Introductory overview: Recommendations for approaching scientific visualization with large environmental datasets

Christa Kelleher[1*] and Anna Braswell[2]

* Corresponding author (ckellehe@syr.edu)

1. Department of Earth and Environmental Sciences, Syracuse University, Syracuse, NY, USA 13206
2. School of Forest Resources and Conservation, University of Florida, Gainesville, FL, USA 32653

Abstract

Scientific visualizations are the foundation for communicating results and findings to a variety of audiences. As the creation of novel and large environmental datasets has grown, this has necessitated new schemes and recommendations for creating effective visualizations. In this overview, we review the foundations of scientific visualization and considerations for visualization of large datasets within the context of the four Vs of big data (volume, variety, veracity, and velocity). Using big datasets requires making decisions as to whether to aggregate or preserve details, approaches for grouping to enable comparisons, and considering how best to show complex data in many-dimensional space. To enable more effective visualizations, we provide several considerations regarding common decisions faced during the visualization process. These recommendations are accompanied by examples applied to existing large datasets. While our recommendations are just that, they encourage intentionality and awareness of the choices faced when visualizing scientific datasets.

Keywords: scientific visualization, visual communication, plots, graphics, multidimensional, visual analytics

Research highlights

- We discuss the challenges of visualizing large environmental datasets
- We outline choices faced when creating scientific visualizations for datasets with large volume or variety
- We present approaches for approaching and improving large volume or multi-dimensional visualizations
- We provide several examples using publicly available datasets and open code

1.0 Introduction

Visualization is one of the foundational mechanisms used to communicate science. Visuals help us make sense of complex problems and interact with information (Kirsh, 2010; Liu and Stasko, 2010; Scaife and Rogers, 1996). More specifically, visuals aid in decision making (Deitrick and Edsall, 2006; Kinkeldey et al., 2014, 2017), learning (Gordin and Pea, 1995; Höffler, 2010; Höffler and Leutner, 2007; Yang et al., 2003) and science communication (Desnoyers, 2011).

In the past several decades, the creation of environmental datasets skyrocketed. This trend emerged for several reasons. In general, large datasets are more widely available because of technological advances resulting in constantly improving computing abilities, enabling analysis and modeling to be performed at higher spatial and temporal resolutions over broader spatio-temporal domains. These technological improvements contribute to growing volumes of data and shrinking costs of in situ (Alam et al., 2020; Murphy et al., 2015; Parra et al., 2018; Wickert, 2014; Wickert et al., 2019) and remote sensing technologies (Zhang et al., 2019), and new (often open source) analysis tools (Gorelick et al., 2017; Vos et al., 2019). In addition to the generation of new data, support for providing public access to datasets used in publications has also increased. The scientific community continues to show broad interest and support for reproducibility and open science (Baker, 2016; Munafò et al., 2017; Sandve et al., 2013; Stagge et al., 2019). Journals and funding agencies are precipitating these efforts through the creation and maintenance of online repositories and requirements to store data of various types. Finally, collaboration has spurred the generation of new large datasets through model intercomparison experiments (Baroni et al., 2019; Best, 2019; Krysanova et al., 2017; Maxwell et al., 2014; Smith et al., 2004), open source coding packages (DeCicco et al., 2020; Fuka, DR et al., 2018; Slater et al., 2019; Souza, 2017), new journals aimed at publishing large and unique datasets (e.g., Scientific Data, Earth System Science Data), community-based data collection (e.g., AmeriFlux, PhenoCam), and citizen science datasets (e.g., CrowdWater, Stream Tracker). All of this amounts to a diverse, sometimes overwhelming, and altogether impressive collection of data now at the fingertips of the earth, ecological, and environmental science and engineering communities.

The acceleration of data availability entails the growth of the spatial, temporal, and uncertainty dimensions of environmental data contained in publications and presentations. To borrow a buzzword, this means many publications are now making use of and visualizing ‘big data’. While there are numerous definitions of ‘big data’, the criteria for defining big data generally is associated with dataset size and complexity, as well as the need for advanced tools or technologies to interact with such datasets (Chang and Grady, 2019; Ward and Barker, 2013). While the line where data becomes ‘big’ is unclear, any dataset, by virtue of its volume (e.g., size), variety (e.g., different types of data or variables), veracity (e.g., uncertainty), and velocity (e.g., speed at which data is collected) may fall under the heading of ‘big data’ (Farley et al., 2018). These different attributes, termed ‘The Four V’s’ of big datasets (and introduced by IBM in the 2000s; IBM), can complicate visualizations and visualization goals (Yang and Huang, 2013).

Though many recommendations exist for how to best use scientific visualizations in publications and presentations (Few, 2009; Kelleher and Wagener, 2011; Rougier et al., 2014; Tufte, 2001, 1990; Weissgerber et al., 2019), the growing volume and variety of data synthesized by

researchers necessitates augmenting existing recommendations to consider the technical and aesthetic challenges associated with the visualization of large datasets. As highlighted by Liu et al. (2017), there are numerous decisions to be made, especially when visualizing high-dimensional datasets. Large datasets are cumbersome and present technical challenges to data wrangling, the transformation of raw values into a form that can be leveraged to address research objectives. Though many of general principles that were famously introduced by Edward Tufte in the 1970s and 1980s still apply to a visualization regardless of the amount of data contained within, how best to meet those recommendations as well as how to approach decision-making when creating visualizations with large datasets remains a common challenge. Colloquially, visualizations produced over the last decade include more raw data, data points, data series, and more variables. Visualizations that move beyond 2D, into 3D and higher dimensional space, are now common.

To date, there has been much attention given to computational processing, analysis, and user interfacing and interaction (Liu et al., 2017). However, there has been much less attention given to how best to effectively convey information in visual form. To address this need, we deliver a set of approaches and recommendations, paying heed to potential pitfalls, for visualizing large environmental datasets. Existing recommendations regarding scientific visualization generated over the last few decades serve as a sound basis for evaluating the effectiveness of any visualization. Our recommendations serve to augment these sound principles in the context of big data visuals. Analysis and presentation of large datasets in many ways stretch the limits of traditional recommendations for visualization; therefore, our focus is specifically on visualizing large volumes and varieties of data to assist in the analysis, synthesis, and comparison of large datasets for presentation and publications.

2.0 Challenges posed by large environmental datasets

Large environmental datasets present major challenges when it comes to developing succinct, easily interpretable, and visually aesthetic plots. These difficulties arise from two sources: technical challenges introduced by computational constraints when visualizing a large dataset, and the decision-making that is involved in how to best display and convey large datasets. These challenges are best expressed when considering the major characteristics of big data, also known as the 4 Vs. Large datasets often have large volume (many values), large variety (many types of data), and inherent (but challenging to communicate) veracity. The fourth V, velocity, we describe in more detail in a later section; here, we interpret this fourth V to refer to the dynamic nature of many large datasets, that may often be best conveyed using animated or interactive approaches. However, the majority of our overview focuses on static visualizations, as these are still the major currency of visual communication. Below, we outline the major challenges introduced by three of the four Vs when it comes to approaching data visualization with large datasets that may fall into one or multiple of these categories.

2.1 Challenge 1: Large datasets are (unsurprisingly) big

The sheer volume of large environmental datasets introduces several considerations for visualizations, beyond posing technical challenges. While there are many examples of voluminous visualizations, there is a tension between ensuring a visualization shows broad patterns and the distribution of the data while at the same time allowing a reader to identify all of the data or the most important data. Too often, we synthesize and remove key pieces of

information through aggregation. Yet this may also be a necessary step to pursue scientific questions that span spatio-temporal scales.

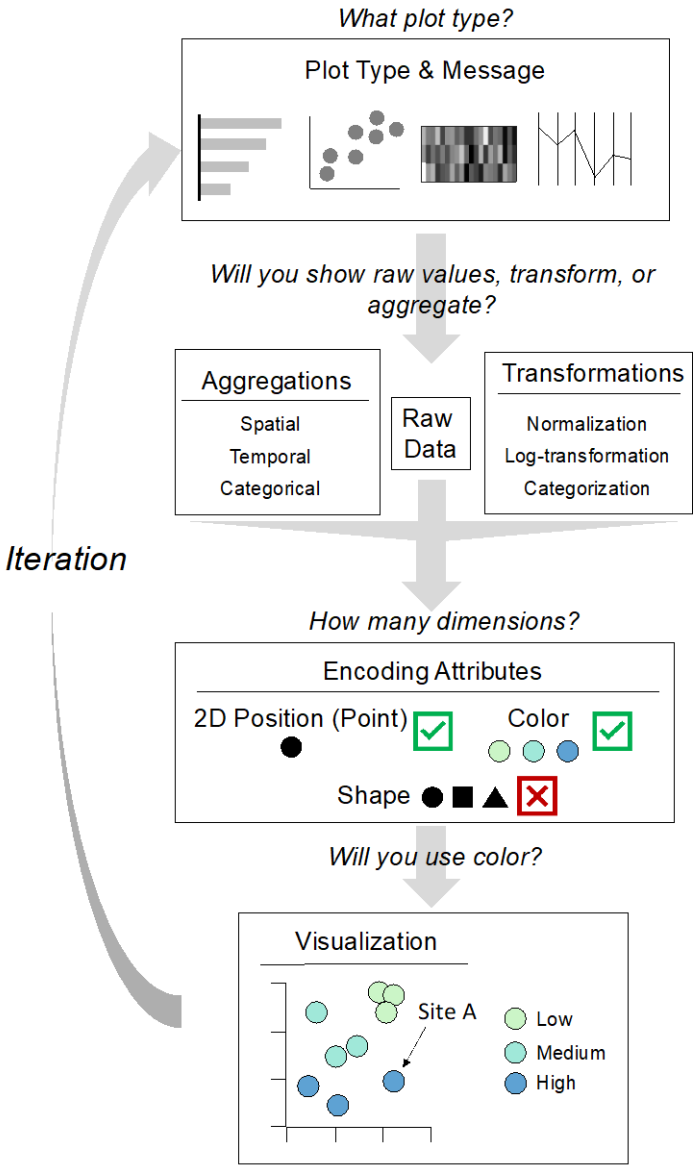

Figure 1: *Key questions faced when creating a visualization. Within the 'how many dimensions?' box, the green check mark corresponds to an affirmative decision to use certain encoding attributes (e.g., 2d position, color), while a red 'x' box corresponds to a negative decision to not use shape as an encoding attribute.*

2.2 Challenge 2: Large datasets often contain variety

Variety in large datasets refers to the inclusion of different types of data, categories of data, or different variables or descriptors. A common challenge in large datasets with exceptional variety is how to best display multi-dimensional data to show broad relationships across many variables or descriptors. Likewise, plots that highlight variety often deal with multiple categories and comparisons. Complexity should not be avoided when creating such visualizations, though it

can be challenging to walk the line between clean visuals and overcomplicated visuals when displaying datasets with large variety.

2.3 Challenge 3: Large datasets are frequently used to communicate veracity

Veracity is often interpreted as data uncertainty; here we broadly interpret this term to refer to all types of uncertainty, variability, and comparisons between values to determine veracity. Plots concerned with veracity may be used to show aggregated metrics such as objective functions (Jackson et al., 2019), uncertainty, error, probabilities, or confidence. These approaches often rely on comparison to a baseline (e.g., modeled uncertainty applied to a timeseries plot; error bars applied to bar chart or dot plot) or feature error as a derived value (e.g., boxplot or violin plot of errors; bar chart of difference from 'true' or zero). Communication of veracity can be especially challenging (Spiegelhalter et al., 2011), as emphasized by the misinterpretation of common graphics used to communicate uncertainty, such as the Hurricane 'cone of uncertainty' (Boone et al., 2018).

3.0 Decision-making for visualizing large datasets

The term visualization can be ambiguous. It may refer to a tool being used to create or generate a visualization, to the process of creating a visualization, to the analysis of data, or to a generated visual (Parsons and Sedig, 2014). In this article, we use the term scientific visualization to refer to visual representations of datasets.

In the literature, two common types of visualizations exist: glyphs and plots. Glyphs (e.g., multidimensional icons) combine multiple encoding attributes into symbols or graphical representations (e.g., Chernoff Faces, Chernoff, 1973, or infographics). In contrast, plots display datasets using coordinate systems. We focus specifically on the creation of scientific visualizations as plots, though note that many of our recommendations also apply to glyphs.

Generating a visualization from a large dataset introduces both technical challenges as well as several (often somewhat subjective) decisions that must be made to generate a visual display. When considering how best to approach visualization of a large dataset, there are four central questions that must be answered when creating a visualization (Figure 1):

- Plot type (or the decision to use multiple plots): Which visualization(s) will you use to display your data?
- Raw values or aggregation: Is aggregation needed or should viewers see raw values?
- Dimensionality: How many dimensions do you need to display?
- Color: Are you using color, and are you using color wisely?

In the sections that follow, we present common challenges or pitfalls when using traditional visualization techniques, and considerations and recommendations for how to re-envision these plots in the context of these four key decisions. We also envision these decisions in Figure 1 as a series of steppingstones to arriving at a final plot. Amongst these recommendations, we qualify that this overview is by no means represents an exhaustive list of all considerations when plotting datasets, whether small or large, but serves as a starting point for thinking about visualizations in the context of large datasets. Importantly, these recommendations are not

intended to be applied in isolation; instead, they are complimentary ideas that should be used to identify how visualizations of large datasets may be approached or improved.

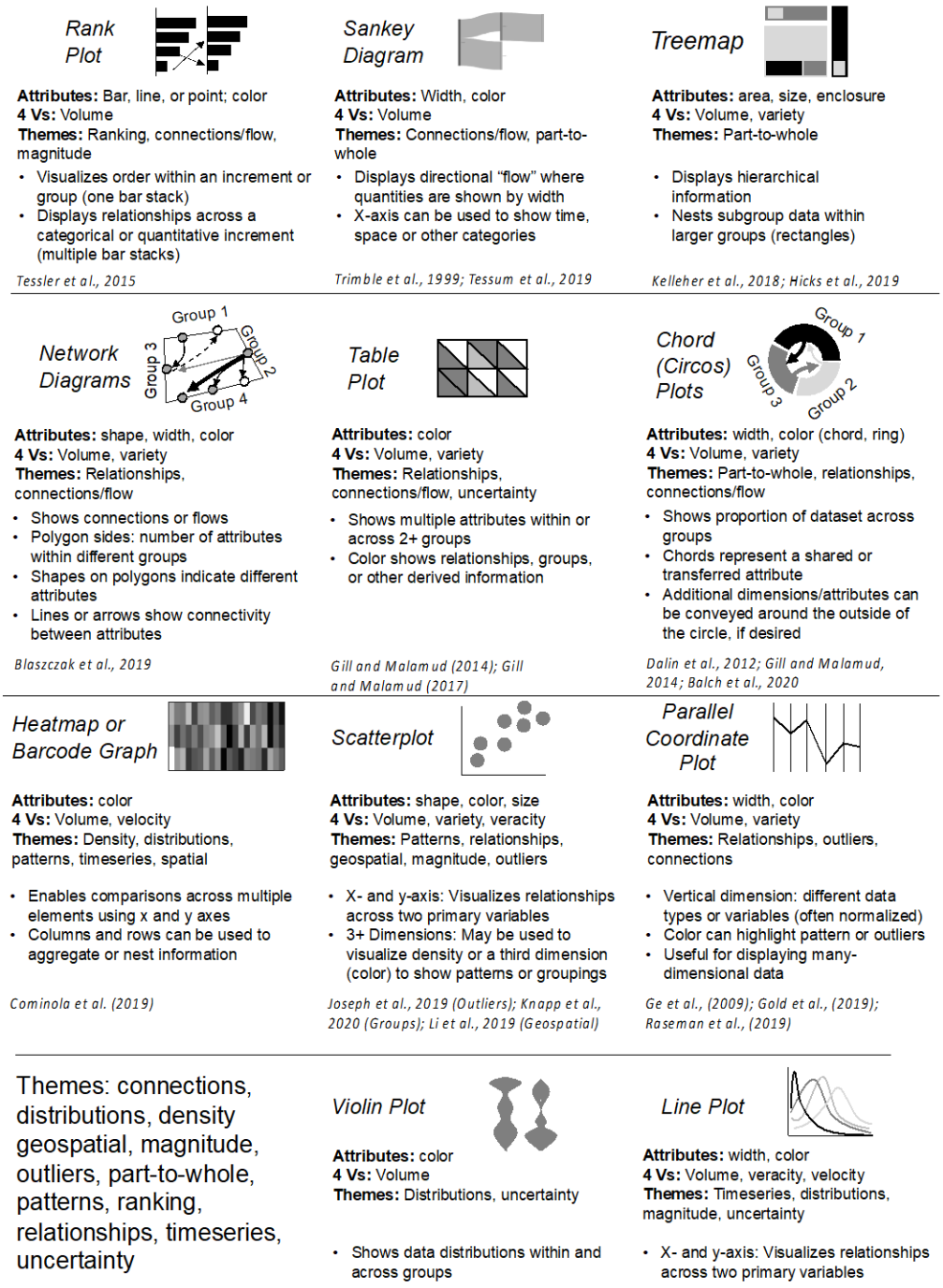

Figure 2: *Examples of common multidimensional visualizations, with associated attributes that can be used to display additional dimensions, which of the 4 Vs the plotting supports, and key themes that can be communicated by each visualization. Below each visualization, we also*

summarize pertinent details, and point to citations from the literature that make strong use of these visualizations. Literature examples include: Balch et al. (2020), Blaszczak et al., (2019), Cominola et al. (2019), Dalin et al. (2012), Ge et al. (2009), Gill and Malamud (2014, 2017), Gold et al. (2019), Hicks et al. (2019), Joseph et al. (2019), Kelleher et al. (2018), Knapp et al. (2020), Li et al., (2019), Raseman et al. (2019), Tessler et al. (2015), Tessum et al. (2019), Trimble (1999).

Using these decisions as a guide, we include examples created from existing large, environmental datasets. These include the GAGESII dataset (Falcone, 2011; Falcone et al., 2010), land cover change data for Alaska, from the National Land Cover Dataset (Homer et al., 2015; National Land Cover Dataset), and the Continuously Updated Digital Elevation Model dataset (Cooperative Institute for Research in Environmental Sciences (CIRES) at the University of Colorado, Boulder, 2014). All visualizations were created within RStudio (v. 3.6.3) and code is available on Github.

3.1 Choosing a plot type, encoding attributes, and overall visualization approach

At a basic level, a visualization is composed of encoding attributes, scales, and coordinate systems (Wickham, 2010). Scientific visualizations rely on the selection of encoding attributes, also known as visual encodings or visual marks. These attributes are used to convey quantitative and qualitative information within the context of a visualization. As summarized by Few (2009), attributes include those associated with form (e.g., length, width, orientation, size, shape, curvature, enclosure, and blur), color (hue, value, saturation, transparency), spatial position (2-d position, spatial grouping, or density) and motion (direction, path). Scales are used to encode information using attributes associated with form, size, and color. They may be quantitative (e.g., color, size) or categorical (color, shape). Coordinate systems provide a means of assessing spatial position. Coordinate systems may be cartesian, logarithmic (on one or multiple axes), polar (r, θ), or multidimensional.

These building blocks of scientific figures ultimately come together in a visualization. While it is sometimes helpful to think about these individual pieces, perhaps more important is to consider the overall plot type, as this is one of the most crucial choices faced in the visualization of a large, multi-dimensional dataset. This decision is an inherently subjective choice but can benefit from keeping in mind the overarching plot goal or message (what is the main message you wish to convey?). This choice will ultimately determine how many dimensions you seek to encapsulate within your plot, which then will help to identify what plot types are at your disposal.

Regardless of the big picture selection of a plot type, the details associated with the plot building blocks are equally important. Within the open source programming language R, these components are often described and implemented as the 'grammar of graphics' (Wilkinson et al., 2005; Wickham, 2010). Though not all programming languages or tools implement a graphical grammar, the grammar introduced by Wilkinson et al. (2005) and refined by Wickham (2010) is helpful for identifying the choices faced and the refinements that can be used in the process of generating a graphic. Such details are crucial for refining visualizations.

Consideration 1: Match your plot type and encoding attributes to your key message

Visualizations are built upon the selection of encoding attributes and the choices made in the selection of components of a figure. While we are all aware of the components that are used to build a visualization, the selection of these components is a key development step in creating a visualization. Visualizations of any type should begin with identifying a key message (aimed to be conveyed with the visualization). From this message, we can select a particular plot type, scale, and coordinate system, built on the selection of encoding attributes to display quantitative information or qualitative groupings (Kelleher and Wagener, 2011). Writing out a key message or the visualization take-away can be a good place to begin, especially when parsing components of a complex visualization. For example, does the reader need to compare groups or categories to determine the key message? Likewise, revisiting these choices during revision of a given visualization can help to clarify the message conveyed by a particular plot.

Though visualizations are unique to the dataset and creator, there are several common key messages that visualizations seek to highlight. These include: connections between values (i.e., flow), the distribution of a dataset, data density (including spatial density), geospatial position, magnitude, outliers, part-to-whole (i.e., hierarchical or layered datasets), patterns, rankings, relationships (i.e., correlation), timeseries, and uncertainty. These common themes may represent a starting point for designing a plot to convey a key message. In Figure 2, we show several cartoon examples of multidimensional visualizations and highlight common key messages (or themes) that may be conveyed using each of these plots. We do also highlight that many figures may be composed of multiple plots, aimed at showing groupings relative to the larger dataset or other groups (also called facets, Wickham, 2010) or groupings relative to subsetted data groups.

We encourage visual creators to remember that there are a multitude of different approaches and types of plots that can be used to visualize data. For inspiration, we direct you to several impressive summaries including the Visual Vocabulary (Smith et al., 2016), The Data Visualisation Catalogue (Ribecca, 2020), and The Graphic Continuum (Schwabish and Ribecca, 2014). In particular, The Graphic Continuum highlights six key plot groups: distributions, time, comparing categories, geospatial, part-to-whole, and relationships.

Consideration 2: Pay heed to overall composition as you finetune your visualization

As discussed above, visualizations inherently consist of many different components that must work together to tell a story. How best to arrange these components such that they most clearly articulate a key message can be thought of as composition. The composition of a plot includes selection of a color palette, the use of annotation through legends, direct labeling, and other words included on the visualization including the caption, and the choice of plot and how the plot is designed.

Visualizations often include annotations – text or enclosures used to highlight or explain features of the visualization. Beyond the caption, annotations are a way to use text or other visual cues to direct the eye of the reader and to aid interpretation. Annotations also encompass the figure legend that is used to describe a qualitative and/or quantitative scale. Ensuring a strong composition requires attention to annotations, which enhance a viewer's understanding of a given visualization.

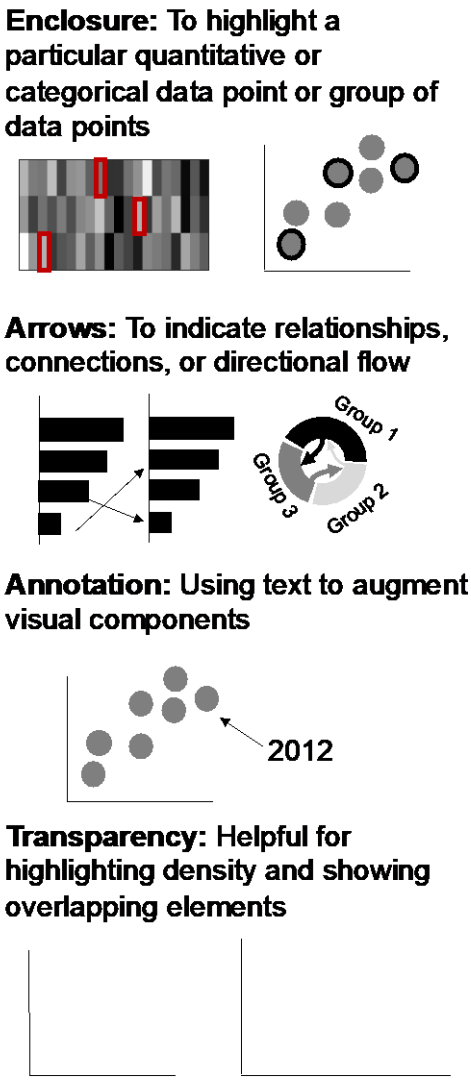

Figure 3: *Aesthetics (beyond color) that can be used in most or all plots.*

Composition also includes the creation of what we will call 'mega-figures', composed of many subplots or facets. Though a single visualization may feature one plot, visualizations (particularly those of complex datasets) may also include a composite of many small multiples (Tufte, 1990) also known as subplots (Matlab) or facets (R). The combination of small multiples may be used to provide additional detail regarding a component of a dataset and can be especially useful for parsing and displaying subsets of a large dataset. In the literature, small multiples are commonly used to parse a single dataset often using a repeated coordinate system, encoding attributes, and scale but varying the data displayed, enabling visualization of high-dimensional data. However, these subplots or small multiples can also be superimposed on larger plots, to display different types of data (e.g., spatial versus temporal versus categorical) or to visualize data at different scales. In composing a plot, we encourage readers to think beyond generating a single plot to producing an integrative visualization that may be composed of many plots and plotting elements.

Consideration 3: Give thought to how you can simplify and clarify for your key message

Across large environmental datasets that may exhibit large volume or variety, there are several common approaches to simplifying such visualizations that may clarify the overarching message of a particular plot:

- Aggregating large volumes into simple distributions or statistics, multiple values into indices, single points into footprints
- Combining multiple types of data into multidimensional plots
- Highlighting outliers, certain groups of data, trends, a single observation

When designing a plot, it is important to consider these options for creating a clear and concise visualization. Finally, ensuring that key message is clear and perceivable by others is one the most important considerations when creating a visualization. Generating a useful and effective visualization not only requires that you have in mind what the goal of your plot is, and how you want to use encoding attributes (color, shape, width/size, orientation; Few, 2009; Kelleher and Wagener, 2011) to convey key messages, but also that this key message is perceivable by others.

Consideration 4: Aesthetics are important – think beyond just color

Visualizations are as much science as art. Often, we associate color with aesthetics (so much so, that we have dedicated an entire section of this overview to the discussion of color). However, aesthetics of visualizations go far beyond color alone. During the visualization process, give thought and attention to the details – annotations, font size, font type, legend placement, axis widths, tick mark spacing. For publication quality graphics, many journals may have recommendations for particular font sizes or types to use and may specify the location (inside the axis or outside the axis) for tick marks. Helvetica and Arial are often preferred fonts when creating visualizations.

In addition, there are several details that can be used to improve the overall interpretation of your visualization (Figure 3). These include enclosure (e.g., to highlight data points that meet a certain p-value), arrows (e.g., to show directional connections), annotation (e.g., to explain or label an unusual or exceptional data value or sets of data values), and transparency (e.g., when elements overlap). Overall, attention to these small details can be used to improve the overall aesthetics of your visualization.

3.2 Preserving individual values versus transformation or aggregation

Often, the analysis of a large dataset begins with visualization of raw, untransformed, unaggregated data. On the path to presentation and publications, this data is often repackaged in different ways within visualizations. This re-packaging often includes the use of transformations and the use of aggregations.

Transformations, depending on the visualization tool, may be applied to the data, to the scale, or to the coordinate system (Wickham, 2010). When applying transformations to scales or coordinate systems, clarity and communication is key. This requires attention to and use of tick marks, legends, and even the figure caption. Visualizations may also rely on statistical transformations that aggregate or alter data in some way. This includes data binning (as is done when plotting distributions or density), data jittering, data smoothing, or categorial or other groupings applied to datasets. While transformations and aggregations are a necessary part of visualizing large datasets, they can also alter the perception of the data and the visualization.

One existing tension in the visualization of large datasets is whether or not it is important to show all values in a given visualization, or whether these values should be aggregated. This decision depends on a few factors, particularly the size of the dataset (Consideration 1) and the approach to aggregation (Consideration 2), but should also be viewed in the context of which approach produces a clear visualization that enables viewers to perceive the overarching plot message.

Several plot types, including scatter plots, spatial scatter plots, and parallel coordinate plots, are used to enable readers to quantitatively perceive all values within a dataset. Humans have a remarkable ability to lump or categorize visual information, so often preserving information while highlighting the main or macro pattern is key for effective visualization (Tufte, 1990). As stated by Tufte (1990), "Clutter and confusion are failures of design, not attributes of information".

Yet, displaying all raw values may overwhelm or obscure trends, variation, or groups. When it comes to large datasets, showing all values may not be possible for high volume datasets (e.g., a long timeseries or for many raw values). For these situations, aggregation is often necessary. However, it is important to keep in mind that aggregation can subsume extensive variability in raw values (which can challenge interpretation of veracity). In this section, we highlight two considerations when making the decision regarding whether to aggregate or preserve raw data.

Consideration 1: Can raw values be distinguished?

Preserving the visualization of all points is particularly challenging for large datasets as the information contained in the plot may become obscured (Figure 4; Figure 5). For instance, plotting many sites or locations, or plotting dense datasets, can produce overlapping values that may be poorly visualized. To combat this, the most commonly used strategy is to plot the shape outline with an empty interior (Figure 4a; Figure 5a). While this strategy may be effective for intermediately sized data, the intended outcome of ensuring that all values can be visually interpreted can be difficult as the number of values to be visualized increases. As an alternative, there are several ways to preserve visibility of all data points in figures displaying large datasets. Plotters can vary the size of attributes, transparency (e.g., Kelleher and Wagener, 2011), or create inset figures where individual points can be distinguished from one another. However, transparency may not be a solution for displaying density across large volumes of data (Figure 4b). Plotting that does not enable the viewer to distinguish all points or values should be avoided, as this approach may obscure outliers, density, or the interpretation of overarching relationships within a dataset.

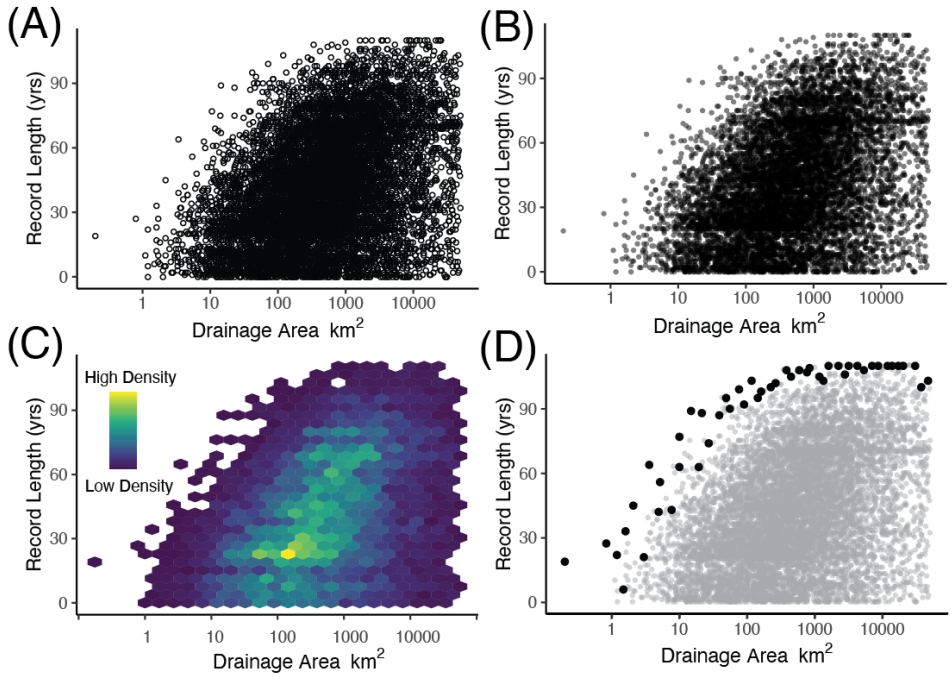

Figure 4: *Aggregating record lengths (years of data) and drainage areas for discharge measurement locations within the GAGESII dataset (Falcone, 2011; Falcone et al., 2010). For this large dataset that includes more than 9000 sites, using (a) using unfilled points obscures any perception of data density. Though (b) using transparency still conveys some aspects of data density, a plot that conveys (c) a continuous bivariate distribution may be a better alternative to highlight higher and lower frequency combinations of drainage area and record length. As an alternative to showing the density pattern in subplot (c), subplot (d) highlights the outliers – the sites with the longest record lengths across different drainage areas. In this particular example, (d) highlights the disparity between record lengths in watersheds with very small versus large drainage areas.*

Preserving raw values (encoded as points or lines) can be especially useful when the goal is to highlight outliers or a particular subset of observations within a particular dataset. From a data science perspective, outliers are often an important source of information. Using a strong color contrast, or changing size or shape, enables perception of this group or set of outliers as compared to the rest of the dataset (Figure 4d). Such an approach can also be used with subplots or facets to highlight multiple sub-groups and to emphasize how they relate to the larger dataset. As we show in Figure 4c, aggregation can be useful for conveying where values are concentrated (such as the conclusion from Figure 4c that most streamflow records occur in moderately sized rivers with record lengths of between 30 and 60 years). However, as shown in Figure 4d, when this information is aggregated, the individual data points are lost; instead, our plotting of outliers shows how streamflow record length varies with watershed drainage area, aiding in the conclusion that larger watersheds typically have longer record lengths.

It can be especially challenging to visualize raw values when all data points are plotted along a single axis (e.g., boxplots or violin plots, parallel coordinate plots). Jittering data values, which creates slight offsets, can be helpful when points are used as an encoding attribute. When lines

are used as an encoding attribute (e.g., parallel coordinate plots), de-cluttering strategies may include use of transparency or bundling (Raseman et al., 2019).

Consideration 2: Aggregation to emphasize patterns

Enabling perception of all values may not be possible for visualization of large datasets. In this case, aggregation may be used to summarize values. Aggregation can enter the visualization process either after a plot type is selected, prior to selecting a plot type, or as part of the iteration when selecting a plot type to use. Approaches to data aggregation will depend on the type of data you are using and in what way you seek to aggregate. When aggregating a dataset for visualizations, you must first decide how you would like your output data to be organized. This requires considering how you will group your values: quantitatively or categorically. Second, you must decide what statistic you will use to transform many values to one value within your groups.

Aggregation may occur during plot creation (such as with a density-based plot, Figure 4c) but often happens prior to plot generation, with the goal of condensing data to be visualized. In these contexts, aggregation may be used to address technical challenges encountered when trying to plot a large volume dataset, and/or may be an approach to simplify the plot itself and the overarching message (such as when summarizing spatio-temporal datasets). In these cases, the choice of a statistic for aggregation will depend on the overall plot message. For instance, frequency is used to highlight density. Statistics available for aggregation include but are not limited to the frequency or count, mean, median, maximum, minimum, and variance of a dataset. During this process, decisions regarding how to group data are especially important. Sometimes these groups may be evident within the dataset (such as countries, cities, watersheds, or species), while others may require choices. In these instances, we encourage transparency to describe such choices and justification in the figure caption.

When working with spatially distributed data, additional decisions are required during aggregation. Aggregation requires the selection of a window or "footprint" size and shape (as we chose to do in Figure 5b). It may be easy to assume a certain footprint size (e.g., municipalities, counties, or other geographic boundaries) or more challenging in some cases (e.g., geolocated reports of flooding, area of hurricane cover). We note that the subject of how best to represent and visualize a footprint is also an interesting and open-ended question. These selections can bias the interpretations gathered from a particular dataset and should be clearly indicated in the figure caption. Similar decisions are encountered when using non-spatial, bivariate plots aimed at highlighting density as a third dimension. Plots that aim to highlight density have commonly used transparency (e.g., Kelleher and Wagener, 2011; Raseman et al., 2019), but this approach falls short for very large and/or very dense datasets (Figure 4c). One option that can be used to visualize density in large datasets is the use of color to indicate density (Figure 4c; Figure 5c), or to show density groups that highlight the fraction of a dataset across the figure space (see example from Harrison, 2017).

Consideration 3: When possible, show raw values AND aggregated information

One of the most common ways to visually contextualize or compare large datasets to use a plot that shows distributions. These types of plots represent succinct ways to summarize large volume datasets while preserving the dataset statistical properties. Of the many plot types that

exist for showing distributions, two of the most common are bar plots and box plots (Figure 6a; Krzywinski and Altman, 2014; McGill et al., 1978; Tukey, 1977). However, there is growing evidence that shows both of these plot types can be misleading (Matejka and Fitzmaurice, 2017; Weissgerber et al., 2015). This confusion arises because different dataset distributions may contain similar or even equivalent summary statistics. Given bar plots and box plots primarily show summary statistics – medians, interquartile ranges, and 95th and 5th percentiles for box plots, and median or mean plus standard error or confidence intervals for bar plots – two similar plots may incorrectly suggest that dataset distributions are equivalent. This problem is even more pronounced with bar plots that use a bar to represent the mean or median of the data, and lines to indicate standard error or confidence intervals (Weissgerber et al., 2019).

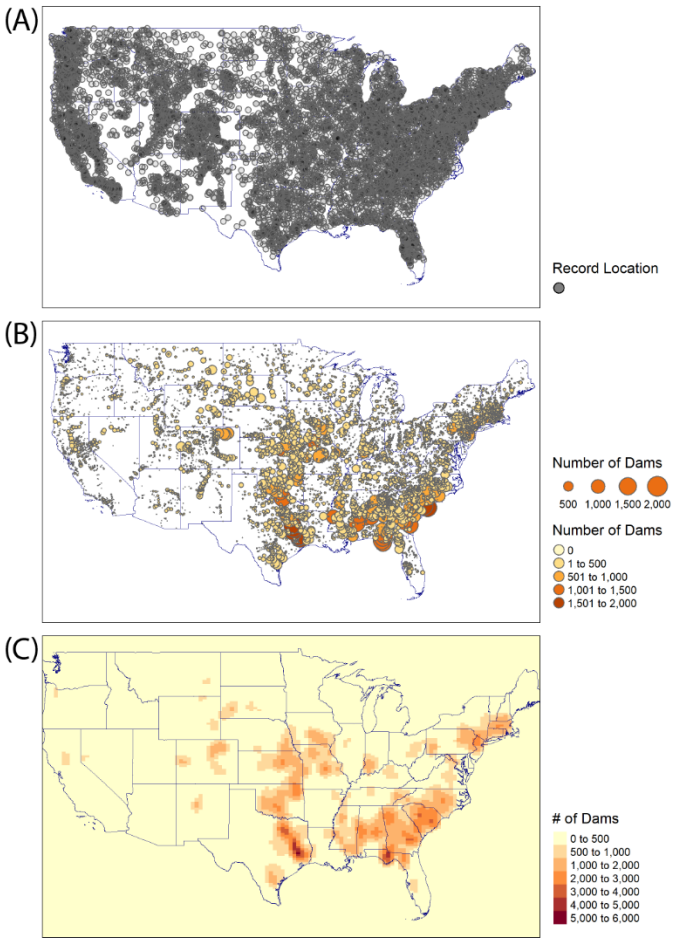

Figure 5: *In geospatial visualizations, the ability to discern the spatial distribution of data is important for visualization. This figure uses the GAGES II dam location dataset (Falcone 2010, 2011) to examine at the spatial distribution of dams across the U.S. using three different visualizations. In Panel A, all records of dams are illustrated with transparent points, this produces a cluttered figure with little available information. Panel B aggregates records of dams within a certain distance to bubbles of varying size and color. Although more detail is available in this figure, there are still areas (Southeast U.S.) that are cluttered, and it is hard to distinguish separate bubbles. Using kernel density estimation to create a heat map, data was aggregated to raster grid cells in Panel C. Although this map shows the "hottest spots" for dams in the most*

easily interpretable way, it does lose information on the location of dams in less dam dense areas (i.e., Colorado River).

Three alternative plots for large datasets that preserve distribution shape are density plots, violin plots, and a new combined approach termed 'raincloud plots' (Allen, 2018; Allen et al., 2019). Density plots can be strong alternatives to boxplots when the goal of a visualization is to show volume but not variety (e.g., multiple groups). Overlaid density plots can summarize density for a small number of groups; however, it may become hard to distinguish between groups for more than three to four categories (Wickham, 2010), depending on the degree of difference between the distributions. As the number of series comparisons grows, subplots should be used to break out individual groups.

One alternative to density plots for comparing multiple groupings with large volumes are violin plots, which are essentially mirrored density plots (Figure 6b). The myriad of violin plot iterations also enables encoding summary statistics alongside the distribution, to preserve both types of information. However, there is an argument to be made that violin plots may include redundant information through mirroring (Allen et al., 2019). Raincloud plots are a different type of approach that combine visualization of the distribution showing an aggregated distribution and individual data values (Allen et al., 2019). While these three approaches represent endmembers in the visualization of distributions, many other iterations of these types of plots exist. For instance, one iteration is to combine a barplot and violin plot (Figure 6c; Hintze and Nelson, 1998), enabling interpretation and comparison of summary statistics and overall distribution. In addition to the variant shown in Figure 6c, other variants include beeswarm plots (Eklund, 2016, 2015) – a re-envisioning of the dot plot (Wilkinson, 1999), and beanplots (Kampstra, 2014, 2008).

One question that may arise when considering plotting distributions: if the goal of a plot is to highlight the distribution of the data, should we just be plotting the raw data? The answer here is an emphatic "no". Estimating distributions and statistics from raw data is notoriously challenging (Bobko and Karren, 1979; Spence et al., 2016).

3.3 Decision-making in the context of dimensionality

Large datasets are often high-dimensional, either in terms of the variables they contain, or in terms of how those variables are categorically or quantitatively grouped. Therefore, selecting the number of dimensions to display within a given plot is often challenging. With so many potential encoding attributes to add – spatial location, shape, width/size, and color, to name a few – it is easy to overcomplicate. At the same time, as the volume and variety of data encapsulated within scientific visualizations grows, plot complexity (in terms of dimensionality, volume of data encoded, and composition) is certainly growing. Though simplicity should still be the ultimate goal of any visualization, this does not have to be in conflict with employing a visualization that exceeds three dimensions, that shows an exceptional volume of data, or that combines multiple subplots into a single visualization.

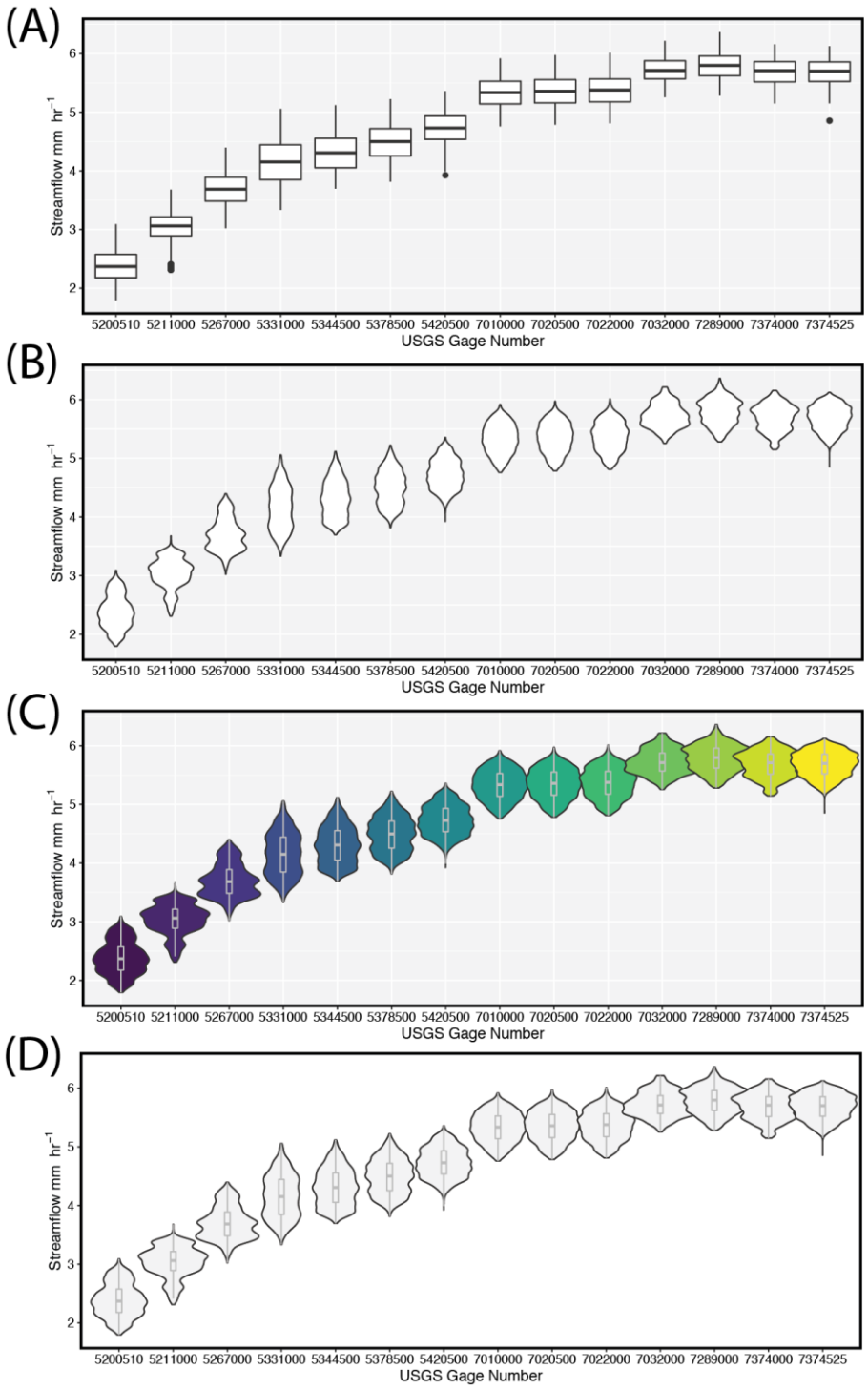

Figure 6: *Visualizing distributions enables comparisons amongst groups, such as USGS streamflow observations. Here, we visualize distributions of daily streamflow (in mm hr^{-1}; Oct 1 2008 through Sept 30 2018) across 14 United States Geological Survey stream gages stretching from Minnesota (drainage area of 1579 km^2) to Louisiana (drainage area of 2,926,687 km^2). These distributions are shown as boxplots (Panel A), violin plots (Panel B), and as combined violin and boxplots (Panel C), adding more information about the distribution moving from A to C.*

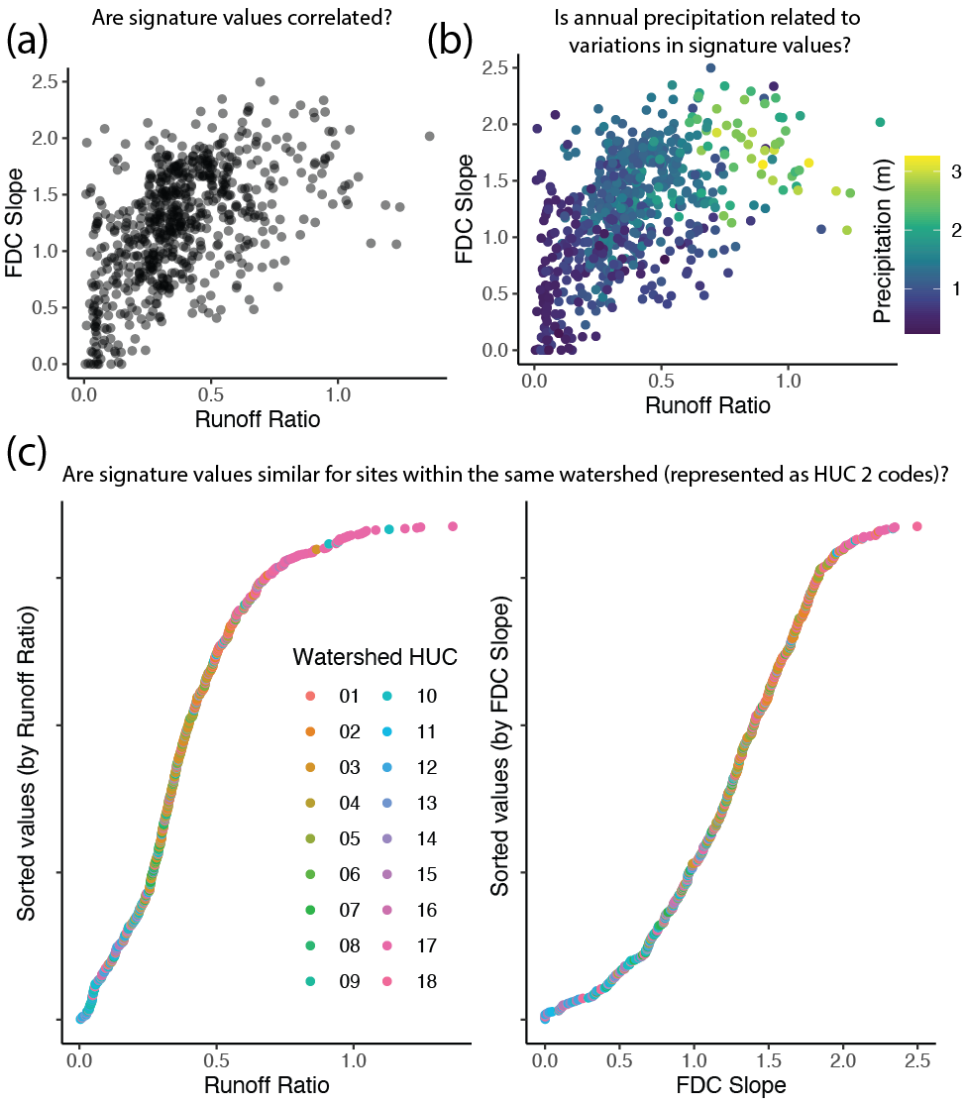

Figure 7: *Organization of and relationships between two hydrologic signatures, runoff ratio and flow duration curve (FDC) slope from the CAMELS database (Addor et al., 2017a, 2017b). These are shown (a) plotted against one another, (b) with precipitation, and (c) sorted by value and colored by HUC2 watershed. Each point represents a watershed. We note the major questions being asked and answered above each subplot.*

Consideration 1: Balance the number of dimensions you show with overall plot simplicity

Decision-making surrounding the choice of a plot, the number of dimensions to display, and a key plot message are inherently linked. Giving thought to how these pieces work together from an early stage is therefore important to creating an effective visualization. When making decisions regarding the number of dimensions you seek to display, it is important to remember that encoding attributes inherently limit us to just a few dimensions – two continuous variables for positions on a bivariate plot, one continuous or categorical dimension for color, and/or size, and one categorical dimension for shape. Therefore, many plot types support displaying anywhere between two and five dimensions, though some plots, such as the parallel coordinate plot and the rose plot, can display many more dimensions. However, as the old adage goes, "just

because you can doesn't mean you should". Additionally, the goal with any plot should be to avoid redundancy (as shown in Figure 6c – color is redundant with labeling on the x-axis).

In Figure 7 and Figure 8 we explore the Catchment Attributes and Meteorology for Large-Sample Studies (CAMELS; Addor et al., 2017a; Addor et al., 2017b) dataset using a series of figures moving from two-dimensional (Figure 7a, c) and three-dimensional (Figure 7b) visualizations to higher-dimensional visualizations such as parallel coordinate plots (Figure 8). In Figures 7 and 8 below, adding higher dimensions and showing variety across hydrological signatures (Figure 8) presents a clearer picture of how watershed behavior is organized across the US (Figure 8a) and to what extent behavior is similar within a larger basin (Figure 8c). However, as additional elements are added to the plot, such as shown in Figure 8b, it becomes difficult to extract useful patterns and to compare across multiple dimensions.

Consideration 2: Is the number of groups or series in a single visualization manageable and discernable?

While our plotting is often limited by dimensions, it is not inherently limited by quantitative or categorical groupings. These groupings are regularly used when visualizing large datasets to emphasize comparisons between quantitative or qualitative groups. Comparison is at the heart of understanding trends or differences in data; visualization must make comparison between groups easy to interpret (Tufte et al., 1990). Grouping is an approach for reducing dimensionality that enables assessment of similarities and differences across dataset subsets. When plotting large datasets, we often use grouping – with colors, symbols, or sometimes both – to show organization within a complex, multi-dimensional, and/or large volume visualizations. However, when the focus is on comparison of these different groups, it can be easy to overwhelm when the number of groups shown plotted concurrently – not side-by-side - begins to exceed three (Wickham, 2010). In these types of plots aimed at showing many groups, all values may be plotted within the same visualization, or may be separated into subplots or insets. The latter can be an effective way to highlight a subset of the data in the context of the broader dataset. In Figure 9, we show examples of hydrologic simulations produced from four different models and compared to observed streamflow for one watershed. When plotted together, the timeseries are hard to distinguish from one another (Figure 9a), even on a logarithmically transformed axis (Figure 9b). Separating each of these comparisons into subplots more clearly illuminates the periods when simulated streamflow is in agreement with observed values (Figure 9c).

In line with ensuring the number of groups or series to be compared are interpretable is giving thought to how these groups are organized within a visualization. When you must specify the order of such groups (e.g., in a heatmap, a parallel coordinate plot, or a distribution-based plot), the choice of how to order components of your plot matters. This ordering should be done intuitively – such as from small drainage areas to large when comparing watersheds (e.g., Figure 6), by magnitudes (in rank plots, Figure 7c), or potentially by mean values (in heat maps).

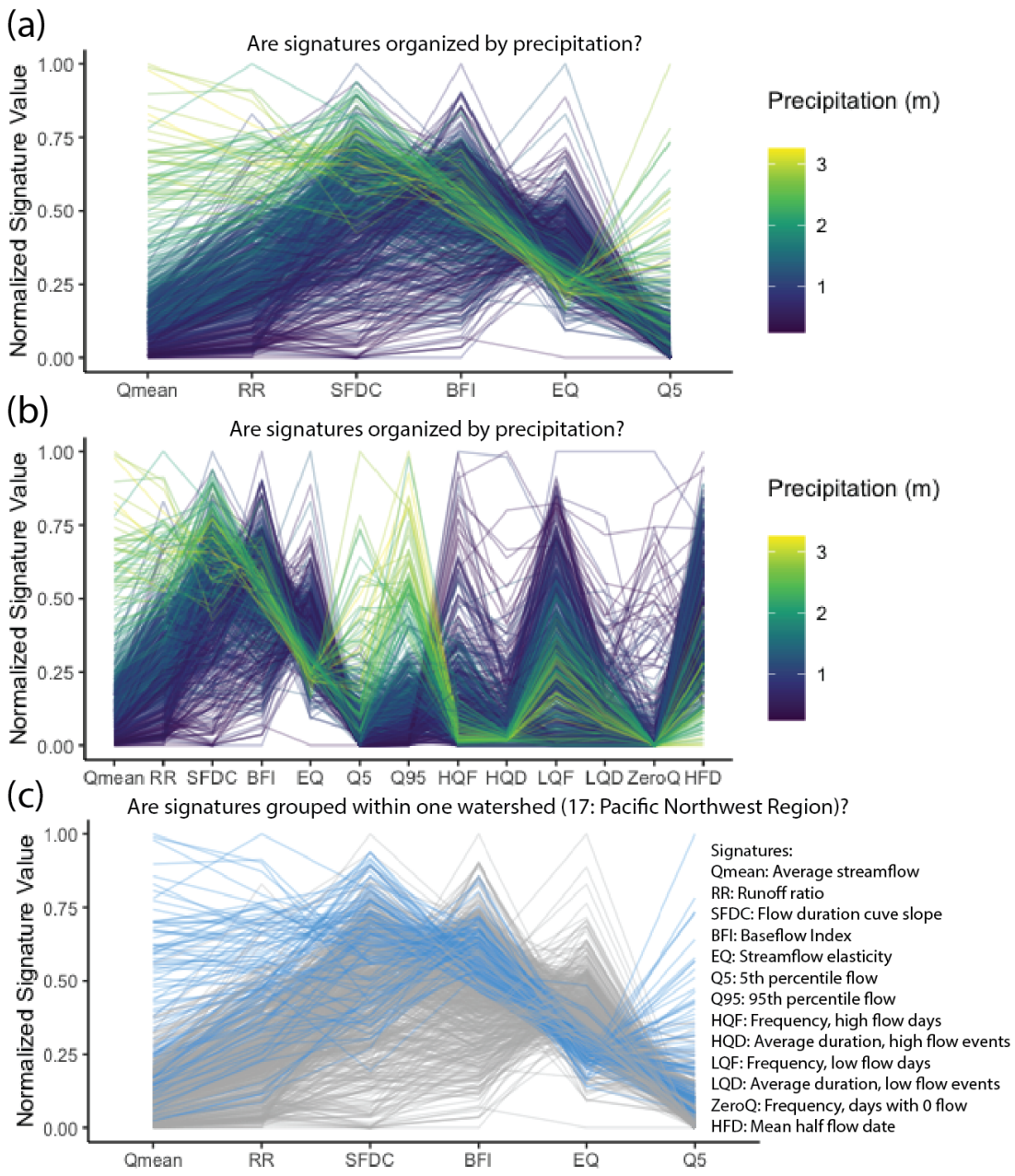

Figure 8: *Visualization of multiple hydrologic signatures from the CAMELS database (Addor et al., 2017a, 2017b) as parallel coordinate plots. Each line represents a watershed. These are shown as a function of precipitation for (a) six hydrologic signatures and (b) 13 hydrologic signatures. Though color can be used to highlight patterns (a, b), it's also useful for highlighting groups (c), such as the signature values from watersheds within the Pacific Northwest Region, shown in light blue.*

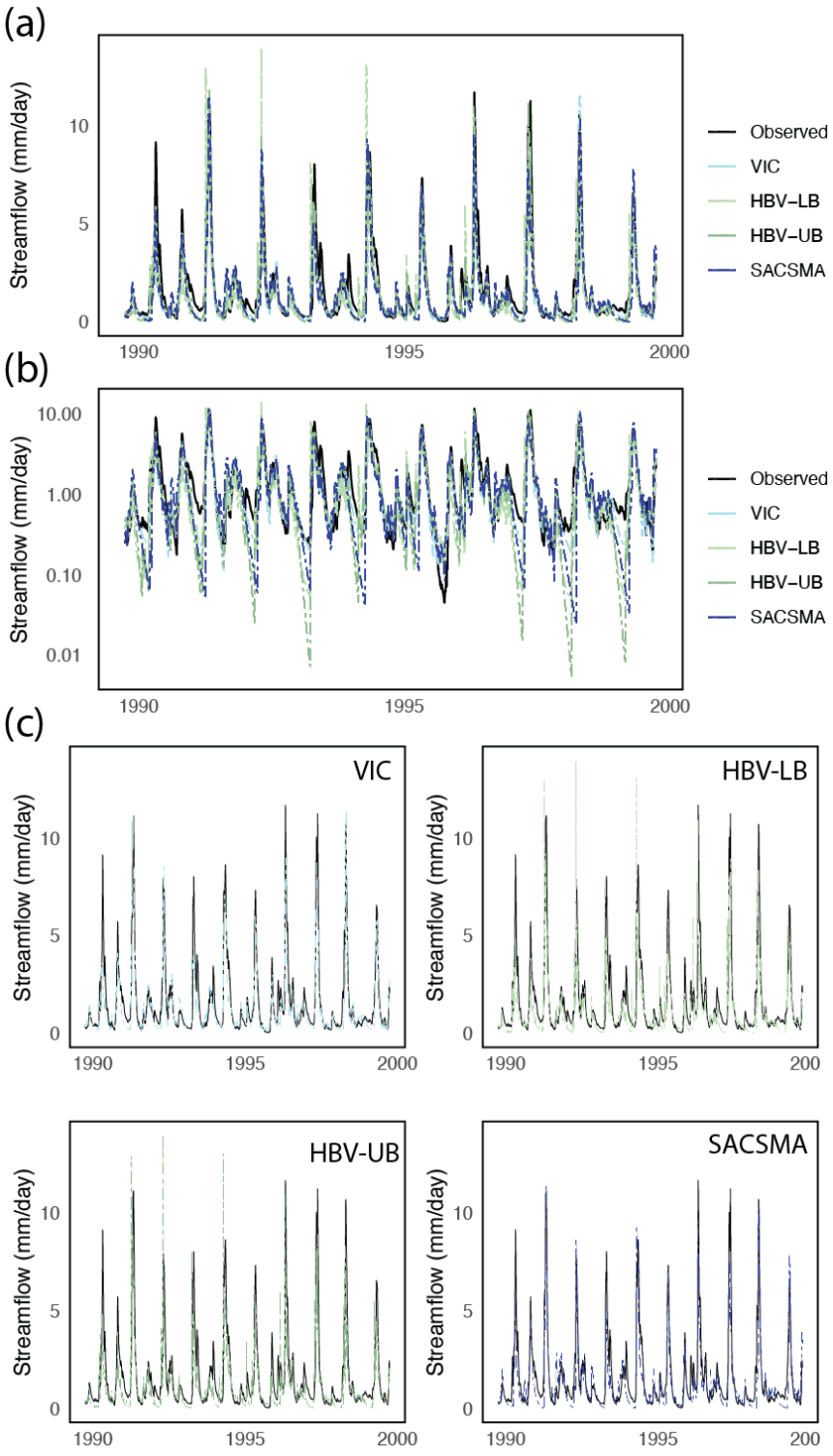

Figure 9: *Plots of model-predicted and observed timeseries of streamflow for the Fish River (USGS Gage #01013500) shown as (a) multiple series, (b) transformed on a logarithmic scale, and (c) as subplots or facets. Modeling observations originate from Kratzert (2019) and Kratzert et al., (2019) and show results for the Variable Infitration Capacity Model (VIC), the HBV model as calibrated to an upper benchmark (HBV-UB) and lower benchmark (HBV-LB), and the Sacramento Moisture and Accounting Model (SACSMA).*

Consideration 3: Make complexity work for you

One of the best ways to simplify large volumes and varieties of data is by using synthesis plots. Here, we define a synthesis plot as any type of plot that combines multiple graphical approaches and encoding attributes. By this definition, many of the best visualizations today combine multiples of plot types to convey key messages. By nature, synthesis plots are complex, summarizing multidimensional datasets with multiple encoding attributes (points, lines, color, arrows). They may incorporate symbols, often make heavy use of color, and include strong labeling and text throughout. Exceptional examples of such plots include circos diagrams (Figure 10a), Sankey diagrams, table-based diagrams, treemaps (Figure 10b), and ordered bar charts. While these visualizations are complicated, what often makes them successful is that the creators give narrative to these plots (Tufte et al., 1990), through captions, labeling, and either verbal (e.g., during a presentation or described in an animation) or written descriptions (e.g., captions or manuscript or article text). If all else fails, your caption should be able to explain your figure. If it can't, your figure is probably too complicated and needs revision. Complexity does not necessarily detract from interpretability; synthesis plots represent an exceptional opportunity for visual communication.

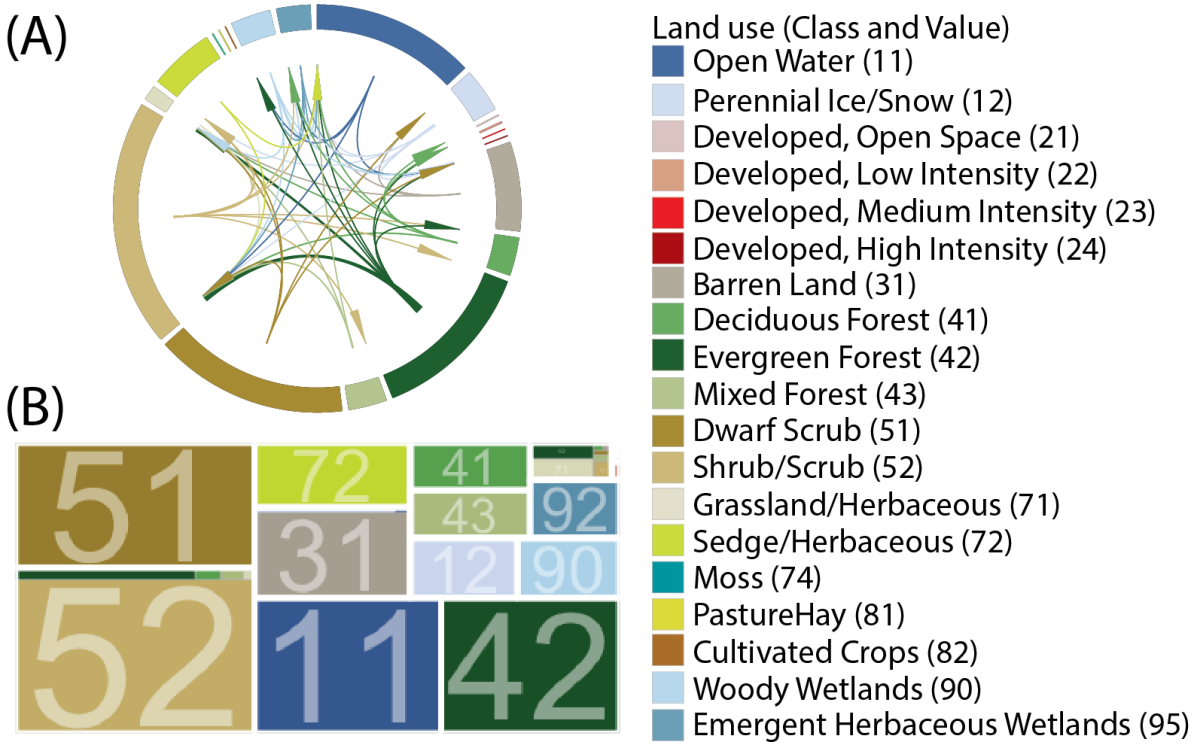

Figure 10: *This figure visualizes over 1 billion pixels of land cover change from 2001 to 2011 across Alaska (Homer et al., 2015; National Land Cover Dataset) using two more complex plots, a circos plot (Panel A) and a treemap (Panel B). The circos plot displays changes in land cover, with the outer circle proportion roughly corresponding to 2001 land cover, and the arrows sized to indicate changes in land cover from 2001 to 2011. The color of these arrows corresponds to 2001 land cover, with arrows pointing to the observed land cover in 2011. In contrast, areas shown in Panel B correspond to 2011 land cover, with inset areas colored according to 2001 land cover.*

3.4 Challenges and opportunities in the use of color

Color is central to the creation of scientific visualizations. (Zeller and Rogers, 2020). While color can mislead the reader when interpreting figures (Ware et al. 2008; Samsel et al., 2018), removing color as an encoding attribute is not always feasible. When using color, there are several approaches that can be used to make visualizations clear and easy to interpret.

Table 1: *Resources for creating color palettes.*

Resource	Description
Color Brewer, https://colorbrewer2.org	Color Brewer is a great website for choosing color blind friendly options with both gradients and categorical variables. The system also has packages that interface with R software and ggplot, as well as Python through seaborn.
Viridis, https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html	Viridis is a color palette and package that provides a color blind and grey scale friendly gradients with several options of colors. The Viridis color palette interfaces through matplotlib in Python and a package in R (viridis)
Cividis, https://github.com/marcosci/cividis;	Cividis is an optimized version of Viridis for all forms of color blindness. You can find Cividis in a function on R called cividis as well as matplotlib in Python.
Color Moves, https://sciviscolor.org/home/colormoves/	Color Moves is an interactive tool that can be used to tailor colorbars to datasets.
Chroma.js Color Palette Helper, https://gka.github.io/palettes/	Choma.js Color Palette helper lets you select colors and then creates a sequential or divergent scheme. The great part about helper is it tells you how different the resulting palette is in lightness, saturation and hue, allowing for creation of unique color palettes that also are easy to distinguish.
Colorgorical, http://vrl.cs.brown.edu/color	Colorgorical is a tool to choose a categorical palette, letting the user specify the distance between colors for better visualization.

Consideration 1: Use care with color-based pattern plots and color-based comparisons

Color is often used in data visualization to differentiate between groups, outliers, and across scales. Although color can be useful in the right context, it is difficult to differentiate between colors of the same intensity or saturation (Samsel et al., 2018; Ware, 2008). Additionally, assigning a scale to a color gradient or understanding how far apart two colors are on a scale are difficult for readers to interpret (i.e., gradients force the readers to do “visual math”). In this vein, using color contrast can distort the reader’s view of the data displayed when not used properly in visualization (Samsel et al., 2018). For example, in gradients of data, ratios of hues red to green

are about equal, but hues orange to blue are about 1:2 for an equal gradient of data. This color inequality is particularly important when areas are displayed because our brains will try to bring the complementary colors into balance and misjudge the aerial extent of a color (Samsel et al., 2018) and are not appropriate for displaying categorical variables (i.e., categories of vegetable should not be displayed on a white to red color gradient).

When it comes to visualizations that rely on color to support interpretation, the widely used heat map is particularly difficult to assess because colors can look different depending on what colors are surrounding them (Albers, 1975). In addition, colors like red can dominate other colors, leading to interpretation that there is more or extensive red in relation to other colors (Figure 6a; often referred to as 'contrast of extension'). Alternatively, it is difficult to compare colors with increasing distance or blank space between the colors. To top all of these issues off, color blindness is a serious limitation in interpreting figures, including but not limited to heatmaps.

Because color can mislead readers, we suggest that color is used carefully and after other alternatives in big data visualization. There are many other encoding attributes that can be used for differentiation between groups or scale. Alternatives include shape, line weight, size and length.

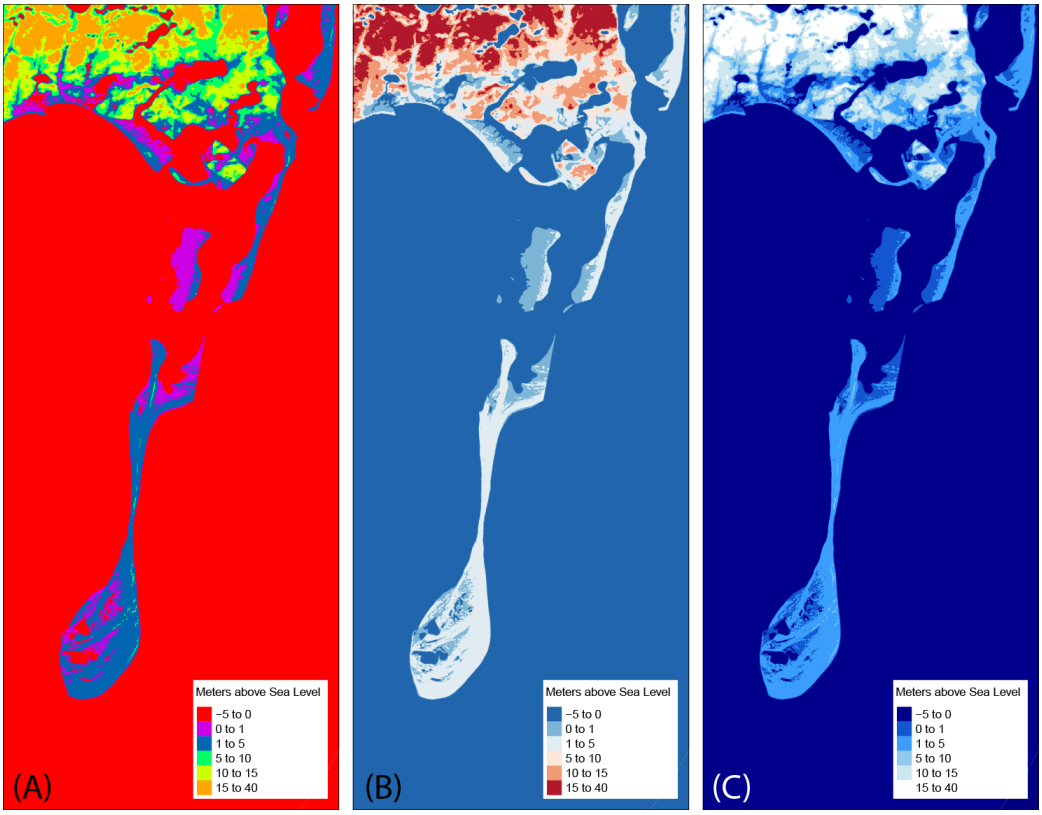

Figure 11: *Color choice is an important decision in any visualization, particularly when colors will touch each other. Using the NOAA DEM dataset (CIRES, 2014), this figure displays three different color schemes to illustrate these choices along the coast of Cape Cod in Massachusetts (USA). Panel A employs a rainbow color scheme where colors vary by hue but not saturation or value. The gradient does not reflect the continuous nature of the data, and the red dominates the rest of the colors, leading viewers to overestimate the area of ocean (contrast of extension).*

Additionally, this color palette does not work for color-blind individuals or gray-scale printing. Panel B displays a diverging, complementary color scheme with blue and red on the opposite sides. Although this color scale is good for visualization, it might not be the most effective for displaying elevation, as the high elevations seem like hotspots of something in red. Finally, Panel C illustrates a sequential color scheme with blue colors. This palette is probably the most effective of the three for this dataset (coastal DEM), where the ocean is deep blue and the higher elevations are white. The blue colors are expected for the archetypal imagery of the ocean.

Consideration 1: Is it useful to add color to highlight an overall pattern or to a distinct group of variables?

Using color to convey an additional dimension within visualizations of large volumes of data can be an effective way to convey additional information. Typically, the intent of using color in these types of visualizations is to either (1) convey an overarching pattern (e.g., Figure 4c, Figure 5b, c; Figure 7b, c; Figure 8a, b) or (2) to display outliers or a categorical group (Figure 4d; Figure 8c). However, a plot that tries to convey both patterns and outliers may easily overwhelm perception of key messages or interpretation of the visualization.

The major challenge to using color to visualize patterns for a large dataset is that there are many aspects of a visualization that can obscure our ability to discern a given pattern (Albers, 1975; Samsel et al., 2018). Care should be given to ensure that all values can be distinguished. Importantly, the choice of a color bar should enable interpretation of the perceived pattern. An alternative to using color may be to use another type of differentiator (e.g., shape) that is easier to interpret.

An alternative to visualizing patterns is to use color to highlight distinct groups of variables (e.g., details). This type of visualization may be employed to identify a statistical grouping, such as outliers, a categorical grouping, such as different types of values (e.g., watersheds within the same sub-basin, Figure 8c), or a research-based outcome, such as groupings from hierarchical clustering or the like. As opposed to a typical pattern-based plot, this type of plot may simplify the number of colors being quantitatively perceived across a large volume of data.

Consideration 2: When using color, choose color schemes or gradients that allow for contrast in hue, value, and saturation.

Color theory represents an extensive study of how colors interact, how best to mix colors, and how best to use colors within visualizations of any kind (Albers, 1975). The basis of color theory is the color wheel, which is used to identify colors that may work together to make visuals appealing and to provide contrast (Rhyne, 2012). As defined by color theory, there are three aspects of color: hue (color wavelength – e.g., blue or green), saturation (depth of the color), and value (grayness of the color). These aspects of color may be used to generate a color scheme or gradient for a particular plot.

Many different combinations of colors that incorporate color theory already exist. A complementary color scheme corresponds to colors that exist on opposite sides of the color wheel. An alternative to a complementary color scheme is a split complementary color scheme, which uses a base color along with two colors on opposites sides of the first color complement (i.e., the color directly across from the base color on the color wheel). An analogous color

scheme is when three colors are adjacent to each other, and a triadic color scheme is when three colors are equally spaced around the color wheel. Color schemes employing complementary color schemes as end members of a gradient are often called divergent color schemes (frequently with white or a neutral color in the middle). Schemes with colors of a similar hue or within an analogous color palette are commonly called sequential color schemes (i.e., light green, grass green, dark green).

The best way to make or select a color scheme that is easier to differentiate and distinguish is to leverage the three different aspects of color: hue, saturation and value. By employing at least two aspects of color, such as saturation and value, visualizations can appear more interpretable (Ware, 2008). Likewise, combining colors of different saturation or value with changes in hue can make colors more distinguishable, particularly for viewers with color blindness.

Colors appropriate for visualization are frequently difficult and overwhelming to choose. For those who may be new to the principles of color theory, there are several web-based resources to assist with selecting a gradient or categorical color scheme. Table 1 includes a list of these resources that can be used for developing a color palette.

Consideration 3: Avoid rainbow color scales and limit the number of categories to enable interpretation

While this recommendation is widely known, it still bears repeating. Despite their broad proliferation, rainbow color ramps use constant saturation and value, only varying in hue (Borland and Taylor, 2007). Therefore, it is very difficult for colorblind individuals to distinguish between colors on a rainbow scale. In addition, the gradient includes all colors on the visible spectrum, making it hard for viewers to perceive which colors correspond to a positive or negative value. This aspect also makes it challenging to interpret which part of the scale the color on the gradient represents (Borland and Taylor, 2007). To emphasize the problems with this particular color scale, we include a particularly horrific example in Figure 11a.

To avoid the rainbow color scale, pick a type of contrast that best fits the visualization (Samsel et al., 2018). Cool to warm contrasts, such as blue to red or yellow, are often good for scales with gradients from low to high. Additionally, using single color gradients is a strong approach for producing easy-to-discern gradients. Humans are biased to prefer balanced gradients with either cool to warm or complementary contrast in a gradient (Albers, 1975). Limiting colors to seven categories or less helps the viewer interpret gradient or categorical color scale (e.g., Figure 5c; Figure 11).

Consideration 4: Think beyond just colorblindness

Colorblindness is not the only visual disability that affects the interpretation of visualizations. Low vision individuals may find it difficult to read or interpret details in a visualization. Often increasing the contrast of the colors used in the visualization or increasing font size can make the figure more accessible to those with low vision (Power and Jergensen, 2010). In addition, rasters are inaccessible to the visually impaired because the components of the figure are hard to separate (Choi et al., 2019). Often those with visual impairments will use computer reading software to get information from a manuscript. To provide more information, make sure the caption in the figure is accessible to reading programs and fully describes the graph and the

results. Accessibility can be increased by considering creative choices to explain visualizations, such as by adding supplemental movies or audio that describe the study or the visualizations (Power and Jergensen, 2010). Finally, there are new frontiers opening up to make visualizations more accessible. For instance, machine learning approaches are being tested to identify individual elements of a visualization to pass to a reading software (Choi et al., 2019).

Consideration 5: Use color to tell a story – go beyond 'best practices' for a single visualization

Throughout a text – whether a manuscript, presentation, interactive video, blog post, or other published article – colors not only tell the story of one visualization but the larger narrative of the manuscript. To keep the message and narrative consistent, it is important to keep a consistent color scheme throughout, particularly for visualizations that use the same components (e.g., variables, parameters, groups). As discussed above, picking colors that are easy to distinguish, that visually represent the figure's trends, and that can be interpreted by all audiences is the most important aspect of choosing a color palette for a presentation or manuscript.

In particular, it's important to select color palettes that do not violate or challenge cultural or archetypal expectations, and that support and correspond to the data being shown (e.g., Figure 10b versus Figure 10c). For example, do not use warm colors for water or snowfall or cool colors for fire frequency. When using a dataset that is known to have an associated color palette, employ this same color scheme (e.g., Figure 11, displaying land cover data with a previously created and widely used color palette). If choosing a color palette fills you with dread, the easiest direction is to pick an already existing color palette that is good for color blind persons and printing grey scale (i.e., viridis). There are also really fun ways to dream up color schemes, such as using movie color palettes (see twitter account @cinemapalettes or Movies in Color: https://moviesincolor.com/).

4.0 From static to interactive

The fourth V of big data, velocity, refers to the speed and temporal resolution at which data is collected, both of which are accelerating, and are a challenge to visualize and represent. We can work with this type of data through two methods: summarizing such information in static formats, or creative visualization techniques such as animation. In this particular piece, we focus less on velocity as a visualization need, though note that the importance of visualizing velocity will likely continue to grow in the near future.

Common approaches to visualize velocity in static formats include visualizing time using color, or grouping values and using time as an organizing dimension. For instance, heat maps or barcode plots are commonly used to convey temporally resolute datasets at one or many locations. Other plotting options include the use of spatial snapshots (e.g., showing temporal plots for a particular point, slice, or volume) or temporal snapshots (e.g., showing spatial plots for a particular time). However, it can be quite challenging to convey velocity in static visualizations.

At current, more and more journals offer the option to upload animations as supplementary material. Journals are also beginning to develop visualization-based submissions (e.g., HPEye, within the journal *Hydrologic Processes*). For those interested in such options, Table 2 lists

several packages in R and Python that can be used for creating animations or interactive graphics.

Table 2: *A list of commonly used packages and libraries for R and Python that enable animation or interactive graphics.*

Programming Language	Packages or Libraries
R	animate
	gganimate
	plotly
	googlevis
	shiny
Python	Plotly Express
	Matplotlib & seaborn
	Bokeh
	Altair
	nbinteract

While the scientific literature largely draws on static visualizations, interactive visualizations are becoming increasingly common for science communication to a range of technical and non-technical audiences. Though the focus of our article is primarily static visualizations, we briefly summarize considerations for interactive visualizations. Interactive visualizations can be used for a variety of purposes: they may be helpful in the data exploration phase, or may accompany peer-reviewed manuscripts. As most journals do not support interactive formats, it is important to remember that creating a static version of the figure is often necessary for a manuscript. However, such interactive visualizations are an option for supplemental or accompanying websites.

Interactive graphs lead to more complexity and additional decisions. While there are many types of interactive visualizations, the most common examples involve: 1) Changing/choosing a dataset or a unit of analysis in a visualization, 2) filtering or querying a dataset in the visualization, 3) toggling visualization features such as variables or colors, 4) combining or separating visualizations, 5) interactive annotations that provide more information, and 6) zooming in and out or changing the level of detail (Wash, 2020). Approaches to parse interactive graphic components are now being described through grammars of interaction (e.g., Vega, https://vega.github.io/vega/; Vega-Lite, https://vega.github.io/vega-lite/), which provide a framework for approaching the creation of interactive visualizations. In addition, many common scientific programming languages are facilitating the creation of interactive web apps (e.g., R Shiny, https://shiny.rstudio.com/).

With these common types of interactive visualizations come pitfalls. The following are a brief list of recommendations to consider when creating interactive visualizations: 1) Ensure that the static version is explanatory, as most people will not interact with the figure; 2) Consider whether interactive features can be engaged and used on all types of devices (e.g., individual points can be difficult to interact with on touch devices); and 3) Make the interactive components of the visualization well-paced and manageable (as opposed to too cumbersome or slow) to maintain interest with your audience (Wash 2020). It is likely that the tools and

recommendations for best practices concerning interactive visualizations will continue to evolve, especially given the growing interest in the creation and use of interactive visualizations across the scientific community.

5.0 Conclusions:

Conveying large datasets within publications or presentations is challenging, especially when it comes to visualizing large amounts of data. The visualization of large, complex datasets requires rethinking our common assumptions and approaches for creating plots. In particular, when visualizing large datasets, researchers must make many decisions before arriving at a final plot. These decisions include whether to show all values or to aggregate, how to convey multiple categories or comparisons, and how to display multiple dimensions, all in succinct, easily interpretable ways.

Our introductory overview provides several recommendations, such as choice of plot type, encoding attributes, and groupings, to aid in the creation of clear visualizations. Thinking carefully about these choices can enhance visualization quality and messaging. However, whether (or not) our recommendations apply in any particular case will ultimately depend on the overarching message conveyed by each visualization and the size and character of the associated dataset. Though we often focus on encoding attributes and plot choice when approaching visualizations, equally important is to give narrative to our plots using notations, legends, and a clear caption.

Above all, when creating plots of large datasets, iteration is key. Be sure to ask for input along the way, ensuring that the key takeaways intended for a particular visualization are clear to others. Scientific visualization is something that we all have very strong feelings about. We therefore emphasize that our recommendations are just that - not hard and fast rules that should be unilaterally applied in all scenarios. Be creative, thoughtful, and intentional with your designs, and use your best judgment along the way.

7.0 Acknowledgements: The authors acknowledge funding from the CUSE grant program at Syracuse University to CK and a National Science Foundation award (#1940006) to AB. All data used in the creation of scientific visualizations are publicly available and cited in the text. All code for visualizations are included in a Github repository at: https://github.com/abraswell/BigDataViz. The article was improved by thoughtful comments from two anonymous reviewers.

8.0 References:

Addor, N., Newman, A. J., Mizukami, N. and Clark, M. P., 2017a. Catchment attributes for large-sample studies. Boulder, CO: UCAR/NCAR. https://doi.org/10.5065/D6G73C3Q

Addor, N., Newman, A. J., Mizukami, N. and Clark, M. P. 2017b. The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, doi:10.5194/hess-21-5293-2017.

Alam, A.U., Clyne, D., Jin, H., Hu, N.-X., Deen, M.J., 2020. Fully Integrated, Simple, and Low-Cost Electrochemical Sensor Array for in Situ Water Quality Monitoring. ACS Sens. 5, 412–422. https://doi.org/10.1021/acssensors.9b02095

Albers, J., 1975. Interaction of Color: 50th Anniversary Edition. Yale University Press, New Haven, CT.

Allen, M., 2018. Introducing Raincloud Plots! Neuroconscience. URL https://micahallen.org/2018/03/15/introducing-raincloud-plots/ (accessed 8.14.20).

Allen, M., Poggiali, D., Whitaker, K., Marshall, T.R., Kievit, R.A., 2019. Raincloud plots: a multi-platform tool for robust data visualization. Wellcome Open Res. 4. https://doi.org/10.12688/wellcomeopenres.15191.1

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nat. News 533, 452. https://doi.org/10.1038/533452a

Balch, J.K., Iglesias, V., Braswell, A.E., Rossi, M.W., Joseph, M.B., Mahood, A.L., Shrum, T.R., White, C.T., Scholl, V.M., McGuire, B., Karban, C., Buckland, M., Travis, W.R., 2020. Social-Environmental Extremes: Rethinking Extraordinary Events as Outcomes of Interacting Biophysical and Social Systems. Earths Future 8, e2019EF001319. https://doi.org/10.1029/2019EF001319

Baroni, G., Schalge, B., Rakovec, O., Kumar, R., Schüler, L., Samaniego, L., Simmer, C., Attinger, S., 2019. A Comprehensive Distributed Hydrological Modeling Intercomparison to Support Process Representation and Data Collection Strategies. Water Resour. Res. 55, 990–1010. https://doi.org/10.1029/2018WR023941

Best, J., 2019. Anthropogenic stresses on the world's big rivers. Nat. Geosci. 12, 7–21. https://doi.org/10.1038/s41561-018-0262-x

Blaszczak, J.R., Delesantro, J.M., Zhong, Y., Urban, D.L., Bernhardt, E.S., 2019. Watershed urban development controls on urban streamwater chemistry variability. Biogeochemistry 144, 61–84. https://doi.org/10.1007/s10533-019-00572-7

Bobko, P., Karren, R., 1979. The Perception of Pearson Product Moment Correlations from Bivariate Scatterplots. Pers. Psychol. 32, 313–325. https://doi.org/10.1111/j.1744-6570.1979.tb02137.x

Boone, A.P., Gunalp, P., Hegarty, M., 2018. Explicit versus actionable knowledge: The influence of explaining graphical conventions on interpretation of hurricane forecast visualizations. J. Exp. Psychol. Appl. 24, 275–295. https://doi.org/10.1037/xap0000166

Borland, D., Taylor, R.M., 2007. Rainbow Color Map (Still) Considered Harmful. IEEE Comput. Graph. Appl. 27, 14–17. https://doi.org/10.1109/MCG.2007.323435

Chang, W.L., Grady, N., 2019. NIST Big Data Interoperability Framework: Volume 1, Definitions.

Chernoff, H., 1973. The Use of Faces to Represent Points in k-Dimensional Space Graphically. J. Am. Stat. Assoc. 68, 361–368. https://doi.org/10.1080/01621459.1973.10482434

924 Choi, J., Jung, S., Park, D.G., Choo, J. Elmqvist, N., 2019.Visualizing for the Non-Visual:
925 Enabling the Visually Impaired to Use Visualization. In *Computer Graphics Forum*,
926 38(3), pp. 249-260.
927 Cominola, A., Nguyen, K., Giuliani, M., Stewart, R.A., Maier, H.R., Castelletti, A., 2019. Data
928 Mining to Uncover Heterogeneous Water Use Behaviors From Smart Meter Data. Water
929 Resour. Res. 55, 9315–9333. https://doi.org/10.1029/2019WR024897
930 Cooperative Institute for Research in Environmental Sciences (CIRES) at the University of
931 Colorado, Boulder, 2014. Continuously Updated Digital Elevation Model (CUDEM) -
932 1/9 Arc-Second Resolution Bathymetric-Topographic Tiles.
933 CrowdWater, n.d. URL https://crowdwater.ch/en/welcome-to-crowdwater/ (accessed 8.11.20).
934 Dalin, C., Konar, M., Hanasaki, N., Rinaldo, A., Rodriguez-Iturbe, I., 2012. Evolution of the
935 global virtual water trade network. Proc. Natl. Acad. Sci. 109, 5989–5994.
936 https://doi.org/10.1073/pnas.1203176109
937 DeCicco, L., Hirsch, R., Lorenz, D., Read, J., Walker, J., Carr, L., Watkins, D., 2020.
938 dataRetrieval: Retrieval Functions for USGS and EPA Hydrologic and Water Quality
939 Data.
940 Deitrick, S., Edsall, R., 2006. The Influence of Uncertainty Visualization on Decision Making:
941 An Empirical Evaluation, in: Riedl, A., Kainz, W., Elmes, G.A. (Eds.), Progress in
942 Spatial Data Handling: 12th International Symposium on Spatial Data Handling.
943 Springer, Berlin, Heidelberg, pp. 719–738. https://doi.org/10.1007/3-540-35589-8_45
944 Desnoyers, L., 2011. Toward a Taxonomy of Visuals in Science Communication. Tech.
945 Commun. 58, 119–134.
946 Eklund, A., 2016. beeswarm: The Bee Swarm Plot, an Alternative to Stripchart.
947 Eklund, A., 2015. beeswarm: an R package. http://www.cbs.dtu.dk/~eklund/beeswarm/ (accessed
948 8.14.20).
949 Falcone, J., 2011. GAGES-II: Geospatial Attributes of Gages for Evaluating Streamflow.
950 https://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml#stdorder
951 (accessed 8.14.20).
952 Falcone, J.A., Carlisle, D.M., Wolock, D.M., Meador, M.R., 2010. GAGES: A stream gage
953 database for evaluating natural and altered flow conditions in the conterminous United
954 States. Ecology 91, 621–621. https://doi.org/10.1890/09-0889.1
955 Farley, S.S., Dawson, A., Goring, S.J., Williams, J.W., 2018. Situating Ecology as a Big-Data
956 Science: Current Advances, Challenges, and Solutions. BioScience 68, 563–576.
957 https://doi.org/10.1093/biosci/biy068
958 Few, S., 2009. Now You See It: Simple Visualization Techniques for Quantitative Analysis, 1st
959 edition. ed. Analytics Press, Oakland, Calif.
960 Fuka, DR, Walter, M., Archibald, J., Steenhuis, T., Easton, Z., 2018. EcoHydRology: A
961 Community Modeling Foundation for Eco-Hydrology.
962 Ge, Y., Li, S., Lakhan, V.C., Lucieer, A., 2009. Exploring uncertainty in remotely sensed data
963 with parallel coordinate plots. Int. J. Appl. Earth Obs. Geoinformation 11, 413–422.
964 https://doi.org/10.1016/j.jag.2009.08.004
965 Gill, J.C., Malamud, B.D., 2017. Anthropogenic processes, natural hazards, and interactions in a
966 multi-hazard framework. Earth-Sci. Rev. 166, 246–269.
967 https://doi.org/10.1016/j.earscirev.2017.01.002
968 Gill, J.C., Malamud, B.D., 2014. Reviewing and visualizing the interactions of natural hazards.
969 Rev. Geophys. 52, 680–722. https://doi.org/10.1002/2013RG000445

970 Gold, D.F., Reed, P.M., Trindade, B.C., Characklis, G.W., 2019. Identifying Actionable
971 Compromises: Navigating Multi-City Robustness Conflicts to Discover Cooperative Safe
972 Operating Spaces for Regional Water Supply Portfolios. Water Resour. Res. 55, 9024–
973 9050. https://doi.org/10.1029/2019WR025462
974 Gordin, D.N., Pea, R.D., 1995. Prospects for Scientific Visualization as an Educational
975 Technology. J. Learn. Sci. 4, 249–279. https://doi.org/10.1207/s15327809jls0403_1
976 Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google
977 Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sens. Environ.,
978 Big Remotely Sensed Data: tools, applications and experiences 202, 18–27.
979 https://doi.org/10.1016/j.rse.2017.06.031
980 Harrison, P., 2017. Scatter plots with density quartiles [WWW Document]. URL
981 http://www.logarithmic.net/pfh-files/blog/01509162940/scatter.html (accessed 8.14.20).
982 Hicks, A., Barclay, J., Chilvers, J., Armijos, M.T., Oven, K., Simmons, P., Haklay, M., 2019.
983 Global Mapping of Citizen Science Projects for Disaster Risk Reduction. Front. Earth
984 Sci. 7. https://doi.org/10.3389/feart.2019.00226
985 Hintze, J.L., Nelson, R.D., 1998. Violin Plots: A Box Plot-Density Trace Synergism. Am. Stat.
986 52, 181–184. https://doi.org/10.1080/00031305.1998.10480559
987 Höffler, T.N., 2010. Spatial Ability: Its Influence on Learning with Visualizations—a Meta-
988 Analytic Review. Educ. Psychol. Rev. 22, 245–269. https://doi.org/10.1007/s10648-010-
989 9126-7
990 Höffler, T.N., Leutner, D., 2007. Instructional animation versus static pictures: A meta-analysis.
991 Learn. Instr. 17, 722–738. https://doi.org/10.1016/j.learninstruc.2007.09.013
992 Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.,
993 Wickham, J., Megown, K., 2015. Completion of the 2011 National Land Cover Database
994 for the Conterminous United States – Representing a Decade of Land Cover Change
995 Information. Photogramm. Eng. Remote Sens. 81, 345–354.
996 Jackson, E.K., Roberts, W., Nelsen, B., Williams, G.P., Nelson, E.J., Ames, D.P., 2019.
997 Introductory overview: Error metrics for hydrologic modelling – A review of common
998 practices and an open source library to facilitate use and adoption. Environ. Model.
999 Softw. 119, 32–48. https://doi.org/10.1016/j.envsoft.2019.05.001
1000 Joseph, M.B., Rossi, M.W., Mietkiewicz, N.P., Mahood, A.L., Cattau, M.E., Denis, L.A.S.,
1001 Nagy, R.C., Iglesias, V., Abatzoglou, J.T., Balch, J.K., 2019. Spatiotemporal prediction
1002 of wildfire size extremes with Bayesian finite sample maxima. Ecol. Appl. 29, e01898.
1003 https://doi.org/10.1002/eap.1898
1004 Kampstra, P., 2014. beanplot: Visualization via Beanplots (like Boxplot/Stripchart/Violin Plot).
1005 Kampstra, P., 2008. Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. J.
1006 Stat. Softw. 28, 1–9. https://doi.org/10.18637/jss.v028.c01
1007 Kelleher, C. A., Scholz, C. A., Condon, L., Reardon, M., 2018. Drones in Geoscience Research:
1008 The Sky Is the Only Limit [WWW Document]. Eos. URL
1009 https://doi.org/10.1029/2018EO092269 (accessed 9.3.20).
1010 Kelleher, C., Wagener, T., 2011. Ten guidelines for effective data visualization in scientific
1011 publications. Environ. Model. Softw. 26, 822–827.
1012 https://doi.org/10.1016/j.envsoft.2010.12.006
1013 Kinkeldey, C., MacEachren, A.M., Riveiro, M., Schiewe, J., 2017. Evaluating the effect of
1014 visually represented geodata uncertainty on decision-making: systematic review, lessons

learned, and recommendations. Cartogr. Geogr. Inf. Sci. 44, 1–21. https://doi.org/10.1080/15230406.2015.1089792
Kinkeldey, C., MacEachren, A.M., Schiewe, J., 2014. How to Assess Visual Communication of Uncertainty? A Systematic Review of Geospatial Uncertainty Visualisation User Studies. Cartogr. J. 51, 372–386. https://doi.org/10.1179/1743277414Y.0000000099
Kirsh, D., 2010. Thinking with external representations. AI Soc. 25, 441–454. https://doi.org/10.1007/s00146-010-0272-8
Knapp, J.L.A., Freyberg, J. von, Studer, B., Kiewiet, L., Kirchner, J.W., 2020. Concentration–discharge relationships vary among hydrological events, reflecting differences in event characteristics. Hydrol. Earth Syst. Sci. 24, 2561–2576. https://doi.org/10.5194/hess-24-2561-2020
Kratzert, F. (2019). CAMELS benchmark models, HydroShare, https://doi.org/10.4211/hs.474ecc37e7db45baa425cdb4fc1b61e1
Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019.
Krysanova, V., Vetter, T., Eisner, S., Huang, S., Pechlivanidis, I., Strauch, M., Gelfan, A., Kumar, R., Aich, V., Arheimer, B., Chamorro, A., Griensven, A. van, Kundu, D., Lobanova, A., Mishra, V., Plötner, S., Reinhardt, J., Seidou, O., Wang, X., Wortmann, M., Zeng, X., Hattermann, F.F., 2017. Intercomparison of regional-scale hydrological models and climate change impacts projected for 12 large river basins worldwide—a synthesis. Environ. Res. Lett. 12, 105002. https://doi.org/10.1088/1748-9326/aa8359
Krzywinski, M., Altman, N., 2014. Visualizing samples with box plots. Nat. Methods 11, 119–120. https://doi.org/10.1038/nmeth.2813
Li, D., Lettenmaier, D.P., Margulis, S.A., Andreadis, K., 2019. The Role of Rain-on-Snow in Flooding Over the Conterminous United States. Water Resour. Res. 55, 8492–8513. https://doi.org/10.1029/2019WR024950
Liu, S., Maljovec, D., Wang, B., Bremer, P.-T., Pascucci, V., 2017. Visualizing High-Dimensional Data: Advances in the Past Decade. IEEE Trans. Vis. Comput. Graph. 23, 1249–1268. https://doi.org/10.1109/TVCG.2016.2640960
Liu, Z., Stasko, J., 2010. Mental Models, Visual Reasoning and Interaction in Information Visualization: A Top-down Perspective. IEEE Trans. Vis. Comput. Graph. 16, 999–1008. https://doi.org/10.1109/TVCG.2010.177
Matejka, J., Fitzmaurice, G., 2017. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, CHI '17. Association for Computing Machinery, New York, NY, USA, pp. 1290–1294. https://doi.org/10.1145/3025453.3025912
Maxwell, R.M., Putti, M., Meyerhoff, S., Delfs, J.-O., Ferguson, I.M., Ivanov, V., Kim, J., Kolditz, O., Kollet, S.J., Kumar, M., Lopez, S., Niu, J., Paniconi, C., Park, Y.-J., Phanikumar, M.S., Shen, C., Sudicky, E.A., Sulis, M., 2014. Surface-subsurface model intercomparison: A first set of benchmark results to diagnose integrated hydrology and feedbacks. Water Resour. Res. 50, 1531–1549. https://doi.org/10.1002/2013WR013725
Mcgill, R., Tukey, J.W., Larsen, W.A., 1978. Variations of Box Plots. Am. Stat. 32, 12–16. https://doi.org/10.1080/00031305.1978.10479236

Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J.J., Ioannidis, J.P.A., 2017. A manifesto for reproducible science. Nat. Hum. Behav. 1, 1–9. https://doi.org/10.1038/s41562-016-0021

Murphy, K., Heery, B., Sullivan, T., Zhang, D., Paludetti, L., Lau, K.T., Diamond, D., Costa, E., O′Connor, N., Regan, F., 2015. A low-cost autonomous optical sensor for water quality monitoring. Talanta 132, 520–527. https://doi.org/10.1016/j.talanta.2014.09.045

National Land Cover Dataset, n.d. 2001 from-to 2011 Land Cover Change Pixels (ALASKA). URL https://www.mrlc.gov/data/nlcd-2001-2011-land-cover-change-pixels-alaska (accessed 8.14.20).

Parra, L., Sendra, S., García, L., Lloret, J., 2018. Design and Deployment of Low-Cost Sensors for Monitoring the Water Quality and Fish Behavior in Aquaculture Tanks during the Feeding Process. Sensors 18, 750. https://doi.org/10.3390/s18030750

Parsons, P., Sedig, K., 2014. Common Visualizations: Their Cognitive Utility, in: Huang, W. (Ed.), Handbook of Human Centric Visualization. Springer, New York, NY, pp. 671–691. https://doi.org/10.1007/978-1-4614-7485-2_27

Potter, M.C., Wyble, B., Hagmann, C.E., McCourt, E.S., 2014. Detecting meaning in RSVP at 13 ms per picture. Atten. Percept. Psychophys. 76, 270–279. https://doi.org/10.3758/s13414-013-0605-z

Power, C. and Jürgensen, H., 2010. Accessible presentation of information for people with visual disabilities. *Universal Access in the Information Society*, *9*(2), pp. 97-119.

Raseman, W.J., Jacobson, J., Kasprzyk, J.R., 2019. Parasol: an open source, interactive parallel coordinates library for multi-objective decision making. Environ. Model. Softw. 116, 153–163. https://doi.org/10.1016/j.envsoft.2019.03.005

Rhyne, T.-M., 2012. Applying Artistic Color Theories to Visualization, in: Dill, J., Earnshaw, R., Kasik, D., Vince, J., Wong, P.C. (Eds.), Expanding the Frontiers of Visual Analytics and Visualization. Springer, London, pp. 263–283. https://doi.org/10.1007/978-1-4471-2804-5_15

Ribecca, S., The Data Visualisation Catalogue. https://datavizcatalogue.com/ (accessed 8.14.20).

Rougier, N.P., Droettboom, M., Bourne, P.E., 2014. Ten Simple Rules for Better Figures. PLOS Comput. Biol. 10, e1003833. https://doi.org/10.1371/journal.pcbi.1003833

Samsel, F., Bartram, L., Bares, A., 2018. Art, Affect and Color: Creating Engaging Expressive Scientific Visualization, in: 2018 IEEE VIS Arts Program (VISAP). Presented at the 2018 IEEE VIS Arts Program (VISAP), pp. 1–9. https://doi.org/10.1109/VISAP45312.2018.9046053

Sandve, G.K., Nekrutenko, A., Taylor, J., Hovig, E., 2013. Ten Simple Rules for Reproducible Computational Research. PLOS Comput. Biol. 9, e1003285. https://doi.org/10.1371/journal.pcbi.1003285

Scaife, M., Rogers, Y., 1996. External cognition: how do graphical representations work? Int. J. Hum.-Comput. Stud. 45, 185–213. https://doi.org/10.1006/ijhc.1996.0048

Schwabish, J., Ribecca, S., 2014. The Graphic Continuum: A Poster Project for Your Office. Policy Viz. URL https://policyviz.com/2014/09/09/graphic-continuum/ (accessed 8.14.20).

Slater, L.J., Thirel, G., Harrigan, S., Delaigue, O., Hurley, A., Khouakhi, A., Prosdocimi, I., Vitolo, C., Smith, K., 2019. Using R in hydrology: a review of recent developments and future directions. Hydrol. Earth Syst. Sci. 23, 2939–2963. https://doi.org/10.5194/hess-23-2939-2019

Smith, A., Campbell, C., Bott, I., Faunce, L., Parrish, G., Ehrenberg-Shannon, B., McCallum, P., Stabe, M., n.d. ft-interactive/chart-doctor [WWW Document]. Financ. Times Vis. Vocab. URL https://github.com/ft-interactive/chart-doctor (accessed 8.14.20).
Smith, M.B., Seo, D.-J., Koren, V.I., Reed, S.M., Zhang, Z., Duan, Q., Moreda, F., Cong, S., 2004. The distributed model intercomparison project (DMIP): motivation and experiment design. J. Hydrol., The Distributed Model Intercomparison Project (DMIP) 298, 4–26. https://doi.org/10.1016/j.jhydrol.2004.03.040
Souza, R., 2017. Ecohydmod: Ecohydrological Modelling.
Spence, M.L., Dux, P.E., Arnold, D.H., 2016. Computations underlying confidence in visual perception. J. Exp. Psychol. Hum. Percept. Perform. 42, 671–682. https://doi.org/10.1037/xhp0000179
Spiegelhalter, D., Pearson, M., Short, I., 2011. Visualizing Uncertainty About the Future. Science 333, 1393–1400. https://doi.org/10.1126/science.1191181
Stagge, J.H., Rosenberg, D.E., Abdallah, A.M., Akbar, H., Attallah, N.A., James, R., 2019. Assessing data availability and research reproducibility in hydrology and water resources. Sci. Data 6, 190030. https://doi.org/10.1038/sdata.2019.30
Stephanie Zeller, Daniel Rogers, 2020. Visualizing Science: How Color Determines What We See 101. https://doi.org/10.1029/2020EO144330
Stream Tracker Project, Streamtracker. https://www.streamtracker.org (accessed 8.11.20).
Tessler, Z.D., Vörösmarty, C.J., Grossberg, M., Gladkova, I., Aizenman, H., Syvitski, J.P.M., Foufoula-Georgiou, E., 2015. Profiling risk and sustainability in coastal deltas of the world. Science 349, 638–643. https://doi.org/10.1126/science.aab3574
Tessum, C.W., Apte, J.S., Goodkind, A.L., Muller, N.Z., Mullins, K.A., Paolella, D.A., Polasky, S., Springer, N.P., Thakrar, S.K., Marshall, J.D., Hill, J.D., 2019. Inequity in consumption of goods and services adds to racial–ethnic disparities in air pollution exposure. Proc. Natl. Acad. Sci. 116, 6001–6006. https://doi.org/10.1073/pnas.1818859116
The Four V's of Big Data [WWW Document], n.d. . IBM Big Data Anal. Hub. https://www.ibmbigdatahub.com/infographic/four-vs-big-data (accessed 8.11.20).
Thorpe, S., Fize, D., Marlot, C., 1996. Speed of processing in the human visual system. Nature 381, 520–522. https://doi.org/10.1038/381520a0
Trimble, S.W., 1999. Decreased Rates of Alluvial Sediment Storage in the Coon Creek Basin, Wisconsin, 1975-93. Science 285, 1244–1246. https://doi.org/10.1126/science.285.5431.1244
Tufte, E.R., 2001. The Visual Display of Quantitative Information, 2nd edition edition. ed. Graphics Press, Cheshire, Conn.
Tufte, E.R., 1990. Envisioning Information. Graphics Pr, Cheshire, Connecticut.
Tukey, J.W., 1977. Exploratory data analysis. Addison-Wesley, Reading, MA.
Vos, K., Splinter, K.D., Harley, M.D., Simmons, J.A., Turner, I.L., 2019. CoastSat: A Google Earth Engine-enabled Python toolkit to extract shorelines from publicly available satellite imagery. Environ. Model. Softw. 122, 104528. https://doi.org/10.1016/j.envsoft.2019.104528
Walsh, N. 2020. Adding interactivity to Data visualization on the web. Envy Labs, posted 22 July 2020. https://envylabs.com/insights/how-to-create-interactive-data-visualizations/ (accessed January 20, 2021).

Ward, J.S., Barker, A., 2013. Undefined By Data: A Survey of Big Data Definitions. ArXiv13095821 Cs.
Ware, C., 2008. Visual Thinking for Design. Morgan Kaufmann Publishers, Burlington, MA.
Weissgerber, T.L., Milic, N.M., Winham, S.J., Garovic, V.D., 2015. Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. PLOS Biol. 13, e1002128. https://doi.org/10.1371/journal.pbio.1002128
Weissgerber, Tracey L., Winham, Stacey J., Heinzen, Ethan P., Milin-Lazovic, Jelena S., Garcia-Valencia, Oscar, Bukumiric Zoran, Savic Marko D., Garovic Vesna D., Milic Natasa M., 2019. Reveal, Don't Conceal. Circulation 140, 1506–1518. https://doi.org/10.1161/CIRCULATIONAHA.118.037777
Wickert, A.D., 2014. The ALog: Inexpensive, Open-Source, Automated Data Collection in the Field. Bull. Ecol. Soc. Am. 95, 166–176. https://doi.org/10.1890/0012-9623-95.2.68
Wickert, A.D., Sandell, C.T., Schulz, B., Ng, G.-H.C., 2019. Open-source Arduino-compatible data loggers designed for field research. Hydrol. Earth Syst. Sci. 23, 2065–2076. https://doi.org/10.5194/hess-23-2065-2019
Wickham, H., 2010. A Layered Grammar of Graphics. J. Comput. Graph. Stat. 19, 3–28. https://doi.org/10.1198/jcgs.2009.07098
Wilkinson, L., 1999. Dot Plots. Am. Stat. 53, 276–281. https://doi.org/10.1080/00031305.1999.10474474
Wilkinson, L., Anand, A., Grossman, R., 2005. Graph-theoretic scagnostics, in: IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., pp. 157–164. https://doi.org/10.1109/INFVIS.2005.1532142
Yang, C., Huang, Q., 2013. Spatial Cloud Computing: A Practical Approach. CRC Press.
Yang, E., Andre, T., Greenbowe, T.J., Tibell, L., 2003. Spatial ability and the impact of visualization/animation on learning electrochemistry. Int. J. Sci. Educ. 25, 329–349. https://doi.org/10.1080/09500690210126784
Zhang, B., Chen, Z., Peng, D., Benediktsson, J.A., Liu, B., Zou, L., Li, J., Plaza, A., 2019. Remotely sensed big data: evolution in model development for information extraction. Proc. IEEE 107, 2294–2301. https://doi.org/10.1109/JPROC.2019.2948454