

EVALUATING GENERALIZATIONS OF HYDROGRAPHY IN DIFFERING TERRAINS FOR *THE NATIONAL MAP OF THE UNITED STATES*

Cynthia A. Brewer
Department of Geography
302 Walker Building, Pennsylvania State University
University Park, PA 16802, USA
cbrewer@psu.edu

Barbara P. Battenfield
Department of Geography, UCB 260
University of Colorado—Boulder
Boulder, CO 80309, USA
babs@colorado.edu

E. Lynn Usery
Center of Excellence for Geospatial Information Science
U.S. Geological Survey,
1400 Independence Rd., Rolla, MO 65401, USA
usery@usgs.gov

Abstract

Introduction: This collaborative effort between the U.S. Geological Survey's (USGS) Center of Excellence for Geospatial Information Science (CEGIS) and two universities addresses research problems in generalization and electronic topographic map design. The work supports the USGS web-based data delivery service called *The National Map*, and most of the geospatial data available in the service is intended for display scales ranging from 1:20,000 to 1:1,000,000 (20K – 1M).

Objectives: Our work is based on two premises. First, a single generalization sequence with uniform tolerance parameters will not produce generalized hydrography data of consistent quality across the country. Second, automating as much as possible of the generalization processing will reduce workloads for the national mapping agency and improve consistency and quality of the results.

Methodology: We have generalized four hydrographic subbasins of the USGS National Hydrography Dataset (NHD) to demonstrate that knowledge about terrain and climate can inform decisions about generalization processing. Work on four additional subbasins is in progress. We work with symbol redesign and elimination decisions for feature classes (display changes) throughout the scale range of 20K to 1M and produce generalized Level of Detail database versions (LoDs) only at scales where display changes fail to produce successful topographic map displays. Results

are evaluated numerically and visually in the context of full-featured digital topographic maps.

Results: We find that the subset of algorithms selected and parameters which guide them vary by landscape regime. Likewise, the sequence of generalization processing is different for different regions. An upstream drainage area attribute plays an important role for selecting major centerlines, for collapsing and labeling wide river channels (area polygons), and for symbolizing all flowlines with graduated line symbols.

Conclusions: As in other countries that have large geographic extents, variations in physiography and climate have differential impact on the amount and types of detail in data that represent the landscape on topographic maps. Hydrographic generalization solutions must characterize rather than homogenize these differences.

Introduction

The emergence of national and global spatial data infrastructures has improved abilities to manage and monitor natural resources and human infrastructure, to analyze and understand environmental processes, and to anticipate and mediate consequences of natural hazards, storms and disasters. Integrating data models and data infrastructures is a necessary prerequisite to numerical analyses and GIS-based modeling (Stoter et al. 2004) and to services such as *The National Map*, a web-based geospatial data delivery program for the United States.

The National Map program is lead by the U.S. Geological Survey (USGS) and integrates data from local, state, federal, and additional sources. Recently, new versions of selected USGS topographic maps are available for download as GeoPDF files in traditional 7.5-minute quadrangle format using current data. In 2009, the features for maps in this new program, *Digital Map – Beta*, include orthoimagery, roads, and geographic names. Hydrography and contours will be added as current data are improved. GeoPDF quadrangle files may be downloaded free using the *Map Locator* link on this page of *The National Map* site: nationalmap.gov/digital_map.

A critical prerequisite for integrating data infrastructures is to generalize and manage data representations across multiple scales. Data modeling has important implications for cartographic design and analysis, since users of *The National Map* expect that downloaded data will "fit together" graphically. Analytical uses of downloaded data carry additional needs and requirements, to support reliable data measurements at any resolution, and to insure that features integrate horizontally (within layers) as well as vertically (between layers) (Bobzien et al. 2008; Buttenfield 2006, 1999; Spaccapietra et al. 2000).

The research reported in this paper builds upon recent work examining the interaction of symbol design and feature generalization (Brewer and Buttenfield 2009, 2007;

Brewer et al. 2007). Brewer and Bittenfield have developed a conceptual tool called ScaleMaster (www.ScaleMaster.org) for cataloging design changes and database changes across a wide range of map scales, such as 1:5,000 to 1:1,000,000 (5K – 1M). Mackaness et al. (2007) report current research from around the world on multi-representation databases (MRDB) and generalization operators. European national mapping agencies (NMAs) are making progress on integrating topographic data across boundaries and through scales, and much of the research reported in the Mackaness et al. volume is based on European MRDB progress.

Our work presents an approach more suited to the United States NMA context, which differs in several respects to many European NMAs. First, USGS compiles data at multiple scales, rather than working from an extremely detailed database compiled at large scale to map the whole country at smaller scales. Second, some databases are compiled exclusively by the NMA. Others, including the National Hydrography Dataset (NHD) utilized in this project, comprise contributions by field workers from many agencies and organizations. Third, the sheer size of the target mapping area (contiguous U.S. plus Alaska, Hawaii, and Puerto Rico) results in distinct physiographic and climatic differences which must be preserved during generalization and data modeling. This third point forms a core emphasis in the work reported in this paper, because no single data processing strategy will generalize data nationwide with uniform success.

We prototype and test automated data generalization methods from finer to coarser resolutions, to meet the demands for multi-scale mapping and analysis. We investigate efficient methods to design, to implement, and to evaluate automated methods for cartographic generalization of datasets for delivery by *The National Map*. We also work toward redesign of U.S. topographic mapping symbology to produce attractive and effective copyright-free digital mapping products for the entire country.

Multi-Scale Hydrography

The majority of the research effort reported in this paper focuses on generalizing hydrography, because it is the most scale-sensitive of all vector layers. As such, it is expected to generate the most difficult data modeling problems in generalizing vector data for *The National Map* (Bittenfield et al. 2008). Hydrography is a vector layer commonly utilized in cartographic base mapping, and will be in high demand by users of *The National Map*. As of 2009, hydrography data are not yet served with GeoPDF maps offered in the *Digital Map – Beta* program, attesting to the sensitivity of these in quality cartography.

For recent work, we designed, implemented, and tested automatic generalization methods to accommodate topographic mapping at a range of scales from 20K to 200K. Because generalization is computationally intensive, it is inefficient to process "on-the-fly." We store the processing output as intermediate scale hydrographic datasets, called Level of Detail (LoD) databases, in correct NHD format, at several

output scales. During the first year, we created draft LoDs for mapping at scales of roughly 50K to 200K. The term "roughly" applies to complicating factors. High-resolution NHD data varies in source scale from 4.8K (1:4,800) down to 24K, but for modeling purposes, we assumed input data to be a uniform source scale of 24K. Our work demonstrates that symbol design and elimination of features (display changes) support scale changes from 24K down to 50K. LoD development is intended for scale jumps of approximately four to five times the larger scale. For example, hydrography data which has been modeled suitably for 50K should be appropriate for mapping and analysis down to 200K. For smaller scales, additional generalizations will be required to maintain analytic and cartographic validity. This thinking led us to choose initial target scales for LoDs at 50K, 200K, and 1M.

We were challenged to automate as much geoprocessing as possible, and to minimize subsequent manual editing to 15 hours per quad sheet, as specified by USGS (Binder and Cooley, 2008). We utilized procedures that are readily available either in existing GIS software such as ESRI ArcGIS, or in existing code written in-house at USGS. Once tested, our generalization strategies are coded as ESRI ModelBuilder sequences; after final validation, we port the code to Python scripts.

Sampling U.S. Landscapes

We generalized four subbasins of the high-resolution National Hydrography Dataset (NHD) selected to crosstabulate dry-humid climates with flat-hilly-mountainous terrains (Table 1). Subbasins were selected by USGS staff (Stanislawski et al. 2007) for a previous study to estimate upstream drainage areas in varying physiographic landscapes.

<i>Subbasin name</i>	<i>State</i>	<i>NHD subbasin</i>	<i>Regime</i>	<i>Physiographic province and division</i>
Upper Suwannee	FL, GA	03110201	Flat Humid	Coastal Plain of Atlantic Plain
Lower Beaver	UT	16030008	Flat Dry	Basin and Range of Intermontane Plateaus
Pomme de Terre	MO	10290107	Hilly Humid	Ozark Plateaus of Interior Highlands
Lower Prairie Dog Town Fork Red	TX	11120105	Hilly Dry	Great Plains and Central Lowland of Interior Plains
South Branch Potomac	WV	02070001	Mountainous Humid	Valley and Ridge of Appalachian Highlands
Piceance-Yellow	CO	14050006	Mountainous Dry	Colorado Plateaus of Intermontane Plateaus

Table 1. Four NHD subbasins processed during the first year of the project are highlighted.



Figure 1. Physiographic divisions (white outlines) and provinces (varied color fills) of the contiguous United States with overlay of NHD basin outlines in gray. The extent of NHD subbasins for the research project are shown by black fills. Two urban subbasins (St. Louis and Atlanta) and two additional rural subbasins (Texas and Colorado) that will be added to the project are labeled in gray text. A group of four Iowa subbasins (North Raccoon, South Raccoon, Middle Des Moines, and Lake Red Rock) are the focus of related research reported in Stanislawski et al. (2009).

Sampling from varied U.S. landscapes as we refine a national approach to generalization is important because the U.S. has widely varying geomorphology. One way to characterize this variation is using physiographic regions based on shared terrain texture, rock type, and geologic history and structure. Figure 1 summarizes the major physiographic divisions (outlined in white with upper-case labels) and provinces (in multiple colors) for the contiguous United States. The U.S. may be further divided at a third tier of detail with physiographic sections (not shown; see tapestry.usgs.gov). Gray outlines for all U.S. NHD basin boundaries are shown overlaid on the physiographic province colors in Figure 1.

We find that three generalization components must be varied with landscape regime (Buttenfield et al. 2008; Wendel et al. 2009). First, the algorithms and parameters must vary. For example, proximal waterbodies in a swamp region in the Florida subbasin are most effectively aggregated with an eye to preserving overall texture and

coverage, but contiguous polygons such as lakes and inundation areas in the Missouri subbasin require boundary simplification and topologic reintegration.

Second, the sequence of processing steps varies with terrain and climate. In principle, the generalization processing sequence prunes flowlines, delineates continuous centerlines, then simplifies flowlines. In practice, centerline delineation requires extra processing in dry regions (e.g., the Utah subbasin), where most channels are intermittent, sometimes are below ground level, and often are disconnected.

Generalization Methods

The generalization sequences we developed convert stream vectors into a geometric network, and prune on thresholds using the ratio:

$$\frac{[\text{Summed remaining stream channel length (kilometers)}]}{[\text{Summed original catchment area (kilometers squared)}]}$$

Pruning is necessary in some but not all landscape regimes. It is needed for example in humid areas that are hilly or mountainous, but not in flat humid subbasins. In related work presented in this proceedings (Stanislawski et al. 2009), pruning is applied to generalize four contiguous subbasins in Iowa which span the boundary between glacial till and a lakebed. Where stream channels display local density differences (Figure 2), the two density types must be isolated and pruned differentially, while retaining channel continuity. If this step is not included, the resulting generalization is characterized by stream networks of uniform density. Local density is an important component for both hydrologic analysis and hydrographic mapping. At present, the isolation of dense networks is accomplished manually, and selection of pruning tolerances is guided by Töpfer and Pillewizer's Radical Law (1966). Efforts to automatically scan a subbasin and detect localized changes in drainage density are underway.

As described above, the generalization strategy for any LoD is initially automated in a ModelBuilder environment. This permits fast prototyping, and easy refinement of tolerance parameters and overall sequencing. Figure 3 shows the 50K ModelBuilder for the Pomme de Terre subbasin, a humid and hilly region of Missouri. Starting from the upper left, two classes of flowline density have been isolated and pruned differentially. The pruned data are merged into a temporary layer. Inside the red box (enlarged in Figure 4) these flowlines are intersected with large water polygons to delineate stream centerlines, which will be symbolized by upstream drainage area, and which will be substituted for stream channel polygons at smaller mapping scales. Flowlines which do not intersect water polygons are retained for symbolization as intermittent or perennial streams.

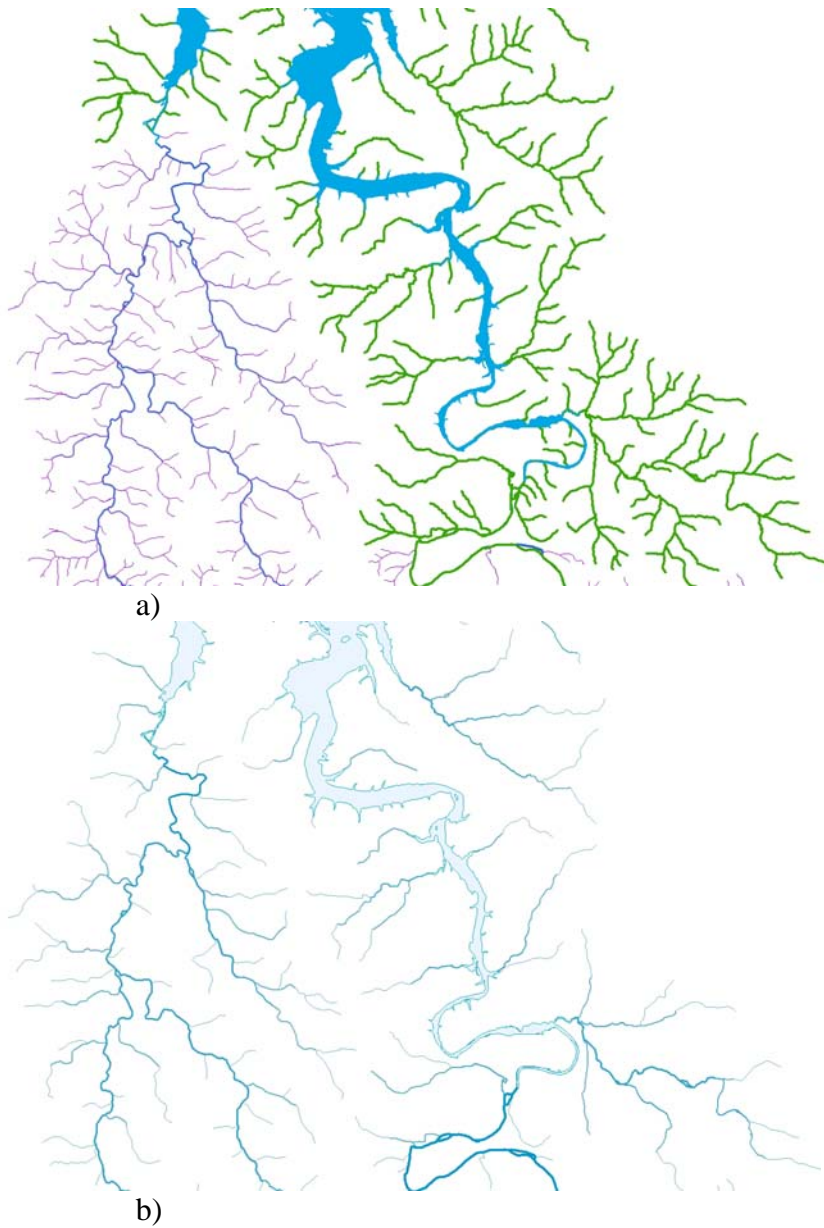


Figure 2. A section of the Pomme de Terre subbasin (Missouri) showing sparse and dense stream networks. Map a) identifies denser parts of the network (thin purple lines) that were selected and which retain channel connectivity. The less dense drainages are shown in thicker green lines (right side of Figure 2a). Map b) shows the 50K LoD with sparse and dense drainages pruned differentially. Lines are symbolized by width classes based on upstream drainage area. (Maps are exported at 100K and reduced to 48% for the paper.)

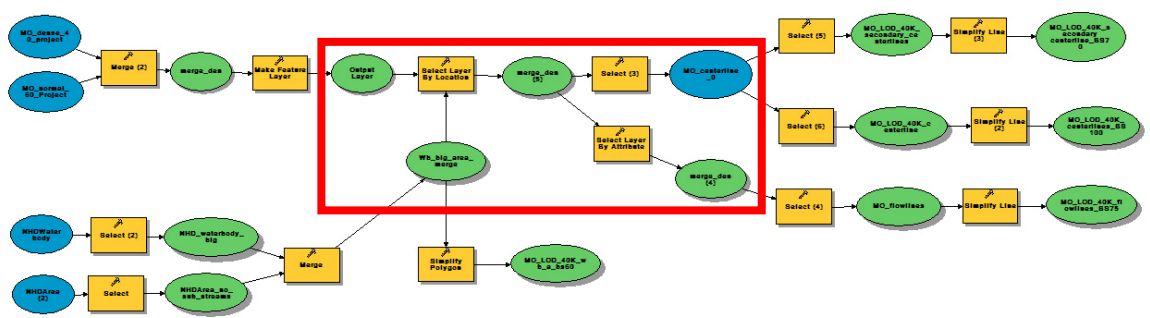


Figure 3. Missouri 50K LoD full Modelbuilder

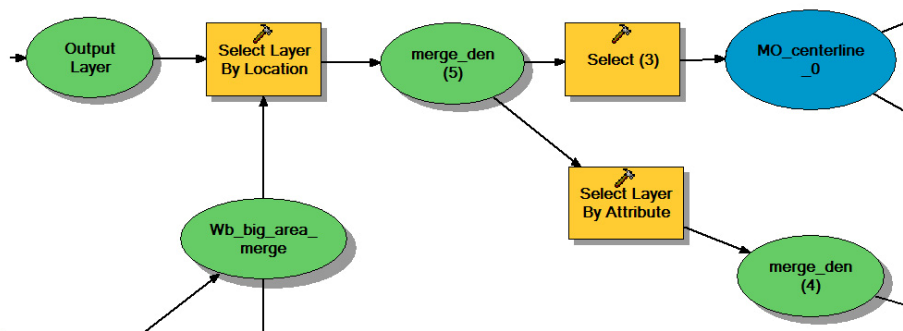


Figure 4. A portion of the Missouri 50K LoD Modelbuilder, highlighting the workflow where centerlines are formed and isolated from other flowlines.

The processing to the right of the red box (Figure 3) isolates named and unnamed stream channels, and simplifies all three flowline categories using the Bend Simplify algorithm developed by Wang and Muller (1998).

The lower part of the ModelBuilder diagram in Figure 3 shows selection of larger water polygons and elimination of submerged streams from the NHD areas feature class. Waterbodies and areas are merged prior to their intersection with flowlines described above. Waterbodies include lakes and ponds, while water areas include wide stream channels with two banks drawn. The merged feature class is subsequently simplified. Each simplification may be run to different tolerances, for example the flowlines, centerlines and secondary (unnamed) centerlines are simplified to 75, 100, and 75 meters, respectively.

Refinements to LoD production are underway and intended to refine the generalized data for mapping, as well as to further automate the processing. For example, a modification currently under investigation is to incorporate upstream drainage area into the delineation of centerlines to comprehensively protect topology. Current centerline solutions display occasional discontinuities which are difficult to predict and more difficult to resolve using our current methods. We also want to validate our final processing sequence (currently in the form of a Python script) by running it

against a different hilly-humid subbasin. This will help us to refine minimum size tolerances as well as simplification parameters. The original hydrography and 50K LoD results are compared in Figure 5.

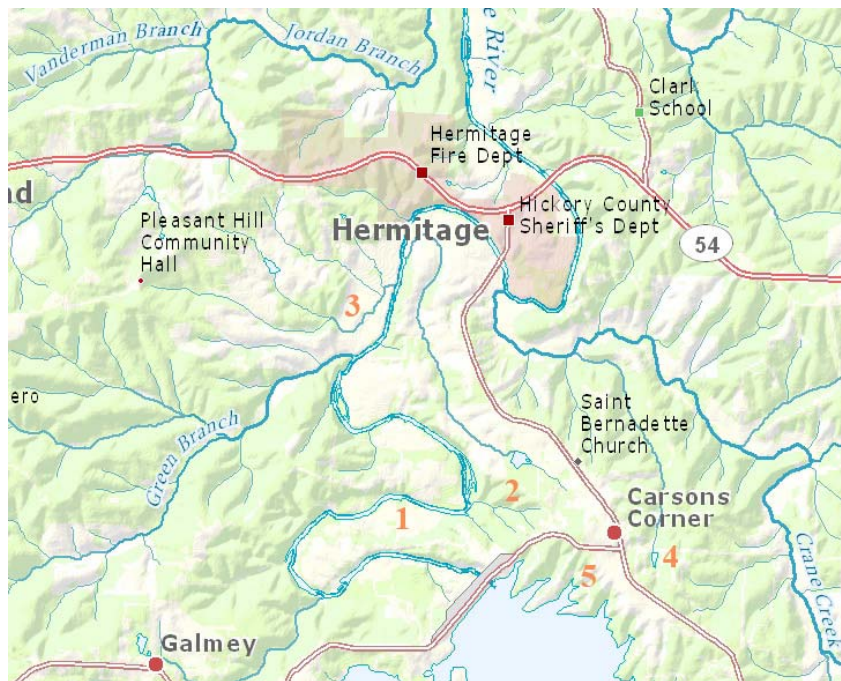
Map Design to Evaluate Generalization

The two maps in Figures 5 are identical except for hydrography. Note the collapse of the polygonal stream channels to centerlines (compare example noted by orange '1' in Figure 5a to 5b), fewer minor tributaries (orange 2), simplified flowlines (3), removal of small lakes (4), and simpler crenulations on reservoir shorelines (5). Stream centerlines are also used to position labels on elongated water polygons.

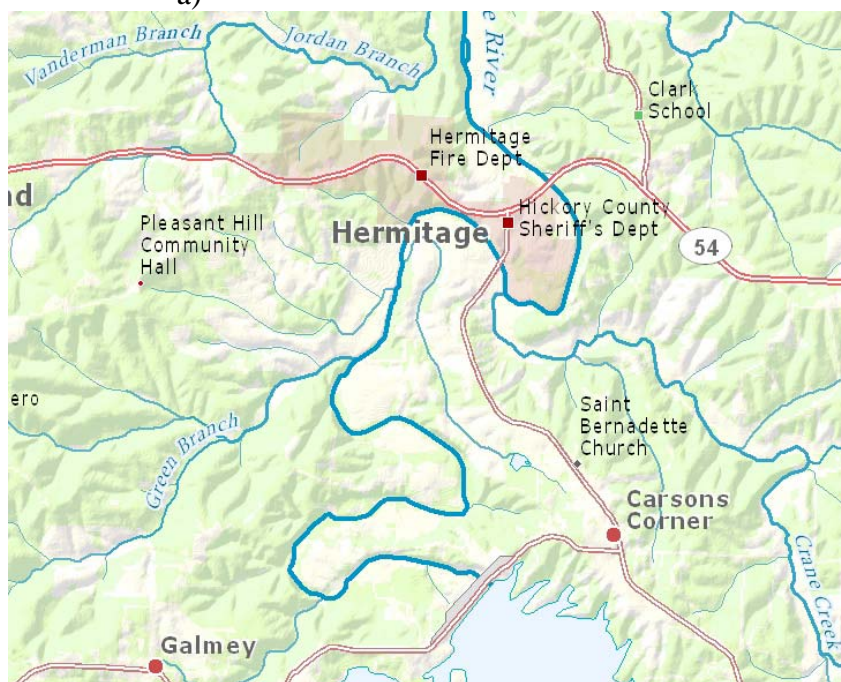
Flowlines on both example maps in Figure 5 are symbolized using five widths representing classes of upstream drainage area to produce a tapered appearance. Thus, small tributaries are fine and major centerlines are wide. Stanislawski et al. (2007) estimate upstream drainage area using Thiessen polygon analysis, and that estimate forms a surrogate for stream order suitable for cartographic production. The high-resolution NHD does not include stream order or upstream drainage area attributes in the files distributed by USGS because of the volume of data and short irregular update cycle (nhd.usgs.gov).

Methods of visual evaluation of map samples include confirmation of similar content and similar densities of hydrographic detail compared to existing topographic maps (printed U.S. topographic maps for the sample subbasins exist at 24K, 100K, and 250K, with widely varied dates). The appearance of draft map designs we have been using to evaluate generalization results are akin to current USGS symbols for printed maps (rockyweb.cr.usgs.gov/nmpstds/acrodocs/qmaps/5psym202.pdf) because we are comparing generalization results to existing topographic maps. Visual evaluations also include judging vertical integration of generalized hydrography with other layers, such as transportation and terrain (represented with contours and hillshading).

Map design affects generalization decisions in multiple ways. For example, features represented with coarse pattern symbols without visible outlines, such as swamps, can be generalized more roughly to suit a wider range of demanded scales. Layer order adjustments allow some features to be masked at larger scales, such as centerlines of area water features, so that processing needs are reduced. Line weights adjusted to a thinner set at smaller scales allow adjacent features to be retained without geometric adjustment as well.



a)



b)

Figure 5. Section of Pomme de Terre NHD subbasin (hilly-humid) showing a) original NHD streams (prepared for 1:24,000 display) and b) generalized 50K LoD streams. (Figures exported at 80K and printed at 90% for the paper.)

Conclusions

This paper reports research that extends work integrating feature generalization with map design for multi-scale topographic mapping in support of *The National Map* of the United States. We discuss preliminary results from generalizing four hydrographic subbasins of the USGS National Hydrography Dataset to demonstrate that knowledge about terrain and climate can inform decisions about generalization processing. We prepare generalization sequences with algorithm choices, sequences, and parameters which are specifically attuned to general landscape regimes. We also apply automated methods to reduce workloads for the NMA and to improve consistency and quality of results. We evaluate hydrographic generalizations in the context of fully-designed maps throughout the scale range of 20K to 1M, and produce generalized hydrography (LoDs) only at scales where the design solutions begin to fail. Future directions for our work incorporate four additional subbasins, automation of local density selection, and validation of the current LoD target scales (50K, 200K and 1M).

Acknowledgements

This ongoing research is funded by the U.S. Geological Survey's Center of Excellence for Geospatial Information Science (CEGIS) through Cooperative Ecosystem Studies Unit (CESU) grants: 09HQPA1000, 08HQPA0040, 08HQPA0039, 07HQPA0047, and 07HQPA0046. Thank you to Mamata Akella at Penn State, and Chris Anderson-Tarver and Jochen Wendel at Boulder, for research assistance; Charlie Frye (ESRI) for advising; Larry Stanislawski (USGS) for pruning code; and Tom Hale (USGS) for data production.

References

- Binder, S., and M. Cooley. 2008. E-Topo 7.5-Minute Maps, *The National Map Standards*, Version 0.0.4. Reston Virginia: U.S. Geological Survey. (Draft standards document: E-Topo Spec v.0.0.5 10-10-08.doc)
- Bobzien, M., D. Burghardt, I. Petzold, M. Neun, and R. Weibel. 2008. Multi-Representation Databases with Explicitly Modeled Horizontal, Vertical and Update Relations. *Cartography and GIS*, 35(1): 3-16.
- Brewer, C.A., and B.P. Buttenfield. 2009. Mastering Map Scale: Balancing Workloads Using Display and Geometry Change in Multi-scale Mapping. *Geoinformatica*, (in press). DOI 10.1007/s10707-009-0083-6.
- Brewer, C.A., and B.P. Buttenfield. 2007. Framing Guidelines for Multi-Scale Map Design Using Databases at Multiple Resolutions. *Cartography and GIS*, 34(1): 3-15.
- Brewer, C.A., B.P. Buttenfield, C. Frye, and J. Acosta. 2007. ScaleMaster: Multi-scale Mapmaking from Multiple Database Resolutions and for Multiple Map Purposes. Proceedings of 23rd International Cartographic Conference (ICC2007), Moscow Russia.

- Buttenfield, B.P. 2006. The Fallacy of the "Golden Feature" in MRDBs: Data Modeling Versus Integrating New Anchor Data. Proceedings of International Cartographic Association Symposium on Map Generalization, Vancouver Washington.
- Buttenfield, B.P. 1999. Progressive Transmission of Vector Data on the Internet: A Cartographic Solution. Proceedings of 19th ICA/ACI Conference, Ottawa Canada. pp. 581-590.
- Buttenfield, B. P., J. Wendel, and C. Frye. 2008. Simplifying Water Feature Details by Shape and Texture. North American Cartographic Information Conference, Missoula Montana, October (presentation).
- Mackaness, W.A., A. Ruas, and L.T. Sarjakoski (editors). 2007. *Generalisation of Geographic Information: Cartographic Modelling and Applications*. Amsterdam: Elsevier. 356 pp.
- Spaccapietra, S., C. Parent, and C. Vangenot. 2000. GIS Databases: From Multiscale to Multirepresentation. In: B.Y. Choueiry and T. Walsh (eds). SARA 2000, LNAI 1864. Berlin: Springer-Verlag. pp. 57-70.
- Stanislawski, L., M. Finn, E.L. Usery, and M. Barnes. 2007. Assessment of a Rapid Approach for Estimating Catchment Areas for Surface Drainage Lines. Proceedings of ACSM-IPLSA-MSPS 2007, St. Louis Missouri.
- Stansilawski, L.V., B.P. Buttenfield, M.P. Finn, and K. Roth. 2009. Stratified Database Pruning to Support Local Density Variations in Automated Generalization of the United States National Hydrography Dataset. Proceedings of 24th International Cartographic Conference (ICC2009), Santiago Chile.
- Stoter, J.E., M.J. Kraak, and R.A. Knippers. 2004. Generalization of Framework Data: A Research Agenda. Proceedings of the ICA Workshop on Generalisation and Multiple Representation. Leicester UK.
- Töpfer, F., and W. Pillewizer. 1966. The Principles of Selection, a Means of Cartographic Generalization. *The Cartographic Journal*, 3(1): 10-16.
- Wang, Z., and J-C. Muller. 1998. Line Generalization based on Analysis of Shape Characteristics. *Cartography and GIS*, 25(1): 3-15.
- Wendel, J., B.P. Buttenfield, C. Anderson-Tarver, and C.A. Brewer. 2009. Cartographic Generalization by Physiographic Region. Association of American Geographers Annual Meeting, Las Vegas Nevada, March (presentation).