*International Environmental Modelling and Software Society (iEMSs)*
*7th Intl. Congress on Env. Modelling and Software, San Diego, CA, USA,*
*Daniel P. Ames, Nigel W.T. Quinn and Andrea E. Rizzoli (Eds.)*
*http://www.iemss.org/society/index.php/iemss-2014-proceedings*

# Cyberinfrastructure for Scalable Access to Stream Flow Analysis

**Tyler Wible[a], Wes Lloyd[ab], Olaf David[ab], and Mazdak Arabi[a]**

*[a]Dept. of Civil and Environmental Engineering, Colorado State University, Fort Collins, Colorado*
*(email address: tcwible@rams.colostate.edu)*
*[b]Dept. of Computer Science, Colorado State University, Fort Collins, Colorado;*

**Abstract:** Traditionally the various components of flow analysis including flooding, drought, base-flow, pollutant loading, and duration curves have been examined independently by various analysis methods or software packages. A better approach would be to combine these multiple packages into a single web-tool to improve access. Infrastructure-as-a-Service (IaaS) cloud provides a scalable infrastructure for model implementation, which is a necessity of web services due to the characteristics of web traffic. IaaS centralizes the computational burden and overhead of multiple model runs from local computers to online servers. This paper demonstrates the scalability benefits of the Comprehensive Flow Analysis (CFA) tool in an IaaS environment. The CFA tool is available through the Environmental Risk Assessment Management System (eRAMS) website. eRAMS facilitates GIS data manipulation, visualization, and preparation of input information for models lik CFA. eRAMS uses the Cloud Services Innovation Platform (CSIP) to request runs of the analyses within CFA. CSIP is an IaaS cloud modeling framework designed for executing various environmental models. This paper summarizes a scalability analysis of the analysis methods within CFA using CSIP in a cloud server environment.

**Keywords:** flow analysis; CFA; CSIP; cyberinfrastructure; eRAMS; scalability.

## 1. INTRODUCTION

Stream flow data has become an increasingly important tool for assessing current stream conditions as well as a basis for predicting future conditions and supply. However, there are many different aspects of flow analysis; floods, droughts, base-flow, frequency-duration, pollutant loading, and more. Previously each aspect of stream flow was approached independently resulting in multiple desktop software packages like BFLOW (Arnold et al., 1995, Arnold and Allen, 1999), HYSEP (Sloto and Crouse, 1996), LOADEST (Runkel et al., 2004), WHAT (Lim et al., 2005), and PEAKFQ (Flynn et al., 2006).

The problem with this style of software implementation is its burden on local computer resources and the fact that multiple software packages are necessary to analyse the entire scope of a topic like stream flow. A logical step would be to remove the local burden of computations and move the analysis to the web or cloud-based servers. Some have taken the first step and built web-based analysis tools like the hydrograph separation tool WHAT (Lim et al., 2005) and the web-accessible version of SPARROW, SPARROW-DSS (Booth et al., 2011). Furthering this trend towards web-software, Govindaraju et al. (2009) conceptualized the necessary cyberinfrastructure to support an end-to-end approach to environmental modelling.

The purpose of this work is to develop and demonstrate a scalable cloud-computing web-tool that facilitates access to and analysis of stream flow data. The specific objectives of this study are to unify the various stream flow analysis topics into a single tool, include access to national stream flow databases, and to demonstrate the cloud-based scalability of the unified flow analysis tool.

## 2. MODEL DEVELOPMENT

### 2.1. Comprehensive Flow Analysis (CFA) Overview

The CFA tool was developed by creating and integrating stream flow analysis methods for the various aspects of river flow. The combination of these multiple independent analyses into the same tool reduces the need to switch programs and re-format input data. CFA contains six flow analysis methods, as listed in Table 1. The use of many of these flow analysis methods is not new but the implementation and scalability benefits of the tool are.

**Table 1.** Comprehensive Flow Analysis Methods

| Method | Description |
|---|---|
| Time Series and Statistics | Graphical and statistical summary of daily, monthly, or annual stream flow data |
| Flood Analysis | Log-Pearson Type-III Regression fitted to existing data and analysed |
| Drought Analysis | Auto-Regressive or Auto-Regressive-Moving-Average Model fitted to existing data and analysed |
| Flow Duration Curve | Weibul Plotting Position Tied-Rank-Max recurrence intervals |
| Load Duration Curve | Weibul Plotting Position Tied-Rank-Max recurrence intervals combined with water quality data and target pollutant concentrations for a daily stream load |
| Base-flow Separation | Base-flow separation from stream flow using the BFLOW tool developed by Arnold et al. (1995; Arnold and Allen 1999) |
| Load Estimation | Nutrient Load estimation using the LOADEST tool developed by Runkel et al. (2004) |

The flood analysis in CFA exemplifies this implementation benefit to existing analysis processes. The flood analysis performs an automated Bulletin 17B Log-Pearson Type-III regression on available annual flood data, which is currently the standard practice for analysing floods on gauged U.S. rivers and streams (IACWD 1982, and WRC-HC 1967). However, the flood analysis outline by the Inter-Agency Committee on Water Data, the Bulletin 17B (1982), requires flood regressions to include the use of a regionalized skewness value, provided in the nation scale Plate I map of the documentation. Since Bulletin 17B's conception there have been a number of local-scale improvements to the Plate I map by state agencies (e.g. Parrett and Johnson, 2004, Soong et al., 2004, Cooper, 2005, Atkins et al., 2009, Olson, 2009, Pomeroy and Timpson, 2010). To assist in the access to this key map for flood analysis, the Plate I map and available state agency report maps were digitized and unified to create a base-layer for the flood analysis in CFA. Whenever a stream gauge station's flood data is analysed by CFA, on eRAMS (see below), the generalized regional skewness is automatically pulled from the compiled database for the station's location.

There are numerous ways to analyse droughts and because of this very few stream flow analysis tools have drought assessment capabilities, which is why it was decided that CFA would include a drought analysis. The approach chosen was that by Salas et al. (2005) in which an auto-regressive (AR) or auto-regressive-moving-Average (ARMA) model is automatically fitted to annualized natural stream flows and used to statistically simulate a larger dataset than the observed one. The new dataset retains the statistical properties of the observed dataset and contains a greater number of "droughts" of varying severity. This larger dataset is then used to determine average recurrence intervals of various drought magnitudes and lengths, a capability not in any existing flow analysis models.

### 2.2 Web Infrastructure

An important part of the novelty of the CFA tool is its cyberinfrastructure. Primary access to the CFA tool is available through on the Environmental Risk Assessment and Management System (eRAMS) website (www.erams.com/flowanalysis), which facilitates geospatial manipulation of data and access to environmental modelling. eRAMS utilizes a PostgreSQL database for spatial data management while the maps are generated using Mapserver (www.mapserver.org) and displayed using OpenLayers (www.openlayers.org). The map-style structure of eRAMS allows location based selection of stream flow monitoring points and their relevant information like drainage area, elevation, and generalized regional flood skewness. This location-based analysis also facilitates downstream-analyses; where a topic of interest is investigated at a site (e.g. pollutant loadings) and then the downstream sites are investigated by the same methodology and standards for the same topic of interest.

### 2.3 CSIP Integration

Web access to CFA on eRAMS is possible by the development and inclusion of the CFA tool into the Cloud Services Innovation Platform (CSIP). CSIP deploys a modelling engine using Eucalyptus 3.x private Infrastructure-as-a-Service Cloud (IaaS) using XEN 4.x for virtualization of multiple virtual machines (VMs) (Lloyd et al. 2012). Eucalyptus is used for its ability to provide an elastic compute cloud which enables management of launching, destroying, and modifying VMs. The original CSIP development was done on the OMS-Cloud which consists of eight physical servers 70 cores, 82 threads with 256GB of physical memory and 12TB of disk space. However, the CSIP services accessed through eRAMS are hosted on the eRAMS Cloud which consists of ten physical computers with a total of 80 physical cores and 160 hyper-threads with 720 GB of physical memory and 12TB of disk space. The scalable cloud infrastructure provided by CSIP provides a key requirement of the new unified flow analysis tool. CFA was implemented as a web-service project in CSIP similar to the RUSLE2 (David et al., 2013). An outline of the integration between eRAMS, CSIP, and CFA is shown below in Figure 1.

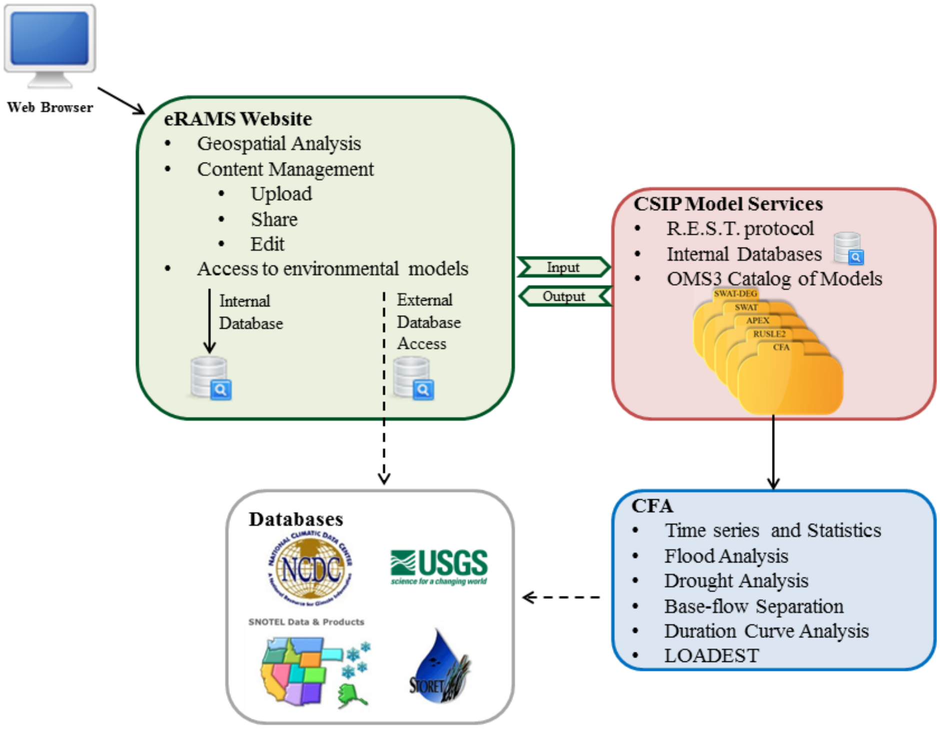

**Figure 1.** Diagram of eRAMS, CSIP, CFA Interaction

## 3. RESULTS AND DISCUSSION

### 3.1 Scalability

In order to test the resulting scalable aspects of the CFA tool, a thousand test cases were generated for each of the analyses in the JavaScript Object Notation (JSON) format used by CSIP. These test cases were then sent for execution to a cloud environment with a set infrastructure; the average execution time of the analyses based on the number of available VMs was then examined. All testing was performed using Eucalyptus clouds from Amazon with 2-core 3.25GHz ECUs.

Based on a preliminary examination of the analyses within CFA, the base-flow separation, load duration curve (LDC), and drought analysis were decided to be tested at a request rate of 2 requests per second while LOADEST was tested at 4 requests per second. Due to the simplicity of the flow duration curve (FDC), time series and statistics, and flood analyses it was decided to test them at rates of 10, 11, and 25 requests per second respectively. The different request rates mean that the models should not be compared to each other, but scalability trends for the tool as a whole can be determined from the results shown below in Figure 2.

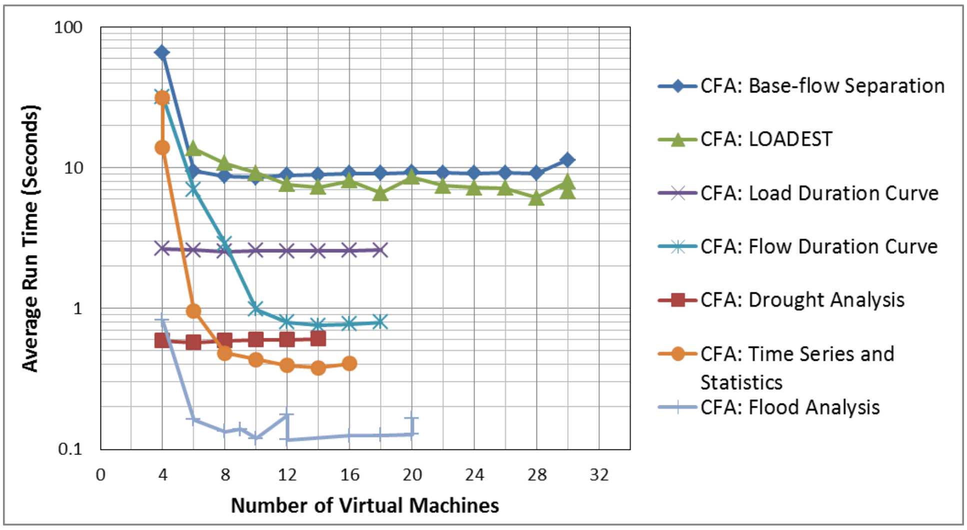

**Figure 2.** Load Testing Results 2-4 requests/s

It is evident from the testing that the LDC and drought analysis are simple and quick enough that there is very little change in average run time regardless of the available VMs. The irregular result in the flood analysis was found to be due to the fact that the analyses ran so quick, new VMs were unable to launch properly for the next testing cycle. The simplicity of the analyses and high request rates of the FDC and time series and statistics resulted in a nice minimization of execution time as more VMs were available. The base-flow separation and LOADEST analyses appeared to quickly reach a run time limit regardless of the number of VMs. This testing supports the claim of scalable infrastructure behind the CFA tool as well as illustrates this implementation's superior performance compared to local software which would require a grand total of 1000 tests times the average execution of the analysis amount of time to complete the same test cases.

### 3.2 Limitations and Future Prospects

A clear limitation of the testing is that some of these analyses; LDC, flood, and drought analyses, are very simple, requiring a large, possibly unrealistic request rates to identify the benefits of scalability with VMs. Therefore, the next step in the continued analysis of the scalability of the CFA tools is to determine realistic maximum request rates for each of the analyses. Even on a global scale website, is a rate of 25 requests per second even realistic? If not, perhaps the next step is to focus on the run time limit of the base-flow and LOADEST analyses rather than the poor results provided by the flood analysis testing. Another topic of future research would be to quantify the benefits of accessibility to the data and analyses of the CFA tool on eRAMS as opposed to the existing USGS NWIS and U.S. EPA STORET/WQX data retrievals and other available flow analysis software packages.

## 4. SUMMARY AND DISCUSSION

As water continues to be an important component in cities, agriculture, and industry, better understanding of the stream systems that are out there requires more tools and analyses techniques.

The existing method of creating a desktop flow analysis software package ignores the benefits of modern computing like cloud-infrastructure. As such, it has been demonstrated that combining many of the available flow analysis methods into a single tool implemented with cyberinfrastructure rooted in CSIP results in a more scalable tool which can additionally access existing national stream flow databases.

## 5. REFERENCES

Arnold, J.G., Allen, P.M., Muttiah, R., Bernhardt, G., 1995. Automated base flow separation and recession analysis techniques. Ground Water, 33(6): 1010-1010.

Arnold, J.G., Allen, P.M., 1999. Automated methods for estimating baseflow and ground water recharge from streamflow records. Journal of the American Water Resources Association 35(2): 411-424.

Atkins, Jr., J.T., Wiley, J.B., Paybins, K.S., 2009. Generalized Skew Coefficients of Annual Peak Flows for Rural, Unregulated Streams in West Virginia. U.S. Geological Survey: Open-File Report: 2008-1304.

Booth, N.L., Everman E.J., Kuo, I. Sprague, L., Murphy, L., 2011. A Web-based Decision Support System for Assessing Regional Water-Quality Conditions and Management Actions. Journal of the American Water Resources Association 47(5): 1136-1150.

Cooper, R.M., 2005. Estimation of Peak Discharges for Rural, Unregulated Streams in Western Oregon. U.S. Geological Survey: Scientific Investigations Report 2005-5116.

David, O., Ascough II, J.C., Lloyd, W., Green, T.R., Rojas, K.W., Leavesley, G.H., Ahuja, L.R., 2013. A software engineering perspective on environmental modelling framework design: The Object Modeling System. Environmental Modelling & Software 39: 201-213.

Flynn, K.M., Kirby, W.H., Hummel, P.R., 2006. User's Manual for Program PeakFQ, Annual Flood-Frequency Analysis Using Bulletin 17B Guidelines. U.S. Geological Survey: Techniques and Methods 4-B4.

Gobindaraju, R.S., Engel, B., Ebert, D., Fossum, B., Huber M., Jafvert, C., Kumar, S., Merwade, V., Niyogi, D., Oliver, L., Prabhakar, S., Rochon, G., Song, C., Zhao, L., 2009. Vision of Cyberinfrastructure for End-toEnd Environmental Exploration (C4E4). Journal of Hydrologic Engineering American Society of Civil Engineers, 12, 1:56-64.

Interagency Advisory Committee on Water Data (IACWD). 1982. "Guidelines for determining flood flow frequency." Bulletin No. 17B (revised and corrected), Hydrology Subcommittee, Washington, D.C.

Lim K.J., Engel, B.A., Tang, Z. Choi, J., Kim, K., Muthukrishnan, S., Tripathy, D., 2005, Automated Web GIS Based Hydrograph Analysis Tool, WHAT. Journal of the American Water Resources Association 41(6): 1407-1416.

Lloyd, W., David, O., Lyon, J., Rojas, K.W., Ascough II, J.C., Green, T.R., Carlson, J.R., 2012. The Cloud Services Innovation Platform Enabling Service-Based Environmental Modelling Using Infrastructure-as-a-Service Cloud Computing. International Environmental Modelling and Software Society 2012 Conference Proceedings.

Olson, S.A., 2009. Estimation of Flood Discharges at Selected Recurrence Intervals for Streams in New Hampshire. U.S. Geological Survey: Scientific Investigations Report 2008-5206.

Parrett, C., Johnson, D.R., 2004. Methods for Estimating Flood Frequency in Montana Based on Data through Water Year 1998. U.S. Geological Survey: Water-Resources Investigations Report 03-4308.

Pomeroy, C.A., Timpson, A.J., 2010. Methods for Estimating Magnitude and Frequency of Peak Flows for Small Watersheds in Utah. U.S. Geological Survey: Report No UT-10.11.

Runkel, R.L., Crawford, C.G., Cohn., T.A., 2004. Chapter A5: Load Estimator (LOADEST): A FORTRAN Program for Estimating Constituent Loads in Streams and Rivers. U.S. Geological Survey Techniques and Methods Book 4.

Salas, J.D., Fu, C., Cancelliere, A., Dustin, D., Bode, D., Pineda, A., Vincent, E., 2005. Characterizing the Severity and Risk of Drought in the Poudre River, Colorado. Journal of Water Resources Planning and Management 131(5): 383-393.

Sloto, R.A, Crouse, M.Y. 1996. HYSEP: A Computer Program for Streamflow Hydrograph Separation and Analysis. U.S. Geological Survey: Water-Resources Investigations Report 96-4040.

Soong, D.T., Ishii, A.L., Sharpe, J.B., Avery, C.F., 2004. Estimating Flood-Peak Discharge Magnitudes and Frequencies for Rural Streams in Illinois. U.S. Geological Survey: Scientific Investigations Report 2004-5103.

Water Resources Council, Hydrology Committee (WRC-HC). 1967. "A Uniform Technique for Determining Flood Flow Frequencies." Bulletin No. 15, Washington, D.C.