A MIXED-METHODS EXPLORATION OF COLOR ENCODINGS FOR TWO-DIMENSIONAL SCALAR FIELDS

by

Philip Samuel Quinan III

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computing

School of Computing

The University of Utah

May 2020

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Philip Samuel Quinan III**
has been approved by the following supervisory committee members:

Miriah Meyer ,	Chair(s)	**12/03/19** Date Approved
Christopher R. Johnson ,	Member	**12/03/19** Date Approved
Robert M. Kirby ,	Member	**12/03/19** Date Approved
Alexander Lex ,	Member	**12/03/19** Date Approved
Sarah H. Creem-Regehr ,	Member	**12/03/19** Date Approved

by **Ross T. Whitaker** , Chair/Dean of
the Department/College/School of **Computing**
and by **David B. Kieda** , Dean of The Graduate School.

ABSTRACT

Over the last several decades, visualization research has generated an extensive body of literature on the subject of color encodings for two-dimensional (2D) scalar data. When working with experts from other domains, however, visualization designers commonly encounter real-world practices that conflict with visualization guidance, especially where color is concerned. This dissertation focuses on the problem of empirically interrogating conflicting domain guidance related to color utilization in 2D scalar visualizations through a mixed-methods research program. As a primary contribution, this work uses guidance conflicts observed during a formative, qualitative design study conducted with meteorologists to motivate subsequent quantitative studies that revise and refine longstanding visualization guidance regarding color usage. As a secondary contribution, we also briefly touch on the subjects of ecological validity, generalizability, and experimental control.

The dissertation begins with two pieces of formative work. The first is our meteorological design study, which provides a grounded understanding of color usage in meteorological practice. The second is a pilot study aimed at improving ecological validity in assessing the impact of uncertainty visualizations in meteorological forecasting. This formative work subsequently acts as a lens for two additional studies investigating specific color guidance conflicts between visualization and other domains: a web-based study looking into the impact of discretized color on meteorological tasks, and a laboratory study assessing the potential role of hue-based banding in implicitly discretizing 2D rainbow color-mapped datasets. Lastly, we discuss the broader implications of our mixed-methods research program and outline potential directions for future work.

To my parents,

Phil and Cyndy Quinan.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to extend my deepest thanks to my advisor and supervisory committee: Miriah Meyer, Sarah H. Creem-Regehr, Christopher R. Johnson, Robert M. Kirby, and Alexander Lex. Without their involvement, feedback, and patience this dissertation would not exist. Miriah taught me how to conduct visualization research and provided me with the opportunity to pursue the projects described in this dissertation. I will be eternally grateful for her guidance and direction.

Along with Miriah, I would like to extend a special thanks to Lace M. Padilla and Sarah H. Creem-Regehr as my primary co-authors. Sarah and Lace were integral to growing my understanding of experimental design and quantitative analysis. Their influence has left an indelible mark on this dissertation and my identity as a researcher.

I am grateful for the feedback and support of the members of the Vis Design Lab at the University of Utah, particularly Sean McKenna, Alex Bigelow, Ethan Kerzner, Nina McCurdy, Carolina Nobre, Jimmy Moore, and Jen Rodgers. Additionally, I would like to thank Varun Shankar, Mark Kim, Konstantin Shkurko, Eleanor Wong, Elena Vasiou, and Fiona Weathersby for their friendship and unfailing support, especially during the darker moments of graduate school. All of the individuals mentioned in this paragraph played a critical role in my completing this degree.

I also need to thank my parents, siblings, and extended family for their encouragement and support throughout the course of my education. My parents also deserve special thanks for their assistance in pushing this dissertation across the finish line.

The research contained in this dissertation benefited from the world-class directors, staff, and facilities at the SCI Institute. It also benefitted from the feedback of the many colleagues on the National Science Foundation (NSF) grant IIS-1212806, including William B. Thompson, Ross T. Whitaker, Robert M. Kirby, Donald H. House, Mary Hegarty, Mahsa Mirzargar, Mukund Raj, Ian Ruginski, and Le Liu.

This dissertation is also indebted to the collaborators, participants, and research assis-

tants involved in the studies that we conducted, especially: Shelby Law, Chuck Maxwell, Lance Avey, Jim Steenburgh, Trevor Alcott, and the members of the Ute Weather Center.

This dissertation extends published work that we have permission to reuse. ©2016 IEEE. Reprinted, with permission from Philip S. Quinan and Miriah Meyer; "Visually Comparing Weather Features in Forecasts;" *IEEE Transactions on Visualization and Computer Graphics*; Jan. 2016 [145]. ©2017 IEEE. Reprinted, with permission from Lace M. Padilla, Philip S. Quinan, Miriah Meyer, and Sarah H. Creem-Regehr; "Evaluating the Impact of Binning 2D Scalar Fields;" *IEEE Transactions on Visualization and Computer Graphics*; Jan. 2017 [137]. ©2019 John Wiley and Sons. Reprinted, with permission from Philip S. Quinan, Lace M. Padilla, Sarah H. Creem-Regehr, and Miriah Meyer; "Examining Implicit Discretization in Spectral Schemes;" *Computer Graphics Forum*; Jul. 2019 [146].

This dissertation also reprints figures courtesy of the United States (U.S.) Environmental Protection Agency (EPA), the U.S. Census Bureau, the U.S. National Oceanic and Atmospheric Association (NOAA) National Centers for Environmental Information (NCEI), the NOAA National Center for Environmental Protection (NCEP), NCEP's Environmental Modeling Center, the U.S. National Weather Service (NWS) Storm Prediction Center (SPC), the University Corporation for Atmospheric Research (UCAR), the U.S. National Center for Atmospheric Research (NCAR) and Unisys Weather.

Funding for this dissertation was provided, in part, by the National Science Foundation via NSF grant IIS-1212806.

CHAPTER 1

INTRODUCTION

This dissertation considers the problem of empirically interrogating conflicts between existing visualization guidance and widespread domain practices related to the use of color in visualizing two-dimensional (2D) scalar data. Specifically, the work presented in this dissertation addresses various questions raised during an initial two-year design study conducted with meteorologists [145]. Observations from that design study spawned a variety of projects, including a pilot study focused on maintaining ecological validity when assessing the impact of visualizations on expert decision-making [147], an online user study designed to assess the impact of using binned color maps to visualize continuous 2D scalar fields across a variety of tasks commonly performed by meteorologists [137], and a laboratory study assessing the potential role of hue-based banding in implicitly discretizing 2D rainbow color-mapped datasets [146]. Together, these projects represent a mixed-methods approach to research, where a qualitatively grounded understanding of how visualizations are used in practice acts as the foundation for quantitative experimental research. This dissertation will outline how this mixed-method approach led us to challenge implicit upstream assumptions, enabling the refinement of multiple long-standing visualization guidelines regarding color usage, including core aspects of the visualization community's understanding of why rainbow color maps are problematic. Additionally, it will discuss the broader implications of these research findings, reflecting on both how guidance conflicts can motivate empirical research and the importance of better reconciling qualitative and quantitative research results.

1.1 Motivations

This dissertation addresses guidance conflicts related to the usage of color in visualizing 2D scalar fields. Although the use of color is both more complicated and, in some cases, less effective than the use of other encoding channels (e.g., compared to position for

encoding quantitative data) [30],[126], it nevertheless remains a core aspect of visualization in a variety of domains [13], [18], [25], [38], [89], [145], [214]. Indeed, color use is especially prevalent when visualizing data such as scalar fields, where the *spatialization* of the data is predetermined [24], [153]. Over the years, the continued prevalence of color encodings has led to a large body of work examining how visualizations can employ color more effectively. Many of the visualization community's findings, however, are not well known in other scientific domains, leading to real-world practices that sometimes conflict with visualization guidance.

This dissertation seeks to understand the real-world impacts of such guidance conflicts using a mixed-methods approach. Mixed-methods research, which is becoming increasingly commonplace in the social, behavioral, and health sciences, represents a third major research paradigm that combines aspects of the already established quantitative and qualitative research paradigms [40], [181]. Driven by a combination of philosophical pragmatism and practice-driven need, mixed methods emphasize that both quantitative and qualitative viewpoints are useful for addressing questions, either in the course of a single study or over the course of a multiphase program of inquiry [79]. Mixed-methods approaches offer several advantages, including producing an enhanced understanding of results through the use of complementary methods, avoiding biases intrinsic to single-method approaches, and enabling the discovery of paradoxes that lead to a reframing of underlying research questions [40], [59]. This last advantage, in particular, drove the use of mixed methods in this dissertation.

The work included in this dissertation is grounded in a formative design study focused on improving visualization tools for meteorologists in decision support roles. During this design study, we encountered a variety of guidance conflicts between the domains of meteorology and visualization that could not be explained by insufficient education or a lack of adequate tools [14], [120]. Instead, these conflicts suggested that existing visualization guidance may not have appropriately accounted for these design choices.

The meteorologists we worked with *always* visualized continuous 2D scalar fields with discrete encodings: sets of isocontours, discretized color scales, or sets of glyphs based on sampled locations. This behavior arguably violates the visualization principle of *expressiveness*, which states that a visual encoding should express all of the relevant relationships

in the data and only the relevant relationships in the data [107], [126]. Additionally, like experts in many other domains [13], [25], [38], [89], [214], meteorologists maintain strong conventions that rely on rainbow color maps, even in the face of the widely enumerated deficiencies of spectral schemes [8], [13], [14], [37], [83], [100], [119], [194]. In both cases, however, there is conflicting literature that appears to support the existing meteorological conventions. Work in cartography and geographic information systems (GIS), for example, suggests that contouring aids the interpretation of continuous features [62], [76]. Similarly, there is evidence to support the idea that segmented or discretized rainbow color maps may represent a judicious use of color [18].

This dissertation uses the conflicts between visualization guidelines and meteorological practice to motivate the reexamination of foundational research regarding the use of color for visualization. Based on a deep understanding of how meteorologists utilize color in practice, which we developed during this formative design study, we conducted a series of quantitative studies designed to shed new light on the guidance conflicts that we encountered. This dissertation provides new knowledge by refining multiple aspects of the visualization community's understanding of how color can and should be used to encode 2D scalar data. It is grounded in an extensive literature review of the principles of effective color-scale design, derived from our knowledge of both human perception and visualization practice. Due to its grounding in the real-world domain practices in meteorology, this dissertation also touches upon questions related to experimental control, ecological validity, and the generalizability of research results both to and from highly specialized, real-world contexts.

1.2 Contributions

The primary contribution of this dissertation is the revision and refinement of long-standing visualization guidelines regarding the use of explicitly and implicitly discretized color maps for encoding 2D scalar fields. This contribution is framed by a formative design study, which provides insight into how color is commonly employed to visualize 2D scalar fields in meteorological forecasting processes [145]. Subsequently, we cover how guidance conflicts highlighted by this design study led to novel insights for our understanding of color. One such conflict led to revised guidance regarding the role of discretized color in

visualizing 2D scalar fields for a variety of tasks [137]. Another led to the refinement of our understanding of the implicit discretization in rainbow color maps [146].

This dissertation also includes two secondary contributions. We discuss the trade-offs among experimental control, ecological validity, and the generalizability of research results through the lens of the various pieces of work comprising the dissertation: our meteorological design study [145], a pilot study aimed at improving ecological validity in assessing the impact of uncertainty visualizations in meteorological forecasting [147], a web-based study looking at the impact of discretization across routine meteorological tasks [137], and a laboratory-based exploratory study designed to better understand the implicit discretization caused by popular spectral schemes [146]. Additionally, since this work collectively constitutes a mixed-methods research program, we discuss how mixed-methods approaches can efficaciously reconcile qualitative and quantitative visualization research results. In particular, we believe that this dissertation illustrates that using qualitative contributions from design studies to inspire and contextualize quantitative experimental research it is not only possible but arguably advantageous.

1.3 Dissertation Structure

This dissertation collectively summarizes the requisite background for discussing its contributions and then presents the individual pieces of formative and explanatory work comprising it as an emergent multiphase mixed-methods research program. In Chapter 2, we provide an overview of the color usage guidelines for 2D scalar fields found across the research domains of visualization, cartography, and geospatial information systems (GIS); and we briefly outline both the origins of the mixed-methods research paradigm and how this paradigm relates to qualitative and quantitative research. Chapter 3 discusses the formative qualitative work for this dissertation, detailing our meteorological design study and what that design study taught us about the real-world practices of meteorologists regarding color use for visualization [145]. Chapters 4 and 5 then outline two pieces of explanatory follow-up quantitative work: a pilot study aimed at increasing ecological validity in assessing the impact of uncertainty visualizations on the meteorological forecasting process [147] and a web-based study we designed and conducted looking into the impact of discretized color on meteorological tasks. In Chapter 6, we then discuss our

final piece of explanatory work: an exploratory laboratory study conducted to help us understand the role of implicit discretization or *hue banding* in our perception of rainbow color-mapped data [146]. Finally, the dissertation concludes in Chapter 7, where we summarize our contributions, reflect on various insights garnered from our mixed-methods research program as a whole, and outline potential directions for future work.

CHAPTER 2

RELATED WORK

This dissertation represents a mixed-methods research program focused on empirically interrogating conflicts between existing visualization guidance and widespread domain practices related to the use of color in visualizing 2D scalar fields. It is grounded in an understanding of color-usage practices in operational meteorological visualizations, the development of which we discuss in Chapter 3.

This chapter summarizes the related work for this research program, loosely dividing it across two areas: our choice of subject and our choice of methodology. The primary body of related work for this dissertation, discussed in Section 2.1, centers around the use of color for visualizing scalar fields. Section 2.2, on the other hand, additionally situates this dissertation within the body of mixed-methods research.

2.1 Color in Visualization

In addressing the subject of color encodings for 2D scalar fields, this dissertation builds on an extensive body of work related to the use of color in visualizing data. To begin, Section 2.1.1 outlines how the domains of visualization and cartography respectively use color to represent 2D scalar data. Section 2.1.2 then discusses the general principles for effectively designing discrete and continuous color scales. Section 2.1.3 highlights research conducted regarding the operationalization of the visualization community's color guidelines, and Section 2.1.4 reviews the comparatively limited research that has attempted to validate those guidelines.

2.1.1 Color Encodings For Scalar Fields

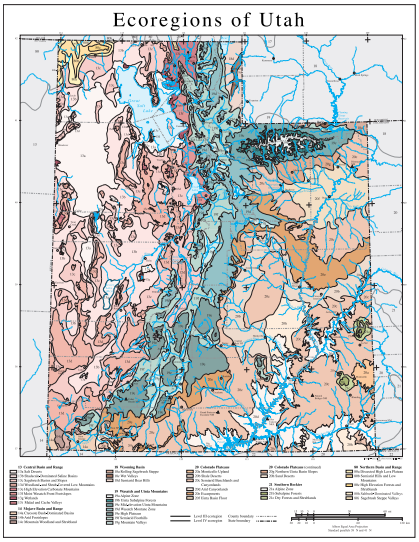
A *scalar field* is a univariate mapping, where different points in space take different values associated with some nonspatial attribute [126]. Scalar fields represent phenomena that occur continuously over space and are commonly found in a diverse set of scientific

domains, including geography (e.g., digital elevation maps), atmospheric sciences (e.g., temperature distributions), medicine (e.g., computed tomography scans), and more. For spatial datasets, where spatial position acts as a key associated with different values for some nonspatial attribute, color remains one of the most widely used and accepted ways of conveying value information [24]. When used properly, color is extremely powerful, acting as a simplifying and clarifying agent in communicating the intended information [41]. Although this statement holds true for the fields of cartography and visualization, how color is used in those respective domains looks markedly different.

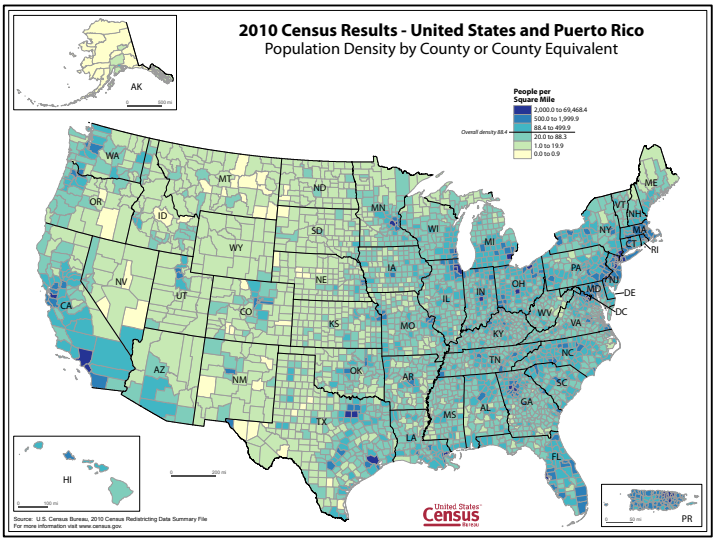
Cartographic map use is commonly separated into two distinct realms: *private* and *public* [41]. The private realm of cartographic map use involves *visual thinking* with maps and includes visualization activities, such as sense-making and data exploration, that help the cartographer understand the relationships in their data through the use of unencumbered and unstructured symbolization [106]. The public realm of cartographic map use, on the other hand, focuses on *visual communication*, emphasizing the use of abstraction (e.g., classification) and structured symbolization to reduce complexity and organize information for presentation to others [41], [106]. In map design, the questions of whether and how to classify data are often core to what the cartographer is trying to communicate [41]. As a result, the use of discretized or classed color maps remains prevalent in cartography [17], [41], [171], [188].

The cartographic emphasis on classification also leads to fundamentally different map types based on the way that the underlying data are classified, as shown in **Fig. 2.1**. *Chorochromatic* maps, for example, are used to show geographic areas that share the same qualitative data classes [92]. *Choropleth* maps, on the other hand, are used to show how values are aggregated or classified over a set of predefined areal regions or *enumeration units* [92], [171].

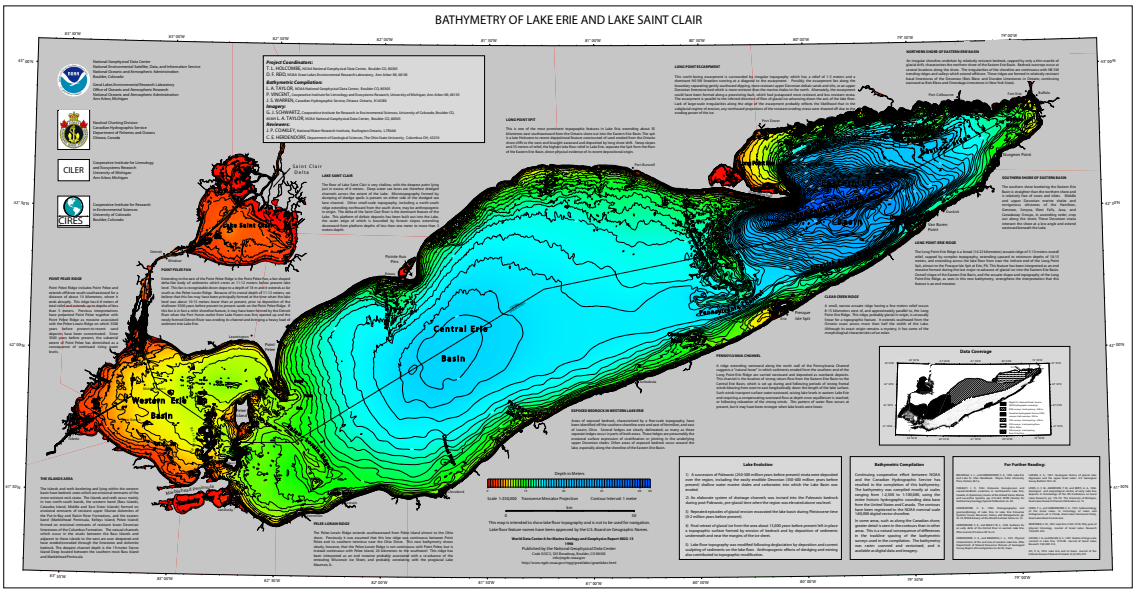
For encoding the sorts of continuous phenomena represented by scalar fields, the relevant map type is the *isarithmic* or contour map. Isarithmic maps use a series of isolines, each individually connecting points of the same value within the map, to provide an impression of a continuous surface [41], [171]. The addition of filled contours, also known as *hypsometric tints*, is relatively common in isarithmic maps, despite the fact that this practice can provide the misleading impression of a stepped surface [41], [171]. An isarithmic map



(a) *Chorochromatic* map of Utah's ecoregions. Reprinted from [208].



(b) *Choropleth* map of population density by county or county equivalent. Reprinted from [189].



(c) *Isarithmic* map (with hypsometric tints) of bathymetry. Reprinted from [131].

Fig. 2.1. Examples of the (a) chorochromatic, (b) choropleth, and (c) isarithmic cartographic map types. These maps are reproduced courtesy of the United States (U.S.) Environmental Protection Agency (EPA), the U.S. Census Bureau, and the U.S. National Oceanic and Atmospheric Association's (NOAA's) National Centers for Environmental Information (NCEI), respectively.

with hypsometric tints is, in effect, a discretized or segmented color encoding of a 2D scalar field. Continuous color encodings technically exist in cartography (e.g., *continuous-tone maps* [93] and *unclassed choropleths* [33]), but they are not commonly employed.

In contrast, the field of visualization tends to focus on sense-making and data exploration tasks, where understanding a wide variety of relationships in the data becomes important [126],[210]. This emphasis motivates the foundational visualization design principle of *expressiveness*, which states that a visual encoding should express all of the relevant the relationships in the data, and only the relevant relationships in the data [107], [126]. For color encodings of continuous phenomena, such as those represented by 2D scalar fields, a strict reading of the expressiveness principle implies that a continuous color map should be used. Admittedly, portions of the visualization community advocate for the more nuanced claim that "significant" data differences need to be preserved by the color map [152], where the notion of significance depends on the data and tasks. Regardless of whether the data are truly continuous or quantized in some meaningful way, however, the literature generally recognizes that the mapping from data to colors itself should be continuous [95], [155]. Moreover, other work specifies that for *isomorphic* representations of continuous data, where the goal is to faithfully reflect the structure in the data, the visual dimension must appear continuous to the user [8]. As a result, continuous color maps are far more prevalent in visualization than in cartography.

Bergman et al. [8] provide one of the only explicit discussions about encoding continuous data discretely in visualizations. They argue that using a discretized color map on continuous data is useful for *segmentation* tasks, where the goal is explicitly to show multiple features. This goal is precisely the aim of cartographic abstraction. Similar to other work in cartography [61], recent work by Görtler et al. investigates the potential benefits of further applying discretization to the spatialization of 2D scalar fields via *stippling* [63]. Although stippling represents an interesting research avenue, this dissertation focuses specifically on color encodings for continuous spatial representations of scalar fields. Notably, however, we also see an emphasis on segmented color encodings for continuous data in the literature on transfer functions.

Transfer function design explores methods and techniques to specify different optical properties, including both color and opacity, for different portions of volumetric datasets.

Transfer functions are predominantly used to classify materials based on features in three-dimensional (3D) scalar fields [49]. The transfer function's implicit role as a classifier stems from necessity. In a 3D scalar field, it is simply not possible to show "all the relevant relationships in the data"; some data relationships will become occluded. These classifications, however, are time-consuming to generate manually [139]. This reality has led to a variety of attempts to automate or optimize data classification based on the underlying data features [54], [85], [193]. These efforts parallel work in cartography focused on generating contour sets driven by the features in 2D scalar fields [62]. We also see parallels in the methods associated with representing either classification errors or uncertainty information within both the transfer function [104] and cartographic [44] literature.

As this dissertation aims to reevaluate certain aspects of the existing guidelines regarding color usage in visualizations of 2D scalar fields, we focus our remaining discussion on the literature related to the creation and evaluation of both discrete and continuous color maps.

2.1.2 Principles of Effective Color-Scale Design

When used effectively, color is is both compelling and advantageous. It acts as a simplifying and clarifying agent in communicating the intended information [41]. Additionally, as Bertin notes, "color exercises an undeniable *psychological attraction*," creating richer experiences that both capture and hold attention and improve information retention [10, p. 91]. When used ineffectively, however, color is known to obscure data and even mislead users [157].

For the most part, the notion of effective color usage traces back to a core set of underlying principles. Cartographers initially developed many of these principles [17], [171], but research in the domains of perception, design, and visualization has, since, refined them [155], [158], [195]. The language used to describe these principles, however, has not always been consistent [24].

As Bujack et al. point out, the literature regarding color guidance not only has authors using the same term to refer to different color-map properties but also has authors using different terms to refer to the same concept [24]. Trumbo originally proposed *order* and *separation* as standard rules in the construction of both univariate and bivariate maps [188].

Trumbo's principle of order states that if the underlying data values have an order, then the colors used to represent those values should be perceived as ordered. Similarly, his principle of separation declares that any important differences in data values should be perceived as different. Around the same time, Pizer argued that color maps should be perceptually linearized so that equal changes in data value are perceived as equal changes in the color map [140]. Pizer's notion of *equal changes*, however, is logically implied by Trumbo's principles of order and separation, and vice versa. Both order and separation are preserved if equal differences from distinct values are perceived as equivalent in distinct portions of the color map. Notably, Levkowitz and Herman also proposed a similar set of principles (*order*; *uniformity and representative distance*; and *no boundaries*), which they used to devise a linearized optimal color scale [95]. Levkowitz and Herman's linearization, however, enforces a straight line in color space, whereas Pizer's linearization establishes equal perceptual distances between colors.

Some of these differences stem from the specific contexts for which these guidelines were created. Compare, for example, the contexts surrounding Trumbo's principle of *separation* and Pizer's notion of *equal values*. Trumbo made his recommendations for map design [188], which predominantly focuses on communicating abstracted relationships [41]. The nature of abstraction emphasizes preserving and communicating important differences. Pizer, in contrast, designed his recommendations for a visualization context that emphasizes extracting patterns and relationships. Separating signal from noise within data, however, requires searching for similar values.

Given the redundancies and ambiguities present in the existing language surrounding color-scale design, Bujack et al. propose a new standardized taxonomy for the core principles of color-scale design based mainly on mathematical definitions [24]. Bujack et al. summarize the perceptual rules associated with effective color-map design according to: 1) local and global discriminative power; 2) local and global uniformity; 3) local and global legend-based order; 4) local and global intuitive order; and 5) smoothness. In this taxonomy, *discriminative power* relates to color distinguishability, *uniformity* relates to the perceived distances between colors, *legend-based order* means that the legend has a logical order, and *intuitive order* means the perceived order does not require looking at the legend [24]. *Smoothness* is the only concept in the taxonomy without a mathematical defi-

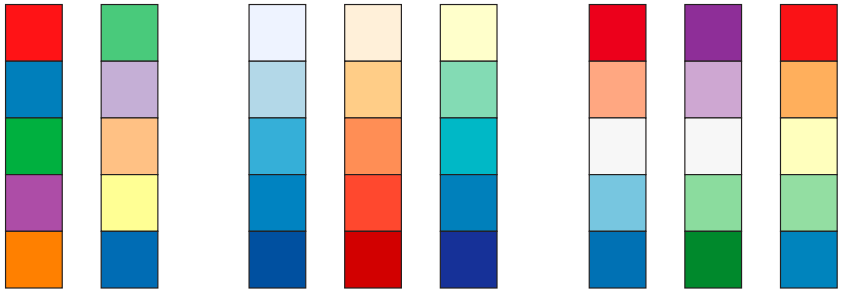
nition [24], but the basic idea of smoothness is that color maps should not have artificially perceived color boundaries or hue bands.

2.1.3 Enabling Effective Color Usage

Over the years, the principles of effective color usage have become the backbone for a variety of guidelines, techniques, and systems designed to help users employ color more effectively in visualizations [169], [213]. Several common guidelines center around specific data conceptualizations, resulting in distinct behaviors for encoding *categorical* (i.e., *qualitative*), *sequential*, and *diverging* data [8], [17], [119], [126], [196], [212]. It is commonly recommended, for example, that categorical or qualitative data are encoded using differing hues with similar levels lightness [17], [19], [126]. For encoding quantitative data, on the other hand, a sequential or diverging color scheme is recommended depending on whether the data range has a critical midpoint of interest [8], [17], [119]. Because the human visual system is incredibly sensitive to changes in lightness [196], it is most often recommended that both sequential and diverging color schemes maintain ordered steps in lightness [17], [19], [126] (c.f., Bergman et al. argue that ordered changes in saturation are more appropriate for data with predominantly low-spatial-frequency information [8]). These guidelines lie at the core of many of the proposed techniques and generative systems for creating color maps [8], [57], [119], [203], [212].

A significant amount of work has specifically focused on helping users choose effective discrete or segmented color palettes. Notable examples include ColorBrewer.org, which suggests a set of expertly designed color palettes based on the user's needs and data characteristics [66] (see **Fig. 2.2** for examples); a technique, outlined by Wijffelaars et al., for automatically generating a larger set of "Brewer-like" palettes using intuitive parameters [203]; Colorgorical, a tool that automatically generates discriminable and aesthetically preferable categorical color palettes by combining perceptual optimizations with user design constraints [57]; and an interactive constraint-based system for palette-space explorations created by Mellado et al. [114].

A variety of publications have also focused on improving or optimizing discrete color palettes. Both Healy [67] and Heer and Stone [68] leverage the notion of color categories to create color palettes that are more differentiable. Lin et al. [97] and Setlur and Stone [166]



(a) qualitative (b) sequential (c) diverging

Fig. 2.2. Example ColorBrewer [66] palettes illustrating Brewer's design guidelines[1] [17], [19] for a) categorical, b) sequential, and c) diverging color scales.

introduce techniques for taking advantage of semantic resonance for generating palettes by selecting colors based on their appropriateness for specific categories. Recent work by Bartram et al. shows how color and palette properties can be manipulated to achieve *affective* expressiveness [7]. Additionally, Walden et al. further expand the notion of a discretized color palette by introducing an adaptive, hierarchical color-mapping technique that interactively alters the color map based on semantic zooming [192].

Even though discrete color palettes are sometimes used to visualize 2D scalar fields in the visualization community (for example, when dealing with ordinal data or segmentation tasks [8]), visualization practitioners generally use continuous color scales. Several techniques and systems have explicitly focused on helping users select or create appropriate continuous color scales. The PRAVDA color system created by Bergman et al., for example, uses a similar set of guidelines to those proposed by Brewer [17] in order to create a recommender system for color maps based not only on data type, but also on the spatial frequency of the data and users' tasks [8]. Ken Moreland created an algorithm for generating perceptually uniform diverging color maps that largely follows Brewer's guidelines and that he used to produce a *cool-warm* color map specifically designed for visualizing scalar fields in scientific visualizations [119]. Further, Yoo et al. present a novel system for reassigning continuous color scales in existing visualizations of 2D scalar fields using an order-preserving method for decolorization [211].

[1] Brewer's design guidelines [17] represent the oldest reference for many of the guidelines now commonly

Because the underlying principles for effective color usage are the same, a handful of work applies equally to both discrete and continuous color maps. Tominski et al., for example, developed recommendations for task-driven color-scale design that apply to the creation of both discrete and continuous color scales [186]. Additionally, Mittelstädt et al. created a methodology and tool to assist users in designing continuous and discrete color maps for combined analysis tasks [117]. A series of papers have also focused on using engineering models for approximating color differences to adapt both discrete color palettes and continuous color scales to account for web-based viewing conditions [180], color-size effects [176], and common visualization mark types [179].

2.1.4 Evaluating Color Usage

A variety of empirical methods have been proposed to assist in the process of evaluating or improving color-map quality. In the "Which Blair Project," Rogowitz et al. introduced a quick visual method for evaluating the luminance monotonicity of a given color map [158]. Kindlmann et al. extended this idea to allow for luminance matching on uncalibrated monitors and showed how their method could be used to generate new color maps with desired luminance properties [86]. Additionally, Ware et al. recently introduced a method for evaluating a color map's capacity for feature resolution and ran a crowd-sourced study to assess discriminative power and uniformity within various commonly used color maps [194], [198].

A handful of studies have also expressly focused on evaluating color-mapping choices. Brewer empirically evaluated several of her proposed schemes using choropleth maps [18], [20], although the extent to which the results generalize either to visualizations of scalar fields or to the use of continuous color maps is not clear. Ware ran a set of experiments designed to assess the impact of different color maps on both metric reading and form comprehension tasks in 2D scalar fields, finding evidence to support the idea that a color map can reduce simultaneous contrast effects by varying nonmonotonically in at least one

---

considered good practices by the visualization community [119], [126]. Although Brewer specifically outlined her guidelines in relation to discrete color scales [17], soon after, Bergman et al. proposed similar guidelines for continuous color scales [8]. Interestingly, the conflation of Brewer's guidance with common visualization practice has led to Brewer's hand-designed palettes [17], [66] being used as a basis for "good" continuous color scales [13], [100], [150], under the *untested* assumption that the color scales' desired properties are maintained through linear interpolations in the RGB color space.

color-opponent visual channel [195]. Kalvin et al. conducted an experiment evaluating people's ability to detect modulations in luminance, hue, and saturation across different spatial frequencies and color spaces [83]. Their results suggest that luminance variation is essential for conveying high-spatial-frequency information [83]. Additionally, Borkin et al. evaluated the efficacy of a rainbow color map compared to a red-black diverging color scheme in both 2D and 3D visualizations using a real-world, medical-diagnostic task [13].

In the last couple of years, there has been a resurgence of work evaluating color maps. Reda et al. recently ran a set of crowd-sourced experiments to determine how the perception of 2D scalar fields is affected by color-map characteristics and spatial frequency [150]. Liu and Heer recently conducted experiments designed to assess the impact of single and multihue color schemes on relative distance judgments within various color maps [100]. Also, Dasgupta et al. recently supervised a study assessing the impact of both luminance monotonicity and hue banding on climate modeling tasks [37].

Collectively, the results of these experiments support many of the visualization community's existing guidelines. Liu and Heer's results [100], for example, further support the recommendation to employ color maps that both increase monotonically in luminance and cycle through a range of hues [126],[195]. Not all of the results, however, are consistent with prior visualization guidance. Reda et al., for example, found no differences between hue-, saturation-, and luminance-varying ramps for low-spatial-frequency data [150], contrary to longstanding guidance [8], [83], [160]. Inconsistencies like this underscore the fact that we still do not fully understand how people perceive and use color. This dissertation further refines our understanding of color perception and color usage by explicitly investigating conflicts between visualization guidance and real-world practices.

2.2 An Introduction to Mixed Methods

Whereas the previous section provided context regarding the subject of this dissertation, here, we focus on methodology. This section provides a primer for mixed-methods research and outlines how the work presented in this dissertation fits into the mixed-methods paradigm.

Mixed methods are increasingly considered a third major paradigm for research in the social and health sciences, in addition to the longstanding quantitative and qualitative

research paradigms. Over the years, diverse viewpoints regarding the definition of mixed methods have led to inconsistencies and variation in the way that scholars conceptualize mixed methods [32], [40]. Although there is not currently a single accepted definition of mixed-methods research, a few accepted criteria have been deemed important in designing and conducting a mixed-methods study or research program [32], [79].

The first criterion for mixed-methods research is that it involves the collection and analysis of both qualitative and quantitative data, and specifies a sequencing of both the qualitative and quantitative elements across the data collection and analysis [32]. This involvement could mean concurrently collecting both types of data, merging them into the same analysis; sequentially collecting each type of data, allowing one set of results to build upon the other; or even embedding one set of methods within the other. Additionally, mixed methods may prioritize one or both types of data, depending on the research emphases [32]. They are always, however, explicit about the manner in which the quantitative and qualitative procedures relate [40]. Mixed methods may be employed in either a single study or in multiple phases of a program of study [32], [79]. Finally, the procedures may or may not be framed within a particular philosophical world-view or theoretic lens, such as philosophical pragmatism [32], [40].

Mixed-methods approaches are commonly used to improve accuracy, to produce a more complete picture by combining information from complementary kinds of data or sources, as a means of avoiding biases intrinsic to single method approaches, or as a way of developing one's analysis and building upon initial findings using contrasting kinds of data or methods [40], [59]. According to Creswell and Plano Clark:

> The research problems best suited for mixed methods are those in which one data source may be insufficient, results need to be explained, exploratory findings need to be generalized, a second method is needed to enhance a primary method, a theoretical stance needs to be employed, [or] an overall research objective can be best addressed with multiple phases or projects [32].

Because researchers are allowed to use all of the data collection tools available to them, they can begin to answer questions that they could not or would not have answered through the use of either quantitative and qualitative methods alone.

Mixed methods also encourage the use of multiple world views, paradigms, and ontologies [32]. Quantitative research builds heavily on a *postpositivist* ontology, where there

is a shared sense of truth or reality and where knowledge is objective. This foundation leads to a focus on top-down or *deductive* methodologies, where theory leads to a hypothesis, which leads to data collection, which finally leads to the confirmation or rejection of that hypothesis. Qualitative research, on the other hand, is often based on a *constructivist* ontology, where all knowledge is inherently subjective. Because of this different basis, qualitative methodologies tend to be bottom-up or *inductive*, where participants' responses are used to build broader themes and where theory is subsequently generated to connect those themes. Mixed methods, by comparison, are often considered to be based on a *pragmatic* philosophy, which focuses on examining practical consequences and empirical findings as a way to facilitate understanding and knowledge generation [78]. This underpinning leads to methodological choices oriented toward "what works" in practice.

Mixed-methods research, however, has its own set of challenges. Interested researchers need expertise in both quantitative and qualitative data collection and analysis techniques, as well as an understanding of how those methods provide knowledge through rigor, reliability, validity, experimental control, and generalizability [32]. Mixed methods also have increased demands for time and resources compared to an individual quantitative or qualitative study [32]. The most substantial challenge, however, is that the comparative novelty of mixed methods can make it difficult to convince specific scholarly communities of the benefits of mixed-methods approaches [32]. In healthcare fields, for example, qualitative research results are sometimes accepted only to the extent that they support or fit into the field's predominantly quantitative paradigm [79].

Despite these challenges, however, the mixed-methods paradigm can be useful for visualization research. Mixed methods have been growing in popularity in HCI [190] and have successfully been employed in diverse research projects ranging from human factors and ergonomics research in healthcare [26] to the evaluation of gameplay [96] to studies of technology in the home [115]. Because of this popularity, we are now beginning to see mixed-methods approaches used in visualization research. Fonseca et al. conducted a mixed-methods study to evaluate the motivation and satisfaction of architecture students with interactive and collaborative tools for visualizing 3D architectural models [53]. Also, recent work by Hung and Parsons outlines how a mixed-methods approach was used to assess user engagement with information visualizations [74]. None of this work, how-

ever, provides any real guidance regarding the specific type of mixed-methods program described in this dissertation.

In this dissertation, we outline our use of an emergent, multistudy mixed-methods research program and describe how this mixed-methods program has led to the revision of multiple longstanding guidelines regarding the use of color in visualizing 2D scalar fields. Although we are not claiming this dissertation makes contributions to mixed-methods research, we do believe that there is a clear need for better integration of qualitative and quantitative research within visualization and that the work presented in this dissertation demonstrates that mixed methods are one possible way to address this need.

CHAPTER 3

VISUALLY COMPARING WEATHER FEATURES IN FORECASTS

In Chapter 1, we outlined how the contributions of this dissertation are grounded by a formative design study conducted with meteorologists. Here, we present a detailed overview of that initial design study and its role in the emergent mixed-methods research program that we employ to address guidance conflicts related to the use of color in visualizing 2D scalar fields.

Faced with the problem of quickly extracting relevant information from vast quantities of simulated data, meteorologists primarily rely on visualizations for locating and relating weather features, which they use to make subsequent forecasts. We set out to conduct a design study focused on correcting common visualization challenges in this forecasting processes. During this design study, however, conflicts between visualization and meteorological conventions left us with open questions related to the generalizability of the guidance from both domains. Were these meteorological conventions entrenched artifacts of domain history and culture, or did they, instead, highlight potential gaps in the understanding of visualization researchers? The need to better understand these guidance conflicts became the motivation for this dissertation.

In this chapter, we discuss the original motivations and contributions of our design study; we outline the problem characterization that we developed for meteorological forecasting and how that problem characterization influences the use of color in meteorological visualizations; and, finally, we reflect on how this design study informed the subsequent work detailed in this dissertation.

3.1 Design Study Motivation

A wide variety of domains depend on weather predictions for making critical decisions, such as wildfire response, avalanche prediction, and hurricane evacuation. Mete-

orologists working in these domains make predictions based on numerically simulated forecasts, the outputs of which include many different variables and time points. Furthermore, as with most numerical simulations, various sources of error lead to inherent uncertainty in the resulting forecasts. In order to account for some of this uncertainty, forecasts can use multiple simulations to sample the space of possible outcomes, creating an ensemble of results for each variable and time point.

Because the simulations are large and computationally expensive, only a few big governmental and intergovernmental agencies run and distribute the majority of the forecast simulations. The quantity of resulting data is substantial, on the order of hundreds of gigabytes for a single day's forecasts, which has fostered a prevalence of tools and third-party organizations that create static visualizations of the forecasts for use and distribution. The challenge for the meteorologists who use these visualizations is that these tools and third-party organizations produce visualizations with vastly different visual conventions, many of which go against well-known visualization principles, and seldom offer support for exploring the uncertainty in the simulations.

To better understand these challenges, we conducted a 2-year design study that involved meteorologists in decision-making contexts across a variety of application areas ranging from wildfire prediction to air quality assessment. We identified two common visualization challenges. The first challenge stemmed from meteorologists attempting to mentally integrate information from sets of visualizations with inconsistent and even conflicting visual encodings. The second challenge involved the limited support for working with ensemble forecasts, which at the time included no effective methods for the direct comparison of multiple features across an ensemble.

In addressing these two challenges, this work included several contributions: a characterization of both the problems and data associated with meteorological forecasting, a concise treatment of which we were unable to locate in the existing body of visualization research; a data-driven formulation of **informed, default** encoding choices that integrate existing meteorological conventions with sound visualization principles; and, as a secondary contribution, the extension of state-of-the-art-techniques for visualizing ensembles to enable the direct comparison of multiple meteorological features. We evaluated these contributions with our collaborators by integrating them into WeaVER, a proof-

of-concept system for weather visualization encoding recommendations packaged into an open-source tool for visualizing weather forecasts. As a final contribution, we also reflected on the challenges and pitfalls that we encountered while working in the domain of meteorology.

3.2 Design Process

Over the course of this design study, we worked closely with meteorologists from a variety of domains. Our collaborators included two meteorologists supporting resource allocation decisions for wildfire emergency management in the southwest region of the United States; a meteorologist working with air quality simulations for Utah's state-level regulatory Department of Air Quality; a meteorologist in the Science and Technology Infusion Division of the National Weather Service; and a professor in the Department of Atmospheric Sciences at the University of Utah who specializes in cold-weather mountain climatology and runs a popular public-facing blog about winter weather and ski conditions along the Wasatch Front.

Our design process was heavily influenced both by the nine-stage framework of Sedlmair et al. [165] and by the processes and recommendations for human-centered design in geo-visualization outlined by Lloyd and Dykes [101]. We began with a series of contextual interviews where we observed the daily workflows of several of our collaborators. These formative interviews shaped our initial domain problem characterization. Due to our lack of background knowledge in meteorology, we sought assistance from the Atmospheric Sciences Department at the University of Utah, which led us to several other collaborators. A series of additional interviews with these individuals led to insights that added significantly to our understanding of the tasks and workflows of meteorologists.

We then moved on to a prototyping phase that began with the parallel development of multiple wireframe prototypes. After an internal review, we integrated the designs from our parallel prototyping session into a full-scale paper prototype that we presented to our collaborators for feedback. We then created additional digital prototypes that focused separately on the ideas of informed, default encodings and directly visualizing multiple features across ensembles. We subsequently presented the digital prototypes to our collaborators in three separate feedback sessions and integrated the resulting feedback into

the development of WeaVER, the open-source tool that we created as a full-scale proof of concept for evaluating our proposed ideas, as described in Section 3.7. WeaVER was, in turn, presented to our collaborators for additional tweaks and final evaluation, as we discuss in Section 3.8. We also include an expanded overview of our early wireframe, paper, and digital prototypes in the supplemental materials[2] for this dissertation.

3.3 Foundations in Weather Forecasting

From the outset of our design study, it was clear that many of the meteorological visualizations used in operational forecasting settings employ conventions that conflict with visualization principles. Our collaborators were attempting to integrate information from collections of visualizations from disparate sources that employ conflicting and often ineffective encoding choices, especially regarding color. The reason for this stems, in large part, from the way that meteorological visualizations are used in the forecasting process. In this section, we provide a problem characterization and data abstraction for weather forecasting, based on the tasks and workflows of our collaborators in meteorology. We also discuss the role of visualization in these workflows at the time of the design study, outlining the needs that we identified for improved visualization methods.

3.3.1 Problem Characterization

In general, a *weather forecast* refers to one or more outputs of numerical weather prediction simulations. The simulations are run using various models, parameterizations, input conditions, and boundary conditions in order to simulate different sets of variables over varying geographical extents, predictive time frames, and grid resolutions. To ensure that meteorologists have the most up-to-date information, many of these simulations are run multiple times throughout the day.

There are two main types of forecasts: deterministic forecasts and ensemble forecasts. *Deterministic forecasts* are the output of a single simulation. *Ensemble forecasts*, on the other hand, are a collection of two or more simulation outputs that cover overlapping geographical extents and predictive timeframes, but use different models, different parameteriza-

[2] All supplemental materials for this dissertation have been archived in a Github repository, which can be found at: `https://github.com/samquinan/DissertationSupplement`.

tions, or different initial or boundary conditions. The simulations comprising an ensemble are generally run at lower resolutions than their deterministic counterparts due to limited computational resources. Because each simulation, or *member*, within an ensemble represents a possible forecast outcome, the ensemble as a whole can be thought of as sampling the space of possible forecast outcomes, providing a mechanism for approximating uncertainty in the simulation process. This uncertainty may be due to the propagation of initial error through the simulation, the use of approximations in modeling the underlying physical processes, or numerical error within the computation itself. These various sources of error grow over the predictive time frame, propagating to successively larger scale features, thereby increasing the divergence among the different simulations [170].

Due to the associated computational complexity, many organizations that rely on forecasts have neither the time nor the resources to run weather simulations. Consequently, the majority of the weather forecast simulations used today are run by a small number of large governmental and intergovernmental organizations such as the United States' National Oceanic and Atmospheric Administration (NOAA) and the European Centre for Medium-Range Weather Forecasts (ECMWF) — we collectively refer to these organizations as gatekeepers [165]. For the many meteorologists who rely primarily or solely on the forecasts run and disseminated by these gatekeepers, predefined choices of models, variables, and grids become limiting factors in all weather predictions and decisions.

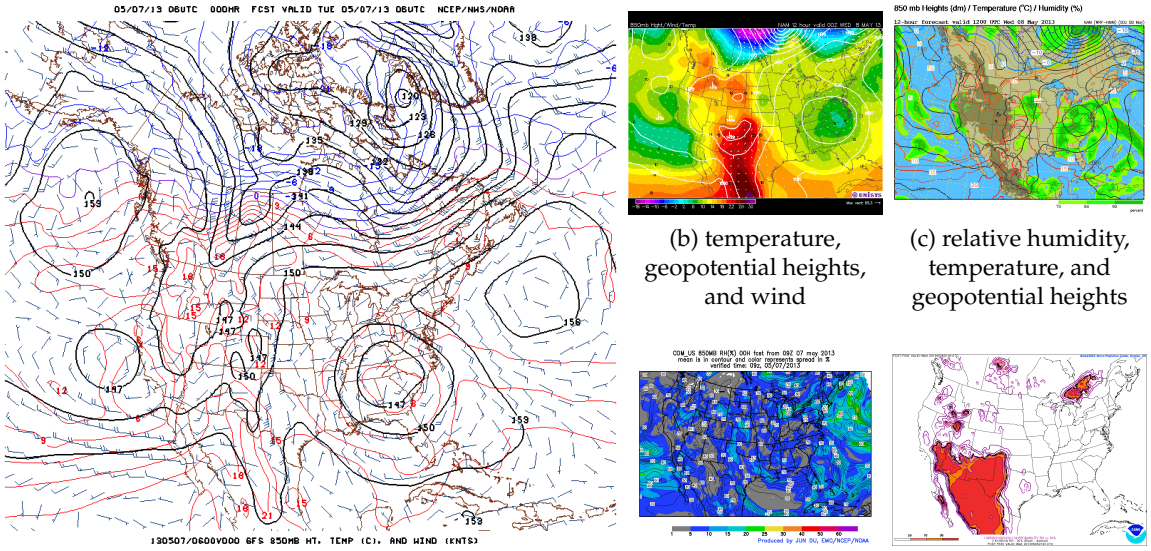
The workflows of most meteorologists center around the tasks of locating and relating specific *features*. Features generally correspond to events, trends, or boundaries: a cold front, a low-pressure system, the freezing point temperature boundary, etc. Meteorologists rarely base their decisions on a single feature; rather, they look for the intersection of multiple features. The threshold for critical wildfire conditions, for example, equates to surface temperatures above 60°F, surface relative humidity under 20%, and sustained surface winds over 20 mph [182]. Meteorologists also rely heavily on experiential knowledge in making judgments based on features. When a simulation's resolution does not accurately resolve mountainous terrain, for example, the resulting forecasts will fail to show major precipitation events; yet one of our collaborators can often still accurately predict those precipitation events. By looking at the same forecasts day after day, meteorologists begin to recognize patterns in how the simulations' biases relate to their specific problems and

begin to account for those biases in their decision-making process.

We observed that meteorologists generally use visualizations in order to understand the big-picture status of a forecast. Many meteorological visualizations, however, fail to present the forecast data in a way that enables good visual comprehension of feature relationships. As we will go on to discuss, this is due to both the problematic usage of visual encodings and unsatisfactory methods for relating features across ensembles. As a result, meteorologists are often forced to look at a large number of visualizations to locate all the features relevant to their current problem, mentally aggregate those features, and then use that aggregated mental model to make their judgments [187]. As one collaborator summarized, *"Forecasters are dealing with a firehose of data, and they need to be able to process it fast."* Pattern recognition plays a critical role in this regard, but only if an individual can look at a consistent set of visualizations over time.

Various tools and systems exist for visualizing forecast data, but many third-party organizations also take the same data, derive additional data products, and release static visualizations over the web. These visualizations are often created using the same or similar tools to those to which meteorologists already have access. In some cases, however, these third-party visualizations provide access to new or experimental research not yet integrated into available forecasting tools, such as derived probabilistic predictions for dry lightning [15]. Other times the data have been postprocessed or bias-corrected to target specific meteorologists' problems, as is the case for a large number of visualizations generated by the Storm Prediction Center in Norman, OK [177]. Sometimes the third-party visualizations are more straightforward to use or more readily available outside the office. Whatever the reason, bookmarks and browser tabs for accessing a combination of these static visualizations factor heavily into the workflows of the majority of our collaborators.

As **Fig. 3.1** illustrates, different variables within a forecast are visualized using a variety of encodings. The visualizations often layer multiple encodings to allow for the comparison of multiple variables and their constituent features of interest. Unfortunately, when showing multiple variables simultaneously, these visualizations often combine encodings in problematic ways, as evidenced by the usage of color in **Fig. 3.1c**. Additionally, the static nature of the visualizations often leads to significant visual clutter, like in **Fig. 3.1a** and **Fig. 3.1d**. An expanded discussion of these issues is presented in Section 3.5.1.



(a) temperature, geopotential heights, and wind

(b) temperature, geopotential heights, and wind

(c) relative humidity, temperature, and geopotential heights

(d) mean and spread

(e) probability

Fig. 3.1. Examples of existing static forecast visualizations that show various meteorological conventions for encoding combinations of variables. Examples (a), (d), and (e) are reproduced courtesy of various subsidiaries of NOAA and the US National Weather Service: the National Center for Environmental Protection (NCEP), NCEP's Environmental Modeling Center, and the Storm Prediction Center, respectively. Example (b) is reproduced courtesy of Unisys Weather, and example (c) is reproduced courtesy of the University of Corporation for Atmospheric Research (UCAR) and US National Center for Atmospheric Research (NCAR).

The notion of visually relating features and feature relationships is further complicated when dealing with ensembles. An ensemble of possible simulated outcomes provides an ensemble of possible instantiations for each feature of interest. Visualization methods for *directly* examining the space of possible features across an ensemble are currently limited to looking at isocontour-based features. Moreover, these visualization methods are ineffective for tasks that involve relating multiple features. Meteorologists consequently examine interactions among multiple features across the ensemble *indirectly*, looking at how the features relate under derived-data transformations. The transformations are usually statistical; examples include averaging the possible outcomes, taking the maximum or minimum possible values, and calculating the probability of an event based on how many ensemble members predict it. These transformations, however, often mask outliers, gradient-based features, and edge cases that are otherwise significant [182]. Further, the resulting derived forecasts often depict physically impossible features and feature relationships.

3.3.2 Data Abstraction

Forecast simulations are multidimensional and multivariate. Each forecast simulation is a time-varying set of height layers, where each height layer is a 2D grid with variable values at every grid point. Each height layer may have tens to hundreds of variables. Ensemble forecasts can analogously be thought of as multivalued, with the set of simulation members defining possible values for a particular variable at a specific point in simulated space-time [143].

Notably, although a set of 2D height layers can be characterized as a single 3D dataset, the meteorologists we spoke with generally do not think about them as such. The simulations do cover a continuous 3D extent, but the height layers are far enough apart that any dependencies between them are considered negligible.

A *field* refers to a particular variable at a particular height: a 2D grid of generally scalar, but possibly vector, values. Our contextual interviews and subsequent conversations indicated that, in practice, a given meteorologist only cares about a subset of around 50 fields out of the hundreds often available within a given forecast. Which 50 fields, however, ranges from meteorologist to meteorologist according to their particular problems and prior experience. Meteorologists additionally incorporate many *derived fields* into their workflows. Derived fields represent a mapping from a set of fields to a single field. This mapping may combine multiple instances of the same field from different simulations, as would be required to calculate the average field over an ensemble. Alternatively, the mapping may combine multiple fields from a single simulation, as in the case of calculating the Haines Index. The Haines Index takes values between two (low risk) and six (high risk) indicating the potential for accelerated wildfire growth based on temperature and dew point differences across various atmospheric height layers [142]. For all the derived fields we discuss, values at different grid locations are treated as entirely independent. This notion of grid-location independence is neither a necessary condition for a derived field nor a reflection of the physical processes being simulated. It is an artifact of how the meteorological community computes derived quantities in practice.

A *feature* refers to a significant artifact, usually an event, trend, or boundary, that can be located within a given field. For scalar fields, features are generally visually represented by isocontours or gradients. For vector fields, features generally correspond to topological

structures such as a source, sink, saddle-point, or closed orbit.

3.4 Related Work in Meteorological Visualization

The related work for this design study loosely falls into two categories: work that addresses the visualization of meteorological data and work that addresses the visualization of ensembles and uncertainty.

3.4.1 Visualizing Weather

A number of software systems have been designed for the visual exploration of meteorological data in both operational and research environments [1], [3], [46], [71], [112], [127], [135], [143], [149], [201]. These systems provide varying levels of user control for selecting which fields to visualize and how to encode them. Some rely entirely on scripting-based interfaces [1], [46], [201], as opposed to offering a graphical UI; some deal solely with gridded information, such as forecasts, and others integrate point-based observational data [46], [201], as well as real-time satellite and radar data [3], [112], [127], [149]. As Table 3.1 illustrates, however, these systems have done very little with either informed, default encodings or the direct visualization of multiple features across an ensemble.

The majority of the aforementioned systems leave encoding choices to users who lack training in visualization principles. So far, initial steps toward addressing this issue have taken orthogonal approaches. Unidata's Integrated Data Viewer (IDV) [127] provides a set of color-map recommendations for different types of fields, although most appear to be spectral and ignore effective visualization practices. The Ensemble Vis framework of Potter et al., on the other hand, reuses a single set of two sequential and two categorical color maps based on visualization principles for all fields of interest [143]. Taking an entirely different tack, the Nocke et al. tool SimEnvVis allows for the creation of rules that assign encodings based on field metadata and user preferences; however, users' lack of training regarding visualization principles remains an issue. Alternatively, Ware and Plumlee propose a set of perceptually motivated design alternatives for encoding weather data [197].

As previously mentioned, various websites also provide users with a range of static forecast visualizations. The Short-Range Ensemble Forecast (SREF) website run by the

Table 3.1. Overview of visualization mechanisms in common meteorological systems.

Tool	2D	3D	Deterministic	Single Feature Direct Ensemble	Multiple Feature Indirect Ensemble	Multiple Feature Direct Ensemble	Movement Toward Informed Defaults
AWIPS [112], [149]	●	●	●	●	●		
Metview/ Magics++ [3]	●		●	●	●		
GrADS [46]	●		●	●	●		
IDV [127]	●	●	●	●	●		■
Vis5D [71]		●	●	▲		▲	
Ensemble Vis [143]	●	●		●	●		●
SimEnvVis [135]	●	●	●		◊		●
FERRET [1]	●	●	●	●	⧫		
GMT [201]	●	●	●		◊		

●: Full support

■ : Colormaps created for specific variables

▲: Small multiple 3D views only

◊: Can theoretically visualize any prederived data using available deterministic mechanisms

⧫: Like ◊, but also supports a data aggregation operation

Storm Prediction Center [177] is one example, but we include a list of others in the supplemental materials archive for this dissertation.

The static visualizations provided on these websites are primarily created using tools or systems like those described at the beginning of this section. Users of these web-based systems, however, exercise even less control over which fields they can look at and no control over how the information is encoded.

3.4.2 Uncertainty

A significant amount of work addresses the visualization of uncertainty — a taxonomy by Potter et al. [144] contains a thorough overview of the state-of-the-art approaches. In terms of research relevant to visualizing uncertainty in meteorological data, the early work focuses primarily on uncertainty in Geospatial Information Systems (GIS). Works by both Pang [138] and MacEachran [105] outline various issues and recommendations with regard to the visualization of geospatial uncertainty, whereas Thomson et al. present a typology for uncertainty in geospatially referenced data for intelligence analysis [184].

Attempting to better address the needs of meteorologists, a number of papers focus on understanding uncertainty from simulation ensembles. Vis5D [71] employs small multiples as a way of looking at multiple ensemble members, whereas SimEnvVis [135] explores various comparative techniques for investigating differences among ensemble members. EnsembleVis [143], on the other hand, enables the generation of visual summaries for ensemble forecasts that specifically emphasize the probabilistic characteristics of the ensemble. Additionally, the Noodles tool created by Sanyal et al. [163], the probabilistic marching cubes technique introduced by Pöthkow et al. [141], and the contour boxplot technique of Whittaker et al. [202] present methods for directly summarizing ensembles of isocontours in visualizations. None of this work, however, has addressed the task of relating multiple features across an ensemble.

3.5 Informed Defaults

Many meteorologists turn to visualization as a way to quickly locate relevant weather features within the forecast data. These meteorologists often mentally integrate information from multiple visualizations, which make use of a wide range of encoding conven-

tions. The design flaws in many of these visualizations can lead to misinterpretation, inaccuracy, and inefficiency, among other consequences [38]. In an attempt to address this issue, we present a set of informed, default encoding recommendations that integrate existing meteorological conventions with visualization principles.

3.5.1 Understanding Meteorological Encoding Conventions

We collected representative samples of visualization products from a variety of sources used extensively by our collaborators. In total, we compiled 41 images representative of the encoding conventions used by 9 meteorological sources — we include these images in our supplemental materials archive. This sampling was intended to be representative rather than exhaustive, encompassing the most popularly referenced visualization products along with those specifically referenced by our collaborators. We coded these images based on the forecast variables shown and the encoding channels used. Using these codes, we then grouped the images based on their similarity.

In general, we found that visualizations generated by a given source have consistencies, but those consistencies rarely extend across sources. The encodings employed in these visualizations include contours, color maps, texture maps, streamlines, glyphs, and wind barbs [197]. Contours, color maps, and wind barbs are favored heavily, with contours and color maps showing up in 85% and 78% of the samples, respectively, and wind barbs comprising over 80% of all glyph usage. For two-thirds of the samples, either two or three fields are encoded in a single visualization, though we did find isolated examples of attempts to encode up to four fields simultaneously. For both contours and color maps, common intervals exist for certain variables: relative humidity and probabilities usually step by 10%; temperature usually steps by 3°, 4°, 5°, or 10°; and geopotential heights usually step by 30, 40, 60, or 120 m. Due to the static nature of these images, a label is required for each connected component of each contour in a given image. As a result, increasing the number of contours in a given visualization, either by using a smaller interval step-size or by layering additional fields, quickly leads to increasingly cluttered visualizations like those in **Fig. 3.1a** and **Fig. 3.1d**.

In terms of color usage, there are several problematic trends. It is common for colored contours to be overlaid on top of color maps, or for colored contours and color maps

with transparent portions to be overlaid on top of multicolored backgrounds, both of which add difficulty to resolving features — this latter case is exemplified in **Fig. 3.1c**. Additionally, rainbow color maps, which are widely considered a poor choice by the visualization community [14], comprise nearly 70% of the color maps in our sample set. We also noted multiple cases of cyclical color-map design, including the cyclical rainbow color map shown in **Fig. 3.1b**.

Other observed color trends, however, do not necessarily violate visualization principles. For example, meteorologists almost always use discrete color maps, rather than continuous color maps; in fact, our sample set did not include a single continuous color map. As another example, when a mean field is displayed in concert with a variational measure such as standard deviation or spread for an ensemble, the variational metric is virtually always color mapped and the mean overlaid as a set of contours, as exemplified in **Fig. 3.1d** — we encountered only one example where this configuration was reversed. We also noted a handful of trends in color use tied to specific variables. Geopotential heights are almost never color mapped and are usually shown as black contours, like in **Fig. 3.1a** and **Fig. 3.1c**. When color mapped, relative humidity is nearly always represented using three shades of green, usually denoting the 70%–80%, 80%–90%, and 90%–100% intervals, with the lower intervals being entirely ignored; **Fig. 3.1c** shows an example of this. When relative humidity is contoured, on the other hand, isocontours are generated for every interval of 10%. The color map for temperature is always a rainbow color map with blues on the low end and reds on the high end, although we also noticed a secondary trend where temperature is displayed using a colored contour scheme with blue specifying values below freezing and red specifying values above freezing, as seen in **Fig. 3.1a** and **Fig. 3.1c**. Additionally, although probability fields are usually encoded using either contours or a color map, there were notable examples, such as the one reproduced in **Fig. 3.1e**, where the probabilities are dual encoded using both. The contours provide a constant overview of the entire field, while the color map highlights value intervals of significance.

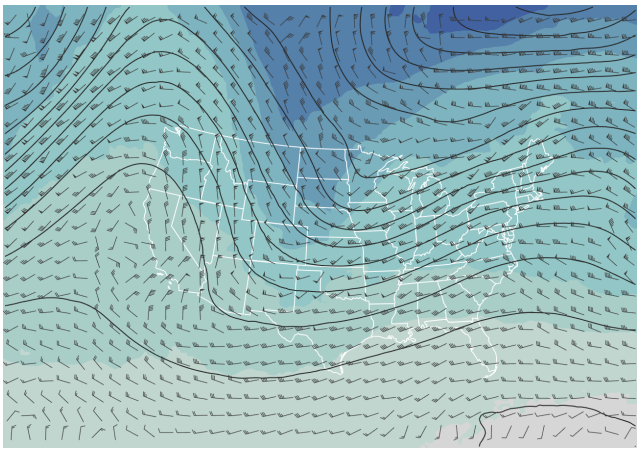
3.5.2 Proposed Defaults

Based on the trends we observed, we propose the following set of informed defaults outlined in three cases. The first case involves the simultaneous display of independent

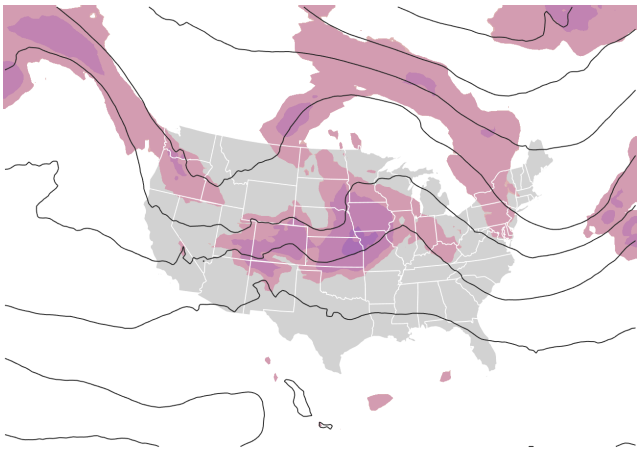
fields. These may be original or derived fields from either a deterministic or ensemble forecast. The notion of *dependence* used here is based on whether the interpretation of a given field necessarily depends on knowledge of some other field. The only dependent fields that arise within the context of this work are standard deviation fields derived from an ensemble, as a measure of standard deviation is generally meaningless without the corresponding mean. Thus, our second case involves the simultaneous display of an ensemble-derived, dependent variation field with its corresponding mean field. Our third case involves the display of the uncalibrated probability of a given event or condition derived from the ensemble of predicted outcomes. The *uncalibrated* probability refers specifically to the percentage of the ensemble members that predict the event or condition of interest. It is important to distinguish this from the actual expected frequency of the condition or event, because most ensembles underestimate the actual range of possible forecast outcomes [170]. For each of these three cases, we have abstracted a recommended set of encoding choices, examples of which can be seen in **Fig. 3.2**.

For simultaneously displaying independent fields, we recommend staying within combinations of three encoding choices: a base color map, a set of contours, and a set of glyphs or texture map (although we do not currently support texture maps in WeaVER). This set of encoding choices allows for the simultaneous display of two 2D scalar fields along with a third 2D field of either scalar or vector values. Ware and Plumlee propose an alternative configuration that uses color, texture with optional contour boundaries, and animation to encode a similar set of fields [197]; however, this configuration is a departure from the conventions at the core of current meteorological visualizations. Given that meteorologists, especially those in decision-making contexts, have significant training and experience using current meteorological conventions, we tried to incorporate those conventions as much as possible while remaining within the bounds of accepted visualization practices.

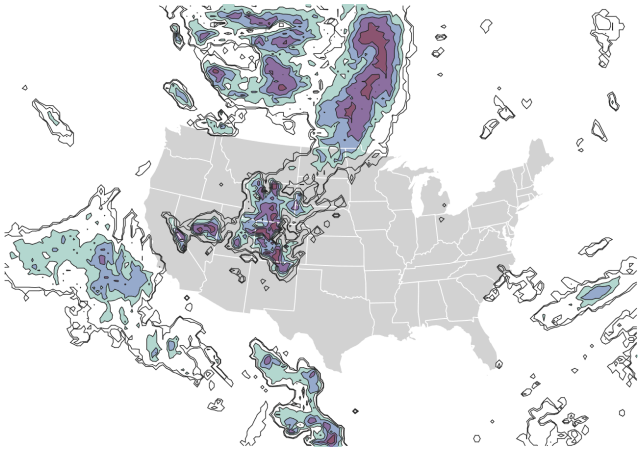
For the other two cases, our recommendations mirror the existing meteorological conventions. For the simultaneous display of an ensemble-derived mean field with its dependent variation field, we recommend a single combined choice that encodes the variation metric as a color map and overlays the mean field using a set of contours. For encoding the uncalibrated probabilities of a condition, we recommend a single dual-encoded choice



(a) layering independent fields



(b) mean and variation



(c) probabilities

Fig. 3.2. Example visualizations generated by each of our three cases for informed defaults.

where value intervals of significance are highlighted by a color map, while all other value intervals are represented only by contours. The notion of a *condition* used here may refer to either a single individual condition (e.g., relative humidity less than 10%) or the joint condition of a set of individual conditions. Although multiple user choices may go into specifying a set of individual conditions, that set still represents a single joint condition choice.

We additionally provide a set of recommendations for low-level encoding behaviors for specific variables or fields. For both the color map and the contour intervals, the trends discussed in Section 3.5.1 represent reasonable choices. For variables without common trends in interval spacing, we use spacings specified by our collaborators. Wind barbs are evenly spaced to prevent glyph intersection, although other placement strategies may also be acceptable [12].

We propose a set of color maps for several variables, shown in **Fig. 3.3**. Aside from the categorical map for representing the Haines index in **Fig. 3.3g**, these color maps are designed to approximate perceptually uniform steps in luminance. We do not claim these color maps guarantee perceptually equivalent luminance steps, since that would require a calibrated monitor, control over the ambient light in the viewing environment, etc. [86]. The color maps are, instead, defined in device-dependent HSV space, but use a nonlinear function in order to vary brightness in a way that creates convincingly uniform perceptual steps. Initial versions of the color maps, which can be found in the supplemental materials archive for this dissertation, were generated algorithmically using a cube-root approximation to the Munsell value scale [125] in combination with linearly varying hue and saturation. Additional hand-tweaking, however, was still required to create the final proposals.

Integrating an interactive routine to approximate a monitor's gamma value, such as that outlined by Kindlmann et al. [86], would allow for more accurate control of luminance variation across devices. We wanted to ensure, however, that the color maps were specified such that they could also be used in other meteorological tools.

The proposed color maps cover similarly large portions of the luminance spectrum, but we explicitly did not use the entire spectrum in order to allow for gray values for both contours and wind barbs that would remain distinguishable from the color-mapped

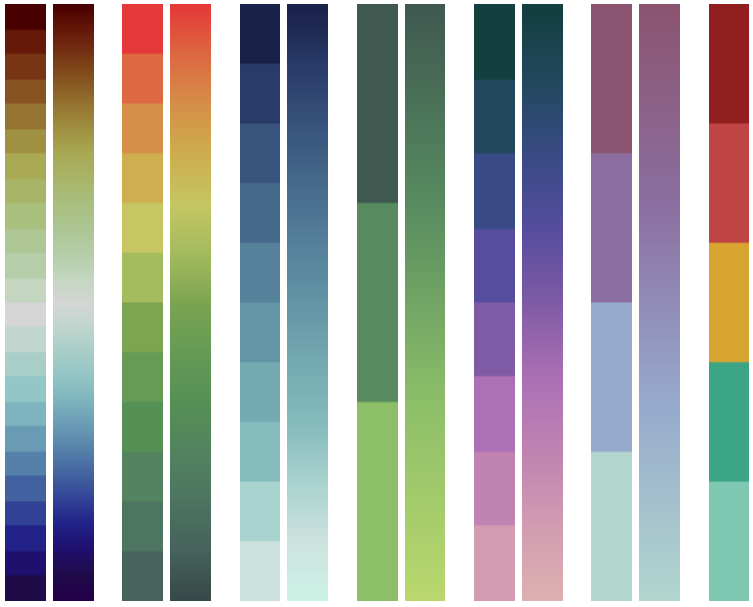


Fig. 3.3. Discrete and continuous color map proposals for: (a) temperature, (b) wind speed, (c) precipitation, (d) relative humidity, (f) standard deviation, (e) probability, and (g) Haines index.

background when overlaid. Based on the color usage trends that we observed, we created a green color map covering the 70%–100% value range for the display of relative humidity and a diverging spectral color map for temperature, as shown in **Fig. 3.3d** and **Fig. 3.3a**. Also included in **Fig. 3.3** are separate, distinguishable color maps for other common variables: accumulated precipitation, high wind speeds, ensemble standard deviation, and ensemble probabilities. Although these additional color maps are based solely on principles of effective color usage [119], rather than any meteorological convention, the unique color maps for each variable within a system are meant to improve users' efficiency in dealing with multiple visualizations. With this goal in mind, the majority of the color maps are defined on an absolute scale covering the range of values taken by that variable across all height layers. The color map for standard deviation is the notable exception, since meaningful thresholds of uncertainty change for different variables.

We recognize that our choice of a spectral color map for temperature is generally considered a poor choice for univariate color-map design [14], but there is a history of justified use of spectral color schemes for visualizing weather [143], [197]. Although hue does not have an inherent perceptual ordering [195], its familiarity and widespread use have led

to an expected ordering within meteorology. This ordering is further strengthened by matching the intuitive mappings that exist for certain colors in spectral schemes (e.g., red is hot, blue is cold) [197]. Further, by controlling for luminance variation, the proposed spectral color map does not suffer from the perceptual irregularities that traditionally plague spectral schemes [14].

We employ a diverging luminance scheme in order to emphasize the nature of the freezing point boundary as a critical center point for temperature. Moreover, well-designed diverging spectral schemes behave just as well as widely accepted two-hue diverging color schemes for modeling data distributions with a critical mid-range value [20]. Given all this, we believe that spectral-nature alone is an insufficient reason to go against one of the strongest meteorological conventions that we encountered.

The low-level encoding behaviors for the variables presented in this section do not comprise a comprehensive solution for meteorological visualization, but we believe they represent a reasonable solution for some portion of the meteorological community. Our early interviews indicated that most meteorologists refer to only 50 or so fields and that, in the case of our collaborators, those fields cover only a handful of different variables. This observation made the specification of unique color maps on a per-variable basis possible for our prototype. Given the large number of variables present in forecasts, however, having separate, distinguishable, absolute color scales for every possible variable will never be feasible. Similarly, while our diverging spectral color scheme for temperature results in clear visual differences at intervals of 10° or even 5°, if our collaborators had needed to be able to resolve steps of 3°, the luminance differences become too small for smoothly varying hue and saturation to result in sufficient visual differences. As such, it is essential to remember that enabling effective defaults will always require understanding the needs of the target users. The specifications we have outlined here are simply one possible configuration that results in effective visualizations for the majority of the cases handled by our collaborators.

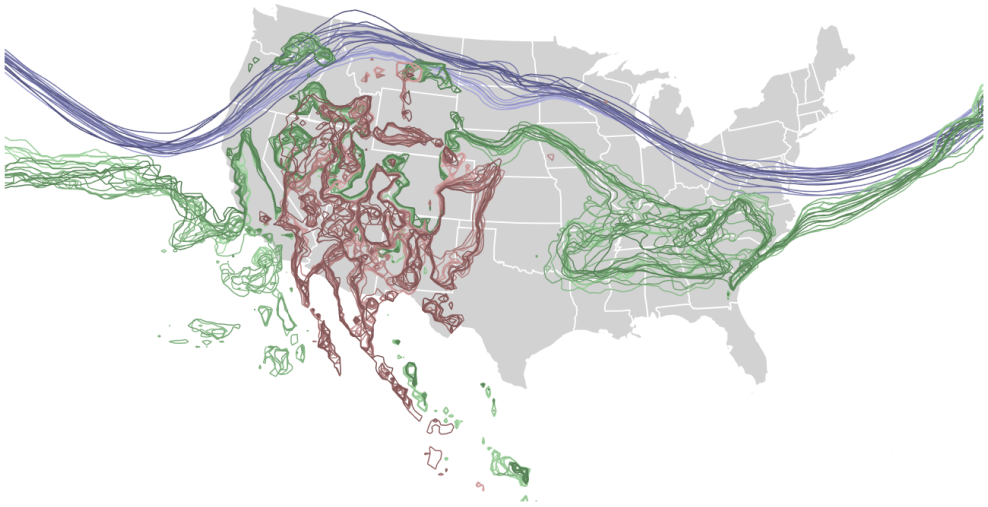
3.6 Ensembles of Features

Understanding the variability and associations among features across ensembles is often critical to decision-making. This statement is especially true in areas such as wildfire

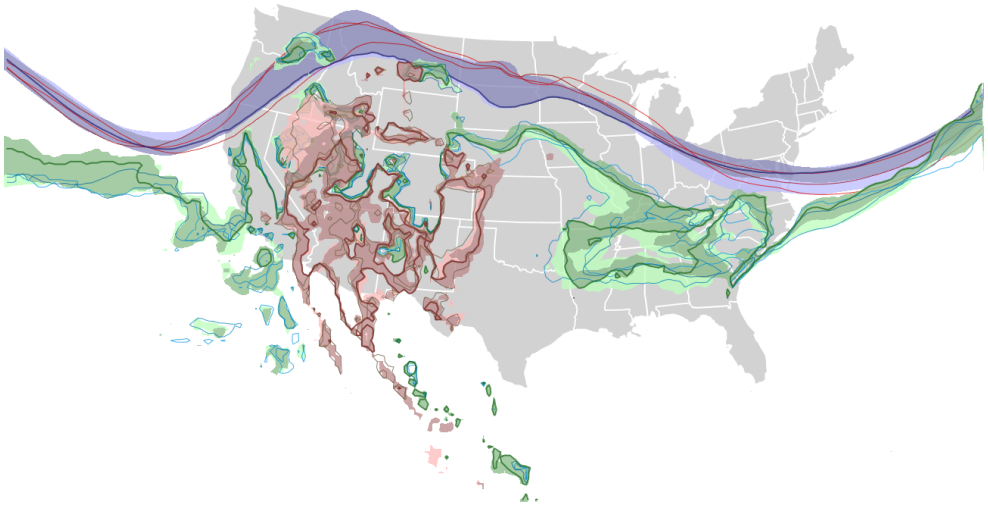
emergency management that rely on forecasts multiple days into the future when the forecasts remain incredibly uncertain. Many meteorologists rely heavily on visualization to develop this understanding, yet no existing visualization methods allow users to explore the variability of feature relationships across an ensemble effectively. Given that most decisions in meteorology are based on the intersection of multiple features, this represents a significant gap in visualization's current support for meteorological tasks.

In meteorology, the only conventional technique for directly examining the distribution of behaviors for a given feature across an ensemble is the use of *spaghetti plots*. Spaghetti plots display the set of isocontours associated with a specific value, one for each ensemble member, within the same, generally static plot. Color is routinely used to distinguish which ensemble member a particular contour is derived from. Spaghetti plots for multiple features, especially intersecting or overlapping features, are rarely overlaid within visualizations because the results do not effectively present the range of feature relationships in the ensemble. If color is used on a per-member basis, visually separating overlapping or intersecting contour sets into their respective features becomes incredibly tricky. Alternatively, when color is used to differentiate feature contour sets explicitly, it becomes nearly impossible to pick out the contours associated with a particular ensemble-member. These problems are compounded as the number of ensemble members or the number of different features increases. In both cases, picking distinguishable colors becomes more challenging, and the occlusion of contours by one another, regardless of which feature they belong to, becomes increasingly problematic.

Although the use of small multiples for each ensemble member in Vis5D [71] represents one possible step toward addressing these deficiencies, the idea has never been generalized back to 2D plots. We opted for a different approach, developing a modified formulation of the spaghetti plot technique that uses interactive highlighting as the primary mechanism for distinguishing between members. We refer to these modified spaghetti plots as **interactive spaghetti plots**. As shown in **Fig. 3.4a**, this formulation frees up the color encoding channel, allowing the use of color to differentiate and enable direct comparisons among the distributions of multiple isocontour features within a single plot. It also mitigates contour distinguishability and occlusion issues by bringing the highlighted contour(s) to the front of the view and decreasing the visual saliency of other features' contours. Users are able



(a) interactive spaghetti plots



(b) interactive contour boxplots

Fig. 3.4. The simultaneous display of three isocontour features using modified ensemble visualization techniques.

to look at multiple features simultaneously and interactively highlight contour sets at both the member and feature level. These capabilities support a variety of both exploratory and investigatory tasks.

Even with interactive highlighting, spaghetti plots still do not scale well, quickly becoming visually cluttered. Given conversations with our collaborators heralding the advent of *super ensembles*, ensembles with hundreds of members, we also wanted to provide users with a technique that could scale. To this end, we integrated *contour boxplots* [202] as a state-of-the-art encoding technique that directly summarizes an ensemble of isocontour

features using a boxplot-like summarization of a set of 2D contours.

There are several benefits to using contour boxplots for simultaneously visualizing multiple isocontour-based feature sets. Being analogous to box-and-whisker plots, they provide a commonly understood statistical framework for summarizing a distribution of isocontour-based features across the ensemble. Because contour boxplots were motivated in part by the idea of aggregation preserving shape, the characteristic details of physically plausible features are still present in the summarizations, which is often not the case for the corresponding features derived using the ensemble-mean. Moreover, contour boxplots are scalable because their visual representation remains consistent for any arbitrary number of ensemble members.

As with the interactive spaghetti plots, we use a modified formulation of **interactive contour boxplots**. Color is, again, used to differentiate between the summarizations of different features over the ensemble, as shown in Fig. 3.4b. We duplicate the functionality for feature-level highlighting, bringing that feature's contour boxplot to the front of the view and decreasing the visual saliency of any other contour boxplots in the display, mitigating distinguishability and occlusion issues. Additionally, we allow users to query and highlight contours from the ensemble on a per-member basis, thereby retaining the capability for interactive exploration of the entire distribution of feature relationships across the ensemble.

We chose to limit users to looking at up to three features simultaneously, but the interactive mechanisms we have outlined could allow for the simultaneous inclusion of a larger number of features in a single view. The added benefit to simultaneously visualizing a fourth or even fifth feature, however, is unclear, when users are provided with a mechanism for easily swapping out features.

3.7 WeaVER

In this section, we provide an overview of WeaVER, an open-source tool developed to test our informed defaults and extended ensemble visualization techniques. WeaVER is not intended to be a fully viable alternative to current operational forecasting tools; it is a proof-of-concept, designed to allow our collaborators to evaluate the proposed ideas.

3.7.1 Data Processing

WeaVER is designed to visualize the NCEP Short-Range Ensemble Forecast (SREF), which at the time of our project contained 21 member simulations: 7 different sets of initial conditions run over 3 models. The simulations were run four times per day (at 03, 09, 15, and 21 UTC) and included predictions at 3-hour intervals out to 87 hours into the future. We used the version of the SREF run on NCEP's 212 Grid [130], which is a Lambert Conic Conformal grid over the continental United States with approximately 40 km grid spacing. The forecast data were retrieved from the NOAA Operational Model Archive and Distribution System (NOMADS) server, `nomads.ncep.noaa.gov`, where they are released in the binary GRIB2 [130] format. The wgrib2 command-line utility [5], which is made publicly available by NCEP's Climate Prediction Center, is used to parse the data into a csv format for preprocessing.

A significant amount of data preprocessing is required for WeaVER to achieve interactive rates. We generate a number of derived fields, such as the Haines index, for each ensemble member, along with various statistical derived fields (max, min, mean, and standard derivation) across the ensemble. We also precompute the statistical quantities required to generate interactive contour boxplots for various isovalues. Additionally, we derive various *condition fields* for calculating the uncalibrated probabilities of arbitrary joint conditions on the fly. Because an individual condition applied to a single field results in a Boolean value at each grid point, we can concisely represent a condition applied to an ensemble of fields as a bit-set at each grid point. This representation reduces the computation of both arbitrary joint conditions and uncalibrated probabilities to a small number of per grid-point operations. The bit-set representing any joint condition can be computed using a series of bitwise AND (&) operations, and calculating the uncalibrated probability requires only dividing the Hamming weight of a bit-set by the total number of ensemble members.

3.7.2 System Overview

WeaVER consists of five interchangeable views: a *deterministic* view, an ensemble statistic (*stat*) view, an ensemble mean and standard deviation (*mnsd*) view, an ensemble *probability* view, and a *direct ensemble* view. The *deterministic* and *stat* views both handle layering of

multiple independent fields, the first of the three cases discussed in Section 3.5. The ensemble *mnsd* and *probability* views handle the second and third cases, respectively. The *direct ensemble* view, on the other hand, allows users to switch back and forth between interactive spaghetti plots and interactive contour boxplot summarizations using the mechanisms described in Section 3.6. The specific fields, isovalues, and conditions supported across these views were explicitly requested by our collaborators.

Across all five views, fields may be dragged from a library on the right-hand side and dropped onto various encoding targets at the top of the view, as illustrated in **Fig. 3.5**. The currently configured visualization is automatically updated according to the user's choices. Mouseover of any of the drop targets reduces the saliency of the other encodings in the visualization. When color maps and contours are controlled by the same layer (as in the ensemble *mnsd* and *probability* views), the visual saliency of the contours is reduced in favor of showing the color map more clearly. In all views, a tab-like mechanism allows users to create and quickly switch back and forth between multiple configured visualizations by using either the GUI interface or the numeric keys on the keyboard. Every view also offers independent time manipulation through animation, a slider-based interface, and forward or backward time stepping using either the GUI interface or the arrow keys on the keyboard. We include a video overview of the interactive features in the supplemental materials archive for this dissertation.

Contrary to meteorological convention, we take a detail-on-demand approach to contours labels. There are no labels on the contours by default; instead, labels appear on mouseover. This action simultaneously highlights the full isocontour, which is especially helpful for contours with multiple connected components. Clicking when a contour is highlighted creates a persistent *sticky label* that moves with the contour across time steps. At any time, a user can reposition a sticky label along the contour by dragging or delete it by double-clicking. Limiting the labels to those requested by the user reduces not only clutter but the number of targets for visual search.

3.7.3 Implementation

All of the data fetch and preprocessing code was implemented using a combination of bash scripts and C++ programs that have been tested and run on both Mac and Linux.

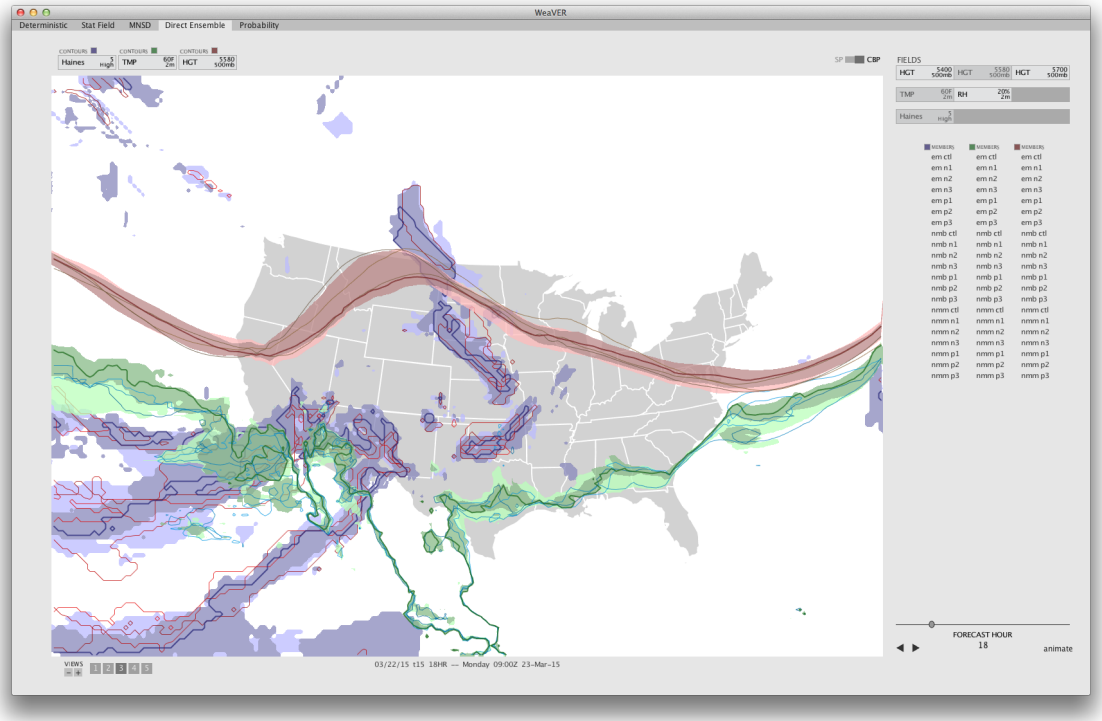


Fig. 3.5. An overview of the interface for WeaVER, an open-source tool developed for supporting meteorological analysis. Here, WeaVER is being used to visually relate multiple isocontour features across an ensemble using interactive contour boxplots.

This code has several dependencies: we use wgrib2 [5] to transform the data from their original binary GRIB2 format, and the contour boxplot implementation relies on ITK [2].

WeaVER itself is implemented entirely in Processing [4]. The resulting Java application has similarly been tested and runs on both Mac and Linux. We are making the source code for WeaVER publicly available at `samquinan.github.io/WeaVER/`. Specifications of the proposed default color maps are included within this source code, along with both example preprocessed data and the data processing code.

We note that the current handling of geographic projection within WeaVER is an approximation of a Lambert Conic Conformal projection. As NCEP's SREF is run on a Lambert Conic Conformal grid, we found that treating the forecast data as an equidistant grid overlaid on top of a preprojected Lambert Conic Conformal map provided a sufficient visual approximation to the proper projection.

Additionally, we based our design decisions to use the 40 km resolution version of the SREF and to visualize the forecast over the entire continental US on the specific needs of our collaborators. In general, meteorological phenomena exist at a range of different geographic scales, ranging anywhere from under a single kilometer to thousands of kilometers. The appropriate geographic scale for visualizing weather data depends both on the scale of the features of interest and on the resolution of the underlying forecast.

3.8 Validation

We validated WeaVER through a series of semistructured interviews with our collaborators. In an initial set of interviews, we elicited feedback from various collaborators after walking them through a full demo of WeaVER. We additionally provided custom installations of the software for an extended evaluation period to our two collaborators working in wildfire prediction, allowing them to evaluate the tool in the context of their work environment. In this section, we describe how one of our collaborators used WeaVER with historical forecast data and provide informal feedback from our other collaborators.

3.8.1 The Diego Fire

For one of our wildfire prediction collaborators, we supplied archived forecasts from three mornings leading up to the Diego fire, a lighting-strike fire that began in northern New Mexico during June 2014 — our collaborator requested these specific forecasts in order to retrace his predictions of the Diego fire using WeaVER. He began by investigating the forecasts through the *deterministic* and *stat* views in order to gain a sense of the big-picture atmospheric conditions over the forecasts. He was particularly excited about the ability to easily create visualizations with combinations of fields tailored to his problems and preferences. He also noted that the resulting visualizations were more visually appealing and easier to read than what he usually looks at. Using these views, he was able to determine that the forecasts showed the expected signals for a threat of a lighting-started fire: moisture, indicating lighting potential, on the front end of the forecasts, followed by windy, dry, unstable conditions for a day or so after.

After gaining a sense of the big picture three days out, our collaborator began investigating the forecasts using the *direct ensemble* view. He stated that he needed a sense

of the spread or variation across the ensemble, but more importantly, he also needed to be able to understand how that variation differs from a particular model or member. He explained that organizations such as the National Weather Service still key their recommendations off of a deterministic forecast, so understanding how the rest of the ensemble compares to that particular member is incredibly important. Consequently, he appreciated being able to highlight a particular member from the ensemble interactively. He was also particularly impressed by the contour boxplot summarizations. He stated that, while it would take training for forecasters to understand what they are looking at, the contour boxplots provide the same visual cues of the forecast as spaghetti plots, but much more quickly and concisely. When pressed to look at multiple features simultaneously, our collaborator noted that he could see expected behaviors and interactions. He also noted that interactive spaghetti plots had significantly decreased utility compared to the interactive contour boxplots for contour-sets generated from non-well-behaved fields such as the Haines Index.

Finally, our collaborator looked at the *probability* view in order to determine which areas had a high likelihood of a critical combination of dry, windy, and unstable conditions in the latter portions of the forecast. As we have reproduced in **Fig. 3.6**, the combined condition of surface temperatures greater than 60°F, surface wind speeds greater than 20 mph, and a Haines index of 5 or greater highlighted the area over northern New Mexico as favorable for fire spread after lighting ignition. Our collaborator noted that this highlighted area, which he would have been worried about, is where the Diego fire originated.

3.8.2 Informal Feedback

We also received a significant amount of supporting feedback from the interviews with our other collaborators. In general, the addition of interactivity was well received. Our collaborators commented that it was straightforward to create a wide variety of views and that the sticky labels and interactive highlighting were a significant improvement over the traditional static visualizations. Several of our collaborators noted that, in existing tools, they must either choose to label every contour or none, which leaves them in a dilemma. They have to either obscure information in order to provide context or forego context entirely. The sticky labels and interactive highlighting, alternatively, made contour

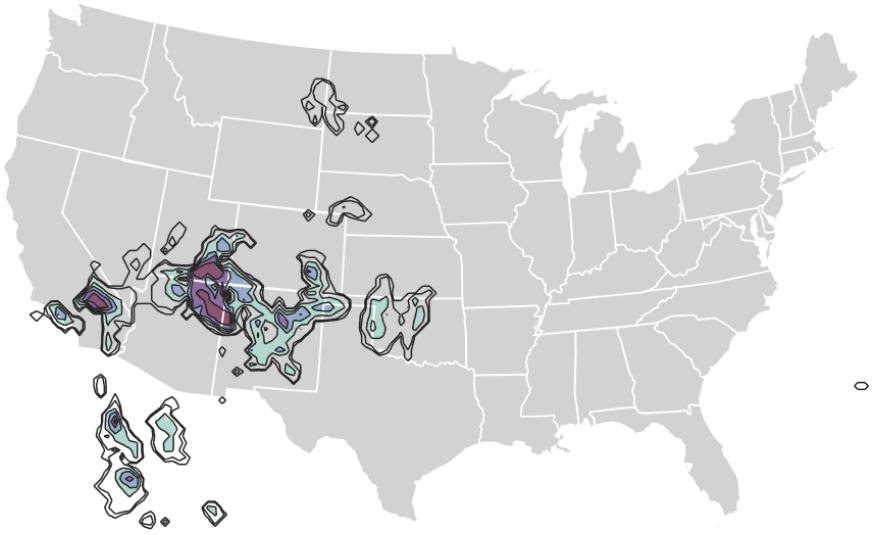


Fig. 3.6. The ensemble-derived probability of conditions favorable for wildfire growth forecasted during the time frame of the Diego fire outbreak, highlighting the area over northern New Mexico where the fire originated.

labels visible only where needed.

We received multiple independent confirmations of the meteorological encoding conventions we derived in Section 3.5.1. Several of our collaborators commented that the informed defaults in WeaVER highlight the fact that effective color usage is something that forecasters generally struggle with. They also noted, however, that some of the meteorological conventions encoded in the informed defaults were not always sufficient for their specific problems. For example, in wildfire prediction, low relative humidity is the primary concern, making a color map highlighting the 70%–100% value range virtually useless. Similarly, although our default temperature increments of 5°C are acceptable for an overview, when forecasting the rain-snow transition line for winter weather predictions, meteorologists need to see a rapid change localized around 0°C with 0.5°C increments.

According to our collaborators, interactive highlighting was the primary key to understanding feature relationships in both the interactive spaghetti plot and interactive contour boxplot techniques. They were particularly intrigued by contour boxplots, which provide a fundamentally different way to get at the statistics or uncertainty of a feature's behavior across the ensemble. One collaborator observed how the quantile bands immediately indicate that half of an ensemble's members are located within a specific geographic region, an observation that would have required him to count individual members in a spaghetti plot.

The collaborator also noted that the latter procedure does not scale, stating that scalability is becoming increasingly important as the number of ensemble members is expected to increase by an order of magnitude over the next couple decades.

3.9 Discussion

The feedback from our collaborators indicates that our informed defaults were mostly a success. Our collaborators, having been provided with simple, transparent visualization encoding choices, were able to interactively generate a wide range of effective visualizations tailored to their individualized needs and problems. Limiting the user choices to practical combinations of encodings and specifying low-level behaviors regarding color usage and contour spacing at the system level resulted in the users' visualizations maintaining several meteorological conventions, while widely being considered easier to read. Labels do not clutter the view, masking important features; instead, meteorologists can interactively place labels exactly where needed based on the current context, regardless of how that context shifts.

The feedback regarding both interactive spaghetti plots and interactive contour boxplots was mixed. Interactivity was the key to enabling the exploration of feature relationships across the ensemble but not in the way that we had anticipated. We figured that interactivity would alleviate differentiation and occlusion issues, but even in static visualizations where those issues exist, we expected both techniques would convey some understanding of the distribution of feature relationships. Instead, interactively highlighting the relationships for each ensemble member did a better job showing the range of feature interactions than the static display of either technique. Additionally, interactive highlighting allows users to efficiently compare the behavior of a given member to the rest of the ensemble, which turns out to be a critical task for meteorologists who need to be able to resolve their predictions against the recommendations of others. Both techniques offer a definite improvement over standard spaghetti plots with regard to relating multiple uncertain features.

Our feedback also suggests that contour boxplots generally represent an improvement over spaghetti plots. Contour boxplots can show physically meaningful statistical variation within the spread of features more simply and more concisely, improving forecasting

speed. They can also create meaningful summaries for complex fields, such as the Haines index, where a lack of grouping among the contours makes it extremely difficult to visually extract a meaningful summary from spaghetti plots. Additionally, contour boxplots are significantly more scalable, which will likely become important in the future.

That said, several aspects of our evaluation were not particularly successful. For reasons we discuss in Section 3.10, we were unable to provide several of our collaborators with their preferred forecast. Additionally, WeaVER does not support investigating ensembles of nonisocontour features, such as cold fronts. These limitations forced our collaborators to make judgments out of context regarding the efficacy of WeaVER.

More generally, our evaluation of informed defaults suffers from a mismatched scope between our designs and validation. We designed the informed defaults using general meteorological conventions in the hope that they would apply to a wide range of meteorologists' needs. We then attempted to validate them with a relatively small set of meteorologists with differing needs and problems. Unsurprisingly, our low-level encoding choices were not always deemed appropriate by all our collaborators. Several alternative options exist. One option, common in design studies, would be to gear the default encoding behaviors to a specific subset of meteorologists, such as those forecasting wildfires in the southwestern US. Another option would be to provide users with the ability to interactively modify low-level encoding behaviors in order to support a broader range of tasks. The first option reduces the generalizability of the informed defaults, whereas the second option provides the opportunity to create increasingly ineffective visualizations. Neither alternative represents a best-case scenario. This conundrum raises an interesting question: what *is* the proper way to design for a set of experts who have similar goals but individualized processes and domains? As we discuss further in Section 3.11, from an evaluation standpoint, such designs may require exploring new models for validation.

3.10 Challenges Working With Weather Data

Several challenges and hurdles made this design study more complicated than we initially anticipated. Here we reflect on a number of these issues to provide guidance for others working with meteorological data in the future.

Because our collaborators were purely consumers of meteorological visualizations, it

was difficult to get direct answers about how existing visualizations were created. For example, confirming that the data processing for derived fields assumed grid-location independence required going through multiple levels of contacts. Similarly, none of our collaborators could answer our questions concerning geographic projections or the smoothing and down-sampling of forecast data. As a result, it took us a long time to understand certain critical complexities.

Of these complexities, we spent significant time understanding the issues with geographic projections. Because NCEP's grids use a nonstandard geodetic datum [130], we needed to understand what issues underlie simultaneously projecting simulations and maps based on different geodetic data. Although there *are* errors that arise from simply modifying the projection equations to use different geodetic data, we were eventually able to confirm that these differences are primarily significant when dealing with high-resolution local-scale simulations. Additionally, despite the prevalence of interpolated contours and color maps, it remains unclear both how the associated latitude and longitude values needed for projection should be interpolated and how existing tools are interpolating those values in practice. For visualizations designed for deployment in decision-making environments, these sorts of considerations need to be accounted for.

As another example, only one of our collaborators was aware that many calculations of the uncalibrated probabilities of joint conditions over an ensemble assume the individual conditions are independent. As the individual conditions derive from the same ensemble, however, this assumption is not only unfounded but generally misleading. Although the resulting differences in predicted probability at a given grid point can be substantial, we generally found fairly subtle differences when visualizing the entire field. Still, such a disconnect between the data presented in visualizations that meteorologists look at and the meteorologists' interpretation of those data is problematic.

We also ran into some practical and engineering challenges that forced us to scale back our designs. The scope of the forecast data was significantly more sizable than we anticipated and initially designed for, and it is only going to expand in the future. Luckily, related work on climatological systems [204] could offer insight into dealing with this data increase. Additionally, internet connectivity issues and a lack of install permissions in several of our collaborators' work offices became nontrivial design hurdles.

We also experienced issues getting the data desired by our collaborators. Several of our collaborators primarily look at long-range ensembles, such as the Global Ensemble Forecast System (GEFS) ensemble. While the GEFS is both run and used internally by the National Weather Service at a 40 km resolution [64], it is made publicly available at only less than half that resolution. These circumstances forced us to test WeaVER with the SREF, which does not provide significant enough lead times for many of our collaborators' decisions. Moreover, a lack of public research archives for the SREF made it challenging to track down meaningful data for our collaborators to look at when eliciting feedback during their off-seasons.

3.11 The Evolution of a Mixed-Methods Research Program

In light of the complications that we ran into during this design study, we became interested in whether there might be a more direct way to validate individual aspects of our proposals. We were particularly excited about the idea of designing a user study to assess the impact of different ensemble-visualization techniques on meteorological decision-making processes. The existing methods used to evaluate uncertainty visualizations primarily involve tasks simplified to focus on low-level details, such as value retrieval, aggregation, comparison, and search [88]. These simplifications are poorly matched to the weather-forecasting process, where "simply showing a complex visualization, expecting a user to extract the necessary information, and to be finished is an oversimplification of how complex visualizations are used" [187, p. 347]. To assess the true impact of different ensemble-visualization techniques, one would need to support the highly individualized, complex reasoning process described in Section 3.3.1 in a controlled way. Our efforts to create a user study that does this are detailed in Chapter 4.

We also wondered whether this complex reasoning process had a larger impact that might begin to explain some of the conflicting guidance we saw between meteorology and visualization. During the design study, we noticed that meteorologists almost always use discrete color maps to encode continuous fields. This practice appears to violate the *expressiveness principle* for visualization design [107], [126], which generalizes to a longstanding recommendation from the visualization community that continuous data types should be

visualized with a continuous encoding in order to show all the data relationships [8], [95]. Given the general types of features meteorologists look for, however, is it possible explicit discretization is conferring some advantage? Chapter 5 outlines the study we conducted looking into this particular guidance conflict. The results, in turn, inspired us to reexamine the role of discretization in another major conflict we observed between meteorological practice and visualization guidance: the continued prevalence of rainbow color maps. That follow-up study is presented in Chapter 6.

In pivoting toward these follow-up studies, we effectively recontextualized this design study as the first part of a more extensive mixed-methods program. From a mixed-methods standpoint, the value of this design study became associated with what *we*, as visualization researchers, learned from the process. The reality of operational weather forecasting meant that, for the proposed contributions of this design study to have any real impact, we would need the backing of the individuals building the meteorological software and generating the most commonly used visualization products. Obtaining that backing, however, requires having empirical support for the theories that we inductively built throughout this initial qualitative course of study. Building that empirical support became the broad goal of our subsequent work.

3.12 Summary

In this chapter, we presented a design study addressing a set of common problems in meteorological visualizations, including the inconsistent and ineffective uses of color. We outlined a system for informed defaults that allows meteorologists without visualization expertise to generate a wide variety of effective visualizations based on current meteorological conventions and visualization principles. We extended state-of-the-art visualization techniques to allow users to relate multiple isocontour features effectively. Moreover, we put our various proposed solutions into a proof-of-concept system, WeaVER, which we validated with collaborators through a series of semistructured interviews.

We also outlined how the insights into the meteorological forecasting process garnered during this design study became the basis for the mixed-methods program at the heart of this dissertation. In particular, this design study provided us with in-depth knowledge of how color is employed in meteorological visualizations and highlighted guidance conflicts

that illuminated potential gaps in our understanding of the best practices for using color in visualizations of 2D scalar data.

CHAPTER 4

TOWARD ECOLOGICAL VALIDITY IN EVALUATING UNCERTAINTY

This chapter provides an overview of the second piece of formative work for this dissertation: a pilot study aimed at improving *ecological validity*, or realism, in a user study investigating the role of uncertainty visualizations in weather forecasting. As we discussed in Chapter 3, user studies evaluating uncertainty visualizations have traditionally focused on simplified, low-level tasks [87], [88] that do not adequately reflect how meteorologists use visualizations. After our design study [145], we wanted to evaluate some of our design proposals further, but we also wanted the results to generalize to meteorological practices. In particular, we began to focus on assessing the impact of different ensemble-visualization techniques across the individualized, complex meteorological reasoning processes that we had observed.

Working with both psychologists and an atmospheric scientist, we designed and ran a 5-week longitudinal study to assess the impacts of adding uncertainty visualization products into meteorologists' forecasting processes. The study treated each individual's process as a black box. We collected data on what information was used in each forecast but did not limit the information a given individual might choose to access or use. The results appear to show that the addition of uncertainty visualization products resulted in some significant differences in forecast accuracy compared to our participants' baseline forecasting ability. We are concerned, however, that uncontrolled weather variability may be confounding these results. In our attempts to address this experimental confound, the expanded full-scale study grew beyond the scope of our resources.

This chapter outlines the motivations, methods, and results from our pilot study and reflects on various lessons learned about the tensions between maintaining ecological validity and establishing experimental control. As the following chapters in this dissertation

will discuss, these lessons significantly impacted how we addressed our subsequent evaluations.

4.1 Study Motivation

Uncertainty is an inherent part of making decisions in a broad spectrum of situations such as wildfire management [145], hurricane evacuation [31], and even water use policy [39]. Studying the effects of visualizing uncertainty in decision-making contexts, however, remains a challenge due to the influences of both expertise and individualized, experiential knowledge [187]. These factors can make controlled studies of the effects of uncertainty visualizations on decision making difficult, if not impossible, in many real-world situations.

There is a growing recognition that evaluation in visualization could benefit from a greater variety of empirical methodologies [27], [94], but the field remains dominated by two primary modes of evaluation: quantitatively focused user-studies and qualitative case studies, like those found in design studies. These two modes of evaluation sit on opposite ends of an experimental spectrum that involves trade-offs among generalizability, precision, and realism [113]. Within this spectrum, the concept of *experimental control* positively correlates with precision in measurements and results. The amount of realism, on the other hand, is largely tied to the concept of *ecological validity*. Ecological validity refers to how closely the experimental setting matches the setting in which the results might be applied [27]. Psychologists have long noted that there are often explicit trade-offs between experimental control and ecological validity [102].

In prioritizing experimental control, existing user-studies of uncertainty visualizations have largely dealt with simplified low-level detailed tasks [87], [88]. These simplifications, however, often force expert users to make judgments outside their usual decision-making contexts. Given the lack of proper context, it is not clear how applicable the results of these studies are in real-world situations. Comparatively, qualitative case-study feedback presents its own set of challenges. Although case studies can show *that* an effect exists in a real-world environment, it is extremely difficult to establish causation. Although additional feedback from activities such as contextual interviews and think-aloud protocols can provide some insights regarding particular design decisions [27], there remains a risk

that qualitative feedback conflates users' preferences or demand characteristics with their performance [23], [70], [77], [133]. Neither set of problems is well suited to understanding the role of uncertainty visualizations in a complex decision-making process such as operational weather forecasting.

In this chapter, we describe an initial attempt at designing and running an ecologically valid user study assessing the effects of uncertainty visualizations on weather forecasting. Relative to the other projects in this dissertation, the contributions of this work are modest. We highlight the need for better ecological validity in evaluating the effects of uncertainty visualizations, especially when expertise is a critical part of the decision-making process, and we outline a strategy for designing and running a more ecologically valid visualization user study with meteorologists. The results from our pilot, although promising, are inconclusive given both the study's small sample size and the existence of potential experimental confounds. As we were unable to run an expanded study to address these issues, we do not claim the results as a contribution. Still, what we learned about ecological validity and expertise in the process of conducting this study had a strong influence on the subsequent work in this dissertation. For this reason, we discuss this project as formative work.

4.2 Evaluating Uncertainty in Weather Forecasting

Weather forecasting is a decision-making domain characterized by *high uncertainty* [34]. Atmospheric dynamics are chaotic by nature, and meteorologists are tasked with predicting those dynamics using a combination of sparse, imperfect measurements of current atmospheric conditions and highly sensitive numerical models of future atmospheric conditions based on uncertain initialization conditions [45]. Compared to experts in other domains, however, weather forecasters are remarkably reliable in their assessments of uncertainty and risk [174]. Daipha argues that the reason for this observed difference is, in part, sociological [34].

Weather forecasters are *institutionally* and *culturally* primed to handle uncertainty information in a way that many other experts and the general public are not [34]. Managing uncertainty pervades the entire forecasting process, from meteorologists' choices about which of the myriad of available data to focus on [145], to their decisions regarding how

to resolve conflicting and uncertain information [45], to their choices in communicating their findings to different audiences [34]. Its ubiquity has made uncertainty a central part of the meteorological cultural context through organizational logic, institutional memory, and shared practices in places such as local weather offices [34], [52]. These factors raise questions about the applicability of generalized findings regarding decision-making under uncertainty [28], [31], [161] in weather prediction.

A series of studies comparing domain experts' and novices' interactions with weather maps, for example, demonstrates that an understanding of atmospheric dynamics fundamentally changes the way individuals read and reason about weather maps [103]. Although there is some evidence that training in task-relevant meteorological principles can enable nonexperts to read weather maps effectively [50], [69], reliably forecasting the weather requires the ability to reason about atmospheric dynamics, not just the ability to read a meteorological chart [111], [187]. Indeed, "simply showing a complex visualization, expecting a user to extract the necessary information, and to be finished is an oversimplification of how complex visualizations are used" by meteorologists [187, p. 347].

Expert forecasters use visualizations to create an aggregate mental model of what is happening in the atmosphere and then use that mental model as the primary source of information for their judgments and decisions [145]. This process is highly individualized [145] and often changes based on "climate, season, experience, and a host of other factors" [187, p. 343]. These complexities not only make it difficult to assess the role of uncertainty visualizations in the decision-making processes of operational weather forecasters but also raise questions about the generalizability of prior work, where evaluation methods are ad hoc [87], [88]. In systematic reviews of the visualization community's uncertainty literature [87], [88], Kinkeldey et al. note that details like the type and level of expertise among participants are often unclear. Additionally, tasks are often oversimplified, focusing on low-level operations such as value retrieval, aggregation, comparison, or search without any explicit justification.

Although a handful of studies have evaluated the impact of uncertainty on decisions using complex real-world tasks (see the survey by Kinkeldey et al. [87] for a list), we found only a single study examining at the impact of uncertainty on a complex meteorological decision [129]. Specifically, Nadav-Greenberg et al. present a study assessing the effect

of different uncertainty visualizations on participants' ability to decide whether to post a high-wind warning advisory [129]. In line with prior work conducted with nonexperts [81], the results indicate that the inclusion of uncertainty can be beneficial but that those benefits depend critically on how the uncertainty is expressed [129].

It is important to note, however, that the task tested by Nadav-Greenberg et al. is not entirely reflective of operational forecasting environments. The set of products a given meteorologist chooses to use in an operational forecast is often incredibly personalized [145]. Because of this fact, requiring different meteorologists to make judgments based on a limited, controlled set of products forces them out of their normal decision-making context. There remains a need for evaluations that more accurately reflect the macrocognitive model of meteorological forecasting [187]. This need is what we set out to address.

Our goal is to evaluate uncertainty visualizations while accounting for the complexities of forecasting the weather in operational environments. Prior studies have attempted to deal with these complexities in a variety of ways. Stewart et al. analyze the accuracy of meteorologists' forecasts in a naturalistic environment by using archived forecasting data [174]. Daipha employs a qualitative ethnographic approach to understand how meteorologists collect and use information in an operational forecasting office [35]. Smallman and Hegarty use a quasi-controlled, quantitative study to identify the factors that influence operational forecasters' selection of, and performance with, complex weather visualizations [172]. Additionally, Wilson et al. demonstrate the efficacy of eye-tracking for providing insight into the impacts of visualization display changes in an operational forecasting office [205]. Like Smallman and Hegarty, we attempt to balance ecological validity and experimental control in a quantitative study. Our study, however, focuses specifically on assessing the predictive impact of how uncertainty information is displayed.

4.3 Methods and Aims

Prior work indicates that the utility of uncertainty information in visualizations depends heavily on how well the presentation matches experts' mental models for the associated task [81], [90], [129]. To our knowledge, however, no prior evaluation of uncertainty visualization techniques has attempted to fully account for the complex mental models used in operational weather forecasting. We set out to address this evaluation gap.

The primary goal of our pilot is to assess the experimental efficacy of reducing experimental control in favor of improving ecological validity. To this end, we selected a hypothesis that we believed would be straightforward to confirm. In line with prior work [81], [129], we hypothesize that different uncertainty visualization products should have varying impacts on forecast accuracy. We hope that, by validating that differences can still be detected in a less controlled operational environment, our experimental design can act as a template for subsequent evaluations of novel uncertainty visualization techniques in forecasts (e.g., how contour boxplots [202] compare to spaghetti plots with respect to forecast accuracy).

We conducted a 5-week longitudinal study centered around the skills and workflows of a small team of *quasi-expert* student forecasters at the University of Utah. The study employed a mixed design, where each week, the participants were asked to issue sets of 3-day forecasts for daily high temperatures at multiple locations. During 3 of the 5 weeks, the participants were also asked to integrate additional uncertainty visualizations into their forecasts. We then compared the accuracy of the participants' forecasts with and without these additional uncertainty products. The following subsections provide further details about the participants, the uncertainty products, and the experimental procedure.

4.3.1 Participants

Our participants were all members of the Ute Weather Center, an undergraduate forecasting club at the University of Utah. The Ute Weather Center operates as a local weather office for the University of Utah community, issuing daily 5-day forecasts of temperature highs and lows, cloud cover, and precipitation for the campus and various locations around the Salt Lake Valley. It was created to be a safe learning environment where undergraduates could gain experience applying class-based skills to weather forecasts in an operational context. At the time of our study, the Ute Weather Center's forecasts were reliable enough that various members of the campus community, including the University of Utah baseball team, frequently utilized them.

The five individuals (1F, 4M) we worked with comprised the team staffing the Ute Weather Center over the summer of 2015. The individuals were all undergraduate students, and they had similar levels of expertise. Although their skills did not yet match

those of professional forecasters, each had at least 6 months' worth of experience regularly issuing forecasts for the Ute Weather Center and 1–3 years of study as an atmospheric sciences major. Professors in the Department of Atmospheric Science confirmed that this level of expertise was sufficient for basic meteorological analyses, such as our selected task of forecasting daily high temperatures.

Our goal in this study is to keep participants as close as possible to their actual decision-making process. To this end, we shadowed two of our eventual participants' daily forecasts while we were designing the study. Both participants' workflows mirrored the ones we observed in our prior design study [145]. Each forecaster began by establishing the big-picture status of what was going on in the atmosphere (e.g., "There is a low-pressure system coming through [the region]"). That understanding was then used to reason about the expected effects at the regional level or *mesoscale* while accounting for interactions with other regional effects, such as the atmospheric phenomena caused by the presence of the Great Salt Lake. The forecasters would then make their local, point-based forecasts keeping all of this information in mind. In line with our prior observations [145], this process was highly individualized, with both student forecasters utilizing different sets of visualization products to arrive at similar forecasts.

In order to account for the individualized process differences among forecasters, our study treats the forecasting process as a black box. We asked the forecasters to tell us which products they used in a given forecast but allowed them to conduct their forecast using whatever information they wanted. We recognize that such a policy represents a notable loosening of experimental control compared to prior work [129], but it is absolutely necessary for ecologically valid observations of the impact of different uncertainty visualizations on decisions in an operational forecasting context.

4.3.2 Uncertainty Stimuli

The uncertainty visualization products that we generated for the experiment were based on visualization techniques commonly used in meteorology. They included plume diagrams, mean and standard deviation (MNSD) plots, and spaghetti plots [148]. In each of these products, the notion of uncertainty derives from an *ensemble* of numerical weather prediction simulations designed to represent a space of possible forecast outcomes [145].

Each product, however, shows a fundamentally different view of the data.

As one atmospheric sciences professor explained to us, meteorologists generally care about four critical perspectives of forecast data:

1) How do things change over time?

2) How do things vary spatially?

3) How do the ensemble members (i.e., simulations) differ?

4) What is the *context*? How do the various meteorological features in the forecast relate?

Plume diagrams, like the one shown in **Fig. 4.1**, emphasize the **time** and **ensemble** perspectives of the data. Individual lines show the predicted values (y-axis) from different simulations (color) for a given variable at a single point in space over time (x-axis). Plume diagrams are primarily used late in the forecast process when the forecaster is making point-based judgments.

Spaghetti plots (see **Fig. 4.2**), on the other hand, emphasize the **space** and **ensemble** perspectives of the data. Each line shows the geographical extent (x- and y- axes) of an *isocontour*, a given value for a single variable in the forecast data, for a different simulation (color) at a given time step. Spaghetti plots are generally used to understand information about uncertainty in the behaviors of large-scale weather features, such as pressure fronts or the freezing point boundary. They tend to be used early in the forecasting process.

MNSD plots (see **Fig. 4.3**) emphasize **space** and **context**, while providing partial information about the variability in **ensemble**. They overlay the mean (black isocontours) and standard deviations (color map) of all the simulated values in the ensemble for a single variable at each grid point for a given time step. MNSD plots utility lies somewhere between the other two plot types we have discussed. Sometimes they are used to relate various weather features; however, as noted in Chapter 3, the averaging operation can result in physically implausible features and relationships. Other times they are simply used as a way to read point values. Where they factor into the forecasting process depends on how the meteorologist is using them.

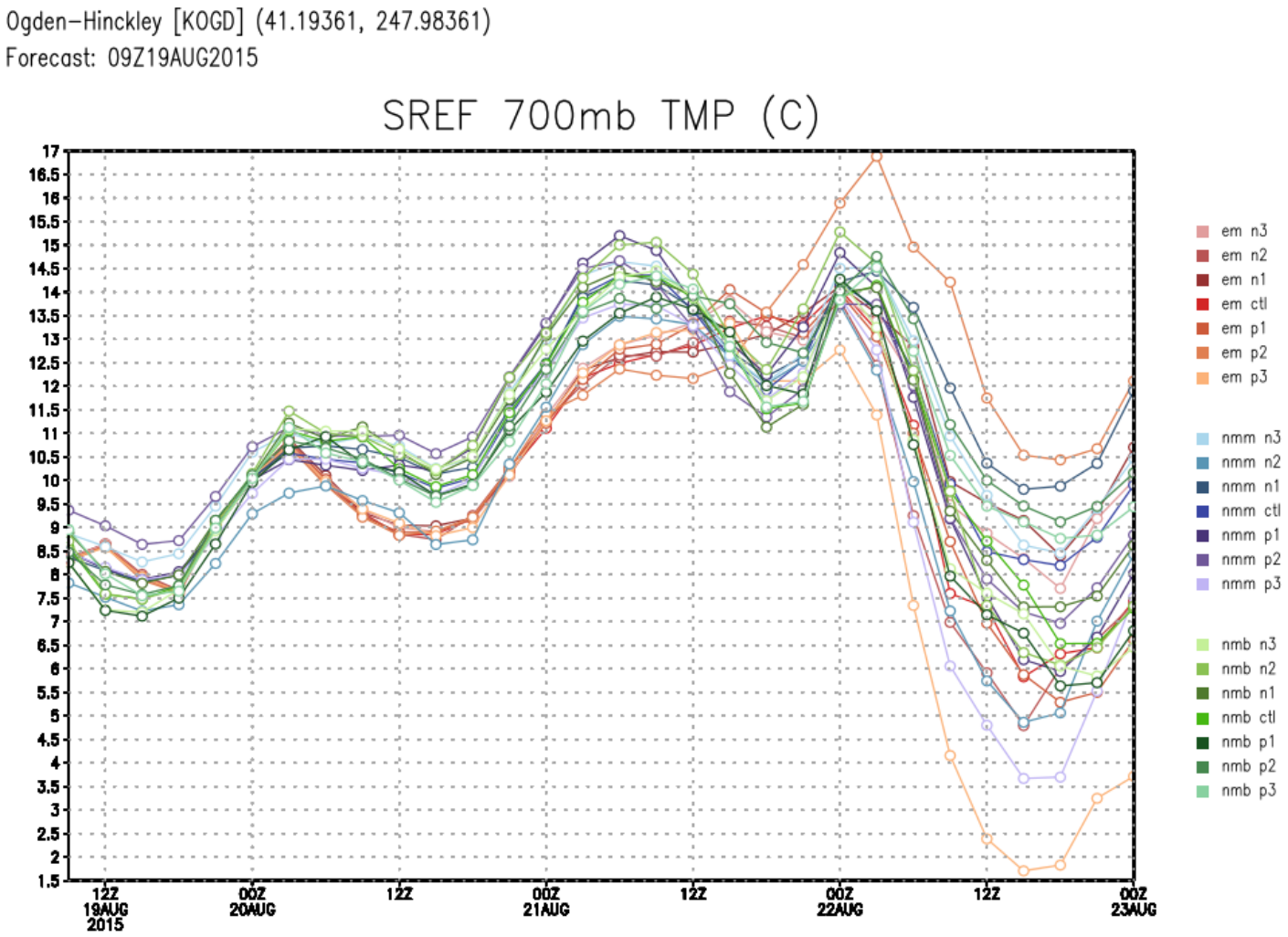


Fig. 4.1. Example plume diagram stimulus from the experiment.

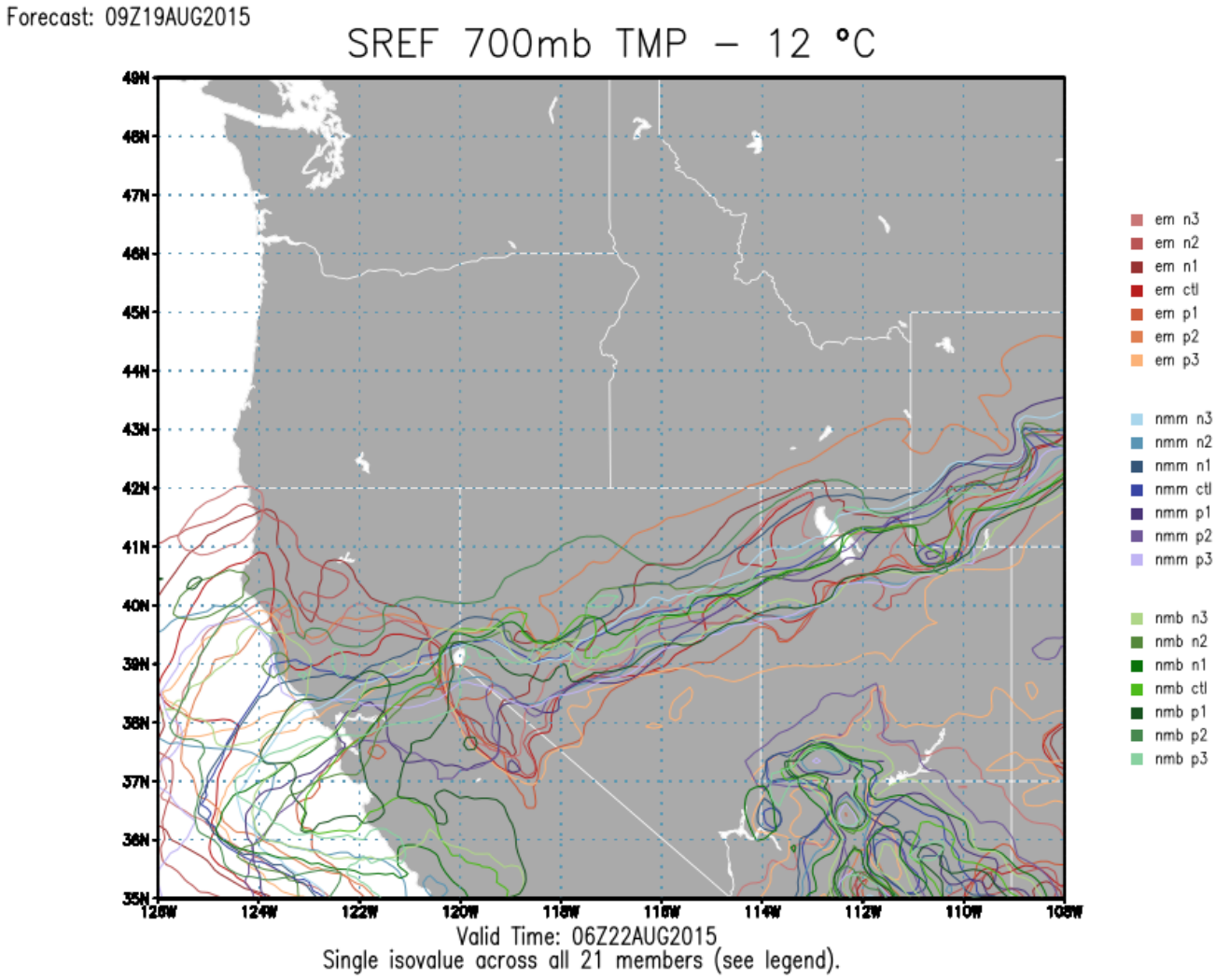


Fig. 4.2. Example spaghetti plot stimulus from the experiment.

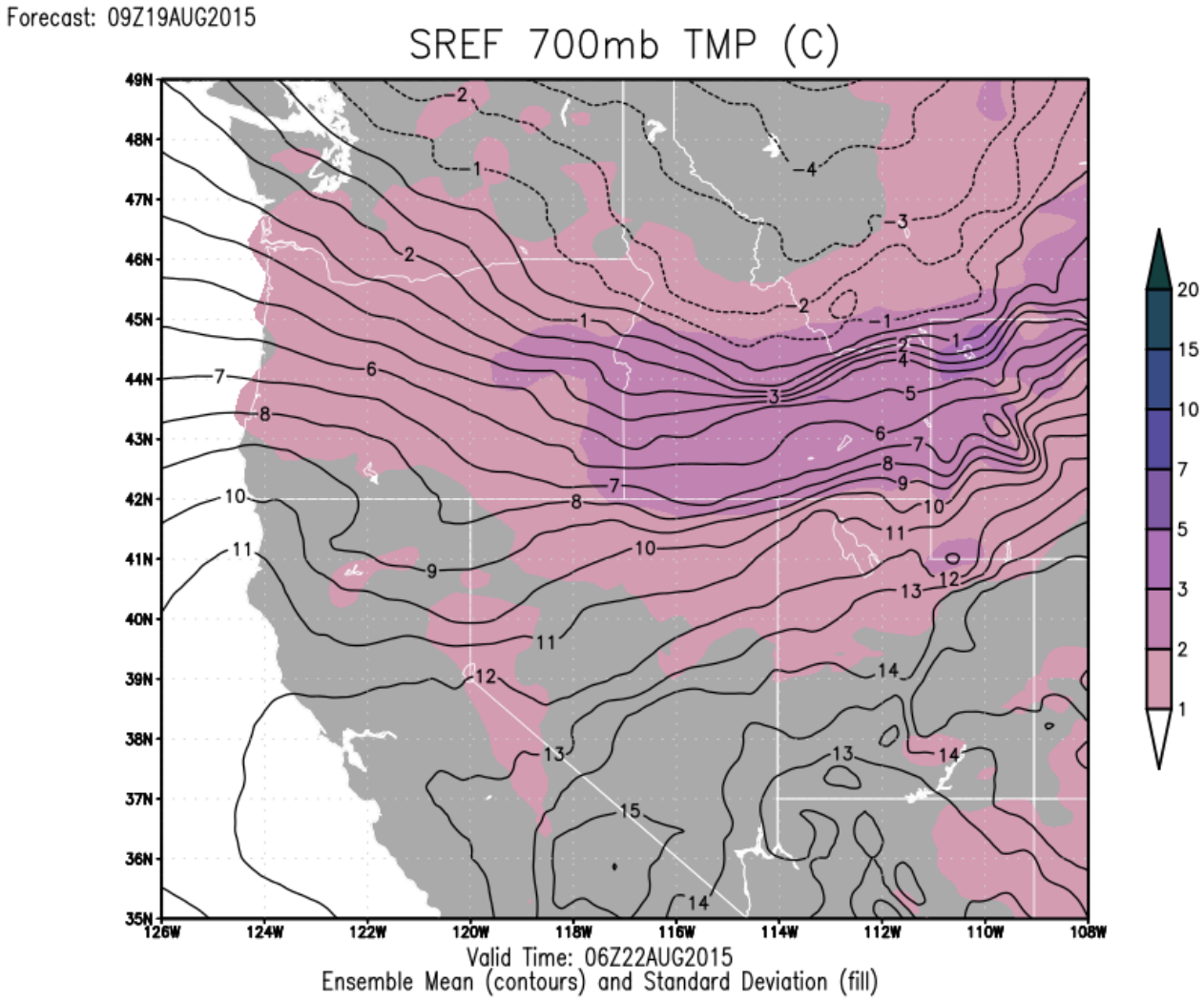


Fig. 4.3. Example MNSD plot stimulus from the experiment.

In each case, the utility of the uncertainty product varies significantly depending on the field shown. As one atmospheric sciences professor explained, plume diagrams are particularly helpful for looking at precipitation, where behaviors tend to be localized, but they are virtually useless for looking at geopotential heights, where features are thousands of kilometers long. By the same token, spaghetti plots are incredibly useful for looking at geopotential heights but often meaningless in the context of precipitation.

The products used in our experiment are all generated using the 700 mb temperature (TMP) field. The reason is that, for areas in the Salt Lake Valley, point values of 700 mb TMP are the primary indicator used in forecasting daily high temperatures. We acknowledge that the expected utility of spaghetti plots for this field is low. The nature of spaghetti plots makes them inefficient for assessing the value and associated uncertainty for specific points in space and time. Given the number of other aspects we are not controlling for in the experimental design, however, we felt it was necessary to use a consistent field across the uncertainty products.

We generated all of the uncertainty products from NCEP's Short-Range Ensemble Forecast (SREF) using the Grid Analysis and Display System (GrADS) software [46]. Updated products were automatically generated with each release of the SREF (at 03, 09, 15, and 21 UTC every day) and published to a web page for the participants to be able to access.

The web page showed only the products that were supposed to be available in any given week. Individual plume diagrams were provided for each of the six locations for which the participants were asked to generate forecasts. For the MNSD plots and spaghetti plots, participants could access any time steps of the forecast and animate across them using an interactive animated stepping interface similar to those found in other publicly available meteorological product pages [177]. Individual sets of spaghetti plots were accessible for every increment of 1°C in the forecast.

Notably, the SREF provides time steps only out to 87 hours, whereas the Ute Weather Center usually issues 5-day forecasts. Unfortunately, we were unable to find a longer range ensemble that was publicly released at a fine enough resolution to allow the forecast to distinguish between Ute Weather's various forecast locations. To account for this mismatch, we asked the participants to report only the first 3 days of their forecasts.

4.3.3 Experimental Procedure

We modeled our experimental procedure off of Ute Weather's normal forecasting operations. Every day, the Ute Weather center issues 5-day forecasts for five locations around the Salt Lake Valley. These forecasts include daily high-temperature predictions. The five team members we worked with were each regularly assigned to forecast one of these five locations. These forecasts usually take team members about an hour to make.

For the study, we asked each participant to provide us with 3-day forecasts of the daily high temperature for two locations: the location for their normal assigned forecast and an additional shared location in the Salt Lake Valley: Ogden-Hinckley airport. Because of the shared mental model, the additional forecast was estimated to require only 10–15 minutes of added work.

We asked the participants to provide forecasts for the study three times each week: on Mondays, Thursdays, and Fridays. While making these forecasts, the participants were instructed to keep track of the URLs for any meteorological products they used and the amount of time the forecast took. They were asked to submit this information along with their forecasts through a Qualtrics survey. The survey also asked participants to rate their confidence in each of their two 3-day forecasts using a 5-point Likert scale. We assessed forecast accuracy by validating each prediction against the observed daily high temperatures measured by weather monitoring stations at each forecast location.

The study ran for 5 weeks. For the middle 3 weeks of the study, we provided the participants with uncertainty visualization products, which we asked them to look at and attempt to incorporate as part of that week's forecasts. At the beginning of these weeks, we provided the participants with a video tutorial and written explanation of how to read and use that week's uncertainty product. We also provided them with a link to a webpage containing up-to-date versions of those products that they could access throughout the week. The rest of the instructions remained the same.

At the conclusion of the study, we conducted in-person interviews with each participant for additional feedback. Participants were compensated $50 for their participation in the study (i.e., at a rate of $10 per hour).

4.4 Analysis

We conducted a linear mixed-effects analysis due to our use of a mixed design with repeated measures. Participants were modeled as a random effect, and the week, location, day, and a potential week:day interaction were all modeled as fixed effects. The weeks of the study were modeled as a categorical variable, with week 1 (no product) as the reference group, weeks 2–4 covarying with the provided uncertainty products (plume, MNSD, and spaghetti plots), and week 5 (no product) acting as a secondary baseline measurement. We modeled the study weeks instead of the uncertainty products directly in order to compare the two baseline measurements and explore whether weather variability might disproportionately be biasing participants' responses. The forecasted location (normal, shared) was also modeled as a categorical variable. The forecasted day, on the other hand, was modeled as an ordered variable ($1 < 2 < 3$) using orthogonal polynomial contrasts, in an attempt to capture the increasing amount of uncertainty depicted in the corresponding visualization products. The full equations and output can be found in the supplemental materials archive for this dissertation, but the core relationships can be seen in **Fig. 4.4**.

As a note, visual inspection of the residuals of our model shows a right-skewed distribution, indicating that the model violates the assumption of normality. Although lin-

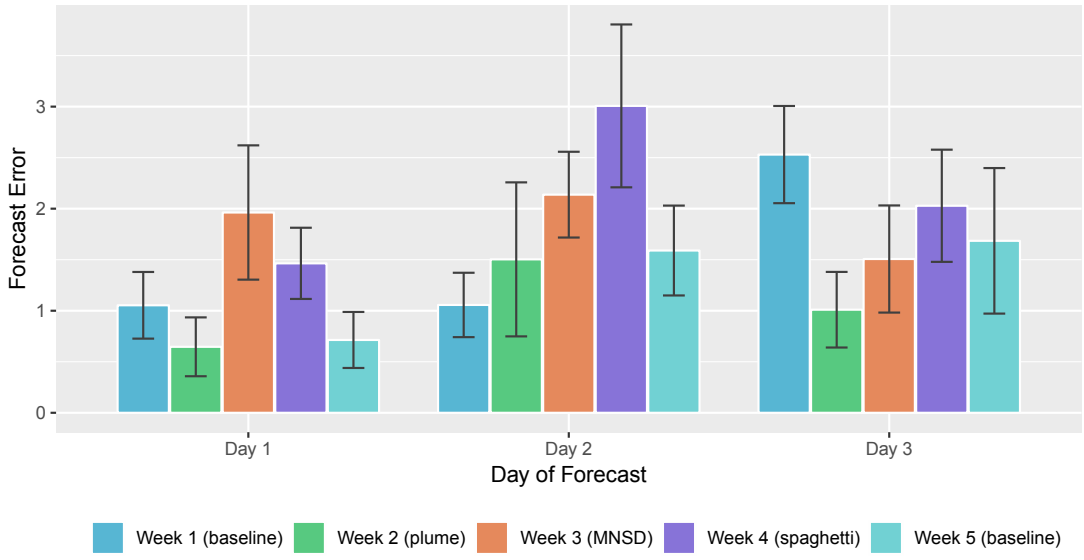


Fig. 4.4. The mean forecast error and 95% confidence interval for each study week and forecasted day. The model shows significant main effects across both weeks and forecasted days, as well as significant week:day interactions.

ear mixed effect models are generally robust to violations of normality [191], this fact, combined with the small number of subjects in our study, suggests a need for caution in interpreting inferences using this model. The implications of this violation are discussed further in Section 4.5.

The model reveals significant main effects where the week that we provided participants with spaghetti plots resulted in significantly less accurate forecasts compared to the original baseline week ($t(414.23) = 3.01$, $p < .01$), and the week that we provided participants with plume diagrams resulted in significantly more accurate forecasts compared to the original baseline week ($t(413.25) = -2.42$, $p < .05$). The planned contrasts also reveal significant linear and quadratic effects across the forecasted days (day.L $t(413.12) = 4.21$, $p < .001$; day.Q $t(413.12) = 2.42$, $p < .05$). These main effects, however, are qualified by several significant interactions between various weeks of the study and both linear and quadratic increases in forecast day (week2:day.L $t(413.32) = -2.22$, $p < .05$; week3:day.L $t(413.19) = -3.92$, $p < .001$; week2:day.Q $t(413.19) = -3.27$, $p < .01$; week3:day.Q $t(413.14) = -2.68$, $p < .01$; week4:day.Q $t(413.12) = -4.56$, $p < .001$; week5:day.Q $t(413.18) = -2.35$, $p < .05$).

We explored these interactions by running the same model on each of the forecasted days individually. This post hoc analysis indicates that for the week the participants were provided plume diagrams there were not significant differences in forecast accuracy compared to the original baseline week for days 1 ($t(135.28) = -1.43$, $p = .15$) or 2 ($t(140) = 1.11$, $p = .27$), but there was a significant increase in forecast accuracy for day 3 ($t(138) = -4.35$, $p < .001$). A similar pattern is observed when participants' forecast accuracy is compared between the second baseline week and the original baseline week ([day1] $t(135.9) = -1.23$, $p = .22$; [day2] $t(140) = 0.80$, $p = .42$; [day3] $t(138) = -2.10$, $p < .05$). For the week that participants were provided with the MNSD plots, on the other hand, forecasts were significantly *less* accurate than the original baseline week for days 1 ($t(135.2) = 3.24$, $p < .01$) and 2 ($t(140) = 2.73$, $p < .01$) and significantly *more* accurate for day 3 ($t(138) = -2.93$, $p < .01$). Finally, during the week participants were provided with spaghetti plots, they were significantly less accurate than the baseline week on day 2 of the forecast ($t(140) = 4.76$, $p < .001$) but showed no significant difference on either days 1 ($t(135.28) = 1.35$, $p = .18$) or 3 ($t(138) = -1.41$, $p = .16$).

4.5 Discussion

As a whole, our analysis suggests that factors unaccounted for in our model contribute to the effects observed in our pilot data. In particular, it seems likely that weather variability impacted forecast difficulty during the study in a way that is confounding the differences found in our analysis. This claim appears to be supported by the fact that a similar pattern of improvement can be seen in both the week that we provided participants with plume diagrams and participants' second 'baseline' week. Additionally, in their poststudy interviews, all of our participants indicated that they largely ignored the provided spaghetti plots. This information suggests that something other than the provided uncertainty product is driving the observed accuracy differences for that week, and the weather is a likely candidate. These observations suggest that weather variability is a larger threat to the validity of our results than either our small number of participants or the normality violation in our model. Without directly addressing this confound, it is impossible to draw any definitive inferences about causation from our data. Accordingly, we do not claim that the results of our pilot study represent a significant contribution.

Our pilot data do, however, contain a couple of trends that deserve closer examination in a follow-up or expanded study. In particular, any subsequent work should look closely at whether plume diagrams actually have a more substantial impact later in the forecast period. During our poststudy interviews, all of our participants noted that they found the plume diagrams incredibly useful in making their forecasts. Given the close alignment of these visualizations' contents with the provided task, there is a strong theoretical basis for why plume diagrams might have more impact when there is more uncertainty.

The multiple indications of a quadratic effect on accuracy for forecasts further in the future should also be investigated. Although we expect that forecast accuracy should vary over the forecast period, there is no intuitive reason why forecasting errors should decrease after increasing. For each of the 3 weeks that we provided participants with uncertainty products, however, the average forecast error peaks on day 2 and improves on day 3. It will be important to assess whether this effect is being driven in some way by the ensemble shown in our uncertainty products. Given that the underlying numerical models are a function of the weather, once again, making that determination will almost certainly require accounting for weather variability.

We see three potential ways to address the problem of weather variability in a follow-up study. One option is to increase the length of the study, averaging out the weather variability through repeated measurements over time. According to a psychology professor familiar with longitudinal analyses, however, we would still need an order of magnitude more participants (around 20) in order to be confident in the stability of the multilevel model estimates. An alternative randomization strategy for handling weather variability would be to stagger the forecasting days across participant groups while counterbalancing when the uncertainty products are provided. Such a study could be run in a shorter amount of time (15 days) but would require even more participants (30–50 participants). The third option is to control weather variability out of the experiment entirely by running a single-day study. The resulting between-subjects design would require the most participants of these three options (around 60 participants).

In all three cases, recruiting the requisite number of participants introduced a cascading set of problems. As a relatively small department, the University of Utah Atmospheric Sciences Department was unlikely to support recruiting more than 10–15 participants with sufficient and similar levels of expertise, meaning that we needed to expand our participant pool. Nothing in our study precludes it from being run online; however, recruiting students from different schools introduces a new problem. Differences in local climatology often lead to different emphases in training. One should expect that student forecasters from the University of Utah may forecast quite differently from student forecasters at Penn State, for example. Climatology, in turn, needs to be accounted for in selecting forecast locations, which also need to be geographically diverse and have similar weather complexity.

After spending several months tweaking and updating our experiment design to account for all of the competing requirements, we eventually realized that running our expanded study was not going to be feasible. We would have had to run a single-day study at multiple universities, each with reasonably large sets of student forecasters (i.e., not the University of Utah). Additionally, the forecast locations and corresponding uncertainty products would need to be picked and generated on the day of the study, in order to properly control for both forecast difficulty and climatology.

4.6 Regarding Increased Realism

Despite the lack of a clear research success story, this project significantly increased our understanding of the various trade-offs between ecological validity and experimental control. Although we argue that there is still a clear need for studies that are more representative of the meteorological forecasting process, there is also a question about the extent to which this need for increased realism in user studies generalizes across domains that involve decision-making. In certain cases, *psychological realism* [21] may present an adequate alternative.

Psychological realism refers to "the extent to which the psychological processes that occur in the experiment are the same as the psychological processes that occur in real life" [21, p. 12]. It is usually contrasted with *mundane realism* or the extent to which the research setting matches the naturalistic environment of the behavior(s) being studied [21]. For testing certain theories, whether an experimental setting matches the naturalistic context of everyday life may be significantly less important than whether the physiological and cognitive processes mediating the causal relationship under investigation are *representative* of what occurs in everyday life.

In this project, our stated goal actually reflects the concept of psychological realism. We were attempting to address the fact that no prior evaluation of uncertainty visualization techniques properly accounts for the complex mental models used in operational weather forecasting or how those mental models mediate the forecasting process. Although it is possible to push even further towards psychological realism, doing so will not necessarily result in a more straightforward experiment. We would still need to ensure, for example, that individuals are extracting information from a visualization, integrating that information into an internal representation using a preexisting framework, and then making judgments based on that internal representation. Still, we have found the basic idea of examining the processes mediating the phenomena one is trying to study incredibly helpful. As the subsequent projects in this dissertation illustrate, focusing on mediating processes can be a useful scaffold for garnering new insights regarding real-world practices.

4.7 Summary

This chapter describes a pilot study aimed at increasing ecological validity in a user study assessing the impact of different uncertainty visualizations on weather forecasts. We discussed the clear need for evaluations of uncertainty visualizations that better reflect complex, real-world decision-making environments, we detailed our attempt at designing and running such an experiment, and we reflected on lessons we learned in the process.

This project does not have a direct impact on the primary contribution of this dissertation; however, the insights about ecological validity and experimental design established during this project significantly influenced studies that do. In this formative role, this project is a core part of a secondary contribution dealing with trade-offs among experimental control, ecological validity, and the generalizability of research results.

CHAPTER 5

EVALUATING THE IMPACT OF BINNING 2D SCALAR FIELDS

This chapter discusses the explanatory quantitative study that we conducted to understand the impact of the meteorological convention to use binned color maps to encode continuous data. As outlined in Chapter 3 Section 3.11, during our formative design study [145], we noticed that the vast majority of meteorological visualizations encode continuous scalar fields using either filled or unfilled contour maps, which conflicts with standard visualization guidance. Given that this is also a common practice in cartography, we were curious whether there might be benefits that were not being accounted for by the existing visualization guidance.

We designed and ran a user study attempting to assess the impact of using discrete color maps to visualize 2D scalar fields across a set of increasingly complex tasks inspired by meteorological practice. The results showed that neither the prediction of the visualization community nor that of the cartographic community was accurate. Instead, the results provided evidence of a speed-accuracy trade-off, suggesting that the choice of a discrete versus continuous color map should be based on a careful consideration of one's tasks.

In this chapter, we provide an overview of the conflicting guidance regarding the use of discrete versus continuous color scales for visualizing 2D scalar fields, we detail our study's methodological choices, and we outline how the results lead to the revision of long-standing guidance in both the visualization and cartographic communities.

5.1 Study Motivation

A foundational design principle in visualization is the *expressiveness principle*, which states that a visual encoding should express all of the relevant relationships in the data, and only the relevant relationships in the data [107], [126]. For visualization contexts centered around sense-making and data exploration tasks, the expressiveness principle implies that

a continuous encoding channel should be used for continuous data. This recommendation is also mirrored elsewhere in the visualization literature [8]. In practice, however, domains such as cartography [171] and meteorology [145] have strong conventions that visualize continuous data with a discrete encoding. These domains rely on visual channels, such as color and saturation, to encode a continuous function defined over two-dimensional space, known as a *2D scalar field*. They commonly do so by employing discrete color maps or contour lines, also called isarithmic maps [171].

The existing literature provides little guidance about either encoding continuous, 2D scalar fields with binned color maps or how this design decision affects data interpretation and decision-making. Research into properties of color maps for encoding continuous data types has largely focused on continuous color maps [8], [117], [152], [186]. This line of research provides guidance on how to capture properties of the data, such as divergence around a center point [186] or emphasis on one end of the data range [8]. These papers go so far as proposing corresponding binned color maps, but do not make claims, or even discuss, their efficacy for continuous data. Work on transfer function design has also proposed methods for binning colors, but with a focus on volumetric scalar fields and with the underlying goal of classifying materials or features [49] as opposed to directly understanding the continuous nature of a scalar field itself.

An exception is work in cartography that explores the efficacy of using binned color maps for encoding continuous values in choropleth maps. This work suggests that bins increase speed and accuracy [66]. However, the tasks that a choropleth supports are different from those of a 2D scalar field, making it unclear whether the results hold more generally across continuous data types. Additional work has focused on contouring 2D scalar fields, such as digital elevation models (DEMs), which involves connecting points of equal value with isolines. Although pragmatic advice is offered for determining distances between isolines [199], no empirical studies have evaluated the impact of contouring on decision-making.

This work aimed to test whether the decision to violate the expressiveness principle when visualizing 2D scalar fields influences decision-making across a broad spectrum of tasks. To this end, we conducted a user study in which participants were provided with continuous or binned grayscale encodings of the same topographical 2D scalar field. Then

we tested participants' abilities to complete a range of increasingly complex tasks in a geospatial context.

The primary contribution of this work is our study results, which demonstrate that decisions made with binned encodings not conforming to the expressiveness principle were either as accurate or more accurate than those that followed the principle and employed a continuous encoding. This finding revises long-standing guidance within the visualization community by showing that strict adherence to the expressiveness principle is not always advisable for visualizing 2D scalar fields. A secondary contribution was the expanded task-set that we created for the study. We selected these tasks based on both Brehmer and Munzner's visualization task typology [16] and actual behaviors observed during our formative design study [145], with the goal of increasing both the tasks' complexity and their relevance to real-world problems.

In the remainder of this chapter, we lay out an overview of existing work on continuous versus binned color maps, a description of our study and results, and a discussion of the findings and their implications for visualization. Section 5.2 includes discussions from cartography regarding the decision to use continuous or segmented mappings and briefly reviews the visualization work on encoding data using color and binning data effectively. We follow this with a detailed description of the experimental setup for our study and the results in Section 5.3. Finally, Section 5.4 discusses the implications of the findings, both generally and in the context of our larger mixed-methods research program.

5.2 Perspectives on Binning Continuous Scalar Data

Work in cartography and visualization offers inconsistent views on binning continuous data. Modern approaches in cartography advocate for binning the data, yet visualization suggests that a continuous channel should encode the data. Given the disagreement in the literature, this section details research from both disciplines as motivation for the study we conducted.

5.2.1 The Cartographic Perspective

A prevalent assumption in cartography regarding choropleth maps is that binned color encodings of quantitative data are more comprehensible than continuous encodings [66],

[199]. This perspective started a lively discussion in 1973 when Waldo Tobler developed a method of creating continuous grayscales using a line plotter [185]. Employing Tobler's method, cartographers had the option between continuous grayscales or discrete tone encodings, and thus a heated debate ensued [171]. Jean-Claude Muller was a key advocate for the new continuous method, conducting various studies demonstrating that on printed maps generated using a facsimile method [124], users could discern similar map features with continuous encodings that are discernible by binned encodings [122], [123]. Muller's and Tobler's work was contested notably by Michael Dobson in a series of commentaries [42], [43] in which he detailed a laundry list of grievances with continuous encodings of choropleth maps, including increased time to perform tasks, reduced ability to control the amount of information presented, decreased pattern discrimination, and increased mental workload [42]. Although these specific findings may not necessarily generalize directly to either 2D scalar field visualizations or modern display technologies, the perspectives exemplify relevant considerations for the trade-offs between the richness of presented data and usability.

Many of the tasks performed with choropleth maps are fundamentally different from those performed with 2D scalar field visualizations, as the latter are often used to make judgments about continuous features in the data. Thus, it is unclear to what extent standard practices used for binning choropleth maps extend to 2D scalar fields. Other work in cartography and geospatial information systems (GIS), however, has focused on specifications for contouring 2D scalar fields such as digital elevation models [199]. This work suggests that contouring aids in interpreting continuous features [62], [76]. Contouring involves defining spans between isolines and, in the case of contour maps, encoding the span between isolines with a discrete encoding such as color or grayscale. Since the resulting visualizations can be similar between contouring and discrete binning, there is likely a relationship between tasks that both methods support. Defining the appropriate relative distance between isolines is domain-specific and commonly a judgment call by the designer. Watson recommends that "by looking at the range of heights in your dataset, you will get an idea of the levels that suit your data" [199, p. 11]. More advanced techniques in GIS include methods for generating contours for a given task based on classifications of data features such as peaks or valleys (for a review see [62]).

Various techniques for representing surface topology as graphs have also been proposed, notably the *Reeb graph* [167] and the *Morse-Smale complex* [36]. These approaches represent the relationship of critical topological features as nodes in a network visualization [99]. However, to our knowledge, no user studies in GIS have been conducted to assess the influence of discretization techniques on decision-making.

5.2.2 The Visualization Perspective

As outlined in Chapter 2, an extensive body of research within the visualization literature considers the role of color in data encoding. Here, we summarize work concerning decisions and issues with encoding continuous data with color.

Previous work in visualization concerning color-encoding channels primarily recommends a binned color map for categorical and ordinal data and a continuous one for quantitative data [8], [117], [186]. Much of the work on discrete color maps centers around Brewer's perceptually optimized color maps [19], [20], [66]. These color maps are widely adopted [213] and have inspired guidelines [212] and generative systems [203]. Alternatively, Healey outlines a systematic method for generating effective isoluminant color schemes with up to seven colors by controlling for color distance, linear separation, and color category [67].

For continuous color maps, the literature primarily focuses on establishing and using principles for color-map design to create better defaults [119] or domain-specific recommendations [145]. However, it is generally accepted that there is no one best color scale and that a variety of factors dictate effective color usage [152]. For example, Ware offers experimental evidence that hue variation is better for extracting absolute values, whereas luminance change enables enhanced form comprehension in continuous 2D scalar fields [195]. Bergman et al. provide continuous color-map recommendations based on data class, task, and spatial frequency [8]. Rheingans outlines various design considerations for creating successful continuous color maps: goals, the nature of the data, the intended audience, the visualization as a whole, and cultural connotations [152]. Additionally, Tominski et al. [186] provide a formalized set of task-driven continuous color-map recommendations based on characteristics of the data, tasks and goals, the user, and the display.

Where principles for continuous color-map design have been generalized to corre-

sponding binned color maps, the discussion is limited to applying those color maps to nominal or ordinal data [117], [186]. Bergman et al. [8] provide one of the only explicit discussions about encoding continuous data discretely. They argue that using a binned color map on continuous data is useful for *segmentation* tasks, where the explicit goal is to show multiple features. There is, however, no discussion of what effect a binned encoding might have on *isomorphic* tasks, where the goal of the representation is to reflect the structure in the data. Additionally, as Mittelstädt et al. point out, many of the existing guidelines for creating effective color maps are geared toward a single elementary task, whereas real-world scenarios typically require more complex, combined tasks [117]. The goal of our work is to look at the influence of binned color maps across a wide range of both segmentation and isomorphic tasks for 2D scalar fields.

Work on visualizing continuous functions in three-dimensional (3D) space using color mainly focuses on techniques for volume visualization. As outlined in Chapter 2 Section 2.1.1, transfer functions for volume visualization play a classification role in 3D scalar fields that closely resembles the Bergman et al. notion of segmentation tasks [49], since the goal in both cases is to elucidate features. Although various attempts have been made to automate or optimize the data classification step in transfer functions [54], [85], [193], the results do not readily support isomorphic tasks, which we also wanted to test.

5.3 Methods and Aims

Even though the expressiveness principle suggests that continuous data should be encoded with a continuous channel, some visualization practitioners have likely discovered, through experience, that binning continuous data types is beneficial in various situations. We argue that violations of the expressiveness principle should be used to calibrate efficacy. Visualization efficacy is domain-specific, but can generally be gauged by speed, accuracy, and confidence in task performance. Focusing on the geospatial domain, we predict that forgoing the expressiveness principle for the 2D scalar field data used in the current study will likely increase accuracy and confidence in task performance, which is consistent with work suggesting that binned choropleth maps are preferable. However, binning may reduce speed, showing a classic speed-accuracy trade-off.

To test our hypothesis, we employed a between-subjects design, with five groups of

participants each performing tasks with one of five binning conditions. Each of the binning groups was comprised of participants from Amazon's Mechanical Turk with participation criteria set to master-class workers in the US. The average screen size used by the population in this study was 1600 x 900 cm, the median was 1280 x 1024 cm, and the mode was 1366 x 768 cm. The five binning conditions, shown in **Fig. 5.1**, **Fig. 5.2**, **Fig. 5.3**, **Fig. 5.4**, and **Fig. 5.5**, include one perceptually linear continuous grayscale encoding, along with four approximately perceptually linear binned grayscale encodings. These were based on binning intervals ranging from fine to coarse. Type of task was a within-subject variable, with each participant completing 12 randomly ordered tasks ranging in complexity.

5.3.1 Stimuli Generation

We created the experimental stimuli from a DEM projected into the continuous range 0–100. This range allowed us to start with a dataset where the data range explicitly matched the range of the lightness channel, L^*, in the commonly used CIELAB color space. CIELAB was designed to be perceptually uniform when evaluating small color differences [175]. These differences are defined based on a Euclidean distance metric, where one unit of Euclidean distance ($\Delta\ E^*_{ab} = 1$) was designed to equal one just-noticeable color difference (JND). Other work, however, suggests that $\Delta\ E^*_{ab} = 2.3$ is a more accurate benchmark for just-noticeable color differences in the CIELAB color space [109]. Although more accurate color difference spaces and color appearance models exist [121], [175], the simplicity of CIELAB has made it the most practical choice for designers [51], [180].

We generated the perceptually linear continuous grayscale encoding by mapping the elevation data values to CIELAB lightness, L^*, while leaving $a^* = b^* = 0$. We utilized the Mahy et al. [109] benchmark of $\Delta\ E^*_{ab} = 2.3$ to create four binning conditions. The smallest bin was *nearly-continuous* ($\Delta\ E^*_{ab} = 2.3$), which resulted in a set of bins approximately 1 JND apart. The three coarser binnings were then generated by doubling ($\Delta\ E^*_{ab} = 4.6$), tripling ($\Delta\ E^*_{ab} = 6.9$), and quadrupling ($\Delta\ E^*_{ab} = 9.2$) the initial bin size. The resulting color-mapped stimuli are shown in **Fig. 5.1**, **Fig. 5.2**, **Fig. 5.3**, **Fig. 5.4**, and **Fig. 5.5**. As none of these quantities neatly divides into the data range, the lowest elevation bins were smaller and variable in size. Therefore, we specifically avoided asking questions regarding the lower elevations as this inconsistency could have influenced participants' responses.

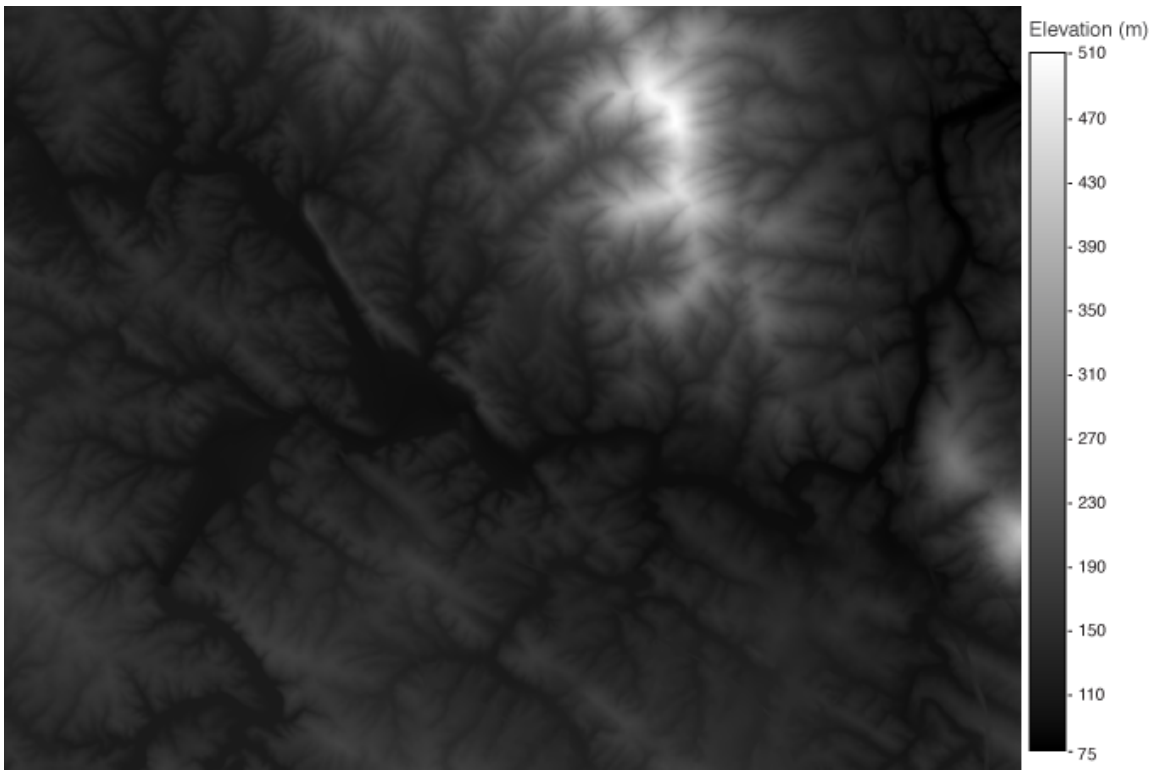


Fig. 5.1. Experimental stimulus showing the Continuous binning condition.

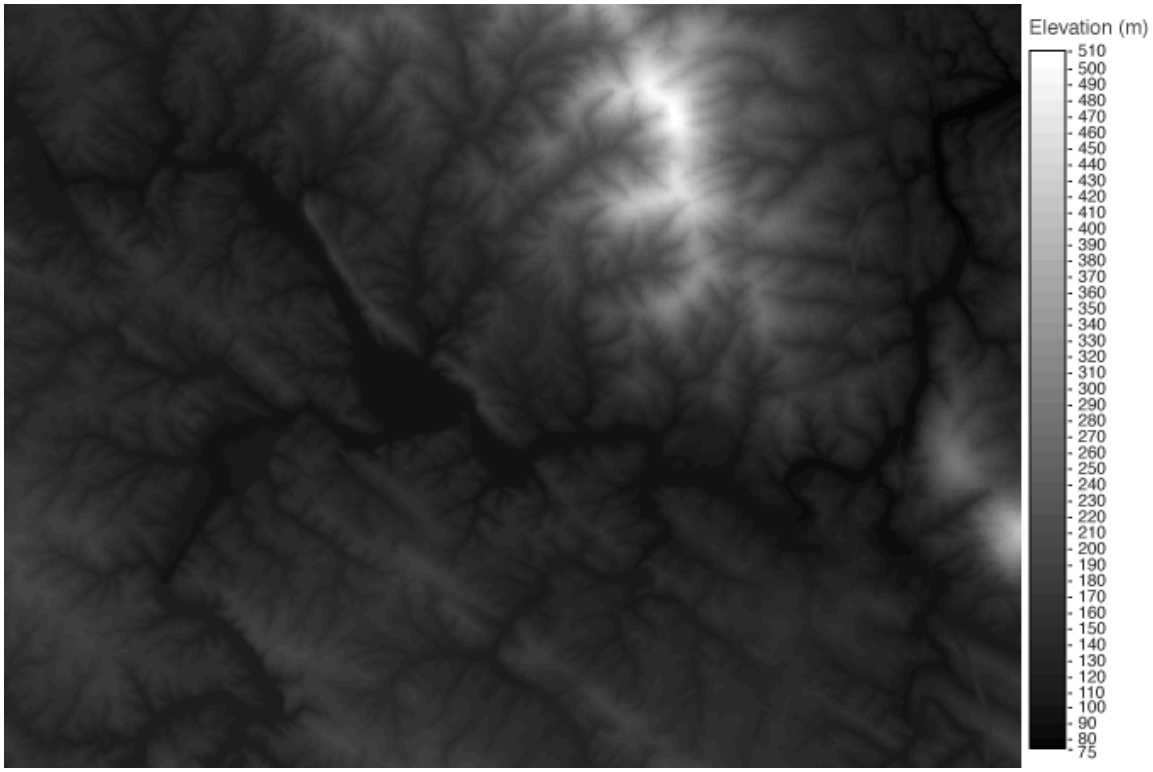


Fig. 5.2. Experimental stimulus showing the 10m binning condition.

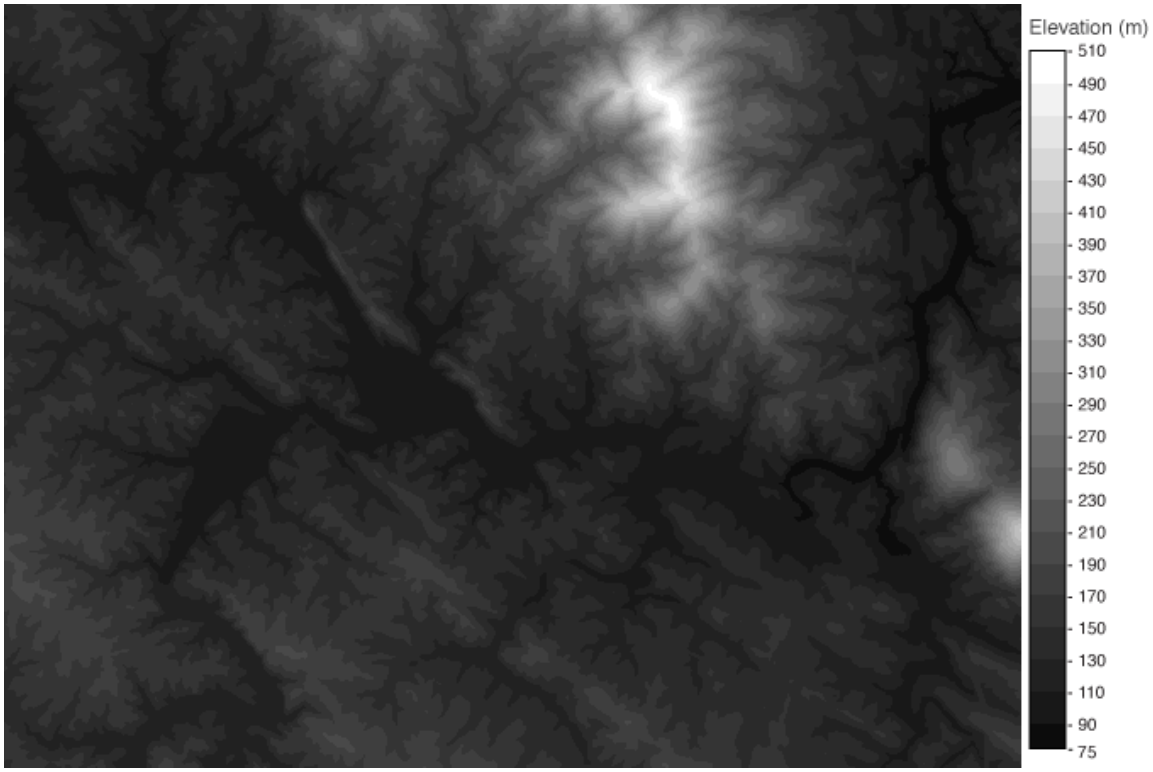


Fig. 5.3. Experimental stimulus showing the 20m binning condition.

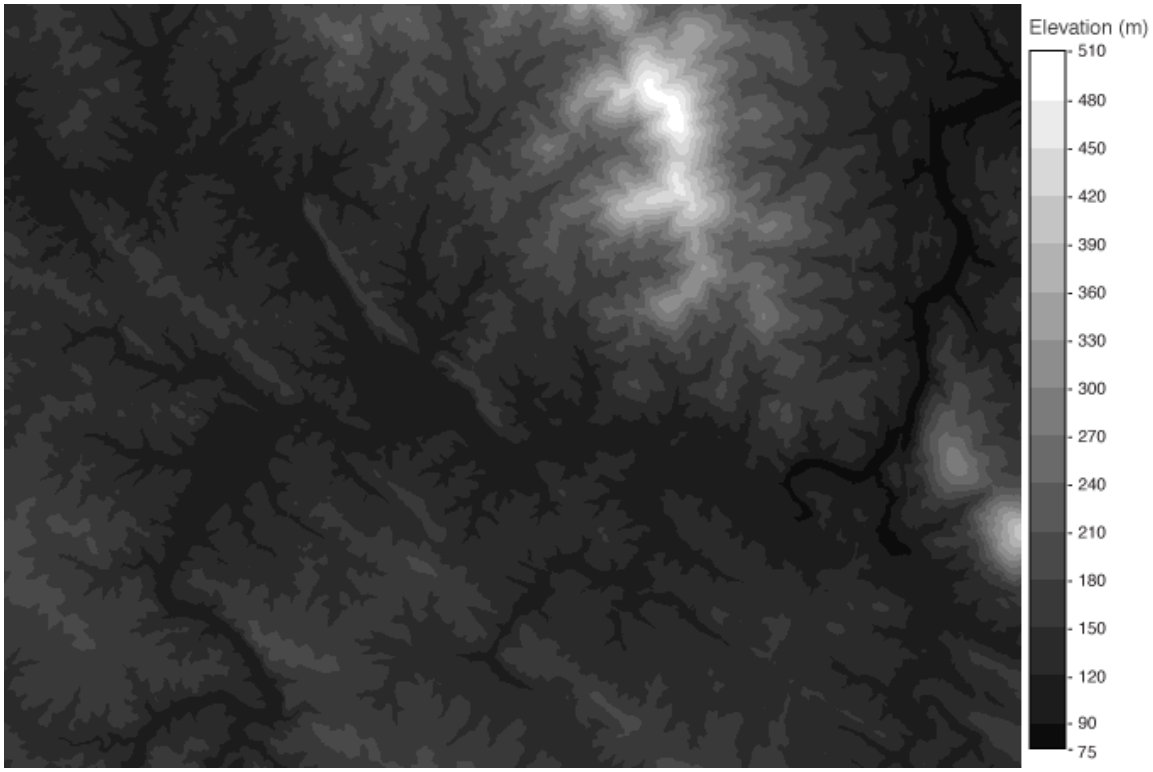


Fig. 5.4. Experimental stimulus showing the 30m binning condition.

Fig. 5.5. Experimental stimulus showing the 40m binning condition.

Grayscale is generally considered to have a limited perceived dynamic range, and approximations for the number of JNDs vary [95], [152]. The data range of our continuous encoding matches the theoretical maximum of 100 JNDs [152], and we base our finest binning on the most conservative estimate of 43 JNDs [109].

It should be noted that we did not account for perceptual effects such as *simultaneous contrast* that have been well documented for grayscale [195]. Recent work by Mittelstädt et al. has proposed and subsequently improved upon postprocessing methods that counter contrast effects [116], [118]. However, prioritizations based on task are required. Consequently, we could not appropriately optimize a single image for all tasks using this technique.

The CIELAB pixel values were converted to sRGB to create the images used in the study. The sRGB color space is nonlinear standardized RGB specification, which is widely accepted as the default RGB specification in web-imaging [175]. Given that we ran the study online, sRGB was a logical choice.[3]

For those stimuli requiring defined regions, the regions were layered on the sRGB images via postprocessing (e.g., **Fig. 5.6**). Additionally, to ensure that our visual stimuli more accurately matched the nature of our questions, we linearly scaled the legend values such that $\Delta\ E^*_{ab} = 2.3$ equates to a 10-meter difference in elevation. The five binning techniques or test conditions are named based on this scale (10m n = 100, 20m n = 99, 30m n = 99, 40m n = 100, and Continuous n = 101). From the participants' point of view, our questions regarded topographical data spanning a range of 75–510 meters. We collected responses using this range and for appropriate tasks transformed the response values back to the original 0–100 value range for analysis.

5.3.2 Tasks

This study examines how visualization techniques that forgo (i.e., binned encodings) and conform (i.e., continuous encoding) to the expressiveness principle influence the general public's interpretations of data. Brehmer and Munzner [16] provide a thorough typology of visualization tasks that we selected various tasks from to test this aim. We focused

[3] An alternative choice would have been to try to utilize the parametric color difference model proposed by Szafir et al. [180], which was designed to incorporate the real-world perceptions of target populations on services such as Amazon's Mechanical Turk.

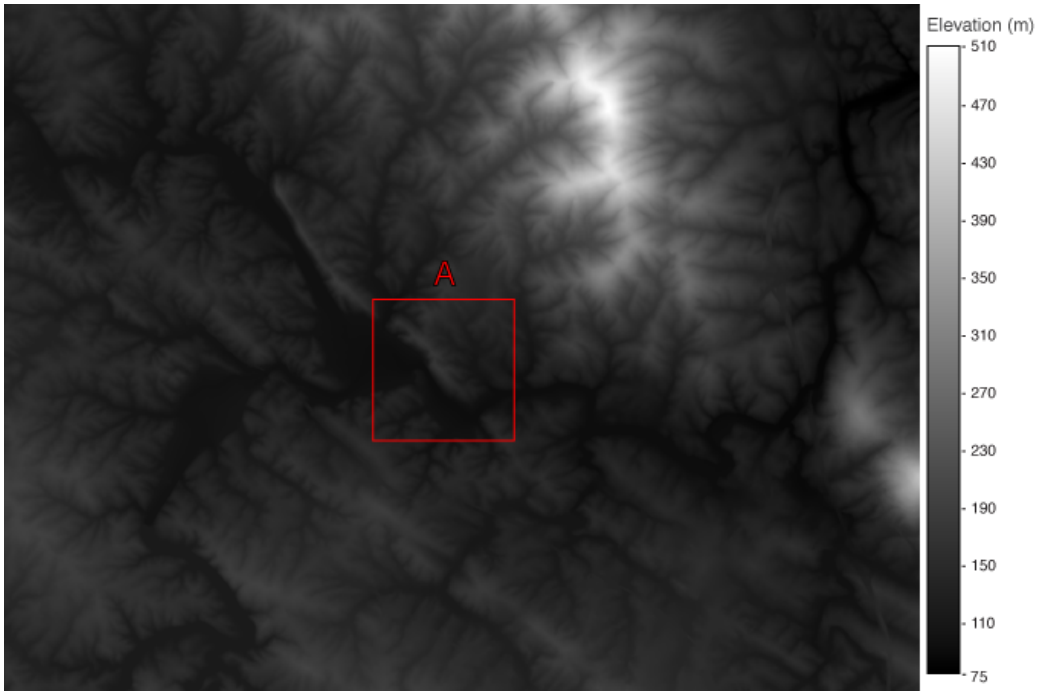


Fig. 5.6. Example Continuous condition stimulus for the Highest Point in Region task.

on discovery-based tasks, which involve generating predictions and testing theories in the data. This categorization of tasks incorporates *search* and *query* tasks. Search tasks can be categorized as *lookup* (when both the target and location are known), *browse* (when the target is unknown, but the location is known), *locate* (when the target is known, but the location is unknown), and *explore* (when neither the target nor the location is known). When the target or set of targets has been located, then a query step is performed, where the participants *identify*, *compare*, or *summarize* the targets. We did not address the *how* and *what* sections of Brehmer and Munzner's typology [16] since these are most relevant to interactivity, which falls outside of the scope of the current work. Further, we were interested in assessing performance on both elementary (on single values) and synoptic (on sets of values) tasks [6],[186]. We utilized comparisons, rankings, and summarizations of regions to this aim.

In an extensive initial pilot study, we tested six tasks that were chosen to represent a broad spectrum of possible visualization tasks ranging in complexity. We were particularly interested in including tasks that incorporated the integration of multiple features since we

had observed users attempting such tasks in applied settings in our previous work [145]. In the simplest task, participants were instructed to click on the highest location on a map, which was intended to test the participants' ability to locate relevant values with the different binning techniques. Prior work has found that participants are quite accurate at mouse cursor location precision, with errors ranging from 2–3% [11], making mouse clicks a reasonable choice of response type. In the most complex task, participants were asked to rank locations according to flood risk, which involved integrating both elevation information and distance from a river. This task was both complex and modeled after real-world scenarios. Although our pilot study was informative, the problem was that glyphs were placed on the visualizations to indicate specific points, which occluded relevant values. For example, two glyphs were placed on the visualizations, and participants were asked to compare the elevation of each location indicated by the glyphs. The present study addressed this issue by asking participants to make relative judgments about regions rather than occluded points. Additionally, the present work uses mouse-click responses and comparisons because they were the most successful task types in the pilot study. We further expanded the spectrum of tasks to include previously untested variants of the aforementioned tasks along with asking participants to summarize fields of data.

For all tasks, participants were shown one topographical map with a single binning technique. Time to complete each task was recorded along with a subjective measure of confidence in task performance. After each task, participants were presented with a 1–7 Likert scale to measure their confidence (1 = not at all confident, 7 = very confident). Additionally, we asked participants questions about monitor display capacities, including black level, white saturation, and gradient banding. We include screenshots of each task and the display questions in the supplemental materials for this dissertation.

5.3.2.1 Locate or Explore and Identify Tasks

Depending on a participant's approach, the following tasks could be completed by either a *locate* or *explore* strategy followed by an identification step, according to Brehmer and Munzner [16]. For example, if a participant identified a relevant value on the legend (e.g., the highest elevation) and then looked for that value, this would be a locate task — seeking for a known target in an unknown location. If, instead, the participant searched

for characteristics in the data that are unknown targets (e.g., relative height or slope), it would be an exploration task [16]. In either case, the participants subsequently identified the target via mouse click.

1) **Highest Point**. Participants were asked to "Click on the highest point in the image."

2) **Highest Point in Region**. Participants were asked to "Click on the highest point inside box A." This task was intended to be a more complex version of the former task, by prohibiting participants from distinguishing the highest elevation on the legend and then identifying that location on the map. Instead, participants had to locate the highest point within a region, which was outlined in red and overlaid on the map (see **Fig. 5.6**). All tasks that required responses in regions used the same type of red outline overlay. Although participants could have looked up values in the legend, this task did not require this strategy and likely primarily depended on participants' ability to make relative comparisons between elevations within the region.

3) **Locate 275m**. Participants were asked to "Click on any point in the image that has an elevation of 275 meters." This task was intended to require participants to use the legend then search in the visualization for the relevant value.

5.3.2.2 Locate or Explore and Compare Tasks

For these tasks, participants could use either a locate or explore strategy, followed by a comparison of targets.

4) **Compare Heights**. Participants were asked to "Find the highest point in each box. Comparing these locations, which box contains the higher location?" This question was a comparison task that could be implemented without occluding the relevant locations on the map (which would have occurred if specific points were overlaid on the map). Although allowing participants to remove and display a glyph interactively would have been one solution to this occlusion issue, we felt that it was important to limit the scope of these tasks to static images. Thus, we created a task requiring participants first to identify the highest location in two regions and then to compare those values.

5) **Rank Heights**. Participants were asked to "Find the highest elevation in each box. Then rank each box according to the elevation of its highest location." This task was intended to increase complexity from the Compare Heights task. Rather than comparing the highest locations in two regions, the participants were instructed to rank the highest locations in three regions.

5.3.2.3 Explore and Identify Task

Exclusively explore tasks require participants to make judgments using characteristics of the data rather than known targets. These are some of the most difficult and rarely employed tasks with the general public.

6) **Steepest**. We tested whether participants could identify the relationship between points by using the colloquial understanding of slope. Participants were provided the following description of slope: "Terrain is steeper where there is a larger change in elevation between adjacent points." In this task, participants were then asked to "Click on the steepest point in the map."

5.3.2.4 Lookup and Identify Task

Lookup tasks are characterized by both the target and location being known.

7) **Lookup**. Participants were asked to identify the height values adjacent to the point they selected in the Steepest task. This question was presented on the same page as the Steepest task, and asked, "What are the highest and lowest elevations adjacent to the point you clicked on?"

5.3.2.5 Explore and Compare Task

These tasks were intended to assess participants' ability to compare relationships in the data through exploration, using the Steepest task outlined above.

8) **Compare Steepness**. Participants were provided with the instructions, "In each box, find the steepest point. Comparing these locations, which box contains the steepest point?"

9) **Rank Steepness**. This task was one of the most difficult tasks that we tested. Participants were given the instructions, "Inside each box, find the steepest point. Rank each box according to their steepest points."

5.3.2.6 Browse and Summarize Task

Browse tasks are characterized by the location being known and the target being unknown. The following browse task used specific regions within which the participants were asked to determine certain characteristics of the region:

10) **Average Height**. This task was intended to test participants' ability to integrate a large set of relationships in the data. We asked participants to provide a summary statistic from a specified region. Participants were asked, "What is the average elevation of the area in box A?" All participants were screened at the start of the experiment for a basic understanding of how to calculate an average from a string of numbers. Although calculating numeric averages is fundamentally different from visual-spatial aggregation, we wanted to ensure that participants were familiar with the term average and related concepts.

5.3.2.7 Browse and Compare Task

In this task, participants summarized two regions and then compared the regions.

11) **Compare Average Height**. Participants were provided with the instructions, "Find the average elevation in each box. Comparing these locations, which box contains the higher average elevation?"

5.3.2.8 Combination Task

The final combination task involved a multistep iterative process that could include all or a subset of the previously defined processes in any order, depending on a participant's strategy.

12) **Steepest at 355**. This question was intended to increase complexity and relevance by combining a locate task and a browse task. Many tasks that are performed in real-world scenarios require combinations of tasks. Participants were asked to "Click on the steepest point in the map at an elevation of 355 meters."

5.3.3 Analysis

The major takeaway from the analysis was that in no cases was there a clear advantage in terms of accuracy with the continuous encoding. The results, described in more detail below, show that for some tasks there was no difference in accuracy between binned and continuous encoding (e.g., Compare Heights and Rank Heights tasks), but in other tasks, binned encodings showed higher accuracy than the continuous encoding (e.g., Highest Point and Lookup tasks). Interestingly, however, we found that participants were generally faster to complete the tasks with the continuous encoding than with some of the binned encodings. These findings suggest that although participants with continuous encodings appear to complete tasks more quickly, continuous encodings did not improve accuracy. The following sections detail the analyses and results for each task.

For the following ANOVA analyses, post hoc Tukey HSD comparisons were used to test significant differences between binning techniques for each task, correcting for multiple comparisons. In addition, all analyses controlled for effects of display parameters (see the supplemental materials archive for the full output). The task accuracy analyses and results are presented for each task separately in Sections 5.3.3.1 through 5.3.3.8, since the nature of the task determined the analyses used. Analyses and results for task completion time and confidence ratings, summarized across tasks, are reported in Sections 5.3.3.9 and 5.3.3.10.

5.3.3.1 Locate or Explore and Identify Tasks

For the **Highest Point** task, participants were more accurate when using the 40m binning display compared to the continuous display. This result was revealed by a one-way between-subjects ANOVA testing the effect of binning technique on accuracy of click location. Click accuracy was calculated by subtracting the L* value at the clicked location from the L* value of the highest location (100), creating an error score in CIELAB space distance. Outliers that were two standard deviations (SD) above the mean were removed (6% of trials). The analysis revealed there was a significant effect of binning technique on click location accuracy, $F(4, 463) = 2.634$, $p = .033$, $\eta_p^2 = .021$. The mean error score for continuous $(M = 4.93, SD = 4.55)$ was significantly higher than 40m $(M = 3.25, SD = 3.06)$ at a $p < .05$ criterion (see Fig. 5.7).

For the **Highest Point in Region** task, participants were more accurate using the 20m

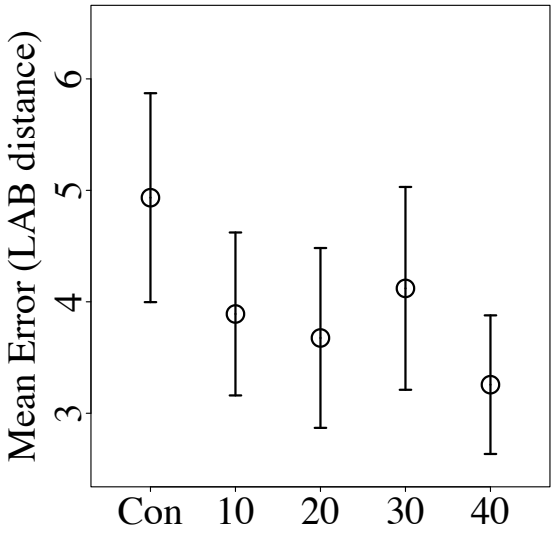


Fig. 5.7. Highest Point task accuracy. Error bars represent 95% confidence intervals.

binning technique, specifically, when compared to the 30m binning. This result was revealed by a one-way between-subjects ANOVA testing the effect of binning technique on accuracy of click location within a region. Click accuracy was calculated by subtracting the L* value at the clicked location from the L* value of the highest location in the region (22.40), creating an error score in CIELAB space distance. Outliers $2SD >$ mean were removed (3.8 % of trials). The analysis revealed there was a significant effect of binning technique on click location accuracy, $F(4, 472) = 2.744, p = .028, \eta_p^2 = .023$. The mean error score for 20m $(M = 2.10, SD = 2.04)$ was significantly lower than 30m $(M = 2.86, SD = 1.89)$ binning, $p < .05$ (see **Fig. 5.8**).

For the **Locate 275m** task, participants were more accurate using the 40m binning technique, specifically when compared to the 30m binning. This result was revealed by a one-way between-subjects ANOVA testing the effect of binning technique on accuracy of clicking a location at an elevation of 275 meters. Click accuracy was calculated by subtracting the L* value at the clicked location from the L* value corresponding to an elevation of 275 meters (45.95), creating an error score in CIELAB space distance. The analysis revealed that there was a significant effect of binning technique on click location accuracy, $F(4, 491) = 2.763, p = .027, \eta_p^2 = .021$. Mean error score for 40m $(M = 18.77, SD = 12.46)$ was significantly lower than 30m $(M = 24.12, SD = 12.87)$ binning, $p < .05$ (see **Fig. 5.9**).

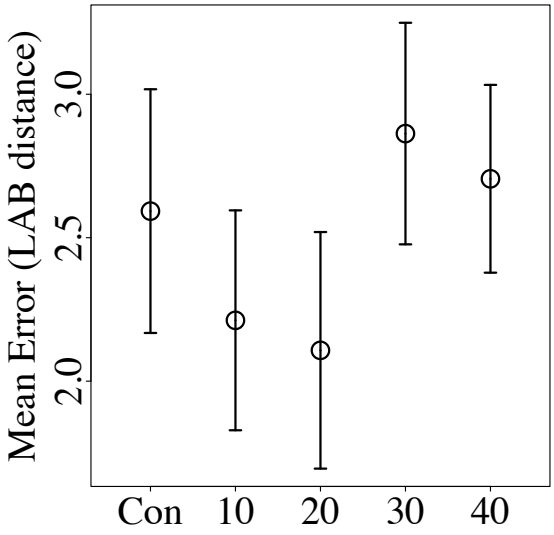


Fig. 5.8. Highest Point in Region task accuracy. Error bars represent 95% confidence intervals.

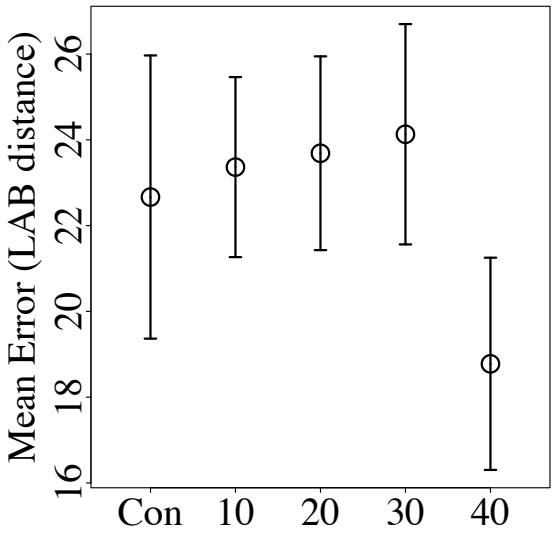


Fig. 5.9. Locate 275m task accuracy. Error bars represent 95% confidence intervals.

5.3.3.2 Locate or Explore and Compare Task

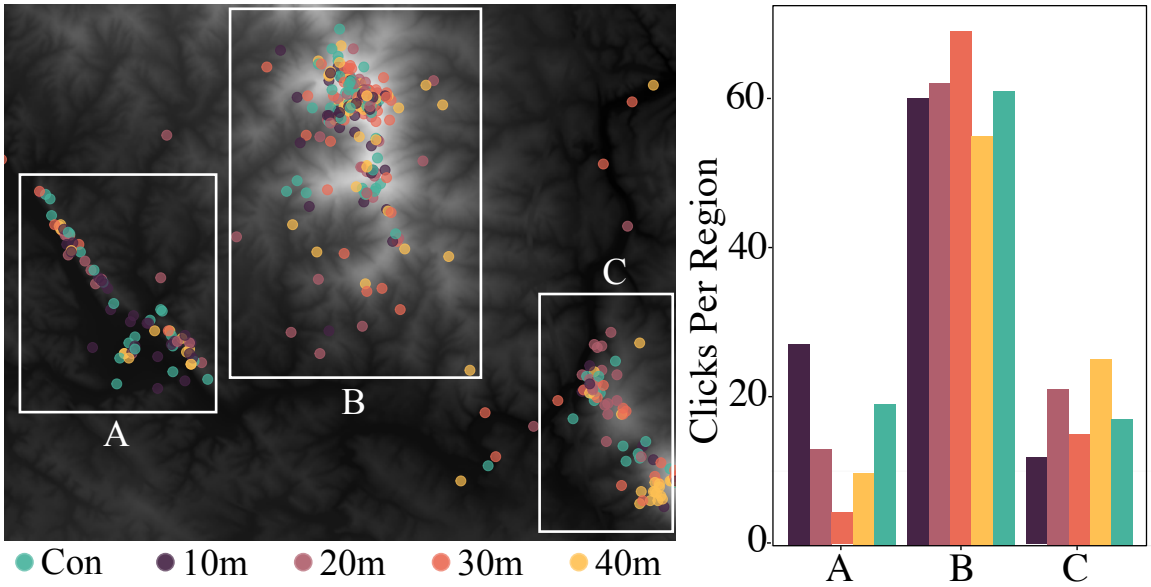
For the **Compare Heights** task, a binomial logistic regression found that there was no significant effect of binning technique on height comparisons, $\chi^2(df = 7) = 7.189, p = .40$. Participants were overall more likely to correctly choose the higher elevation (73% responded correctly), and this did not vary across the binning conditions.

For the **Rank Heights** task, an ordinal logistic regression was used to test the effect of binning technique on rankings of regions by height. Participants responded to this question by entering rankings of 1-3 (3 indicating the highest region and 1 the lowest) for regions A, B, and C. The ordinal logistic regression equation did not significantly predict rankings when using binning technique and region as predictors, $\chi^2(df = 9) = 10.63, p = 0.30$. In general, participants performed poorly on this task: 35% correctly ranked B as the highest, 52% correctly ranked A second, and 35% correctly ranked C as lowest, which makes it difficult to compare the binning conditions.

5.3.3.3 Explore and Identify Task

For the **Steepest** task, preliminary diagnostic checks of the data revealed that participants predominantly clicked in three separate areas, as shown in **Fig. 5.10a**. We assessed steepness by looking at the DEM's gradient, which we approximated using a Sobel operator. The point with the highest gradient magnitude (44.33) was located in area A. However, this point was located proximal to a low elevation depression in the map. The vast majority of clicks were in area B, which was proximal to a peak feature. This pattern of responses suggests that participants' click decisions may have been influenced by an assumption about the relationship between steepness and peaks. Because of this initial pattern, we decided to provide a qualitative description of the number of clicks per binning condition in each area separately. As shown in **Fig. 5.10b**, there are different proportions of total clicks in regions A and C across binning conditions. We can speculate that different binning techniques influenced some participants' incorrect assumptions relating elevation and steepness; however, future studies are needed to understand these effects fully.

An additional complicating factor when asking participants to make judgments of steepness comes from the high spatial frequency of the gradient magnitude (see **Fig. 5.11**). In the areas with the highest gradient magnitude, relatively close points (i.e., only a few



(a) Click locations overlaid on the continuous encoding, showing three main regions.

(b) Number of clicks per region by binning technique.

Fig. 5.10. Breakdown of participants' Steepest task click responses.

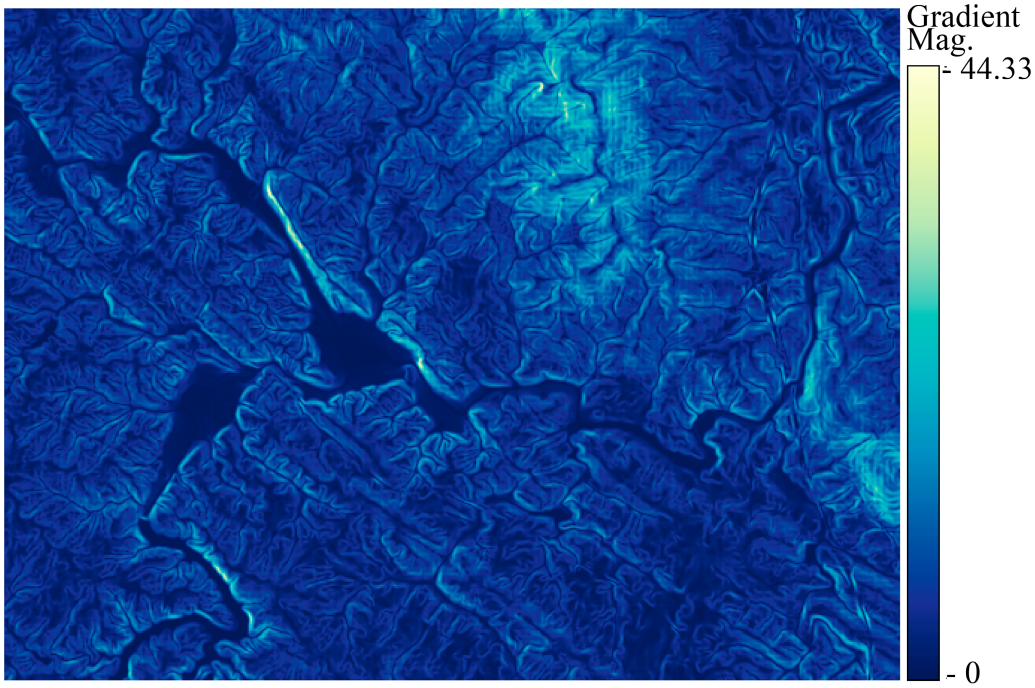


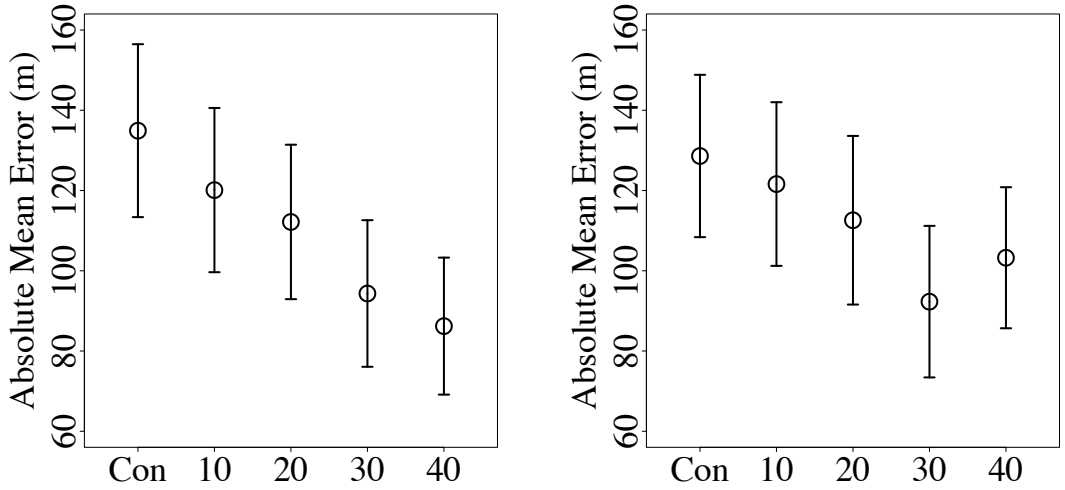
Fig. 5.11. A visualization of the spatial frequency of the DEM used in the study.

pixels apart) could have vastly different gradient magnitude values. Future work should investigate strategies to account for any potential problems caused by this issue.

5.3.3.4 Lookup and Identify Task

The **Lookup** task followed the Steepest task and asked participants to report the lowest and highest values adjacent to their click; thus, the accuracies of the lowest and highest points were analyzed separately. For the lowest value, participants were the least accurate using the continuous encoding, specifically when compared to the 30m and 40m binnings. Accuracy was calculated by subtracting the reported lower adjacent elevation from the actual lower adjacent elevation, creating an error score in CIELAB space distance. Outliers greater than 2 SD above the mean were removed (7.8% of trials). A one-way between-subjects ANOVA controlling for reported higher adjacent elevation showed there was a significant effect of binning technique, $F(4, 451) = 4.418, p = .001, \eta_p^2 = .057$. The mean error score for the continuous encoding $(M = 134.91, SD = 105.20)$ was significantly less accurate than the 30m binning $(M = 94.33, SD = 87.21)$ and 40m binning $(M = 86.20, SD = 82.88)$, $p < .05$ (see **Fig. 5.12a**).

A similar analysis was performed on the reported upper adjacent elevation. Outliers greater than 2 SD above the mean were removed (2% of trials). There was a significant effect of binning technique on reporting the upper adjacent elevation, $F(4, 479) =$



(a) Lowest adjacent elevation

(b) Highest adjacent elevation

Fig. 5.12. Lookup task accuracy. Error bars represent 95% confidence intervals.

2.602, $p = .03, \eta_p^2 = .044$. However, post hoc Tukey HSD comparisons did not reveal significant differences between the conditions at the $p < .05$ level when accounting for multiple comparisons. To understand the main effect of binning, we generated planned contrast codes that compared the continuous encoding to the binned encodings. Similar to the lower elevation analysis above, using a one-way between subjects ANOVA, we found that the continuous encoding $(M = 128.60, SD = 100.90)$ was less accurate than the 30m binned encoding $(M = 107.36, SD = 97.16)$, $F(1, 490) = 7.79, p = 0.005$ (see **Fig. 5.12b**).

5.3.3.5 Explore and Compare Task

For the **Compare Steepness** task, a binomial logistic regression found that there was no significant effect of binning technique on a steepness comparison, $\chi^2(df = 7) = 1.35, p = .98$. Area A contained the steepest point with a magnitude gradient of 44.33, and Area B contained the second steepest point with a magnitude gradient of 41.65 (see **Fig. 5.13**). However, 62% of participants incorrectly selected Area B as containing the steeper point. Similar to the Steepest task, these findings suggest that participants' prior understanding

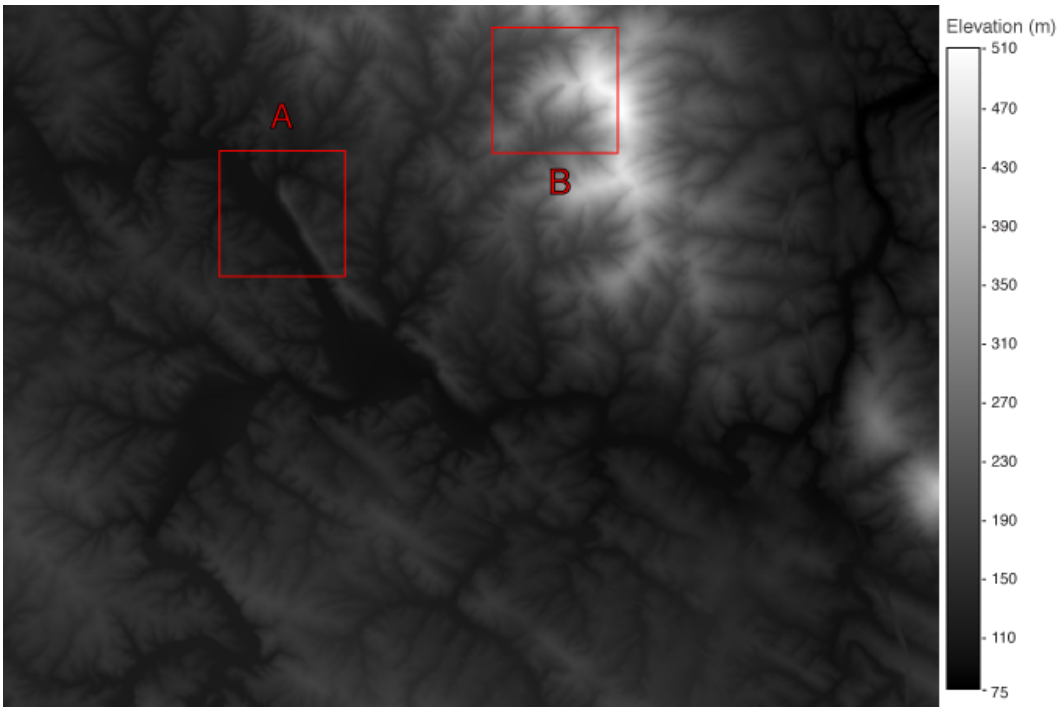


Fig. 5.13. Example Continuous condition stimulus for the Compare Steepness task, showing areas A and B.

of topography and an assumption about a connection between steepness and peaks could have biased the incorrect responses. Additionally, these findings may be influenced by the issues related to the high spatial frequency of the gradient magnitude noted in Section 5.3.3.3.

The **Rank Steepness** task used an ordinal logistic regression to test the effect of binning technique on rankings of the highest gradient magnitude between regions. Participants responded to this question by entering rankings of 1-3 (3 indicating the highest gradient magnitude region and 1 the least) for regions A, B, and C. These regions were selected because they contained the 3rd, 4th, and 5th steepest points (see Fig. 5.14). The ordinal logistic regression equation did significantly predict rankings when using binning technique and regions as predictors $\chi^2(df = 9) = 137.79, p < .00$, but binning technique did not affect gradient magnitude rankings. Both the regions and the order of rankings were significant predictors of rankings.

This task and the prior tasks relating to steepness judgments suggest that multiple different factors likely influenced a reduced effect of binning, such as prior assumptions

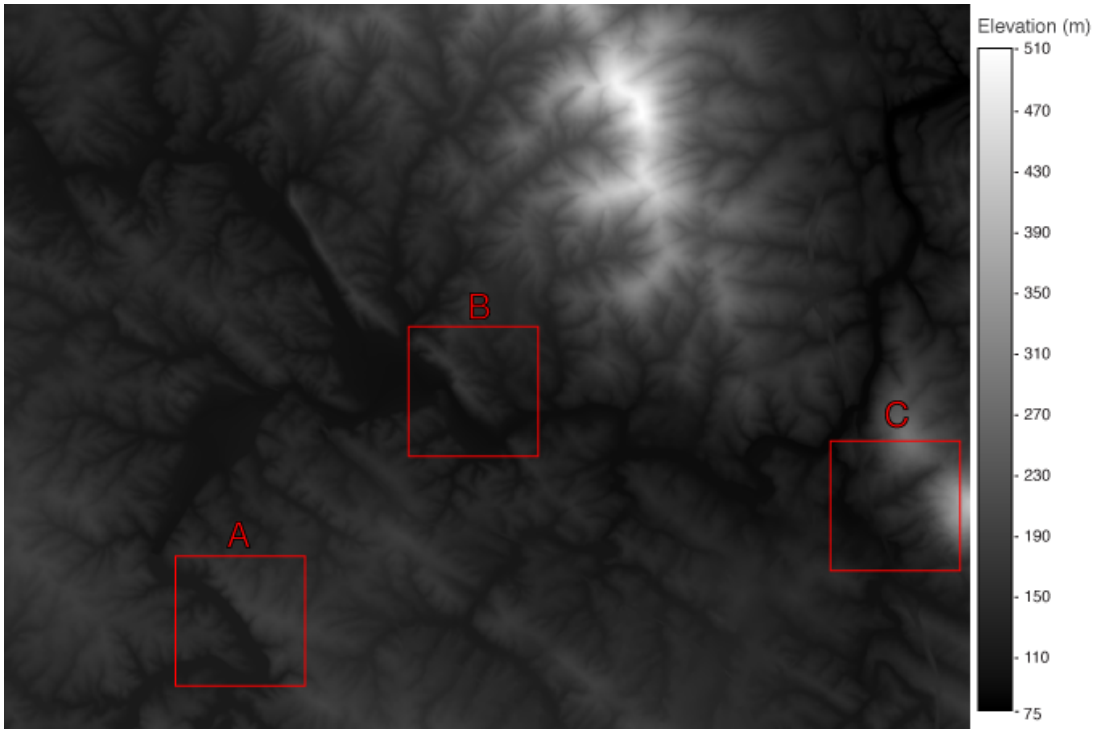


Fig. 5.14. Example Continuous condition stimulus for the Rank Steepness task.

about how elevation peaks and slopes relate and variable gradient magnitudes.

5.3.3.6 Browse and Summarize Task

For the **Average Height** task, a one-way between-subjects ANOVA testing the effect of binning technique on reported average height in a region found no effect of binning on response accuracy $F(4, 474) = 1.241, p = .293$, after removing outliers greater than 2 SD above the mean error score (3.4% of trials). This finding may be due to the large magnitude of errors in all of the binning conditions $(M = 101.105, SD = 47.44)$.

5.3.3.7 Browse and Compare Task

Notably, for the **Compare Average Height** task, *participants were least accurate using the 40m binning technique*. This result was revealed by a logistic mixed model, testing the effect of binning technique on a comparison of steepness. Comparison accuracy was calculated by coding correct responses as 1 and incorrect rankings as 0. The analysis revealed that the 40m binning displayed a significantly lower proportion of accurate responses (63%) compared to the 10m (82%), 20m (79%), 30m (87%) binnings and the continuous encoding (86%). The model accounted for a significant proportion of variance compared to a baseline model with no predictors, $\chi^2(df = 7) = 22.80, p = .001$ (see Fig. 5.15).

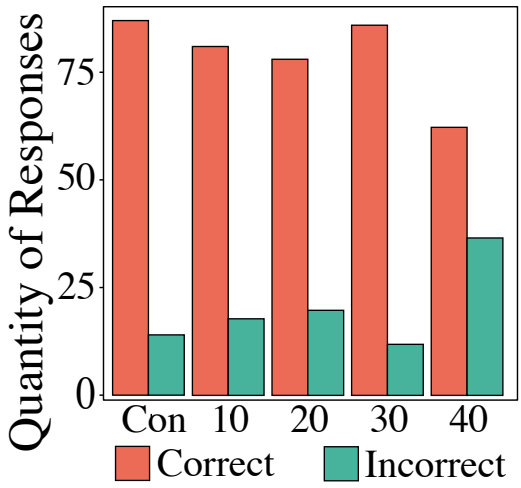


Fig. 5.15. Proportion of correct and incorrect responses in the Compare Average Height task for each encoding condition.

5.3.3.8 Combination Task

For the **Steepest at 355** task, preliminary diagnostic checks of the data revealed that the majority of participants were clicking at elevations much lower than 355 meters as instructed (see Fig. 5.16). We defined regions based on relative proximity in elevation to 355 meters, making three groups (below 255 meters, 255–455 meters, above 455 meters) and qualitatively assessed the frequency of clicks in each region by binning condition. Many participants did not correctly identify the 355 elevation, as the greatest number of clicks fell in the "low" category. Overall, the wide distribution of click responses makes it difficult to assess differences between the binning conditions and suggests that the task was difficult for users.

5.3.3.9 Task Completion Time

Eight of the 12 tasks showed significant differences in task completion time between binning techniques, with the continuous encoding displaying consistently faster performance than some of the binnings. These results were revealed by one-way between-subjects ANOVAs computed for each task, examining the effect of binning condition on task completion time. Table 5.1 includes mean, standard deviation statistics, and Tukey HSD post hoc tests of significant differences accounting for multiple comparisons for each binning technique in all 12 tasks. Note that, for the first task, binning technique signifi-

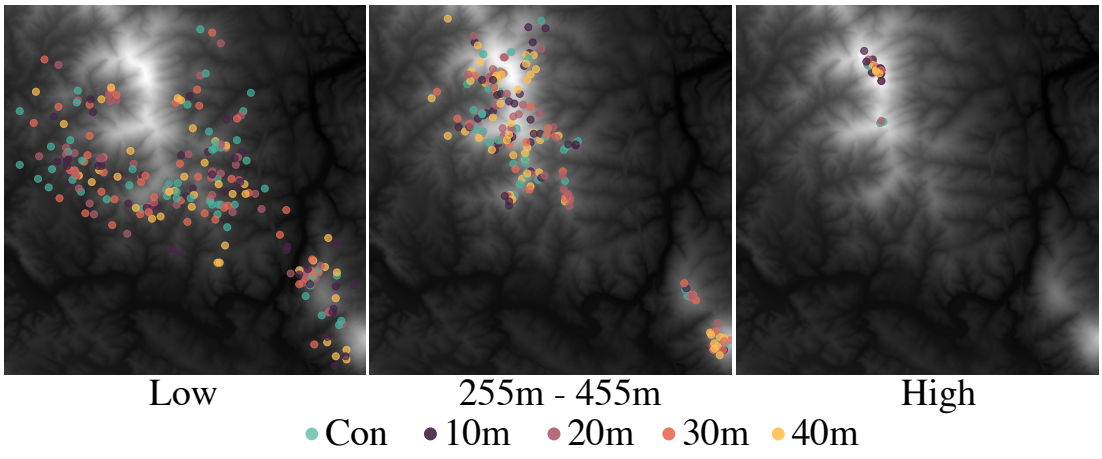


Fig. 5.16. Distribution of click locations for low, middle, and high elevation groups, for the Steepest at 355 task.

Table 5.1. Task completion time, with ** = $p < .000$, * = $p < .05$. Outliers refers to the percent of outliers two standard deviations beyond the mean that were removed.

Task	Continuous		Binned Encodings		F	p	outliers	figure
	μ	σ	μ	σ				
1	**10.90***	5.75	13.95*	10.54	7.79	**.005****	1.4	**Fig. 5.17a**

Task	Continuous		10m		20m		30m		40m		F	p	outliers	figure
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ				
2	**11.99***	6.83	**13.35***	7.14	**11.81***	6.71	**15.74***	10.20	**16.87***	9.0	7.679	**.000****	2	**Fig. 5.17b**
3	**19.10***	10.96	20.84	10.76	**19.95***	11.30	**20.16***	9.48	**25.00***	12.50	7.679	**.000****	4.4	**Fig. 5.17c**
4	**16.56***	9.88	**17.14***	10.00	**16.54***	10.25	**21.63***	11.17	20.49	12.24	4.88	**.000****	2.4	**Fig. 5.17d**
5	**29.24***	15.85	30.43	17.34	**29.71***	22.10	**37.94***	21.16	36.75	22.10	4.652	**.001****	3.2	**Fig. 5.17e**
6/7	**44.83***	23.32	53.42	32.57	48.37	28.41	**56.41***	31.31	53.87	29.72	2.522	**.04***	2.4	**Fig. 5.17f**
8	20.35	15.25	19.16	11.78	20.17	12.43	23.17	16.00	22.82	15.52	1.462	.213	2.6	
9	31.71	14.81	32.01	16.86	29.56	16.02	36.32	18.64	33.86	18.72	1.5	.201	2.8	
10	**21.34***	11.81	**22.72***	12.63	24.81	16.86	**28.60***	17.04	26.38	15.33	3.629	**.006****	3.6	**Fig. 5.17g**
11	**16.66***	9.57	18.81	12.18	16.15	9.36	17.90	10.34	**20.82***	11.09	2.977	**.019***	3.4	**Fig. 5.17h**
12	23.72	15.44	27.93	17.60	25.20	15.59	26.87	16.60	30.03	19.07	2.011	.09	3.9	

cantly affected time. However, post hoc comparisons did not reveal significant differences between the conditions at the $p < .05$ level when accounting for multiple comparisons. To understand the main effect of binning, we generated contrast codes that compared the continuous encoding to the binned encodings, which is shown in the first row of the table. It should also be noted that independent task-completion times were not collected for the Steepest and Lookup tasks. Because they were presented in concert, we collected and analyzed the aggregate time data for those two tasks. Fig. 5.17 shows the mean times for each task that demonstrated significant differences between binning techniques.

5.3.3.10 Task Confidence

Four tasks showed significant differences in task confidence between binning techniques. However, no consistent relationship between binning and confidence was observed (see Table 5.2 and Fig. 5.18).

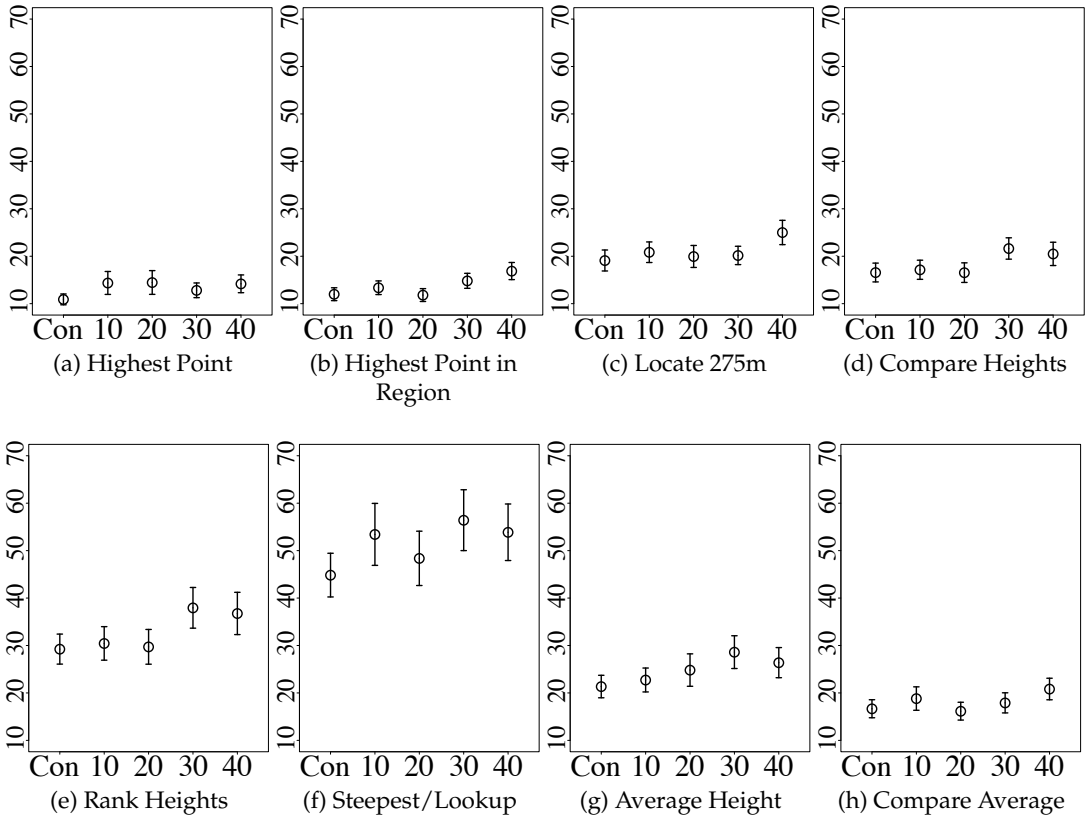


Fig. 5.17. Mean task completion time (s) for tasks with significant differences across binning conditions. Error bars represent 95% confidence intervals.

Table 5.2. Task confidence, with ** = $p < .000$, * = $p < .05$.

Task	Continuous		10m		20m		30m		40m		F	p	figure
	μ	σ	μ	σ	μ	σ	μ	σ	μ	σ			
1	5.68	1.35	5.95	1.28	5.76	1.31	5.87	1.21	6.12	1.36	1.66	.15	
2	5.01	1.34	4.87	1.45	4.92	1.49	5.17	1.27	4.90	1.52	0.73	.57	
3	4.12	1.55	**3.78***	1.42	4.06	1.49	**4.62***	1.32	**4.59***	1.63	5.91	**.000****	Fig. 5.18a
4	**4.99***	1.45	**4.89***	1.43	**4.85***	1.49	4.57	1.67	**3.98***	1.63	7.05	**.000****	Fig. 5.18b
5	4.52	1.49	4.42	1.47	4.25	1.58	4.50	1.47	4.04	1.68	1.72	.14	
6	4.36	1.73	4.49	1.74	4.67	1.52	5.01	1.48	4.55	1.71	2.22	.06	
7	3.89	1.57	3.93	1.49	4.00	1.47	4.11	1.56	4.02	1.72	0.29	.88	
8	**4.85***	1.47	5.26	1.58	**5.39***	1.57	**5.53***	1.41	5.33	1.67	2.77	**.02***	Fig. 5.18c
9	4.61	1.46	4.53	1.52	4.66	1.53	4.86	1.34	4.97	1.58	1.5	.20	
10	3.83	1.37	3.59	1.27	3.61	1.43	3.98	1.28	3.93	1.44	1.75	.13	
11	4.66	1.54	4.78	1.26	4.67	1.49	**5.17***	1.39	**4.40***	1.63	3.59	**.006****	Fig. 5.18d
12	3.85	1.47	3.81	1.68	3.67	1.56	4.11	1.57	4.23	1.71	2.00	.09	

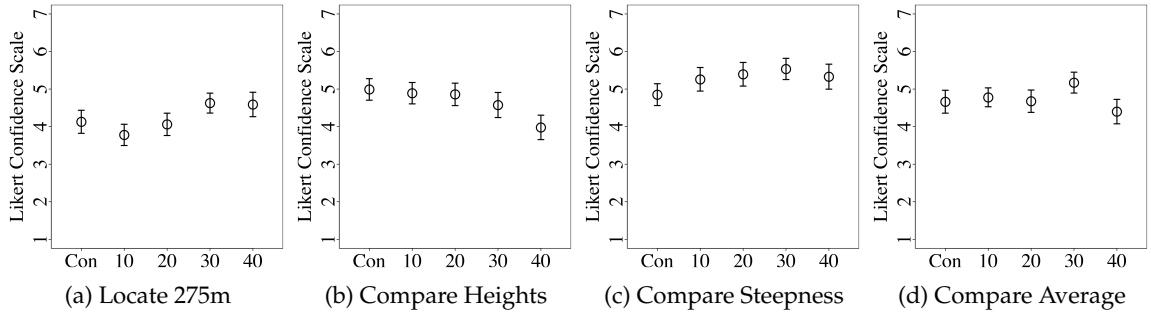


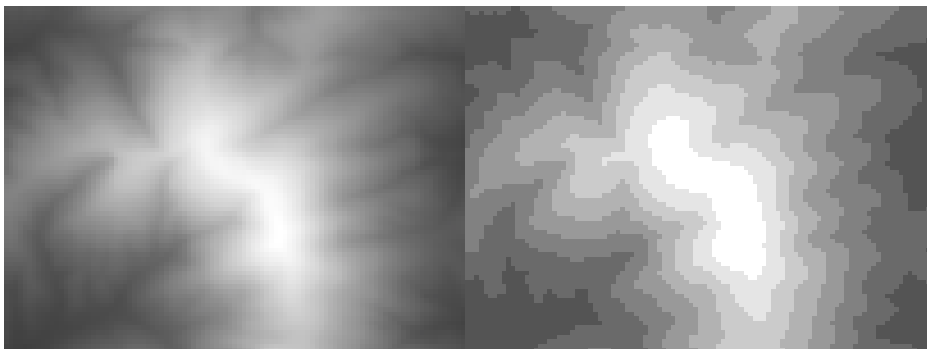
Fig. 5.18. Mean task confidence for tasks with significant differences across binning conditions. Error bars represent 95% confidence intervals.

Further, overall confidence ratings were relatively consistent between tasks. These results were revealed by one-way between-subjects ANOVAs computed for each task, examining the effect of binning condition on task confidence ranking. These results suggest that confidence ratings were not a very sensitive measure of differences in encoding conditions and that future work might use other subjective ratings such as preference or ease of task.

5.3.3.11 Results Summary

Significant differences in accuracy of decision-making, response speed, and response confidence were found between continuous and binned encodings. The results were task-specific, but overall, our findings suggest that the continuous encoding increased speed but rarely accuracy in many tasks. The most substantial differences among binning conditions were found in the Highest Point and Lookup tasks, suggesting that the continuous encoding technique was the least effective in those situations. Our findings suggest that the choice of binning versus continuous encoding of 2D scalar fields is important to consider as an influence on user task performance.

More specifically, we found that for the Locate or Explore and Identify Tasks (Section 5.3.3.3), the binning technique influenced task performance. For example, those with the continuous encoding had worse performance than those with the 40m binning in the Highest Point task, which may have been due to a pop-out effect that the 40m binning produced (see **Fig. 5.19**). The other Locate or Explore and Identify tasks varied based on



(a) Continuous (b) 40m

Fig. 5.19. The highest portion of the DEM shown using the Continuous versus 40m encodings, illustrating the pop-out effect created by the latter.

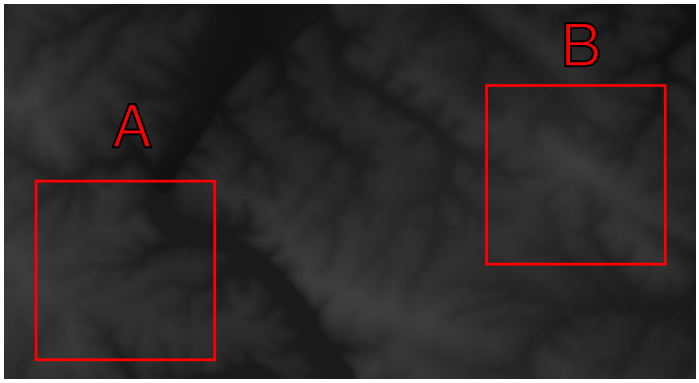
binning interval.

Responses for all tasks that utilized a steepness judgment (e.g., Explore and Identify task (Section 5.3.3.3), Explore and Compare tasks (Section 5.3.3.5) and the Combination task (Section 5.3.3.8)) appear to be highly influenced by prior assumptions about the relationship between elevation peaks and slopes and variable gradient magnitudes as described in Section 5.3.3.3.

In the Lookup and Identify task (Section 5.3.3.4), those with the continuous encoding also performed the worst. This effect may have been due to the difficulty of locating values on the continuous legend. Work by Brewer [19],[20],[66] suggests that identifying relevant values on a choropleth map legend with many bins (over 12) is difficult. Indiscernibility of specific values on the continuous legend may have contributed to the inefficacy of this encoding.

For the Browse tasks (Sections 5.3.3.6 and 5.3.3.7), task difficulty likely led to large errors when participants were asked to report the average elevation of a region. However, participants were reasonably effective at comparing the average height of two regions, and those with the 40m binning were the least effective at this comparison. This result may have been due to the 40m binning lacking sufficient detail for an accurate average height comparison (see **Fig. 5.20** for a comparison).

The response time data were the most consistent across tasks. Overall, these results show that the continuous encoding was significantly faster than some of the binning tech-



(a) Continuous

(b) 40m

Fig. 5.20. Portions of the continuous versus 40m stimuli for the Compare Average Height task, illustrating the loss of detail in the 40m encoding.

niques for many tasks. This finding is not consistent with work that suggests that binning data facilitates task completion speed [42], [43], [66], [199].

Finally, although significant differences in task confidence were observed, confidence ratings did not consistently vary across binning conditions.

5.4 Discussion

Although the expressiveness principle implies that encoding continuous data continuously is advantageous, some domains have strong conventions that advocate for a discretized encoding of continuous data. The existing literature lacks systematic evaluation of the effects of binning 2D scalar fields on human performance. Our goal in this work was to directly compare performance using several different binning intervals and a continuous encoding across a range of tasks for 2D scalar fields. Consistent with our

initial hypothesis, we found that accuracy in performance was superior using binning in some tasks. Although, in other tasks, we did not show differences between binning and continuous representations, no tasks showed a clear advantage for the continuous encoding. This result stands in contrast to one of the dominant principles in visualization and is a finding that should be pursued further in both controlled laboratory studies and real-world contexts.

Our findings on task completion time are also notable. In contrast to assertions in cartography, the current study reveals that a binned encoding does not facilitate faster task completion times, compared to a continuous encoding. Instead, we found that performance on the continuous encoding condition was the fastest across a majority of the tasks, even when accuracy was worse. One possible explanation for this speed-accuracy trade-off is that tasks that require more effort can activate a slower and more deliberative cognitive process [73]. Judgments made using fast cognitive processes are more error-prone and subject to cognitive biases than judgments made using slower, more contemplative cognitive processes [82]. The continuous encoding may have produced quick responses that did not allow for careful inspection of the visualization or thoughtful consideration of the task.

Additionally, one of the aims of this study was to create a methodology for testing increasingly complex and ecologically valid tasks, which are traditionally difficult to assess and, therefore, rarely utilized. Inspired both by our prior observations of meteorological workflows [145] and our prior experiences trying to increase ecological validity with respect to meteorological tasks [147], we chose to increase task complexity by having participants make judgments concerning features and summary statistics in addition to points and values, by increasing the number of comparisons, and by requiring multistep judgments. Multistep visual comparison tasks are highly dependent on spatial working memory (SWM), and differences in SWM may account for the inconsistent performance observed. In particular, we suspect that ranking tasks may not be appropriate for spatial decision-making as limits on participants' SWM may confound this type of task. Further work is needed to disambiguate the demands on SWM from the increased task complexity. Another issue is found in the multistep judgment tasks (i.e., the Steepest at 355m task), where it is unclear how to disambiguate participants' strategies from the influence of

the binning technique. The inconsistent findings of the Steepest at 355m task may have resulted from participants prioritizing clicking on the correct elevation or on a steepest location rather than a combination of the two. Each strategy would produce vastly different responses, which may have influenced the uninterpretable variability in our findings.

As discussed briefly in Sections 5.3.3.3 and 5.3.3.5, interpretations of the questions that involved steepness were difficult. We believe that this is, in part, because the image gradient we used as the basis for steepness had high-spatial-frequency information. Gradient approximations are generally known to be sensitive to high-frequency information such as noise [49], [56]. There are a variety of gradient estimation methods we could have pursued, but we chose to use the Sobel operator, which includes a smoothing component [49], [56]. Smoothing obfuscates more of the data, and the influence of smoothing on decision-making requires further study. Given that our results also left open the possibility of other factors influencing the results, such as the conflation of steepness with height, it was not clear what amount of smoothing would have been appropriate. Although our findings are specific to 2D scalar fields embedded in $\mathbb{R}^2$, we suspect that our results are also relevant for 2D scalar fields embedded in $\mathbb{R}^3$, such as those found in bioengineering [108]. This study motivates future work aimed at assessing the influence of individual components of spatial data, such as spatial frequency, and determining appropriate smoothing for specific conditions.

As we were interested in performance across a spectrum of tasks, we did not attempt to compare an optimized binning technique. There are a variety of data-informed optimization strategies for binning in the visualization literature [48], [164], [183]. Additionally, standardized binning approaches for choropleth maps from cartography [168] might also be generalizable. However, it is not clear how these various optimizations interact with tasks. Given that existing research suggests that binning strategies that are highly effective for a particular task are likely inadequate for a larger set of tasks [117], optimizing binning for a variety of tasks remains an open problem. We believe a particularly fruitful area of inquiry would be investigating the effects of interactive interfaces for helping a user customize bins according to evolving task demands. In such a system, to what extent does the *act* of binning, itself, become a vehicle for exploration of the data? What sorts of default settings are desirable? Many potential avenues for future work exist in this space.

The primary implications of this work are for topographical analysis, geography, cartography, and meteorology, suggesting that designers should consider selectively employing the expressiveness principle to promote accuracy or speed. If accuracy is the main objective of the visualization, then binning may be more effective. If speed is of interest, then following the expressiveness principle and using a continuous encoding may be desirable. In many cases, designers will want to strike a balance between speed and accuracy, requiring a fine tuning of the binning intervals based on task demands and data. Further, the tasks developed in this study should be applicable to any domain using 2D scalar fields. In future work, designers should also consider the influence of both legend labels [110] and even coarser bins (e.g., three or five) that are commonly used with choropleth maps [66]. Previously studied labeling effects may interact with very coarse binnings, adding to cognitive load. The possible influences of varying spatial frequencies and topological characteristics in the underlying data also warrant future study.

A secondary implication for this work involves the use of rainbow color maps. Rainbow color maps have been marked as a poor choice for encoding data in part because they are not perceptually continuous, resulting in perceptual bands that hide details [14], [119],[157] (cf., [20]). Even though large portions of the visualization community reject this technique, as we noted in Chapter 3, rainbow color maps are still widely used by experts, even as defaults. This study's results call into question the extent to which the natural discretization produced by rainbow color maps should truly be considered a deficiency. We also wondered if the continued appeal of rainbow color maps might, in part, derive from a general preference for discretization. As we detail in Chapter 6, these questions became the basis for the final piece of work in our emergent mixed-methods research program.

5.5 Summary

This chapter describes a study in which we explicitly tested the efficacy of binned versus continuous color maps for visualizing 2D scalar fields, in order to better understand the impact of the meteorological convention of using binned color maps to encode continuous data. Contrary to the expressiveness principle, no cases were found in which a continuous encoding of 2D scalar field data was advantageous for task accuracy, and for

some tasks, specific binned encodings facilitated accuracy. Furthermore, we found support for the counterintuitive finding that decisions with binned encoding were slower than those made with continuous encoding. Given this apparent trade-off between accuracy and speed of performance, it is important to consider specific task-goals when deciding whether to adhere to the expressiveness principle. We also developed and tested several different tasks that ranged in complexity and relevance to real-world problems, and we offered insights on specific types of tasks and data characteristics to avoid.

Our approach and findings contribute to a greater understanding of the use of binning for 2D scalar data and lead to many possibilities for future research. Further, as this study empirically tests a guidance conflict observed during our formative design study [145], it provides initial validation of the mixed-methods research approach selected for this dissertation. At the same time, this study also provided us with new insights into an even larger conflict between visualization guidance and real-world practice: the continued, widespread use of rainbow color maps.

CHAPTER 6

EXAMINING IMPLICIT DISCRETIZATION IN SPECTRAL SCHEMES

This chapter discusses the final explanatory project in our mixed-methods research program: an exploratory study we conducted to understand whether rainbow color maps implicitly discretize data, and if so, how and how consistently across different individuals, datasets, and spectral schemes. During our initial meteorological design study [145], one of the strongest conventions we encountered was the use of rainbow color maps. Indeed, despite longstanding efforts by the visualization community to enumerate the many problems with rainbow color maps [8], [13], [14], [119], [126], [158], [196], the use of these color maps remains prevalent in many scientific domains [13], [25], [38], [89], [145], [214]. As discussed in Chapter 5, the results of our binning study provided us with a new avenue for exploring this apparent conflict: the role of discretization.

Two of the primary reasons rainbow color maps are considered harmful stem from the idea that they implicitly discretize the encoded data into *hue-based bands* [8], [14], [119]. Our binning study results, on the other hand, raise the question of whether the implicit discretization in rainbow color maps might afford similar benefits to those we found for explicit discretization (i.e., binning) [137]. In attempting to examine this question, however, we discovered a lack of foundational research establishing either that people perceive rainbow color maps as implicitly discretized, or how that discretization varies across individuals, datasets, and rainbow color map variants. This chapter outlines the exploratory study we conducted to establish those foundations. We outline both our experimental design and results, and we discuss those results in the broader context of our understanding of color usage in visualizations of 2D scalar fields.

6.1 Study Motivation

As mentioned in the introduction, two of the primary reasons rainbow color maps are considered harmful stem from an argument that they implicitly discretize encoded data into hue-based bands [8],[14],[119]. The literature argues that this perceived banding both highlights nonexistent relationships in the data through the creation of false boundaries and masks real relationships within a given band [14]. Our current understanding of the implicit discretization in rainbow color maps, however, is based on a combination of generalized knowledge about how humans perceive the visible spectrum [159] and anecdotal evidence that has yet to be empirically tested [14].

To our knowledge, no work has empirically evaluated the perceived banding in rainbow color maps or characterized the potential differences across color scales, datasets, or individuals. Understanding and characterizing any perceived banding in rainbow color maps is important, as precise knowledge of how people perceive these bands is critical for leveraging implicit discretization in color-map design to improve the performance of some tasks [37],[137]. An improved understanding of implicit discretization also ensures that visualization guidance regarding rainbow color maps has proper scientific foundations [91].

To better understand the perceived banding in rainbow color maps in color displays, we conducted an exploratory study aimed at assessing both whether rainbow color maps implicitly discretize data and how implicit discretization varies across different individuals, datasets, and spectral schemes. Participants were shown sets of color-mapped visualizations and asked first to count the color categories/boundaries that they perceived and then to delineate those categories/boundaries interactively. We then compared participants' delineations against potential boundary locations derived from variation in the perceptual dimensions of color for each spectral scheme. Although previous work has attributed some perceived bands in rainbow color maps to variation in luminance [8],[14],[119], luminance alone cannot explain the banding perceived in all spectral schemes [86]. This study expands the investigation of banding effects to variations across all three perceptual dimensions of color: lightness, chroma, and hue.

As with the binning project discussed in Chapter 5, the results of our study are the primary contribution of the work presented in this chapter. The study results suggest that rainbow color maps *are* implicitly discretized with consistency across individuals. Addi-

tionally, the results show correspondences between participants' responses and variation in each perceptual dimension of color. The results also indicate that the discretization produced by a given color map varies in unexpected and unpredictable ways across different datasets, revealing practical challenges for common tasks like drawing comparisons across datasets. Further, the findings suggest that the visualization community's current understanding of both rainbow color maps, and more generally effective color usage, remain incomplete.

The remainder of this chapter is outlined as follows: Section 6.2 briefly summarizes related work conducted in the visualization, vision science, and cognitive science communities; Section 6.3 discusses the wide range of definitions for the term *rainbow color map*; Section 6.4 details both the study's aims and methods; and Section 6.5 then outlines the results of the study, which we discuss further in Section 6.6.

6.2 Foundations

This work in this chapter builds on two distinct bodies of literature: the visualization community's prior work regarding rainbow color maps and work regarding the categorical perception of color, conducted predominantly by the vision and cognitive science communities. In this section, we highlight closely related work from both.

6.2.1 Rainbow Color Maps

Rainbow color maps are claimed to be harmful for three primary reasons. First, because hue is not inherently ordered [196], ordered relationships in data visualized using rainbow color maps are not necessarily preserved [14], [119]. The other two reasons are that rainbow color maps both mask and overaccentuate small data differences, biasing our understanding of the underlying data relationships [14], [83]. Insufficient luminance variation in certain portions of rainbow color maps can obscure small details, and hue bands can introduce artifacts such as false boundaries that actively mislead users. Current thought further links both problems to the irregular nature of the implicitly perceived discretization.

Several studies provide empirical evidence for parts of these claims. Color matching experiments by Kalvin et al. [83] and feature discrimination experiments by Ware et

al. [194], [198] support the idea that insufficient luminance variation hides data variation in the green region of the traditional rainbow color map. Rogowitz and Kalvin show that, even though local subsections of the traditional rainbow color map maintain a luminance-based ordering, the overall color map is not inherently ordered [158]. Recent work by Liu and Heer indicates that, when asked to compare the relative distances of colors within a color map, respondents are slower and more error-prone with the *jet* color map compared to both single-hue and nonrainbow multihue schemes. Further, a study that Borkin et al. conducted with medical experts shows that the traditional rainbow color map is ineffective for real-world tasks [13]. This result is echoed in a recent study Dasgupta et al. ran with climate scientists, where the jet color map produced larger errors in average magnitude comparisons of geospatial maps compared to color maps with monotonic luminance [37].

Other work suggests that rainbow color maps are not always a bad choice. Experiments by both Ware [195] and Reda et al. [150] show that rainbow color maps are accurate for quantity estimation tasks and provide support for some form-comprehension and gradient-estimation tasks. Additionally, studies by Brewer [18] and Gresh [60] show that modified rainbow color maps are interpreted accurately when used as multihue diverging schemes. Rainbow color map variants that control luminance variation to avoid many of the problems commonly attributed to rainbow color maps also exist [58], [86]. Collectively, this work leaves a variety of open questions about if, when, and why rainbow color maps are harmful.

Several papers argue that rainbow color maps are harmful because they implicitly discretize the encoded data into hue-based bands [8], [14], [119]. These bands introduce false boundaries and obscure data variation, thereby leading users "to infer structure which is not present in the data and to miss details that lie completely within a single color region" [8, p. 119]. These arguments, however, are predicated on empirical evidence that visible light is perceptually discretized when diffracted through a prism [159] and on anecdotal examples [14]; no work has empirically tested whether rainbow color maps are perceived as banded. Moreover, recent work raises questions about whether and why implicit discretization is problematic. Our prior work shows that, in grayscale color maps, regularly spaced discretization does not negatively impact and sometimes improves accuracy across various tasks [137], suggesting that discretization can be benefi-

cial. Further, Dasgupta et al. show that although hue banding negatively impacts average magnitude comparisons, it enables more accurate difference comparisons across geospatial datasets [37].

No prior work, however, has empirically investigated the implicit discretization in visualizations using rainbow color maps. We address this gap by exploring the questions of whether rainbow color maps implicitly discretize encoded data into perceived bands and how that discretization varies across individuals, data characteristics, and spectral schemes. Our broad goal is to understand whether any perceived banding in rainbow color maps can be predicted using the perceptual dimensions of color: lightness, chroma, and hue.

6.2.2 Categorical Perception of Color

The idea that humans perceive continuous color as discretized traces back centuries to experiments by Issac Newton and Hermann von Helmholtz [173]. The phenomenon is currently believed to be an effect of *categorical perception*, where viewers are faster and more accurate at discriminating colors in different categories (e.g., green and blue) compared to colors in the same category (e.g., different shades of blue) [65]. Categorical perception represents one possible explanation for why and how people might implicitly discretize rainbow color maps.

An extensive body of literature has investigated categorical perception both in general [55] and specifically as it applies to color [22],[151], but little work directly investigates the effects of categorical perception across continuous color ranges. Most existing research focuses on probing individuals' perception at a small set of established color-category boundaries in order to test theories about the perceptual or cognitive underpinnings of color categorical perception, often using pair-wise color judgments [72], [136], [200]. This focus likely stems from an ongoing debate about the roles of language and perception in forming color categories. One side argues that a perceptual phenomenon gives rise to the formation of consistent categories across languages; the other argues that prior knowledge of language biases individuals' perception. The literature has yet to reach a consensus regarding what drives categorical perception [151], [206].

Two studies that examine how people perceive color categories across the visible spec-

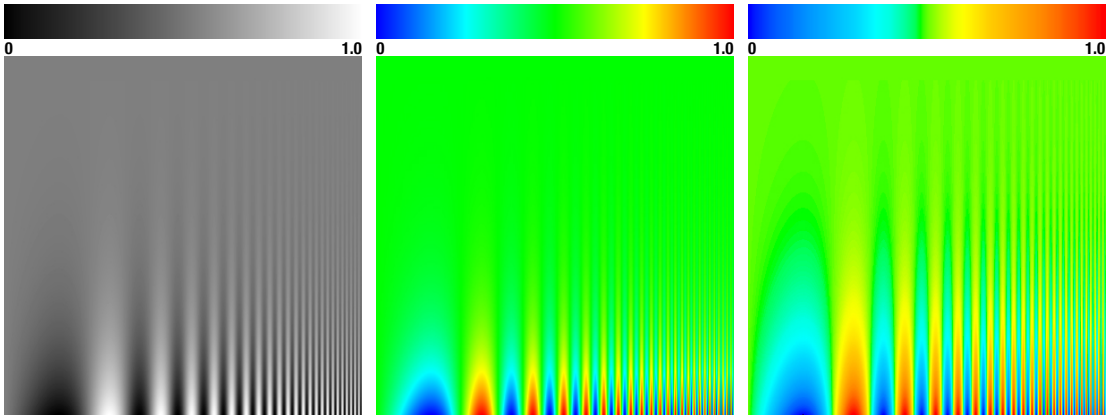
trum or approximations thereof present some evidence of consistency in subjects' perception of color category boundaries [173], [207]. Small sample sizes and specific experimental design choices, however, create questions about the generalizability of these results. Smeulders et al. asked 5 participants to delineate a diffracted spectrum into a specified number of categories [173], and Witzel and Gegenfurtner asked 10 participants to name isolated color samples derived from isoluminant hue circles presented on computer monitors using basic color terms [207]. It is not clear that either set of results reflects the potential banding expected in rainbow color-mapped visualizations. Moreover, neither study addresses the question of whether categorical perception creates an implicit discretization that might affect how a user completes a given task. A handful of visualization papers leverage categorical perception either to create more effective visualizations [29], [67], [68] or to model participant responses [100], but no work directly addresses the potential role of categorical perception in discretizing continuous color scales.

6.3 Defining the Rainbow

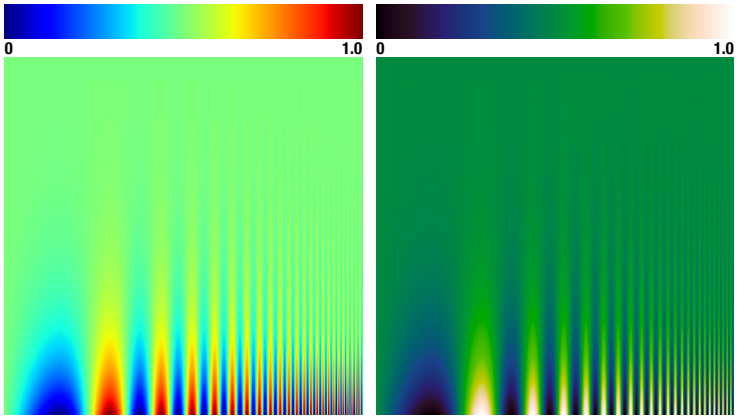
Within the visualization community, a variety of color maps that approximate the visible spectrum are described broadly using the term *rainbow color map*. Specific research results, however, often relate to particular spectral schemes [13], [37], [100], [158], [195], raising questions about the generalizability of those results to the larger class of rainbow color maps. This ambiguity is problematic because not all rainbow color maps suffer from the same problems to the same extent.

Fig. 6.1 shows Campbell-Robson contrast sensitivity charts encoded using a variety of rainbow color maps.[4] Contrast sensitivity charts are commonly used to illustrate that rainbow color maps hide data variation [14], [89], [119], although prior comparisons are limited to grayscale (**Fig. 6.1a**) and the *traditional* rainbow color map (**Fig. 6.1b**). The traditional rainbow color map is commonly defined by tracing the boundary of the device-dependent RGB gamut from blue to red [83], although a variant that cycles from magenta to red also exists [195]. **Fig. 6.1** also includes comparisons to the *jet* color map from MATLAB [47] (**Fig. 6.1d**); Gresh's perceptually linearized rainbow [60] (**Fig. 6.1c**); and the *Kindlmann*

[4]Due to the color distortions introduced by printing processes, **the color-mapped images in this chapter should be viewed on a computer monitor**, optimally using the ICC-profile color-managed copies in the supplemental materials archive for this dissertation.



(a) grayscale (b) traditional (c) Gresh



(d) jet (e) Kindlmann

Fig. 6.1. Campbell-Robson contrast sensitivity charts visualized using (a) grayscale, (b) the traditional rainbow color map, (c) Gresh's perceptually linearized rainbow, (d) the jet color map, and (e) the Kindlmann color map show pronounced differences in the extents to which rainbow color maps capture data variation. In each image, spatial frequency increases left to right, and contrast decreases bottom to top.

color map [86], [120] (**Fig. 6.1e**), which modifies the traditional rainbow color map to linearly increase in perceived luminance.

Fig. 6.1 highlights pronounced differences in the extents to which these color maps capture data variation. Similar differences can also be observed across prior work. Both Kalvin et al. [83] and Ware et al. [194], [198] found evidence of low discriminability in the isoluminant, green region of the traditional rainbow color map, whereas Liu and Heer [100] found evidence of high discriminability in the corresponding isoluminant, green region of the jet color map. These differences illustrate a need for increased precision in discussions regarding rainbow color map research. Within the context of this chapter, the term *rainbow*

color map refers to any of the general class of spectral schemes shown in Fig. 6.2. Individual color maps, such as those referenced in our study, are referred to using specific names: the traditional rainbow color map, the jet color map, etc.

6.4 Methods and Aims

The primary objective of this research is to understand the nature of implicit discretization in rainbow color maps. If the perceived banding is linked to categorical perception, how much variation should we expect across individuals, and how does that variation change across different rainbow color maps? Moreover, given the known effects of spatial frequency on our perception of color-mapped data [83], [150], to what extent should we expect any perceived banding to be affected by the encoded data? To better understand these relationships, we conducted an exploratory study designed to generate empirical observations about how individuals perceive hue bands across different datasets visualized using various rainbow color maps. We focus solely on implicit discretization in electronic displays while controlling for expected real-world confounds such as gamut differences and viewing conditions. The following subsections detail our hypotheses, stimuli, experimental apparatus, tasks, procedures, and participant demographics.

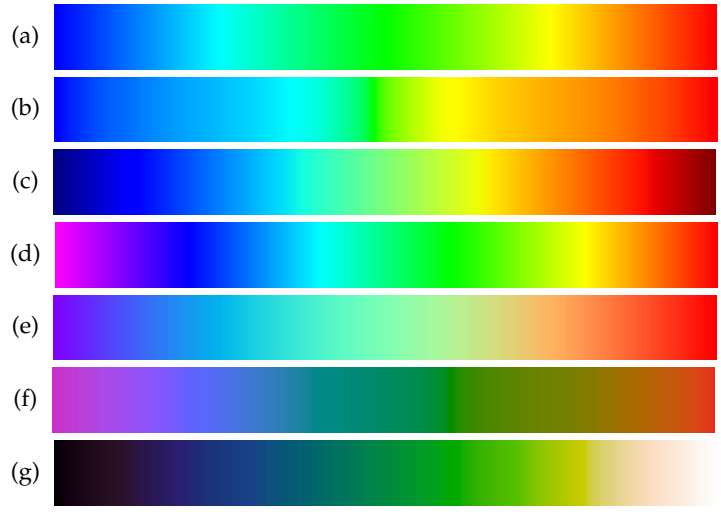


Fig. 6.2. Examples of spectral schemes that are commonly referred to as **rainbow color maps**: (a) the traditional rainbow color map (truncated at blue), (b) Gresh's perceptually linearized rainbow, (c) the jet color map popularized by MATLAB, (d) the traditional rainbow color map (cycling to magenta), (e) the rainbow color map specified by matplotlib, (f) Kindlmann's isoluminant rainbow, and (g) the Kindlmann color map.

6.4.1 Hypotheses

In this study, we explored three main hypotheses.

H1 *In line with the long-standing suppositions of the visualization community [8], [14], [119], we expect that data visualized using rainbow color maps is perceived as implicitly discretized into hue-based bands. Additionally, we predict that this implicit discretization will vary across different rainbow color maps.*

The categorical perception literature provides evidence that humans perceive specific sets of colors as grouped or categorized according to various color terms. When white light is diffracted into a prismatic spectrum, people consistently perceive continuous ranges of wavelengths as individual bands of uniform color [173]. Similar categories also exist in how people perceive and group individual samples of uniform color [72], [200], [207]. Therefore, we have reason to expect that this same phenomenon might affect visualizations encoded using continuous color maps.

The literature also suggests that the categorical perception of color is inherently tied to a relatively small set of basic color terms [9], [151] that effectively partition color space. By definition, different spectral schemes trace fundamentally different paths through color space, suggesting that they also trace different paths through the regions of color space associated with these color terms. Thus, we expect that the widths of the perceived bands associated with a given basic color term should vary across different rainbow color maps. It is not immediately apparent, however, to what extent individual differences might also impact this variation.

H2 *We predict that the implicit discretization produced by a given color map is based on more than just luminance variation.*

The idea that sudden shifts in luminance should cause visible discontinuities in rainbow color maps is well documented in prior work [8], [119], but luminance alone cannot explain all the banding seen in common rainbow color maps, which can be surmised from **Fig. 6.3** and **Fig. 6.4**. **Fig. 6.3** shows the CIELCh lightness (L*), chroma (C*), and hue (h) profiles for four color maps: a perceptual grayscale, the traditional rainbow color map, jet, and the Kindlmann color map. **Fig. 6.4** visualizes three datasets using these

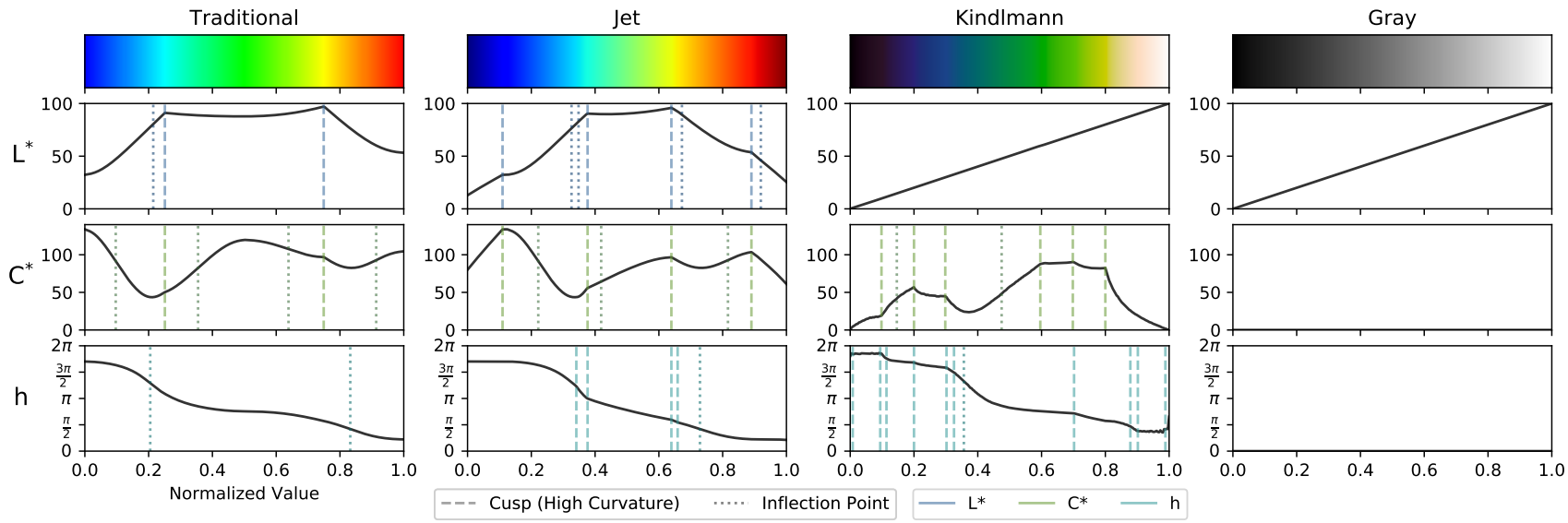


Fig. 6.3. The CIELCh lightness (L*), chroma (C*), and hue (h) profiles for the four color maps used in our exploratory study, with dotted and dashed lines showing the derived locations of cusps and inflection points.

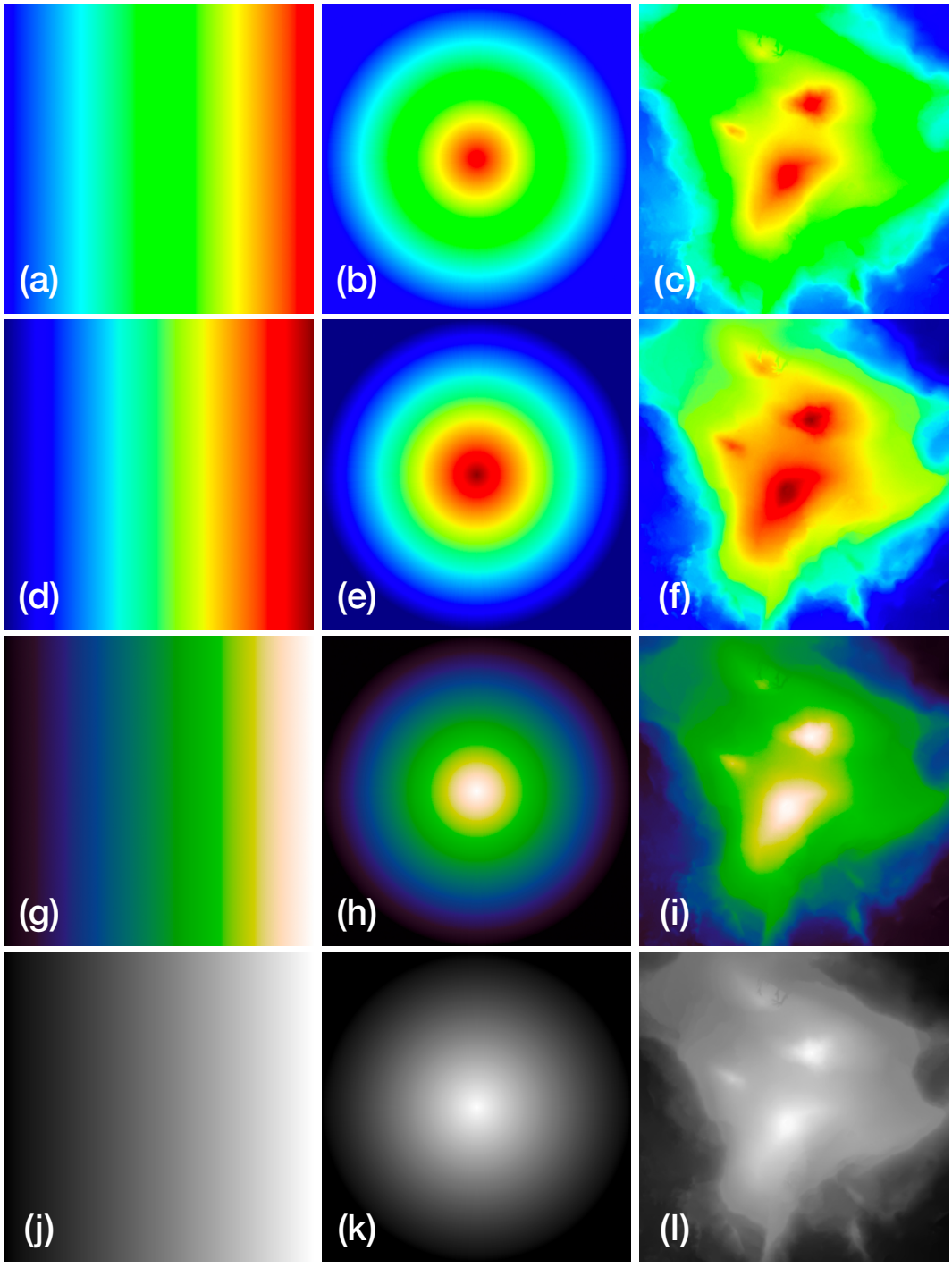


Fig. 6.4. Experimental stimuli encoding a linear ramp, a radial gradient, and a complex 2D geospatial dataset using four color maps: (a)-(c) the traditional rainbow, (d)-(f) *jet*, (g)-(i) the Kindlmann color map, and (j)-(l) perceptual grayscale.

same color maps. *Lightness* is a measure of perceived luminance, judged relative to a comparably illuminated white [51]. Being based on pair-wise color judgments, CIELCh is not a perfect metric of perceived color differences in continuous color fields; however, a better alternative does not currently exist [179].

Despite encoding linear data, **Fig. 6.4g** and **Fig. 6.4h** both appear to show banding induced by the Kindlmann color map. This banding cannot be explained by the Kindlmann color map's linear lightness profile, seen in **Fig. 6.3**. **Fig. 6.3** also indicates that the traditional rainbow color map's lightness profile has only two sudden changes (i.e., cusps). If luminance perception alone drove this phenomenon, **Fig. 6.4a** and **Fig. 6.4b** should contain only three distinct hue bands.

We suspect that variation in chroma and, to a lesser extent, hue also contributes to the banding perceived across the first two columns of **Fig. 6.4**. Chroma, like saturation, is a relative measure of colorfulness, the distinction being that chroma, similar to lightness, is measured relative to the brightness of a comparably illuminated white, whereas saturation is measured relative to the stimulus' brightness [51]. **Fig. 6.3** indicates that the lightness and chroma profiles for both the traditional rainbow and jet color maps have *cusps*, sharp features where two curves intersect.[5] The chroma profile for the Kindlmann color map also contains cusps, which could explain perceived banding independent of luminance. Further, *inflection points* or concavity changes in the chroma profiles of each color map appear to loosely correspond to additional hue-band boundaries, such as potential red:orange and blue:light-blue boundaries in both the traditional rainbow and jet color maps. Weaker evidence suggests that hue variation may also be contributing to this phenomenon. **Fig. 6.3** indicates that *some* of the cusps in chroma correspond to cusps in hue for both the jet and Kindlmann color maps. Additionally, there are two inflection points in the hue profile for the traditional rainbow color map that might also impact perceived banding. Given these observations, we hypothesize that the implicit discretization in rainbow color maps might be explained by a combination of cusps and inflection points in the perceptual dimensions of color.

---

[5]Here, we do not use the term cusp in a pure, mathematical sense. Mathematically, some of these features might be more accurately characterized as *corners* (i.e., discontinuities in the first derivative) rather than cusps, or *spinodes*, where two branches of a curve meet such that they share a common tangent.

H3 *We predict that the implicit discretization perceived in a given visualization will depend on the data being visualized.*

Any banding perceived in a given color-mapped visualization will be related to the color variation in the resulting image. When encoding linear functions, any perceived banding in the resulting image space should be the result of color-map artifacts. Thus, assuming rotational invariance, we anticipate that the 1D linear ramps and the 2D radial gradients shown in **Fig. 6.4** should have similar perceived bands. With real-world datasets, on the other hand, we expect that the set of the perceived boundaries in image space will reflect a combination of not only color-map artifacts but also underlying data features that will vary across datasets. This potential conflation of data features with perceptual artifacts is a core part of why the existing literature argues that rainbow color maps are misleading [14]. The larger goal of this hypothesis is to empirically explore the differences between linear functions and real-world datasets in the hope of garnering insights that might allow us to begin to model the perceived banding in rainbow color-mapped visualizations in subsequent work.

6.4.2 Anticipated Indicators

To explore hypotheses H2 and H3, we derived the locations of cusps and inflection points from the CIELCh lightness, chroma, and hue profiles for each of the rainbow color maps in our study. The resulting locations – which we call *indicators* – are represented as vertical dotted lines in **Fig. 6.3**. Cusps are modeled as locations of high curvature such that curvature magnitude surpasses a specified threshold, whereas inflection points reflect zero crossings in curvature.

These indicators were derived using standard numerical methods. Using the 256 colors in each color map, we constructed interpolating cubic splines that approximate the CIELCh profiles for each color map. The roots of the second and third derivatives of those splines correspond to the zeros and maxima/minima in curvature, respectively. We threshold the curvature maxima/minima using the absolute value of the second derivative (i.e., curvature magnitude) to generate a set of local maxima/minima with arbitrarily high curvature. Inflection points are derived similarly, using the locations of zero curvature and thresholding based on gradient magnitude. This process is illustrated in **Fig. 6.5**.

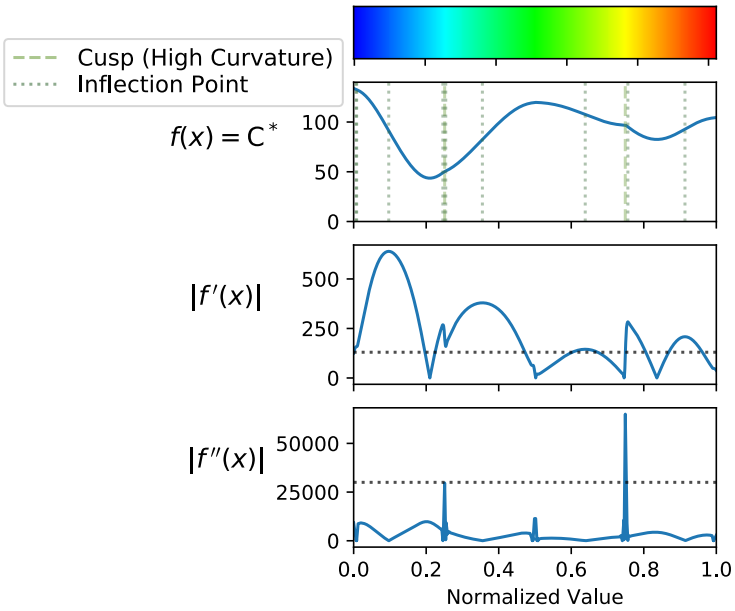


Fig. 6.5. Deriving the chroma (C*) indicators for the traditional rainbow color map: (upper) the cubic spline approximation of the chroma profile, (center) the derived gradient magnitude, and (bottom) the derived curvature magnitude. Horizontal lines show the thresholds used to isolate the cusps and inflection points, which are represented by the vertical lines overlaid on the chroma profile.

Because numerical differentiation is known to be highly sensitive to small changes [84], we employed both Gaussian smoothing on the CIELCh profiles and thresholding of the derivatives to eliminate numeric artifacts generated by noise. We also manually removed any spurious indicators that could be traced to numeric artifacts, such as boundary conditions. An expanded discussion that includes the smoothing and thresholding parameters used to derive the indicators in **Fig. 6.3** is included in the supplemental materials archive for this dissertation.

6.4.3 Stimuli

During the study, each participant was presented with 12 stimuli generated by encoding 3 univariate datasets with 4 different color maps, each shown in **Fig. 6.4**. The datasets included a 1D linear ramp, a 2D radial gradient, and a complex real-world 2D geospatial dataset. The linear ramp and radial gradient datasets are functionally defined as affine transformations of $f(x,y) = x$ and $f(x,y) = x^2 + y^2$, respectively. The complex dataset is a 3-second resolution coastal relief model of Hawaii Island sourced from the

National Oceanic and Atmospheric Administration's National Center for Environmental Information [132]. The stimuli shown in **Fig. 6.4j**, **Fig. 6.4k**, and **Fig. 6.4l** show these three datasets encoded using a perceptual grayscale color map, which was created by linearly interpolating from black to white in CIELCh. This color map was chosen as a baseline to enable separating data features from artificial boundaries created by the rainbow color maps that we anticipate according to hypothesis H3.

The remaining experimental stimuli were generated from three rainbow color maps. The traditional rainbow and jet color maps were chosen as well-known and commonly used rainbow color maps. We generated the traditional rainbow color map by linearly interpolating between equally spaced blue, cyan, green, yellow, and red control points in sRGB. For the jet color map, we utilized the implementation included in matplotlib [75].

As a spectral scheme with a linear lightness profile, the Kindlmann color map was chosen to facilitate comparisons between the other rainbow color maps and grayscale. We chose this color map over various other spiral color maps, such as the cubehelix [58] or black body [120] color maps, because it traverses a similar distribution of hue values to the traditional rainbow and jet color maps, it exhibits banding that cannot be explained by luminance variation [37], and it has an established pattern of use by the visualization community [37], [120], [178], [211], [213]. As no accepted device-independent definition of the Kindlmann color map currently exists, we reconstructed the color map directly from the original paper figure, modifying the lightness channel to ensure linearity in CIELCh. An extended discussion of this implementation choice is included in the supplemental materials archive.

6.4.4 Apparatus

The study was conducted in a controlled laboratory setting. All trials were conducted in a windowless room with the lights turned on, using two identically set-up workstations with Dell U2412M monitors. An experimenter ran contrast and gamma monitor tests [134] prior to the trials, to ensure display constancy. We did not, however, use external color measurement to verify that colors appeared the same on both monitors. At each workstation, the chair, monitor, and keyboard were placed in the same locations for all trials, with a viewing distance of 60 cm and a monitor size of 61 cm (16:10 aspect

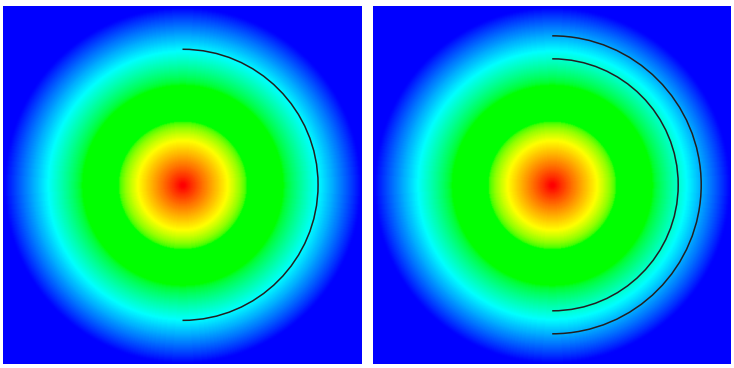
ratio). Each stimulus was centered full-screen on a medium-gray background, subtending approximately 19.3° in visual angle (768x768 pixels).

6.4.5 Tasks

For every stimulus, each participant was asked to perform two tasks according to one of two assigned instruction conditions. The first task was to count the number of color categories or color boundaries that they saw. The second was then to delineate those color categories or color boundaries interactively.

Early in the experimental design process, internal discussions revealed that salient features perceived in the yellow and cyan regions of the traditional rainbow and jet color maps could result in two fundamentally different response patterns. As illustrated in Fig. 6.6, an individual could decide to treat these features either as explicit boundaries or as prototypes within broader color categories.

In an effort to ensure that instructions did not disproportionately bias individuals' responses, we developed two separate sets of instructions to capture different ways of completing the tasks. In one set of instructions, participants were asked to make judgments related to *color categories*, which we defined as "continuous subsets of the color map where colors within the subset are more similar to one another than colors outside the subset." In the other set of instructions, participants were specifically asked about *color boundaries*, defined as "the locations where colors on the same side of the boundary are considered



(a) feature as boundary (b) feature as prototype

Fig. 6.6. Two fundamentally different response patterns illustrated for the salient cyan feature in the traditional rainbow color map using the study's boundary placement interface: (a) treating the feature as an explicit boundary versus (b) treating the feature as a prototype subsumed by a broader color category.

more similar to one another than colors on the opposite side of the boundary." Each participant was assigned a single, consistent instruction set for all trials.

In early experimental prototypes, we also noted that the inclusion of black lines as explicit delimiters appeared to influence judgments about the underlying color category boundaries. This observation is not entirely surprising given that prior work showed that black line delineations increased the number of distinct colors perceived in a diffracted spectrum [173]. It did, however, present a challenge in terms of interface design. Our goal was to understand where people perceive bands, yet the most direct interface for interrogating that question influences the perception of that phenomenon.

We opted to provide an interface where the delimiters covered only part of the underlying experimental stimuli, also illustrated in **Fig. 6.6**. By requiring participants to count the color categories or boundaries before delineation, we prime each individual's delineation responses. The interface then allows participants to attempt to line up the edge of each delimiter with the boundaries perceived in the undelineated portion of the stimuli. Delimiters can be placed or reselected by clicking, moved by dragging, or deleted with a double click. Although this design does not entirely control for the potential confounding effects of explicit delimiters, we felt it was satisfactory for an exploratory study.

6.4.6 Procedure

Each participant provided informed consent before beginning the study. The participants were then assigned one of the two instruction conditions and given a corresponding training module designed to familiarize them with the definitions and interactions in the study.

Upon completion of the training module, the participant was presented with the 12 experimental stimuli using a randomized block scheme. We used 4 blocks, each containing the 3 stimuli encoded using a given color map. Each participant encountered these 4 blocks in a different random order, and within each block, the 3 stimuli were presented in a different random permutation. This procedure resulted in a counterbalanced randomization scheme where each participant encountered exactly 1 of the 24 permutations of the 4 color-map blocks and 4 of the 6 possible dataset permutations across those blocks. For each stimulus, the participants were asked to, first, count and, subsequently, interactively

delineate the color categories or color boundaries that they saw according to their assigned instructions. Working versions of both the study and the training modules are included in the supplemental materials archive for this dissertation.

After completing the main study, the participants were asked to fill out a survey in which they answered questions about their judgments during the study and provided demographic information. Although the participants had been prescreened for color vision deficiencies, this survey included explicit secondary checks of their color vision using Ishihara plates along with questions regarding other potential confounds such as prior familiarity with the geography of Hawaii Island. Response times were unconstrained, but the study took most participants about 25 minutes to complete.

6.4.7 Participants

Participants were recruited from both the University of Utah's psychology participant pool and the University of Utah campus community. They were prescreened for either color vision deficiencies or significant prior exposure to rainbow color maps through the nature of their area of study, and they were compensated for their time at either a rate of $10/hour or via course credit.

We collected data from 62 participants across both instruction conditions, although we excluded the responses of 6 individuals who placed more than two standard deviations above the mean number of boundaries from the final analysis. The excluded participants were evenly distributed across our two instruction sets. No additional exclusions were made based on the postexperiment survey responses. Of the 56 participants included in the analysis, 42 were female and 14 were male, and the mean age was 21.55 years (SD = 5.26). The *category* instructions were assigned to 25 participants (23F, 2M), and the *boundary* instructions to 31 participants (19F, 12M). Although there is currently no clear consensus regarding either the presence or absence of sex-related differences in human color vision [80], [128], [156] and testing for such differences was outside our intended scope, we recognize that the sex imbalance among our participants is a potential limitation.

6.5 Results

Given our hypotheses that implicit discretization is occurring and is influenced by both color map and dataset characteristics, we were primarily interested in analyzing where individuals perceive and delineate banding in rainbow color-mapped visualizations. Although our open-ended boundary placement task directly examines this question, we have no way of knowing a priori which subset of participants' delimiters is supposed to correspond to a particular perceived boundary. That correspondence would require prior knowledge of the very facts we are attempting to establish: that individuals perceive bands and where they perceive the boundaries of those bands. It is not clear how one would perform quantitative analyses on predicted cusps and inflection points without prior knowledge of these facts. As a result, much of our analysis relies on qualitative visual analysis methods, which provide a structured way of exploring both participants' response trends and our hypotheses about what drives those trends, free from any assumptions about the existence or nature of hue banding. As a descriptive analysis, we also tested if the color maps influenced the number of color boundaries or categories that participants counted and placed. The statistical analyses are discussed in Section 6.5.1, and the remaining subsections provide an overview of our visual analyses as they pertain to each of our three hypotheses. We have also included a variety of interactive tools and expanded discussions as supplemental materials to assist readers in better assessing the validity of our claims.

6.5.1 Descriptive Statistical Analysis

To get an initial understanding of the relationships present in the results, we conducted statistical analyses on the number of boundaries/categories participants perceived. We used a linear mixed-effects analysis due to the mixed design with unbalanced sample sizes. Participants were modeled as a random effect, and color map (traditional, grayscale, jet, Kindlmann), dataset (1D, 2D, complex), instruction condition (category, boundary), task/response-method (counted, delineated), and potential color-map:dataset and instruction:response-method interactions were all modeled as fixed effects. Additionally, the grayscale color map and complex dataset were used as reference groups, given that we specifically hypothesized differences compared to these groups.

Although the full equations and outputs can be found in the supplemental materials archive for this dissertation, Fig. 6.7 illustrates the core relationships in this linear mixed-effects analysis. The analysis revealed main effects where each of the rainbow color maps elicited significantly more delineations than grayscale, and the 1D dataset elicited significantly more delineations than the complex dataset. It also showed significant color-map:dataset interactions, with subsequent post hoc analysis revealing that, for both the jet and Kindlmann color maps, participants perceived significantly more boundaries for the 2D dataset compared to the complex dataset but no significant difference in the number of boundaries for grayscale. Additionally, both before and after accounting for these color-map:dataset interactions, neither the instruction condition nor the task/response-method had a significant effect on the number of perceived boundaries.

These statistically significant effects support the idea that both color map and dataset influence how people perceive boundaries in a given color-mapped visualization, but proving or disproving our hypotheses hinges on showing differences in the distributions of the perceived boundaries. Consider the grayscale results shown in Fig. 6.7, for example.

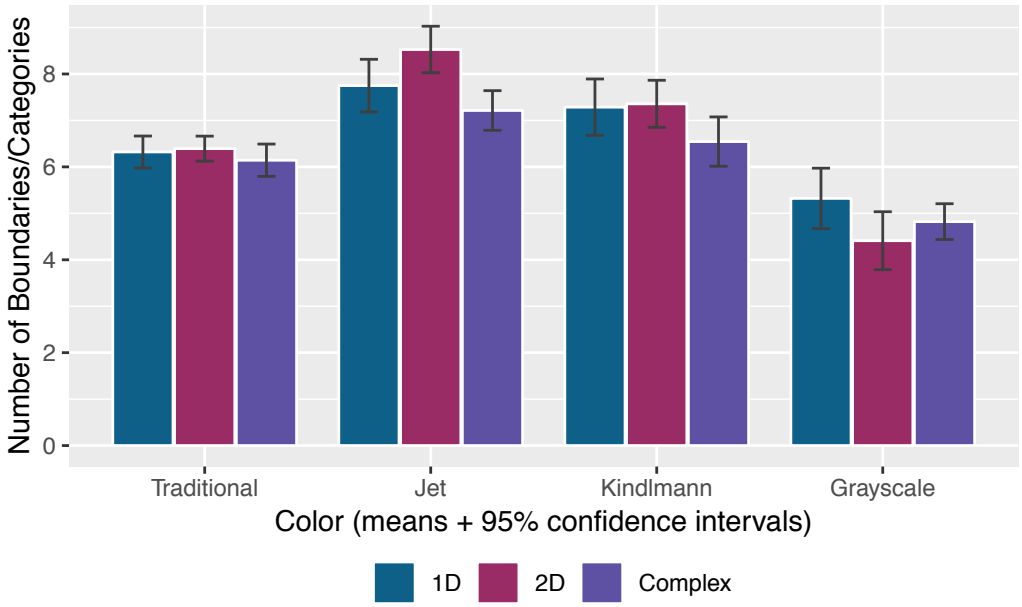


Fig. 6.7. The mean number of delineations that participants perceived and/or placed along with the 95% confidence interval for each color map and dataset. Descriptive statistical analysis indicates that each rainbow color map elicited significantly more delineations than grayscale, that the 1D dataset elicited significantly more delineations than the complex dataset, and that there were significant color-map:dataset interactions.

As ”black,” ”white,” and ”gray” are all basic color terms in English [9], meaning grayscale contains multiple color categories, we expect participants to count and delineate boundaries in the grayscale stimuli. We further expect, however, that any delineated boundaries will be randomly distributed in the 1D and 2D stimuli, but centered around data features in the complex stimuli. Likewise, for each rainbow color map, we anticipate that the delineated boundaries will center around color-map artifacts in the 1D and 2D stimuli, but be confounded by data features in the complex stimuli. In each case, understanding the distribution of participants' delineations is critical.

6.5.2 H1: Evidence of Implicit Discretization

Fig. 6.8 provides an overview of the distribution of participants' placed delimiters. For each color map, dataset, and instruction condition, we use kernel density estimation (KDE) to calculate a probability density function (pdf) from the participants' collective delimiter placements. The pdfs shown use different bandwidths, each computed from the associated delimiter placements through multiple iterations of *leave-subject-out* Monte Carlo cross-validation (CV), utilizing a train-test split of 90% to 10%. Leave-subject-out CV is an established blocked CV approach with theoretic optimality that accounts for dependencies within subject responses [98], [154], [162], [209]. Peaks in the resulting pdfs highlight consistencies across participants' placed delimiters. **Fig. 6.8** also illustrates that the distributions of participants' delimiters are largely similar across both instruction sets.

For each of the three rainbow color maps, participants' delimiter placements are clustered around distinct locations, but those locations vary both across the color maps for a given dataset and across the datasets for a given color map. In each case, however, the clusters are irregularly spaced, confirming that the perceived bands in rainbow color maps are not uniform in size. By comparison, participants' delimiter placements for the grayscale stimuli are more uniformly distributed across the normalized data value range. Some of the patterns in the grayscale responses, however, can be explained either by artifacts caused by mapping the perceptual grayscale color map into 24-bit RGB color or by data features in the case of the complex dataset. The former is illustrated in the 1D grayscale results presented in **Fig. 6.9**, where breaking out participants' delimiter placements by individual shows responses clustered around a series of doubled values in the color map.

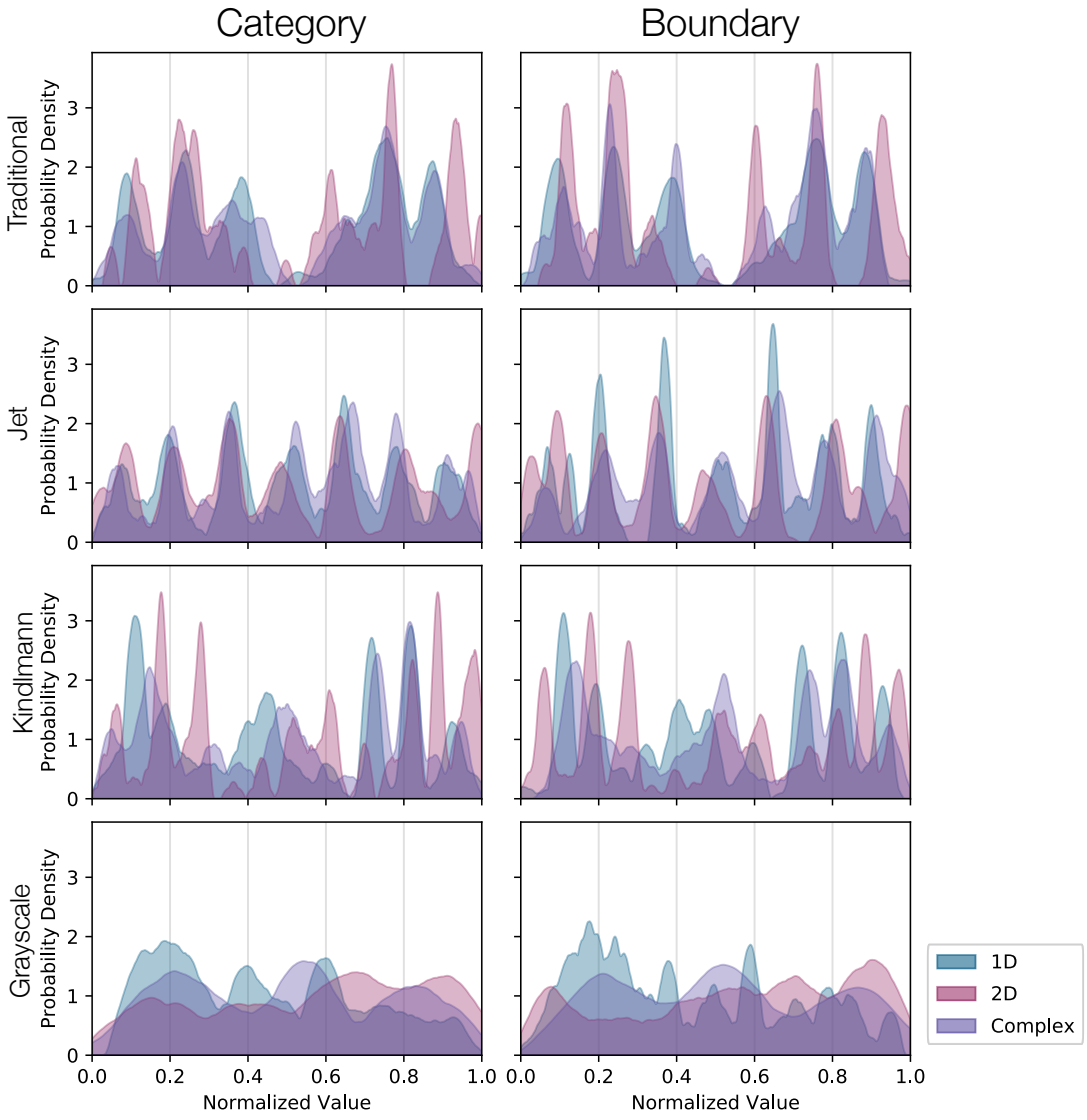
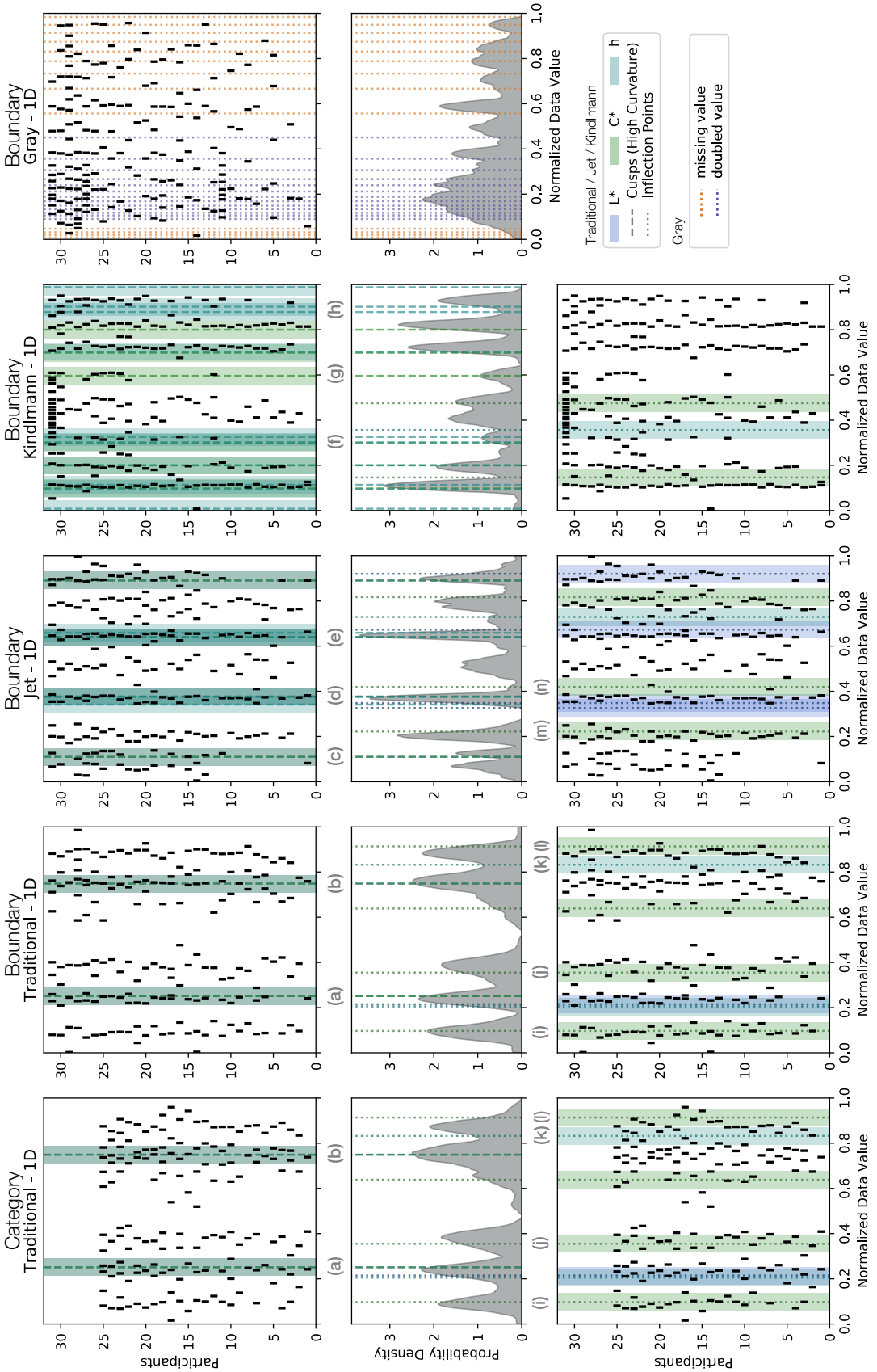


Fig. 6.8. Probability density functions fit to the participants' collective delimiter placements, partitioned by wording condition, color map, and dataset. Peaks highlight clusters in participants' responses for all three rainbow color maps across all three datasets. Participants' grayscale responses also show a few larger clusters but, overall, are more uniformly distributed.

Fig. 6.9. An overview of participants' delimiter placements in the 1D experimental stimuli. In the top and bottom plots, each row of marks contains the delimiters placed by a single participant with participants ordered along the y-axis by the average number of delimiters they placed overall. For each rainbow color-mapped stimuli, dotted lines show the locations of cusps (top) and inflection points (bottom), with corresponding bands showing the expected individual variation for color category boundaries [200]. The same indicators are also overlaid on the pdfs (center) estimated from the delimiters. Convenience labels (a)-(n) are included for indicators referred to in the text. For the grayscale stimuli, dotted lines mark the locations of color-map artifacts, with doubled values corresponding to a large response cluster.


Category
Traditional - 1D
Boundary
Traditional - 1D
Boundary
Jet - 1D
Boundary
Kindlmann - 1D
Boundary
Gray - 1D
Participants
Probability Density
Normalized Data Value
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
(l)
(m)
(n)
Traditional / Jet / Kindlmann
L*
C*
h
Cusps (High Curvature)
Inflection Points
Gray
missing value
doubled value

Further discussion of grayscale patterns is included in the supplemental material archive for this dissertation.

When taken together, these results provide empirical support for H1. Participants appear to implicitly discretize rainbow color-mapped datasets with marked consistency across individuals. Moreover, as hypothesized, this discretization varies across the different rainbow color maps tested. Given the provided task, it is possible individuals may have also attempted to use color categories when reasoning about grayscale; however, the clustering in participants' responses is less consistent than for the rainbow color maps.

6.5.3 H2: Clear Correspondences Beyond Luminance

In comparing participants' responses to the derived indicator sets, the results support hypothesis H2. As we hypothesized, luminance does play a role in the implicit discretization observed in the rainbow color-mapped stimuli, but so does chroma and, to a lesser extent, hue. As shown in **Fig. 6.9**, the majority of participants' response trends correspond to cusps or inflection points in the CIELCh profiles of each color map. Not every indicator predicts a response trend, however. Here, we provide an overview of participants' responses related to the 1D dataset, where any clusters or trends should be artifacts of the color maps themselves. The supplemental material archive for this dissertation contains similar findings across each of the datasets and instruction conditions in the study.

For each perceptual dimension of color (lightness, chroma, and hue), cusps in the CIELCh profiles of a given color map exhibit some correspondence with participants' response trends. As prior work predicts [8], [119], the cusps associated with the salient cyan and yellow features in both the traditional rainbow (**Fig. 6.9a** and **Fig. 6.9b**) and jet color maps (**Fig. 6.9d** and **Fig. 6.9e**) correspond to strong participant response trends in 1D. In addition, strong response trends align with some cusps in the chroma and hue profiles of the Kindlmann color map. In other cases, however, cusps in the perceptual dimensions of the color maps have weak or no correspondence with participants' delimiter placements. The coincident lightness and chroma indicators corresponding to the dark-blue:blue boundary in the jet color map (**Fig. 6.9c**), for example, capture only the right-hand side of a split response trend. Additionally, the 1D Kindlmann results reveal

chroma and hue cusps that either correspond to weak trends that only emerge when participants place a large number of boundaries (Fig. 6.9f and Fig. 6.9g) or fail to correspond to any response trends (Fig. 6.9h).

The correspondences between participants' response trends and inflection points in each color map's CIELCh profiles are similarly mixed. Inflection points in chroma capture a number of strong response trends that are not predicted by cusps, such as those corresponding to potential blue:light-blue boundaries in the traditional rainbow (Fig. 6.9i) and jet color maps (Fig. 6.9m). Again, however, not every inflection point corresponds to a response trend. Certain inflection-point indicators exhibit pronounced offsets from their associated response trends (Fig. 6.9j to Fig. 6.9l), whereas others have no corresponding response trend (Fig. 6.9n).

6.5.4 H3: Unexpected Patterns in Data-Driven Variation

As shown in Fig. 6.10 and Fig. 6.11, for each of the rainbow color maps tested, the different datasets show shifts in the locations and consistency of clusters in participants' responses. Consequently, our results provide support for our H3 hypothesis that implicit discretization depends on the dataset. The variation that we found, however, differs from what we originally anticipated. We expected that a complex stimuli based on real-world data would result in an implicit discretization different from a smoothly varying 1D or 2D stimuli. This reasoning, however, neither predicts nor explains the observed differences in our participants' response trends for the 1D versus 2D stimuli for each rainbow color map. Further, for the complex stimuli, we found no clear indication that the underlying data features impacted participants' response trends. Neither Fig. 6.10 nor Fig. 6.11 provides evidence of data features creating or accentuating perceived boundaries.

Fig. 6.10 provides an overview of participants' delimiter placements across the three experimental stimuli encoded using the traditional rainbow color map. This overview contains several notable differences in the strengths and locations of participants' response trends. The response trend associated with the leftmost inflection-point indicator (Fig. 6.10a) shifts to the right in the 2D stimuli compared to 1D stimuli but dissipates into a weaker trend in the complex stimuli. The other end of the color map (Fig. 6.10d) exhibits a pronounced shift in the location of participants' 1D and 2D response trends. Also, toward

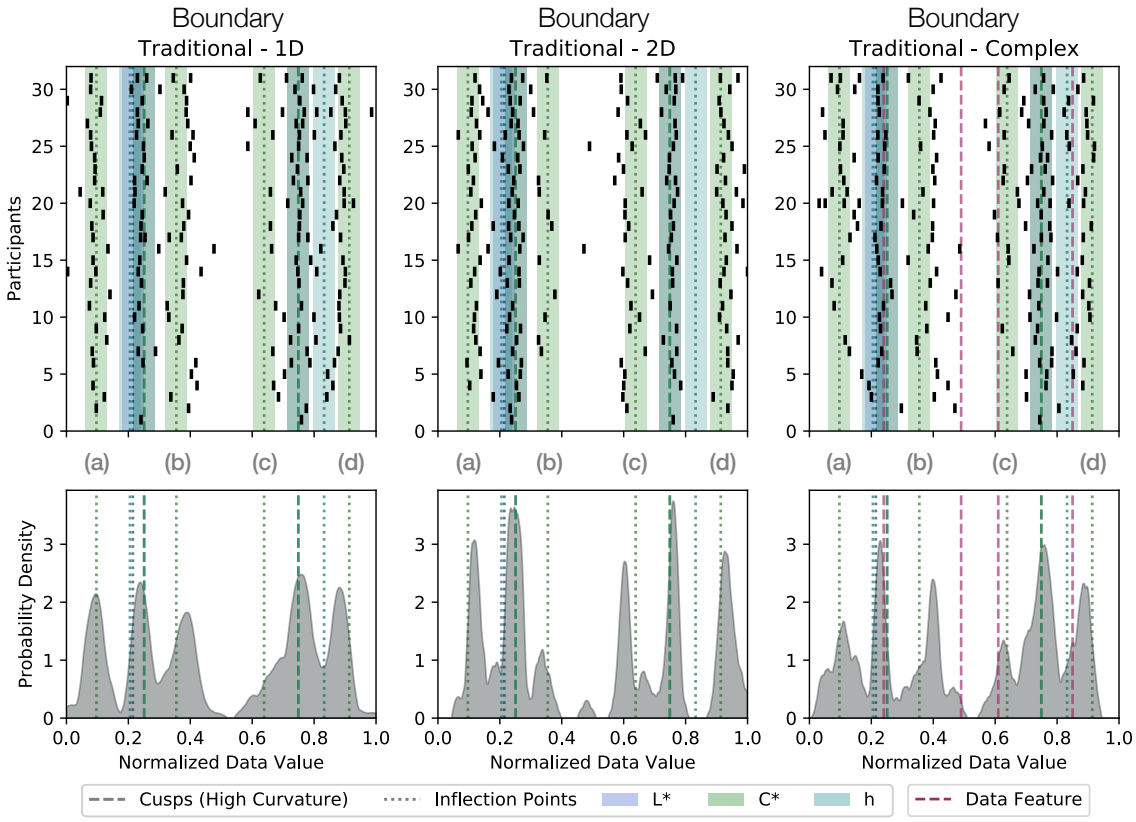


Fig. 6.10. An overview of participants' boundary placements within the traditional rainbow color map stimuli, showing changes in participants' response trends across the different datasets. Contrary to expectations, the changes include noticeable shifts in responses between 1D and 2D. Further, the underlying data features in the complex dataset appear to have had minimal impact on participants' responses. Convenience labels (a)-(d) are included for specific indicators referred to in the text.

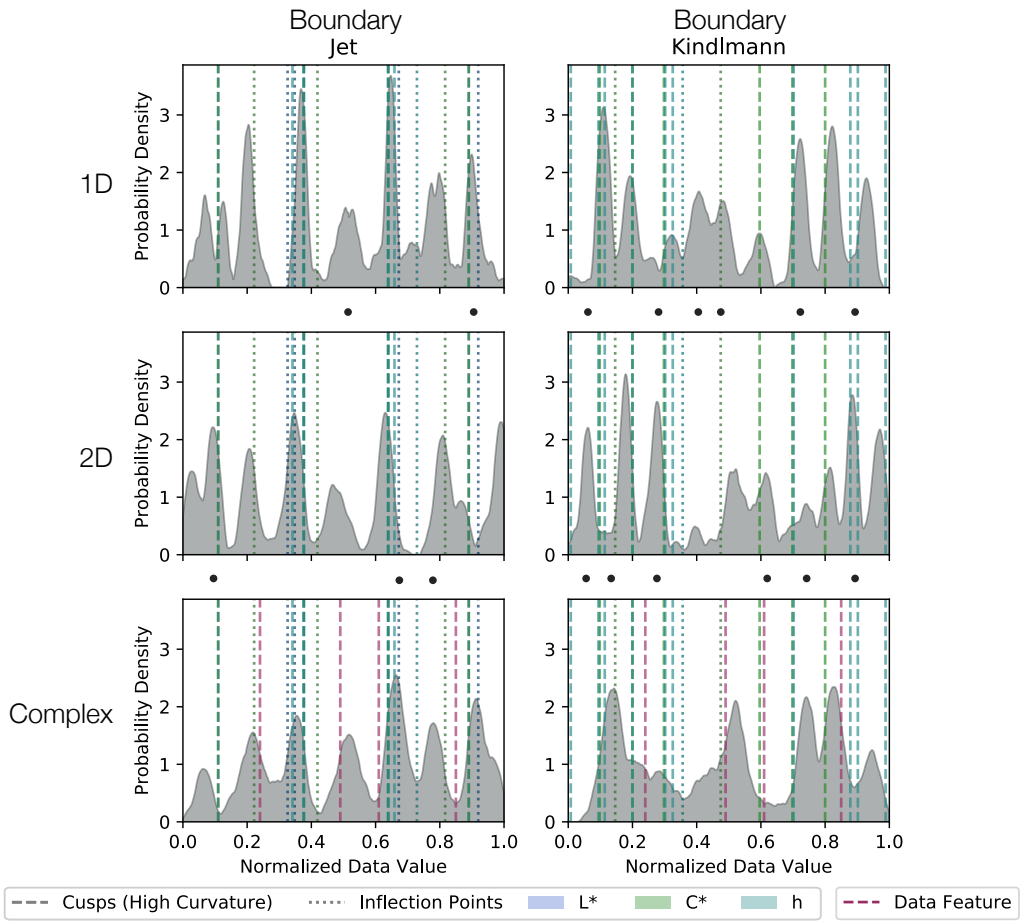


Fig. 6.11. The probability density plots of participants' delimiter placements for both the jet and Kindlmann stimuli, which also exhibit significant variation in participants' response trends across all three datasets. Notable differences between the adjacent plots are marked with a • symbol.

the center of the color map, two more trends (near **Fig. 6.10b** and **Fig. 6.10c**, respectively) vary in strength across the three datasets. Even though each of these trends happens to correlate with an inflection point in chroma, these shifts do not appear specific to trends associated with either chroma variation or inflection point indicators. **Fig. 6.11** exhibits similar variations across both the jet and Kindlmann stimuli that affect trends corresponding to a wide variety of indicator types.

In summary, the results do support H3, but they also highlight questions about the nature of the interaction between color maps and datasets. Additional research is needed to determine which dataset characteristics produce variation in implicit discretization and

what the underlying perceptual mechanisms for this effect are.

6.6 Discussion

The perceived banding in rainbow color-mapped visualizations depends on the data being encoded, but in a way that is neither predicted nor readily explained by existing theory. The results presented in this chapter show differences in participants' discretizations of smoothly varying linear and radial data gradients that cannot be immediately explained by the human visual system's decreased chromatic sensitivity to high-spatial-frequency information [83], [150], [194], [198]. Given that both the 2D and complex datasets contain varying gradient magnitudes, known interactions between color and size might account for the data-driven variation we observed. Models for color-size effects [179], however, have not yet been extended to handle the complexities of continuous scalar fields.

The results also provide evidence that implicit discretization is driven by more than just luminance. We illustrate correspondences between the perceived banding in rainbow color maps and both cusps and inflection points in each of the perceptual dimensions of those color maps. The rainbow color maps that we tested, however, contain coincident and proximately located indicators, making it challenging to fully separate the effects of luminance, chroma, and hue. Assessing what truly drives many of the individual response trends that we observed would require more systematic control than was present in our exploratory study.

The results further indicate that the implicit discretization caused by rainbow color maps is relatively consistent across individuals. Although the nature of the study's tasks did not allow us to directly assess the amount of individual variation across participants' perceived hue-bands, estimates of individual variation from prior color category experiments [200] approximate the variation in many of the response trends found in the results. The results also confirm that the perceived hue-bands in rainbow color maps are, indeed, irregularly spaced. We observed no apparent confounding impacts from sex differences or intermonitor variation, although both are potential limitations that should be addressed in subsequent work.

The study's findings show that different datasets create unpredictable variation in the perceived hue bands in rainbow color-mapped visualizations. This unpredictability pres-

ents challenges for experts in a variety of scientific fields [13], [25], [37], [89], [145], [214], where the implicit discretization in rainbow color maps is used either for classification or as a heuristic for quick visual comparisons. Moreover, given that the results show similar data-driven inconsistencies in the Kindlmann color map, which follows the visualization community's core guidelines regarding effective color usage [8], these same practical challenges may apply to a broader set of multihue continuous color scales. This variation could also explain the recent finding of Dasgupta et al. that hue banding negatively impacted magnitude estimation [37].

Despite the visualization community's promotion of more perceptually appropriate alternatives [8], [13], [60], [86], [95], [119], [188], rainbow color maps remain commonplace in a variety of scientific domains, including medicine [13], atmospheric and climate sciences [38], [145], bioengineering [25], aerospace [89], and astronomy [214]. Although domain convention is often used to justify the inclusion of rainbow color map variants in visualization systems [143], [145], [197], we still do not understand *why* experts continue to gravitate to spectral schemes. Cited reasons include familiarity [13], [145], aesthetic preference [13], [18], [120], and ease of use [14], [120], but evidence also suggests that rainbow color maps may be a *satisficing* design choice for specific types of tasks, such as locating and quantifying extreme values [38], [150], [194], [195], [198]. Improving our understanding of both how rainbow color maps are used and the ways in which they are ineffective could lead to improved guidance regarding effective color usage more broadly.

6.7 Summary

In this chapter, we presented an exploratory study investigating the nature of hue banding in rainbow color maps. The results represent a necessary first step in addressing open questions, including whether rainbow color maps implicitly discretize encoded data into hue-based bands and how that discretization varies across different individuals, datasets, and spectral schemes. The results presented in this chapter also suggest that the visualization community's current understanding of how rainbow color maps are perceived and used remains incomplete.

The results begin to address gaps in our understanding of the nature of implicit discretization in common spectral schemes, but they also leave open questions. Rainbow

color maps appear to discretize data into hue-based bands, but we currently have an insufficient understanding of the mechanisms that drive this phenomenon and no method for modeling or predicting the banding. Additional work is needed to explore whether the gradient variation in the encoded datasets plays a role in the response trend variations we observed and to examine how to minimize implicit discretization in multihue color maps. Moreover, the results do not fully deal with the questions established in the course of our larger mixed-methods research program. They do not explicitly address either whether implicit discretization affords similar benefits to explicit discretization or whether implicit discretization plays a role in the continued widespread prevalence of rainbow color maps.

We argue, however, that the work presented in this chapter still provides additional validation of mixed-methods research approaches. Even though the scope of this study differs from what we set out to explore, the results provide new insights into the real-world impacts of the conflicting domain guidance regarding the use of rainbow color maps for visualizing 2D scalar data. Moreover, this particular study used qualitative visual analysis methods to analyze quantitive data; and in doing so, highlighted major open questions regarding the core arguments about *why* rainbow color maps are harmful. We believe that addressing the knowledge gaps highlighted by these results will result in improved guidelines for color encodings of 2D scalar fields.

CHAPTER 7

REFLECTIONS AND FUTURE WORK

This dissertation comprises an emergent multiphase mixed-methods research program, designed to generate new knowledge through the revision and refinement of visualization guidelines regarding the use of color in encoding 2D scalar fields. We use a grounded understanding of how color is used in meteorological practice, developed throughout a 2-year design study, as a lens for investigating two major guidance conflicts regarding color usage between visualization and other domains. We examined longstanding, contrasted claims from both the visualization and cartographic communities regarding the use of discretized color maps in visualizing 2D scalar data, and we found empirical evidence of a speed-accuracy trade-off that conflicts with both domains suppositions about best practices [8], [66], [107], [117], [145], [171], [186], [199]. Based on those results, we then conducted an empirical examination of the role of *implicit* discretization in individuals' perceptions of rainbow color maps, an encoding choice that remains prevalent in a diverse set of scientific domains [13],[25],[38],[89],[145],[214] despite decades of vilification within the visualization community [8], [13], [60], [86], [95], [119], [188]. The results of this second study indicate that, although rainbow color maps do implicitly discretize data, existing theory [83], [150], [159], [179], [198] does not adequately explain how or why.

These two sets of findings improve our understanding of color encodings for 2D scalar fields, and together, constitute the primary contribution of this dissertation. The findings also validate our emergent mixed-methods research program, as subsequent studies build upon and contextualize prior studies. As secondary contributions, however, the various pieces of research in this dissertation also touch on two ancillary topics: the inherent limits of reductionist quantitative and qualitative insight-based evaluations and the potential for mixed-methods research programs to reconcile quantitative and qualitative research results.

The various studies in this dissertation highlight the inherent trade-offs among experimental control, ecological validity, and generalizability. Qualitative insight-based evaluations, such as the design study discussed in Chapter 3, offer a strong sense of ecological validity but lack the necessary control to understand the impact of particular design decisions. In contrast, the prioritization of experimental control in reductionist quantitative evaluations allows us to establish causation but can leave open questions about either the generalizability of the results or their applicability in complex real-world decision-making processes. Chapter 4 illustrates the complexities and pitfalls involved in attempting to prioritize ecological validity in a quantitative study. Chapters 5 and 6, on the other hand, show quantitative experimental designs geared toward psychological realism and a simplified focus on a general perceptual phenomenon. Each successive study makes greater compromises to ecological validity but also improves our ability to draw inferences.

As an emergent mixed-methods research program, however, these studies contextualize one another, allowing us to reconcile certain qualitative observations with quantitative explanations. Like orthogonal views, qualitative and quantitative studies can offer different perspectives of the same underlying phenomenon that, when taken together, provide a more accurate understanding of reality. As a case study, this dissertation demonstrates that mixed-method research programs are a powerful tool for reconciling qualitative and quantitative research contributions. In particular, we argue that using qualitative contributions from design studies to inspire and contextualize quantitative experimental research can be advantageous. The scope of these advantages, however, needs to be explored and further validated in future work.

Another broader implication of the work presented in this dissertation is that it repeatedly illustrates the importance of examining implicit assumptions contained in prior research. Furthermore, it suggests that examining conflicts between widespread domain practices and visualization guidance can be a useful way to locate such assumptions. Viewed as a case study, we believe that this dissertation demonstrates that domain guidance conflicts deserve to be more closely examined by the visualization community moving forward.

This dissertation also makes it clear that the visualization community's understanding of color encodings for 2D scalar fields remains far from complete. Throughout this disser-

tation, we highlighted several important avenues for future work regarding color. We need research exploring the effects of interactive and customizable color scales, particularly in contexts with evolving task demands. We need to better understand how specific characteristics in the encoded data – such as topological features, spatial frequencies, and gradients – impact our ability to understand visualizations that employ color maps with either explicit or implicit discretization. We also need to resolve the perceptual mechanisms driving implicit discretization and investigate whether those mechanisms can be leveraged to confer similar effects to explicit discretization. Arguably, this dissertation raises more questions than it answers regarding color. We contend, however, that this observation points to a large potential impact for this work.

This dissertation generates new knowledge about color encodings for 2D scalar fields while illustrating the benefits of emergent mixed methods research programs. We anticipate that it will not only inform future work on color encodings but also, as a case study, influence a wider variety of research attempting to combine qualitative insights with quantitative explanations.

REFERENCES

[1] *Ferret*. (1995) Pacific Marine Environmental Laboratory, NOAA. Accessed: Mar. 2015. [Online]. Available: http://ferret.pmel.noaa.gov/Ferret/.

[2] *Insight Segmentation and Registration Toolkit (ITK)*. (Revision 4.7.2) Kitware. [Online]. Available: http://www.itk.org/.

[3] *Metview*. (Version 4.5) European Centre for Medium-Range Weather Forecasts. Accessed: Mar. 2015. [Online]. Available: https://confluence.ecmwf.int/display/METV/Metview.

[4] *Processing*. (Version 2.2.1) [Online]. Available: https://processing.org/.

[5] *Wgrib2*. (2005) National Weather Service Climate Prediction Center. Accessed: Mar. 2015. [Online]. Available: https://www.cpc.ncep.noaa.gov/products/wesley/wgrib2/.

[6] N. Andrienko and G. Andrienko, *Exploratory Analysis of Spatial and Temporal Data: A Systematic Approach*. Berlin/Heidelberg, Germany: Springer-Verlag, 2006.

[7] L. Bartram, A. Patra, and M. Stone, "Affective color in visualization," in *Proc. 2017 CHI Conf. on Hum. Factors in Computing Syst.*, ser. CHI '17. New York, NY, USA: ACM, 2017, pp. 1364–1374.

[8] L. D. Bergman, B. E. Rogowitz, and L. A. Treinish, "A rule-based tool for assisting colormap selection," in *Proc. IEEE Conf. Visualization, 1995: Visualization '95.* Washington, DC, USA: IEEE Comput. Soc., Oct. 1995, pp. 118–125.

[9] B. Berlin and P. Kay, *Basic Color Terms: Their Universality and Evolution*. Stanford, CA, USA: CSLI Publications, Mar. 1999.

[10] J. Bertin, *Semiology of Graphics: Diagrams, Networks, Maps*, W. J. Berg, Ed. Madison, WI, USA: Univ. Wisconson Press, 1983.

[11] B. Bied Sperling and T. S. Tullis, "Are you a better "mouser" or "trackballer"?: A comparison of cursor-positioning performance," *SIGCHI Bull.*, vol. 19, no. 3, pp. 77–81, Jan. 1988.

[12] R. Borgo, J. Kehrer, D. H. S. Chung, E. Maguire, R. S. Laramee, H. Hauser, M. Ward, and M. Chen, "Glyph-based visualization: Foundations, design guidelines, techniques and applications," in *Eurographics 2013 - State of the Art Reports*, M. Sbert and L. Szirmay-Kalos, Eds. The Eurographics Association, 2012.

[13] 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 Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2479–2488, Dec. 2011.

[14] D. Borland and R. M. Taylor, "Rainbow color map (still) considered harmful," *IEEE Comput. Graph. Appl.*, vol. 27, no. 2, pp. 14–17, Mar. 2007.

[15] P. D. Bothwell and L. M. Richardson, "Forecasting lightning using a perfect prog technique applied to multiple operational models," in *15th Int. Conf. Atmospheric Electricity*, Norman, OK, USA, Jun. 15–20, 2014.

[16] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2376–2385, Dec. 2013.

[17] C. A. Brewer, "Color use guidelines for mapping and visualization," in *Visualization in Modern Cartography*, ser. Modern Cartography Series, A. M. MacEachren and D. F. Taylor, Eds. Oxford, U.K.: Pergamon Press, 1994, vol. 2, pp. 123–147.

[18] ——, "Spectral schemes: Controversial color use on maps," *Cartography Geographic Inform. Syst.*, vol. 24, no. 4, pp. 203–220, Jan. 1997.

[19] ——, *Designing Better Maps: A Guide for GIS Users*, 2nd ed. Redlands, CA, USA: Esri Press, Dec. 2015.

[20] C. A. Brewer, A. M. MacEachren, L. W. Pickle, and D. Herrmann, "Mapping mortality: Evaluating color schemes for choropleth maps," *Ann. Assoc. Amer. Geographers*, vol. 87, no. 3, pp. 411–438, Sep. 1997.

[21] M. B. Brewer, "Research design and issues of validity," in *Handbook of Research Methods in Social and Personality Psychology*, H. T. Reis and C. M. Judd, Eds. New York, NY, USA: Cambridge Univ. Press, 2000, pp. 3–16.

[22] B. Brogaard and D. E. Gatzia, "Cortical color and the cognitive sciences," *Topics Cogn. Sciences*, vol. 9, no. 1, pp. 135–150, Jan. 2017.

[23] B. Brown, S. Reeves, and S. Sherwood, "Into the wild: Challenges and opportunities for field trial methods," in *Proc. SIGCHI Conf. Hum. Factors Computing Syst.*, ser. CHI '11. Vancouver, BC, Canada: ACM, 2011, pp. 1657–1666.

[24] 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 Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 923–933, Jan. 2018.

[25] B. M. Burton, J. D. Tate, W. Good, and R. S. Macleod, "The role of reduced left ventricular, systolic blood volumes in ST segment potentials overlying diseased tissue of the ischemic heart," *Comput. Cardiology*, vol. 43, pp. 209–212, Sep. 2016.

[26] P. Carayon, S. Kianfar, Y. Li, A. Xie, B. Alyousef, and A. Wooldridge, "A systematic review of mixed methods research on human factors and ergonomics in health care," *Applied Ergonomics*, vol. 51, pp. 291–321, Nov. 2015.

[27] S. Carpendale, "Evaluating information visualizations," in *Inform. Visualization*, ser. Lecture Notes in Computer Science, A. Kerren, J. T. Stasko, J.-D. Fekete, and C. North, Eds. Berlin/Heidelberg, Germany: Springer, 2008, vol. 4950, pp. 19–45.

[28] 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," *Int. J. Geographic Inform. Sci.*, vol. 30, no. 7, pp. 1377–1404, Jul. 2016.

[29] J. Chuang, M. Stone, and P. Hanrahan, "A probabilistic model of the categorical association between colors," *Color Imaging Conf.*, vol. 2008, no. 1, pp. 6–11, Jan. 2008.

[30] W. S. Cleveland and R. McGill, "Graphical perception: Theory, experimentation, and application to the development of graphical methods," *J. Amer. Statistical Assoc.*, vol. 79, no. 387, pp. 531–554, Sep. 1984.

[31] J. Cox, D. House, and M. Lindell, "Visualizing uncertainty in predicted hurricane tracks," *Int. J. Uncertainty Quantification*, vol. 3, no. 2, pp. 143–156, 2013.

[32] J. W. Creswell and V. L. Plano Clark, *Designing and Conducting Mixed Methods Research*, 2nd ed. Los Angeles, CA, USA: SAGE, Jun. 2010.

[33] R. G. Cromley, "Classed versus unclassed choropleth maps: A question of how many classes," *Cartographica: Int. J. Geographic Inform. Geovisualization*, vol. 32, no. 4, pp. 15–27, Oct. 1995.

[34] P. Daipha, "Weathering risk: Uncertainty, weather forecasting, and expertise," *Sociology Compass*, vol. 6, no. 1, pp. 15–25, Jan. 2012.

[35] ——, "From bricolage to collage: The making of decisions at a weather forecast office," *Sociological Forum*, vol. 30, no. 3, pp. 787–808, Sep. 2015.

[36] E. Danovaro, L. De Floriani, L. Papaleo, and M. Vitali, "A multi-resolution representation for terrain morphology," in *Geographic Inform. Sci.*, ser. Lecture Notes in Computer Science, M. Raubal, H. J. Miller, A. U. Frank, and M. F. Goodchild, Eds. Berlin/Heidelberg, Germany: Springer, 2006, pp. 33–46.

[37] A. Dasgupta, J. Poco, B. Rogowitz, K. Han, E. Bertini, and C. T. Silva, "The effect of color scales on climate scientists' objective and subjective performance in spatial data analysis tasks," *IEEE Trans. Vis. Comput. Graphics*, pp. 1–1, 2018.

[38] A. Dasgupta, J. Poco, Y. Wei, R. Cook, E. Bertini, and C. T. Silva, "Bridging theory with practice: An exploratory study of visualization use and design for climate model comparison," *IEEE Trans. Vis. Comput. Graphics*, vol. 21, no. 9, pp. 996–1014, Sep. 2015.

[39] S. Deitrick and R. Edsall, "The influence of uncertainty visualization on decision making: An empirical evaluation," in *Progress in Spatial Data Handling*. Berlin/Heidelberg, Germany: Springer, 2006, pp. 719–738.

[40] M. Denscombe, "Communities of practice: A research paradigm for the mixed methods approach," *J. Mixed Methods Res.*, vol. 2, no. 3, pp. 270–283, Jul. 2008.

[41] B. D. Dent, J. Torguson, and T. W. Hodler, *Cartography: Thematic Map Design*, 6th ed. New York, NY, USA: McGraw-Hill Education, Aug. 2008.

[42] M. W. Dobson, "Choropleth maps without class intervals?: A comment," *Geographic Anal.*, vol. 5, no. 4, pp. 358–360, Oct. 1973.

[43] ——, "Commentary: Perception of continuously shaded maps," *Ann. Assoc. Amer. Geographers*, vol. 70, no. 1, pp. 106–107, Mar. 1980.

[44] M. A. Dooley and S. J. Lavin, "Visualizing method-produced uncertainty in isometric mapping," *Cartographic Perspectives*, no. 56, pp. 17–36, Mar. 2007.

[45] C. A. Doswell, "Weather forecasting by humans — heuristics and decision making," *Weather and Forecasting*, vol. 19, no. 6, pp. 1115–1126, Dec. 2004.

[46] B. E. Doty and J. L. Kinter III, "Geophysical data analysis and visualization using the grid analysis and display system," in *Visualization Techniques in Space and Atmospheric Sciences*, E. Szuszczewicz and J. Bredekamp, Eds. Washington, DC, USA: NASA, Report NASA-SP-519, 1995, pp. 209–219.

[47] S. Eddins, "Rainbow color map critiques: An overview and annotated bibliography," MathWorks, Tech. Rep. 92238v00, Sep. 2014.

[48] M. Eisemann, G. Albuquerque, and M. Magnor, "Data driven color mapping," in *Int. Workshop Vis. Analytics (EuroVA 2011)*, S. Miksch and G. Santucci, Eds. The Eurographics Association, 2011, pp. 5–8.

[49] K. Engel, M. Hadwiger, J. M. Kniss, C. Rezk-Salama, and D. Weiskopf, *Real-Time Volume Graphics*. Natick, MA, USA: A K Peters, 2006.

[50] S. I. Fabrikant, S. R. Hespanha, and M. Hegarty, "Cognitively inspired and perceptually salient graphic displays for efficient spatial inference making," *Ann. Assoc. Amer. Geographers*, vol. 100, no. 1, pp. 13–29, Jan. 2010.

[51] M. D. Fairchild, *Color Appearance Models*, 3rd ed., ser. Wiley-IS&T Series in Imaging Science and Technology. Chichester, U.K.: Wiley, 2013.

[52] G. A. Fine, *Authors of the Storm: Meteorologists and the Culture of Prediction*. Chicago, IL, USA: Univ. Chicago Press, Oct. 2009.

[53] D. Fonseca, E. Redondo, and S. Villagrasa, "Mixed-methods research: A new approach to evaluating the motivation and satisfaction of university students using advanced visual technologies," *Universal Access Inform. Soc.*, vol. 14, no. 3, pp. 311–332, Aug. 2015.

[54] I. Fujishiro, T. Azuma, and Y. Takeshima, "Automating transfer function design for comprehensible volume rendering based on 3D field topology analysis," in *Proc. Conf. Visualization '99: Celebrating Ten Years*, ser. VIS '99. Washington, DC, USA: IEEE Comput. Soc.., Oct. 1999, pp. 467–470.

[55] R. L. Goldstone and A. T. Hendrickson, "Categorical perception," *WIREs Cogn. Sci.*, vol. 1, no. 1, pp. 69–78, Jan. 2010.

[56] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 3rd ed. Upper Saddle River, NJ, USA: Pearson, Aug. 2007.

[57] C. C. Gramazio, D. H. Laidlaw, and K. B. Schloss, "Colorgorical: Creating discriminable and preferable color palettes for information visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 521–530, Jan. 2017.

[58] D. A. Green, "A colour scheme for the display of astronomical intensity images," *Bull. Astron. Soc. India*, vol. 39, pp. 289–295, Jun. 2011.

[59] J. C. Greene, V. J. Caracelli, and W. F. Graham, "Toward a conceptual framework for mixed-method evaluation designs," *Educational Evaluation Policy Anal.*, vol. 11, no. 3, pp. 255–274, Sep. 1989.

[60] D. Gresh, "Self-corrected perceptual colormaps," IBM, Tech. Rep. RC24542 (W0804-104), 2008.

[61] R. E. Groop and P. Smith, "A dot matrix method of portraying continuous statistical surfaces," *Amer. Cartographer*, vol. 9, no. 2, pp. 123–130, Jan. 1982.

[62] E. Guilbert, "Multi-level representation of terrain features on a contour map," *Geoinformatica*, vol. 17, no. 2, pp. 301–324, Apr. 2013.

[63] J. Görtler, M. Spicker, C. Schulz, D. Weiskopf, and O. Deussen, "Stippling of 2D scalar fields," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 6, pp. 2193–2204, Jun. 2019.

[64] T. M. Hamill, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau, Y. Zhu, and W. Lapenta, "NOAA's second-generation global medium-range ensemble reforecast dataset," *Bull. Amer. Meteorological Soc.*, vol. 94, no. 10, pp. 1553–1565, Feb. 2013.

[65] J. R. Hanley, "Color categorical perception," in *Encyclopedia of Color Science and Technology*, M. R. Luo, Ed. New York, NY, USA: Springer, 2016, pp. 239–243.

[66] M. Harrower and C. A. Brewer, "ColorBrewer.org: An online tool for selecting colour schemes for maps," *Cartographic J.*, vol. 40, no. 1, pp. 27–37, Jun. 2003.

[67] C. G. Healey, "Choosing effective colours for data visualization," in *Proc. 7th Annu. IEEE Conf. Visualization: Visualization '96.* IEEE Comput. Soc.., Oct. 1996, pp. 263–270.

[68] J. Heer and M. Stone, "Color naming models for color selection, image editing and palette design," in *Proc. SIGCHI Conf. Hum. Factors Computing Syst.*, ser. CHI '12. New York, NY, USA: ACM, 2012, pp. 1007–1016.

[69] M. Hegarty, M. S. Canham, and S. I. Fabrikant, "Thinking about the weather: How display salience and knowledge affect performance in a graphic inference task," *J. Exp. Psychol.: Learning, Memory, Cogn.*, vol. 36, no. 1, pp. 37–53, 2010.

[70] M. Hegarty, H. S. Smallman, A. T. Stull, and M. S. Canham, "Naïve cartography: How intuitions about display configuration can hurt performance," *Cartographica: Int. J. Geographic Inform. Geovisualization*, vol. 44, no. 3, pp. 171–186, Oct. 2009.

[71] B. Hibbard and D. Santek, "The VIS-5D system for easy interactive visualization," in *Proc. 1st IEEE Conf. Visualization, 1990 (Visualization '90).* Washington, DC, USA: IEEE Comput. Soc., 1990, pp. 28–35.

[72] Z. Hu, J. R. Hanley, R. Zhang, Q. Liu, and D. Roberson, "A conflict-based model of color categorical perception: Evidence from a priming study," *Psychonometric Bull. Rev.*, vol. 21, no. 5, pp. 1214–1223, Oct. 2014.

[73] J. Hullman, E. Adar, and P. Shah, "Benefitting InfoVis with visual difficulties," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2213–2222, Dec. 2011.

[74] Y.-H. Hung and P. Parsons, "Assessing user engagement in information visualization," in *Proc. 2017 CHI Conf. Extended Abstracts on Hum. Factors in Computing Syst.*, ser. CHI EA '17. New York, NY, USA: ACM, 2017, pp. 1708–1717.

[75] J. D. Hunter, "Matplotlib: A 2D graphics environment," *Comput. Sci. Eng.*, vol. 9, no. 3, pp. 90–95, Jun. 2007.

[76] E. Imhof, *Cartographic Relief Presentation*. Redlands, CA, USA: Esri Press, Jun. 2007.

[77] M. J. Intons-Peterson, "Imagery paradigms: How vulnerable are they to experimenters' expectations?" *J. Exp. Psychol.: Hum. Perception Perform.*, vol. 9, no. 3, pp. 394–412, Jun. 1983.

[78] R. B. Johnson and A. J. Onwuegbuzie, "Mixed methods research: A research paradigm whose time has come," *Educational Researcher*, vol. 33, no. 7, pp. 14–26, Oct. 2004.

[79] R. B. Johnson, A. J. Onwuegbuzie, and L. A. Turner, "Toward a definition of mixed methods research," *J. Mixed Methods Res.*, vol. 1, no. 2, pp. 112–133, Apr. 2007.

[80] G. Jordan and J. D. Mollon, "A study of women heterozygous for colour deficiencies," *Vision Res.*, vol. 33, no. 11, pp. 1495–1508, Jul. 1993.

[81] S. Joslyn and J. LeClerc, "Decisions with uncertainty: The glass half full," *Current Directions Psychological Sci.*, vol. 22, no. 4, pp. 308–315, Aug. 2013.

[82] D. Kahneman, *Thinking, Fast and Slow*, 1st ed. New York, NY, USA: Farrar, Straus and Giroux, Oct. 2011.

[83] A. D. Kalvin, B. E. Rogowitz, A. Pelah, and A. Cohen, "Building perceptual color maps for visualizing interval data," in *Proc. SPIE, Hum. Vision Electr. Imag. V*, B. E. Rogowitz and T. N. Pappas, Eds., vol. 3959. SPIE, 2000, pp. 323–335.

[84] D. R. Kincaid and E. W. Cheney, *Numerical Analysis: Mathematics of Scientific Computing*, 3rd ed. Pacific Grove, CA, USA: Brooks Cole, Oct. 2001.

[85] G. Kindlmann and J. W. Durkin, "Semi-automatic generation of transfer functions for direct volume rendering," in *Proc. 1998 IEEE Symp. Volume Visualization*, ser. VVS '98. New York, NY, USA: ACM, 1998, pp. 79–86.

[86] G. Kindlmann, E. Reinhard, and S. Creem, "Face-based luminance matching for perceptual colormap generation," in *Proc. IEEE Conf. Visualization '02*, ser. VIS '02. IEEE Comput. Soc.., 2002, pp. 299–306.

[87] C. Kinkeldey, A. M. MacEachren, M. Riveiro, and J. Schiewe, "Evaluating the effect of visually represented geodata uncertainty on decision-making: Systematic review, lessons learned, and recommendations," *Cartography Geographic Inform. Sci.*, vol. 44, no. 1, pp. 1–21, Sep. 2015.

[88] C. Kinkeldey, A. M. MacEachren, and J. Schiewe, "How to assess visual communication of uncertainty? A systematic review of geospatial uncertainty visualisation user studies," *Cartographic J.*, vol. 51, no. 4, pp. 372–386, Nov. 2014.

[89] M. Kirby, T. Etiene, and C. Silva, ""Flow visualization" juxtaposed with "visualization of flow": Synergistic opportunities between two communities," in *51st AIAA Aerosp. Sci. Meeting New Horizons Forum Aerosp. Expo.*, Grapevine, TX, USA, Jan. 7–10, 2013.

[90] S. S. Kirschenbaum, J. G. Trafton, C. D. Schunn, and S. B. Trickett, "Visualizing uncertainty: The impact on performance," *Hum. Factors*, vol. 56, no. 3, pp. 509–520, May 2014.

[91] R. Kosara, "An empire built on sand: Reexamining what we think we know about visualization," in *Proc. 6th Workshop Beyond Time Errors: Novel Eval. Methods Visualization*, ser. BELIV '16. New York, NY, USA: ACM, 2016, pp. 162–168.

[92] M.-J. Kraak and F. Ormeling, *Cartography: Visualization of Spatial Data*, 3rd ed. London, U.K.: Taylor & Francis, 2013.

[93] M. P. Kumler and R. E. Groop, "Continuous-tone mapping of smooth surfaces," *Cartography Geographic Inform. Syst.*, vol. 17, no. 4, pp. 279–289, Jan. 1990.

[94] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical studies in information visualization: Seven scenarios," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 9, pp. 1520–1536, Sep. 2012.

[95] H. Levkowitz and G. T. Herman, "Color scales for image data," *IEEE Comput. Graph. Appl.*, vol. 12, no. 1, pp. 72–80, Jan. 1992.

[96] A. Lieberoth and A. Roepstorff, "Mixed methods in game research: Playing on strengths and countering weaknesses," in *Game Research Methods*, P. Lankoski and S. Björk, Eds. Pittsburgh, PA, USA: ETC Press, 2015, pp. 271–289.

[97] S. Lin, J. Fortuna, C. Kulkarni, M. Stone, and J. Heer, "Selecting semantically-resonant colors for data visualization," *Comp. Graph. Forum*, vol. 32, pp. 401–410, Jul. 2013.

[98] M. A. Little, G. Varoquaux, S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "Using and understanding cross-validation strategies. Perspectives on Saeb et al." *GigaScience*, vol. 6, no. 5, May 2017.

[99] S. Liu, D. Maljovec, B. Wang, P. T. Bremer, and V. Pascucci, "Visualizing high-dimensional data: Advances in the past decade," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 3, pp. 1249–1268, Mar. 2017.

[100] Y. Liu and J. Heer, "Somewhere over the rainbow: An empirical assessment of quantitative colormaps," in *Proc. 2018 CHI Conf. Hum. Factors Computing Syst.*, ser. CHI '18. New York, NY, USA: ACM, 2018, paper 598, pp. 1–12.

[101] D. Lloyd and J. Dykes, "Human-centered approaches in geovisualization design: Investigating multiple methods through a long-term case study," *IEEE Trans. Vis. Comput. Graphics*, vol. 17, no. 12, pp. 2498–2507, Dec. 2011.

[102] J. M. Loomis, J. J. Blascovich, and A. C. Beall, "Immersive virtual environment technology as a basic research tool in psychology," *Behavior Res. Methods, Instrum., Comput.*, vol. 31, no. 4, pp. 557–564, Dec. 1999.

[103] R. K. Lowe, "Components of expertise in the perception and interpretation of meteorological charts," in *Interpreting Remote Sensing Imagery: Human Factors*, R. R. Hoffman and A. B. Markman, Eds. Boca Raton, FL, USA: CRC Press, 2001, pp. 185–206.

[104] C. Lundström, P. Ljung, A. Persson, and A. Ynnerman, "Uncertainty visualization in medical volume rendering using probabilistic animation," *IEEE Trans. Vis. Comput. Graphics*, vol. 13, no. 6, pp. 1648–1655, Nov. 2007.

[105] A. M. MacEachren, "Visualizing uncertain information," *Cartographic Perspectives*, no. 13, pp. 10–19, Jun. 1992.

[106] ——, "Visualization in modern cartography: Setting the agenda," in *Modern Cartography Series*, ser. Visualization in Modern Cartography, A. M. MacEachren and D. R. F. Taylor, Eds. Oxford, U.K.: Pergamon Press, Jan. 1994, vol. 2, pp. 1–12.

[107] J. Mackinlay, "Automating the design of graphical presentations of relational information," *ACM Trans. Graph.*, vol. 5, no. 2, pp. 110–141, Apr. 1986.

[108] R. S. MacLeod, C. R. Johnson, and M. A. Matheson, "Visualization blackboard - visualizing bioelectric fields," *IEEE Comput. Graph. Appl.*, vol. 13, no. 4, pp. 10–12, Jul. 1993.

[109] M. Mahy, L. V. Eycken, and A. Oosterlinck, "Evaluation of uniform color spaces developed after the adoption of CIELAB and CIELUV," *Color Res. Appl.*, vol. 19, no. 2, pp. 105–121, Apr. 1994.

[110] J. Matejka, M. Glueck, T. Grossman, and G. Fitzmaurice, "The effect of visual appearance on the performance of continuous sliders and visual analogue scales," in *Proc. 2016 CHI Conf. Hum. Factors Computing Syst.*, ser. CHI '16. New York, NY, USA: ACM, 2016, pp. 5421–5432.

[111] P. J. McCarthy, D. Ball, and W. Purcell, "Project Phoenix - optimizing the machine-person mix in high-impact weather forecasting," in *22nd Conf. Weather Anal. Forecasting / 18th Conf. Numer. Weather Prediction*, Park City, UT, USA, Jun. 25–29, 2007.

[112] P. T. McCaslin, P. A. McDonald, and E. J. Szoke, "3D visualization development at NOAA Forecast Systems Laboratory," *SIGGRAPH Comput. Graph.*, vol. 34, no. 1, pp. 41–44, Feb. 2000.

[113] J. E. McGrath, "Methodology matters: Doing research in the behavioral and social sciences," in *Human-Computer Interaction*, R. M. Baecker, J. Grudin, W. A. S. Buxton, and S. Greenberg, Eds. San Francisco, CA, USA: Morgan Kaufmann, 1995, pp. 152–169.

[114] N. Mellado, D. Vanderhaeghe, C. Hoarau, S. Christophe, M. Brédif, and L. Barthe, "Constrained palette-space exploration," *ACM Trans. Graphics*, vol. 36, no. 4, article 60, pp. 1–14, Jul. 2017.

[115] V. Mitchell, K. L. Mackley, S. Pink, C. Escobar-Tello, G. T. Wilson, and T. Bhamra, "Situating digital interventions: Mixed methods for HCI research in the home," *Interacting Comput.*, vol. 27, no. 1, pp. 3–12, Jan. 2015.

[116] S. Mittelstädt, A. Stoffel, and D. A. Keim, "Methods for compensating contrast effects in information visualization," *Comput. Graph. Forum*, vol. 33, no. 3, pp. 231–240, Jun. 2014.

[117] S. Mittelstädt, D. Jäckle, F. Stoffel, and D. A. Keim, "ColorCAT: Guided design of colormaps for combined analysis tasks," in *Eurographics Conf. Visualization (EuroVis) - Short Papers*, E. Bertini, J. Kennedy, and E. Puppo, Eds. The Eurographics Association, 2015.

[118] S. Mittelstädt and D. A. Keim, "Efficient contrast effect compensation with personalized perception models," *Comput. Graph. Forum*, vol. 34, no. 3, pp. 211–220, Jun. 2015.

[119] K. Moreland, "Diverging color maps for scientific visualization," in *ISVC 2009: Advances Vis. Comput.*, ser. Lecture Notes in Computer Science, vol. 5876. Berlin/Heidelberg, Germany: Springer-Verlag, 2009, pp. 92–103.

[120] ——, "Why we use bad color maps and what you can do about it," *Electron. Imag.*, vol. 2016, no. 16, pp. 1–6, Feb. 2016.

[121] N. Moroney, M. D. Fairchild, R. W. G. Hunt, C. Li, M. R. Luo, and T. Newman, "The CIECAM02 color appearance model," in *IS&T/SID 10th Color Imag. Conf. (CIC '02)*, Scottsdale, Arizona, USA, Nov. 12–15, 2002. [Online]. Available: https://scholarworks.rit.edu/other/143

[122] J.-C. Muller, "Perception of continuously shaded maps," *Ann. Assoc. Amer. Geographers*, vol. 69, no. 2, pp. 240–249, Jun. 1979.

[123] ——, "Visual comparison of continuously shaded maps," *Cartographica: Int. J. Geographic Inform. Geovisualization*, vol. 17, no. 1, pp. 40–52, Spring 1980.

[124] J.-C. Muller and J. L. Honsaker, "Choropleth map production by facsimile," *Cartographic J.*, vol. 15, no. 1, pp. 14–19, Jun. 1978.

[125] A. H. Munsell, *A Color Notation*, 11th ed. Baltimore, MD, USA: Munsell Color Co., 1961.

[126] T. Munzner, *Visualization Analysis and Design*, 1st ed., ser. AK Peters Visualization Series. New York, NY, USA: A K Peters/CRC Press, Dec. 2014.

[127] D. Murray, J. McWhirter, S. Wier, and S. Emmerson, "The Integrated Data Viewer: A web-enabled application for scientific analysis and visualization," in *19th Conf. Int. Interactive Inform. Process. Syst. (IIPS) for Meteorol., Oceanogr., Hydrol*, Boston, MA, USA, Feb. 10–13, 2003, pp. 13.2:1–13.2:5.

[128] I. J. Murray, N. R. A. Parry, D. J. McKeefry, and A. Panorgias, "Sex-related differences in peripheral human color vision: A color matching study," *J. Vision*, vol. 12, no. 1, article 18, pp. 1–10, Jan. 2012.

[129] L. Nadav-Greenberg, S. L. Joslyn, and M. U. Taing, "The effect of uncertainty visualizations on decision making in weather forecasting," *J. Cognitive Eng. Decis. Making*, vol. 2, no. 1, pp. 24–47, Mar. 2008.

[130] National Centers for Environmental Prediction, "Office note 388 - GRIB 1, revised," Feb. 2005, [Online]. Available: http://www.nco.ncep.noaa.gov/pmb/docs/on388/.

[131] National Geophysical Data Center, "Bathymetry of Lake Erie and Lake Saint Clair," National Geophysical Data Center, Boulder, CO, World Data Center A for Marine Geology and Geophysics Report MGG-13, 1998.

[132] ——, "U.S. coastal relief model - Hawaii," National Geophysical Data Center, NOAA, 2005, doi: 10.7289/V5RF5RZZ.

[133] J. Nielsen and J. Levy, "Measuring usability: Preference vs. performance," *Commun. ACM*, vol. 37, no. 4, pp. 66–75, Apr. 1994.

[134] H.-K. Nienhuys, "The Lagom LCD monitor test pages," Lagom.nl. http://www.lagom.nl/lcd-test/ (accessed Jan. 5, 2016).

[135] T. Nocke, M. Flechsig, and U. Böhm, "Visual exploration and evaluation of climate-related simulation data," in *Proc. 39th Conf. Winter Simul.: 40 Years! The Best is yet to Come*, ser. WSC '07. Washington, DC, USA: IEEE Comput. Soc., 2007, pp. 703–711.

[136] E. Özgen and I. R. L. Davies, "Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis," *J. Exp. Psychol.: General*, vol. 131, no. 4, pp. 477–493, Dec. 2002.

[137] L. Padilla, P. S. Quinan, M. Meyer, and S. H. Creem-Regehr, "Evaluating the impact of binning 2D scalar fields," *IEEE Trans. Vis. Comput. Graphics*, vol. 23, no. 1, pp. 431–440, Jan. 2017.

[138] A. Pang, "Visualizing uncertainty in geo-spatial data," in *Proc. Workshop Intersections Geospatial Inform. Inform. Technol.* Washington, DC, USA: National Academies Committee of the Computer Science and Telecommunications Board, 2001.

[139] H. Pfister, B. Lorensen, C. Bajaj, G. Kindlmann, W. Schroeder, L. S. Avila, K. M. Raghu, R. Machiraju, and J. Lee, "The transfer function bake-off," *IEEE Comput. Graph. Appl.*, vol. 21, no. 3, pp. 16–22, May 2001.

[140] S. M. Pizer, "Intensity mappings to linearize display devices," *Comput. Graph. Image Process.*, vol. 17, no. 3, pp. 262–268, Nov. 1981.

[141] K. Pöthkow, B. Weber, and H.-C. Hege, "Probabilistic marching cubes," in *Proc. 13th Eurographics / IEEE - VGTC Conf. Visualization*, ser. EuroVis '11. The Eurographics Association & Wiley, 2011, pp. 931–940.

[142] B. E. Potter, J. A. Winkler, D. F. Wilhelm, R. P. Shadbolt, and X. Bian, "Computing the low-elevation variant of the Haines index for fire weather forecasts," *Weather and Forecasting*, vol. 23, no. 1, pp. 159–167, 2008.

[143] K. Potter, A. Wilson, P.-T. Bremer, D. Williams, C. Doutriaux, V. Pascucci, and C. R. Johnson, "Ensemble-Vis: A framework for the statistical visualization of ensemble data," in *Proc. 2009 IEEE Int. Conf. Data Mining Workshops*, ser. ICDMW '09. Washington, DC, USA: IEEE Comput. Soc., 2009, pp. 233–240.

[144] K. Potter, P. Rosen, and C. Johnson, "From quantification to visualization: A taxonomy of uncertainty visualization approaches," in *Uncertainty Quantification in Scientific Computing*, ser. IFIP Advances in Information and Communication Technology, A. M. Dienstfrey and R. F. Boisvert, Eds. Berlin/Heidelberg, Germany: Springer, 2012, vol. 377, pp. 226–249.

[145] P. S. Quinan and M. Meyer, "Visually comparing weather features in forecasts," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 389–398, Jan. 2016.

[146] P. S. Quinan, L. Padilla, S. Creem-Regehr, and M. Meyer, "Examining implicit discretization in spectral schemes," *Comput. Graph. Forum*, vol. 38, no. 3, pp. 363–374, Jun. 2019.

[147] P. S. Quinan, L. M. Padilla, S. H. Creem-Regehr, and M. Meyer, "Towards ecological vailidty in evaulating uncertainty," in *2015 Workshop Visualization Decis. Making Under Uncertainty (at IEEE VIS '15)*, Chicago, IL, Oct. 26, 2015.

[148] M. Rautenhaus, M. Böttinger, S. Siemen, R. Hoffman, R. M. Kirby, M. Mirzargar, N. Röber, and R. Westermann, "Visualization in meteorology — a survey of techniques and tools for data analysis tasks," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 12, pp. 3268–3296, Dec. 2018.

[149] Raytheon, "AWIPS CAVE-D2D user's manual: AWIPS II operational build 13.4.1," Raytheon Document AWP.MAN.UM.A2-OB13.4.1, Jul. 2013.

[150] K. Reda, P. Nalawade, and K. Ansah-Koi, "Graphical perception of continuous quantitative maps: The effects of spatial frequency and colormap design," in *Proc. 2018 CHI Conf. Hum. Factors Computing Syst.*, ser. CHI '18. New York, NY, USA: ACM, 2018, paper 272, pp. 1–12.

[151] T. Regier and P. Kay, "Language, thought, and color: Whorf was half right," *Trends Cogn. Sciences*, vol. 13, no. 10, pp. 439–446, Oct. 2009.

[152] P. L. Rheingans, "Task-based color scale design," in *Proc. SPIE, 28th AIPR Workshop: 3D Visualization Data Exploration Decis. Making*, W. R. Oliver, Ed., vol. 3905. SPIE, May 2000, pp. 35–43.

[153] T.-M. Rhyne, M. Tory, T. Munzner, M. Ward, C. Johnson, and D. H. Laidlaw, "Information and scientific visualization: Separate but equal or happy together at last," in *Proc. 14th IEEE Conf. Visualization, 2003*, ser. VIS '03. Washington, DC, USA: IEEE Comput. Soc., Oct. 2003, pp. 611–614.

[154] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, no. 8, pp. 913–929, Aug. 2017.

[155] P. K. Robertson and J. F. O'Callaghan, "The generation of color sequences for univariate and bivariate mapping," *IEEE Comput. Graph. Appl.*, vol. 6, no. 2, pp. 24–32, Feb. 1986.

[156] M. Rodríguez-Carmona, L. T. Sharpe, J. A. Harlow, and J. L. Barbur, "Sex-related differences in chromatic sensitivity," *Vis. Neuroscience*, vol. 25, no. 3, pp. 433–440, May 2008.

[157] B. E. Rogowitz and L. A. Treinish, "Data visualization: The end of the rainbow," *IEEE Spectrum*, vol. 35, no. 12, pp. 52–59, Dec. 1998.

[158] B. E. Rogowitz and A. D. Kalvin, "The "Which Blair Project": A quick visual method for evaluating perceptual color maps," in *Proc. IEEE Conf. Visualization, 2001*, ser. VIS '01, T. Ertl, K. I. Joy, and A. Varshney, Eds. Washington, DC, USA: IEEE Comput. Soc., 2001, pp. 183–190.

[159] B. E. Rogowitz, D. T. Ling, and W. A. Kellogg, "Task dependence, veridicality, and preattentive vision: Taking advantage of perceptually rich computer environments," in *Proc. SPIE, Hum. Vision, Vis. Process., Digit. Display III*, B. E. Rogowitz, Ed., vol. 1666. SPIE, 1992, pp. 504–513.

[160] B. E. Rogowitz and L. A. Treinish, "Using perceptual rules in interactive visualization," in *Proc. SPIE, Hum. Vision, Vis. Process., Digit. Display V*, B. E. Rogowitz and J. P. Allebach, Eds., vol. 2179. SPIE, 1994, pp. 287–295.

[161] I. T. Ruginski, A. P. Boone, L. M. Padilla, L. Liu, N. Heydari, H. S. Kramer, M. Hegarty, W. B. Thompson, D. H. House, and S. H. Creem-Regehr, "Non-expert interpretations of hurricane forecast uncertainty visualizations," *Spatial Cognition Comput.*, vol. 16, no. 2, pp. 154–172, Apr. 2016.

[162] S. Saeb, L. Lonini, A. Jayaraman, D. C. Mohr, and K. P. Kording, "The need to approximate the use-case in clinical machine learning," *GigaScience*, vol. 6, no. 5, May 2017.

[163] J. Sanyal, S. Zhang, J. Dyer, A. Mercer, P. Amburn, and R. J. Moorhead, "Noodles: A tool for visualization of numerical weather model ensemble uncertainty," *IEEE Trans. Vis. Comput. Graphics*, vol. 16, no. 6, pp. 1421–1430, Nov.-Dec. 2010.

[164] P. Schulze-Wollgast, C. Tominski, and H. Schumann, "Enhancing visual exploration by appropriate color coding," in *Proc. 13th Int. Conf. Central Europe Comput. Graph., Visualization, Comput. Vision, 2005*, ser. WSCG '05, Plzen, Czech Republic, Jan. 31–Feb. 4, 2005, pp. 203–210.

[165] M. Sedlmair, M. Meyer, and T. Munzner, "Design study methodology: Reflections from the trenches and the stacks," *IEEE Trans Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2431–2440, Dec. 2012.

[166] V. Setlur and M. C. Stone, "A linguistic approach to categorical color assignment for data visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 1, pp. 698–707, Jan. 2016.

[167] Y. Shinagawa and T. L. Kunii, "Constructing a Reeb graph automatically from cross sections," *IEEE Comput. Graph. Appl.*, vol. 11, no. 6, pp. 44–51, Nov. 1991.

[168] T.-K. Shyy, I. Azeezullah, I. Azeezullah, R. J. Stimson, and A. T. Murray, "Classification for visualizing data: Integrating multiple attributes and space for choropleth display," in *Handbook of Research Methods and Applications in Spatially Integrated Social Science*, ser. Handbooks of Research Methods and Applications, R. J. Stimson, Ed. Cheltenham, U.K.: Edward Elgar, 2014, pp. 265–286.

[169] S. Silva, B. Sousa Santos, and J. Madeira, "Using color in visualization: A survey," *Comput. Graph.*, vol. 35, no. 2, pp. 320–333, Apr. 2011.

[170] J. K. Sivillo, J. E. Ahlquist, and Z. Toth, "An ensemble forecasting primer," *Weather and Forecasting*, vol. 12, no. 4, pp. 809–818, Dec. 1997.

[171] T. A. Slocum, *Thematic Cartography and Visualization*, 1st ed. Upper Saddle River, NJ, USA: Pearson, Dec. 1998.

[172] H. S. Smallman and M. Hegarty, "Expertise, spatial ability and intuition in the use of complex visual displays," *Proc. Hum. Factors Ergonomics Soc. Annual Meeting*, vol. 51, no. 4, pp. 200–204, Oct. 2007.

[173] N. Smeulders, F. W. Campbell, and P. R. Andrews, "The role of delineation and spatial frequency in the perception of the colours of the spectrum," *Vision Res.*, vol. 34, no. 7, pp. 927–936, Apr. 1994.

[174] T. R. Stewart, P. J. Roebber, and L. F. Bosart, "The importance of the task in analyzing expert judgment," *Organizational Behav. Hum. Decis. Processes*, vol. 69, no. 3, pp. 205–219, Mar. 1997.

[175] M. Stone, *A Field Guide to Digital Color*. New York, NY, USA: A K Peters/CRC Press, Apr. 2016.

[176] M. Stone, D. A. Szafir, and V. Setlur, "An engineering model for color difference as a function of size," *Color Imag. Conf.*, vol. 2014, no. 1, pp. 253–258, Nov. 2014.

[177] Storm Prediction Center, "Short Range Ensemble Forecast (SREF) products," Storm Prediction Center, NOAA/National Weather Service, Norman, OK, USA [Online]. Available: http://www.spc.noaa.gov/exper/sref/.

[178] B. Summa, J. Tierny, and V. Pascucci, "Visualizing the uncertainty of graph-based 2D segmentation with min-path stability," *Comput. Graph. Forum*, vol. 36, no. 3, pp. 133–143, Jun. 2017.

[179] D. A. Szafir, "Modeling color difference for visualization design," *IEEE Trans. Vis. Comput. Graphics*, vol. 24, no. 1, pp. 392–401, Jan. 2018.

[180] D. A. Szafir, M. Stone, and M. Gleicher, "Adapting color difference for design," *Color Imag. Conf.*, vol. 2014, no. 1, pp. 228–233, Nov. 2014.

[181] S. Tariq and J. Woodman, "Using mixed methods in health research," *JRSM Short Reports*, vol. 4, no. 6, article 2042533313479197, pp. 1–8, May 2013.

[182] S. J. Taylor, R. Bright, G. Carbin, P. Bothwell, and R. Naden, "Using short range ensemble model data in national fire weather outlooks," in *5th Symp. Fire and Forest Meteorol.*, Orlando, FL, USA, Nov. 16–20, 2003.

[183] D. Thompson, J. Bennett, C. Seshadhri, and A. Pinar, "A provably-robust sampling method for generating colormaps of large data," in *2013 IEEE Symp. Large-Scale Data Analy. Visualization (LDAV)*, Oct. 2013, pp. 77–84.

[184] J. Thomson, E. Hetzler, A. MacEachren, M. Gahegan, and M. Pavel, "A typology for visualizing uncertainty," in *Proc. SPIE, Visualization Data Anal. 2005*, R. F. Erbacher, J. C. Roberts, M. T. Grohn, and K. Borner, Eds., vol. 5669. SPIE, 2005, pp. 146 – 157.

[185] W. R. Tobler, "Choropleth maps without class intervals?" *Geographic Anal.*, vol. 5, no. 3, pp. 262–265, Jul. 1973.

[186] C. Tominski, G. Fuchs, and H. Schumann, "Task-driven color coding," in *Proc. 12th Int. Conf. Inform. Visualisation, 2008*, E. Banissi, L. Stuart, M. Jern, G. Andrienko, F. T. Marchese, N. Memon, R. Alhajj, T. G. Wyeld, R. A. Burkhard, G. Grinstein, D. Groth, A. Ursyn, C. Maple, A. Faiola, and B. Craft, Eds. Washington, DC, USA: IEEE Comput. Soc., Jul. 2008, pp. 373–380.

[187] J. G. Trafton and R. R. Hoffman, "Computer-aided visualization in meteorology," in *Expertise Out of Context: Proc. 6th Int. Conf. Naturalistic Decision Making*, ser. Expertise: Research and Applications Series, R. R. Hoffman, Ed. New York, NY, USA: Psychology Press, 2007, p. 337–357.

[188] B. E. Trumbo, "A theory for coloring bivariate statistical maps," *Amer. Statistician*, vol. 35, no. 4, pp. 220–226, Nov. 1981.

[189] U.S. Census Bureau, "Population density by county: 2010," 2010, accessed on: Apr 26, 2019. [Online]. Available: https://www.census.gov/library/visualizations/2010/geo/population-density-county-2010.html.

[190] K. van Turnhout, A. Bennis, S. Craenmehr, R. Holwerda, M. Jacobs, R. Niels, L. Zaad, S. Hoppenbrouwers, D. Lenior, and R. Bakker, "Design patterns for mixed-method research in HCI," in *Proc. 8th Nordic Conf. Hum.-Comput. Interaction: Fun, Fast, Foundational*, ser. NordiCHI '14. New York, NY, USA: ACM, 2014, pp. 361–370.

[191] G. Verbeke and G. Molenberghs, "Inference for the marginal model," in *Linear Mixed Models for Longitudinal Data*, ser. Springer Series in Statistics. New York, NY, USA: Springer, 2000, pp. 55–76.

[192] N. Waldin, M. Waldner, M. Le Muzic, E. Gröller, D. S. Goodsell, L. Autin, A. J. Olson, and I. Viola, "Cuttlefish: Color mapping for dynamic multi-scale visualizations," *Comp. Graph. Forum*, vol. 38, no. 6, pp. 150–164, Mar. 2019.

[193] L. Wang, X. Chen, S. Li, and X. Cai, "General adaptive transfer functions design for volume rendering by using neural networks," in *Neural Inform. Processing (ICONIP '06)*, ser. Lecture Notes in Computer Science, I. King, J. Wang, L.-W. Chan, and D. Wang, Eds., vol. 4233. Berlin/Heidelberg, Germany: Springer, Oct. 2006, pp. 661–670.

[194] C. Ware, T. L. Turton, R. Bujack, F. Samsel, P. Shrivastava, and D. H. Rogers, "Measuring and modeling the feature detection threshold functions of colormaps," *IEEE Trans. Vis. Comput. Graphics*, vol. 25, no. 9, pp. 2777–2790, Sep. 2018.

[195] C. Ware, "Color sequences for univariate maps: Theory, experiments and principles," *IEEE Comput. Graph. Appl.*, vol. 8, no. 5, pp. 41–49, Sep. 1988.

[196] ——, *Information Visualization: Perception for Design*, 3rd ed. San Francisco, CA, USA: Morgan Kaufmann, 2012.

[197] C. Ware and M. D. Plumlee, "Designing a better weather display," *Inform. Visualization*, vol. 12, no. 3-4, pp. 221–239, Jul. 2013.

[198] C. Ware, T. L. Turton, F. Samsel, R. Bujack, and D. H. Rogers, "Evaluating the perceptual uniformity of color sequences for feature discrimination," in *EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3)*, K. Lawonn, N. Smit, and D. Cunningham, Eds. The Eurographics Association, 2017.

[199] D. F. Watson, *Contouring: A Guide to the Analysis and Display of Spatial Data*, 1st ed., ser. Computer Methods in the Geosciences. Oxford, U.K.: Pergamon Press, 1992.

[200] M. A. Webster and P. Kay, "Color categories and color appearance," *Cognition*, vol. 122, no. 3, pp. 375–392, Mar. 2012.

[201] P. Wessel, W. H. F. Smith, R. Scharroo, J. Luis, and F. Wobbe, "Generic Mapping Tools: Improved version released," *Eos, Trans. Amer. Geophysical Union*, vol. 94, no. 45, pp. 409–410, Nov. 2013.

[202] R. T. Whitaker, M. Mirzargar, and R. M. Kirby, "Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 2713–2722, Dec. 2013.

[203] M. Wijffelaars, Vliegen, Roel, J. J. Van Wijk, and Erik-Jan van der Linden, "Generating color palettes using intuitive parameters," *Comp. Graph. Forum*, vol. 27, no. 3, pp. 743–750, Sep. 2008.

[204] D. Williams, T. Bremer, C. Doutriaux, J. Patchett, S. Williams, G. Shipman, R. Miller, D. Pugmire, B. Smith, C. Steed, E. Bethel, H. Childs, H. Krishnan, P. Prabhat, M. Wehner, C. Silva, E. Santos, D. Koop, T. Ellqvist, J. Poco, B. Geveci, A. Chaudhary, A. Bauer, A. Pletzer, D. Kindig, G. Potter, and T. Maxwell, "Ultrascale visualization of climate data," *Computer*, vol. 46, no. 9, pp. 68–76, Sep. 2013.

[205] K. A. Wilson, P. L. Heinselman, and Z. Kang, "Exploring applications of eye tracking in operational meteorology research," *Bull. Amer. Meteorological Soc.*, vol. 97, no. 11, pp. 2019–2025, Feb. 2016.

[206] C. Witzel, "Misconceptions about colour categories," *Rev. Philos. Psychol.*, vol. 10, no. 3, p. 499–540, Sep. 2019.

[207] C. Witzel and K. R. Gegenfurtner, "Categorical sensitivity to color differences," *J. Vision*, vol. 13, no. 7, article 1, Jun. 2013.

[208] A. J. Woods, D. A. Lammers, S. A. Bryce, J. M. Omernik, R. L. Denton, M. Domeier, and J. A. Comstock, "Ecoregions of Utah," Color poster with map. 1:1,175,000. U.S. Geological Survey, Reston, VA, 2001, accessed on: Apr 26, 2019. [Online]. Available: https://www.epa.gov/eco-research/ecoregion-download-files-state-region-8.

[209] G. Xu and J. Z. Huang, "Asymptotic optimality and efficient computation of the leave-subject-out cross-validation," *Ann. Statist.*, vol. 40, no. 6, pp. 3003–3030, Dec. 2012.

[210] M. A. Yalçin, N. Elmqvist, and B. B. Bederson, "Cognitive stages in visual data exploration," in *Proc. 6th Workshop on Beyond Time Errors: Novel Eval. Methods Visualization*, ser. BELIV '16. New York, NY, USA: ACM, 2016, pp. 86–95.

[211] M.-J. Yoo, I.-K. Lee, and S. Lee, "Color sequence preserving decolorization," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 373–383, May 2015.

[212] A. Zeileis, K. Hornik, and P. Murrell, "Escaping RGBland: Selecting colors for statistical graphics," *Comput. Statist. Data Anal.*, vol. 53, no. 9, pp. 3259–3270, Jul. 2009.

[213] L. Zhou and C. D. Hansen, "A survey of colormaps in visualization," *IEEE Trans. Vis. Comput. Graphics*, vol. 22, no. 8, pp. 2051–2069, Aug. 2016.

[214] A. Zolotov, A. Dekel, N. Mandelker, D. Tweed, S. Inoue, C. DeGraf, D. Ceverino, J. R. Primack, G. Barro, and S. M. Faber, "Compaction and quenching of high-z galaxies in cosmological simulations: Blue and red nuggets," *Monthly Notices Roy. Astron. Soc.*, vol. 450, no. 3, pp. 2327–2353, Jul. 2015.