Graph-based Analysis for Large-scale Hydrological Modelling

Lorne Leonard, *Member, IEEE*, Kamesh Madduri, Member, IEEE and Chris Duffy

Abstract—Graph based algorithms play an important role in large-scale hydrological modelling. This article explains why graphs are required for hydrology and outlines the spatial data scale to create models anywhere in the continental United States (CONUS) using heterogeneous national data products. We discuss two resolutions (scales) at which graphs are created. The first represents level-12 Hydrological Unit Code watersheds as nodes, to rapidly share essential terrestrial variable datasets using HydroTerre data workflows. Lessons learnt from creating graphs at this scale with heterogeneous datasets illustrate the data issues that cannot be addressed without high-resolution datasets and expert intervention. Expert intervention aided with visual analytical tools is necessary to address edge directions at the second graph scale: subdividing CONUS streams as edges (851,265,305) and nodes (683,298,991) for large-scale hydrological modelling.

Index Terms—Large graphs, Hydrological modelling, Visual Analytics, Data sharing, Distributed Systems, Spatial databases, GIS,

INTRODUCTION

HydroTerre (www.hydroterre.psu.edu) [1, 2] serves *Essential Terrestrial Variables* (ETVs) data for *hydrological modelling* anywhere in the *continental United States of America (CONUS)*, using distributed computing resources and web services. This data includes elevation, soils, land cover, and atmospheric forcing. This data requires hundreds of terabytes of disk storage, and is used as input for hydrological models. Our current cyber-infrastructure supports small watersheds and to scale up to CONUS scale watersheds, large graph based abstractions and algorithms are essential, both to serve data and to create large-scale hydrological models. The first section introduces the graph abstractions and the context in which they are used in hydrology and HydroTerre. We discuss how heterogeneous national datasets are used to create directed graphs. The second section discusses how we created graph abstractions for large-scale hydrological modelling at two spatial scales. At the first scale, we explain how graphs were created using watersheds as nodes. During this process, it became evident that expert intervention is essential to assign the correct direction to edges in the graph, as current national data products have missing or incomplete information. At the second scale, using stream segments as graph edges and nodes, both the graph and errors in edge directions are significantly larger than the first scale. With existing CONUS elevation data products, expert intervention is essential to verify millions of edge directions for large-scale watersheds. Our vision is to develop new visual analytic tools incorporated into the HydroTerre workflows [1, 2], to enable expert users to identify, reduce in scale, and correct edge directions. As well as incorporating changes automatically, by keeping the process data-driven, using geodatabases, so that the HydroTerre cyber-infrastructure shares improvements, creates versions, provenance, and reproducibility with its non-expert users.

1 CREATING GRAPHS FOR HYDROLOGY USING HETEROGENEOUS NATIONAL DATA PRODUCTS

HydroTerre is a heterogeneous distributed compute environment tuned to retrieve ETVs at a Hydrological Unit Code (HUC) spatial scale [1, 3]. The United States Geological Survey (USGS) partitions the land surface into watersheds, from *level-12* HUCs (the smallest unit, ~40,000 square meters) up to *level-2* HUCs (millions of square kilometres) [4,5,6]. A watershed represents an area of land constructed by elevation features, with all the surface water that passes a defined cross-section of a river or stream [7]. When web users visit HydroTerre, data workflows are performed via web services to query ETVs for an individual HUC-12. This paper focuses on combining HUC-12s for hydrological modelling anywhere in the CONUS. To do this requires large graphs at two scales. The ***first scale*** is to assign each HUC-12 as a graph node and scale from an individual HUC-12 to 10,000s of HUC-12s (Mississippi watershed) to access 100s of Terabytes of data. Currently, HydroTerre consists of 250 Terabytes of ETV data, and **Table 1** summarizes the essential data resources for hydrological modelling. **Table 2** quantifies the amount of data generated from one HUC-12 (~1 Gigabyte) up to the Mississippi scale (2.6 Terabytes). The ***second graph scale*** is using National Hydrograph Dataset (NHD) flow-lines or streams [8]. There are 29,560,501 named NHD v220 stream objects with hundreds of attributes required for hydrological modelling. However, when each stream object is divided into nodes (polyline vertex) and edges (polyline connections), the resultant CONUS graph has 683,298,991 unique nodes and 851,265,305 edges using the existing digitized geometry. As will be explained in section two, the size of the graph to address data issues will be larger dependent on other needs in hydrological modelling.

TABLE 1
The Essential Terrestrial Variables (ETV) that presently exists as national products for support of catchment models anywhere in the United States of America. We also indicate the data resolution used and the data disk size for databases (including processing) stored by the HydroTerre ETV Data Web Service.

Category	Variables	Nat. Products	Resolution
Atmospheric Forcing	precipitation, albedo, temperature , etc.	NLDAS II [9], NAAR [10]	8km, hourly, 5 TB per year
Digital Terrain	DEM, LiDAR	NED [11]	30m, 10TB
HUCs, Streams	Discharge etc.	NHD+ [8]	5 TB
Land Cover	Leaf area etc.	NLCD [12],	30m, 5TB
Soils	Sand, Silt etc.	SSURGO [13]	30m, 15TB

1.1 Graphs for data workflows

HydroTerre provides free access to data services via web applications to provide ETVs for an individual HUC-12. Data workflows are executed when a user selects a HUC-12 [1, 2]. Dependent on the model needs, data is often required for watershed boundary conditions, in other words, connecting graph edges. This creates a problem with our limited resources, as when a web user selects a

- *L. Leonard is with the Department of Civil & Environmental Engineering, The Pennsylvania State University, 406 Sackett Building, University Park, PA 16802, USA. E-mail: lnl3@ psu.edu.*
- *K.Madduri is with the Department of Computer Science and Engineering, The Pennsylvania State University, 0343E INFO SCI & TECH BL. University Park, PA 16802, USA. E-mail: madduri@cse.psu.edu.*
- *C.Duffy. is with the Department of Civil & Environmental Engineering, The Pennsylvania State University, 212 Sackett Building, University Park, PA 16802, USA. E-mail: cxd11@psu.edu.*

TABLE 2
ETV data requirements for modelling catchments at level 12 HUC scales. For example at the Juniata watershed, 10 GB of data is required as model inputs with 30 years of NLDAS [9] forcing data.

Hardware Scale	Watershed (Spatial scale)	Area (sq. km)	Number of level-12 HUC	ETV Data for 30 years (GB)
Servers	Shale Hills PA USA	0.2	<1	<1
Clusters	Juniata PA USA	8,800	146	10
	Susquehanna PA USA	71,215	1,038	150
Super Computer	Chesapeake Bay, USA	488,060	6,305	700
	Mississippi USA	3,789,449	40,425	2,600

HUC-12 node near the outlet of the Mississippi river, the entire upstream 40,425 nodes will be selected, resulting in 10s of Terabytes of ETV data (**Table 2**). Clearly, this occurs to all major river basins. Not only does this constrain HydroTerre HPC resources, the majority of end users cannot handle this amount of data for their modelling needs. Therefore, a graph can be used to not only select connecting nodes (HUC-12s), but also for trimming graph branches appropriate to both HPC limitations and user requirements. Our vision is to provide a data web service that scales appropriately to select and retrieve ETV data at any scale within the CONUS. To do this requires CONUS scale graphs as inputs to our data workflows [2, 3, 14].

1.2 Visual Analytic tools for expert intervention

There are issues with the NHD data products. One issue is the data structures have only one designated node [14, even though spatial analysis indicates 1000s of nodes with multiple edge connections. Furthermore, there are erroneous designated node connections and edge directions are not provided by NHD [14]. In hydrology, edge directions are important for determining the water flow direction of a stream. In practice, we would expect flow direction to be consistent in one direction, from higher (mountains) to lower elevations (oceans). For models such as the *Penn State Integrated Hydrological Model (PIHM)* [15, 16], when edge directions meet, the model solvers will either not converge, or be inefficient causing simulations to be magnitudes slower. Based on the analysis of executing workflows 1.5 million times with CONUS HUC-12s, data-model workflows (responsible for transforming ETV data to PIHM inputs) failed 26.91%, due to poor graphs [2]. The stream graph is essential for generating PIHM meshes, incorrect graphs account for 70.66% of model workflow failures within CONUS HUC-12s [2].

Evidently, HydroTerre, PIHM and other hydrological models require verified graphs at both HUC-12 node and flow-line (edge) scales. However, the entire CONUS graph requires intervention and verification. As we detail in the next section, our vision is to implement visual analytic tools to correct these graphs with expert intervention. Current heterogeneous national data products are not sufficient to correct NHD edges automatically. However, manual intervention is not feasible due to the magnitude of data issues and the HydroTerre ethos of data workflows that encourages reproducibility and provenance.

2 Discussion

Section 1 introduced two graph scales required for large-scale hydrological modelling within the CONUS using national data products. This section discusses our strategy towards developing software tools to create and verify these large graphs. Section 2.1 describes the process to create a CONUS HUC-12 node based graph. Lessons learnt from this scale made it evident the difficulty to address graphs using edges from CONUS NHD stream flow-line products. Section 2.2 demonstrates these issues at the Juniata and larger spatial watershed scales and our vision is to use visual analytic tools to guide expert users to not only verify the graphs, but also encourage users to share and correct graphs via HydroTerre for large-scale hydrological modelling anywhere in the CONUS.

2.1 Creating a graph with CONUS HUC-12s

The method to create a graph using CONUS HUC-12s from NHD geodatabases (v2.1) [8] started by developing a dictionary containing unique HUC-12 identifications (keys). By treating each HUC-12 geometry as a node, 83016 keys were identified. Checking the designation keys identified by NHD, it was discovered that 38 unique keys did not exist that were assigned to 221 HUC-12s [14]. With a relatively low number of HUC-12s, it appeared that manual intervention was feasible. However, by creating a list of neighbouring HUC-12s that touch the borders of a HUC-12, it was discovered that 1940 HUC-12s had designated keys (target) that did not touch the source [14]. Additionally, NHD datasets do not include multiple connecting HUC-12s. Using our strategy, not only *includes multiple edges*, but HUC-12s called closed basins were also selected. Closed basins are watersheds that do not have surface streams (edges) that connect to another watershed, but some models need to include these watersheds for groundwater and lake analysis.

To determine edges and direction between HUC-12 nodes, intersection points (derived dataset) were created at each HUC-12 watershed boundary and stream. Each point was then buffered at 15, 45, 60, 75, and 90-meter intervals, creating more points along the stream geometry. Using the National Elevation Dataset (NED) which has a 30-meter resolution, elevation values were assigned to these points with the goal to determine high and low elevation values. This technique was *not effective* due to inadequate elevation resolution, short stream reaches, lack of smooth stream reaches and/or watershed boundaries, and the large number of flat-sloped areas at these intersections. We assume this approach is feasible if we had access to LiDAR products that are currently not available at CONUS scale [17].

Instead, the chosen solution was to assume the digitized direction is from down-stream to up-stream. Using a buffer distance of 2 millimetres (a polygon enclosing the intersection point) to capture rapid change in stream direction, the intersection points had their position along the stream calculated; it was then possible to determine the edge direction [14]. Unfortunately, it is unclear and unlikely that all professionals involved with the stream flow-line generation used the same approach. However, using this strategy minimized the number of manual corrections required. **Figure 1** demonstrates the effectiveness of this strategy at the Juniata watershed and verified programmatically by checking for inner-hole topography within the watershed boundary. For further details with major river basins, we refer the reader to [9].

This graph is used to create a selection list of HUC-12 nodes in HydroTerre with Depth First Search (DFS) and Breadth First Search (BFS) algorithms when a user selects a HUC-12 to select all upstream datasets (**Figure 1**). Now it is possible to pre-determine the amount of input and output data for models such as PIHM before consuming any HPC resources. It has been demonstrated with a DFS graph that HydroTerre data bundles corresponding to 83.46% of the CONUS can be prepared and served under 4 minutes by reducing the amount of overlapping data inputs per individual HUC-12 [3].

This process, with a relatively small graph, demonstrates the need for new tools to simplify correcting graph edge connections when using spatial heterogeneous national data products. These new tools need to identify not only cases that require expert intervention, but provide data that scales up from the site, such as the intersection of a watershed boundary and stream outlet. Then to provide spatial context, the user has access to the entire stream (edge) length with elevation values. As we discuss in the next section, these issues become more difficult as we increase the graph size.

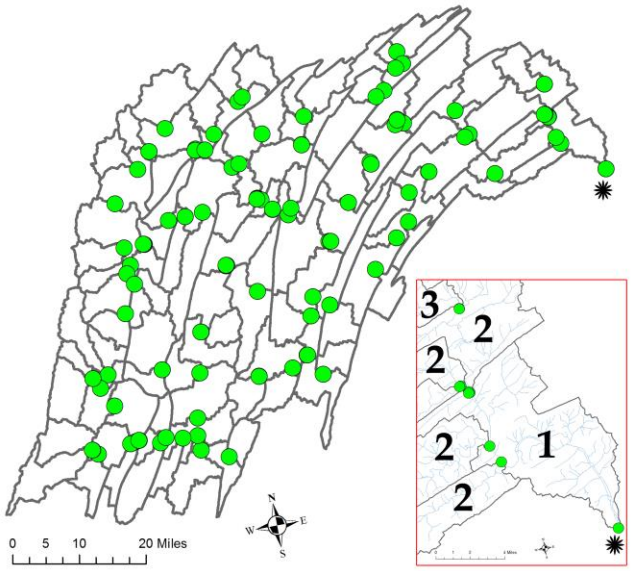

Fig. 1. (a) By selecting the HUC-12 node marked with an asterisk, using DFS or BFS algorithms, 148 HUC-12s are selected that form the Juniata watershed located in Pennsylvania USA. The green circles mark the locations where edges (streams) connect to adjacent HUC-12s. In Hydrology, these circles represent watershed outlets. (b) The insert illustrates the stream geometry (light blue) and the numbers denote the graph distance from the selected node (asterisk).

2.2 Subdividing CONUS streams as edges and nodes

The previous subsection focused on treating each HUC-12 as a node and investigating edges directly where streams overlap HUC-12s. Models such as PIHM also require the NHD stream geometry within each HUC-12. However, the stream geometry throughout the NHD CONUS national product are not consistent in direction. **Figure 2** illustrates the 15,255 edges (streams) within the Juniata catchment (**Table 3**). The arrowheads indicate the direction of the edge. Clearly, by visual inspection, determining which edge directions are correct or not is a time consuming and tedious task. When the edges are not consistent, the physic solvers used in PIHM will not converge, or the simulations will not produce any results. Thus, both HydroTerre and PIHM require strategies to identify and automatically fix the entire CONUS graph that consists of 851,265,305 edges when split into individual segments with existing digitized vertices and 683,298,991 nodes (**Table 3**).

Some of the most common and addressable issues with edge direction include *edges that meet* (identified as **A and B** in **Figure 2** insert), which can be solved using a simple graph traversal algorithm of searching for edges that have the same terminating node. Then we can reverse the incorrect edge by checking connecting edges, and validating the graph. Another example, are edges that have *no connections* to other edges, or are sub-graphs. *Difficult tasks* include edges that are topologically correct and appear valid from visual inspection (identified as **C** in **Figure 2**). However, when inspected with elevation data, these two edges actually need to be reversed. By including elevation attributes to all nodes will solve edges like this when topography differs significantly.

However, as discussed in Section 2.1, elevation resolution is not sufficient and with so many nodes sharing the same elevation value (same raster grid cell), determining edge direction requires traversing the graph both up and down from the edge of interest. This is problematic, as the graph up or down has yet to be validated but indicates that edge correction needs to start either from the root nodes (major river basin outlets) or from leaf nodes. For catchments at the Juniata scale, this amounts to 1000s of edits, and many magnitudes more with larger major river basin scales (**Table 3**). Therefore, new tools are required that are data-driven, but can be user-driven to reduce the analysis, scalable, and encourages navigation and searching of multilevel resolutions [17].

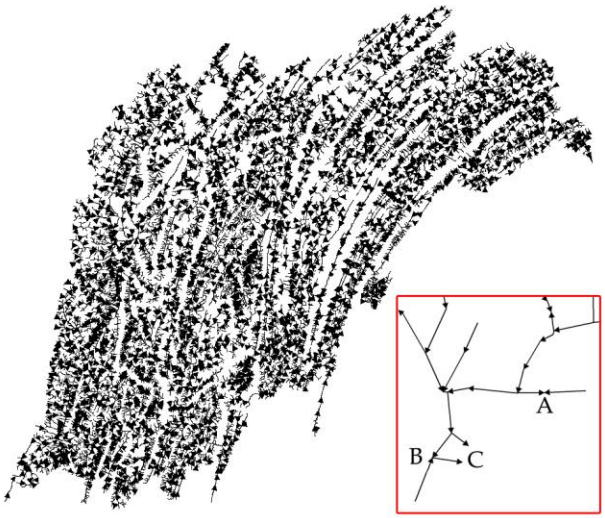

Fig. 2. (a) The Juniata watershed has 15,355 stream segments. The black lines represent the edge (stream) geometry. The arrowheads represent the edge direction (water flow direction). (b) The insert demonstrates edge directions that meet (A, B) and need to be reversed (C).

TABLE 3
Graph sizes for different watershed scales. The number of edges represents NHD v220 [8] digitized streams, while the number segments are these same streams split by digitized vertices. The number of nodes are the unique digitized vertices.

Watershed (Spatial scale)	Number of level-12 HUCs	Number of edges (Streams)	Number of edges split into segments by digitized vertices	Number of Unique Nodes
Juniata PA USA	146	15,255	310,940	310,616
Susquehanna PA USA	1,038	151,605	4,993,721	3,513,452
Chesapeake Bay, USA	6,305	751,910	16,740,974	9,644,155
Mississippi USA	40,425	12,950,292	309,308,615	446,693,961
CONUS	83,016	33,338,895	851,265,305	683,298,991

By creating a topologically correct directed graph using CONUS flow-lines will address the HUC-12 node graph discussed in section 2.1. However, determining edge direction for the CONUS graph has further implications for hydrological modelling. Models like PIHM, SWAT [19], and HEC-RAS [20] need stream cross-section geometry for stream flow analysis that are dependent on both existing nodes, but will require further insertion of vertices along edges. This could potentially increase the CONUS graph size by a factor of two to three. The directed graph is also necessary to assign missing soil attributes. PIHM requires sand, silt, clay, percentages that are not available for soil under streams. One approach is to average values adjacent to each edge. To do this, again requires more inserted vertices from the 220,641,540 SSURGO [13] soil polygons that exist within the CONUS national soil databases.

We have demonstrated the difficulty and large-scale of using heterogeneous national data products (NHD, NED, SSURGO) to create CONUS scale directional graphs for hydrology. Our vision is to develop visual analytical tools capable of handling spatial graphs with billions of edges and nodes for expert users to verify edge

directions (stream flow) without overwhelming users. By developing new tools that simplify these processes, are reproducible, and encourage provenance and sharing, we can achieve these goals [18, 21]. However, expert intervention is necessary to access edge directions as supporting datasets, in particular elevation (NED), do not have enough resolution to resolve direction. In addition, these national datasets are dynamic products with many improvements happening on a regular basis. Therefore, new tools that address graph edges will also need to be rapid, database driven, and be adaptable for future data products (such as LiDAR) as national data resolution increases. These tools need to help, and reduce edge quantities to the overwhelmed user, represent uncertainty (elevation) and encourage the user to collect and share supporting evidence. Otherwise, without accurate CONUS scale graphs, large-scale hydrology is not feasible for high-resolution (centimetre to 1 meter) watershed analysis.

3 Conclusion

Large graphs are essential to improve hydrological science. We demonstrated two scales of graphs using heterogeneous national datasets. The *first scale* discussed represented level-12 National Hydrological Datasets HUCs as nodes and focused on the connections between watersheds. At this graph scale, HydroTerre provides data bundles for hydrological modelling anywhere in the CONUS using depth and breadth first search algorithms. However, data assumptions were necessary about flow directions to minimize manual editing. Lessons learnt at this scale, demonstrated the need for new software tools to aid expert users to improve data from national datasets.

At the *second scale*, using CONUS NHD flow-lines (streams) as graph edges and nodes, it is not feasible without expert intervention with visual analytic tools to identify, simplify, and address flow directions. It was demonstrated at the Juniata watershed scale (8,800 square kilometres), a user is required to verify 310,940 edges due to elevation resolution not being sufficient using current national data products. The graph is clearly larger at the CONUS scale, with 851,265,305 edges and to create cross-section profiles and fill-in missing soil parameters, this graph will be significantly larger. Without addressing these CONUS scale graphs, it is more difficult to develop large-scale hydrological models with high-resolution national data products.

Acknowledgments

This research was supported in part by the National Science Foundation through XSEDE resources provided by the XSEDE Science Gateways program (TG-EAR120019), NSF EarthCube (GEO-44417482), NSF IN-SPIRE (IIS-1344272), EPA (96305901), NOAA (NA10OAR4310166), CC-NIE (1245980). The authors would like to acknowledge the support from the Institute for CyberScience Director Padma Raghavan and Penn State Institutes for Energy and the Environment Director Tom Richard at The Pennsylvania State University.

References

[1] Leonard, L., and Duffy, C. J. (2013). "Essential Terrestrial Variable data workflows for distributed water resources modeling." Environmental Modelling & Software, Elsevier Ltd, 50, 85–96.

[2] Leonard, L., and Duffy, C. J. (2014). "Automating data-model workflows at a level 12 HUC scale: Watershed modeling in a distributed computing environment." Environmental Modelling & Software, Elsevier Ltd, 61, 174–190.

[3] Leonard, L., Madduri, K., and Duffy, C. (2015). "Tuning Heterogeneous Computing Platforms for Large-scale Hydrology Data Management." IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, IN REVIEW

[4] Seaber, P. R., Kapinos, F. P., and Knapp, G. L. (1987). Hydrologic unit maps. U.S. G.P.O. ; For sale by the Books and Open-File Reports Section, U.S. Geological Survey, [Washington]; Denver, CO, 66.

[5] Simley, J. (2013). USGS National Hydrography Dataset Newsletter. 6.

[6] USGS. (2013). "USGS HUC." http://water.usgs.gov/GIS/huc.html.

[7] Dingman, S. L. (2002). *Physical hydrology*. Prentice Hall, Upper Saddle River, N.J.

[8] NHD. (2013). "USGS National Hydrography Dataset." http://nhd.usgs.gov/.

[9] NLDAS. (2011). "North American Land Data Assimilation System." http://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing.php.

[10] NARR. (2011). "North American Regional Reanalysis Homepage." http://www.emc.ncep.noaa.gov/mmb/rreanl/ (Dec. 5, 2010).

[11] USGS. (2011). "National Elevation Dataset." http://ned.usgs.gov/.

[12] NLCD. (2011). "National Land Cover Database, Multi-Resolution Land Characteristics Consortium." http://www.mrlc.gov/ (Oct. 7, 2012).

[13] SSURGO. (2011). "Soil Survey Geographic Database." http://soils.usda.gov/survey/geography/ssurgo (Aug. 1, 2010).

[14] Leonard L., Duffy C., 2014, "HydroTerre: Selecting Up-Stream Level-12 HUCS Using Depth-First Graphs Anywhere In The Continental USA", 11th International Conference on Hydroinformatics HIC 2014, New York City, USA.

[15] PIHM. (2014). "Penn State Integrated Hydrologic Model." http://www.pihm.psu.edu.

[16] Qu, Y., and Duffy, C. J. (2007). "A semidiscrete finite volume formulation for multiprocess watershed simulation." Water Resources Research, 43(8).

[17] USGS. (2013). "USGS Center for LIDAR Information Coordination and Knowledge." http://lidar.cr.usgs.gov/

[18] Wong, P. C., Shen, H.-W., Johnson, C. R., Chen, C., and Ross, R. B. (2012). "The Top 10 Challenges in Extreme-Scale Visual Analytics." Computer Graphics and Applications, IEEE.

[19] SWAT. (2013). "Soil and Water Assessment Tool." http://swat.tamu.edu/

[20] Brunner, G. (2010). "HEC-RAS river analysis system, Hydraulic reference manual, Version 4.1." US Army Corps of Engineers Hydrologic Engineering Center, Davis CA, (January), 1–790.

[21] Johnson, C. R., and Ross, R. B. (2007). Visualization and Knowledge Discovery: Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale.

Lorne Leonard received his PhD in Civil and Environmental Engineering and is a Research Programmer with Penn State Institutes for Energy and the Environment at The Pennsylvania State University. He received his Masters from The University of Illinois at Champaign-Urbana and undergraduate degrees from The University of Western Australia. His research interests include large-scale hydrology, high-performance scientific computing, visualization, and geographic information sciences.

Kamesh Madduri is an Assistant Professor in the Computer Science and Engineering department at The Pennsylvania State University. He received his PhD from the College of Computing at Georgia Institute of Technology, and his undergraduate degree from the Indian Institute of Technology Madras. His research interests include high-performance scientific computing, graph computations, and scientific data analysis.

Chis Duffy is a Professor in Civil and Environmental Engineering at The Pennsylvania State University. He received his PhD, Masters, and undergraduate degree from New Mexico Institute of Mining and Technology. His research interests include stochastic and numerical modelling of groundwater flow and solute transport, modelling large-scale hydrologic systems.