2011 Seventh IEEE International Conference on e-Science Workshops

Watershed Reanalysis

Towards a National Strategy for Model-Data Integration

Christopher Duffy, Lorne Leonard, Gopal Bhatt, Xuan Yu
Dept. of Civil & Environmental Engineering
Pennsylvania State University
University Park, PA, USA
cxd11@psu.edu

Lee Giles
College of Information Science and Technology
Pennsylvania State University
University Park, PA, USA
giles@ist.psu.edu

***Abstract*— Reanalysis or retrospective analysis is the process of re-analyzing and assimilating climate and weather observations with the current modeling context. Reanalysis is an objective, quantitative method of synthesizing all sources of information (historical and real-time observations) within a unified framework. In this context, we propose a prototype for automated and virtualized web services software using national data products for climate reanalysis, soils, geology, terrain and land cover for the purpose of water resource simulation, prediction, data assimilation, calibration and archival. The prototype for model-data integration focuses on creating tools for fast data storage from selected national databases, as well as the computational resources necessary for a dynamic, distributed watershed prediction anywhere in the continental US. In the future implementation of virtualized services will benefit from the development of a cloud cyber infrastructure as the prototype evolves to data and model intensive computation for continental scale water resource predictions.**

Keywords-component; Climate Reanalysis, Data Analytics, Distributed Hydrologic Model, Numerical Watershed Prediction (NWP), PIHM, Software as a Service (SaaS), Visual Analytics, Web Services

I. INTRODUCTION

"*The challenge of water scarcity (or flooding) has to be confronted watershed by watershed according to the local physical and political conditions…*" (Stephen Solomon, Water and Civilization). There is a clear national need to provide researchers, educators, resource managers and the general public seamless and fast access to essential geo-spatial/geo-temporal data, physics-based numerical models, and data fusion tools that are necessary to understand, predict and manage the nations surface and groundwater resources. The evaluation of ecosystem and watershed services such as the detection and attribution of the impact of climatic change on floods and drought provides one of many examples of the pressing need for high resolution, spatially explicit resource assessments. At the present time, there is no unified cyber infrastructure for supporting watershed models, and the data resource itself (weather/climate reanalysis products, stream flow, groundwater, soils, land cover, satellite data products, etc.) resides on many federal servers with limited access. It is clear that fast and efficient access to the data during model development, analysis and simulation remains the challenge. It is also important to state that computation for terrestrial watershed modeling is a data-intensive process, requiring extensive data libraries for climate, soils, geology land-use and land-cover, etc. Once acquired, each of these data sources must be processed before it is useful for constructing the physical watershed model and successive versions of the data and model are desirable. Model-data versioning would also include retrospective simulations, real time forecasting as well as future scenarios for climate and landuse change simulations [7].

Predicting the spatial and temporal distribution of water on complex landscapes begins with a multi-physics model for water and energy that couples surface and subsurface flows, with a community model for land surface moisture and energy fluxes [12]. In our research we have designed and developed the Penn State Integrated Hydrologic Model (PIHM). The hydrological processes in PIHM are fully coupled on a spatially-distributed unstructured grid. The unstructured grid and domain decomposition allows the user to construct quality numerical grids that can be constrained to follow or preserve important features of the model domain (e.g. watershed boundaries, soil, geology, political boundaries, etc.). Once the model domain is formed, the process of acquiring and projecting the geo-spatial and geo-temporal data on the model grid is perhaps the most time consuming and difficult process in model development. The web-based strategy described here focuses on implementation of the cyber infrastructure and workflow facilitating model prototyping through rapid data access, model input generation, model-data archival and versioning, and visualization of the results. In principal, the strategy discussed here would enable World Wide Web users to have seamless access to all necessary data products and to be able to carry out a simulation from archived data for any watershed or HUC (Hydrologic Unit Code) in the United States. Figure 1 illustrates the 1248 HUC-8 product for the USA as an example.

II. SCALE OF COMPUTATION

Beyond the problem of access to national data, the scale of computation for both data processing and model computational represents a major hurdle. This predicament is especially true since our application promotes a national approach to watershed prediction. All data resources must be processed before they are useful as parameters, inputs, boundary and

Research funded by: NSF EAR 0725019 Critical Zone Observatory, NOAA NA10OAR4310166 Role of Groundwater in land surface processes and European Commission SoilTrEC Project.

978-0-7695-4598-1/11 $26.00 © 2011 IEEE
DOI 10.1109/eScienceW.2011.32

Figure 1. *The 1248 HUC-8 watersheds in the USA. There are 103,444 HUC-12 watersheds in the USA with average area of 100-200 km^2.*

initial conditions for integrated hydrologic models. Dynamic data such as atmospheric variables interact with the terrestrial model in both space and time during the simulation hydrological model execution. This makes the model-data handling (communication) computationally expensive. To enable rapid prototyping of watershed models requires a transparent workflow while minimizing user intervention in low-level details, which can be achieved through a virtualization of data, modeling and web-based software services within an HPC environment. Ideally, the web service should allow the user to simply select the area of interest via a web-based application and efficiently compile the required data support for the modeling task. At present, there is no central data store or gateway for the range of variables needed for watershed simulation. Data exists across multiple federal and state data servers and retrieving data from these individual servers in real-time is generally not feasible for most users due to network connections, bandwidth restrictions, security, scheduled maintenance, etc.

Our strategy for watershed simulation utilizes virtualized web services where the user has access to standard tools for generating model input data from national data sets, initializing the model, running the code and carrying out data and visual analytics. The prototype we are developing will ultimately be deployed in a "cloud" environment as the prototype evolves to a national scale. The prototype utilizes time series services as well as geospatial data sets. Watershed reanalysis extends climate reanalysis to include new processes such as the role of groundwater and baseflow to streams (see Knowledge Products: NSF Critical Zone Observatory http://www.czo.psu.edu/data.html). From our point of view the scale of data processing is the limiting step to the success of national watershed reanalysis and archival. Table 1 provides some initial estimates on the data requirements for supporting a national watershed simulation archive. At this point we conservatively estimate that the problem will ultimately needs ~500 TB of fast storage and ~2000 CPU-hrs/year for data processing, model runs and info-vis for the 103,444 HUC-12 watersheds in the coterminous US (average-100km^2).

	Storage (TB)	CPU-HRS/ YR
National Data Products*(1)	60	748
Digital terrain models*(2)	20	170
Atmospheric forcing*(3)		
Reanalysis: NLDAS-2, NARR	5	100
Climate Scenario: IPCC	1	20
Soils (SSURGO)	3	<1
Land cover/use (NLCD 2001)	20	336
Landuse Scenario: (-)	5	100
Digital geology (-)	3	<1
Observations data (CUAHSI) *(4)	~2	20
Nat. Hydrogrsphy Data (NHD) *(5)	< 1	<1
National Data Processed *(6)	8	505
HUC 12 watershed/stream network	1	5
Watershed Climate data (NLDAS-2/)	2	450
Soil hydraulic properties (PTF)	< 1	20
Land cover param.'s (LAI, albedo)	< 1	30
Hydrogeologic properties (-)	< 1	-
Stream hydraulic geometry (-)	~2	-
Watershed Model Prototype	300	1.8E6
HUC-12 Model run: reanalysis 30yr	0.5GB/ HUC12	40/HU C12
HUC-12 Model run: IPCC scenario 30yr	0.5GB/ HUC12	40/HU C12
Model Code Versions *(6)	2GB/H UC12	80/ HUC12
US – 103,444 HUC-12's *(7)	300TB	1.8E6
Model-Data InfoVis HUC-12 Prototype	40	25
Space-time analysis products		
Reanalysis	0.2GB/ HUC12	
Scenario	0.2GB/ HUC12	
US– 103,444 HUC-12's *(7)	40	25
Totals for US	408	1.8E6

Table 1. Estimated storage (TB) and cpu-hrs/yr for data processing, model runs and info-vis for the 103,444 watersheds in the coterminous US. Notes for Table 1: *(1) National data includes downloading and maintaining digital terrain, soils, hydrogeology, NLCD land cover/use, hydroclimatic data from NLDAS-2, NARR, and MODIS satellite data. *(2) Digital terrain includes 1m-lidar, 3m, 10m, 30m, 90m DEM. *(3) National data 1979-present: precipitation, net radiation, wind speed/direction, soil moisture, etc. on a 4km grid. *(4) National historical streamflow, soil moisture, weather station point gauging data.*(5) National hydrographic data for watersheds, streams, channel geometry in GIS formats. *(6) Processing of all National Data Products to generate input and model parameter database for all 103,444 watersheds, To be used to run the watershed model PIHM. *(7) US National coverage. There are 103,444 HUC12 watersheds in the continental US.

III. MODEL-DATA INTEGRATION

The objective of the model-data integration framework is to provide watershed modeling tools and via a world wide web user interface that enables researchers, water managers and stakeholders the ability to perform complex workflows leading to efficient watershed model prototyping and simulation. These workflows are saved (versioned), in a database for online interrogation and visualization with access to online data mining and visual analytic tools. PIHM Web Services have been built as a web front-end middle-ware layer using Representational State Transfer, REST [21] and Simple Object Access Protocol, SOAP [30] protocols.

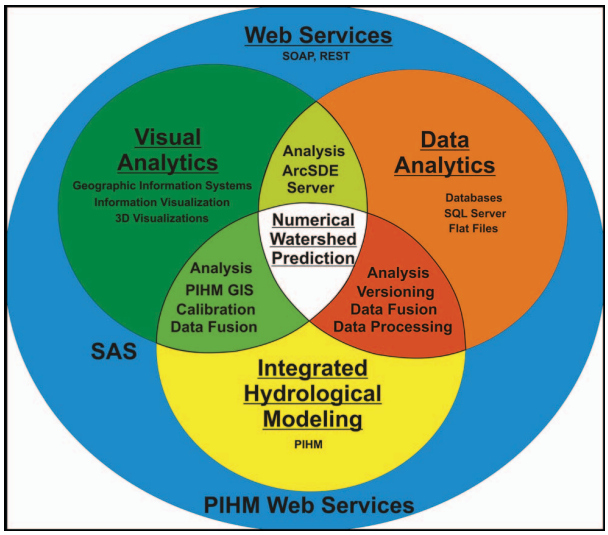

Figure 2: The elements of the model-data integration framework as a virtualized software web service.

These protocols enable sharing and analysis of data and model results within PIHM Web Services. Figures 2 illustrates the three elements of PIHM Web Services: 1) Data Analytics 2) Integrated Hydrologic Modeling, and 3), Visual Analytics. These are examined next.

IV. Watershed Web Services

The prototype web service requires the integration of GIS (geographical information systems) as an essential step in model building for watershed systems. It also supports visualization for users to interact with the PIHM model inputs, predictions and scenarios. The PIHM Web Services **Visual Analytics** element organizes the tools and techniques designed to synthesize information and derive insight from computationally large and expensive geo-spatial and geo-temporal datasets (e.g. multi-frequency satellite data, real-time environmental sensors, and the synthesis of ambiguous and conflicting national datasets such as soil, geology, land cover). These tools minimize the amount of time to interpret results, help detect conflicts, recommend strategies, and provide spatial/temporal scenarios for the assessment [33]. The **Visual Analytics** element and underlying water resources cyber infrastructure virtualizes the GIS and visualization techniques. *ArcGIS*™ server is our primary GIS system and is used to display both processed and simulation data sets from both flat files and databases. *GDAL* is used for raster manipulations of the raw national datasets and *TauDEM* [3] for processing lidar and digital elevation model formats. Three-dimensional rendering and visualizations are customized from *Silverlight* and *OpenGL*. Data analytics also includes multi-channel singular spectrum analysis, principal components, and uncertainty analysis.

The underlying database facilitates interaction between Data, Physical Model and Visual Analytics. Figure 3 shows the current implementation of the PIHM Web Services Workflow. Data Analytics provides access to several virtualized Data Management services such as raster processing, vector processing, domain decomposition etc. Visual Analytics includes virtualization of services related to space-time query, uncertainty analysis, and data assimilation. In general, each element of PIHM Web Services hides the unnecessary details of data handling, hydrologic model development and visualization from users by virtualizing tools that empower dynamic data discovery or resource assessment. The simplified process of data extraction and processing for PIHM has three basic steps: (1) watershed selection (2) State variable selection and (3) method of delivery. The complex workflow is executed, and the results are returned on-line as images (maps), data tables along with notifications for download access, meta-data, etc. Preliminary customized visualizations provide the ability to explore model simulation results and focus on analysis. In the next two sections, we demonstrate some of the features of the workflow and discuss a national application.

V. Application

The prototype web service initially focused on the NSF-funded watershed research testbed known as the Shale Hills-Susquehanna Critical Zone Observatory (CZO) in central PA. Watershed reanalysis at the site involved reprocessing and assimilation of historical observational data collected at various periods over a 40+ year span, into a fully coupled integrated hydrologic model. In the 1970's observations at the site consisted of a spatial array of 40 groundwater level sites measured daily, daily soil moisture records, and 10 minute streamflow records. The early data was used for empirical studies by forest hydrologists to resolve the role of antecedent moisture in peak flows within the forested watershed. Over the last 3 years the NSF Critical Zone Research effort has extended the early experimental research by deploying a real-time and spatially distributed embedded sensor network for soil moisture, soil temperature, soil conductance, groundwater levels, temperature, conductance, matric potential, snow depth. Shale Hills has a 30m tower with eddy covariance, net

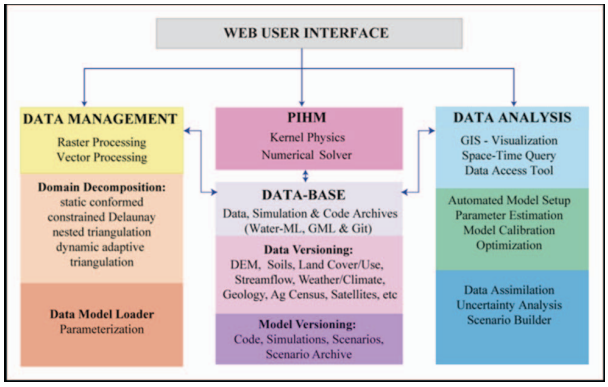

Figure 3: Current PIHM web services workflow and tools for model-data analysis and management [4,5].

radiation, infra-red surface temperature. Precipitation monitoring uses a disdrometer or laser precipitation monitor for

rain type and rain drop distribution, load cell gauge for accurate precipitation rates and a network of tipping bucket rain gauges (10 minute data).

The watershed model at the CZO testbed links atmospheric observations, land-surface vegetation, soil moisture and subsurface observations into a fully-coupled distributed system [1,2,8,9,10,11,20]. The model was calibrated using 2009 data and validated with 1974 soil moisture, groundwater levels and streamflow records. A 30 year reanalysis product of hourly land surface and subsurface states was carried out and posted on the CZO website http://www.czo.psu.edu/data.html. Figure 4. Illustrates the watershed numerical grid superimposed on high resolution digital terrain model from 1-m resolution lidar. Figure 5 illustrates observed versus predicted runoff for an artificial rainfall experiment carried out in the 1974 by forestry researchers [21]. Figure 6 illustrates the integrated water budget for the testbed based on 2009 data. We note that very few if any sites in the world have the kind of intensive data available at Shale Hills.

VI. SUMMARY

In this article, we have outlined out prototype for integrating models and national data focusing on numerical simulation anywhere in the US, with virtualized access to national datasets. The software infrastructure automatically manages data processing, and the simulation workflow uses a physics-based model, the Penn State Integrated Hydrologic Model (PIHM). The prototype for Web Services enables shared discovery of the impact of climate change at scales relevant to real impacts of flooding, drought etc. The approach establishes a framework to support numerical watershed prediction on a national basis . Some of the reasons why this is desirable include: To test our understanding of the terrestrial water cycle (including atmospheric, hydrological, bio-geochemical and energy balances) and especially determining the effects of feedback or amplification mechanisms with the ecological and climate components; To provide seamless and rapid access to historical and real-time data and data assimilation products for evaluating and predicting change and extending the impact of change on surface and groundwater resources; To develop future scenarios and model projections to access the impacts of climate and land use changes and the potential for extreme events such as floods, droughts and sea-level rise; To develop and implement access to real-time monitoring, data assimilation and water resource forecasting; To evaluate water supply and manage current and future water uses, demands and costs at all scales with explicit spatial identification of sources, sinks, and flows; To assess impacts on human and ecosystem services risk and vulnerability, water access, food production, food security, and sustainable development; To implement and monitor water policy across political boundaries.

Our vision is that the prototype must evolve as data and computational requirements grow nationally and globally in the future. This growth will demand a greater degree of virtualization such as those offered by cloud computing environments, and that can handle the predicted peta-scale computation described in this paper. The research attempts to address some of the hurdles involved with data-intensive computation, for physics based watershed models through web-based workflow and HPC resources. The concept of numerical watershed prediction on a national basis is long overdo.

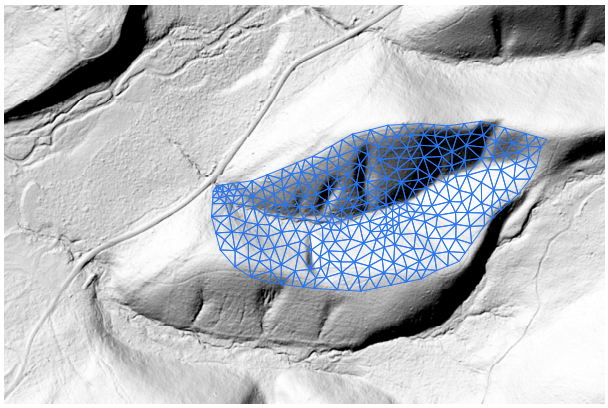

Figure 4. The 8 hectare Shale Hills CZO testbed showing lidar terrain and the numerical unstructured grid used to complete a 32 year reanalysis of the land-surface, soil moisture, groundwater and streamflwo at the site.

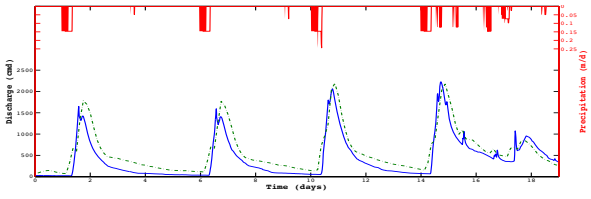

Figure 5. The simulated and observed runoff at the Shale Hills CZO testbed for controlled precipitation experiments in 1974.

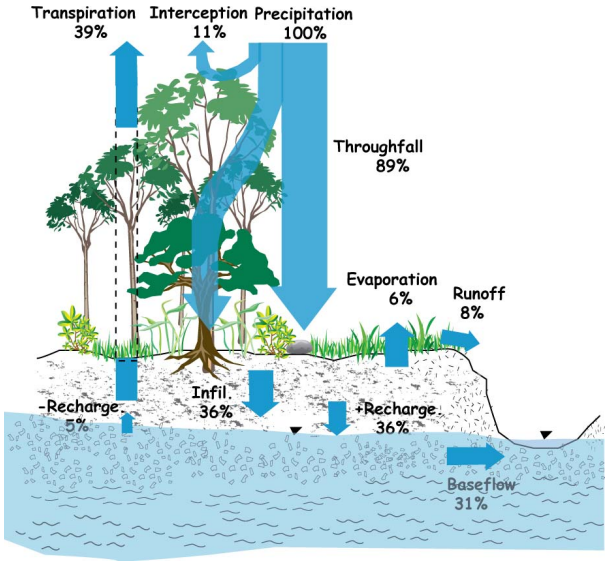

Figure 6. The simulated overall annual water budget for Shale Hills CZO testbed for 2009.

ACKNOWLEDGMENT

The authors would like to acknowledge the support from the Institute for CyberScience and Director Padma Ragavan at Penn State University, and the Research Computing Cyberinfrastructure group at Penn State led by Vijay Agarwala.

REFERENCES

[1] Bhatt, G., Kumar, M., Duffy, C.J., Weller, D., 2007. Using Integrated Hydrologic Models to Trace the Source and Dynamics of Fresh Water Discharge in a Coastal Watershed, American Geophysical Union, Fall Meeting

[2] Bhatt, G., Kumar, M., Duffy, C.J., 2008. Bridging the Gap between Geohydrologic Data and Distributed Hydrologic Modeling, International Environmental Modelling and Software Society (iEMSs)

[3] CZO, 2011. Susquehanna Shale Hills CZO. Retrieved May 13, 2011 from http://www.czo.psu.edu/

[4] ESRI, 2011. ArcGIS, Mapping Software and Data. Retrieved February 15, 2011 from http://www.esri.com/

[5] GDAL, 2011. Geospatial Data Abstraction Library. Retrieved August 1, 2010 from http://www.gdal.org/

[6] GDEM, 2011. Global Digital Elevation Map. Retrieved December 5, 2010 from http://asterweb.jpl.nasa.gov/gdem.asp

[7] IPCC, 2011. Intergovernmental Panel on Climate Change. Retrieved May 13, 2011 from http://www.ipcc.ch/

[8] Kumar, M., Duffy, C.J., 2007. Domain Partitioning for Implementation of Large Scale Integrated Hydrological Models on Parallel Processors. In Preparation.

[9] Kumar, M., Bhatt, G., Duffy, C.J., 2008. An efficient domain decomposition framework for accurate representation of geodata in distributed hydrologic models, International Journal of Geographical Information Science, IJGIS-2008-0069.R1

[10] Kumar, M., Duffy, C.J., 2010. An Object Oriented Shared Data Model for GIS and Distributed Hydrologic Models, IJGIS

[11] Kumar, M., Bhatt, G., Duffy, C.J., 2011. The Role of Physical, Numerical and Data Coupling in a Mesoscale Watershed Model (PIHM), In preparation.

[12] Mitchell, K.E., Lohmann, D., Houser P.R., Wood, E.F., Schaake, J.C., Robock, A., Cosgrove, B.A., Sheffield, J., Duan, Q., Luo, L., Higgins, R.W., Pinker, R.T., Tarpley, J.D., Lettenmaier, D.P., Marshall, C.H., Entin, J.K., Pan, M., Shi, W., Koren, V., Meng, J., Ramsay, B. H., and Bailey, A.A., 2004. The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system, J. Geophys. Res., 109, D07S90, doi:10.1029/2003JD003823.

[13] MODIS, 2011. Moderate Resolution Imaging Spectroradiometer, Retrieved December 5th, 2010 from http://modis.gsfc.nasa.gov/about/

[14] NARR, 2011. North American Regional Reanalysis Homepage. December 5th, 2010 from http://www.emc.ncep.noaa.gov/mmb/rreanl/

[15] NLCD, 2011. National Land Cover Database, Multi-Resolution Land Characteristics Consortium. Retrieved on Month XX, TODO from http://www.mrlc.gov/

[16] NLDAS, 2011. North American Land Data Assimilation System. December 5th, 2010 from http://ldas.gsfc.nasa.gov/nldas/NLDAS2forcing.php

[17] NOAA, 2011. National Oceanic and Atmospheric Administration. December 5th, 2010 from http://www.noaa.gov/

[18] OpenGL, 2011. The Industry's Foundation for High Performance Graphics. August 1st, 2010 from http://www.opengl.org/

[19] PIHM, 2011. Penn State Integrated Hydrological Model. August 1st, 2010 from http://www.pihm.psu.edu

[20] Qu, Y., Duffy, C.J., 2007. A semi-discrete finite volume formulation for multiprocess watershed simulation, Water Resource Res., 43, W08419, doi:10.1029/2006WR005752

[21] Richardson, L., Sam R., 2007. RESTful web service. O'Reilly Media.

[22] Silverlight, 2011. The Official Microsoft Silverlight. Retrieved August 1st, 2010 from http://www.silverlight.net/

[23] SQL Server , 2011. Microsoft SQL Server. Retrieved August 1st, 2010 from http://www.microsoft.com/sqlserver/en/us/default.aspx

[24] SSURGO, 2011. Soil Survey Geographic Database. Retrieved August 1st, 2010, from http://soils.usda.gov/survey/geography/ssurgo/

[25] STATSGO, 2011. State Soil Geographic Database. Retrieved August 1st, 2010, from http://soils.usda.gov/survey/geography/statsgo/

[26] Tarboton, D.G., 2011. TauDEM: Hydrology Research Group. Retrieved February 3rd 2011, from http://hydrology.usu.edu/taudem/taudem5.0/index.html

[27] Thomas J., Cook K., 2005. Illuminating the Path: The Research and Development Agenda for Visual Analytics. IEEE Press.

[28] USDA/NRCS, 2010. National Cartography & Geospatial Center: Geospatial Data Gateway. Retrieved Month XX, TODO from http://datagateway.nrcs.usda.gov/

[29] USGS, 2011. Seamless Data Warehouse. Retrieved August 1st, 2010, from http://seamless.usgs.gov/

[30] W3C, 2007. SOAP, Version 1.2 Part 1: Messaging Framework (Second Edition).

[31] Wosten, J.H.M., Lilly, A., Nemes A., Le Bas, C., 1998. Using existing soil data to derive hydraulic parameters for simulation models in environmental studies and in land use planning, Final Rep. Eur. Un., Wageningen.

[32] Yu, X., Leonard, L., Bhatt, G., Duffy, C.J., 2011. Shalehills Critical Zone Observation Repository. Retrieved March 10, 2011 from http://cataract.cee.psu.edu/czo/Reanalysis/