Environmental Modelling & Software 26 (2011) 822–827

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Short communication

Ten guidelines for effective data visualization in scientific publications

Christa Kelleher*, Thorsten Wagener

Department of Civil and Environmental Engineering, The Pennsylvania State University, 212 Sackett Building, University Park, PA 16802, USA

ARTICLE INFO

Article history:
Received 20 April 2010
Received in revised form
10 December 2010
Accepted 15 December 2010
Available online 19 January 2011

Keywords:
Data visualization
Scientific visualization
Visual analytics

ABSTRACT

Our ability to visualize scientific data has evolved significantly over the last 40 years. However, this advancement does not necessarily alleviate many common pitfalls in visualization for scientific journals, which can inhibit the ability of readers to effectively understand the information presented. To address this issue within the context of visualizing environmental data, we list ten guidelines for effective data visualization in scientific publications. These guidelines support the primary objective of data visualization, i.e. to effectively convey information. We believe that this small set of guidelines based on a review of key visualization literature can help researchers improve the communication of their results using effective visualization. Enhancement of environmental data visualization will further improve research presentation and communication within and across disciplines.

1. Introduction

Visualization is one of the most important components of research presentation and communication due to its ability to synthesize large amounts of data into effective graphics (Ware, 2000). It is easier for the brain to comprehend an image versus words or numbers (Cukier, 2010), making effective graphics an especially important part of academic literature. The increasing accessibility and quantity of data (Cukier, 2010; Szalay and Gray, 2006) requires effective ways to analyze and communicate the information that datasets contain in simple, easy-to-understand formats. Visualization serves two major purposes, data analysis (Rebolj and Sturm, 1999; Jeong et al., 2006; Kollat and Reed, 2007; Wagener and Kollat, 2007; Xu et al., 2010) and data presentation. The latter is the focus of this paper, assuming that analysis is completed.

Data visualization refers to any graphic that examines or communicates data in any discipline (Few, 2009), whereas scientific visualization is a term that describes visualization of physical and scientific data (Card et al., 1999). As a field of research, scientific visualization explores the effectiveness of different types of graphics to display data. Despite interdisciplinary research advancements in recent years, common pitfalls in scientific visualizations do remain and regularly limit the effective communication through graphics. The topic of visualization has been explored in a range of books (Cleveland, 1994; Ware, 2000; Spence, 2001; Few, 2004b; Tufte, 2006; Strange, 2007) and journal articles, where the scientific visualization discussion can be either discipline-specific (e.g. Puhan et al., 2006), general (Kosslyn and Chabris, 1992), or written from a theoretical or psychological perspective (e.g. Spence and Lewandowky, 1991; Cleveland and McGill, 1984, 1987; Kosslyn, 1989). In this commentary, we primarily survey books on information or scientific visualization for helpful guidelines, as these books represent comprehensive surveys of basic guidelines for scientific visualization.

The ten guidelines summarized here represent a general list of suggestions that can enhance the effectiveness of scientific visualization across a range of disciplines. The guidelines are intended to address common pitfalls or provide simple ideas to be used by researchers when creating graphics for publications or presentations.

2. Ten guidelines

The ten guidelines for effective data visualization are presented in Figs. 1(a), (b), and 2 and discussed in detail below. Each guideline contains references to books or to journal articles which contain more information and specific examples of each issue. In the context of this paper, we intend the term 'guideline' to be a general principle that can be applied most of the time, but to which there are exceptions.

2.1. Guideline 1: create the simplest graph that conveys the information you want to convey (Tufte, 1983 [pp. 91–137])

The reason for including a graphic in a scientific publication is to explain something or to support an argument. Redundant plot attributes or excess ink can overcomplicate displays and confuse the plot's purpose (Tufte, 1983 [p. 93]). To simplify visualizations, remove redundancy in properties, while ensuring that the reader

* Corresponding author. Tel.: +1 503 913 8953; fax: +1 814 863 7304.
E-mail address: cak307@psu.edu (C. Kelleher).

1364-8152/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2010.12.006

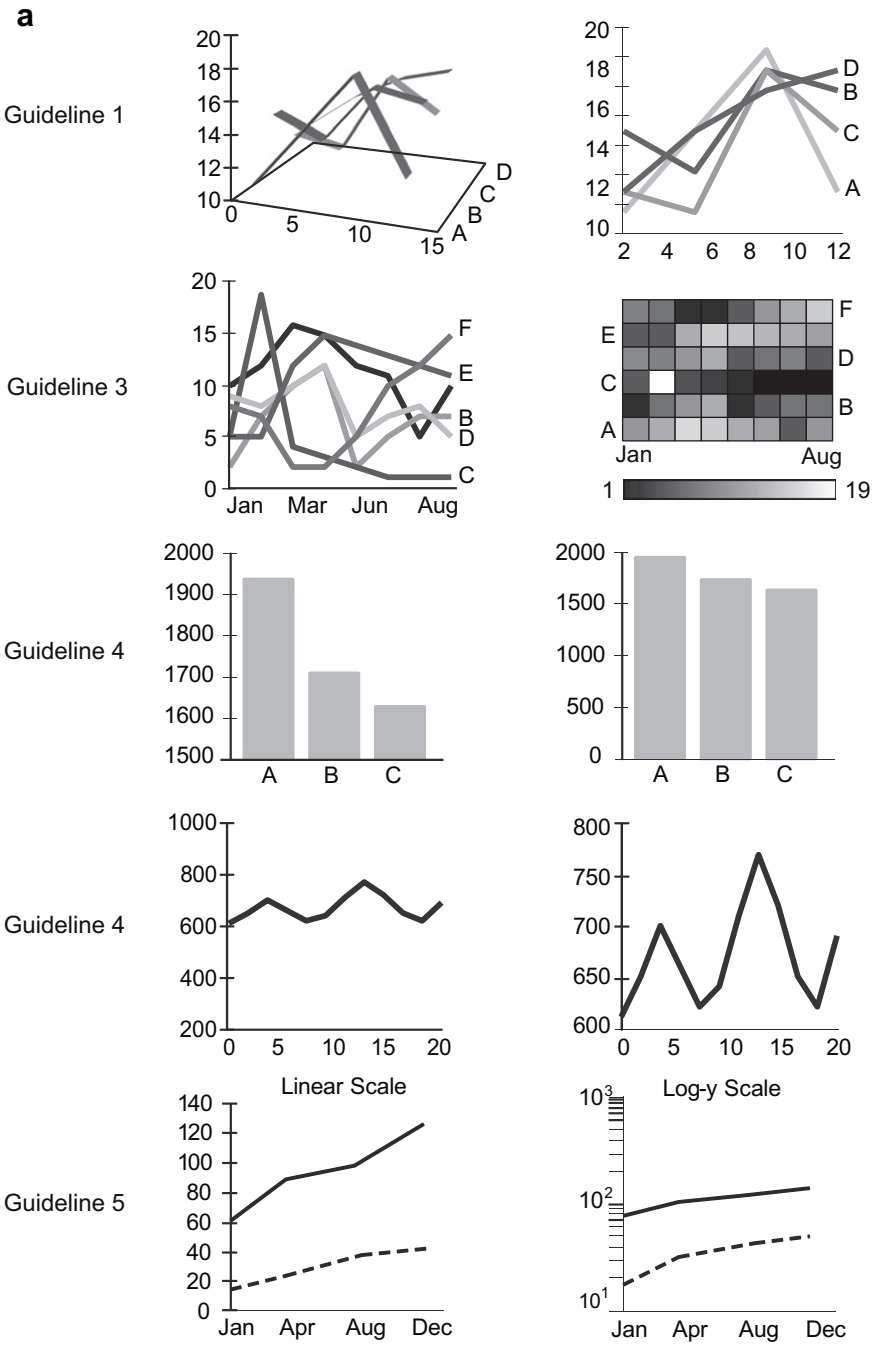

Fig. 1. Visual examples of the guidelines, excluding guideline 2. (a) illustrates guidelines 1, 3, 4 and 5. (b) illustrates guidelines 6 through 10.

can discriminate between the different visualization properties, such as shape, color, and thickness (Cleveland, 1984). Graph simplicity can be improved by minimizing the so-called '*data-ink ratio*', defined as the amount of ink used to present non-redundant information (or 'data ink') versus the total ink of the graphic (Tufte, 1983 [p. 93, 96]).

The visualization ability of software packages can tempt the user to use impressive rather than sensible plots. Three-dimensional graphics for example are often not helpful for visualization (though they can be helpful during analysis) because they make it difficult to compare datasets and to distinguish values (Few, 2004b [p. 170]). When necessary, multi-dimensional data can be visualized in 2D space by changing colors, shapes, and sizes to represent other data dimensions (e.g. contour plots) or by slicing the dataset, though too much variation can overcomplicate the plot. Other alternatives for displaying multi-dimensional data are coplots (Cleveland, 1994), which visualize three variables in 2D space, and scatter plot matrices, which display a matrix of 2D scatter plots for any number of variables (Cleveland, 1994).

2.2. Guideline 2: consider the type of encoding object and attribute used to create a plot (Chambers et al., 1983 [pp. 137–140]; Cleveland and McGill, 1984)

Graphical encoding objects (points, lines, and bars) and their value-encoding attributes (point position, line length, color) are

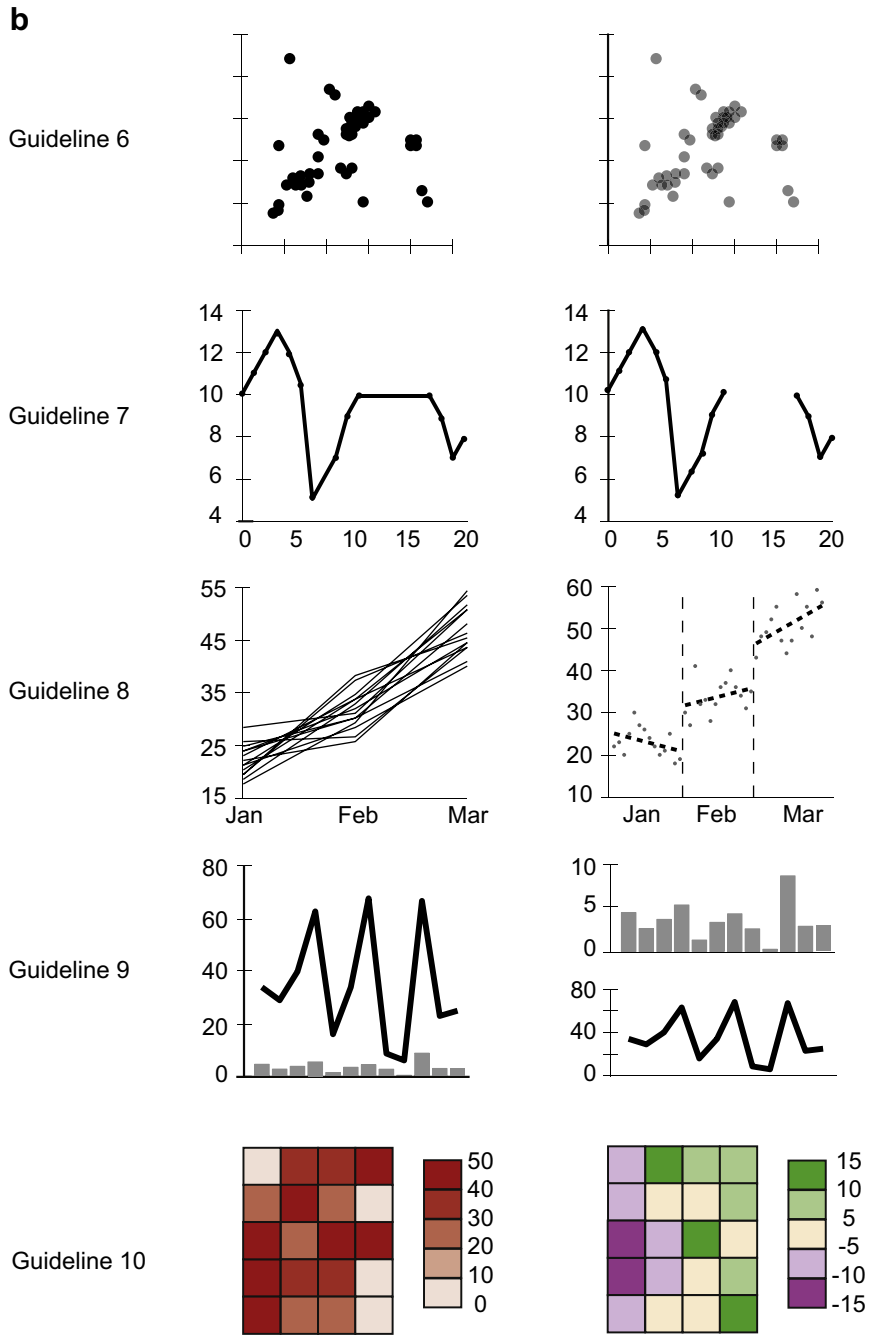

Fig. 1. (*continued*).

used to display different pieces of information from a dataset (see Fig. 2). Selecting attributes to use within a plot is especially important, because humans can quantify certain graph attributes better than others (Cleveland and McGill, 1984). Both length and position (2D) are better quantitatively perceived than other attributes, meaning that the data values that they represent and how those values compare to other values are easily determined (Cleveland and McGill, 1984). These types of attributes should be used when displaying the actual values of a dataset is important. Attributes that are difficult to quantitatively perceive, like line width, color hue or tint, or marker size (area), should be used for plots that show relative comparisons or general patterns (Cleveland and McGill, 1984).

2.3. Guideline 3: focus on visualizing patterns or on visualizing details, depending on the purpose of the plot (Few, 2004a; Kosslyn and Chabris, 1992)

A basic choice when selecting a plot is between displaying patterns or details. This choice requires the selection of a type of plot as well as the objects used to encode values within the plot (see Guideline 2). When searching for patterns, heatmaps (Wilkinson and Friendly, 2009) or bubble plots (Few, 2009 [p. 159]) can be effective even though extracting actual differences between values is difficult. Bar or line graphs should be used when individual values are important, as length and position are easily quantitatively perceived (Cleveland and McGill, 1984). A comparison between

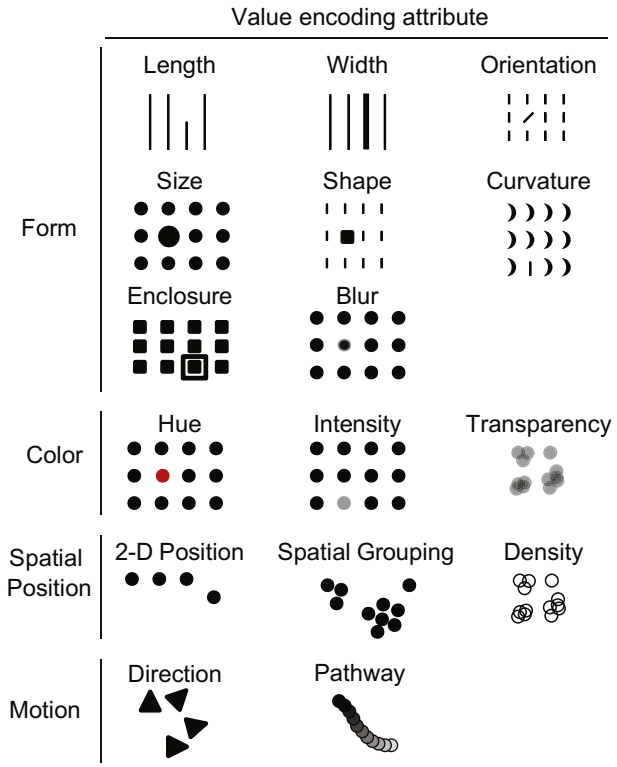

Fig. 2. The following attributes corresponding to different encoding objects can be used in data visualization to highlight contrasts between different parts of a graphic, to represent quantitative differences, or to categorize and group information. Adapted from Few (2009, p. 39).

a detail plot, shown as a line graph (left) and a pattern plot, represented by a heatmap (right) is illustrated in Fig. 1(a). A heatmap represents single or multiple datasets using a sequence of squares, where each square represents a data value and color represents the data point's magnitude. Pattern plots usually make use of color scales, which can strongly influence graph perception.

An alternative to heatmaps is horizon graphs (e.g. www.panopticon.com), which display multiple time-series in parallel. Horizon graphs are similar to a time-series plot, but use color to highlight differences and extreme values within and across time-series (Heer et al., 2009). While horizon graphs may convey more information, they are better used for data analysis and comparison since their effective use requires some level of familiarity with this type of plot.

2.4. Guideline 4: select meaningful axis ranges (Robbins, 2005 [pp. 239–241; 285]; Tufte, 2006 [p. 60]; Strange, 2007 [p. 89])

Selecting a range for the vertical axis depends on a graph's purpose and type. When absolute magnitudes are important, the vertical axis should begin at zero (Robbins, 2005 [pp. 239–241]; Strange, 2007 [p. 89]). Displaying data along a vertical axis that does not include zero misrepresents the data range and exaggerates the relative magnitude between values. This can be seen in Fig. 1(a), where the bar chart that begins at a nonzero value (left) shows a much larger difference between the three values compared to the bar chart with an axis that begins at zero (right). Displaying the absolute magnitude of the datasets, referenced to zero, ensures that the relative difference in size represented by the graphic matches the actual relative difference in values. Bar charts, Cleveland dot plots (Cleveland, 1994), histograms, and line or point plots where absolute magnitudes are important represent plots that should include zero on the vertical axis. Dot plots, which use a dot to visualize what would otherwise be the top of the bar chart, are a good alternative to bar charts, especially for simultaneously visualizing multiple variables (Cleveland and McGill, 1984).

When relative magnitudes are important, plots should exhibit a 'lumpy profile' (Robbins, 2005 [p. 285]; Tufte, 2006 [p. 60]). A 'lumpy profile' is created by setting the limits on the plot as close as possible to the magnitude of the dataset range, and improves visualization of variability and eliminates white or wasted space. As shown in Fig. 1(a), it is easier to see variations in a dataset when the plot limits are closer to the data range (right) versus when they are further removed (left). Scatter plots can also appear more correlated when plotted at scales much greater than the dataset range (Cleveland et al., 1982). Depending on the dataset, the vertical axis can be plotted using a few percent more/less than the maximum/minimum values. This guideline is appropriate for visualization of line and scatter plots (Cleveland and McGill, 1984).

2.5. Guideline 5: data transformations and carefully chosen graph aspect ratios can be used to emphasize rates of change for time-series data (Cleveland, 1994 [p. 66, 95, 103])

Visualizing rate of change of a time-series, which refers to the difference in values between time steps, can be enhanced or hindered by vertical axis transformations. The decision to use a transformation should depend on the dataset(s) and the intent of the plot, as transformations can change the impression of a graphic and hence the information conveyed.

Plotting on a logarithmic vertical axis can remove skewness in datasets with ranges that include very large and small values (Cleveland, 1994). Logarithmic scales also visualize the rate of change normalized to an initial value, such that the slope of the dataset represents the percentage change of a slope or trend. As is shown in Fig. 1(a), the linear plot (left) illustrates a larger rate of change between January and April for the solid line versus the dashed line. However, a vertical log scale transformation shows that the percentage rate of change between January and April is larger for the dashed line versus the solid line (Fig. 1(a) right). Rate of change comparisons can also be achieved by transforming the dataset (e.g. normalizing values to a mean). Different transformations highlight distinct aspects of the data, and are helpful in different contexts.

The aspect ratio or shape parameter of the graph, which is the ratio of a graph's height to width, can also improve visualization of rates of change (Cleveland and McGill, 1987; Cleveland et al., 1988; Strange, 2007 [p. 147]). Rates of change are judged based on the slope of graphed data, which is a function of the shape of the graph (Cleveland et al., 1988). As a rule of thumb, Cleveland (1994) recommends selecting an aspect ratio by 'banking to 45°', a rule that sets a graph's aspect ratio by banking line segments of a graphed dataset to 45°. Refer to Cleveland (1994) for a discussion of how to apply this rule to different datasets.

2.6. Guideline 6: plot overlapping points in a way that density differences become apparent in scatter plots (Few, 2009 [p. 121]; Cleveland, 1994 [p. 159])

In scatter plots where points are opaque (Fig. 1(b) left), density differences are obscured or even invisible as multiple points plotted in the same location appear as one point. Changing plotted points from opaque to transparent (Fig. 1(b) right) enhances the information conveyed by visualizing density differences. An alternative strategy to achieve a similar effect is to plot unfilled circles (Cleveland, 1994 [p. 159]). For large datasets, density may be better visualized by decreasing point size. The magnitude of transparency will depend on figure content and publisher requirements.

Densities can also be visualized via kernel density estimates, which are generalized probability density functions (pdf) that plot data value versus likelihood (Jones et al., 1996; Delaigle and Gijbels, 2004). Histograms represent a special case of kernel density estimation, with a uniform bandwidth equal to the histogram interval length (Scott, 1992). Though kernel density estimates yield continuous pdfs and avoid problems with bin size, they are subjective to the type of data and selection of the smoothing parameter (Jones et al., 1996; Delaigle and Gijbels, 2004).

2.7. Guideline 7: use lines when connecting sequential data in time-series plots (Strange, 2007 [p. 150])

Plots that connect non-sequential data or values on either side of a period of missing data with a line imply a linear change between the points. This rule is illustrated at the bottom of Fig. 1(b), where a data gap in the time-series is connected with a line (left) and without a line (right). Without indicating the missing period of data, the left panel implies that there is no change in value between $t = 10$ and $t = 15$. Non-sequential data, such as points on a scatter plot or categorical data, data that can be separated into groups and is often qualitative (e.g. location, method, type of plant or animal), should also not be connected with lines.

2.8. Guideline 8: aggregate larger datasets in meaningful ways (Cleveland and Devlin, 1980; Chambers et al., 1983 [pp. 21–24]; Cleveland, 1994 [p. 187])

Simplicity can be difficult to achieve in displays of large sets of quantitative or categorical data. Large quantitative datasets can be simplified via summary plots such as box-and-whisker plots (Chambers et al., 1983 [pp. 21–24]) or through kernel smoothing strategies (Scott, 1992). Dataset characteristics, usually a combination of quantitative and categorical data, can be displayed using Cleveland dot plots (Cleveland, 1994 [pp. 150–151]) or linked micromap plots, which present data referenced to a map or location (Robbins, 2005 [pp. 136–137]), though the ordering of such data can influence the perception of individual points (Cleveland, 1994 [p. 15]). Some summary plots such as pie charts should be avoided altogether, as it is difficult to perceive differences in angles in this particular case (Cleveland, 1994 [p. 262]; Strange, 2007 [p. 85]).

For long time-series, temporal aggregation, averaging values across a larger time step (e.g. averaging daily data to monthly) can be used to reduce the number of data points, but sacrifices data resolution. Cycle plots are one alternative to a traditional time-series plot (Fig. 1(b) left) that preserve data resolution and display trends at a repeating time interval (Cleveland and Devlin, 1980; Cleveland, 1994 [p. 187]). A cycle plot (Fig. 1(b) right) graphs time-series data that repeats at some interval at two different timescales, like monthly data taken across many years, to visualize a long-term trend (e.g. annual) and a short-term trend (e.g. monthly) (Robbins, 2005).

2.9. Guideline 9: Keep axis ranges as similar as possible to compare variables (Cleveland, 1994, [pp.86–87]; Few, 2009 [p. 180])

Display of variables across subplots with different axis ranges hinders comparison of range and variability across datasets (Fig. 1 (b) left). By maintaining the same axis ranges, the datasets can be more easily compared (Fig. 1(b) left). Separating variables with large scale differences into subplots (Fig. 1(b) right) highlights variability within individual datasets, whereas variables with similar ranges can be grouped together in a single plot. Maintaining vertical or horizontal axis ranges across subplots or combining plots for multiple variables enhances data comparison and eliminates misrepresentation of relative differences between data series.

2.10. Guideline 10: select an appropriate color scheme based on the type of data (Brewer, 1994; Harrower and Brewer, 2003)

Using a color scheme that matches the type of data will further support the purpose of a plot. Sequential schemes, made up of intervals of one or two colors graduating from light to dark, should be used for quantitative data, with low values in lighter tints and high values in darker tints (Fig. 1(b) left) (Harrower and Brewer, 2003). Diverging color schemes on the other hand should be used to highlight contrasts between low and high values relative to an average value (Fig. 1(b) right) (Harrower and Brewer, 2003). Diverging schemes use a light, neutral color to represent average values and contrasting dark hues for low and high values (Harrower and Brewer, 2003). Categorical data is best represented with qualitative schemes, which are made up of contrasting colors that show differences without reference to magnitude (Harrower and Brewer, 2003). A range of publications discusses the importance of color scale selection (e.g. Ware, 2000; Light and Bartlein, 2004; Stephenson, 2005; Stone, 2006). Software tools are also available to assist with color scale generation for different schemes (http://colorbrewer2.org/).

3. Discussion and conclusions

The objective of any graphic in the context of scientific publications and presentations is to effectively convey information. The ten guidelines proposed here represent an effort to reduce common pitfalls in the pursuit of this objective. Above all, these guidelines should be taken as general recommendations that can be used to improve visualization design, and not as absolute rules that apply in every case. Adhering to these recommendations will generally improve the presentation of scientific data, and subsequently, the communication of research outcomes. In addition to the guidelines discussed above, there are other good practices that should be generally adhered to. A main practice that has not yet been generally accepted is the inclusion of uncertainty estimates or error bars in the visualization of both observed and modeled data. It has been argued and demonstrated in many places that such estimates improve decision-making and provide a better reflection of our scientific understanding (e.g. Reichert and Borsuk, 2005; Beven, 2006).

Acknowledgements

Support for this research was provided by an EPA STAR Early Career Award (R834196) to Thorsten Wagener and an EPA STAR Graduate Fellowship to Christa Kelleher. The authors would like to thank Joe Kasprzyk, Keith Sawicz, and Riddhi Singh for help with manuscript revisions and Kevin McGuire for literature suggestions. The authors would also like to acknowledge the comments of Lucy Marshall, Felix Andrews, and a third anonymous reviewer that assisted manuscript expansion and improvement.

References

Beven, K.J., 2006. On undermining the science? Hydrol. Proc. 20, 3141–3146.
Brewer, C.A., 1994. Color use guidelines for mapping and visualization. In: MacEachren, A.M., Taylor, D.R.F. (Eds.), Visualization in Modern Cartography. Elsevier Science, Tarrytown, NY, pp. 123–127.
Card, S.K., Mackinlay, J.D., Shneiderman, B., 1999. Readings in Information Visualization: Using Vision to Think. Academic Press, San Diego, CA.
Chambers, J.M., Cleveland, W.S., Kleiner, B., Tukey, P.A., 1983. Graphical Methods for Data Analysis. Duxbury Press, Boston, MA.
Cleveland, W.S., 1984. Graphs in scientific publications. Am. Stat. 38 (4), 261–269.

Cleveland, W., 1994. The Elements of Graphing Data, second ed. Hobart Press, Summit, NJ.
Cleveland, W.S., Devlin, S.J., 1980. Calendar effects in monthly time series: detection by spectrum analysis and graphical methods. J. Am. Stat. Assoc. 75 (371), 487–496.
Cleveland, W.S., Diaconis, P., McGill, R., 1982. Variables on scatterplots look more highly correlated when the scales are increased. Science 216 (4), 1138–1141.
Cleveland, W.S., McGill, R., 1984. Graphical perception: theory, experimentation, and application to the development of graphical methods. J. Am. Stat. Assoc. 79 (387), 531–554.
Cleveland, W.S., McGill, R., 1987. Graphical perception: the visual decoding of quantitative information on graphical displays of data. J. R Stat. Soc. Ser. A 150 (3), 192–229.
Cleveland, W.S., McGill, M.E., McGill, R., 1988. The shape parameter of a two-variable graph. J. Am. Stat. Assoc. 83, 289–300.
Cukier, K., 2010. A special report on managing information. The Economist 394 (8671), 3–18.
Delaigle, A., Gijbels, I., 2004. Practical bandwidth selection in deconvolution kernel density estimation. Comput. Stat. Data Anal. 45, 249–267.
Few, S., 2004a. Eenie, Meenie, Minie, Moe: Selecting the Right Graph for Your Message. Intelligent Expertise. Perceptual Edge. Available at <http://www.perceptualedge.com/articles/ie/the_right_graph.pdf> (accessed on 3.08.10.).
Few, S., 2004b. Show Me the Numbers: Designing Tables and Graphs to Enlighten. Analytics Press, Oakland, CA.
Few, S., 2009. Now You See It. Analytics Press, Oakland, USA.
Harrower, M., Brewer, C., 2003. ColorBrewer.org: an online tool for selecting colour schemes for maps. Cartog J. 40 (1), 27–37.
Heer, J., Kong, N., Agrawala, M., 2009. Seizing the horizon: the effects of chart size and layering on the graphical perception of time series visualizations. ACM Hum. Factor. Comput. Syst. (CHI), 1303–1312.
Jeong, S., Liang, Y., Liang, X., 2006. Design of an integrated data retrieval, analysis, and visualization system: application in the hydrology domain. Environ. Model. Software 21 (12), 1722–1740.
Jones, M.C., Marron, J.S., Sheather, S.J., 1996. A brief survey of bandwidth selection for density estimation. J. Am. Stat. Assoc. 91 (433), 401–407.
Kollat, J.B., Reed, P., 2007. A framework for visually interactive decision-making and design using evolutionary multi-objective optimization (VIDEO). Environ. Model. Software 22 (12), 1691–1704.
Kosslyn, S.M., 1989. Understanding charts and graphs. Appl. Cognit. Psychol. 3, 185–226.
Kosslyn, S.M., Chabris, C.F., 1992. Minding information graphics. Folio, 69–71.
Light, A., Bartlein, P.J., 2004. The end of the rainbow? Color schemes for improved data graphics. EOS 85 (40), 385, 391.
Puhan, M.A., ter Riet, G., Eichler, K., Steurer, J., Bachmann, L.M., 2006. More medical journals should inform their contributors about three key principles of graph construction. J. Clin. Epidemiol. 59, 1017–1022.
Rebolj, D., Sturm, P.J., 1999. A GIS based component-oriented integrated system for estimation, visualization and analysis of road traffic air pollution. Environ. Model. Software 14, 531–539.
Reichert, P., Borsuk, M.E., 2005. Does high forecast uncertainty preclude effective decision support? Environ. Model. Software 20, 991–1001.
Robbins, N., 2005. Creating More Effective Graphs. Wiley-Interscience, Hoboken, NJ.
Scott, D.W., 1992. Multivariate Density Estimation: Theory, Practice, and Visualization. John Wiley, New York, NY.
Spence, I., Lewandowsky, S., 1991. Displaying proportions and percentages. Appl. Cognit. Psychol. 5, 61–77.
Spence, R., 2001. Information Visualization. ACM Press, New York, NY.
Stephenson, D.B., 2005. Comment on “Color schemes for improved data graphics”. EOS 86 (20), 196.
Stone, M., 2006. Choosing colors for data visualization. Business Intelligence Network Available at: <http://www.perceptualedge.com/articles/b-eye/choosing_colors.pdf> (accessed on 3.08.10.).
Strange, N., 2007. Smoke & Mirrors: How to Bend Facts & Figures to Your Advantage. A & C Black Publishers, London, UK.
Szalay, A., Gray, J., 2006. 2020 Computing: science in an exponential world. Nature 440, 413–414.
Tufte, E.R., 1983. The Visual Display of Quantitative Information. Graphics Press, Cheshire, CT.
Tufte, E.R., 2006. Beautiful Evidence. Graphics Press, Cheshire, CT.
Wagener, T., Kollat, J., 2007. Visual and numerical evaluation of hydrologic and environmental models using the Monte Carlo Analysis Toolbox (MCAT). Environ. Model. Software 22 (7), 1021–1033.
Ware, C., 2000. Information Visualization: Perception for Design. Morgan Kaufmann Publishers, San Francisco CA.
Wilkinson, L., Friendly, M., 2009. The history of the cluster heat map. Am. Stat. 63 (2), 179–184.
Xu, B., Lin, H., Chiu, L.S., Tang, S., Cheung, J., Hu, Y., Zeng, L., 2010. VGE-CUGrid: an integrated platform for efficient configuration, computation, and visualization of MM5. Environ. Model. Software 25 (12), 1894–1896.