in heat maps and slope plots, and for organizing the layout of the small multiples, we selected a default alphabetical order and did not perform any layout optimization. This was for two reasons: First, for the small multiple approach, each multiple will have a different ordering and it could be hard to trace model names across many of them. Second, since the tasks were about similarity and dissimilarity (all performed together), any optimization would be based on a pairwise distance metric. Sorting or laying out small multiples based on a pairwise distance metric is non-trivial and may lead to visual complexity. Therefore, for the sake of consistency, we did not optimize the layout for any of the plots. Further, we controlled for the following factors: i) *Ordering*: The visualizations were randomly ordered across participants. To minimize the recall effect, for the same kind of visualization across two data sets, the left and right orders were swapped. For example, we had two instantiations of juxtaposed heatmaps: one where the right metric was correlation and the left one was nRMSE, whereas in the second data set, it was the reverse. ii) *Scales*: The ranges on the axes of the plots were scaled based on the minimum and maximum value of the data; small differences could be highlighted by not scaling between a global minimum and global maximum. ii) *Resolution*: The degree of numerical resolution possible in each visualization was dependent on the visualization type. Heat maps use the lowest resolution defined by the number of color levels, but all other visualizations use spatial encoding, with their effective resolution depending on spatial locations. Equivalence in resolution between visualizations was supported by holding the interval spacing of axis ticks constant across all spatial visualizations. Pilot studies also helped us fine-tune the resolution and reduce clutter.

Questionnaire and Task Selection

In this section, we discuss the questions posed to participants and the tasks they had to perform during the study for each of the three sections. Participants were required to respond to all items in the study for their response to be considered complete, and they were not allowed to skip questions.

Section 1. Demographics, Experience, and Familiarity: In this section, we asked scientists to report their age, gender, years of domain experience, and their primary role as a climate scientist (i.e., global model user, global model developer, regional model user, or regional model developer). These questions were followed by the introduction of all visualizations by showing an example of each visualization, using real data, and describing each visualization in the context of the tasks. Before starting the task performance section, we first asked participants to rate their *familiarity* with each of the four visualizations, and their *frequency of use* of each visualization as part of their analysis routine. By normalizing and averaging their responses on these two questions, we derived a familiarity score for each participant for a particular visualization technique.

Section 2. Analytical Tasks: For the study, we focused on the comparison-based tasks that we derived through our initial

discussions about climate scientists' analytical workflow. We conducted pilot tests to evaluate suitable comparative analysis tasks that could be performed using all four visualization types and the associated difficulty levels of the tasks, and accordingly distilled these into the following three tasks that were posed for each visualization:

Task A: "Identify two models, which are most similar in their correlation and normalized RMSE metrics across most of the output variables."

Task B: "Which two output variables show the highest variability in their correlation and normalized RMSE across all models?"

Task C: "Identify the two models that disagree the most in their correlation and normalized RMSE metrics across most of the output variables."

For each of these tasks, participants were asked to rate their confidence and comfort level with each visualization, on a 5 point Likert scale. The **difficulty level** of Tasks A and C was greater than that of Task B as in the case of the latter, only a few output variables showed significantly greater variability than others. This was easier to detect visually as compared with the subtle similarities and differences among multiple models.

Section 3. Subjective Impressions Rating: At the end of the task performance section, participants were asked to record their subjective impressions of the visualizations they used. We recorded two types of responses: **i) Perceived accuracy, efficiency, comfort, confidence, and preference**: To analyze participants' perceptions about using a particular visualization, they were asked to record, on a Likert scale, their perceived accuracy, efficiency, comfort, confidence, and preference for each visualization. These questions were a subset of the trust-related questions used in McAllister's survey [20], and later adopted by Takayama et al [27] for evaluating analysts' trust in user interfaces for system administrators. Note the redundancy in asking about perceived confidence and comfort levels, for each task, and also amongst post-completion questions. This intentional redundancy allowed us to assess the consistency between participant's subjective perceptions after completion of each task, compared with their post-study, overall evaluations of each visualization. **ii) Preference Ranking**: This was a ranking on a scale of 1 (highest) to 4 (lowest), where participants assigned a rank to their preferred visualization. Preference is a strong indicator of potential adoption of a visualization technique. Following the recommendations of Lam et al. [19] for assessing user experience with visualizations, we also asked participants to comment on the advantages and barriers they perceived in using each visualization. By analyzing these comments we could potentially find the reason behind scientists' strong or weak preference for a visualization.

Participants

Participants were recruited anonymously through different mailing lists that involve collaborations among universities and research laboratories. No personal identifiers about participants were recorded. We only recorded their IP addresses, to ensure that duplicate responses were not submitted from the same machine.Overall, there were 47 participants with a self-reported background in climate modeling. The study participants comprised a very experienced group of climate scientists with a median domain experience of 12 years, and a majority of participants rating themselves as *very familiar* or *extremely familiar* with modeling. Each participant performed three tasks (Task A,B,C) for each of the four visualization types, where two data sets were used as a repeated measure. The total number of trials was thus $47 \times 4 \times 2 = 376$.

Settings

The experiments reported in this study were all web-based. This remotely-based study setting was necessary because the participants in the study were climate scientists spread across different academic institutions and research labs across the United States, Europe, and South America. In our experimental set-up we took several measures to ensure reliability and minimize bias in the results that could arise from the web-based setting. First, to minimize the risk that a participant might not understand the question or be ready for the test, we trained participants about the visualizations and allowed them to them quit the study if they did not understand a question. Participants could not return to previous tasks to compare their responses across multiple visualizations.If a participant stopped the study and returned to it later, they reentered the study from the point where they had left off, preventing unintentional repetition of the tasks by a participant.

The trials were randomized across participants, and they sequentially performed all three analytic tasks using one particular visualization before being presented with the next visualization, reflecting the general analytical workflow of scientists. We asked participants if they had any known color vision deficiency (color-blindness), as the ability to distinguish among colors, was needed to interpret most of the visualizations used in the study. We filtered out results from one participant who self-identified as color-blind.

Hypotheses

Objective Accuracy of Task Performance. Given our design rationale for slope plots, we expected them to support greater objective accuracy in task performance. But we also expected task complexity, the length of domain experience and degree of familiarity with particular visualizations to affect performance accuracy. These expectations led to the following hypotheses: **H1a)** Overall, participants will report a greater level of accuracy with slope plots. **H1b)** Tasks A and C will result in greater differences in performance than Task B due to the greater complexity of these tasks **H1c)** Longer domain experience and greater familiarity with particular visualizations will lead to more accurate performance.

Perceived Confidence, Comfort, Accuracy and Efficiency: We know that familarity with an analytical tool increases trust and confidence in that tool among users [11]. This led to the following hypothesis:**H2)** Participants with longer domain experience will report higher average confidence and comfort levels for familiar visualizations than for relatively unfamiliar visualizations, such as the slope plots. Given the position based encoding of slope plots, we hypothesized that: **H3a)**

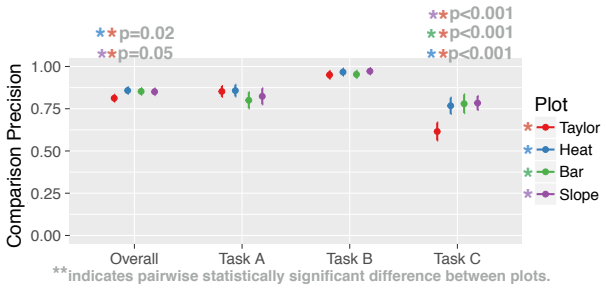

Figure 5. Objective accuracy based on the comparison precision metric. Error bars represent 95% confidence intervals. Overall, the heat map and slope plot showed the highest level of accuracy, with a significant difference with respect to the Taylor plot (p <0.05). While comparison precision values for Tasks A and B were similar, for Task C, we observe that the heat map, slope plot, and bar chart were significantly more accurate than the Taylor plot (p <0.05).

Participants will report greater perceived accuracy with slope plots as compared to heat maps. **H3b)** Participants will appreciate the utility of slope plots, and that will be reflected in their comparable or higher average ranking in terms of perceived accuracy and efficiency than more familiar plots such as Taylor plots, heat maps, and bar charts.

Familiarity Vs Preference: Recent visualization studies with domain experts have also demonstrated that carefully and collaboratively designed visualizations and interfaces can convince experts of their utility and may inspire trust and preference [3, 9]. This led to: **H4a)** Overall, participants will exhibit a level of preference for slope plots that is comparable with that of familiar visualizations. We expected the effect of familiarity on preference to be weaker in participants with fewer years of domain experience as they might be more open to novel approaches. This led to: **H4b**) Participants with greater domain experience will exhibit a stronger preference for familiar visualizations, while participants with less experience will be most likely to report a post-study preference for less familiar visualizations that are of similar or greater perceived effectiveness for completing the tasks.

Objective Accuracy Vs Perceived Accuracy, Preference, Familiarity: Regarding the discrepancy between subjective impressions and precision of scientists' judgment, we have seen in past studies that self-calibrated levels of trust correlate with greater accuracy in task performance [30]. This led to the following hypotheses: **H5a)** Participants' perceived accuracy will match their objective performance accuracy. **H5b)** Participants' familiarity and preference ranking for different visualizations will match with the ranking of the visualizations derived from their objective accuracy.

Metrics

Comparison Precision: In Tasks A and C, participants were asked to identify the two most similar and dissimilar models, respectively. To compute an objective metric characterizing the correctness of each participants' response, we used the the following method: First, we computed the root mean squared difference (RMSD) between each pair of models for a given metric, and then calculated the average between the two RMSD values for each of the two metrics. Next we ranked the model pairs based on the average RMSD values. The correct

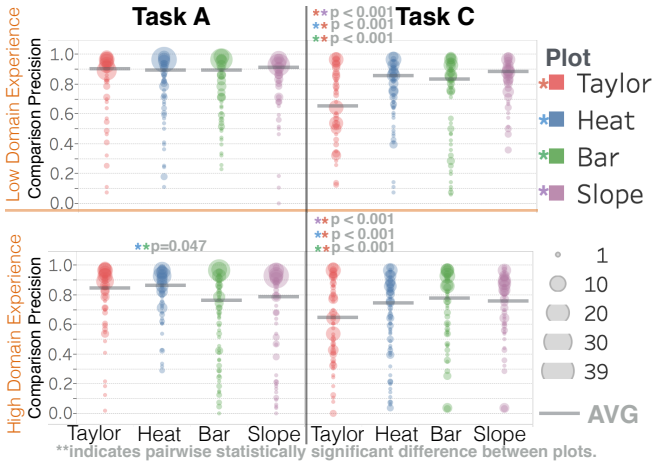

Figure 6. Distributions of comparison precision scores for Tasks A and C for the high and low experience groups. Task A resulted in comparable levels of accuracy across both experience groups except for the significant difference in accuracy between heat map and bar chart for the high experience group. For Task C, we can observe consistently lower accuracy levels for Taylor plots across both experience groups.

response to Task A and to Task C is the pair of models that exhibits the highest or lowest RMSD , respectively.

In Task B, participants were asked to identify the two models with the greatest variability in their fidelity scores. To identify these two models, we first calculated the standard deviation of the variables with respect to the individual metrics and then averaged them for a net variability score. Next we ranked the models by their net variability.

Because the fidelity scores used in this study differ only subtly between models, the differences among the RMSD and variability scores computed for different models were also modest. Hence, instead of categorizing each participant's response as true or false, we developed a combined "comparison precision" metric derived from the relative ranks of the models for the similarity score (Task A), dissimilarity score (Task C) and average variability score (Task B). *Comparison precision* is given by $C_p = \frac{(n-r)}{n}$ where n is the total number of models and r is the rank of the model based on any given individual score,and where the first two models (i.e., the correct answers) are assigned a rank 0, so that the correct response has a precision of 1. We therefore have a C_p score for each of the tasks and for an overall comparison we compute an average of the three scores across all the tasks.

Discrepancy: We computed the discrepancy between how scientists actually performed and their perceptions about their performance and the visualizations using the difference between the rankings of visualizations derived from *comparison precision* scores and their subjective impression ratings, by normalizing them to a common scale. For each participant, by using their average comparison precision score across all the tasks, we rank ordered the visualizations on a scale of 1 (best) to 4 (worst), and compared this ranking with their preference and familiarity based rankings. Next we derived an average discrepancy between the rankings (ΔAccPreference, ΔAccFamiliarity) by computing the difference between them and averaging that across all participants. For analyzing the discrepancy between objective and perceived accuracy, we

rescaled the responses for perceived accuracy to span the range from 0 to 1, by converting $(0,1,2,3,4,5)$ to (0, 0.25, 0.5, 0.75 and 1). We rescaled the objective accuracy ranks, ranging from 1 to 4, using a similar transformation, that is, converting $(1,2,3,4)$ to $(1.0,0.66,0.33,0)$, giving higher weight to higher ranked models. We then obtained the discrepancy (ΔObjSubjAccuracy) by subtracting the rescaled subjective impression scores from those of objective accuracy. A greater difference indicates that the rankings of scientists' performance and their self-rated impression of their own performance did not match: they either overestimated or underestimated their performance accuracy.

Participant grouping criteria: We grouped participants into *high* and *low* groups based on self-reported domain experience. Participants were asked to rate their degree of familiarity (on a 5-point Likert scale) and the number of years of experience in climate modeling. These two measured variables had different scales, so we first normalized them by dividing by the maximum value for each category and calculated the average of the two normalized values. We then assigned 25 participants to the high group (above the average) and 22 to the low group (below the average).

RESULTS

We report the study results by comparing participants' objective accuracy with subjective impressions and analyzing the effects of experience and familiarity. Uncertainty in all results is reported as 95% confidence intervals, estimated by bootstrapping, reflecting the range of uncertainty in the mean value. The fixed effects (i.e., independent variables) were the two climate modeling experience groups (i.e., high and low experience), three tasks (i.e., Task A, B, and C), and four visualization types (e.g., Taylor plot, heat map, bar chart, and slope plot). Confidence and comfort level were assessed using a repeated measures design :participants were exposed to multiple treatment conditions (i.e., visualization tasks) and answered the same set of questions for each condition. To analyze these variables, we fit a mixed effects analysis of variance (ANOVA) model [32], with a normal conditional distribution, and random effects for repeated measures to account for the non-independent nature of the data. We adjusted the p-values for analyses involving multiple comparisons using the Bonferroni correction [15].

Objective Accuracy: By applying the average comparison precision metric, we found that overall, participants were more precise in their judgments using heat maps (Mean: 0.864, Confidence Interval: $[0.841,0.886]$) and slope plots (0.860, $[0.837,0.882]$) with significant differences from Taylor plots (0.806, [0.784, 0.828]) at $p<0.05$ (Figure 5), thereby partially supporting our hypothesis **H1a**. By looking at the taskwise breakdown, we can see that: i) for Task A, performances using heat maps (0.857, $[0.818,0.896]$) and Taylor plots (0.852, $[0.813,0.891]$) were slightly more accurate than for bar charts (0.800 $[0.761,0.839]$) and slope plots (0.823, $[0.784,0.862]$), ii) for Task B, all plots exhibited comparable levels of accuracy, and participants were more accurate than for Tasks A and C, thus supporting **H1b**; and iii) for Task C, heat maps (0.768, $[0.729,0.806]$), slope plots (0.784, $[0.745,0.823]$) and

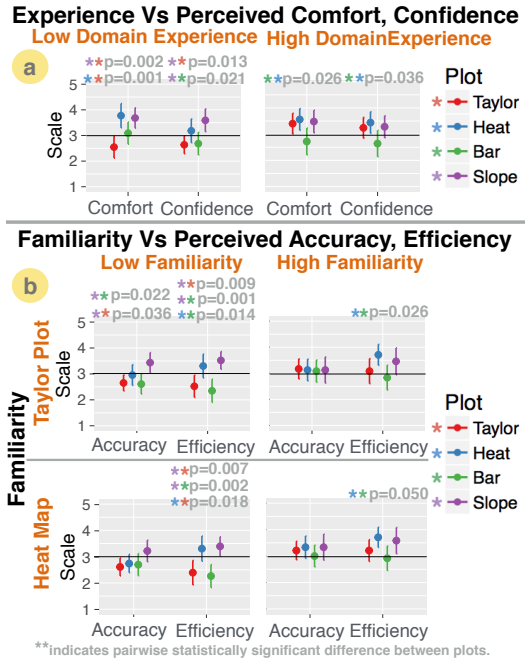

Figure 7. Experience and Familiarity Vs Subjective Impressions. Error bars represent 95% confidence intervals. Significant results $(p < 0.05)$ **annotated on the plots show that low experienced participants were less comfortable and confident with Taylor plots as opposed to slope plots. Across different familiarity groups, slope plots show comparable, if not higher perceived accuracy and efficiency ratings, as opposed to more familiar plots.**

bar charts (0.780, $[0.741,0.818]$) led to significantly greater accuracy than Taylor plots (0.615, $[0.576,0.654]$). Across all tasks, we failed to find an overall effect of familiarity on performance accuracy. By analyzing the effect of domain experience, we found different trends for Task A and Task C. For Task A we found that that performances were comparable across all visualizations in both experience groups except for participants with greater domain experience were more accurate with heat maps (Mean = 0.86) than the bar charts (Mean = 0.76). But f For Task C, we found that across both experience groups, participants performed significantly better using slope plots, heat maps and bar charts, than the Taylor plots. Therefore we found limited support for **H1c**.

Perceived Comfort, Confidence, Accuracy and Efficiency: Figure 7a shows confidence and comfort levels for all four visualizations. The low experience group rated the slope plot highest in comfort (Mean: 3.61, Confidence Interval: $[3.43,3.80]$) and confidence (3.56, $[3.37,3.74]$), with significant differences from the Taylor plot for comfort (2.91, $[2.72,3.10]$; $p<0.05$) and confidence (2.97, $[2.79,3.16]$; $p<0.05$). For the high domain experience group, the perceived levels of comfort and confidence were comparable across the Taylor plot, heat map, and slope plot, with significant differences only with respect to the bar chart. We therefore did not find evidence to support **H2**. Participants tended to rate the slope plot and heat map significantly higher ($p<0.05$) in perceived accuracy and efficiency, compared to the Taylor plot and bar chart. We however could not find support for **H3a**, as perceived accuracy for heat maps and slope plots was comparable. We created low and high familiarity groups for both the Taylor plot and heat map, using the familiarity rankings assigned by the partici-

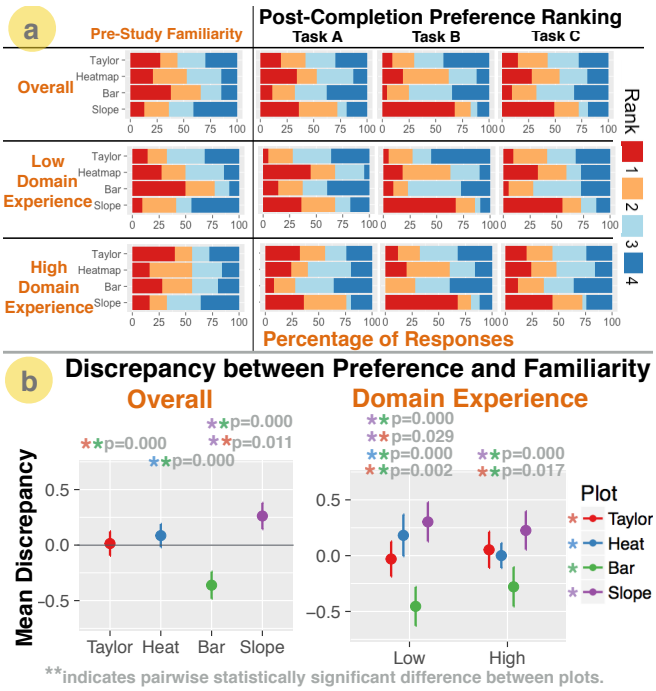

Figure 8. Post-completion preference rankings and the discrepancy between rankings based on familiarity and preference. Error bars represent 95% confidence intervals. We can that irrespective of experience many participants tended to have a high preference for the relatively unfamiliar slope plots.

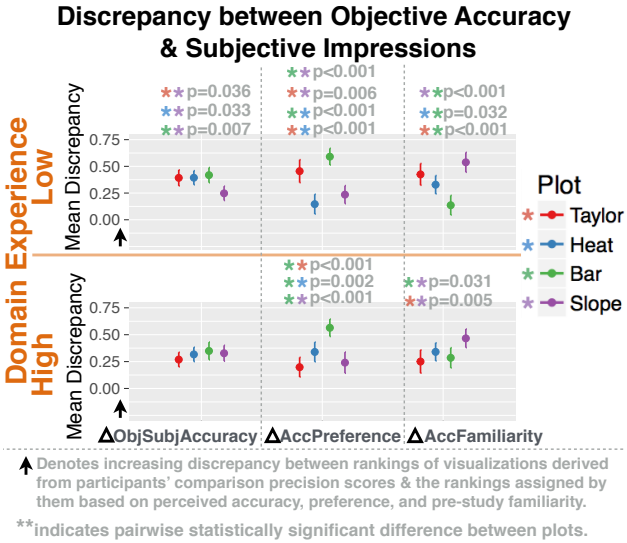

Figure 9. Discrepancy between rankings derived from objective accuracy and perceived accuracy, preference, and familiarity. Error bars represent 95% confidence intervals. The low experience group was least discrepant about their accuracy using slope plots as evidenced by the least mean discrepancy score. Slope plots also showed: i) consistently low discrepancy levels between objective accuracy and preference across both experience groups, signifying that preference for slope plots also matched performance accuracy, and ii) consistently high discrepancy levels between objective accuracy and familiarity, signifying that lack of familiarity was not a barrier in accurate performance using slope plots.

pants. These two visualizations that had a sufficient number of participants in each category to achieve statistical significance. As shown in Figure 7, we found that participants with greater familiarity with a plot type perceived that visualization to have higher accuracy and efficiency. However, for the high familiarity groups we can observe that participants' perceived levels of accuracy and efficiency with slope plots were comparable to either Taylor plots or heat maps. For both the low familiarity groups, in most cases participants expressed greater levels of perceived accuracy and efficiency with the slope plots. These findings supported our hypothesis **H3b**.

Familiarity and Preference: Figure 8a shows the distribution of the familiarity and preference rankings for each visualization. The slope plot clearly ranks much lower in familiarity, being ranked first or second for familiarity by the fewest participants (27%), followed by the Taylor plot (48%), heat map (52%), and bar chart (74%). As observed in Figure 8a, participants' preference rankings differed substantially from their pre-study familiarity rankings for slope plots. Overall, the slope plot was ranked first for Task B and Task C, which both involved dissimilarity identification, and was ranked second for Task A, which involved similarity identification. When identifying their overall most-preferred visualization, 33% of participants selected the slope plot, which is only 4% less than the heat map. Thus our findings did not support **H4a**. We compared the pre-study familiarity and post-study preference rankings using the discrepancy metric. A greater or positive discrepancy score indicates a high preference for a relatively less familiar visualization, while a lower or negative discrepancy score is associated with a low preference for a relatively more familiar visualization. As we observe in Figure 8b, slope plots (0.264, [0.153, 0.376]) attained the greatest positive discrepancy, with significant differences from the Taylor plot (0.011, $[-0.100, 0.123]$) overall ($p < 0.05$) and for low domain experience group ($p < 0.05$) (thus supporting **H4b**), and from heat map for high domain experience group ($p < 0.05$).

Discrepancy between Objective Accuracy and Subjective Impressions: We found that the discrepancy between objective and perceived accuracy (ΔObjSubjAccuracy) was lower for more experienced participants, and that within the high experience group, there was less variability in the discrepancy scores across all visualizations. As shown in Figure 9,for less experienced participants, slope plots (0.247, [0.175, 0.320]) were the least discrepant, with significant differences from all other visualizations, Taylor plots (0.392, [0.319, 0.465]), heat maps (0.393, [0.321, 0.466]), and bar charts (0.418, [0.345, 0.491]), at ($p < 0.05$). We therefore did not find evidence to support **H5a** as there were significant discrepancies. We also found limited support for **H5b**: across both experience groups, there was significantly high discrepancy between objective accuracy-based ranking and subjective preference-based rankings for bar charts ($p < 0.05$). For the low experience group we observed a similarly high discrepancy for the Taylor plots (0.455, [0.362, 0.548]), with significant differences from heat maps (0.146, [0.053, 0.239]) and slope plots (0.235, [0.142, 0.328]). The discrepancy between familiarity and objective accuracy for slope plots (Figure 9) was the highest for both experience groups, implying that irrespective of participants' experience, slope plots ranked lower with respect to familiarity, but ranked higher relative to more familiar visualizations for performance accuracy.

DISCUSSION

In this section, we reflect on our key findings and analyze participants' comments provided at the end of the study.

Effect of Experience and Familiarity: We found that domain experience and familiarity with visualizations did not drive performance accuracy (Figure 5, 6) for multi-way comparison tasks. Across all groups, participants reported similarly high levels of comfort, confidence, accuracy, and efficiency with the relatively unfamiliar slope plots (Figure 7). On the other hand, we found that less experienced scientists tended to underestimate their performance accuracy for Taylor plots. This could possibly be caused by their lower confidence and comfort ratings (Figure 7a) due to the visual complexity of Taylor plots or the associated learning curve involved in their interpretation for scientists who have less experience with this visualization. From previous research, we know high perceived accuracy, comfort, and confidence are antecedents of human-machine trust [20]. These findings lead us to an intuition that using visualization design best practices can improve accessibility, potentially overcome the barrier of low familiarity, and inspire trust in domain experts for use in deriving analytical insights [8].

Implications for Model Fidelity Visualization: In their final comments almost all participants (44 out of 47) mentioned that the slope plot was useful for easily comparing models and variables and for identifying variability across models. Examples of the answers include: *"It's much easier to compare many models on this plot. The other plots I had to compare pairwise in order for it to make sense to me; here, we get many models all at once which is more efficient (P33)."*. We also found that the use of small multiples (Taylor plots, bar charts) or large singles (heat maps, slope plots) [31] did not seem to have an effect on objective accuracy. Small multiples of bar charts were as accurate with the large single approach of slope plots for Task C, while the large single approaches were as accurate than other visualizations for Task A. Further research is needed to explore the relative advantages and disadvantages of these approaches in the context of visual comparison tasks. In the context of subjective impressions, however, there was a stronger preference for the large single approach for comparing many models and many variables as small multiples entail significant comparison complexity. This was implied by the following observation by one of the participants: "Seeing simultaneous patterns in many/many models/variables takes a lot of concentration on shapes of distributions of dots and corresponding colors of the dots - challenging!". Several participants with longer domain experience, however, commented that for a smaller number of output variables and models, they might still prefer Taylor plots or bar charts.

Adoption of Unfamiliar Visualization: Overall, we found that most participants had a high degree of preference for the slope plots irrespective of their familiarity with other plots or experience, which is indicative of potential adoption [25] by scientists in the future. Many participants mentioned the slope plot may complement other visualization approaches, or be well-suited to particular tasks. One participant commented that the *"[s]lope plot does not give an immediate intuitive understanding of the performance of a particular model, as the heatmap does. I would use it in combination with a heatmap as a more accurate but less intuitive visualization with strength in showing variability of variables (P10)." "This is probably the best visualization for looking at the spread amongst models with regards to a specific variable (P4).", etc.* In keeping with this sentiment, our collaborators have begun to adopt the slope plots as part of their workflow, and we are currently planning follow-up research on how different similarity-based visualizations (slope plots, heat maps) can be combined for exploring model fidelity patterns at different levels of detail.

CONCLUSION AND FUTURE WORK

We have presented a comprehensive analysis of objective and subjective measures of the efficacy of alternative visualization designs for climate model fidelity comparison and how these measures are mediated by the scientists' length of domain experience and prior familiarity with each visualization. Our significant findings are as follows: i) Familiarity did not drive objective accuracy or self-reported comfort with, confidence in, or preference for any visualization. ii) Objective accuracy on all tasks was similar across all visualizations, except that scientists, irrespective of the length of their domain experience, were less accurate in identifying dissimilar models using Taylor plots compared to other visualizations. iii) Scientists with greater domain experience were generally less discrepant about their perceived accuracy than less experienced scientists, while the latter group was least discrepant about their perceived accuracy with the unfamiliar slope plots. iv) The unfamiliar slope plot was the most frequently preferred visualization for the tasks presented in this study, and exhibited comparable or higher levels of objective accuracy with respect to familiar visualizations, for both highly experienced and less experienced scientists. An appreciation and preference for the slope plot were also expressed in many of the written comments submitted by the participants. This suggests that collaborative design of optimal visualizations, adapted to scientific tasks, can potentially lead to a broader acceptance and adoption of new designs within a scientific community.

In the future, we would like to extend these findings to other domains, and determine whether greater comfort and confidence levels also lead to greater scientific consensus. Especially in climate modeling, a recurring problem is the lack of well-defined consensus on the features that characterize good or bad models. We believe that effective visualizations can aid scientists in more objectively identifying patterns of model fidelity and their causes, leading to greater consensus regarding model fidelity, and improved efficiency in model calibration and assessment activities. This will ultimately contribute to improved understanding of global climate change patterns.

ACKNOWLEDGMENT

We thank Daniel Tompkins for his assistance in the metrics calculations and plotting. This research was funded by the Laboratory Directed Research and Development Program (LDRD) at the Pacific Northwest National Laboratory, which is operated by Batelle for the U.S. Department of Energy (DOE) under contract DE-AC05-76RLO01830.

REFERENCES

1. Jürgen Bernard, Martin Steiger, Sven Widmer, Hendrik Lücke-Tieke, Thorsten May, and Jörn Kohlhammer. 2014. Visual-interactive Exploration of Interesting Multivariate Relations in Mixed Research Data Sets. In *Computer Graphics Forum*, Vol. 33. Wiley Online Library, 291–300.
2. Jacques Bertin. 1983. *Semiology of graphics: diagrams, networks, maps*. University of Wisconsin press.
3. Matthew Brehmer, Jocelyn Ng, Kevin Tate, and Tamara Munzner. 2016. Matches, Mismatches, and Methods: Multiple-View Workflows for Energy Portfolio Analysis. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 449–458.
4. Sabrina Bresciani and Martin J Eppler. 2009. The benefits of synchronous collaborative information visualization: Evidence from an experimental evaluation. *IEEE transactions on visualization and computer graphics* 15, 6 (2009), 1073–1080.
5. Bertjan Broeksema, Alexandru C Telea, and Thomas Baudel. 2013. Visual Analysis of Multi-Dimensional Categorical Data Sets. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 158–169.
6. Curt Covey, Krishna M AchutaRao, Michael Fiorino, Peter J Gleckler, Karl E Taylor, and Michael F Wehner. 2002. *Intercomparison of climate data sets as a measure of observational uncertainty*. UCRL.
7. Aritra Dasgupta and Robert Kosara. 2010. Pargnostics: Screen-Space Metrics for Parallel Coordinates. *Transactions on Visualization and Computer Graphics* 16, 6 (2010), 1017–26.
8. Aritra Dasgupta, Joon-Yong Lee, Ryan Wilson, Robert A Lafrance, Nick Cramer, Kristin Cook, and Samuel Payne. 2017. Familiarity Vs Trust: A Comparative Study of Domain Scientists' Trust in Visual Analytics and Conventional Analysis Methods. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 271–280.
9. Aritra Dasgupta, Jorge Poco, Yaxing Wei, Robert Cook, Enrico Bertini, and Claudio T. Silva. 2015. Bridging Theory with Practice: An Exploratory Study of Visualization Use and Design for Climate Model Comparison. *IEEE Transactions on Visualization and Computer Graphics* 21, 9 (2015), 996–1014.
10. Michael Friendly and SAS Institute. 2000. *Visualizing categorical data*. Sas Institute Cary, NC.
11. David Gefen. 2000. E-commerce: the role of familiarity and trust. *Omega* 28, 6 (2000), 725–737.
12. Peter J Gleckler, Karl E Taylor, and Charles Doutriaux. 2008. Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres* 113, D6 (2008).
13. Michael Gleicher, Danielle Albers, Rick Walker, Ilir Jusufi, Charles D. Hansen, and Jonathan C. Roberts. 2011. Visual comparison for information visualization. *Information Visualization* 10, 4 (2011), 289–309.
14. Mark Harrower and Cynthia A Brewer. 2003. ColorBrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal* 40, 1 (2003), 27–37.
15. Yosef Hochberg. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75, 4 (1988), 800–802.
16. James W. Hurrell, M.M. Holland, P. R. Gent, S. Ghan, J. E. Kay, P. J. Kushner, J.-F. Lamarque, W.G. Large, D. Lawrence, K. Lindsay, W. H. Lipscomb, M. C. Long, N. Mahowald, D. R. Marsh, R. B. Neale, P. J. Rasch, S. Vavrus, M. Vertenstein, D. Bader, W.D. Collins, J.J. Hack, J. Kiehl, and S. Marshall. 2013. The Community Earth System Model: A Framework for Collaborative Research. *Bulletin of the American Meteorological Society* (2013), 1339–1360.
17. Alfred Inselberg and Bernard Dimsdale. 1990. Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry. In *IEEE Visualization*. IEEE CS Press, 361–378.
18. Kurt Koffka. 2013. *Principles of Gestalt psychology*. Vol. 44. Routledge.
19. Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. 2012. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics* 18, 9 (2012), 1520–1536.
20. Daniel J McAllister. 1995. Affect-and cognition-based trust as foundations for interpersonal cooperation in organizations. *Academy of management journal* 38, 1 (1995), 24–59.
21. T. Munzner. 2009. A nested process model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics* 15, 6 (2009), 921–928.
22. Allan H Murphy. 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review* 116, 12 (1988), 2417–2424.
23. R B. Neale, C. C. Chen, A. Gettelman, P. H. Lauritzen, S. Park, D. L. Williamson, A. J. Conley, R. Garcia, D. Kinnison, J-F. Lamarque, D. Marsh, M. Mills, A. K. Smith, S. TIlmes, F. Vitt, H. Morrison, P. Cameron-Smith, W. D. Collins, M. J. Iacono, R. C. Easter, S. J. Ghan ans X. Liu, P. J. Rasch, and M. Taylor. 2012. *Description of the NCAR Community Atmosphere Model: CAM5.0*. Technical Report NCAR/TN-486+STR. National Center for Atmospheric Research, Boulder, Colorado, USA. 268 pp., URL = `http://www.ccsm.ucar.edu/models/atm-cam`.
24. Intergovernmental Panel on Climate Change. 2015. *Climate Change 2014: Mitigation of Climate Change*. Vol. 3. Cambridge University Press.

25. Catherine Plaisant. 2004. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*. ACM, 109–116.

26. Dominik Sacha, Hansi Senaratne, Bum Chul Kwon, Geoffrey Ellis, and Daniel A Keim. 2016. The role of uncertainty, awareness, and trust in visual analytics. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 240–249.

27. Leila Takayama and Eser Kandogan. 2006. Trust as an underlying factor of system administrator interface choice. In *CHI'06 extended abstracts on Human factors in computing systems*. ACM, 1391–1396.

28. Karl E Taylor. 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres* 106, D7 (2001), 7183–7192.

29. Edward R. Tufte. 2001. *The Visual Display of Quantitative Information* (2nd ed.). Graphics Press.

30. Ananth Uggirala, Anand K Gramopadhye, Brain J Melloy, and Joe E Toler. 2004. Measurement of trust in complex and dynamic systems using a quantitative approach. *International Journal of Industrial Ergonomics* 34, 3 (2004), 175–186.

31. Stef van den Elzen and Jarke J van Wijk. 2013. Small multiples, large singles: A new approach for visual data exploration. In *Computer Graphics Forum*, Vol. 32. Wiley Online Library, 191–200.

32. Edward Vonesh and Vernon M Chinchilli. 1996. *Linear and nonlinear models for the analysis of repeated measurements*. CRC press.

33. C. Ware. 2004. *Information visualization: Perception for Design*. Morgan Kaufmann.