The Effect of Color Scales on Climate Scientists' Objective and Subjective Performance in Spatial Data Analysis Tasks

Aritra Dasgupta, Jorge Poco, Bernice Rogowitz, Kyungsik Han, Enrico Bertini, and Cláudio T. Silva, *Fellow, IEEE*

Abstract—Geographical maps encoded with rainbow color scales are widely used by climate scientists. Despite a plethora of evidence from the visualization and vision sciences literature about the shortcomings of the rainbow color scale, they continue to be preferred over perceptually optimal alternatives. To study and analyze this mismatch between theory and practice, we present a web-based user study that compares the effect of color scales on performance accuracy for climate-modeling tasks. In this study, we used pairs of continuous geographical maps generated using climatological metrics for quantifying pairwise magnitude difference and spatial similarity. For each pair of maps, 39 scientist-observers judged: i) the magnitude of their difference, ii) their degree of spatial similarity, and iii) the region of greatest dissimilarity between them. Besides the rainbow color scale, two other continuous color scales were chosen such that all three of them covaried two dimensions (luminance monotonicity and hue banding), hypothesized to have an impact on task performance. We also analyzed subjective performance measures, such as user confidence, perceived accuracy, preference, and familiarity in using the different color scales. We found that monotonic luminance scales produced significantly more accurate judgments of magnitude difference but were not superior in spatial comparison tasks, and that hue banding had differential effects based on the task and conditions. Scientists expressed the highest preference and perceived confidence and accuracy with the rainbow, despite its poor performance on the magnitude comparison tasks. We also report on interesting interactions among stimulus conditions, tasks, and color scales, that lead to open research questions.

Index Terms—visualization, color maps, rainbow color map, user study

1 Introduction

There is often a mismatch between visualization research and visualization practice in scientific domains. In this paper, we focus on one of the most popular, and often debated, mismatches: the use of the rainbow color scale as a means of communication and analysis of scientific data. In particular, we focus on the use of color scales in climate science, where they play a pivotal role. Climate being an inherently geographical/spatial (and temporal) phenomenon, scientists often produce maps to convey information about how measures of interest distribute spatially across the globe or other regions of interest.

Previous qualitative studies have shown that many climate scientists are in disagreement with, or unaware of, the efficacy of perceptually corrected color scales [16], and prefer to use the rainbow color scale as the de facto standard. This paper is an attempt to analyze and explain potential reasons for this mismatch. We describe the results of a web-based user experiment that studies how different color maps affect performance on a selected set of scientifically-motivated tasks. The study is the result of a long-standing collaboration between the authors and a group of experienced climate scientists who helped us understand the specific set of spatial data analysis tasks performed using continuous geographical maps. Accordingly, we selected stimuli that were carefully designed based on climatological metrics that quantify differences in magnitude and spatial distribution between pairs of maps.

Why conduct another user study for evaluating color scales in practice? The study was motivated by two observations. First, although color scales have been the subject of extensive visualization research, there are very few empirical studies of color scales, on ecologically valid tasks with domain experts and real-world data [5], [42]. Even with some recent studies and theories [25], [35], we lack an understanding of how color scales affect the objective performance of experts in spatial data analysis tasks involving continuous geographical maps. Second, previous research shows that familiarity and experience with an analysis medium influences its use and preference [45]. This web-based study investigates how familiarity with the rainbow color scale influences performance accuracy, experts' subjective impressions (e.g., perceived accuracy, preference, etc.), and the relationships between objective and subjective measures of performance. Studying these factors with a group of highly skilled domain experts, under real-world conditions, can shed light on the use and adoption of visualization best practices in real-world domains.

In close collaboration with climate scientists, for the

- *A. Dasgupta is with New Jersey Institute of Technology. E-mail: aritra.dasgupta@njit.edu*
- *J. Poco is with Fundação Getulio Vargas and Universidad Católica San Pablo. E-mail: jorge.poco@fgv.br*
- *B. Rogowitz is with Visual Perspectives Research and Consulting E-mail: bernice.e.rogowitz@gmail.com*
- *K. Han is with Ajou University. E-mail:kyungsikhan@ajou.ac.kr*
- *E. Bertini, and C. Silva are with New York University. E-mail: {enrico.bertini, csilva}@nyu.edu*

study, we carefully selected three spatial data analysis tasks, stimuli generated by selecting maps that differed in their magnitude and spatial distributions, and color scales that co-varied the perceptual dimensions of hue and luminance. Our three main contributions in this paper are the following: i) Domain and problem characterization for understanding climate scientists' tasks using color-coded maps that encode continuous variables, ii) Design of a user study using those tasks and alternative color scales that address the specific tasks of climate scientists but can be generalized for related spatial data analysis tasks in other domains, and iii) Analysis of scientists' objective performance and their subjective impressions about their perceived accuracy, confidence, and preference of a color scale.

2 Background and Related Work

In this section, we first provide a short introduction to the problem of designing perceptually effective color scales. Instead of using the term color "map" to describe the range of colors mapped onto the range of a scalar variable, we adopt the term color "scale", reserving "map" for signifying geographical maps. While describing the problem in full detail is beyond the scope of the paper, the introduction is meant to help the reader familiarize with the problem. We then briefly describe related work on color scales design and studies conducted to evaluate their effectiveness in the context of spatial data analysis tasks.

2.1 Designing Perceptually Effective Color Scales for Scalar Fields

The problem faced by color scale designers involves defining a principled way of mapping data values to colors that communicate relevant data characteristics effectively and faithfully. In this work, we focus exclusively on color scale design for 2D scalar fields: spatial visualizations in which a numerical value, sampled across a 2D region (typically geographical), is represented using color. This kind of representation is very common in scientific and engineering disciplines, since natural and computed phenomena can often be described as numeric samples over a spatial region..

Color scale design typically involves the use of a particular color specification space (*RGB, HSV, CIE Lab, Munsell*, etc.) and the definition of a mapping function, which determines, for a given range of numeric values, the color each corresponds to. Human color perception has been studied extensively in the vision science literature [21] and the literature has been reviewed recently in the visualization community [2]. In the following, using the same convention used by Munzner [32], we refer to a generic three-dimensional space defined by the following three perceptual channels: *Hue*, the color name, *luminance*, which represents the brightness or value of the color, and *saturation*, which characterizes the vividness. Hue, being perceived categorically, may not be easily ordered, perceptually, and is more suited to the representation of categorical information, and in the segmentation of data points [6], [39].

In our experiments, we represent the perceptual dimension of perceived luminance in terms of $L*$, which is the dimension representing perceived luminance in the LAB color space. There are many color spaces in use in visualization that have a color space dimension for this perceptual dimension. The ones derived from RGB color space, for example, HSV, or HSL, are not perceptually uniform, in the sense that equal steps in V or L do not correspond to equal perceptual steps. Since output devices in visualization, however, are not typically calibrated, there is considerable uncertainty about the actual L* value presented to the observer.

Thus, we use the term "luminance monotonicity" to capture the idea that perceived increments in luminance should be at least monotonic with increase in value. Using many different color scales created in many different color spaces, Rogowitz et al. [38] showed that all color scales with a monotonic luminance component were able to effectively represent the magnitude of spatial information, and that most of the variance was carried by the magnitude of $L*$. Rogowitz and Treinish [39], [40] observed that because the luminance system has higher spatial frequency sensitivity than the opponent color system [31], color scales designed to represent the magnitude of fine resolution detail should contain a monotonic luminance component.

2.2 Examples of Color Scales and Their Properties

Visualization researchers have criticized the rainbow color scale because of potential misrepresentation of the data, predominantly owing to its non-monotonically varying luminance [6], [39]. One of the reasons why climate scientists prefer the rainbow color scale is the transition across multiple hues (hue banding) leading to perceived fine-grained representation of the data. However, in the rainbow color scale, the perceptual transition between hues is not uniform and therefore introduces bands and artifacts which can affect perception of the data. Color scales like the "sequential" color scale from Brewer's Colorbrewer library correct for luminance monotonicity [18], and they may also vary in hue and saturation. Brewer also proposed the "diverging" color scale has also been proposed for scalar data. The diverging color scale has a saturated and low-luminance hue component transitioning to another by passing through an unsaturated, often higher-luminance, value in the middle. For scientific data visualization, Moreland [29] has developed a version of the diverging color scale, which has recently been accepted as the default in ParaView [23]. Rogowitz and Kalvin [37], found that luminance contrast, independent of hue and saturation, drove the effectiveness with which a color scale represented the face. Based on this research, Kindlmann et al. [22] developed a luminance-matching technique, which could be used to create color scales that contained a range of hues, with monotonically varying luminance. Due to the banding effect of multiple hues, it was posited that such color scales could both effectively carry magnitude information while also providing segmentation information.

Bergman et al. [1] introduced a rule-based system that suggested appropriate color scales based on the data type (ordinal, interval, ratio), spatial frequency, and on the task. Tominski et al. [46] extended these ideas by proposing a task taxonomy and appropriate color scales comparison, localization, and data value identification tasks.

We used a sequential color scale from the ColorBrewer library, and the color scale developed by Kindlmann et

al. [22] (henceforth referred to as the Kindlmann scale) for our experiments. We study how continuous spatial distributions affect visual averaging and comparison using color scales co-varying in hue and luminance.

2.3 Empirical Evaluation of Color Scales

Over the years, many researchers have conducted experiments comparing the effectiveness of different color scales, using both artificial stimuli and real-world conditions. In order to isolate and study specific experimental variables, many empirical studies in this field have relied on using synthetic stimuli. Rogowitz et al. [38] constructed many color scales as trajectories in luminance, saturation and hue, in many color spaces, and measured increment thresholds for detecting Gaussian patches visualized with these different color scales. They found that for monotonic luminance and monotonic saturation scales, the threshold for detecting a change in magnitude was proportional to data value, with luminance color scales providing the most sensitive results. With color scales that varied in hue, much larger increments in data value were needed for detection, and perceived changes in magnitude were not proportional to changes in data magnitude. Ware [49] used artificial stimuli to explore users' ability to read magnitude information from a region on a visualization and map it onto a value on a color scale. To emulate real-world medical imaging situations, Tajima et al. [44] and Levkowitz and Herman [24] used the detection of artificial phantom "blobs" in medical images to reveal advantages of the luminance grayscale over other color scales, including the heated-body scale, which is monotonic in luminance but varies in hue. Recently, Borkin et al. [5] studied visual performance using the rainbow and a diverging color scale proposed by Brewer [9], [18] in a real-world setting. In this task cardiologists identified arterial blockages from color-coded medical images. They found a significant advantage of the diverging color scale over the rainbow. Their findings are consistent with the recent study conducted by Liu and Heer [25] where they found that the rainbow color scale performed the worst in terms of both efficiency and accuracy as compared to singe-hue and other multi-hue color scales.

Spatial data analysis using color-coded maps have been largely evaluated based on the task of identifying discrete regions [9], [10], [34] from choropleth maps. Some previous studies look at the effect of diverging color scale on perception of uncertainty for flood risk assessment [42] and the comparison among encodings based on hue and texture for estimating uncertainty involving wildfires [12]. However, there is very little guidance on how to construct effective color scales for continuous maps, because there are many trade-offs involved in making a design decision about a continuous color scale, which have been recently described by Bujack et al. [11]. In a recent work, Ware et al. [50] devise a new way of quantifying perceptual uniformity of color scales and conduct a Mechanical Turk study for understanding how feature resolution is affected by the design choices. Their findings stress the importance of luminance variation in influencing task performance. Reda et al. [35] conducted a study with continuous color scales for understanding the effect of spatial frequency on color perception based on alternative scales are chosen according to the properties of luminance monotonicity and hue variation. While their design criteria of luminance monotonicity and hue variation are consistent with ours, they base their findings on a set of value retrieval based tasks from a single map. Our study is grounded in ecologically valid tasks and real data used by climate scientists. We focus on visual comparison tasks by juxtaposing *continuous* geographical maps, where scientists make judgments about relative differences in magnitude and identify spatial similarities and dissimilarities. To the best of our knowledge, there exists little empirical evidence in real-world setting evaluating such spatial data analysis tasks based on visual comparison using continuous color scales.

In our study, we also reflect on the relationships between familiarity and perceived levels of accuracy and preference. Our work complements that of Schloss and Palmer who developed metrics for modeling individual preference of color scales [41] and applied ecological valence theory [33] for reasoning about the individual differences in preference levels. We aim to understand if familiarity is a barrier in scientists' acceptance of potentially more effective color scales and if participatory design can help mitigate the effects of familiarity like recent user studies [14], [15] have demonstrated.

3 Domain and Problem Characterization

The findings presented in this paper are a result of a long-standing collaboration between visualization researchers and a group of climate scientists working on the Multi-Scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP) [20], which develops innovative solutions for the comparison of complex climate models. We followed a two-stage process for developing our understanding of the scientists' spatial data analysis goals and methods.

First, we interacted with two direct collaborators, both climate scientists at the Oak Ridge National Laboratory, over a period of six months for collecting examples of maps and their corresponding analysis goals. We followed up our interactions with in-person and online semi-structured interviews, that helped us refine our understanding of these goals. Second, we organized two face-to-face meetings and many remote follow-up meetings with a bigger group of 10 climate scientists to understand in details what specific questions they ask and judgments they make using color-coded geographical maps. During these meetings, we presented examples taken from the scientists' work and asked them to describe what kinds of questions and visual operations they would perform when examining them.

In this section, we first describe the domain-specific spatial data analysis goals and methods, as synthesized from our discussions, and then characterize the problems relevant to color scale selection.

3.1 Types of Spatial Data Analysis Tasks

A common analysis routine performed by climate scientists is using multiple maps for analyzing model outputs. These maps typically represent outputs from different climate models, at different time periods, and allow the scientists

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.2018.2876539, IEEE Transactions on Visualization and Computer Graphics

to compare model behavior. An example of one such output is the Gross Primary Productivity (GPP), which serves as one of the ecosystem health indicators. Scientists generally perform visual comparison tasks through juxtaposition of these maps [17] in a small multiple setting [47]. These visual comparison tasks can be classified into the following types:

Magnitude Comparison: In this task scientists visually estimate the difference in global mean GPP among multiple color-coded geographical maps. In the course of our interactions, visualization researchers pointed out that these are expensive operations, and trivial solution could be to simply calculate the "mean" value that can support the task of quantitative comparison across different maps. However, scientists mentioned that they perform these tasks in a relative context: they use their visual judgment to verify numbers that are computed by a metric and documented in a table, or they visually compare multiple maps of model outputs to a map depicting observation data, whose mean is already known. Webster et al. [51] showed that people can successfully estimate average chromaticity of two hues, and Maule and Franklin [27] showed the human ability to average across several hues and color boundaries for such tasks.

Spatial Distribution Comparison: This class of tasks involves analyzing the shape of global and local spatial distributions. Climate scientists are interested in detecting the degree of similarity among global spatial distributions of different maps, and also in analyzing how different regions contribute to that similarity. They also want to identify regions that are most dissimilar across maps, and this gives them an incentive to further explore the causes behind this dissimilarity. This class of tasks is similar to the visual structure estimation tasks proposed by Szafir et al.([43]) but differ in the comparative nature of the tasks performed in a small multiple setting.

3.2 Perceptual Characteristics of Color Scales

In this paper, we have focused on characteristics of color scales which could have differential effects, depending on the magnitude comparison and spatial distribution comparison tasks. We chose the color scales based on the criteria of luminance monotonicity and hue banding. Perceptual uniformity of color spaces [11], [50] is another criterion we considered. However, equal steps in a uniform color space, do not ensure that the data magnitudes represented by these steps will be equally discriminable when used in a color scale [37], [50]. Equal JNDs in luminance perception (L*) is a good measure to represent human luminance perception, and a good predictor of the ability of the luminance component in a color scale to "carry" magnitude information. It does not, however, work for the other dimensions. Most notably, equal steps along the iso-luminant plane, which have large discriminable JNDs still do not "carry" high spatial-frequency magnitude information. For example, face images produced with iso-luminant variations were not visible ([38]), and high spatial-frequency features produced with stimuli that varied in hue and saturation, but not luminance, were less visible even though they had equal JNDs in a uniform color space [28]. Therefore, we did not compute distances for our color scales in terms of JNDs in a perceptually-uniform color space.

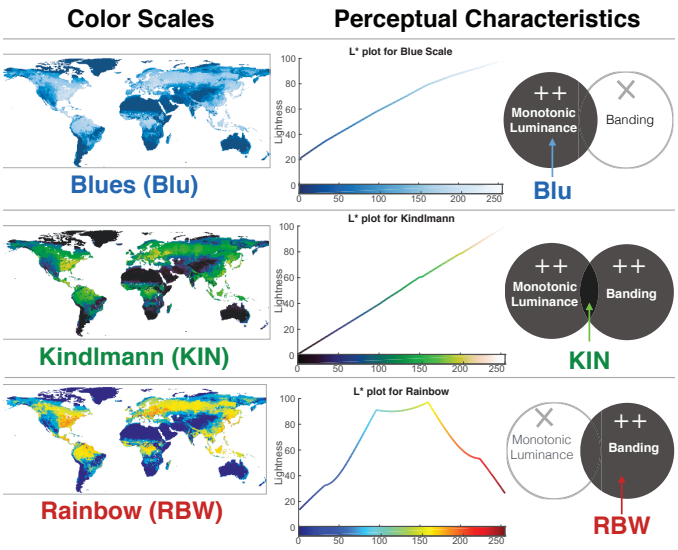


Fig. 1. **Perceptual characteristics of the three color scales in our study**. The first column shows these three scales (Blues, Kindlmann, and Rainbow) mapped onto the same climatological data. The two other columns show the luminance profiles for each scale and whether they display luminance monotonicity or hue-banding or both. By covarying these two perceptual dimensions, this study explores their role for different spatial data analysis tasks, like pairwise magnitude comparison and spatial distribution comparison for maps.

We discuss the rationale behind the chosen color scales below.

Effects of luminance and banding: The first of these characteristics is luminance monotonicity. Previous experiments have shown luminance monotonicity to be critical for carrying magnitude information, and critical for faithfully representing magnitude variations in high spatial-frequency data, such as the stimuli examined in climate modeling. The second feature is what we call "banding". In the rainbow color scale, for example, although the scale value increases monotonically with data value, the perceived hue does not change continuously. As in viewing a spectrum of light through a prism, we perceive bands of hues–blue, cyan, green, yellow, orange and red, which segment the data range into discrete regions.

Since monotonic luminance and banding are both present in the Rainbow (*RBW*), we selected experimental color scales that would allow us to understand their effects. Figure 1 shows three different color scales which vary in luminance monotonicity and banding. *RBW* has a non-monotonic luminance distribution; the luminance increases with data value, bounces around a bit, then decreases. The Rainbow also exhibits spatial banding. The hues do not vary continuously over the range, but describe distinct hue regions.

Alternative Color Scales: For the study we chose the *jet* color scale in MATLAB that our climate scientist collaborators use as a default. We chose two alternative color scales for our study based on their luminance and banding properties.

The blues color scale (*BLU*) is a popular selection from the ColorBrewer library [8], and has its roots in geographical map design. It has a single hue (blue) and is monotonic in luminance. Equal steps in color scale value correspond to measured increasing steps in perceived luminance. There is

minimal hue variation, and no hue banding, and since luminance is monotonic, there are also no *Mach bands* [7], which are perceptual discontinuities that can appear between luminance steps. We did not test the divergent color scale, also from the ColorBrewer library, because the scientists we worked with felt that the divergent color scale would incorrectly imply that there was a zero-crossing in the GPP variable.

The second color scale we selected was suggested by Kindlmann, et al. [22] as an alternative to the *RBW*. It uses vibrant, saturated colors, while also providing monotonic luminance. The scale runs from dark violet, through blue, to green, to yellow to white. Despite the luminance monotonicity, these hues appear as distinct bands, perhaps because of the large hue angle swept by this color scale.

These three color scales (Figure 1) allow us to separate out the effects of luminance modulation and banding, since these parameters covary across the three choices. *BLU* and *KIN* are monotonic in luminance, but differ in banding; *KIN* and *RBW* have hue banding, but differ in luminance monotonicity. Together, they make a useful set of color scales for examining how these two features interact in their effects on user performance in magnitude and spatial similarity tasks.

Pre-attentive Vision: The chosen color scales vary in another important perceptual way, which relates to attention. Bottom-up visual attention can be drawn "pre-attentively" to regions that have a different hue or luminance, that is to say, they have a *pop-out* effect for attracting attention. If the color scale contains such regions, data falling in such a pre-attentive range will be highlighted visually, even if values in this data range are not important. For example, the "yellow" region in the *RBW* has high luminance and appears very bright, so regions in the map that happen to fall within this range will attract attention. This could be a disadvantage, in that the region that is highlighted may not be of importance to the analytical task. Or, it could be an advantage, simplifying the task and focusing attention on a range of the data that could well be important. In fact, many practitioners using the *RBW* will manipulate its range to center the bright yellow or dark red on phenomena of importance.

3.3 Potential Effects of Color Scales on Tasks

Given the existing research on color scales and their use in visualization application, we outline our expectations regarding the effect of color scales on the tasks.

Effect on magnitude comparison: In Section 2, we reviewed several experiments in which luminance monotonicity was critical to making magnitude judgments ([38], [49]): luminance modulation is especially critical where the data has a high spatial frequency ([39]), as in geographical map applications. We expect, therefore, that the *BLU* and *KIN* color scales with this property would be effective in the magnitude comparison tasks faced in climate modeling. It has been suggested that adding a hue-variation would provide additional information to the observer [22], since it provides another channel of information. The Kindlmann color scale, however, not only provides a hue modulation, it also introduces perceived banding, which might reduce the observer's ability to make magnitude judgments. The semantic hue regions might mask the effectiveness of the magnitude cues provided by the monotonic luminance component of the color scale.

Effect on spatial distribution comparison: In practice, scientists often use segmented color scales to see differences in spatial distribution in their data. Since segmented color scales are ordinal, by definition, they do not provide a complete representation of all the data values, which are binned into color categories, but this binning can reveal structures in the data. We do not know of any study that explicitly compares segmented with continuous color scales, but it seems plausible that the banding produced in *RBW* and *KIN* color scales could provide benefit in making spatial distribution comparisons across geographies. These color scales, however, have the distinct problem, in that the size of the bands are not equal, which means that some regions in the data are differentially favored. In *RBW*, for example, the "blue" region occupies a much larger range in the data scale than the "yellow" region. That is, regions of low discriminability will be unevenly spaced over the data range and may lead to misinterpretation. Likewise, regions that are served by more closely-spaced bands may produce higher discrimination, since values in that region are sampled more finely.

4 Study Design

We designed a counterbalanced within-subjects experiment in which each observer performed three tasks on four types of map pairs, using three different color scales. We also collected confidence ratings and concluded the session with a survey that queried observers' subjective impressions about the different color scales. In this section, we describe all the different elements of the study.

4.1 Task Selection

We have crystallized the types of spatial data analysis tasks performed by climate scientists (Section 3.1) into three quantitative tasks. We selected a magnitude comparison task (Task 1) and two spatial distribution comparison tasks (Tasks 2 and 3) for our study, that we describe below. In order to allow all participants to be exposed to all color scales, we used a repeated measures design, where each participant had to perform a given task with all three color scales.

Task 1: In this magnitude comparison task, the participants compared the overall GPP in a reference map with overall GPP in a test map, and made a numerical estimate of the GPP in the test map. In this task, the fine spatial structure of the GPP variations was not considered; the participants simply provide an overall average estimate.

They were asked to answer the question: *"Given the global mean GPP based on one map (A), what is the global mean GPP of map (B)?"*. For providing their answer, participants had to adjust a slider, the range of which was set from the overall minimum to the overall maximum of mean GPP, across all models in the set of stimulus maps. To help participants with an explicit reference point, the initial position of the slider was set to the global mean GPP value of map A.

Task 2: In this spatial distribution comparison task, participants compared two maps, but in this case, focused on the spatial distribution of GPP, and judged the degree to which the comparison and test maps exhibit a similar spatial distribution of GPP. This was a very different task since two maps can have the same overall GPP, but very different spatial distributions.

Participants were given two maps and asked the question: *"How similar are the spatial distributions of the two maps?"* They were provided with a Likert scale that had a continuous range between 1 (most dissimilar) to 5 (most similar). Task 2 was thus about comparing the *degree of similarity* between two maps.

Task 3: This is also a spatial distribution comparison task, however, in this case, when the participant compared the two maps, they identified the region of maximum difference. Unlike the other two tasks, here the observer was not making an overall judgment, pooled across the whole spatial extent, but identified a single spatial region with the greatest difference in GPP.

Participants were given the same pair of maps as in Task 2 and asked to answer the question: *"In map A click on the area that is the most dissimilar from the one on B"*. Only a single click was allowed and they had to select a particular point on the map which they thought was the roughly the center of the region. Task 3 was thus about *dissimilarity identification* based on comparison of two maps.

Tasks 1, 2, and 3, are generally performed by climate scientists in a multi-way comparison setting, where more than two maps are involved. However, to simplify the tasks and make them achievable within a reasonable amount of time, for the study, we focused only on pairwise comparisons with two juxtaposed maps.

Since both Task 2 and Task 3 belong to the same class of structure (i.e., spatial distribution) estimation tasks, we decided to share the same trials between the two tasks, with both judgments made on the same map pair. There were thus 48 trials, 3 color maps x 2 trials for each of four spatial/magnitude quadrants x 2 tasks

Recording scientists' subjective impressions: Participants indicated their level of confidence on a discrete five-point Likert Scale. At the end of all the tasks, participants rated their familiarity, preference, perceived accuracy, and comfort in performing the tasks with the different color scales, and provided comments about the rationale behind their choices.

4.2 Hypotheses

Given the established link in the literature between luminance monotonicity and magnitude judgment, we predicted that performance in Task 1 (overall magnitude comparison) would be better with the two color scales with that property, which are *BLU* and *KIN*.

Task 2 involved comparing spatial distributions of GPP. If observers are simply making a magnitude judgment, then we would expect, again, that color scales with a monotonic luminance profile would enable the best performance. However, if the judgment is based on locating specific regions in the data, or if segmentation helped to reduce the complexity of the judgment, then we would expect that the two color scales with spatial banding would provide the best performance. And, if both segmentation and luminance monotonicity were at play, then the KIN color scale would be predicted to provide the best performance.

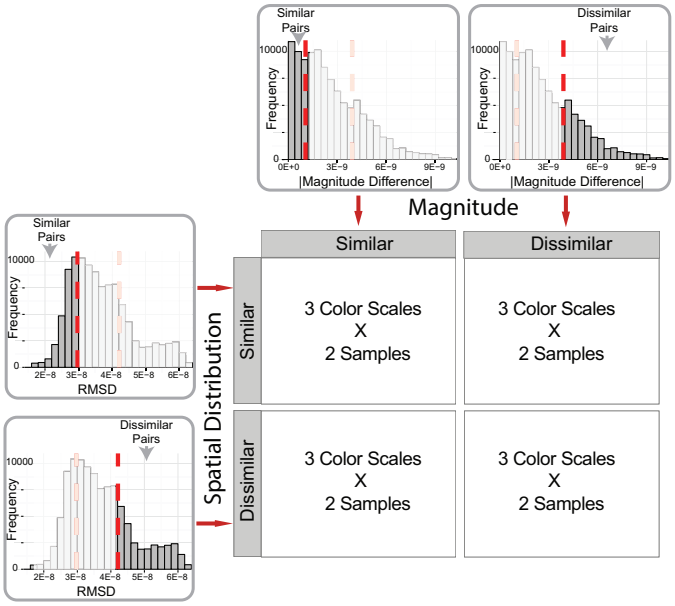


Fig. 2. **Selection of stimuli based on the magnitude and spatial characteristics of maps.** We used a spatial difference metric (RMSD) and a GPP magnitude difference metric (AMD), that climatologists use as part of their analysis, to characterize a large pool of map pairs. For these experiments, we identified the upper and lower quartiles for each metric, as marked by the red dotted lines. We randomly selected 8 pairs of maps, co-varying low vs. high-spatial difference (rows) and low vs. high- magnitude difference (columns). We show examples of map pairs in Figure 3.

Task 3 required the observers to select the region of greatest dissimilarity. In cases where the maps are dissimilar, this would involve identifying regions that are at the lower end of the scale in one map and at the higher end of the scale in the other. The bright white at the top of the KIN range or the saturated red at the top of the Rainbow might attract attention to this disparity, but may not be as effective when the two maps are similar.

Since the climate scientists in this study were most familiar with the Rainbow, since it is the de facto standard in their field, we expected that they would feel more confident using it than the two less familiar color scales.

In the subjective survey, we expected the scientists to express higher familiarity, ease of use, preference and confidence with the Rainbow color scale, since it is their common tool. If self-assessment of their own performance matched their actual performance, we would expect perceived accuracy to follow objective accuracy. This is an important measure, since introspection often guides choices in visualization, and a mismatch between introspection and reality would be a valuable observation.

4.3 Selection of Stimuli

Based on our general hypotheses and comparison based tasks, we aimed at generating pairs of maps that differ with respect to two main factors: *magnitude* and *spatial distribution* (Figure 2). We describe the metrics and the stimuli generation process below. Examples of the map pairs are shown in Figure 3.

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.2018.2876539, IEEE Transactions on Visualization and Computer Graphics

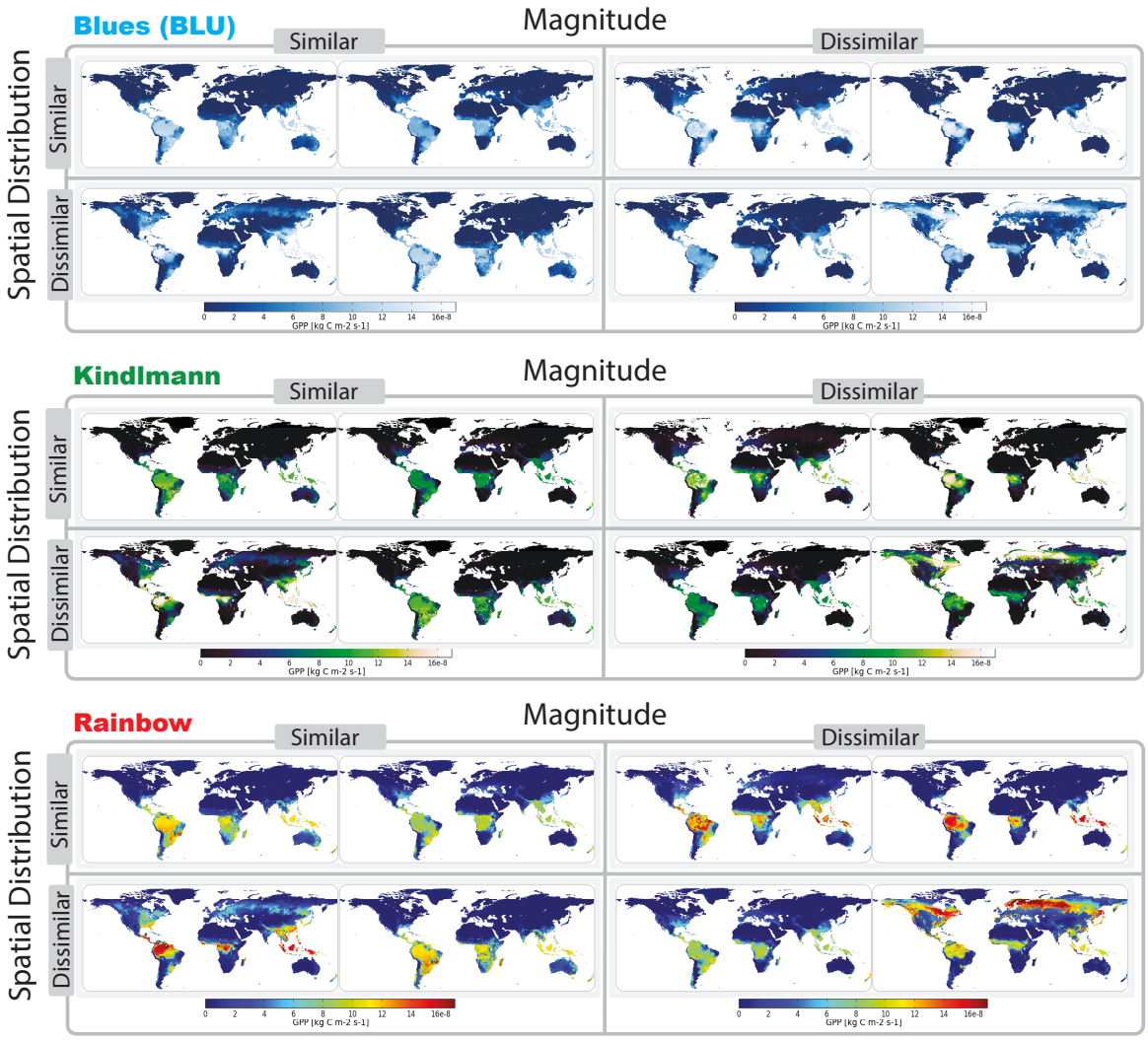


Fig. 3. **Example map pairs illustrating our selection of stimuli based on pairwise differences in magnitude and spatial distribution**: This figure shows color map pairs in each of the four quadrants defined in Figure 2, showing pair that co-vary in spatial and magnitude difference. Examples are shown for Rainbow (*RBW*), Blues (*BLU*) and Kindlmann (*KIN*) in the three sets. The maps were carefully selected based on the distributions of the climatological metrics: RMSD and Absolute Magnitude Difference. These metrics are used by climate scientists for quantifying differences between maps.

Generating similar and dissimilar map pairs: We selected stimuli for the experiment by grouping pairs of maps into four bins according to the scheme: low/high difference in magnitude and low/high difference in spatial distribution. For instance, two maps can have a similar distribution of values across the maps but different overall magnitude. As shown in Figure 3, it is possible to have maps with similar spatial distribution but different magnitude (top right) as well as maps with different spatial distributions but similar magnitude (bottom left). It is important to notice that while these differences may seem hard to understand by a non-expert, climate scientists are highly trained to derive this information from the color-coded maps.

In order to automatically generate map pairs that fall into the four groups outlined above, we leveraged metrics that climate scientists regularly use to quantify the difference between two maps in terms of magnitude and spatial distribution, and derived two measures after consulting our collaborators: *Root Mean-Squared Difference (RMSD)* to quantify the difference between two spatial distributions and *Absolute Magnitude Difference (AMD)* to quantify the difference between two global mean GPP. RMSD is obtained by comparing corresponding intensity values pixel-by-pixel between the two maps using Euclidean distance. Both of these metrics were area-weighted as equatorial regions have higher climatological weight than tropical regions. Map pairs have similar global mean GPP when AMD is low and similar spatial distributions when RMSD is low.

Figure 2 shows how maps were generated systematically from the distribution of these metrics. In order to create effective stimuli we selected, for both measures, map pairs in the lower quartile, to generate cases of high similarity, and those in the upper quartile to generate cases with low similarity. Accordingly we have four bins in the data: similar global mean GPP and similar spatial distribution, similar global mean GPP and dissimilar spatial distribution, dissimilar global mean GPP and similar spatial distribution, and dissimilar global mean GPP and dissimilar spatial distribution. Examples of these map pairs with the three color scales are shown in Figure 3.

Ensuring Variability and Coverage: Our maps pairs are generated using the GPP variable from 6 models (BIOME, GTEC, SIB3, CLM, CLM4VIC, LPJ) [19]. Each model has a spatial resolution of 360×720 and monthly temporal resolution of 360 time steps (30 years). The greatest variability in the model outputs is generally found across different seasons be it a same or different year. However we did not want to pick and choose the data from seasons of a particular year, as some events might affect that GPP for a region in a particular year, and we would not be able to account for that. Instead, to ensure variability we selected 10 random time steps for each model and compare against all the time steps of all the other models. Thus, we have in total 108,000 pairs (6 models $\times$ 5 models $\times$ 10 random time steps $\times$ 360 time steps). This not only ensured variability in the data but also a coverage of the data points. Eventually the map pairs for our study were selected from these pairs, based on our definition of stimuli as described previously.

4.4 Participants and Trials

We selected our participants for the main study anonymously through mailing lists of climate scientists, and 39 participants completed the study. Among them, 24 were male and 15 were female. Since 3 of them were identified as having color deficiencies, we excluded their responses from our analysis. The participants were aged between 24 and 65, with the median experience being 10 years in climate science and 6 years of using color scales with maps. The range of their overall experience was between 0 and 33 years.

Each participant completed all tasks and trials. The tasks were ordered sequentially and the trials were randomized to mitigate learning effects. Since there were 48 trials for each participant, the total number of trials was $48 \times 36 = 1728$. Tasks were always presented in the same order for each participant. However, the order of stimuli was randomized for each participant for a particular task. We did not impose any time restrictions for each trial or task as in a web-based study it is difficult to reliably control and analyze the effect of stimuli on response time.

4.5 Study Setting

Before deciding the final settings for the study, we conducted a pilot study, where we could build confidence in the tasks and color scales, and explore variations in the flow of the study, and check our training method for the participants. Participants who took the pilot tests were excluded from the main study.

The experiments reported in this study were all web-based. This setting was necessary as all our participants in the study are climate scientists spread across different academic institutions and research labs across the United States and Europe. One of the critical issues with our study is to ensure reliability and minimize bias in the results. In our experimental set-up we took several measures to address these. First, we took care of the case where participants did not understand the question or if they were ready for the test. To this effect, we showed them example questions and let them quit the study if they did not understand the question. They could not go back to check there answers or get feedback on the correctness of their responses. The IP address of the participants was recorded, so that we could know if the same participant has responded twice. Even if they stopped the study and took a break, they had to start from where they left off. This ensured prevention of unintentional repetition of the tasks by a participant.

5 Metrics for Judging Accuracy

In this section, we discuss the metrics we used to quantify the accuracy in the participants' judgments across the three tasks. The metrics we used were aimed at quantifying the error in participants' judgment as compared to the ground truth that was generated based on spatial distribution and magnitude difference of maps, which we illustrated in Figure 3.

5.1 Magnitude Comparison (Task 1)

Task 1 was performed by estimating a magnitude, i.e., global mean GPP of one map relative to the global mean GPP of another map. This was similar to the task of comparing the size of bars that was presented in the well-known study by Cleveland and McGill [13]. While in that case participants directly had to mention the degree to which bars were bigger or smaller, in our case participants provided an absolute value for the second map, and we derived the degree of overestimation or underestimation by normalizing the estimated GPP values with respect to reference GPP value. The error metric is thus derived as follows:

$$\text{Judged Percent} = \frac{\text{Estimated } GPP_B}{GPP_A} \times 100$$

$$\text{True Percent} = \frac{\text{True } GPP_B}{GPP_A} \times 100$$

$$\text{Error} = |\text{Judged Percent} - \text{True Percent}|$$

5.2 Degree of Similarity Comparison (Task 2)

The similarity comparison task was performed using a continuous Likert scale where 1 indicated lowest similarity and 5 indicated the highest similarity. As discussed in Section 4, the ground truth for generating maps was based on the difference in spatial distribution and magnitude. The RMSD metric quantified the difference or dissimilarity in spatial distribution between two maps: the larger an RMSD value, the more different were the distributions, while a smaller RMSD value indicated a smaller difference. For gauging the accuracy of this comparison task, we observe the inverse correlation between perceived similarity (Likert scale responses) and computed dissimilarity (based on the RMSD metric). The greater the correlation, the more accurate would be the visual comparison using a particular color scale.

5.3 Dissimilarity Identification (Task 3)

For Task 3, which was the task about identifying the most dissimilar region between two maps, we first compute the difference maps, where each point on the difference map indicates the absolute difference in GPP between two maps. From this difference map, we derive two values: value of the subject's click position in the difference maps ($\text{click}_{\text{diff}}$)

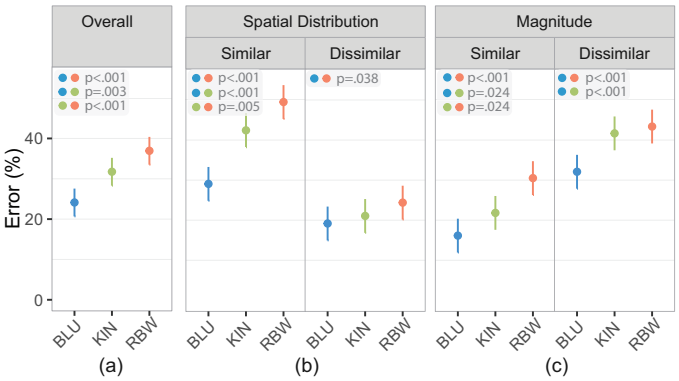


Fig. 4. **Task 1 (Magnitude Comparison) Results**: Percent error in judging GPP magnitude (using metric defined in Section 5.1) is plotted for three color scales (*BLU*, *KIN* and *RBW*) across all conditions (a) and with drill-downs for variations in Spatial Distribution (b) and Magnitude (c). Significant differences between color scales ($p < 0.05$) are annotated in the figure. We found the same ordering for the degree of error across all conditions, i.e., the *BLU* being the best and the *RBW* being the worst and most of the differences being statistically significant.

and maximum difference between map A and B ($\max_{\text{diff}}$). The error in dissimilarity judgment is given by:

$$\text{Error} = 1 - \text{click}_{\text{diff}}/\max_{\text{diff}}$$

The error varies from between 0 (no error) to 1 highest error. This metric captures how well a subject was able to select an area of maximal difference. In order to avoid effect of noise in our data, we compute $click_{diff}$ and max_{diff} using a Gaussian kernel in the pixel' neighborhood with $\sigma = 2$.

6 Results

We measured climate scientists' ability to judge magnitude similarity (Task 1), spatial distribution similarity (Task 2), and maximum difference (3) between two maps. The magnitude similarity and spatial similarity of the map pairs were co-varied in a counterbalanced design, and all pairs were judged using three color scales (*BLU*, *KIN* and *RBW*). We fit a mixed effects analysis of variance (ANOVA) model, with a normal conditional distribution and random effects for repeated measured to account for the non-independent nature of the data [48]. We include visualization type, magnitude, and spatial distribution as fixed effects, and participant as a random effect. We employed a mixed effects model since it is more robust and makes fewer assumptions than a repeated measures ANOVA model: it can cope with missing outcomes, time-varying covariates and relaxes the sphericity assumption of conventional repeated measures ANOVA. For these reasons, it is now becoming more commonplace to use a mixed effects model to analyze data in many domains [4] that used to be done by a repeated measures ANOVA design. We conducted post-hoc comparisons using the t-test with Bonferroni correction.

In this section, we report on the objective performance results in different conditions of magnitude and spatial distribution for three color scales. We then report on the subjective impressions that were recorded through a survey at the end of the study. For all our results we computed the 95% confidence intervals using the bootstrapping method.

6.1 Magnitude Comparison

Task 1 addressed magnitude comparison: judging the global mean GPP in one map with respect to the given reference GPP for another map.

Overall Effect: In Figure 4(a) we plot the absolute % error in judging the magnitude difference between map pairs and 95% confidence intervals for the three color scales (*BLU*, *KIN*, and *RBW*). Significant differences between color scales is indicated by the asterisks below the x-axis labels. The first panel (a) shows overall performance across conditions. Panels (b) shows the break-down by Spatial Distribution; Panel (c) shows the breakdown by Magnitude. Overall (a), the two monotonic luminance scales were more effective in helping the analysts make correct judgments about the global mean GPP than *RBW*. Users had a significantly higher error rate with *RBW* (Mean: 37%, CI: [34.3, 39.5]), and significantly fewer errors with *KIN* (32%, [29.1, 34.3]) and *BLU* (24%, [21.5, 26.7]) , showing F(2,930) = 16.75, $p < .001$. The performance with *BLU* was significantly better than *RBW* ($p < .001$), and *KIN* ($p = .003$), and performance with *KIN* was significantly better than with *RBW* ($p < .001$).

Effect of Spatial Distribution: In Figure 4(b) we drill down with respect to similar and dissimilar spatial distributions between maps. The ordering of results for the three color scales is the same in both conditions, that is, *RBW* produces the highest error, followed by *KIN*, and followed by *BLU*. All these differences are significant when the maps being compared are spatially similar (F(2,465) = 20.29, $p < .001$), where the *BLU* (Mean: 29.0, CI: [25.4, 32.7]) led to significantly fewer errors than *RBW* (49.4, [45.8, 53.1]; $p < .001$) and *KIN* (42.4, [38.7, 46.0]; $p < .001$) and *KIN* led to significantly fewer errors than the *RBW* ($p = .005$). There was a significant, but weaker, effect of color scales when the maps being compared were spatially dissimilar (F(2,465) = 3.38, $p = .035$) with only the difference between *RBW* (24.3, [20.7, 28.0]) and *BLU* (19.1, [15.5, 22.8]) being significant ($p = .038$). Thus, *RBW* color scale affords less accurate comparisons of magnitude, whether the spatial distributions are similar or dissimilar, but the degree to which the monotonic luminance scales outperform is much greater when the maps are similar. This also shows clearly that the task of comparing GPP is much harder with multi-hue color scales when the maps have similar spatial distributions. The amount of error using *RBW* is almost twice that when using the BLU, while the error using *BLU* is comparable for both similar and dissimilar spatial distributions.

Effect of Magnitude: Figure 4(c) shows the break-down by magnitude difference for the Magnitude estimation task. When the comparison maps are similar in magnitude (F(2,465) = 13.12, $p < .001$), performance with *BLU* (Mean: 16.1, CI: [12.4, 19.8]) was significantly better than with *RBW* (30.5, [26.8, 34.1]; $p < .001$), and *KIN* (21.8, [18.1, 25.4]; $p = .024$), performance with *KIN* was significantly better than with *RBW* ($p = .024$). When the comparison maps are dissimilar in magnitude, the performance with *BLU* (32.0, [28.4, 35.7]) was significantly better than the *RBW* (43.3, [39.7, 26.8]; $p < .001$) and *KIN* (41.6, [38.0, 45.3]; $p < .001$).

Another interesting observation in these data is that there were significant differences between conditions. Participants had a very hard time judging magnitude differ-

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.2018.2876539, IEEE Transactions on Visualization and Computer Graphics

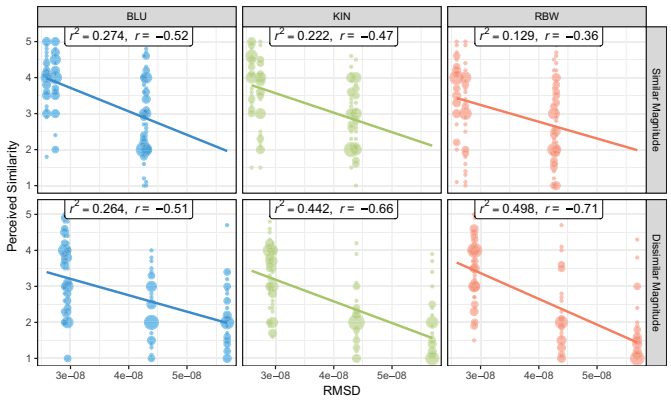


Fig. 5. **Task 2 (Spatial Distribution Comparison) Results.**: Perceived similarity (as rated on a Likert scale) as a function of computed dissimilarity (given by the RMSD metric) for the two magnitude conditions. Inverse correlation is an expected pattern. We can observe that: i) performance with *BLU* was not affected whether magnitude was similar or dissimilar. ii) Both *KIN* and *RBW*'s correlations improved when the maps' magnitudes were dissimilar, and this was especially true for *RBW*, which had an $r = 0.7$ in the dissimilar magnitude case.

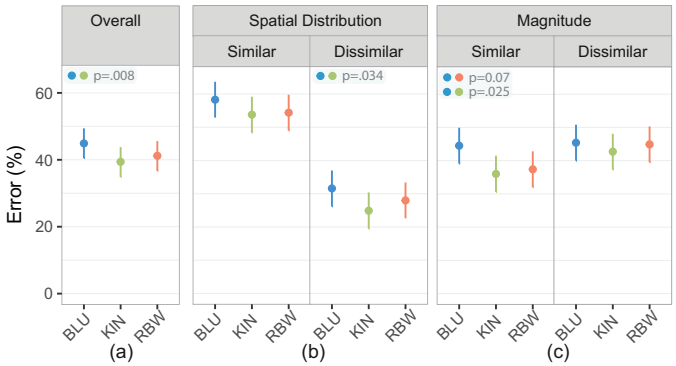


Fig. 6. **Task 3 (Identification of Region of Maximum Difference) Results** Percent error in identifying the most dissimilar region (using metric defined in Section 5.3) is plotted for three color scales (*BLU*, *KIN* and *RBW*) across all conditions (a) and with drill-downs for variations in Spatial Distribution (b) and Magnitude (c). Significant differences between color scales ($p < 0.05$) are annotated in the figure. We found that dissimilarity judgment is affected by the color scales. Performance using the *BLU* color scale was worst in all of the cases, with significant differences from *KIN* in the overall and dissimilar spatial distribution case, and from *KIN* and *RBW* in the similar magnitude case.

ences when the spatial distributions of the pairs were dissimilar or when the magnitudes were similar. Despite large differences in the difficulty of the task, however, the best performance was achieved using *BLU* and *KIN*, the two color scales with monotonic luminance. We can also see that the *BLU* scale afforded better performance than the *KIN*, whose luminance range is higher. This suggests that the hue modulation in *KIN* did not enhance magnitude estimates, and may have had a detrimental effect, countering the benefit of its monotonic luminance component.

6.2 Spatial Distribution Comparison

Task 2 was about judging the degree of similarity between a pair of maps. We evaluate the performance on Task 2 by looking into the inverse correlation between perceived similarity and computed dissimilarity as was described in Section 5.2. As shown in Figure 5, the expected pattern is an inverse correlation between perceived similarity on the Y-axis and computed dissimilarity, using the RMSD metric, on the X-axis: perceived similarity based on the Likert scale responses increases along the Y-axis from 1 (most dissimilar pair) to 5 (most similar pair) and computed dissimilarity based on the RMSD metric increases along the X-axis. We further drill down into the categories of similar and dissimilar GPP magnitude (Figure 5). The correlation values for the *BLU* scale are identical across the two magnitude conditions, but *RBW* and *KIN* show noticeable differences. For map pairs with similar magnitudes, *RBW* and *KIN* exhibit much poorer correlation between perceived and computed ground truth. Also, note the change in orderings: *RBW* performs the best for the dissimilar magnitude case, *BLU* performs the best for the similar magnitude case, while *KIN* is always in the middle.

Task 3 was about identifying the region of maximal difference. Figure 7 shows performance in this task using the error metric derived in Section 5.3. Overall, the performance with *BLU* (Mean: 44.9, CI: [40.5, 49.4]) was worse than with *KIN* (39.4, [34.9, 43.8]) and *RBW* (41.2, [36.7, 45.6]). There was no statistically significant difference in performance between *KIN* and *RBW* (F(2,930) = 2.316, $p = .100$). The difference between the *BLU* and the *KIN*, was significant ($p = .009$). This result is echoed in the Dissimilar Spatial drill-down (panel b, where F(2,465) = 4.70, $p = .009$) and in the Similar Magnitude case (panel c, where F(2,465) = 4.81, $p = .008$). In panel (b), the difference between the *BLU* (31.6, [26.2, 37.0]) and the *KIN* (24.9, [19.5, 30.4]) was significant ($p = .034$) and, in panel (c), the *BLU* (44.5, [39.1, 49.9]) produced significantly more errors than the *KIN* (36.0, [30.6, 41.5]; $p = .025$) and *RBW* (37.4, [32.0, 42.8]; $p = .007$).

In Figure 7 we show an example of the variance in the clicked regions across different color scales. The third row provides a visualization of the actual difference in GPP for that pair. The dots indicate the geographic regions identified as being the most different. In this example, we see generally good agreement between performance using the three color scales. However, this agreement does not necessarily match the regions of greatest actual difference. When using the *RBW* and *KIN* color scales, many observers identified northern South America as containing regions of maximal difference, even though the physical difference between maps in not high in that geography. This error is not as evident with the *BLU* scale.

6.3 Analyzing Subjective Performance Measures

One of the goals of our study was to compare the perceived accuracy and confidence of the scientists with the objective measures from the study. To this effect, we asked participants to rate their level of confidence for each task, and collected their subjective feedback in last section of the study. We collected feedback about their familiarity, preference, confidence, perceived accuracy and ease of use of the color scales, by asking questions such as: "which color scale did you prefer the most", "which color scale were you most confident with", etc. In this section we present an analysis of their task-wise confidence ratings, their subjective impressions, and the effect of domain experience on their ratings.

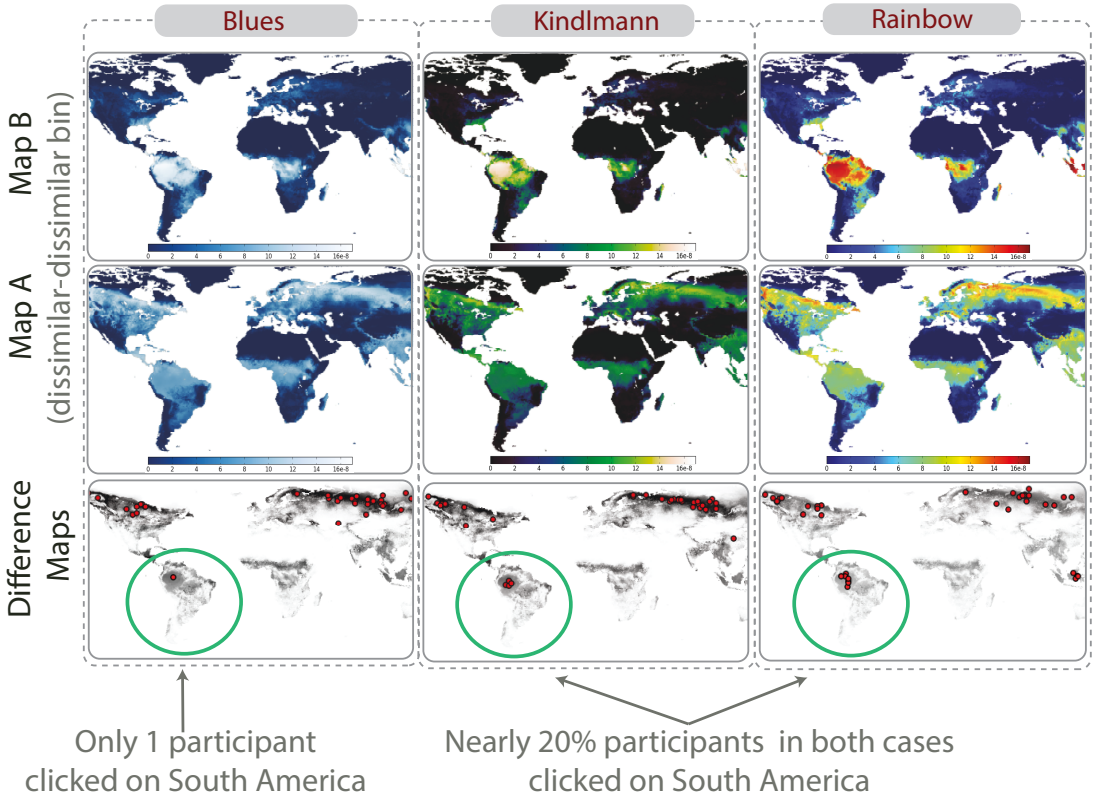


Fig. 7. **A case from Task 3 where the maps had dissimilar spatial distributions and dissimilar magnitude. The top two rows show the two comparison maps used in the study, with the difference map shown in the third row, for *RBW*, *KIN* and *BLU*.** In this case, 20% of the observers clicked on a region of South America that did not have a particularly high difference. This error was not made with the *BLU* map.

6.3.1 Task-wise Confidence Ratings

We analyzed the confidence rating results for all the tasks. Participants seemed to be more confident with Task 2 and 3, with the average confidence levels being higher. Analyzing the effect of color scales on confidence ratings, we found that overall the scientists were least confident with the *BLU*: they were more confident on average with the *RBW* than the *BLU* ($p = .004$) and more confident on average with the *KIN* than the *BLU* ($p < .013$).

We also compared their objective performance, using the error metrics for the different tasks with their perceived confidence levels. We expected scientists to commit fewer errors when they were more confident. However, for Task 1 we found the average confidence level for *BLU* was slightly lower (3.04) than *KIN* (3.26) and *RBW* (3.26), despite the average accuracy being greater. Also, at high self-rated confidence levels (greater than 3), the average degree of error was much higher in *RBW* (37.4%) than *BLU* (24.7%) or *KIN* (31.4%), thereby showing a discrepancy between scientists' confidence and accuracy levels. For Tasks 2 and 3, we did not find any noticeable variability in error with respect to high or low confidence levels.

6.3.2 Post-Study Survey

The results of the post study survey are shown in Figure 8, and we comment on the general trends below.

Perceived Accuracy and Confidence vs. Familiarity: Since the *RBW* is the *de facto* standard in climate science, it was not surprising that over 94% responded that they were most familiar with it. Despite the familiarity with the *RBW* among an overwhelming majority of them, nearly 25% of the participants felt more accurate or confident with either the *KIN* or the *BLU*.

Familiarity vs. Preference: Comparing familiarity to preference, we observe a difference of nearly 40% for the *RBW*, which is compensated by more participants preferring either the *BLU* or the *KIN*. Given the high familiarity with the rainbow color scale, this difference is significant. We can observe that the subjective preferences of a significant number of climate scientists were in favor of a relatively unfamiliar, perceptually corrected color scale for the specific study conditions. Following are some of their comments that demonstrate, although the scientists were overwhelmingly positive about the Rainbow, they could recognize its liabilities and advantages of the other color scales, especially *KIN*:

"Kindlmann works best because it has both good tone contrast AND value contrast across the spectrum, whereas rainbow has good tone contrast but little value contrast and blues has little color contrast and not great value contrast."

"It was easier to see magnitude of change with rainbow, and especially hotspots in red. My concern was that I was overestimating the red areas and not paying enough attention to changes at the other end of the spectrum. I thought my first sense of overall global pattern change was easier with blues but it was much harder to compare changes in spatial pattern or magnitude between different regions. Kindlmann was therefore a compromise for me...not as dramatic, did not highlight the hotspots as much, but allowed me to compare differences more easily across regions".

Effect of Experience: We also looked at the effect of expe-

rience on the subjective impressions of the participants. We wanted to investigate if greater domain experience has any effect on the perceived confidence, accuracy and comfort levels with different color scales. We used the median of the self-reported years of experience of the participants to divide them into *high* and *low* experience groups. However, we failed to find any significant effect of experience. Marginally higher percentage of less experienced participants showed greater post-study preference for color scales other than the *RBW*, but none of these effects were significant.

7 Discussion

In this section, we summarize our key findings about the effect of color scales on objective and subjective performance measures, how they relate to our hypotheses and the open research questions that our findings lead to.

7.1 Magnitude Comparison

In Task 1, we found that visual performance in judging the magnitude difference between climate maps was best using color scales with a monotonic luminance component. Judgments with *BLU* scale had the fewest errors, followed by *KIN*, and last, by *RBW*. This ordering was observed in the average across conditions, and also in the drill-downs by magnitude and by spatial distribution. This is a strong and also statistically significant result.

Effect of luminance monotonicity: Since the two scales with monotonic luminance enabled the best performance, clearly this factor plays a major role. However, closer examination reveals that the situation is more complex, since the luminance modulation for *KIN* was greater than for *BLU* (Figure 1). If luminance modulation were the only driving factor, *KIN* would thus be expected to provide the better performance in magnitude judgment.

Effect of hue banding: The observation that performance is worse with *KIN* than *BLU* also challenges the notion that variations in hue would facilitate magnitude judgments, providing an extra channel for signaling magnitude signal [49]. One possible explanation is that the banding in the *KIN* color scale actually inhibited the judgment of magnitude. If changes in data magnitude are less salient to the user if they occur within a hue band, missing that information would be expected to reduce magnitude judgment performance. Likewise, if bands in the *KIN* or *RBW* color scale artifactually enhanced the magnitude of the data in a particular range, that could also produce errors in judging magnitude.

Scope for future research: In future work, it would be valuable to create a set of color scales that explicitly co-varied luminance monotonicity and hue banding. One method for doing so would be to vary the hue trajectory of the hue-varying monotonic-luminance scale. In the *KIN* color scale, the colors sweep a large angle around the hue circle, from dark purple to blue to green through yellow to white. Several authors [6], [36] have argued about an advantage of the heated-body color scale, which is monotonic in luminance but covers a narrow hue angle, ranging from dark red through orange to yellow and white. To further compare the perceptual effects of monotonic luminance and banding, color scales could be constructed that had identical luminance modulations, and carefully-constructed hue and saturation variations.

7.2 Degree of Spatial Similarity Judgment

In Task 2, if we average across conditions, all three color scales provided equal benefit to the observers. Differences were observed however, when looking at the drill-downs by magnitude similarity. When the comparison maps were similar, the results agreed with those of the magnitude estimation task; the best color scale was *BLU*, followed by *KIN* and then by *RBW*. When the compared maps were dissimilar in overall magnitude, however, the ordering was different. Using the *BLU* color scale produced less correlation between perceived and computed similarity than with *KIN* and *RBW*.

Nature of Spatial Judgments: The nature of spatial judgments can be understood based on Bertin's proposed reading levels of the human vision system [3]. It may be, when two maps have similar spatial profiles, there are fewer salient spatial features. The judgment possibly occurs at an *elementary level*, where scientists look at the relative value encoded by the color of each pixel. Since colored pixels have no spatial extent, the human vision system is good at recognizing the average value of an area, and the size of those areas [26]. The perceptual problem reduces to a magnitude judgment, where monotonic luminance has a clear advantage as shown in Task 1. When the maps are spatially dissimilar, there are fewer local cues, and judging shapes across spatial regions may rely on other mechanisms. The judgment possibly occurs at an intermediate level, where they look at the shapes of distributions formed by the pixels, and the task involves visually segmenting regions and judging their magnitude.

Effect of hue-banding: In case of intermediate levels of judgments, hue-banding can help in segmenting different regions. As we discussed in the introduction, both the Rainbow and *KIN*'s hues segment the data range into regions with semantic color names. This de facto segmentation might help identify regions in maps that had spatially-adjacent regions with similar values, making the spatial judgment easier.

Scope for future research: To study this possible interaction effect (luminance monotonicity better for similarity judgments when the maps are similar, and hue-banding being better when the maps are dissimilar) would require exploring maps with carefully-controlled spatial modulations. The climate map is a very complex, high spatial-frequency stimulus. It has been demonstrated [40] that segmented color scales are more effective for representing changes in low spatial-frequency data, so an interesting experimental manipulation might be to explore this possible interaction effect at a range of spatial frequencies.

7.3 Identification of the Most Dissimilar Region

In each trial in Task 3, the observer identified the region in the test map that was most different from its corresponding region in the comparison map, and this judgment was compared with ground truth. Overall, the *BLU* color scale was the least useful to the observer in making this

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.2018.2876539, IEEE Transactions on Visualization and Computer Graphics

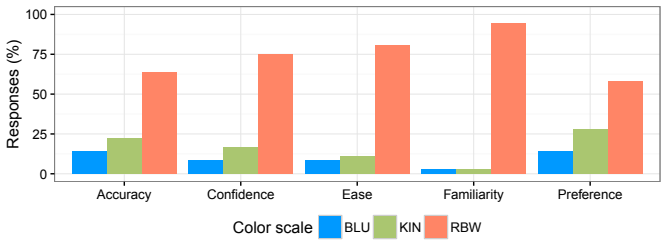


Fig. 8. **Analyzing subjective impressions of participants**. The Y-axis represents the percentage of responses for each category. an overwhelming majority of the participants also expressed preference for the *RBW* despite its lowest performance accuracy for Task 1, and its comparable performance for Task 2 and Task 3 with respect to *BLU* and *KIN*. Also, despite their overall familiarity with the *RBW*, about 43% of the participants preferred the *KIN* or the *BLU* and 33% felt they were more accurate with them in their post-study survey.

judgment, suggesting that the observers were not basing their choices solely on magnitude judgments. The *KIN* color scale enabled significantly better performance than the *BLU*, overall in two of the drill-downs. *KIN* and *RBW* were never significantly different, suggesting that a common characteristic drove their performance.

Effect of Pre-Attentive Vision: One possibility is that the *RBW* and *KIN* color scales are providing visual cues that guide attention to specific regions in the data range where the color is particularly bright or prominent. Figure 7 shows a set of map pairs where observers using *KIN* and *RBW* identify an area in eastern Colombia as being a region of maximum difference when clearly it is not. Looking at the two color scales, we see that even though this is not a region of highest difference, that region in one of the maps just happens to fall in the prominent "yellow" region of *KIN* and the "red" region of RBW. In this case, it is interesting to note, this region was not falsely called out using the *BLU* scale.

Scope for future Research: It may be, thus, that attention is falsely drawn to a region, because those data values just happened to fall on salient colors. To test this hypothesis further, we would want to construct color scales in which the highlighted region could be manipulated with respect to the data, to measure the extent to which an errant highlight could distort judgments.

7.4 Subjective Impressions

In a post-experiment survey, all the participants were asked to judge the color scales on a number of attributes (Figure 8). For subjective measures of accuracy, confidence, ease, familiarity, preference we found that: *RBW*, unsurprisingly, has much higher scores for all of these metrics (that is, all of our participants found *RBW* more accurate, felt more confident in the results, found it easier to use, more familiar and preferred it over the others), but with different proportions. The fact that proportions differ points to interesting interpretations.

Perceived Vs Objective Accuracy: Another very interesting finding is that although 70% of respondents marked *RBW* as being subjectively more accurate, our results do not seem to point to any advantage, in terms of accuracy, of *RBW* over the other color scales. In fact, Rainbow was the least effective color scale for magnitude judgments. This is a very important finding: there is a mismatch between the subjective perception of how accurate one is and how one actually is.

Familiarity Vs Preference: When comparing familiarity to preference we observe a major shift. Many of our respondents prefer *KIN* or *BLU* even being more familiar with *RBW*. Unfortunately, given the setup we used for the experiment, we do not know whether this observed shift has been induced by the study itself or just a prior preference our participants had before participating to our study. In any case this results demonstrate a certain degree of awareness of the potential issues with *RBW* and the fact that for some datasets in some conditions other color scales may be appropriate.

Scope for future research: Recent research has shown: i) visualization tools have the potential to inspire a higher level of trust in analysts as compared to more familiar methods [15], and ii) effective visualization design can lead to greater performance accuracy in visual comparison based judgments, as compared to more familiar visual representations of climate models [14]. However, in this study, we found that an overwhelming majority of climate scientists indicating their preference for *RBW* due to prior familiarity despite their preference not being reflective of their actual performance in any of the tasks. As indicated by Moreland [30], one of the key reasons is that the *RBW* scale is deeply "entrenched in scientific visualization", the default scale for many tools, and therefore, scientists keep using it out of a kind of inertia. This hinders the adoption of potentially better color scales. We believe that an effective way forward is to conduct more studies with domain experts and their data for demonstrating the value of perceptually more optimal color scales. In the future, we will conduct more experiments to this end that will help influence scientists' preference levels and provide them with better design choices for solving their tasks.

8 Conclusion and Future Work

Using color scales to represent data values is one of the most important and ubiquitous operations in visualization. The color scale selected may be guided by conventions in a particular field, by the choices available in the visualization system being used, or by perceptual research, which over the past 20 years, has offered advice on which color maps to use for particular situations, or which color components in the color scale itself best communicate specific features in the data. Some of this guidance has been very simplistic, as in "never use a Rainbow color scale", and some has been very abstract, such as measuring the effectiveness of different color scales on artificial stimuli. Attempts have been made to develop taxonomies that can help the practitioner select appropriate color scales, either building on the data type, perceptual operations, or different tasks. But, these perceptual operations and tasks, so far, have been quite simple, and with a few exceptions, have not been conducted with domain experts [5], and hence, do not capture the complexity of the problem-solving needs of scientists and engineers in real-world settings.

We presented a web-based user study for measuring the effectiveness of different color scales in climate modeling tasks. In a counterbalanced design, climate scientists made three different judgments of map pairs, each judgment capturing a representative task in their real-world analysis environment, using representative stimulus comparisons. In the first task, observers judged the overall magnitude difference between pairs; in the second, they judged the spatial similarity of each pair; and in the third, they clicked on the region that was most dissimilar. In each task, three different color scales were used to represent the data. The color scales co-varied in luminance monotonicity and hue banding. The *BLU* scale was monotonic in luminance and displayed no color banding. The Rainbow color scale (*RBW*) was not monotonic in luminance and displayed visible hue bands. The *KIN* scale was monotonic in luminance and also contained visible hue bands.

Our key findings are the following. i) Monotonic luminance had a positive effect, and hue banding seemed to have a negative effect on magnitude comparison, ii) color scales with hue banding enabled more accurate judgments of differences in spatial distribution, iii) scientists' high confidence levels with the rainbow color scale, did not get reflected in greater performance accuracy, and iv) despite overwhelming familiarity with the rainbow, many scientists expressed post-study preference for the relatively unfamiliar *KIN* color scale.

We expect that our results would generalize to the representation of any scalar variable on across a geographical map, at least at the spatial resolutions we studied. In our experiments, we studied just three color scales, which sampled two theoretically important ideas: luminance monotonicity and banding (with a secondary focus on highlighting). In the future, based on the knowledge gained about the climatological tasks, we will design more experiments to study the effects of luminance dynamic range, the continuous vs segmented nature of color scales, and spatial frequency. We will also continue to pursue the research questions about similarities and differences between performance accuracy and subjective impressions to see how visualization adoption can be impacted by our findings.

9 Acknowledgment

This work was supported in part by: the Pacific Northwest National Laboratory; the Moore-Sloan Data Science Environment at NYU; NASA; DOE; NSF awards CNS-1544753, CNS-1229185, CCF-1533564, CNS-1730396, OAC-1640864. C. T. Silva is partially supported by the DARPA D3M program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of DARPA.

References

[1] L. D. Bergman, B. E. Rogowitz, and L. A. Treinish. A rule-based tool for assisting colormap selection. In *Proceedings Conference on Visualization*, page 118. IEEE Computer Society, 1995.

[2] J. Bernard, M. Steiger, S. Mittelstädt, S. Thum, D. Keim, and J. Kohlhammer. A survey and task-based quality assessment of static 2d colormaps. In *IS&T/SPIE Electronic Imaging*, pages 93970M–93970M. International Society for Optics and Photonics, 2015.

[3] J. Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. Central Asia book series. University of Wisconsin Press, 1983.

[4] M. P. Boisgontier and B. Cheval. The anova to mixed model transition. *Neuroscience & Biobehavioral Reviews*, 68:1004–1005, 2016.

[5] M. Borkin, K. Gajos, A. Peters, D. Mitsouras, S. Melchionna, F. Rybicki, C. Feldman, and H. Pfister. Evaluation of artery visualizations for heart disease diagnosis. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2479–2488, 2011.

[6] D. Borland and R. M. Taylor. Rainbow color map (still) considered harmful. *Computer Graphics and Applications*, 27(2):14–17, 2007.

[7] R. Boynton, M. Hayhoe, and D. MacLeod. The gap effect: chromatic and achromatic visual discrimination as affected by field separation. *Journal of Modern Optics*, 24(2):159–177, 1977.

[8] C. A. Brewer. Color use guidelines for data representation. In *Proceedings of the Section on Statistical Graphics, American Statistical Association*, pages 55–60, 1999.

[9] C. A. Brewer, A. M. MacEachren, L. W. Pickle, and D. Herrmann. Mapping mortality: Evaluating color schemes for choropleth maps. *Annals of the Association of American Geographers*, 87(3):411–438, 1997.

[10] C. A. Brewer and L. Pickle. Evaluation of methods for classifying epidemiological data on choropleth maps in series. *Annals of the Association of American Geographers*, 92(4):662–681, 2002.

[11] R. Bujack, T. L. Turton, F. Samsel, C. Ware, D. H. Rogers, and J. Ahrens. The good, the bad, and the ugly: A theoretical framework for the assessment of continuous colormaps. *IEEE transactions on visualization and computer graphics*, 24(1):923–933, 2018.

[12] L. Cheong, S. Bleisch, A. Kealy, K. Tolhurst, T. Wilkening, and M. Duckham. Evaluating the impact of visualization of wildfire hazard upon decision-making under uncertainty. *International Journal of Geographical Information Science*, 30(7):1377–1404, 2016.

[13] W. S. Cleveland and R. McGill. Graphical perception: The visual decoding of quantitative information on graphical displays of data. *Journal of the Royal Statistical Society. Series A (General)*, pages 192–229, 1987.

[14] 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. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1193–1204, 2017.

[15] A. Dasgupta, J.-Y. Lee, R. Wilson, R. A. Lafrance, N. Cramer, K. Cook, and S. Payne. 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):271–280, 2017.

[16] A. Dasgupta, J. Poco, Y. Wei, R. Cook, E. Bertini, and C. Silva. Bridging theory with practice: An exploratory study of visualization use and design for climate model comparison. *IEEE TVCG*, 21(9):996–1014, 2015.

[17] 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.

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

[19] D. Huntzinger, C. Schwalm, Y. Wei, R. Cook, A. Michalak, K. Schaefer, A. Jacobson, M. Arain, P. Ciais, J. Fisher, D. Hayes, M. Huang, S. Huang, A. Ito, A. Jain, H. Lei, C. Lu, F. Maignan, J. Mao, N. Parazoo, C. Peng, S. Peng, B. Poulter, D. Ricciuto, H. Tian, X. Shi, W. Wang, N. Zeng, F. Zhao, and Q. Zhu. NACP MsTMIP: Global 0.5-deg terrestrial biosphere model outputs (version 1) in standard format. Oak Ridge National Laboratory Distributed Active Archive Center.

[20] D. N. Huntzinger, C. Schwalm, A. M. Michalak, K. Schaefer, et al. The north american carbon program multi-scale synthesis and terrestrial model intercomparison project - part 1: Overview and experimental design. *Geoscientific Model Development Discussions*, 6(3):3977–4008, 2013.

[21] P. K. Kaiser and R. M. Boynton. Human color vision. 1996.

[22] G. Kindlmann, E. Reinhard, and S. Creem. Face-based luminance matching for perceptual colormap generation. In *Proceedings of the conference on Visualization'02*, pages 299–306. IEEE Computer Society, 2002.

[23] Kitware. The Visualization Toolkit (VTK) and Paraview. http://www.kitware.com.

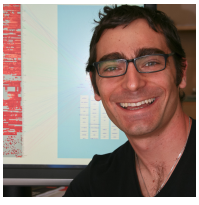
[24] H. Levkowitz and G. T. Herman. Color scales for image data. *IEEE Computer Graphics and Applications*, 12(1):72–80, 1992.
[25] Y. Liu and J. Heer. Somewhere over the rainbow: An empirical assessment of quantitative colormaps. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2018.
[26] A. M. MacEachren. *How maps work: representation, visualization, and design*. Guilford Press, 2004.
[27] J. Maule and A. Franklin. Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of vision*, 15(4):6–6, 2015.
[28] W. H. McIlhagga and K. T. Mullen. Contour integration with colour and luminance contrast. *Vision Research*, 36(9):1265–1279, 1996.
[29] K. Moreland. Diverging color maps for scientific visualization. In *Advances in Visual Computing*, pages 92–103. Springer, 2009.
[30] K. Moreland. Why we use bad color maps and what you can do about it. *Electronic Imaging*, 2016(16):1–6, 2016.
[31] K. T. Mullen. The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *The Journal of physiology*, 359(1):381–400, 1985.
[32] T. Munzner. *Visualization Analysis and Design*. CRC Press, 2014.
[33] S. E. Palmer and K. B. Schloss. An ecological valence theory of human color preference. *Proceedings of the National Academy of Sciences*, 107(19):8877–8882, 2010.
[34] L. W. Pickle. Usability testing of map designs. In *Proceedings of Symposium on the Interface of Computing Science and Statistics*, pages 42–56, 2003.
[35] K. Reda, P. Nalawade, and A.-K. Kate. Graphical perception of continuous quantitative maps: the effects of spatial frequency and colormap design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2018.
[36] P. L. Rheingans. Task-based color scale design. In *28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making*, volume 3905, pages 35–44. International Society for Optics and Photonics, 2000.
[37] B. Rogowitz and A. D. Kalvin. The" which blair project": a quick visual method for evaluating perceptual color maps. In *Visualization, 2001. VIS'01. Proceedings*, pages 183–556. IEEE, 2001.
[38] B. E. Rogowitz, A. D. Kalvin, A. Pelah, and A. Cohen. Which trajectories through which perceptually uniform color spaces produce appropriate colors scales for interval data? In *Color and Imaging Conference*, pages 321–326. Society for Imaging Science and Technology, 1999.
[39] B. E. Rogowitz and L. A. Treinish. How not to lie with visualization. *Computers in Physics*, 10(3):268–273, 1996.
[40] B. E. Rogowitz and L. A. Treinish. Data visualization: the end of the rainbow. *Spectrum, IEEE*, 35(12):52–59, 1998.
[41] K. B. Schloss, L. Lessard, C. Racey, and A. C. Hurlbert. Modeling color preference using color space metrics. *Vision research*, 2017.
[42] S. Seipel and N. J. Lim. Color map design for visualization in flood risk assessment. *International Journal of Geographical Information Science*, 31(11):2286–2309, 2017.
[43] D. A. Szafir, S. Haroz, M. Gleicher, and S. Franconeri. Four types of ensemble coding in data visualizations. *Journal of vision*, 16(5):11–11, 2016.
[44] J. Tajima. Uniform color scale applications to computer graphics. *Computer Vision, Graphics, and Image Processing*, 21(3):305–325, 1983.
[45] L. Takayama and E. Kandogan. Trust as an underlying factor of system administrator interface choice. In *CHI'06 extended abstracts on Human factors in computing systems*, pages 1391–1396. ACM, 2006.
[46] C. Tominski, G. Fuchs, and H. Schumann. Task-driven color coding. In *Information Visualisation, 2008. IV'08. 12th International Conference*, pages 373–380. IEEE, 2008.
[47] E. R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT, USA, 1986.
[48] E. Vonesh and V. M. Chinchilli. *Linear and nonlinear models for the analysis of repeated measurements*. CRC press, 1996.
[49] C. Ware. Color sequences for univariate maps: Theory, experiments and principles. *Computer Graphics and Applications, IEEE*, 8(5):41–49, 1988.
[50] C. Ware, T. L. Turton, F. Samsel, R. Bujack, D. H. Rogers, K. Lawonn, N. Smit, and D. Cunningham. Evaluating the perceptual uniformity of color sequences for feature discrimination. In *EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3). The Eurographics Association*, 2017.
[51] J. Webster, P. Kay, and M. A. Webster. Perceiving the average hue of color arrays. *JOSA A*, 31(4):A283–A292, 2014.

Aritra Dasgupta is an Assistant Professor at the Ying Wu College of Computing in New Jersey Institute of Technology (NJIT). Before joining NJIT, he was a Senior Research Scientist at the Visual Analytics group in Pacific Northwest National Laboratory (2015-2018). He received his Ph.D. in Computing and Information Systems in 2012 from the University of North Carolina at Charlotte, and was a postdoctoral research scholar at New York University (2012-2015). His main research interests are information visualization, visual analytics, and human-data interaction.

Jorge Poco is an Associate Professor at the School of Applied Mathematics at Fundação Getulio Vargas in Brazil. He received his Ph.D. in Computer Science in 2015 from New York University, his M.Sc. in Computer Science in 2010 from the University of São Paulo (Brazil), and his B.E. in System Engineering in 2008 from National University of San Agustín (Peru). His research interests are data visualization, visual analytics, and data science.

Bernice Rogowitz Bernice Rogowitz received her PhD in Experimental Psychology from Columbia University. After a fellowship at Harvard, she worked as a scientist and research manager at IBM Research, and currently runs a research and consulting practice called Visual Perspectives. Her research explores perceptual and cognitive topics related to the representation and exploration of data, including color and color scales, image semantics, and haptic interfaces, grounded real-world problems in science, medicine and finance. She is an instructor at Columbia University, where she designed the Data Visualization course for the Masters Program in Applied Analytics and is the founding co-Editor-in-Chief of the new Journal of Perceptual Imaging.

Kyungsik Han is an Assistant Professor in the Department of Software and Computer Engineering at the Ajou University, Suwon, South Korea. He received his Ph.D. in information sciences and technology from the Pennsylvania State University, USA (2015). He was a research staff member at the Pacific Northwest National Laboratory, USA (2015-2017). His research interests include human-computer interaction, social media analysis, and social computing.



Enrico Bertini is an Associate Professor in the Department of Computer Science and Engineering at NYU Tandon School of Engineering. His research focuses on studying and developing interactive visual analytics methods to help scientists, researchers and domain experts, make sense of complex data sets and models. Professor Bertini earned his PhD degree in Computer Engineering at Sapienza University of Rome in Italy.

Cláudio T. Silva is a professor of computer science and engineering and data science at NYU. His research lies in the areas of visualization, data science, graphics, and geometric computing, and recently has focused on urban and sports data. He is an IEEE Fellow, and was the recipient of the 2014 IEEE Visualization Technical Achievement Award. He has served as the chair of the IEEE Visualization & Graphics Technical Committee since 2016.