This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2934540, IEEE Transactions on Visualization and Computer Graphics

Separating the Wheat from the Chaff: Comparative Visual Cues for Transparent Diagnostics of Competing Models

Aritra Dasgupta, Hong Wang, Nancy O'Brien, and Susannah Burrows

Abstract—Experts in data and physical sciences have to regularly grapple with the problem of competing models. Be it analytical or physics-based models, a cross-cutting challenge for experts is to reliably diagnose which model outcomes appropriately predict or simulate real-world phenomena. Expert judgment involves reconciling information across many, and often, conflicting criteria that describe the quality of model outcomes. In this paper, through a design study with climate scientists, we develop a deeper understanding of the problem and solution space of model diagnostics, resulting in the following contributions: i) a problem and task characterization using which we map experts' model diagnostics goals to multi-way visual comparison tasks, ii) a design space of comparative visual cues for letting experts quickly understand the degree of disagreement among competing models and gauge the degree of stability of model outputs with respect to alternative criteria, and iii) design and evaluation of MyriadCues, an interactive visualization interface for exploring alternative hypotheses and insights about good and bad models by leveraging comparative visual cues. We present case studies and subjective feedback by experts, which validate how MyriadCues enables more transparent model diagnostic mechanisms, as compared to the state of the art.

Index Terms—Visual comparison, Visual cues, Model evaluation, Transparency, Simulation

1 Introduction

Distinguishing between the best and the worst, among a set of competing alternatives, is a pervasive analytical problem. A common instance of this problem is when domain experts want to diagnose which models, among a set of competing alternatives, most appropriately simulate or predict real-world phenomena. This requires significant time and human effort, whereby experts combine their domain knowledge with a data-driven understanding of the trade-offs and nuances involving multiple models. Complexity in such diagnostic evaluation process stems from experts' need to reconcile many outputs, from multiple models, and many ways to evaluate the quality of competing outputs, for ultimately selecting good models.

Depending on the goal for model selection, experts have to consider a suite of domain-specific criteria. These include criteria based on transparency and interpretability for predictive modeling [7, 26] or statistical fidelity criteria based on output-observation matches for simulation modeling [8], which is the focus of this paper. A cross-cutting, domain-agnostic challenge in these modeling scenarios is to develop reliable analytical solutions that experts can adopt for overcoming the inherent complexity of model diagnostics process.

To address this challenge, through a collaboration with climate scientists, we study how visualization can be used for ensuring reliable **post-hoc diagnostics** of climate models. Understanding differences among many climate model outputs is a challenging task. This is usually done (Figure 1) by comparing the simulation outputs to observation data captured from satellites, ground-based sensors, etc. For example, let us say that a climate model simulates cloud cover over a region. This simulated output is compared with the observed cloud cover over a region. The degree to which the simulated and observed output match constitutes the **fidelity** of a model. Fidelity is usually quantified using a suite of metrics, such as correlation or root-mean-squared distance between simulated and observed outputs.

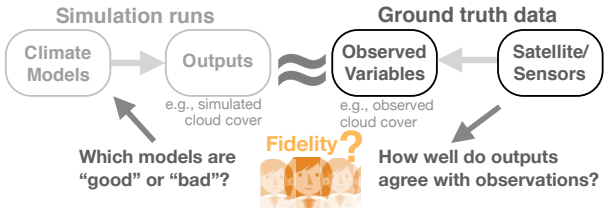

Fig. 1. **A conceptual sketch of post-hoc model diagnostics.** Scientists need to diagnose which models have high or low fidelity. Fidelity is defined by statistical metrics that score model outputs based on how closely they match observation data.

In many real-world applications (e.g., perturbed physics or perturbed initial condition ensembles), scientists compare hundreds of simulation models, tens of model outputs (e.g., cloud cover, temperature, aerosol content, etc.), and many different metrics that quantify the fidelity of a model. Because of the complexity of this task, scientists typically spend weeks or months carefully, and often, manually, verifying multiple aspects of each model output. Adoption of more automated approaches to model evaluation has been hindered in part by two key challenges in determining appropriate overall metrics for systematic model evaluation, as reflected in our previous survey of scientists' model diagnostics practices [5]. This survey demonstrated a lack of consensus within the scientific community about the relative importance of the factors (i.e. outputs, metrics) contributing to the overall fidelity of a model. Additionally, through scientists' subjective comments in the survey, it was recorded that current analytical tools do not provide the flexibility to explicitly capture scientists' assumptions, and to understand how robust their overall evaluation of models is to those assumptions.

To alleviate these problems, we contribute a design study through which we demonstrate model diagnostics processes can be made more reliable using *multi-way visual comparison* techniques. At the core of our solution, are **comparative visual cues**, which facilitate preattentive search for model disagreement patterns thereby reducing the complexity of visual search across many combinations of models, outputs and metrics. This in turn, drastically increases the return on investment of scientists' time and effort for selecting the best models. We have three main contributions as part of this design study. First, we provide a characterization of the model diagnostics problem and identify a set of multi-way visual comparison tasks. Second, we derive a design space of task-driven comparative visual cues using a classifi-

- *Aritra Dasgupta is with New Jersey Institute of Technology. E-mail: aritra.dasgupta@njit.edu.*
- *Hong Wang is with Arizona State University, Inc.. E-mail: hong.wang@asu.edu.*
- *Nancy O'Brien is Pacific Northwest National Laboratory. E-mail: nancy.obrien@pnnl.gov.*
- *Susannah Burrows is with Pacific Northwest National Laboratory. E-mail: susannah.burrows@pnnl.gov.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2934540, IEEE Transactions on Visualization and Computer Graphics

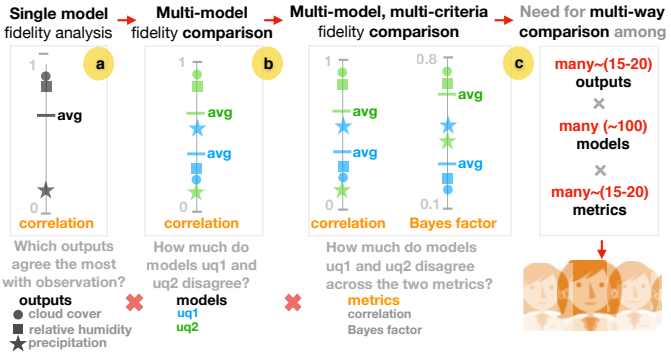

Fig. 2. **Understanding the complexity of multi-way comparisons.** The vertical axes in a,b, and c, represent fidelity scores for different model outputs, such as cloud cover, relative humidity, and precipitation. Scientists need to reconcile information from hundreds of comparisons among many models, outputs, and fidelity metrics, for judging the consistency and robustness of model fidelity levels.

cation scheme. Third, we contribute MyriadCues, a tool for providing climate scientists with an interactive mechanism to build alternative hypotheses about the factors affecting model fidelity levels and make reliable judgments about good and bad models. We provide a detailed case study to demonstrate how the tool helped climate scientists gain insights about the consistency (i.e. the degree to which models agree about an output) and robustness (i.e. the degree to which fidelity levels change under different conditions) of model outputs and outline the lessons learned from expert feedback about the efficacy of MyriadCues.

2 Multi-criteria Model Fidelity Analysis

Climate models are complex computer simulations of the physical, chemical, and biological processes shaping our environment [35]. Models differ in the algorithms and codes that are used, as well as in their parameter configurations and boundary conditions. At climate modeling centers worldwide, development efforts that lead to a new model version are followed by a post-hoc, time-intensive model calibration effort, whereby experts examine statistical model fidelity metrics and other diagnostics to determine which configurations produce credible realizations of different climate phenomena [16]. Statistical fidelity metrics for climate models [9] measure the degree to which model outputs match observation (Figure 1). The greater the fidelity of a model, the closer its agreement with observations of present-day and historical climate phenomena.

Besides the choice of fidelity **metrics**, expert judgment is required to decide which outputs to include in the fidelity calculations, and how much weight should be assigned to each of them. This is, therefore, a multi-criteria fidelity analysis problem, where the overall fidelity of a **model** is given by the weighted average of the fidelity scores for each **output**, for a given metric. In a simplified example, as shown in Figure 2a, the overall fidelity of the model is quantified as the weighted average of the correlation metric for three outputs: *cloud cover*, *relative humidity* and *precipitation*.

Figure 2 demonstrates the complexity of the multi-criteria fidelity analysis problem from left to right. The tasks are simpler when evaluating a single model, `{uq1}`, using a single metric, such as correlation (Figure 2a.). The model has high overall fidelity with the exception of the precipitation output. When multiple models, `{uq1}` and `{uq2}` are compared, ((Figure 2b.) we can observe that `{uq2}` has higher average fidelity, while there are disagreements about the precipitation output between the two models, with respect to the correlation metric. When multiple metrics are compared next (Figure 2a.), we can observe that the Bayes factor metric is more consistent with regards to the precipitation output from both the models. However, with respect to the average, and other outputs, `uq1` still has higher fidelity than `uq2`. In this case, experts, might decide to put less weight on the precipitation output due to the recorded difficulty in simulating precipitation precisely [43]. This will lead them to reliably judge, with respect to both metrics, that `{uq1}` is a better model than `{q2}`. This example illustrates a simplistic scenario of comparison among two models, three outputs, and two metrics. In reality, scientists often grapple with a much larger and complex comparison space, with hundreds of simulations, tens of outputs and metrics. This necessitates an analytical solution that will enable them to efficiently perform multi-way comparison tasks.

Multi-criteria fidelity analysis can be theoretically framed as a multi-criteria decision analysis problem (MCDA) [12, 23], where a human decision needs to be informed by many alternatives and many criteria. However, in practice, it is difficult to apply automated models of MCDA, which require specification of trade-offs among criteria as inputs. These relative trade-offs are not necessarily known a prior, and discovering how best to balance different criteria when comparing models is an open problem. Here we provide tools to support a more efficient process for comparing multiple models on multiple criteria, and improve transparency for scientists seeking to understand the impact of the trade-offs they make between criteria when selecting models.

3 Related Work

Our contributions span two areas of research: i) the design space of visual comparison and ii) visualization-driven model evaluation.

3.1 Design Space of Visual Comparison

Comparative analysis is an integral part of quantitative reasoning [45]. However, in most existing task classification schemes, comparison tasks have been treated as a monolith, with the exception of the recent work by Gleicher [10], where a set of challenges and considerations are presented for reasoning about comparative visualization techniques. Gleicher observed that the comparison methods described in most existing visualization systems focus on pairwise comparison or comparison among very few objects [15, 22, 30, 32]. Recent studies on design implications for visual comparison tasks [42] focus on simple, retrieval-based comparisons among a few categories. In contrast, we focus on the scale and complexity challenges of visual comparisons. We extend the classification scheme proposed by Gleicher [10] and the previously proposed space of encodings based on superposition, juxtaposition, and explicit encoding [11], for reasoning about the visualization design space of multi-way visual comparison tasks.

Multi-way comparisons are challenging because of both the scale (i.e., the number of distinct objects, which are models, metrics, and outputs) and complexity (i.e., the size of the objects, which is given by the number of models, metrics, and outputs) of the tasks. For making the comparison tasks efficient, we use comparative visual cues that are systematically derived based on perceptual principles [19, 41] and that guide experts' attention to salient fidelity patterns of interest.

Comparison mechanisms have been previously used for evaluating topic models [1]. The Buddy plots technique scales to hundred of topics but only supports pairwise comparison between models. In our design space, we consider more complex comparisons, for multi-criteria fidelity analysis, among a combination of many models, many output variables and metrics. We propose and leverage a classification scheme for overcoming the scalability and complexity challenges [10, 46] associated with those tasks, based on a small multiple [45] based design. We realize the design space in MyriadCues, an interactive visualization interface for multi-criteria model fidelity analysis. A key functionality of MyriadCues is to provide visual guidance on key changes to model fidelity levels and the disagreement among metrics in response to expert assigned weights to outputs. Had there been a consensus in the climate science community about the different trade-offs involving the contribution of outputs to fidelity levels, we could have used an approach similar to Pajer et al. [34]. They developed an MCDA tool named Weightlifter that directly visualizes the trade-offs in the decision space based on automated additive weighting strategies after experts have input their preferences in terms of weights or trade-offs. In MyriadCues, due to the unknown nature of these trade-offs, we allow more direct multi-way comparison of what-if scenarios with respect to understanding the effect of the weights on both model rankings and the metrics. An outcome of the use of MyriadCues is a more nuanced

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2934540, IEEE Transactions on Visualization and Computer Graphics

understanding of how different trade-offs could explain variability in model fidelity rankings.

3.2 Visualization-Driven Model Evaluation

Human-centered analysis of simulation models falls into four broad categories: i) analysis of similarities and differences in model outputs [20, 36], ii) analysis of input-output relationships [39], iii) visual communication of model decisions to non-expert users [4] and iv) post-hoc model performance evaluation with respect to ground truth [8, 26, 29]. While this categorization is generally true for any domain, here we focus only on the climate science domain and discuss our contributions in the last category.

Many visual analytic related methods focus on analyzing output similarity or exploring the effect of high-dimensional parameter spaces on model outputs. For example, Kehrer et al. proposed a faceted approach towards similarity analysis of multiple outputs from a model over space and time [21], and this was extended by Poco et al. for supporting similarity-based comparison for multiple models and outputs [36]. However, both approaches are limited by the number of models (< 10). For parameter-space analysis, Wang et al. proposed a nested parallel coordinates based visualization system [47], while Poco et al. used a visual reconciliation method for understanding the effect of input parameters on output similarity [37].

The goal of selecting appropriate model parameters is to achieve optimal performance from climate models, for which visual steering based techniques [39, 48] can be used. However, an open question in climate science is: *which metrics and output variables should be considered for qualifying model performance as good or bad?* Without effective methods to quantify model performance (i.e., fidelity), it is difficult to define an objective cost function for parameter tuning, which explains the time and effort spent by scientists in the tuning process [16].

With the exception of the work from Kothur et al. [24], where they use reference data for assessing performance of ocean models, there is little research using interactive visualization for multi-criteria fidelity analysis for climate models. Existing visualization approaches for model performance analysis are mostly static, suffer from clutter [9], and do not scale beyond a few models and variables [6,44]. In this work, we address this gap by using scalable, interactive visual comparison methods derived through participatory design.

4 Model Diagnostics Task Abstraction

The first phase of our design study was focused on developing a shared understanding of the model diagnostics goals between climate scientists and visualization researchers. We collaborated with two climate scientists from a national laboratory, with an average experience of 15 years between them, for over a period of 2 years. We followed the nested model [31] where a problem characterization phase was followed by the iterative stages of visualization task analysis, design, and evaluation. One of the climate scientists (a co-author of this paper) acted as a liaison [40] between the climate science and visualization research groups and helped us facilitate interviews, build a shared understanding of the state-of-the-art visualization techniques, and conduct participatory design sessions.

4.1 Scientific Goals

We derived the following domain specific goals that were relevant for solving this problem:

G1: Hypothesize about model fidelity. Scientists need to formulate an initial hypothesis about "good" or "bad" models with respect to a preferred metric. Climate science groups working on the model diagnostics problem may have a preferred metric and they use it to understand, with respect to an average fidelity score, which models could be good or bad.

G2: Judge contribution of model outputs. Scientists need to evaluate or refine the hypothesis by inspecting the contribution of many outputs. Scientists are generally looking for cases where fidelity levels are dissimilar for a given model across multiple output variables, and also for different models for a given output variable.

G3: Assess fidelity consistency. Scientists might start with a preferred metric but to test the consistency of the fidelity levels of a model they often use a suite of statistical criteria. Fidelity is a proxy to understand how much disagreement there is among models: two different fidelity scores imply that the the model outputs were different. Often, they need data-driven guidance for selecting a set of metrics for comparison.

G4: Assess fidelity robustness. Scientists need to investigate how assigning different weights to output variables affect the fidelity levels with respect to the chosen set of metrics, and also how they change the agreement or disagreement levels across the metrics. The more invariant the fidelity levels, a model is assessed to have more robust levels of fidelity.

4.2 Comparison tasks

We involved our collaborators for distilling specific comparison tasks for satisfying their analysis goals, as outlined below:

T1: Identify best/worst models. Identify top-ranked and bottom-ranked models and detect small differences. For developing an initial hypothesis about good or bad models (**G1**), scientists first want to rank-order the models and visually identify the best and worst ones. They also want to detect small differences in average fidelity values for models, as given by a metric.

T2: Compare average fidelity. Visualize the magnitude distribution of all models as the context for comparison. While scientists want to focus on the top-ranked or bottom-ranked models, for developing a robust hypothesis (**G1**) they also need to understand how the average fidelity scores are distributed across all models and establish an appropriate context for judging if the scores are reliable or not.

T3: Compare model-output dissimilarity. For addressing **G2**, scientists want to understand if the fidelity scores for individual output variables are consistent with the average/overall fidelity score (e.g., a model having high fidelity score can have poor fidelity on one variable) and also find the degree of variability in fidelity scores for one particular variable (e.g., several models may disagree about the fidelity scores leading to a high interquartile range). Similar levels of fidelity across models and output variables is an expected pattern and the main pattern of interest is the degree of disagreement. Even finding very small differences is of interest to the scientists.

T4: Compare metric-metric disagreements. Visualize which metrics disagree, and by how much, across models and outputs. Scientists usually start with their preferred metric or select a metric based on their intuition. Each metric has different scale and semantics. They need to visually judge the degree to which metrics disagree about both the overall fidelity levels and at the level of each model-output combination (**G3**).

T5: Understand fidelity change. This is a change detection task in a comparative setting where scientists need to know how the average scores and output-specific scores change in response to scientist-defined weights to those variables (**G4**). Scientists are mainly interested in spotting the big changes.

T6: Understand disagreement change. Gauge how weighing outputs affect metric agreement/disagreement. This is also a change detection task in a comparative setting, where scientists need to observe how assigning different weights to variables can make metrics agree or disagree more about the overall and variable level fidelity scores (**G4**).

An important take-away from this task distillation was that the comparison tasks are unlikely to be executed sequentially or in isolation. It is a common scenario for scientists not to have a prior hypothesis about expected model behavior. In that case, they start from task T4 and T6, which are complex multi-way comparison tasks focusing on assessment of consistency and robustness, and subsumes other tasks. To facilitate such composite **multi-way comparison**, we consider a set of unitary tasks for informing the task-driven design space.

5 Comparative Visual Cues

In the second phase of our study, we characterized the design space for achieving the multi-way comparison tasks. We conducted discussion sessions between the climate science and visualization teams where we jointly critiqued existing visualization solutions and as an outcome

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2934540, IEEE Transactions on Visualization and Computer Graphics

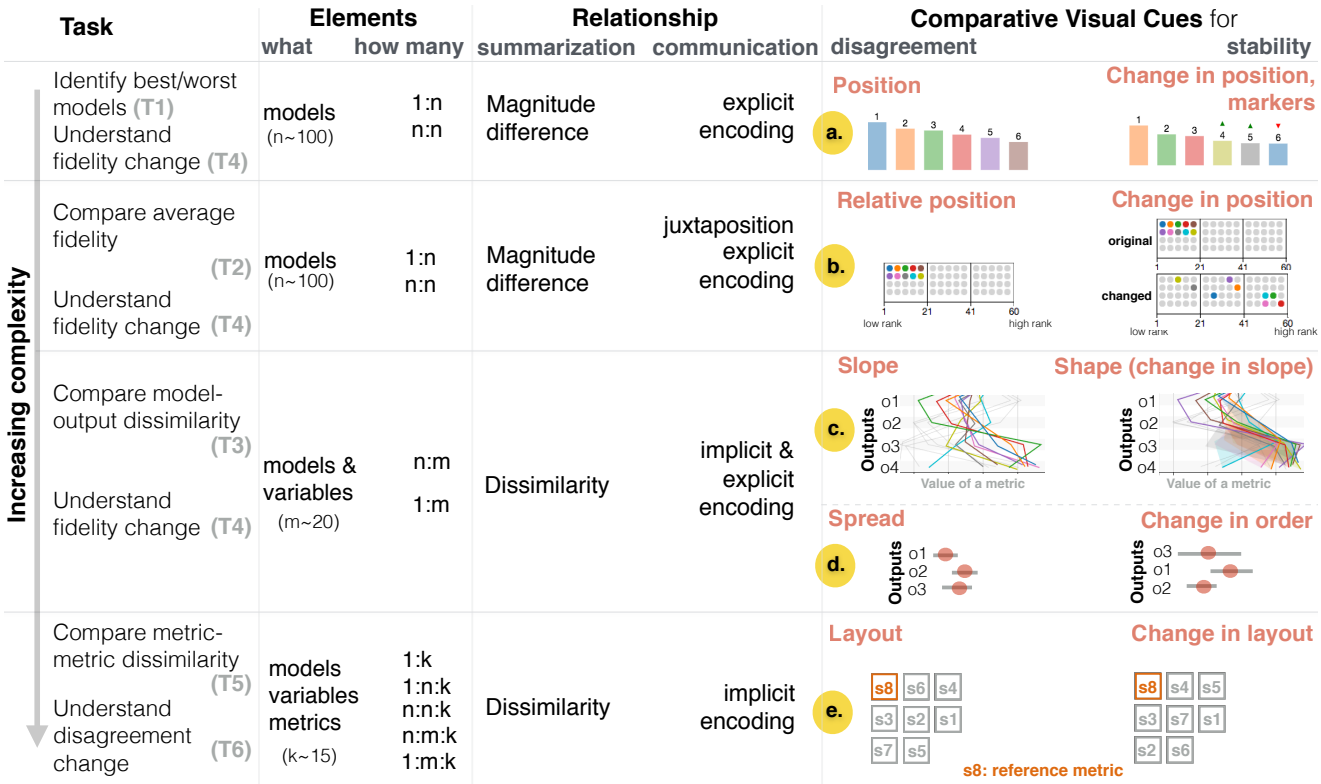

Fig. 3. **A classification scheme for deriving comparative visual cues** that address the tasks (T1, T2, T3, T4, T5, T6) for climate model fidelity analysis. The visual cues, by leveraging the perceptual principles of visual encoding, help minimize comparison complexity by letting scientists readily spot patterns of disagreement and stability across many combinations of models, metrics, and output variables

of many participatory design sessions, we derived a set of comparative visual cues through a classification scheme. Comparative visual cues leverage pre-attentive properties of visual encodings to facilitate efficient search across many combinations of models, variables, and metrics, for spotting small differences while maximizing accuracy of comparisons. In this section, we first discuss the classification scheme and then discuss the task-driven visual cues.

5.1 Classification Scheme

We adapted and extended the classification scheme proposed by Gleicher [10] for deriving a set of task-driven visual cues. In Figure 3, we describe our classification scheme. *Elements* indicate *what*, among models, metrics, and output variables, are being compared; and also *how many* elements are being compared to indicate the complexity of the task. Combining multiple elements, like hundreds of models and tens of variables has a multiplicative effect on the **scale** and **complexity** of comparisons. In terms of *relationships* among the compared elements, scientists are mainly interested in finding small differences in magnitude and also understanding which metrics, models, or output variables, and their combinations, are most dissimilar than others. In our scheme, we describe how we summarize these relationships which ultimately guide how they are visually communicated through comparison designs like juxtaposition or implicit and explicit encoding [11]. When the relationships among data objects are approximately recoverable but not precisely encoded in a visualization, we term this as implicit encoding. A simple scatter plot is a good example, where the degree of correlation between two dimensions is approximately recoverable even without any explicit encoding of the correlation. Other examples include quality metric [3] based reordering of layouts or dimensions where one can gauge how closely related data objects are, using the metrics as the guides. The last part of our classification scheme is about realizing the comparison tasks by optimizing the visual search process. To this end, we first needed to know which patterns we should optimize for and accordingly decide the visual cues necessary for guiding scientists' attention to those patterns. A sequential search for patterns for each of the many possible comparisons would be time-consuming and ineffective. These comparative cues leverage the human vision system's capability to preattentively process patterns [14], thereby leading to a much more efficient parallel visual search. The patterns that scientists are mainly looking for are as follows:

Visualizing disagreement: Fidelity is a lens to understand how much disagreement there is among models: two different fidelity scores imply that the the model outputs were different. Similarity of fidelity levels is the "normal" pattern because simulations, if perfectly parameterized and calibrated, should all produce similar outputs resulting in similar fidelity scores. However, in reality, scientists have to reliably understand, where disagreements occur and exercise their expert judgment to reason about and resolve those.

Visualizing stability: Stability of a model output or a metric is given by the degree to which the fidelity levels are insensitive to different weights assigned to multiple variables. These are important factors for scientists to consider while they come to the final judgment of which models are the best and the worst, and also, which metrics are most effective in capturing the "true" fidelity of a model. Usually, there are inherent trade-offs exploring which scientists can conclude under which specific scenarios or conditions models and metrics are stable.

5.2 Cues for comparing model dissimilarity

To satisfy T1, which is the simplest among all the tasks, a visualization needs to facilitate relative judgment of rank and magnitude (i.e., the average fidelity scores) with a high degree of **accuracy**. To satisfy T2, a point-based visualization is needed for looking at the magnitude distribution of all models: one which is **scalable** with respect to about 100 models while at the same time allowing scientists to readily identify a particular model.

Elements: The comparison ($1:n$) involves many models for understanding small differences in average fidelity scores. T4 involves an $n:n$ comparison as one has to compare two sets of models, with two different weighing schemes.

Relationship: For comparing magnitude difference across all models, we use the average scores of models (across all variables) for ranking, that can be used for sub-setting the top-ranked or bottom-ranked models.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2934540, IEEE Transactions on Visualization and Computer Graphics

For summarizing pairwise (1 : 1) relationships between models, we use the Euclidean distance as a measure of magnitude difference.

Understanding disagreement: We use relative positions of models in terms of their average magnitude and rank. Magnitude difference among models is expressed through a rank ordering of models and the heights of the bars (Figure 3a). represent the average fidelity score for a model. For representing all models, we choose a space-efficient encoding that can represent all the models while at the same time indicating their ordering with respect to their positions. As shown in Figure 3b, the position of a dot indicates the rank of a model and any change in position is quickly reflected by juxtaposing the two views, thus representing sensitivity in rank changes to expert-defined weights. While a box plot or a bean plot [18] could be used as summaries of magnitude differences, here the goal was to directly identify the high or low ranked models and use those as subsets for focusing the analysis.

Understanding stability: Changes in position can be hard to track if multiple models change rank-based position at once. For this reason, we provide explicit cues based on markers (arrow-heads) which indicate upward (green arrow) or downward (red arrow) trend of rankings.

5.3 Cues for comparing model-output dissimilarity

To satisfy both T3 and T4, which are more complex than T1 and T2, a visualization has to **expressive** [28] about dissimilar patterns so that scientists spend minimal effort and time to detect them, among a large number of model-output variable combinations. For T4, we need to carefully consider the trade-off between scalability and effectiveness: if there are many changes, we need to show only the significant ones so that scientists can quickly understand the effect of the weights they assign to variables.

Elements: The comparison involves combinations of many models and many variables, therefore needing both one-to-many ($1 : m$) comparison among variables and many-to-many comparisons $n : m$ among models and variables.

Relationship: For summarizing relationships among models, we use lack of correlation, with respect to the fidelity scores across all variables, as a measure of dissimilarity. For T3, we use implicit encoding for communicating dissimilarity among models and for T4, we explicitly encode salient changes.

Understanding disagreement: We use both position among models and the connection among them, through lines, as cues (Figure 3c). We term this plot as the slope plot [6], a hybrid between a slope graph [45] and parallel coordinates [17]. For encoding the dissimilarity (n:m) among models and variables, we considered two options. One of the options was to use a color scheme to indicate the degree of differences, which could have led to a heatmap based design. But color is less accurate than position [2], especially in communicating small differences, which scientists most interested in. Therefore, we decided to use positions of models along continuous axes that represented different variables, and connect those positions along multiple axes by a polyline. These polylines also added Gestalt effects [19] of continuity, proximity, and connectedness, leveraging which experts could readily integrate the differences for a single model across multiple variables. These Gestalt effects help in implicitly encoding the differences among multiple models, resulting in efficient visual scanning and tracking of the differences across multiple models, variables, and metrics, simultaneously without putting too much cognitive load on the experts. Cues about variables are provided by explicitly encoding the spread in terms of the interquartile range. We also considered a multi-dimensional projection based layout as an alternative design where many-to-many comparison would be possible on a scatter plot where relative distances indicated differences among models. However, since this involved an abstraction over pairwise distances, and the contributions of each variable would be hard to recover, our experts did not prefer this method.

Understanding stability: As shown in Figure 3c, unstable models, in response to differing variable weights, is expressed by shapes: by drawing envelopes around the lines, the side containing the line indicates the current weighted value of the metric and other side indicates previous value. Cues about variables are additionally provided by a change in vertical ordering of the variables, based on the degree of change.

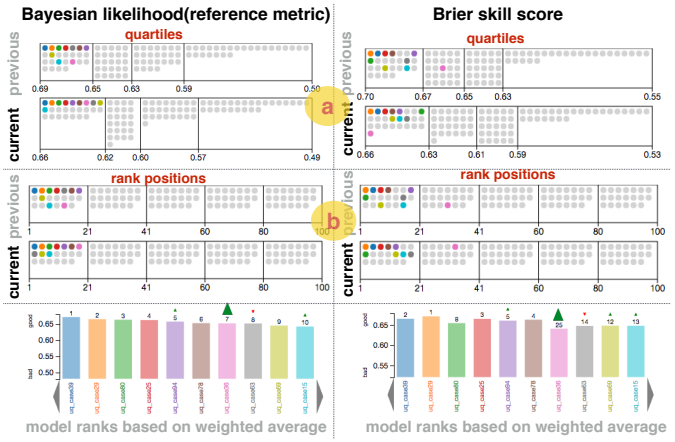

Fig. 4. **Configurable magnitude plots**. Experts can either choose quartiles or rank based bins for understanding the ordering of models computed by their weighted average of fidelity scores. The dot plots complement the bar charts by providing flexibility to select models from any range and also the ability to see changes among multiple models readily when weights are adjusted. Scattered dots provide an immediate cue about high sensitivity of models to the weight changes.

5.4 Cues for comparing metrics

T5 and T6 are the most complex tasks in this set as they involve multi-way comparison across models, variables, and metrics. As mentioned before, these tasks subsume T1, T2, T3, T4. For both these tasks, a visualization has to be **effective in linking and tracking** model-output combinations across multiple metrics.

Elements: The elements under comparison are metrics, models, and variables. A noteworthy point here is that scientists generally choose a particular metric as a reference and compare the outcome of other metrics with respect to the reference, leading to a $1 : k$ comparison first (k being the number of metrics), which is followed by repeating all the other one-to-many or many-to-many comparison tasks k times.

Relationship: Sensitivity of magnitude difference or correlation across models and metrics needs to be computed when experts assign different weights to the variables interactively. By default, all variables are equally weighted. When different weights are applied, for summarization of $1 : k$ relationships across all metrics, we first perform an aggregation over all models based on their correlation or magnitude difference scores. Next we compute the difference between the aggregated scores for each metric to indicate which metrics disagree the most.

Understanding disagreement: We use small multiples for comparing across different metrics and use layout of the small multiples as cues for indicating relative dissimilarity. Dissimilarity or differences with respect to a reference metric (1:k association) is communicated through implicit encoding: adjusting the layout of small multiples, where the proximity of small multiples represent the degree of difference between a reference and an associated plot. The layout can be adjusted either by choosing the magnitude difference (i.e., the average difference in ranks of models for a reference metric and and for an associated metric) or the correlation (average correlation across models for a reference metric and an associated metric).

Understanding stability: A change in layout indicates which metric was the most sensitive to the change in weights. The layout could be ordered in the following ways: juxtapose the most similar or dissimilar metric (with respect to the reference metric) with the reference metric, and juxtapose the metric that changed the most with the reference metric. A sudden change in layout can cause change blindness [41]. Therefore, care is taken to provide options to the user for reordering small multiples, which they can control interactively.

6 Interactive Multi-Way Comparison in MyriadCues

We instantiate the classification scheme through MyriadCues (Figure 5), an interactive, web-based visualization interface that resulted from multiple participatory design sessions involving visualization researchers

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2934540, IEEE Transactions on Visualization and Computer Graphics

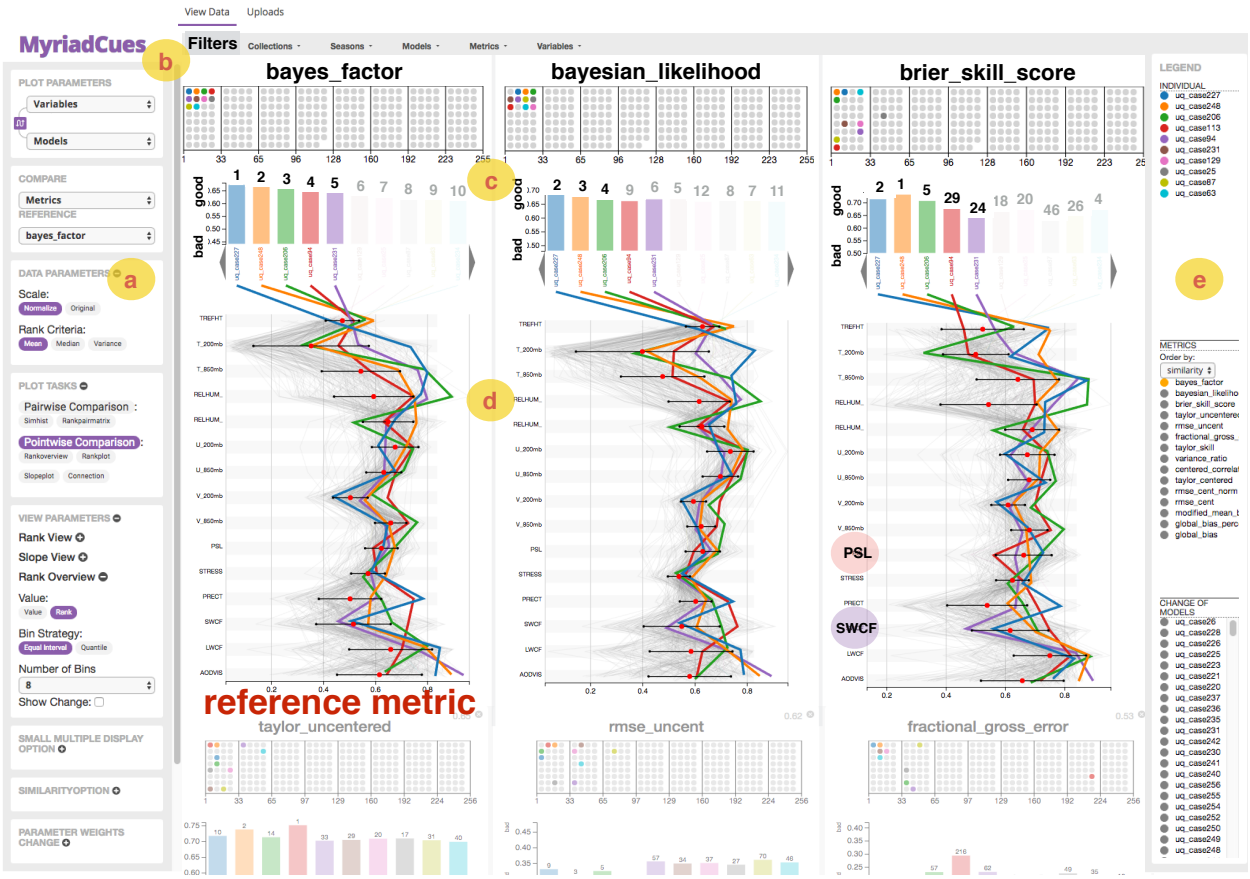

Fig. 5. **MyriadCues comprises**: a) a set of controls for configuring different visualizations by selecting axes, data normalization and ranking strategies, and adjusting the parameters for different views; b) a set of filters for subsetting across different elements; c) Magnitude Plots for showing magnitude differences across models; d) Slope Plots for showing dissimilarities across models and variables; and e) legends and guides for navigating the visualizations. In this view, *bayesfactor* is the reference metric, which means that models are color-coded based on their ranks with respect to *bayesfactor*. These colors are used to link models in other small multiples, where their respective ranks are shown.

and climate scientists. Interactivity is required for providing scientists with the flexibility to reflect their preferences and expert judgment using multi-way visual comparison tasks. In this section, we describe how MyriadCues helps us satisfy the task and design requirements using comparative visual cues and user interaction.

6.1 Reconfigurable small multiples

The core design element of MyriadCues is a set of small multiples (Figure 5c,d) that scientists can configure based on the tasks they want to perform. They can use magnitude plots (explained below) and slope plots in combination or separately. The small multiples (each representing a metric) are laid out sequentially. While we considered a force-based layout [27], the potential visual complexity in interpreting the relative distances and their changes (on weight adjustment) led us to implement the sequential layout.

Magnitude Plots: We use a combination of bar charts and dot plots and collectively call them magnitude plots. As shown in Figure 4, the height of a bar represents the weighted average of a particular metric across all variables for a given model. The bar chart for a reference metric is always rank-ordered (from left to right). Here, the top 10 models are shown with respect to the Bayesian likelihood metric. For additional metrics like the Brier skill score, the same ten models are displayed with their rank with respect to the Brier skill score indicated on the top of the bar. This encoding choice helps link the reference rankings with other rankings and readily observe magnitude differences. When weights are changed, a small green or red arrow indicates whether the ranking of each model improved (green) or degraded (red) in response to the change. The dot plots allow selection of models of interest and help in estimating the degree of rank change on adjustment of weights. Each dot plot can be configured by an expert. As shown in Figure 4a, the models can be divided into quartiles based on the minimum and maximum range of a metric. This view is useful to spot how the ranges and the distribution on models are affected by the weighting of metrics for different outputs. An alternative binning method is to choose equal intervals and a rank based ordering. As shown in Figure 4b, this view helps in quickly spotting which models changed positions, as we can see for the pink model for both the metrics. This view is especially useful when most of the models are sensitive to the weight changes and there are a lot of simultaneous position changes.

Slope Plots: The expressiveness and effectiveness of slope plots in communicating small differences can be observed in Figure 5d. We can see that among the top five models (with respect to Bayes factor), the red and the purple model show significant differences in rankings with respect to the brier skill score. For example, by tracing the lines across the variables, it is immediately obvious that the PSL variable contributes to the poor performance of the red model while the SWCF variable contributes to the poor performance of the purple model. Using this visual cue, an expert can quickly and accurately detect small differences in fidelity. Slope plots also help clearly express a key discrepancy scientists are interested in. As part of T3 (understanding model-output dissimilarity), they would like to quickly find model pairs with similar average fidelity level but with low correlation with respect to individual fidelity levels for specific outputs. The connectedness among models using polylines makes it very quick to spot these discrepancies visually, by identifying line crossings that indicate a lack of correlation. We also let experts interactively select a model and query the system to find the most similar or dissimilar model with respect to a given metric.

6.2 User Interaction

In addition to providing optimal encodings for reducing comparison complexity, MyriadCues incorporates a number of interactive capabilities for experts to further control their search space and exercise their

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2934540, IEEE Transactions on Visualization and Computer Graphics

expert judgment. We describe the key interactive capabilities below.

Selecting a reference metric: At the outset, we made a key design decision. We first let scientists choose a **reference metric** (Figure 5d) for facilitating the $1:k$ comparison task (T5) as part of $T5$ (*Understanding metric-metric dissimilarity*). Without a reference, $k:K$ comparisons (many-to-many comparisons among metrics) would have been needed which would make the search space for disagreement or stability patterns too complex to navigate. Reference selection is also consistent with the scientists' need to compare a set of alternative metrics with a preferred metric.

Selecting and highlighting models: Scientists can select a set of models from any of the small multiples of metrics, but they are rank ordered (T1) based on the reference. We constrain the number of models in the selection set to be 10, as we can use the most distinct categorical colors [13] for these 10 models and avoid color mixing, which can not only occlude the slope plots, but also prevent multi-way comparison among the magnitude and slope plots (T2, T3). These 10 could be the top 10, bottom 10 or any random set of 10 models selected through interaction with the dot plot. Coloring therefore serve as a way to link the same model across the small multiples (Figure 5c, d).

Dynamic ranking: As shown in Figure 5a,b, MyriadCues provides a set of filters for sub-setting models, outputs, or metrics, based on experts' preferences for all the tasks. Once a set of models are filtered, the rankings are automatically updated. The rank criteria can also be changed to using a weighted mean, median, or variance among outputs.

Flexible reordering: For controlling the ordering of outputs in slope plots and the layout of the small multiples, experts can use a number of reordering options. For slope plots, one can order the plots from top to bottom in increasing or decreasing order of mean or variance across all outputs (T3). For small multiples, experts can choose a layout based on *most similar first* or *most dissimilar first* (T5), with respect to a reference metric. As shown in Figure 5c, the $bayes_factor$ and the $bayesian_likelihood$ metric exhibit the most similar rankings.

Finding the most dissimilar models: Within the top ten or bottom ten models, scientists are often interested in finding out which models are most dissimilar with respect to all the output variables and if that dissimilarity changes by varying weights of output variables (T3, T5). In that case, scientists can select a model and then MyriadCues will automatically display the most dissimilar model within the top ten or bottom ten. This lets scientists quickly compare this dissimilarity across all other metrics and assess consistency.

Exploring stability: For visualizing stability of models (T4, T6) from different perspectives, scientists can explore multiple options. They can fix the set of selected models (by assigning colors to a chosen set), adjust weights of outputs and observe how these selected set of models respond to these changes. Using another option, they can also choose to dynamically view which set of models fall within the top or bottom ten by not fixing the colors. This is accomplished using the dot plots, where drastic position changes of models provide a cue for instability. In this case ,colors get assigned to the top or bottom 10 set, the membership of which is a function of weight changes. In our experience of deploying MyriadCues, while scientists appreciated the flexibility to choose these perspectives, they were mostly interested in the first case, where they could fix a set of models and observe their stability.

7 Expert Case Studies and Subjective Feedback

Evaluating MyriadCues was challenging as there is little ground truth data and consensus about which metrics effectively capture model fidelity, implying that focusing on questions around "finding the best or worst model" would have led to a high degree of individual differences among experts. In the light of these challenges, we decided to assess the subjective user experience [25] of experts by centering our evaluation around two factors: if the experts can develop confident judgments about model fidelity using the tool and if they perceive the tool as useful enough to be adopted in their own analysis routine. We focused on the two original goals for this design study: whether multi-way visual comparison can lead to a better understanding of the consistency and robustness of model fidelity levels and the associated factors. For the following case studies, model simulations were taken

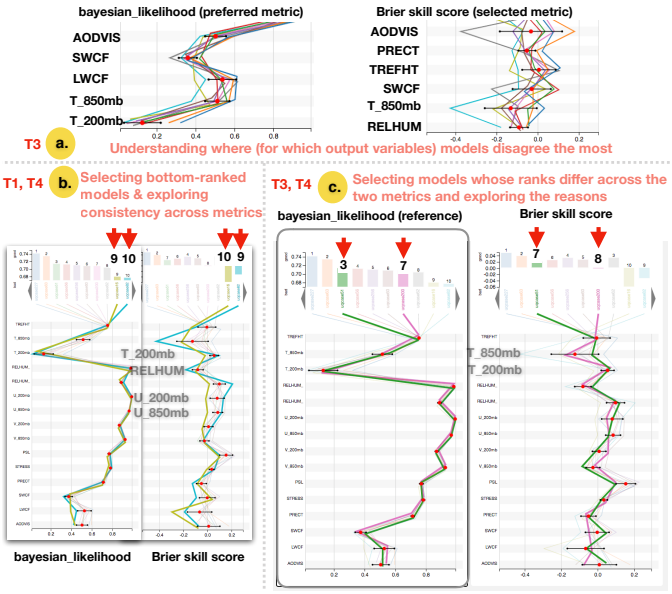

Fig. 6. **Case study for assessing consistency among fidelity metrics.** The different stages include: understanding what causes disagreement among models (a), and inspecting cases where metrics disagree about the fidelity levels across multiple output variables (b, c).

from a 256-member perturbed parameter ensemble of simulations in the Community Atmosphere Model [33], where 16 parameters controlling the emissions of aerosol particles and their interactions with clouds were systematically perturbed. Five-year simulations were performed in an atmosphere-only model with prescribed sea surface temperatures (SSTs) [38]. We initially conducted a three-hour long session with our collaborator for exploring different usage scenarios using MyriadCues. Next, she used MyriadCues by herself over the next few days and derived several case studies, two of which we report below.

7.1 Assessing consistency among alternative metrics

The purpose of this case study was two-fold: i) our collaborator wanted to experience how MyriadCues can fit into her analytical routine, and ii) she wanted to compare her preferred statistical metric to other alternatives and assess the consistency of model fidelity with respect to those metrics. To satisfy these goals, she used MyriadCues in conjunction with the AMWG package, a climate model diagnostics package developed by the Community Atmosphere Model's Atmosphere Model Working Group (AMWG). The AMWG package helps visualize geographical distributions of output and observation data, providing complementary information for expert judgment of fidelity. Our collaborator sub-selected a set of 10 models based on the AMWG maps and in MyriadCues she started her analysis by selecting *Bayesian likelihood score*, her preferred metric, as the reference. By comparing small multiples of magnitude plots (Figure 4a), she found that the Brier skill score provided somewhat similar rankings, but, as observed, there were small differences, because of which she wanted to compare these metrics in greater detail. By looking at relative dissimilarities among variables using the slope plot (Figure 6a), she found that the Bayesian likelihood score exhibited less disagreement among the models (depicted by similar slopes) for different output variables, than exhibited by the Brier skill score (depicted by dissimilar slopes). However, the average fidelity scores for Brier skill score were higher and more consistent with each other than the Bayesian likelihood score (as demonstrated by the more abrupt variation in slopes in Figure 6a, where all models seemed to have low fidelity for `T200mb` and `SWCF` variables. These patterns were surprising but not conclusive. To investigate more, our collaborator selected the top-ranked models. But there was little disagreement between the metrics. Next, she selected the bottom ranked models. or `uqcase16` (light green) and `uq_case82` (light blue), the Brier Skill Score and the Bayesian Likelihood score gave different relative rankings (Figure 6b). The slope plots indicated that this difference

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2934540, IEEE Transactions on Visualization and Computer Graphics

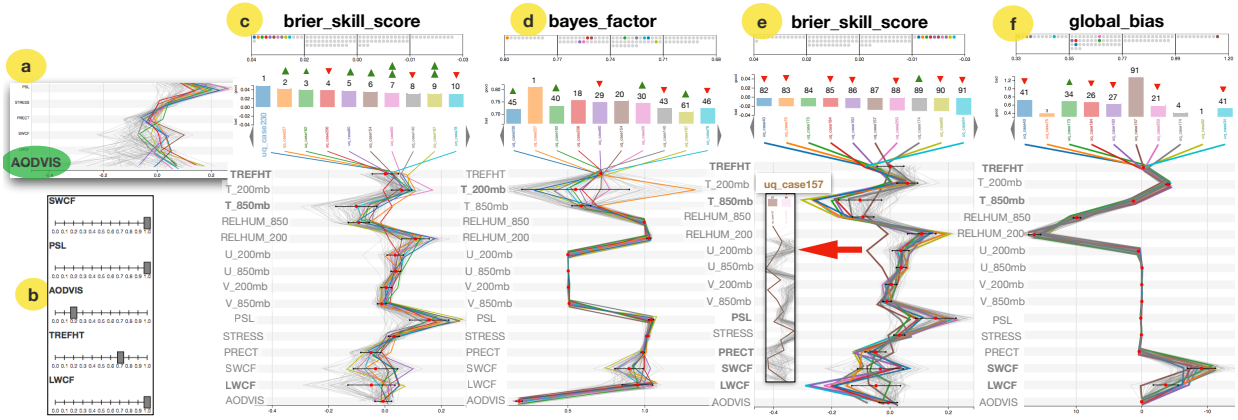

Fig. 7. **Case study for assessing robustness of model rankings.** The different stages include: spotting of model outputs with high variability (a) (T3) that resulted in the adjustment of variable weights (b), followed by comparison across multiple metrics to look at disagreement about fidelity values across different variables (c,d,e,f) (T4, T6).

was attributable to `uqcase82` performing better on Brier Skill Score on the variables `RELHUM200mb`, `U$850$mb`, `U$_200$mb`, and `T200mb`, as compared to the their performance on the the Bayesian Likelihood score. With respect to `T200mb`, both models performed poorly, therefore one could not conclude which model had a better fidelity. However, with respect to `U200mb`, the AMWG maps verified that `uqcase82` had a better fidelity than `uqcase16`, which was consistent with the Brier skill score. Next, our collaborator selected the `uqcase51` (green) and `uqcase200` (magenta) models which exhibited different relative ranks with respect to both the metrics (Figure 6c). Examining the AMWG maps of these fields, she found that the much lower fidelity of uqcase200 for the variable `T$850$mb` (according to the Brier Skill score) was associated with a high-latitude cold bias, high-latitude high pressure bias, and an overly-strong jet stream, particularly in the Northern Hemisphere. This led our expert to assess `uqcase200` as lower in overall fidelity, which was more consistent with the ranking by the mean Brier Skill Score. A similar conclusion was derived in case of the `uqcase51` model as well. As a result of these evaluations, she concluded that the Brier Skill Score seemed to be more consistent with the overall rankings that she would have assigned to these models, and concluded that she would prefer the Brier Skill Score over the others for model ranking, which was a change from her initial preference for the Bayesian likelihood score. Finally, she assigned weights to each of the model variables, to explore how robust the rankings would be to changes in variable weights. Most models did not change their ranking after assigning their weights, when using the Bayesian likelihood, or the Brier Skill Score. Overall, she concluded that *"the Brier Skill Score was most consistent with the rankings she would likely have assigned based on the diagnostics and metrics from the AMWG package"*. She also felt the need to diagnose more carefully, the computation of the fidelity levels using the Bayesian likelihood score. This exercise was a satisfactory experience for our collaborator as she could directly realize the value of MyriadCues in re-assessing her hypothesis and preferences. She commented: "*We currently have only limited tools for performing multi-model comparisons, and none of our current tools employ interactive features in the visual display of data. While I initially was unsure whether interactive features would add value to the tool, many of the interactions are very helpful because they allow users to quickly and intuitively reduce the clutter, focus on specific portions of the data (e.g. by simultaneously comparing models across all the metrics), or quickly access additional details about fidelity. This allows users to explore the data much more reliably and efficiently than would be possible using typical methods where a script is written to generate a static plot, and the script needs to be updated and run again to generate any new display of the data.*"

7.2 Assessing robustness of model rankings

The purpose of this case study was to assess the efficacy of MyriadCues as a standalone tool, by using it to reduce the complexity of possible comparisons to a few important ones and derive hypothesis about the robustness of the model rankings. For this case study, 100 models, 15 output variables, and 15 variables were selected by our collaborator. Next, our collaborator selected the Brier skill score metric and examined the rank and slope plots to understand which factors contributed to models achieving a high or low ranking on this metric, and how sensitive these rankings were to the weighting of individual variables. The variable `AODVIS` (Aerosol optical depth) stood out in displaying a large variance between models in their fidelity, as measured by the standard deviation (Figure 7a). `AODVIS` however, is less important to evaluating climate model behavior than other variables in this collection; aerosols are of less physical importance to the climate system than variables such as precipitation, temperature, and clouds, and `AODVIS` is an imperfect measure of aerosol amount in the atmosphere. Therefore she decreased the weight of `AODVIS`; after this change, while `uqcase230` was still the highest-ranked model, and eight of the top ten highest-ranked models were still the same, the rankings of other models changed. Next, she assigned new weights to some of the other variables to reflect their approximate relative physical importance, and reduced weights of variables that likely contain redundant information (e.g., `TREFHT` and `T850mb`; the temperature at 10 m about the Earth's surface and temperature at 850 hPa, in the lower troposphere)(Figure 7b).By examining the highlighted lines in the slope plots, she identified patterns among the most highly ranked models (Figure 7c) that were responsible for their superior performance on the Brier skill score metric. Next, by look at the small multiples of the magnitude plots for all metrics, our collaborators found the Bayes factor metric to exhibit a uniform distribution of fidelity scores, suggesting high information content. She compared the results between the Brier Skill Scores and the Bayes factor metric (Figure 7d). Comparing the slope plots revealed that the variable `T$200$mb` exhibits far greater variability in the Bayes factor than in the Brier skill score. Our expert speculated that this different behavior might arise from differences in how the two scores are constructed, since the Bayes factor discounts model-observation discrepancies below a pre-defined threshold. Among the ten highest-ranked models, all of them consistently performed above average on the weighted average metrics for the following variables: `LWCF`, `TREFHT`, `T850mb`. Most also performed above average on `RELHUM850mb`, `PRECT`, and `PSL`. Some models compensated for poor performance on one variable by performing well on another variable, for instance, `uqcase206` (green) performed poorly on `T200mb` and `RELHUM200mb`, but better on `T850mb` and `RELHUM850mb` than most other models. The ten lowest-performing models (Figure 7e), by contrast, mostly performed below average on the variables `LWCF`, `TREFHT`, `T850mb`, `RELHUM850mb`, `PRECT`, and `PSL`. Interestingly, `SWCF` did not appear to be a strong predictor of overall model fidelity, although the highest-performing model, `uqcase230`, performed higher than average on this variable. The Brier skill score of these models was particularly poor on the variable `LWCF`. To better understand what caused this low ranking, our collaborator examined the global bias metric (Figure 7f), which provides a information on a complementary aspect of model fidelity. Most of the lowest-performing models exhib-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2019.2934540, IEEE Transactions on Visualization and Computer Graphics

ited a strong negative bias in LWCF, meaning that clouds did not produce enough warming through their long-wave forcing effect, and these models also mostly were colder than other models in the lower and upper troposphere (TREFHT, T850mb, T200mb). An exception to this pattern was uqcase157, which performed similarly poorly on the mean Brier skill score, but exhibited a very different pattern of behavior across the metrics for the individual variables, as shown. The normalized view of the slopeplots revealed a very different behavior for this model on the pattern of global biases; the model had almost no bias in LWCF while almost all other models had a negative bias; uqcase157 was also warmer than most other models, suggesting a strong inverse correlation between mean LWCF and temperature in this simulation ensemble. In summary, through this exercise, our collaborator was able to reason about the outputs contributing to good or poor overall model fidelity, and iteratively flag models and variables for progressively investigating the stability of model fidelity rankings in response to the weighting of different physical variables. Our collaborator's appreciation for the flexibility of the tool is reflected in this comment : *"By enabling us to supply our own weights, the tool flexibly allows us to update the influence of different aspects of model fidelity (e.g., fidelity of different physical variables), incorporating our physical understanding of which aspects of system behavior are most important, and immediately receive feedback on how this influences overall model ranking.*"

7.3 Expert Interview and Feedback

Besides the case studies, we validated the utility of MyriadCues by recording the subjective feedback of scientists. To this end, we used questionnaires and in-person, structured interviews for gaining an understanding of how scientists benefit from using MyriadCues to solve model fidelity probems. We recruited two senior climate scientists, both of them work as senior climate scientists in a national laboratory and have an average of 17 years of research experience between them and were not familiar with the tool. We used the data set from the first case study and recorded their feedback in a 40 minute long session.

The two interview sessions were structured as follows. We first gave them a brief (5 miniutes) introduction to the functionalities of the tool. Next, we asked them to use the tool for reasoning about good and bad models. We instructed them not to look for a correct answer, as we did not have any ground truth data. Instead, we suggested that they should assess if and how the tool can help them in making reliable judgments about choice of metrics and disagreement among model rankings. We encouraged think-aloud protocol during the sessions. They explored the analysis scenarios for about 30 minutes and then filled out the questionnairre. We observed that the participants had different starting points in their analysis process, each of them started with their own preferrend metric and expressed surprise at some of the disagreements about model fidelity across other metrics. They felt that the tool provided them with enough insights to develop alternative hypotheses about model fidelity. From the responses given by our participants and comments given during the interview, we group the feedback into the following categories. i) **Effectiveness.** Our first participant commented that: "*This tools gives you a comprehensive picture across many variables and you can do much more than looking at one or two numbers*". He also observed that taking many factors into account adds to the task complexity and it is essential that we keep the interactions as simple as possible going forward. Our second participant commented that *"this is very nicely designed"*, and the main advantage is that this tool *"integrates many different metrics"* and let him observe model behavior going beyond a few preferred metrics. The feature in the tool they most liked was the pairing of the magnitude plot with the slope plot that gave simultaneous cues about model ranking and contributions of the variables towards the weighted scores. ii) **Flexibility.** Our first participant particularly appreciated the level of flexibility in his analysis that the tool allows:*"the good thing is that this is so flexible in choosing any analysis scenario"* and adapt the selections accordingly for detecting small differences. Our second participant, while being positive about the interactions through which he could assign weights and observe the changes, stressed on the need to have better support for automatically finding or highlighting variables that are correlated so that the weights could be adjusted using that information. iii) **Advantages over the state of the art** Both our participants observed that this tool will help speed up the analysis process as existing techniques mostly involve manual scripting and allow them look at few variables and model at a time. They further observed that this tool can be a nice complement to the existing diagnostic packages that lets them visualize spatial patterns. Our second participant also mentioned that this tool can be very useful in cases where "you need to track model errors over time" and that *"there is no tool for that right now"*. He commented that the small multiples can be easily configured for showing different temporal instances for a particular metric. iv) **Shortcomings.** Both experts observed that while the tool is immediately usable, the tool also has a great potential to solve an open problem in climate science: *how to choose parameters with the knowledge of model fidelity?* This tool currently does not support parameter analysis and that is our planned next step. Our first participant also commented that he should be given the option *"to choose from a list of many different variables"* and that the tool should support automatically supporting NetCDF files. v) **Potential for adoption.** Both experts were enthusiastic about using the tool as part of their own analysis routine and the lack of prior familiarity did not seem to be a barrier. One of them commented that *"you should release the software as soon as possible"* for other scientists to benefit from it. They observed that the tool has a great potential for adoption by the broader climate science community and we should engage in more outreach activities to build awareness about this research.

8 Conclusion and Future Work

The design study reported in this paper is a significant first step towards developing a viable solution for addressing the long-standing need of greater transparency in multi-criteria model fidelity analysis. Through our case studies, we demonstrated that comparative visual cues were effective and MyriadCues was able to inspire confidence in climate scientists, both as a complementary and a standalone tool for performing complex, multi-way comparison tasks. Feedback from the broader community has demonstrated a strong potential for the adoption of this tool by modeling groups. These contributions should be understood in the context of the state of the art in climate model fidelity analysis where interactive visualizations are rarely used and data tables summarizing metric scores for model outputs are often preferred by climate scientists over visualizations for building their hypotheses. Currently, besides integrating parametric analysis methods, we are working on addressing two shortcomings of MyriadCues. First, MyriadCues does not capture analytical provenance. This is important, as scientists want to keep track of different versions of model simulations and their corresponding diagnostics. To this end, we will be developing a provenance-enabled backend that helps build a shared knowledge base about model outcomes. Second, we are working on making the design and implementation of MyriadCues even more scalable, to support the simultaneous analysis of upwards of 500 simulation models. We are also engaging in outreach activities beyond the climate science community: our solution for multi-criteria decision analysis is equally applicable in data science scenarios, where there is a growing need for going beyond traditional accuracy metrics for machine learning models and incorporate metrics about bias, fairness, interpretability, etc. We will apply MyriadCues in these scenarios and thereby establish a domain-agnostic, comparative visualization approach for tackling these cutting edge model diagnostics problems.

9 Acknowledgment

This work was partially supported by the Laboratory Directed Research and Development Program at PNNL, a multi-program national laboratory operated by Battelle. We would like to thank Feng Wang for developing the initial prototypes, and Phil Rasch, Yun Qian, and Po-Lun Ma for their feedback about MyriadCues. We are also grateful to the anonymous reviewers for their constructive comments, which helped refine the discussions in the paper.

REFERENCES

[1] E. Alexander and M. Gleicher. Task-driven comparison of topic models. *IEEE transactions on visualization and computer graphics*, 22(1):320–329, 2015.

[2] J. Bertin. *Semiology of graphics: diagrams, networks, maps*. University of Wisconsin press, 1983.

[3] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: an overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, 2011.

[4] M. Booshehrian, T. Möller, R. M. Peterman, and T. Munzner. Vismon: Facilitating analysis of trade-offs, uncertainty, and sensitivity in fisheries management decision making. In *Computer Graphics Forum*, vol. 31, pp. 1235–1244. Wiley Online Library, 2012.

[5] S. M. Burrows, A. Dasgupta, S. Reehl, L. Bramer, P.-L. Ma, P. J. Rasch, and Y. Qian. Characterizing the relative importance assigned to physical variables by climate scientists when assessing atmospheric climate model fidelity. *Advances in Atmospheric Sciences*, 35(9):1101–1113, 2018.

[6] A. Dasgupta, S. Burrows, K. Han, and P. J. Rasch. Empirical analysis of the subjective impressions and objective measures of domain scientists' visual analytic judgments. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 1193–1204. ACM, 2017.

[7] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *2016 IEEE symposium on security and privacy (SP)*, pp. 598–617. IEEE, 2016.

[8] G. Flato, J. Marotzke, B. Abiodun, P. Braconnot, S. C. Chou, W. Collins, P. Cox, F. Driouech, S. Emori, V. Eyring, et al. Evaluation of climate models. In *Climate change 2013: the physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, pp. 741–866. Cambridge University Press, 2014.

[9] P. J. Gleckler, K. E. Taylor, and C. Doutriaux. Performance metrics for climate models. *Journal of Geophysical Research: Atmospheres*, 113(D6), 2008.

[10] M. Gleicher. Considerations for visualizing comparison. *IEEE Transactions on Visualization and Computer Graphics*, 2017 (in publication).

[11] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Information Visualization*, 10(4):289–309, 2011.

[12] S. Greco, J. Figueira, and M. Ehrgott. *Multiple criteria decision analysis*. Springer, 2016.

[13] 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.

[14] C. G. Healey, K. S. Booth, and J. T. Enns. High-speed visual estimation using preattentive processing. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 3(2):107–135, 1996.

[15] D. Holten and J. J. Van Wijk. Visual comparison of hierarchically organized data. In *Computer Graphics Forum*, vol. 27, pp. 759–766. Wiley Online Library, 2008.

[16] F. Hourdin, T. Mauritsen, A. Gettelman, J.-C. Golaz, V. Balaji, Q. Duan, D. Folini, D. Ji, D. Klocke, Y. Qian, et al. The art and science of climate model tuning. *Bulletin of the American Meteorological Society*, 98(3):589–602, 2017.

[17] A. Inselberg and B. Dimsdale. Parallel coordinates: A tool for visualizing multi-dimensional geometry. In *IEEE Visualization*, pp. 361–378. IEEE CS Press, 1990.

[18] P. Kampstra et al. Beanplot: A boxplot alternative for visual comparison of distributions. 2008.

[19] D. Katz. *Gestalt psychology, its nature and significance*. Greenwood Pub Group, 1979.

[20] J. Kehrer and H. Hauser. Visualization and visual analysis of multifaceted scientific data: A survey. *IEEE transactions on visualization and computer graphics*, 19(3):495–513, 2013.

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

[22] J. Kehrer, H. Piringer, W. Berger, and M. E. Gröller. A model for structure-based comparison of many categories in small-multiple displays. *IEEE transactions on visualization and computer graphics*, 19(12):2287–2296, 2013.

[23] G. A. Kiker, T. S. Bridges, A. Varghese, T. P. Seager, and I. Linkov. Application of multicriteria decision analysis in environmental decision making. *Integrated environmental assessment and management*, 1(2):95–108, 2005.

[24] P. Köthur, M. Sips, H. Dobslaw, and D. Dransch. Visual analytics for comparison of ocean model output with reference data: Detecting and analyzing geophysical processes using clustering ensembles. *IEEE transactions on visualization and computer graphics*, 20(12):1893–1902, 2014.

[25] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012.

[26] Z. C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[27] X. Liu, Y. Hu, S. North, and H.-W. Shen. Correlatedmultiples: Spatially coherent small multiples with constrained multi-dimensional scaling. In *Computer Graphics Forum*. Wiley Online Library, 2015.

[28] J. Mackinlay. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, 5(2):110–141, 1986.

[29] N. M. Mangan, J. N. Kutz, S. L. Brunton, and J. L. Proctor. Model selection for dynamical systems via sparse regression and information criteria. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2204):20170009, 2017.

[30] M. Meyer, T. Munzner, and H. Pfister. Mizbee: a multiscale synteny browser. *IEEE transactions on visualization and computer graphics*, 15(6):897–904, 2009.

[31] T. Munzner. A nested process model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.

[32] T. Munzner, F. Guimbretière, S. Tasiran, L. Zhang, and Y. Zhou. Treejuxtaposer: scalable tree comparison using focus+ context with guaranteed visibility. In *ACM Transactions on Graphics (TOG)*, vol. 22, pp. 453–462. ACM, 2003.

[33] 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, et al. Description of the ncar community atmosphere model (cam 5.0). *NCAR Tech. Note NCAR/TN-486+ STR*, 2010.

[34] S. Pajer, M. Streit, T. Torsney-Weir, F. Spechtenhauser, T. Möller, and H. Piringer. Weightlifter: Visual weight space exploration for multi-criteria decision making. *IEEE transactions on visualization and computer graphics*, 23(1):611–620, 2016.

[35] W. S. Parker. Ii—confirmation and adequacy-for-purpose in climate modelling. In *Aristotelian Society Supplementary Volume*, vol. 83, pp. 233–249. Wiley Online Library, 2009.

[36] J. Poco, A. Dasgupta, Y. Wei, W. Hargrove, C. Schwalm, R. Cook, E. Bertini, and C. Silva. SimilarityExplorer: A visual inter-comparison tool for multifaceted climate data. *Computer Graphics Forum*, 33(3):341–350, 2014.

[37] 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. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1923–1932, 2014.

[38] Y. Qian, H. Yan, Z. Hou, G. Johannesson, S. Klein, D. Lucas, R. Neale, P. Rasch, L. Swiler, J. Tannahill, et al. Parametric sensitivity analysis of precipitation at global and local scales in the community atmosphere model cam5. *Journal of Advances in Modeling Earth Systems*, 7(2):382–411, 2015.

[39] M. Sedlmair, C. Heinzl, S. Bruckner, H. Piringer, and T. Möller. Visual parameter space analysis: A conceptual framework. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2161–2170, 2014.

[40] S. Simon, S. Mittelstädt, D. A. Keim, and M. Sedlmair. Bridging the gap of domain and visualization experts with a liaison. In *Eurographics Conference on Visualization (EuroVis Short Paper)*. The Eurographics Association, 2015.

[41] D. J. Simons and R. A. Rensink. Change blindness: Past, present, and future. *Trends in cognitive sciences*, 9(1):16–20, 2005.

[42] A. Srinivasan, M. Brehmer, B. Lee, and S. M. Drucker. What's the difference?: Evaluating variations of multi-series bar charts for visual comparison tasks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 304. ACM, 2018.

[43] F. J. Tapiador, R. Roca, A. Del Genio, B. Dewitte, W. Petersen, and F. Zhang. Is precipitation a good metric for model performance? *Bulletin of the American Meteorological Society*, 100(2):223–233, 2019.

[44] K. E. Taylor. Summarizing multiple aspects of model performance

in a single diagram. *Journal of Geophysical Research: Atmospheres*, 106(D7):7183–7192, 2001.

[45] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 2nd ed., 2001.

[46] T. von Landesberger. Insights by visual comparison: The state and challenges. *IEEE computer graphics and applications*, 38(3):140–148, 2018.

[47] J. Wang, X. Liu, H.-W. Shen, and G. Lin. Multi-resolution climate ensemble parameter analysis with nested parallel coordinates plots. *IEEE transactions on visualization and computer graphics*, 23(1):81–90, 2017.

[48] J. Waser, H. Ribicic, R. Fuchs, C. Hirsch, B. Schindler, G. Bloschl, and E. Groller. Nodes on ropes: A comprehensive data and control flow for steering ensemble simulations. *IEEE transactions on visualization and computer graphics*, 17(12):1872–1881, 2011.